

# Open Database Searching Enables the Identification and Comparison of Bacterial Glycoproteomes without Defining Glycan Compositions Prior to Searching

## Authors

Ameera Raudah Ahmad Izaham and Nichollas E. Scott

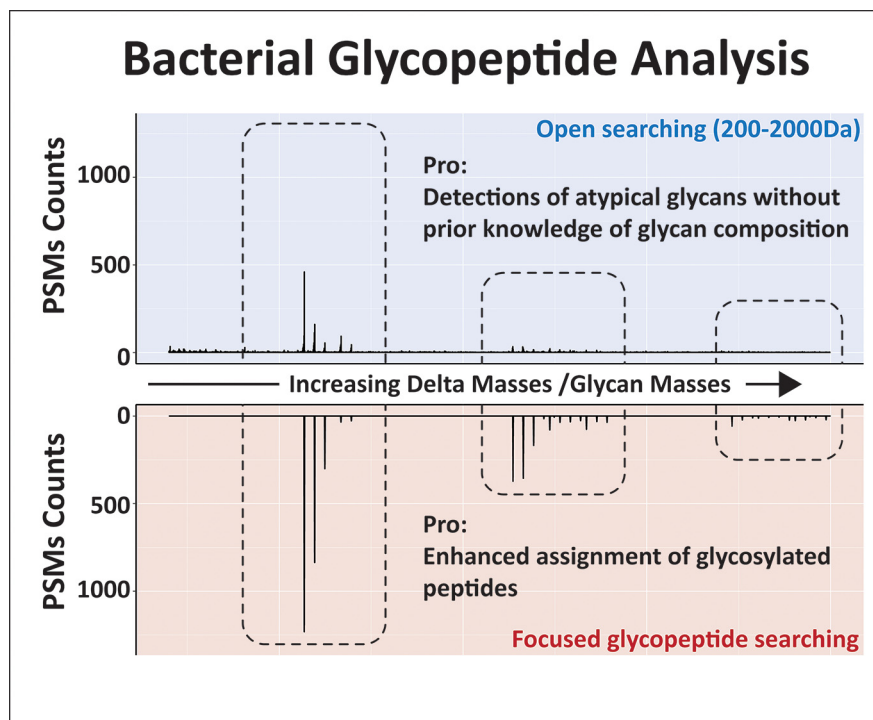
## Correspondence

Nichollas.scott@unimelb.edu.au

## In Brief

We assess the ability of open searching (OS) to analyze bacterial glycoproteomes. We find OS readily identifies known and unknown glycans within samples. Using OS, we compared glycan use across eight *Burkholderia* species confirming the conservation of glycan use and highlight how low frequency glycans can be identified in OS data sets. Importantly, OS is not without caveats; leading to lower assignment scores and under reporting multiply glycosylated peptides. This supports the value of OS followed by focused searching approaches.

## Graphical Abstract



## Highlights

- We demonstrate the use of open searching for the detection of atypical glycosylation in bacterial glycoproteomes.
- We show that although open searching is able to detect unique glycoforms it is less sensitive for the detection of multiply glycosylated peptides than focused searches.
- Using open search delta mass profiles, we demonstrate glycan use across bacterial glycoproteomes can be easily compared.
- For the *Burkholderia* genus we confirm the use of similar glycans for protein glycosylation yet also highlight that species specific glycans do exist.



# Open Database Searching Enables the Identification and Comparison of Bacterial Glycoproteomes without Defining Glycan Compositions Prior to Searching

Ameera Raudah Ahmad Izaham and Nichollas E. Scott\*<sup>1</sup>

Mass spectrometry has become an indispensable tool for the characterization of glycosylation across biological systems. Our ability to generate rich fragmentation of glycopeptides has dramatically improved over the last decade yet our informatic approaches still lag behind. Although glycoproteomic informatics approaches using glycan databases have attracted considerable attention, database independent approaches have not. This has significantly limited high throughput studies of unusual or atypical glycosylation events such as those observed in bacteria. As such, computational approaches to examine bacterial glycosylation and identify chemically diverse glycans are desperately needed. Here we describe the use of wide-tolerance (up to 2000 Da) open searching as a means to rapidly examine bacterial glycoproteomes. We benchmarked this approach using *N*-linked glycopeptides of *Campylobacter fetus subsp. fetus* as well as *O*-linked glycopeptides of *Acinetobacter baumannii* and *Burkholderia cenocepacia* revealing glycopeptides modified with a range of glycans can be readily identified without defining the glycan masses before database searching. Using this approach, we demonstrate how wide tolerance searching can be used to compare glycan use across bacterial species by examining the glycoproteomes of eight *Burkholderia* species (*B. pseudomallei*; *B. multivorans*; *B. dolosa*; *B. humptydoensis*; *B. ubonensis*, *B. anthina*; *B. diffusa*; *B. pseudomultivorans*). Finally, we demonstrate how open searching enables the identification of low frequency glycoforms based on shared modified peptide sequences. Combined, these results show that open searching is a robust computational approach for the determination of glycan diversity within bacterial proteomes.

Protein glycosylation, the addition of carbohydrates to proteins, is a widespread and heterogeneous class of protein modifications (1–3). Within Eukaryotes, multiple glycosylation systems have been identified (1–3) and up to 20% of the pro-

teome is thought to be subjected to this class of modifications (4). Within Eukaryotes, both *N*-linked and *O*-linked glycosylation systems are known to generate highly heterogeneous glycan structures (2, 3) with this glycan heterogeneity important for the function of glycoproteins (5, 6). Although the glycan repertoire used in eukaryotic systems is thought to be large, the diversity within any given biological sample is constrained by the limited number of monosaccharides used in eukaryotic systems (7), as well as the expression of proteins required for the construction of glycans such as glycosyltransferases (8). Experimentally, these constraints lead to only a limited number of glycans being produced across eukaryotic samples (9, 10) despite the large number of potential glycan structures (11, 12). This limited diversity within both eukaryotic *N*-linked and *O*-linked glycans has enabled the development of glycan databases which have facilitated high throughput glycoproteomic studies (13) using tools such as Byonic (14) and pGlyco (15). Unfortunately, these databases are not suitable for all glycosylation systems and fail to identify glycopeptides modified with novel or atypical glycans such as those found in bacterial glycosylation systems.

Within bacterial systems, glycosylation is increasingly recognized as a common modification (16–19). Although glycosylation in bacteria was first identified in the 1970s (20), it is only within the last two decades that it has become clear that this class of modifications is ubiquitous across bacterial genera (16, 18, 21). Unlike eukaryotic systems, which use a relatively small set of monosaccharides, bacterial glycoproteins are decorated with a diverse range of monosaccharides (22) leading to a staggering array of glycan structures (23–32). This glycan diversity represents a significant challenge to the field as it makes the identification of novel bacterial glycoproteins a nontrivial analytical undertaking. Yet, through advancements in mass spectrometry (MS) (28, 30, 33, 34),

From the Department of Microbiology and Immunology, University of Melbourne at the Peter Doherty Institute for Infection and Immunity, Melbourne, Australia

This article contains [supplemental data](#).

\* For correspondence: Nichollas E. Scott, [Nichollas.scott@unimelb.edu.au](mailto:Nichollas.scott@unimelb.edu.au).

This is an open access article under the [CC BY](#) license.

these once obscure modifications are increasingly recognizable and are now known to be essential for bacterial fitness (26, 35–38). Despite our ability to generate rich bacterial glycopeptide data the field still largely uses manual interrogation to identify and characterize novel glycosylation systems (23–32). This dependence on manual interrogation is not scalable, time-consuming and prone to human error, especially in the detection of glycoform heterogeneity. This is exemplified in our own experience characterizing glycosylation in *Acinetobacter baumannii* where our initial analysis overlooked alternative methylated and deacetylated forms of glucuronic acid (26). Thus, new approaches are needed to ensure bacterial glycosylation studies can be undertaken in a robust and high-throughput manner.

Wide precursor mass tolerance database searching, also known as “open” or wildcard searching, is an increasingly popular approach for the detection of protein modifications within proteomic data sets (39–43). The underlying premise of this approach is that by allowing a wide precursor mass tolerance, modified peptides can be detected by the difference in their observed mass from their expected mass. Importantly, this makes the identification of modifications independent of needing to define the modification in the initial search parameters. This approach has been used to examine chemical modifications such as formylation (44) and mis-alkylation (45) as well as large modifications such as DNA-peptide cross-links (43) and eukaryotic glycosylation (46, 47). Although this approach is effective, it is not without tradeoffs being computationally more expensive than traditional searches leading to longer search times (48). To date, these searches have typically been undertaken using  $\pm 500$  Da tolerances (39–43) yet large delta mass windows of  $\pm 1000$  Da (43, 46) and even  $+3000$  Da (46, 47) have been reported. Despite the growing application of open database searching in eukaryotic proteomics, few bacterial studies have used this technique. This said, alternative strategies such as dependent peptide searching have been used in bacteria to track misincorporation of amino acids (49) and identify novel forms of glycosylation such as arginine-rhamnosylation (50).

In this study, we demonstrate that wide mass (up to 2000 Da) open database searching enables the rapid identification of bacterial glycopeptides without the need to assign glycan masses before database searching. Using Byonic™, which enables both glycopeptide and open database searching (14, 51), we benchmark wide mass open searching on three previously characterized bacterial glycosylation systems: the N-linked glycosylation system of *Campylobacter fetus subsp. fetus* NCTC10842 (25); the O-linked glycosylation system of *Acinetobacter baumannii* ATCC17978 (26, 52, 53); and the O-linked glycosylation system of *Burkholderia cenocepacia* J2315 (23, 37). Each of these bacteria have increasingly complex proteomes (ranging from 1600 to nearly 7000 proteins) enabling us to assess the performance of open

database searching across a range of proteome sizes. We find open database searching readily enabled previously characterized glycoforms and microheterogeneity to be identified across all samples. Applying this approach to representative species of the *Burkholderia* genus (23, 37), we provide the first snapshot of glycosylation across this genus. Consistent with the conservation of the biosynthetic pathway responsible for the *Burkholderia* O-linked glycans (23) all *Burkholderia* species examined predominately modify their glycoproteins with two glycan structures of similar composition. Excitingly, we demonstrate that open searching also enables low frequency glycoforms to be detected, highlighting that species-specific glycan structures do exist in *Burkholderia*. Thus, open database searching enables the identification of diverse bacterial glycan structures in a high-throughput manner.

#### EXPERIMENTAL PROCEDURES

**Bacterial Strains and Growth Conditions**—*C. fetus subsp. fetus* NCTC 10842 was grown on Brain-Heart Infusion medium (Hardy Diagnostics) with 5% defibrinated horse blood (Hemostat, Dixon, CA) under microaerobic conditions (10% CO<sub>2</sub>, 5% O<sub>2</sub>, 85% N<sub>2</sub>) at 37 °C as previously reported (25). *Burkholderia pseudomallei* K96243 was grown as previously reported (54) in Luria Bertani (LB) broth. All other bacterial strains were grown overnight on LB agar at 37 °C as previously described (37). Details on the strains, their origins, references and proteome databases used in this study are provided within Table I.

**Generation of Bacterial Lysates for Glycoproteome Analysis**—Bacterial strains were grown to confluency on agar plates before being flooded with 5 ml of pre-chilled sterile PBS (PBS) and bacterial cells collected by scraping. Cells were washed 3 times in PBS to remove media contaminants, then collected by centrifugation at  $10,000 \times g$  at 4°C for 10 min and then snap frozen. Snap frozen cell samples were resuspended in 4% SDS, 100 mM Tris pH 8.0, 20 mM Dithiothreitol (DTT) and boiled at 95°C with shaking at 2000 rpm for 10 min. Samples were clarified by centrifugation at  $17,000 \times g$  for 10 min, the supernatants then collected, and protein concentration determined by a bicinchoninic acid assay (Thermo Fisher Scientific, Waltham, MA). 1 mg of protein from each sample was acetone precipitated by mixing one volume of sample with 4 volumes of ice-cold acetone. Samples were precipitated overnight at -20°C and then spun down at  $16,000 \times g$  for 10 min at 0°C. The precipitated protein pellets were resuspended in 80% ice-cold acetone and precipitated for an additional 4 h at -20°C. Samples were centrifuged at  $17,000 \times g$  for 10 min at 0°C, the supernatant discarded, and excess acetone driven off at 65°C for 5 min. Three biological replicates of each bacterial strain were prepared.

**Digestion of Protein Samples**—Protein digestion was undertaken as previously described with minor alterations (28). Briefly, dried protein pellets were resuspended in 6 M urea, 2 M thiourea in 40 mM NH<sub>4</sub>HCO<sub>3</sub> then reduced for 1 h with 20 mM DTT followed by being alkylated with 40 mM chloroacetamide for 1 h in the dark. Samples were then digested with Lys-C (1/200 w/w) for 3 h before being diluted with 5 volumes of 40 mM NH<sub>4</sub>HCO<sub>3</sub> and digested overnight with trypsin (1/50 w/w). Digested samples were acidified to a final concentration of 0.5% formic acid and desalted with 50 mg tC18 Sep-Pak columns (Waters corporation, Milford, MA) according to the manufacturer's instructions. tC18 Sep-Pak columns were conditioned with 10 bed volumes of Buffer B (0.1% formic acid, 80%

TABLE I  
Strain list

Strains	Source (description, country, year)	Reference	Proteome database
Campylobacter fetus subsp. fetus NCTC 10842	Brain of sheep fetus, France, 1952	(79)	Uniprot Database: UP000001035
Acinetobacter baumannii ATCC17978	Fatal meningitis of a 4-month old infant, 1951	(80)	GenBank assembly accession: GCA_001593425.2
Burkholderia pseudomallei K96243	Human clinical specimen, Thailand, 1996	(81)	Uniprot Database: UP000000605
Burkholderia cenocepacia (LMG 16656/J2315)	Human clinical specimen, United Kingdom, 1989	(82)	Uniprot Database: UP000001035
Burkholderia multivorans MSMB2008	Soil isolate, Australia, 2012	(83)	Burkholderia Genome Database (84), Strain number: 3016
Burkholderia dolosa AUJ0158	Human clinical specimen, USA unknown	(85)	Uniprot database: UP0000032886
Burkholderia humptydoensis MSMB43	Water isolate, Australia, unknown	(83, 86)	Burkholderia Genome Database (84), Strain number: 4072
Burkholderia ubonensis MSMB22	Soil isolate, Australia, 2001	(85)	Burkholderia Genome Database (84), Strain number: 3404
Burkholderia anthina MSMB649	Soil isolate, Australia, 2010	(83)	Burkholderia Genome Database (84), Strain number: 2849
Burkholderia diffusa MSMB375	Water isolate, Australia, 2008	(83)	Burkholderia Genome Database (84), Strain number: 2966
Burkholderia pseudomultivorans MSMB2199	Soil isolate, Australia, 2011	(83)	Burkholderia Genome Database (84), Strain number: 3251

acetonitrile), then equilibrated with 10 bed volumes of Buffer A\* (0.1% TFA, 2% acetonitrile) before use. Samples were loaded on to equilibrated columns then columns washed with at least 10 bed volumes of Buffer A\* before bound peptides were eluted with Buffer B. Eluted peptides were dried by vacuum centrifugation and stored at  $-20^{\circ}\text{C}$ .

**ZIC-HILIC Enrichment of Bacterial Glycopeptides**—ZIC-HILIC enrichment was performed as previously described with minor modifications (28). ZIC-HILIC Stage-tips (55) were created by packing 0.5 cm of 10  $\mu\text{M}$  ZIC-HILIC resin (Millipore, Burlington, MA) into p200 tips containing a frit of C8 Empore<sup>TM</sup> (Sigma, St. Louis, MO) material. Before use, the columns were washed with ultrapure water, followed by 95% acetonitrile and then equilibrated with 80% acetonitrile and 5% formic acid. Digested proteome samples were resuspended in 80% acetonitrile and 5% formic acid. Samples were adjusted to a concentration of 3  $\mu\text{g}/\mu\text{L}$  (a total of 300  $\mu\text{g}$  of peptide used for each enrichment) then loaded onto equilibrated ZIC-HILIC columns. ZIC-HILIC columns were washed with 20 bed volumes of 80% acetonitrile, 5% formic acid to remove nonglycosylated peptides and bound peptides eluted with 10 bed volumes of ultrapure water. Eluted peptides were dried by vacuum centrifugation and stored at  $-20^{\circ}\text{C}$ .

**Reverse Phase LC-MS**—ZIC-HILIC enriched samples were resuspended in Buffer A\* and separated using a two-column chromatography set up composed of a PepMap100 C18 20 mm  $\times$  75  $\mu\text{m}$  trap and a PepMap C18 500 mm  $\times$  75  $\mu\text{m}$  analytical column (Thermo Fisher Scientific). Samples were concentrated onto the trap column at 5  $\mu\text{L}/\text{min}$  for 5 min with Buffer A (0.1% formic acid, 2% DMSO) and then infused into an Orbitrap Fusion<sup>TM</sup> Lumos<sup>TM</sup> Tribrid<sup>TM</sup> Mass Spectrometer (Thermo Fisher Scientific) at 300 nl/minute via the analytical column using a Dionex Ultimate 3000 UPLC (Thermo Fisher Scientific). 185-min analytical runs were undertaken by altering the buffer composition from 2% Buffer B (0.1% formic acid, 77.9% acetonitrile, 2% DMSO) to 28% B over 150 min, then from 28% B to 40% B over 10 min, then from 40% B to 100% B over 2 min. The composition was held at 100% B for 3 min, and then dropped to 2% B over 5 min before being held at 2% B for another 15 min. The Lumos<sup>TM</sup> Mass Spectrometer was operated in a data-dependent mode automatically switching between the acquisition of a single Orbitrap MS scan (120,000 resolution) every 3 s and Orbitrap MS/MS HCD scans of precursors (NCE 30%, maximal injection time of 80 ms, AGC  $1 \times 10^5$  with a resolution of 15,000). HexNAc oxonium ion (204.087  $m/z$ ) product-dependent MS/MS analysis (56) was used to trigger three additional scans of potential glycopeptides; an Orbitrap EThcD scan (NCE 15%, maximal injection time of 250 ms, AGC  $2 \times 10^5$  with a resolution of 30,000); a ion trap CID scan (NCE 35%, maximal injection time of 40 ms, AGC  $5 \times 10^4$ ) and a stepped collision energy HCD scan (using NCE 32%, 40%, 48% for N-linked glycopeptide samples and NCE 28%, 38%, 48% for O-linked glycopeptide samples with a maximal injection time of 250 ms, AGC  $2 \times 10^5$  with a resolution of 30,000). For *B. pseudomallei* K96243 glycopeptide enrichments, duplicate runs were undertaken as above with the Orbitrap EThcD scans modified to use the extended mass range setting (200  $m/z$  to 3000  $m/z$ ) to improve the detection of high mass glycopeptide fragment ions (61).

**Data Analysis**—Raw data files were batch processed using Byonic v3.5.3 (Protein Metrics Inc (14)) with the proteome databases denoted within Table I. Data were searched on a desktop with two 3.00 GHz Intel Xeon Gold 6148 processors, a 2TB SSD and 128 GB of RAM using a maximum of 16 cores for a given search. For all searches, a semi-tryptic N-ragged specificity was set and a maximum of two missed cleavage events allowed. Carbamidomethyl was set as a fixed modification of cystine whereas oxidation of methionine was included as a variable modification. A maximum mass precursor tolerance of 5 ppm was allowed whereas a mass tolerance of

up to 10 ppm was set for HCD fragments and 20 ppm for EThcD fragments. For open searches of *C. fetus fetus* samples (*N*-linked glycosylation), the wildcard parameter was enabled allowing a delta mass between 200 Da and 1600 Da on asparagine residues. For open searches of *O*-linked glycosylation samples, the wildcard parameter was enabled allowing a delta mass between 200 Da and 2000 Da on serine and threonine residues. For focused searches, all parameters listed above remained constant except wildcard searching which was disabled and specific glycoforms as identified from open searches included as variable modifications. To ensure high data quality, separate data sets from the same biological samples were combined using R and only glycopeptides with a Byonic score >300 were used for further analysis. This score cutoff is in line with previous reports highlighting that score thresholds greater than at least 150 are required for robust glycopeptide assignments with Byonic (44, 57). It should be noted that a score threshold of above 300 resulted in false discovery rates of less than 1% for all combined data sets. Pearson correlation analysis of delta mass profiles was undertaken using Perseus (58). Data visualization was undertaken using ggplot2 within R with all scripts included in the PRIDE uploaded data sets.

**Experimental Design and Statistical Rationale**—For each bacterial strain examined three biological replicates were prepared and used for glycopeptide enrichments leading to three LC-MS runs per bacterial strain. Three separate enrichments were prepared and run to ensure an accurate representation of the observable glycoproteome. *B. pseudomallei* K96243 biological replicates were run twice with two different instrument methods to improve the characterization of the 990 Da glycan. For *C. fetus fetus* NCTC 10842 unenriched peptide samples were run with identical methods as those used for glycopeptide analysis to assess the presence of formylated glycans before enrichment.

## RESULTS

**Open Database Searching Allows the Identification of Bacterial *N*-Linked Glycopeptides**—Although open database searching enables the detection of a variety of modifications including eukaryotic glycosylation (46, 47), to our knowledge, it has not been applied to bacterial systems or the study of atypical forms of glycosylation. To enable the identification of glycopeptides with complex glycans, large delta mass windows are needed as even modest glycans (>three monosaccharides) would be larger than the 500 Da window typically used for open searching (39–43). To enable wide mass searching we used Byonic, which uses a peak look up approach for the identification of PSMs (14, 51). This approach is both rapid and overcomes the combinatorial explosion which leads to long search times with database open searching (48, 51). To assess the ability of Byonic based open searching to identify bacterial glycopeptides, we first examined glycopeptide enrichments of *C. fetus fetus* NCTC 10842. *C. fetus fetus* possesses a small proteome (~1600 proteins (61)) and is known to produce two *N*-linked glycans composed of  $\beta$ -GlcNAc- $\alpha$ 1,3-[GlcNAc1,6]-GlcNAc- $\alpha$ 1,4-GalNAc- $\alpha$ 1,4-GalNAc- $\alpha$ 1,3-diNAcBac (1243.507 Da) and  $\beta$ -GlcNAc- $\alpha$ 1,3-[Glc1,6]-GlcNAc- $\alpha$ 1,4-GalNAc- $\alpha$ 1,4-GalNAc- $\alpha$ 1,4-diNAcBac (1202.481 Da) where diNAcBac is the bacterial specific sugar 2,4-diacetamido-2,4,6 triideoxyglucopyranose (25).

We searched ZIC-HILIC glycopeptide enrichments of *C. fetus fetus* allowing a wildcard mass of 200 Da to 1600 Da on asparagine enabling the processing of 3 h LC-MS runs by open searching in less than 2 h (supplemental Fig. S1A). Examination of the detected modifications by binning the observed delta masses in 0.001 Da increments demonstrated a clear cluster of modifications with masses >1000 Da (Fig. 1A). Within these masses 1242.501 Da and 1201.475 Da were the most numerous delta masses observed (Fig. 1A, supplemental Table S1) yet these are one Da off the expected glycoforms of *C. fetus fetus* (25). Close examination of the observed delta masses reveals evidence of mis-assignments of the mono-isotopic masses by the appearance of satellite peaks (42, 48) differing by exactly one Da (supplemental Fig. S2A). The mis-assignment of the mono-isotopic peak of large glycopeptides has been previously noted (62) and within Byonic is combated by allowing isotope re-assignment denoted as the “off-by-x” parameter. Examination of the “off-by-x” masses supports the inappropriate mass correction of the 1243/1202 Da glycans to the observed 1242/1201 delta masses (supplemental Fig. S2B). These results support that the 1243/1202 Da glycans are readily detected in *C. fetus fetus* samples using open database searching despite the splitting of the delta mass observations across multiple masses because of errors in mono-isotope assignments.

Surprisingly, our open search also revealed additional glycoforms corresponding to formylated glycans (+27.99 Da) as well as a modification corresponding to the loss of a HexNAc (−203.079 Da) or Hex (−162.053 Da) from the 1243 Da or 1202 Da glycans respectively (Fig. 1B). MS/MS analysis supports these delta masses as unexpected but bona fide glycoforms (supplemental Fig. S3A to S3J). Formylated glycans have been previously observed (25, 28) during ZIC-HILIC enrichment and are most likely artifacts because of the high concentrations of formic acid (44) used during enrichment. These formylated glycans represented a significant proportion of all potential glycopeptide PSMs (Fig. 1B, supplemental Table S1). Consistent with glycan formylation being artifactual, it is not observed on *C. fetus fetus* glycopeptides within unenriched samples (supplemental Fig. S4). To assess the accuracy of the glycan masses obtained using open searching, we extracted the mean delta mass of the 1243 and 1202 Da glycans using a density based fitting approach (63) (Fig. 1C and 1D). We find the open search defined mass of the 1243 Da and 1202 Da glycans are both within 5 ppm of the known masses (25) supporting that this approach allows high accuracy determination of large modifications. Finally, we assessed the proteome coverage of our open database approach to a traditional search using the seven identified glycoforms (1040.423 Da, 1068.419 Da, 1202.475 Da, 1230.469 Da, 1243.501 Da, 1271.497 Da, 1299.492 Da, Fig. 1B) as a focused search (64). Consistent with previous studies focused searches outperformed open database

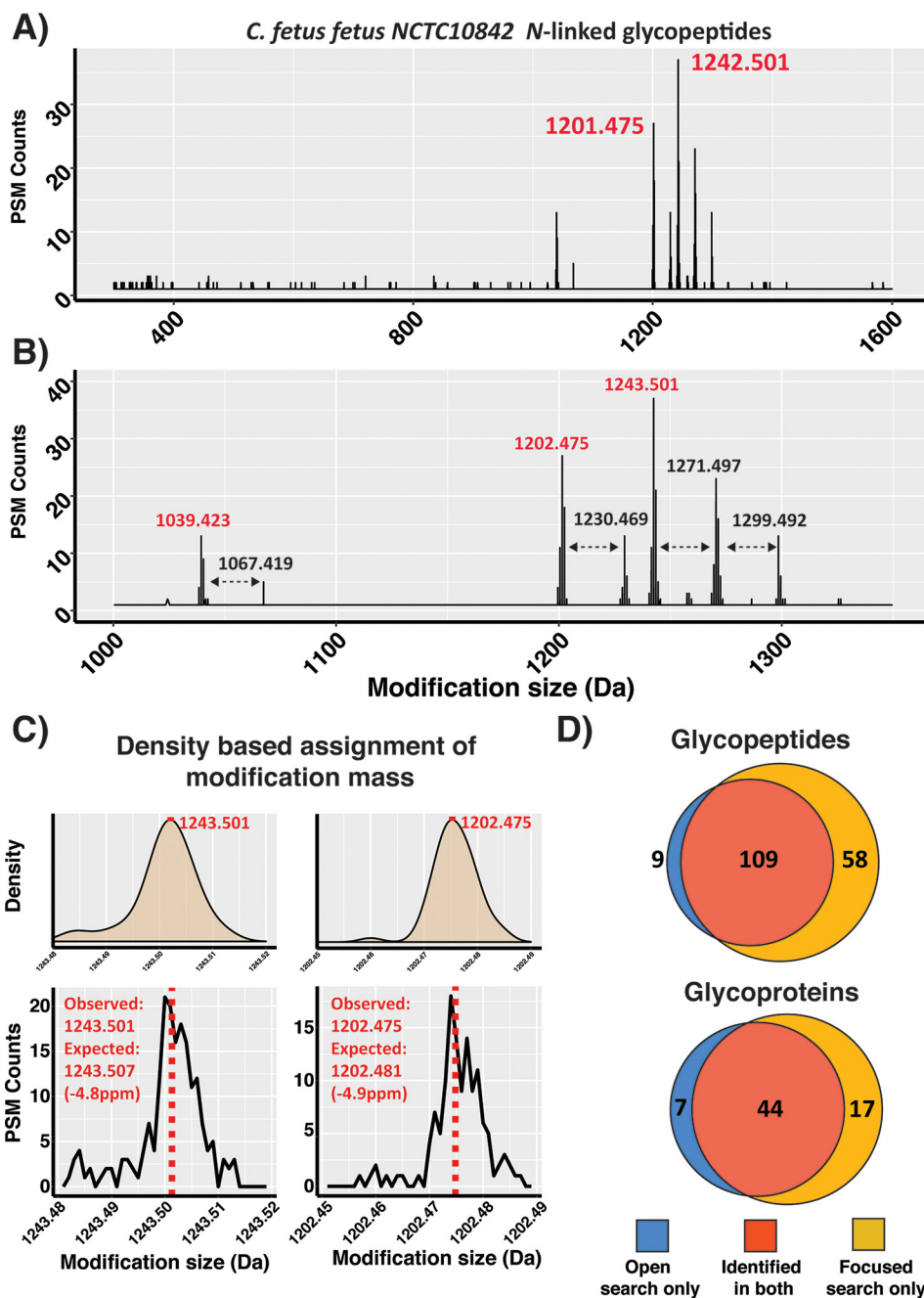
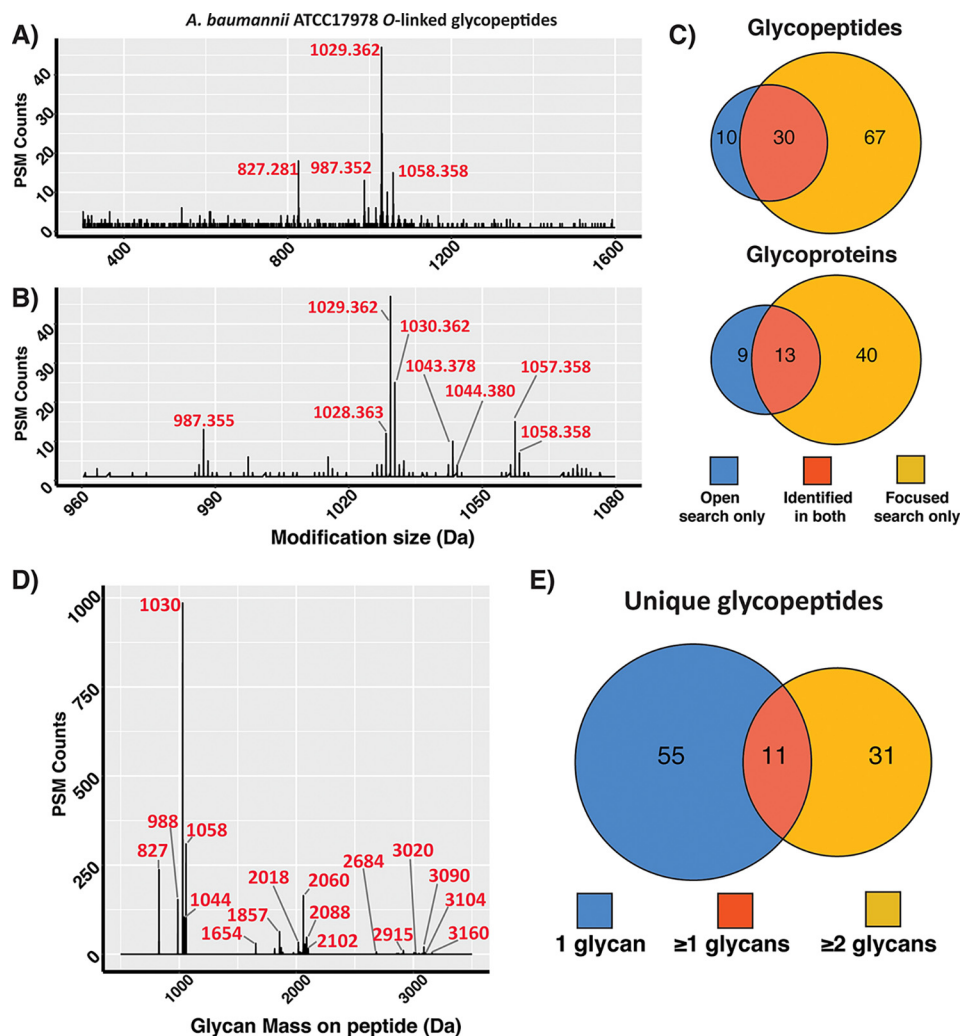


FIG. 1. **Open searching analysis of *C. fetus fetus* NCTC 10842 glycopeptides.** A, *C. fetus fetus* glycopeptide delta mass plot of 0.001 Da increments showing the detection of PSMs modified with masses over 1000 Da. B, Zoomed view of the *C. fetus fetus* glycopeptide delta mass plot highlighting the most numerous observed delta masses; red masses correspond to nonformylated glycans whereas black masses correspond to formylated glycans. C, Density and zoomed delta mass plot of the *C. fetus fetus* glycan masses 1243.507 Da and 1202.481 Da. D, Comparison of the unique glycopeptide sequences and glycoproteins observed between open and focused searches across *C. fetus fetus* data sets.

searches (39, 46, 47) improving the identification of unique glycopeptides by 35% and glycoproteins by 28% (Fig. 1D, supplemental Table S2). This improvement was also associated with an increase in the average Byonic score from 456 to 491 for glycopeptide PSMs with identical MS/MS scans receiving an average 114 Byonic score increase between focused and open search assignments (supplemental Fig.

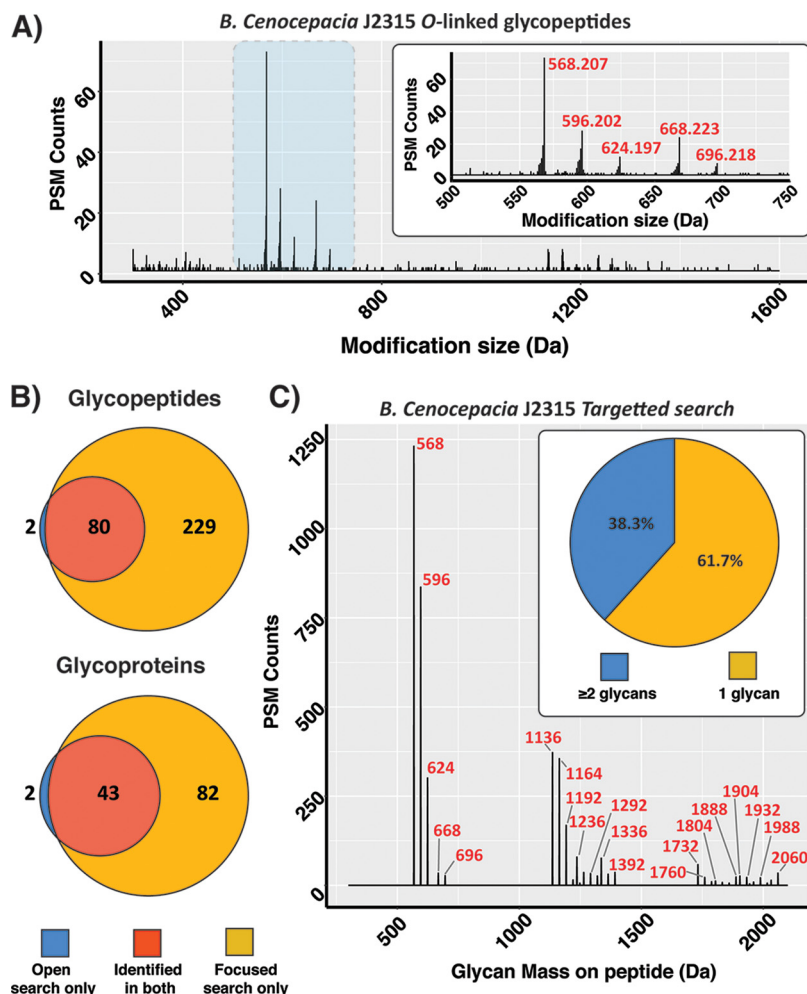
S5). These focused searches also demonstrated that formylated glycans account for nearly a 1/3 of all glycopeptide assigned PSMs (>1246 formylated glycan PSMs out of the total 3824 glycopeptide PSMs, supplemental Table S2). Combined, these results demonstrate open searching allows the detection of heterogeneous bacterial *N*-linked glycopeptides without the need to define glycans before searching.



**FIG. 2. Open searching analysis of *A. baumannii* ATCC17978 glycopeptides.** *A*, *A. baumannii* glycopeptide delta mass plot of 0.001 Da increments showing the detection of PSMs modified with masses over 800 Da. *B*, Zoomed view of *A. baumannii* glycopeptide delta mass plot highlighting the most numerous observed modifications. *C*, Comparison of unique glycopeptide sequences and glycoproteins observed between open and focused searches across *A. baumannii* data sets. *D*, Glycan mass plot showing the amount of glycan (in Da) observed on glycopeptide PSMs within the focused searches. *E*, Venn diagram showing the number of unique glycopeptide sequences grouped based on the number of glycans observed on these peptides.

**Open Database Searching Allows the Identification of Bacterial O-Linked Glycopeptides**—To assess open searching's compatibility with bacterial O-linked glycopeptides, we examined glycopeptide enrichments of *A. baumannii* ATCC17978. The *A. baumannii* proteome is twice the size of *C. fetus fetus* (~3600 proteins (65)) with glycosylation of both serine and threonine residues reported to date (52). Within this system, glycoproteins are modified predominantly with the glycan GlcNAc3NAcA4OAc-4-( $\beta$ -GlcNAc-6-)- $\alpha$ -Gal-6- $\beta$ -Glc-3- $\beta$ -GalNAc where GlcNAc3NAcA4OAc corresponds to the bacterial sugar 2,3-diacetamido-2,3-dideoxy- $\alpha$ -D-glucuronic acid (glycan mass 1030.368 Da (52)). Importantly, this terminal glucuronic acid can also be found in methylated as well as un-acetylated states (corresponding to the glycan masses 1044.383 Da and 988.357 Da respec-

tively (26, 52)). *A. baumannii* glycopeptide enrichments were searched allowing a wildcard mass of 200 Da to 2000 Da on serine and threonine residues. The increased complexity of this search, both in terms of the number of amino acids potentially modified as well as the size of the proteome, resulted in a marked increase in the search times per data files to ~10 h (supplemental Fig. S1B). Within these samples, open searching readily enabled the identification of multiple delta masses of similar sizes to the expected glycans of *A. baumannii* ATCC17978 as well as the unexpected glycoforms of 827.281 and 1058.358 Da (Fig. 2A, supplemental Table S3). These novel glycan masses are consistent with formylation (+27.99 Da) as well as the loss of HexNAc (-203.079 Da) from the 1030 Da glycan with MS/MS analysis supporting these assignments (supplemental Fig. S6).

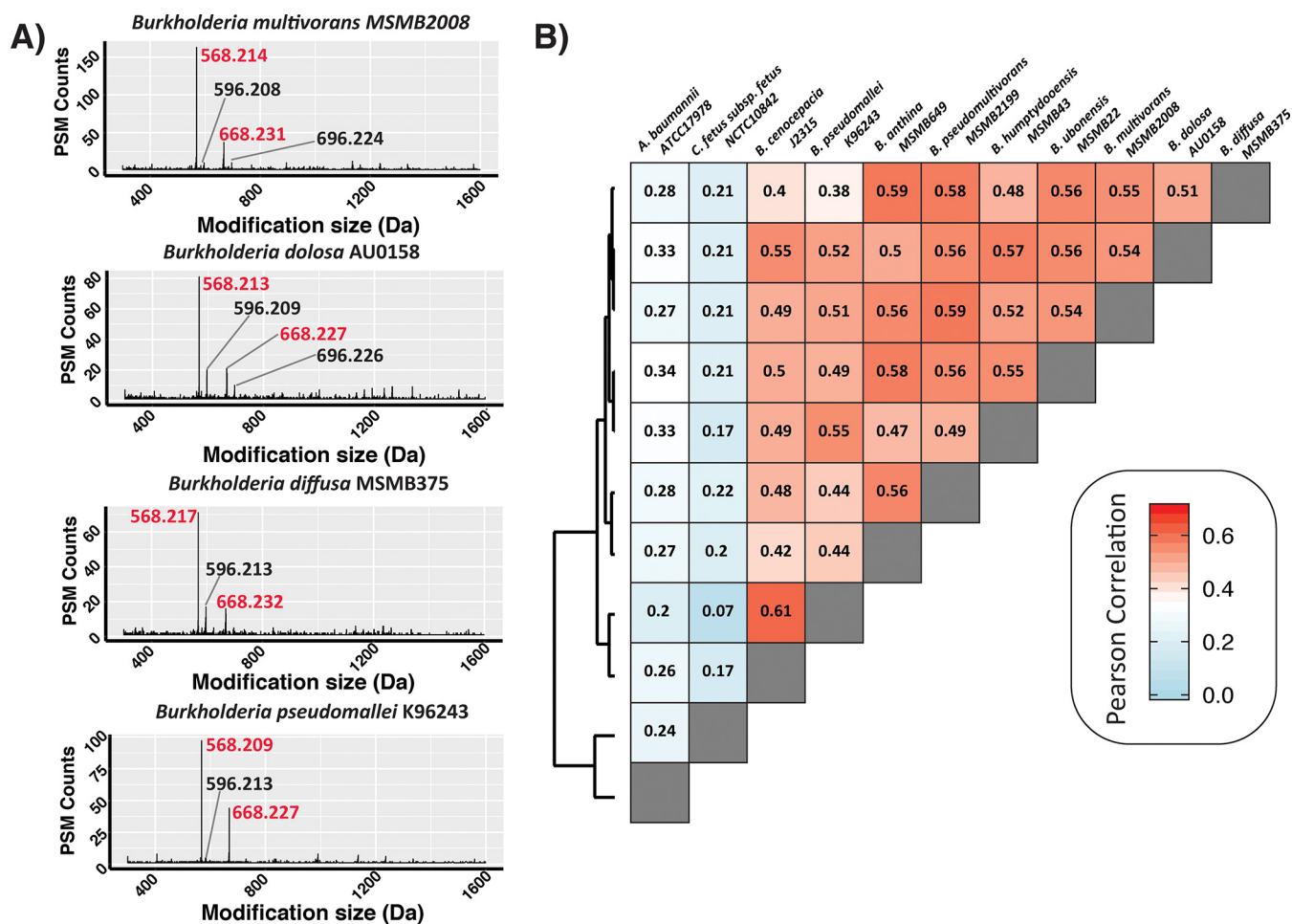


**FIG. 3. Open searching analysis of *B. Cenocepacia* J2315 glycopeptides.** *A*, *B. Cenocepacia* glycopeptide delta mass plot of 0.001 Da increments showing the detection of PSMs modified with masses over 500 Da. Highlighted area shown in zoomed panel. *B*, Comparison of the unique glycopeptide sequences and glycoproteins observed between open and focused searches across *B. Cenocepacia* data sets. *C*, Glycan mass plot showing the amount of glycan (in Da) observed on glycopeptides PSMs within the focused searches. Nearly 40% of all glycopeptide PSMs are decorated with two or more glycans.

Examination of these masses revealed the most numerous delta masses (1029.362 Da, 987.355 Da and 1043.378 Da) were one Da less than the expected *A. baumannii* glycan masses (Fig. 2B (26, 52)). As with *C. fetus fetus*, inspection of these assignments reveals the incorrect application of the “off-by-x” parameter leading to the splitting of delta masses across multiple mass assignments separated by one Da (supplemental Fig. S7). Using the masses 1030.368 Da, 988.357 Da, 1044.383 Da, 827.281 Da and 1058.358 Da, we researched these *A. baumannii* data sets to assess the performance of open to focused searching. In contrast to the ~35% increase in unique glycopeptides observed between open and focused searches in *C. fetus fetus* we noted a dramatic >240% improvement in unique glycopeptides identified within *A. baumannii* using focused searches (Fig. 2C). To understand this dramatic improvement, we examined the 67 glycopeptides unique to the focused searches. Within these

glycopeptides we noted a large proportion of PSMs corresponding to glycopeptides modified with multiple glycans (Fig. 2D, supplemental Table S4). In fact, >20% (494 out of the total 2282 glycopeptide PSMs) corresponded to glycopeptides with greater than one glycan attached. Within these PSMs, 31 unique peptide sequences are only observed with >1 glycan attached (Fig. 2E). The increased numbers of unique glycopeptides identified within focused searches were also associated with an increase in the mean Byonic scores as well as identical MS/MS scans receiving an average 149 Byonic score increase between focused and open search assignments (supplemental Fig. S8). Similar to *C. fetus fetus*, a large proportion of glycopeptide PSMs were identified with formylated glycans (>309 out of the total 2282 glycopeptide PSMs, supplemental Table S4). It is important to note that the delta masses of multiply glycosylated peptides fall outside the 2000 Da window used for open





**FIG. 4. Comparison of Burkholderia glycoproteomes using open searching.** *A*, Representative delta mass plots of four out of the eight Burkholderia strains examined demonstrating the 568 Da and 668 Da glycans are frequently identified delta masses in Burkholderia glycopeptide enrichments. Formylated glycans are denoted in black whereas Burkholderia O-linked glycans are in red. *B*, Pearson correlation and clustering analysis of delta mass plots enable the comparison and grouping of samples.

searching making the inability to detect these glycopeptides an expected limitation of the search parameters. Thus, although open searching enables the rapid identification of glycopeptides, large glycans/multiply glycosylated peptides can be overlooked supporting the value of a two-step (open followed by focused) searching approach.

*Open Database Searching Enables the Identification of Glycosylation within Large Proteomes*—As open searching enabled the identification of both N and O-linked glycopeptides, we sought to explore the compatibility of this approach with larger proteomes using glycopeptide enriched samples from the bacteria *B. Cenocepacia* J2315 as a model. The *B. Cenocepacia* proteome encodes ~7000 proteins (66) and possesses an O-linked glycosylation system responsible for modifying at least 23 proteins (37). Previously, we showed that this glycosylation system transfers two glycans composed of  $\beta$ -Gal-(1,3)- $\alpha$ -GalNAc-(1,3)- $\beta$ -GalNAc and Suc- $\beta$ -Gal-(1,3)- $\alpha$ -GalNAc-(1,3)- $\beta$ -GalNAc where Suc is Succinyl with these glycans corresponding to the masses 568.211 Da

and 668.228 Da respectively (23, 37). As with *A. baumannii*, the increased complexity of this proteome led to an increase in the search time with individual data files taking ~20 h to process (supplemental Fig. S1C). These open searches revealed the presence of the expected glycoforms of *B. cenocepacia* (568.207 Da and 668.223 Da) as well as additional formylated variants (596.202 Da, 624.197, and 696.218 Da) leading to the identification of five unique glycoforms (Fig. 3A, supplemental Table S5). Unlike the large glycans of *C. fetus fetus* and *A. baumannii*, it is notable that the mono-isotopic mass of the known *B. Cenocepacia* glycans (37) were correctly assigned (Fig. 3A). Thus, this supports that for smaller glycans mis-assignment of the mono-isotopic masses during open searches does not appear as problematic.

Incorporating these glycoforms into focused searches again led to a dramatic ~4-fold increase in the number of glycopeptides and a ~2-fold increase in the total number glycoproteins identified compared with open searches (Fig. 3B, supplemental Table S6). Although the improvement in

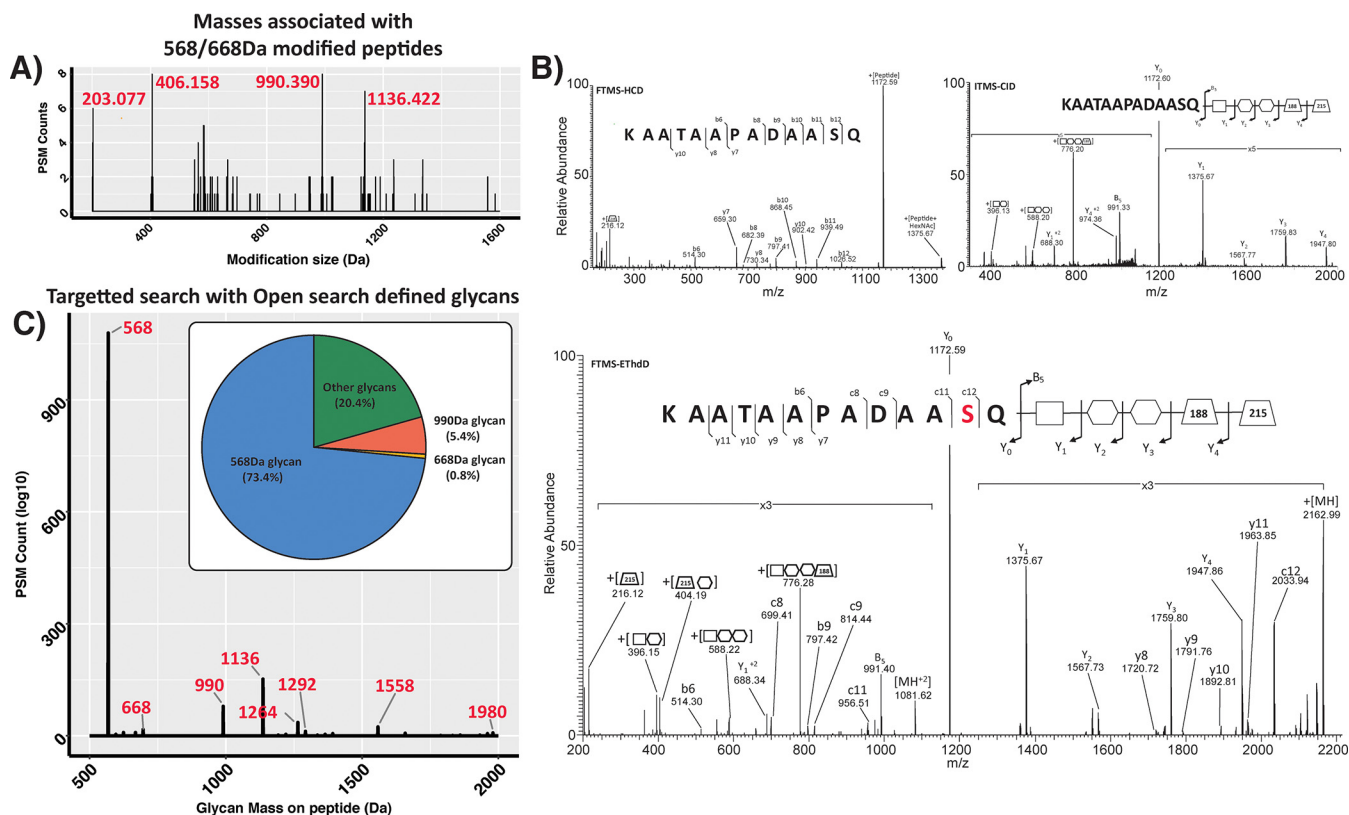


FIG. 5. Identification of minor glycoforms within *B. pseudomallei* K96243. A, Delta mass plot, binned by 0.001 Da increments, showing delta masses observed for peptide sequences also modified with the 568 or 668D glycans. B, MS/MS analysis (FTMS-HCD, ITMS-CID and FTMS-ETHCD) supporting the assignment of a linear glycan of HexNAc-Heptose-Heptose-188-215 attached to the peptide KAATAAPADAASQ. C, Glycan mass plot showing the amount of glycan (in Da) observed on glycopeptides PSMs within focused searches. Only ~6% of all PSMs observed are modified with the 990 Da glycan.

the total number of identifications was associated with a decrease in the mean Byonic score (from 728 to 700) the assigned scores of identical MS/MS scans reveals focused searches lead to an increase in the average Byonic score of 175 compared with assignments from open searches (supplemental Fig. S9). As the dramatic improvement in the glycoproteome coverage of *A. baumannii* was partially driven by the detection of multiply glycosylated peptides we examined the amount of glycosylation within glycopeptide PSMs in *B. Cenocepacia*. As *B. Cenocepacia* glycopeptides modified with multiple glycans would be less than 2000 Da, we were surprised by the limited number of multiply glycosylated peptides identified within our open searches (<10% of all identified glycopeptides, supplemental Fig. S10). In contrast, focused searches identified ~40% of all PSMs (Fig. 3C, 1508 out of 3937 identified glycopeptide PSMs) corresponded to multiply modified peptides. As with *C. fetus fetus* and *A. baumannii* formylated glycans make up nearly 50% of all glycopeptide PSMs (Fig. 3C, supplemental Table S6). These data supports that, although open searching performs well for singly modified peptides, this approach appears to underrepresent multiply glycosylated peptides even if the combined mass of the glycan is within the range of the open search.

*Open Database Searching Allows the Screening of Glycan Use across Biological Samples*—Having established that open searching enables the identification of a range of glycans, we sought to explore if this could also facilitate the comparison of glycan diversity across bacterial samples. Recently, we reported that a single-loci was responsible for the generation of the O-linked glycans in *B. Cenocepacia* and that this loci is conserved across the *Burkholderia* (23). Although these results support that *Burkholderia* species use similar glycans, it has been noted that within other bacterial genera extensive glycan heterogeneity exists (25, 26, 29, 32). As glycan heterogeneity can be challenging to predict, we reasoned that open searching would provide a means to assess the similarities in glycans used across *Burkholderia* species. We examined glycopeptide enrichments from eight species of *Burkholderia* (*B. pseudomallei* K96243; *B. multivorans* MSMB2008; *B. dolosa* AU0158; *B. humptydoensis* MSMB43; *B. ubonensis* MSMB22, *B. anthina* MSMB649; *B. diffusa* MSMB375; and *B. pseudomultivorans* MSMB2199). Examination of the delta masses observed across these eight species demonstrates that the 568 Da and 668 Da glycoforms as well as their formylated variants are present in all strains (Fig. 4A and supplemental Fig. S11, supplemental Tables S7 to S14).

Having generated “delta mass fingerprints” for each species, we assessed if these profiles could enable the comparison of glycan use using Pearson correlation and hierarchical clustering (Fig. 4B and supplemental Fig. S12). Consistent with the similarities in the delta mass fingerprints Pearson correlation and hierarchical clustering resulted in the grouping of all Burkholderia species compared with the delta mass fingerprints of *C. fetus fetus* and *A. baumannii* (Fig. 4B). These results support that consistent with the conservation of the glycosylation loci within Burkholderia, the major glycoforms observed within Burkholderia species, based on mass at least, are identical. It should be noted that as with the above glycopeptide data sets, focused searches significantly improved the identification of glycopeptides and glycoproteins in all Burkholderia species (supplemental Fig. S13, supplemental Table S15 to S22).

**Open Database Searching Allows the Detection of Glycoforms Identified at a Low Frequency Based on Known Glycosylatable Peptides**—In addition to allowing the comparison of glycan diversity across species, we reasoned that open searching would also allow the identification of novel glycans based on the shared use of glycosylatable peptide sequences. Within bacterial glycosylation studies, proteins compatible with different glycosylation machinery are routinely used to “fish” out glycans used for protein glycosylation in different bacterial species (26, 32). Similarly, by focusing on peptides modified with the 568/668 Da glycans we hypothesized this would provide the means to detect alternative glycans used for glycosylation within Burkholderia species. To assess this, we examined the glycopeptide enrichments of *B. pseudomallei* K96243 filtering for delta masses only observed on peptide sequences also modified with either the 568/668 Da glycans (Fig. 5A). Examination of these delta masses readily revealed the presence of PSMs matching the modification of peptides with single (203.077 Da) or double (406.158 Da) HexNAc residues, two 568 Da glycans (1136.422 Da) and an unexpected mass at 990.390 Da (Fig. 5A). Examination of PSMs assigned to this 990 Da delta mass revealed a linear glycan composed of HexNAc-Heptose-Heptose-188-215 where the 188 Da and 215 Da are moieties of unknown composition (Fig. 5B). Incorporation of this unexpected glycan mass into a focused search with the known Burkholderia glycans demonstrated that the 990.390 Da glycan is observed on multiple peptide substrates yet less than 6% of all glycopeptide PSMs correspond to this novel glycan (Fig. 5C). Thus, this demonstrates that open searching provides an effective means to detect unexpected glycoforms which could be overlooked because of the low frequency of their occurrence in glycoproteomic data sets.

### DISCUSSION

MS analysis of glycoproteomic samples typically requires knowledge of both the proteome and possible glycan compositions to facilitate software-based identification (13). As

bacterial glycosylation systems do not use glycans found in eukaryotic glycan databases (23–32), we sought to establish an alternative approach for the high-throughput analysis of bacterial glycoproteomes. Within this work we demonstrate that wide mass open database searching enables the identification of bacterial glycosylation without the need to define glycan masses before searching. This approach overcomes a significant bottleneck in the identification and characterization of novel bacterial glycosylation systems. We demonstrate that a range of diverse glycan structures, both reported (25, 26, 37, 52) and unreported, such as the 990 Da glycan observed in *B. pseudomallei* K96243, as well as glycan artifacts such as formylated glycans can be identified using this approach. In addition, we also demonstrate that open database searches can be used to provide a simple means to compare delta mass profiles across biological samples. This enables a straightforward method to compare and contrast bacterial glycoproteomes, enabling the grouping of Burkholderia profiles from nonsimilar glycan profiles such as those seen in *C. fetus fetus* or *A. baumannii*.

Within this work we used open searching within Byonic, a widely used tool in the glycoproteomic community for the analysis of glycosylation (57, 67, 68). This enabled us to directly compare the performance of open searches to focused glycopeptide searches within the same platform. We observed a marked improvement in glycopeptide and glycoprotein identifications within focused searches, especially for glycopeptides modified with multiple glycans. As several unique considerations are needed for optimal glycopeptide identification, such as accounting for glycan fragments (13, 69), this improvement in performance is unsurprising. Consistent with this we observed an increase in the Byonic scores within focused searches compared with open searches for most data sets (supplemental Figs. S5, S8, and S9). This improvement translates to an increase in the numbers of unique glycopeptides and glycoproteins identified by ~35% to 240%. This is in line with previous studies (46, 47) and the observation that nonoptimized glycopeptide analysis can lead to a reduction in glycoproteome coverage (64). Although Byonic was used within this study it should be noted alternative noncommercial platforms such as MSfragger (43) also allow open searching. In our hands MSfragger performed comparably to Byonic for the identification of glycoforms using open searches (supplemental Figs. S14A to S14F). Yet, as with our open Byonic searches MSfragger did not identify as many unique glycopeptides/glycoproteins as focused Byonic searches (supplemental Figs. S15A to S15F). These results demonstrate that open searching can be used to identify glycopeptides, yet because of the unique challenges associated with glycopeptide identification open searches can be less sensitive than focused searches.

Although we have used open searching with masses up to 2000 Da wider mass ranges can and have been used to identify glycopeptides. In fact, for eukaryotic glycosylation analysis open searches with +3000 Da have been demonstrated

enabling the identification of glycans of up to ~2000 Da in size (46, 47). Although multiple open searching tools allow the maximum delta mass range to be widened this can be at the expense of search times (48). It should be noted for open searching there is likely an upper limit for the maximum delta mass which can be used before no additional high confidence assignments are gained. This maximum limit is not only determined by the analytes being examined but also the MS/MS acquisition parameters. For example, for glycopeptides as the amount of glycan decorating peptides increases the optimal amounts of collision energy and ETD reaction times begin to diverge from standard parameters (70, 71). Thus, even using open searching some specific subsets of peptides may still be unidentifiable without careful tailoring of the data acquisition methods.

At its core, this analytical approach uses a “strength in numbers” based strategy for the detection of glycans. A key strength of this approach is that it does not require the identification of unmodified versions of a peptide for the modified forms to be identified as required in dependent peptide-based approaches (49, 50). This independence of the need for unmodified peptides makes this approach compatible with enrichment strategies such as ZIC-HILIC glycopeptide enrichment. This is important as for optimal performance this approach requires large numbers of PSMs with identical delta masses. It should be noted that within ZIC-HILIC enrichments bacterial glycopeptides are still only a minor proportion of observed peptides being <10% of the identified PSMs (supplemental Fig. S16). Within this work we focused on bacterial glycosylation systems known to target multiple protein substrates (16, 18), ensuring large numbers of unique PSMs/peptide sequences would be identified. We found that, within glycopeptide enrichments, the known glycoforms of *C. fetus fetus*, *A. baumannii* and *B. cenocepacia* were easily detected whereas infrequently observed glycans, such as the 990 Da glycan observed within *B. pseudomallei*, required additional filtering to distinguish this from background signals. This supports that although open searching enables the detection of glycoforms, it is sensitive to the frequency at which modification events are observed within data sets. Although we used filtering based on glycosylatable peptide sequences to identify low frequency events, recently Kernel density estimation based fitting approaches were shown to effectively address this issue in a more general manner (63). Thus, open database searching provides multiple approaches to identify modifications even those which are poorly resolved from background.

A surprising observation within our open searches was the commonality of glycan formylation events within enriched glycopeptide samples. Recently it was shown that formic acid concentrations as low as 0.1% could lead to peptide formylation (44). For the enrichment of bacterial glycopeptides ZIC-HILIC enrichment with 5% formic acid/80% acetonitrile has been extensively used (24, 26–28, 36, 37, 52, 72) yet

the observation of formylation artifacts on a substantial number of glycopeptide PSMs (supplemental Tables S2, S4, S6, S15 to S22) raises concerns about this protocol. Multiple reagents including Tris (73) and alkaline solutions such as ammonium hydroxide/sodium hydroxide (74) can lead to alterations within glycan structures necessitating the judicious use of these chemicals within glycan/glycopeptide sample preparation protocols. Although 5% formic acid improves the selectivity for glycopeptides within ZIC-HILIC enrichments (75) alternative acids can also be used. Previously Mysling *et al.* demonstrated that formic acid could be substituted for TFA to improve the enrichment of glycopeptides (76), whereas Ding *et al.* noted that hydrochloric acid was an effective ion-pairing agent for normal phase enrichment of bacterial glycopeptides (77). Thus, the observation of formylation highlights that alternative buffers should be considered for future glycopeptide studies and that formylated glycans need to be considered when analyzing glycopeptide data sets where formic acid has been used during glycopeptide enrichments.

Although open searching enabled the identification of glycosylation within all bacterial samples, the analysis of *C. fetus fetus* and *A. baumannii* data sets highlighted the commonality at which mono-isotopic masses of glycopeptides with large glycans (>1000 Da) are mis-assigned. This problem has been highlighted previously (62) and is not unique to glycopeptides with the mono-isotopic mass of other large biomolecules such as cross-linked peptide shown to be mis-assigned in 50 to 75% of PSMs (78). This mis-assignment of mono-isotopic masses leads to the splitting of the number of observed PSMs with a specific glycan mass across multiple mass channels. Although we demonstrate that these mis-assigned glycopeptides can be readily identified by examining the “off-by-x” parameter, it should be noted that this splitting dilutes the observable glycopeptides at a specific mass, complicating the analysis of glycoproteomes from open searches. This complication, coupled with the lower sensitivity of glycopeptide identification with open searching compared with focused searches discussed above, supports that open searching is a useful discovery tool yet typically under reports unique glycopeptides and glycoproteins within data sets. The simplest solution to this issue is to use open searching as a means to identify glycans which can then be included as variable modifications within a focused search. As highlighted above, this significantly improves glycoproteome coverage and in our hands provided the flexibility of being able to detect novel glycans yet also ensured optimal identification of glycopeptides. Automated pipelines using multi-step searching have already been demonstrated (42, 48) yet to our knowledge these have not been optimized or implemented for glycopeptide identification. Thus, we recommend a multi-step analysis to enable the identification of atypical glycosylation, using open searching to define glycans which are then incorporated into focused searches.

Finally, it should be noted that although not the subject of this manuscript, the glycoproteins/glycopeptides identified in this work are themselves a useful resource for the bacterial glycosylation community. Previous studies on *C. fetus fetus*, *A. baumannii* and *B. cenocepacia* identified a total of 26, 26, and 23 unique glycoproteins respectively (25, 26, 37) yet the majority of these studies were undertaken on previous generations of MS instrumentation. Within this work, undertaken on a current generation instrument, we observed a marked improvement in the number of glycoproteins identified with 61 (2.3-fold), 53 (2.0-fold) and 125 (5.4-fold) glycoproteins identified in *C. fetus fetus*, *A. baumannii* and *B. cenocepacia* respectively. Similarly, our glycoproteomic analysis of the 8 Burkholderia species highlights that at least 70 proteins are glycosylated within each Burkholderia species. Taken together, this work highlights that the glycoproteomes of most bacterial species are likely far larger than earlier studies suggested with open searching providing an accessible starting point to probe these systems.

DATA AVAILABILITY

All MS proteomics data (Raw data files, Byonic search outputs, R Scripts and output tables) have been deposited into the PRIDE ProteomeXchange Consortium repository (59, 60) with the data set identifier: [PX018587](https://www.ebi.ac.uk/pride/archive/study/PXD018587).

**Acknowledgments**—We thank the Melbourne Mass Spectrometry and Proteomics Facility of The Bio21 Molecular Science and Biotechnology Institute for access to MS instrumentation and Byonic. We would like to thank Christine Szymanski and Justin Duma for the kind gift of *C. fetus fetus* NCTC 10842 lysates; Mitali Sarkar-Tyson and Nicole Bzdyl for providing the *B. pseudomallei* K96243 lysates; Deborah Yoder-Himes, Mark Mayo, Bart Currie and Amy Cain for kindly providing Burkholderia strains for this analysis as well as Chris McDevitt and Saleh Alquethamy for providing *A. baumannii* ATCC17978. We thank Ben Parker and Nick Williamson for their critical feedback on the manuscript.

**Funding and additional information**—This work was supported by a National Health and Medical Research Council of Australia (NHMRC) project grant awarded to NES (APP1100164).

**Author contributions**—ARAI performed research; NES designed research; NES analyzed data; NES wrote the paper.

**Conflict of interest**—Authors declare no competing interests.

**Abbreviations**—The abbreviations used are: MS, mass spectrometry; LB, Luria Bertani; PBS, phosphate-buffered saline; SDS, sodium dodecyl sulfate; NCE, normalized collisional energy; ZIC-HILIC, Zwitterionic Hydrophilic Interaction Liquid Chromatography; Tris, Tris(hydroxymethyl)aminomethane; TFA, Trifluoroacetic acid; DTT, Dithiothreitol; ETHcD, Electron-trans-

fer/higher-energy collision dissociation; HCD, higher-energy collision dissociation; CID, Collision-induced dissociation; AGC, Automatic Gain Control; DMSO, Dimethyl sulfoxide; diNAcBac, 2,4-diacetamido-2,4,6 trideoxyglucopyranose; PSMs, peptide spectrum matches; Glc, Glucose; Gal, galactose; GalNAc, N-Acetylgalactosamine; GlcNAc, N-Acetylglucosamine; HexNAc, N-acetylhexoseamine; Hex, Hexose; 2,3-diacetamido-2,3-dideoxy- $\alpha$ -D-glucuronic acid, GlcNAc3NAcA4OAc.

Received April 21, 2020, and in revised form, May 27, 2020 Published, MCP Papers in Press, June 23, 2020, DOI 10.1074/mcp.TIR120.002100

REFERENCES

1. Struwe, W. B., and Robinson, C. V. (2019) Relating glycoprotein structural heterogeneity to function - insights from native mass spectrometry. *Curr. Opin. Struct. Biol.* **58**, 241–248
2. Brockhausen, I., Stanley, P., et al. (2015) O-GalNAc Glycans. In: Varki A, Cummings RD, Esko JD, Stanley P, Hart GW, editors. *Essentials of glycobiology*. Cold Spring Harbor (NY) 113–123
3. Stanley, P., Taniguchi, N., Aebi, M., et al. (2015) N-Glycans. In: Varki A, Cummings RD, Esko JD, Stanley P, Hart GW, editors. *Essentials of glycobiology*. Cold Spring Harbor (NY) 99–111
4. Khoury, G. A., Baliban, R. C., and Floudas, C. A. (2011) Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. *Sci. Rep* **1**
5. Moremen, K. W., Tiemeyer, M., and Nairn, A. V. (2012) Vertebrate protein glycosylation: diversity, synthesis and function. *Nat. Rev. Mol. Cell Biol.* **13**, 448–462
6. Xu, C., and Ng, D. T. (2015) Glycosylation-directed quality control of protein folding. *Nat. Rev. Mol. Cell Biol.* **16**, 742–752
7. Freeze, H. H., Hart, G. W., Schnaar, R. L., et al. (2015) Glycosylation Precursors. In: rd, Varki A, Cummings RD, Esko JD, Stanley P, Hart GW, editors. *Essentials of glycobiology*. Cold Spring Harbor (NY) 51–63
8. Kawano, S., Hashimoto, K., Miyama, T., Goto, S., and Kanehisa, M. (2005) Prediction of glycan structures from gene expression data based on glycosyltransferase reactions. *Bioinformatics* **21**, 3976–3982
9. Campbell, M. P., Royle, L., Radcliffe, C. M., Dwek, R. A., and Rudd, P. M. (2008) GlycoBase and autoGU: tools for HPLC-based glycan analysis. *Bioinformatics* **24**, 1214–1216
10. Böhm, M., Bohne-Lang, A., Frank, M., Loss, A., Rojas-Macias, M. A., and Lütke, T. (2019) Glycosciences.DB: an annotated data collection linking glycomics and proteomics data (2018 update). *Nucleic Acids Res.* **47**, D1195–D1201
11. McDonald, A. G., Tipton, K. F., and Davey, G. P. (2016) A knowledge-based system for display and prediction of O-glycosylation network behaviour in response to enzyme knockouts. *PLoS Comput. Biol.* **12**, e1004844
12. Akune, Y., Lin, C.-H., Abrahams, J. L., Zhang, J., Packer, N. H., Aoki-Kinoshita, K. F., and Campbell, M. P. (2016) Comprehensive analysis of the N-glycan biosynthetic pathway using bioinformatics to generate UniCorn: A theoretical N-glycan structure database. *Carbohydr. Res.* **431**, 56–63
13. Hu, H., Khatri, K., Klein, J., Leymarie, N., and Zaia, J. (2016) A review of methods for interpretation of glycopeptide tandem mass spectral data. *Glycoconj. J.* **33**, 285–296
14. Bern, M., Kil, Y. J., and Becker, C. (2012) Byonic: advanced peptide and protein identification software. *Curr. Protoc. Bioinformatics* Chapter 13: Unit13.20. Epub 2012/12/21. PubMed Central PMCID: PMC3545648
15. Liu, M.-Q., Zeng, W.-F., Fang, P., Cao, W.-Q., Liu, C., Yan, G.-Q., Zhang, Y., Peng, C., Wu, J.-Q., Zhang, X.-J., Tu, H.-J., Chi, H., Sun, R.-X., Cao, Y., Dong, M.-Q., Jiang, B.-Y., Huang, J.-M., Shen, H.-L., Wong, C. C. L., He, S.-M., and Yang, P.-Y. (2017) pGlyco 2.0 enables precision N-glycoproteomics with comprehensive quality control and one-step mass spectrometry for intact glycopeptide identification. *Nat. Commun.* **8**, 438
16. Nothaft, H., and Szymanski, C. M. (2019) New discoveries in bacterial N-glycosylation to expand the synthetic biology toolbox. *Curr. Opin. Chem. Biol.* **53**, 16–24

17. Szymanski, C. M., and Wren, B. W. (2005) Protein glycosylation in bacterial mucosal pathogens. *Nat. Rev. Microbiol.* **3**, 225–237
18. Koomey, M. (2019) O-linked protein glycosylation in bacteria: snapshots and current perspectives. *Curr. Opin. Struct. Biol.* **56**, 198–203
19. Joshi, H. J., Narimatsu, Y., Schjoldager, K. T., Tytgat, H. L. P., Aebi, M., Clausen, H., and Halim, A. (2018) SnapShot: O-glycosylation pathways across Kingdoms. *Cell* **172**, 632–632e2
20. Schaffer, C., and Messner, P. (2017) Emerging facets of prokaryotic glycosylation. *FEMS Microbiol. Rev.* **41**, 49–91
21. Macek, B., Forchhammer, K., Hardouin, J., Weber-Ban, E., Grangeasse, C., and Mijakovic, I. (2019) Protein post-translational modifications in bacteria. *Nat. Rev. Microbiol.* **17**, 651–664
22. Imperiali, B. (2019) Bacterial carbohydrate diversity - a Brave New World. *Curr. Opin. Chem. Biol.* **53**, 1–8
23. Fathy Mohamed, Y., Scott, N. E., Molinaro, A., Creuzenet, C., Ortega, X., Lertmemongkolchai, G., Tunney, M. M., Green, H., Jones, A. M., DeShazer, D., Currie, B. J., Foster, L. J., Ingram, R., De Castro, C., and Valvano, M. A. (2019) A general protein O-glycosylation machinery conserved in Burkholderia species improves bacterial fitness and elicits glycan immunogenicity in humans. *J. Biol. Chem.* **294**, 13248–13268
24. Harding, C. M., Nasr, M. A., Kinsella, R. L., Scott, N. E., Foster, L. J., Weber, B. S., Fiester, S. E., Actis, L. A., Tracy, E. N., Munson, R. S., and Feldman, M. F. (2015) *Acinetobacter* strains carry two functional oligosaccharyltransferases, one devoted exclusively to type IV pilin, and the other one dedicated to O-glycosylation of multiple proteins. *Mol. Microbiol.* **96**, 1023–1041
25. Nothhaft, H., Scott, N. E., Vinogradov, E., Liu, X., Hu, R., Beadle, B., Fodor, C., Miller, W. G., Li, J., Cordwell, S. J., and Szymanski, C. M. (2012) Diversity in the protein N-glycosylation pathways within the *Campylobacter* genus. *Mol. Cell. Proteomics* **11**, 1203–1219
26. Scott, N. E., Kinsella, R. L., Edwards, A. V. G., Larsen, M. R., Dutta, S., Saba, J., Foster, L. J., and Feldman, M. F. (2014) Diversity within the O-linked protein glycosylation systems of *acinetobacter* species. *Mol. Cell. Proteomics* **13**, 2354–2370
27. Scott, N. E., Nothhaft, H., Edwards, A. V. G., Labbate, M., Djordjevic, S. P., Larsen, M. R., Szymanski, C. M., and Cordwell, S. J. (2012) Modification of the *Campylobacter jejuni* N-linked glycan by EptC protein-mediated addition of phosphoethanolamine. *J. Biol. Chem.* **287**, 29384–29396
28. Scott, N. E., Parker, B. L., Connolly, A. M., Paulech, J., Edwards, A. V. G., Crossett, B., Falconer, L., Kolarich, D., Djordjevic, S. P., Højrup, P., Packer, N. H., Larsen, M. R., and Cordwell, S. J. (2011) Simultaneous glycan-peptide characterization using hydrophilic interaction chromatography and parallel fragmentation by CID, higher energy collisional dissociation, and electron transfer dissociation MS applied to the N-linked glycoproteome of *Campylobacter jejuni*. *Mol. Cell. Proteomics* **10**, M000031–MCP201
29. Hadjineophytou, C., Anonsen, J. H., Wang, N., Ma, K. C., Viburienne, R., Vik, Å., Harrison, O. B., Maiden, M. C. J., Grad, Y. H., and Koomey, M. (2019) Genetic determinants of genus-Level glycan diversity in a bacterial protein glycosylation system. *PLoS Genet.* **15**, e1008532
30. Ulasi, G. N., Creese, A. J., Hui, S. X., Penn, C. W., and Cooper, H. J. (2015) Comprehensive mapping of O-glycosylation in flagellin from *Campylobacter jejuni* 11168: A multienzyme differential ion mobility mass spectrometry approach. *Proteomics* **15**, 2733–2745
31. Jervis, A. J., Wood, A. G., Cain, J. A., Butler, J. A., Frost, H., Lord, E., Langdon, R., Cordwell, S. J., Wren, B. W., and Linton, D. (2018) Functional analysis of the *Helicobacter pullorum* N-linked protein glycosylation system. *Glycobiology* **28**, 233–244
32. Jervis, A. J., Butler, J. A., Lawson, A. J., Langdon, R., Wren, B. W., and Linton, D. (2012) Characterization of the structurally diverse N-linked glycans of *Campylobacter* species. *J. Bacteriol.* **194**, 2355–2362
33. Madsen, J. A., Ko, B. J., Xu, H., Iwashkiw, J. A., Robotham, S. A., Shaw, J. B., Feldman, M. F., and Brodbelt, J. S. (2013) Concurrent automated sequencing of the glycan and peptide portions of O-linked glycopeptide anions by ultraviolet photodissociation mass spectrometry. *Anal. Chem.* **85**, 9253–9261
34. Zampronio, C. G., Blackwell, G., Penn, C. W., and Cooper, H. J. (2011) Novel glycosylation sites localized in *Campylobacter jejuni* flagellin FlaA by liquid chromatography electron capture dissociation tandem mass spectrometry. *J. Proteome Res.* **10**, 1238–1245
35. Cain, J. A., Dale, A. L., Niewold, P., Klare, W. P., Man, L., White, M. Y., Scott, N. E., and Cordwell, S. J. (2019) Proteomics reveals multiple phenotypes associated with N-linked glycosylation in *Campylobacter jejuni*. *Mol. Cell. Proteomics* **18**, 715–734
36. Elhenawy, W., Scott, N. E., Tondo, M. L., Orellano, E. G., Foster, L. J., and Feldman, M. F. (2016) Protein O-linked glycosylation in the plant pathogen *Ralstonia solanacearum*. *Glycobiology* **26**, 301–311
37. Lithgow, K. V., Scott, N. E., Iwashkiw, J. A., Thomson, E. L. S., Foster, L. J., Feldman, M. F., and Dennis, J. J. (2014) A general protein O-glycosylation system within the Burkholderia cepacia complex is involved in motility and virulence. *Mol. Microbiol.* **92**, 116–137
38. Abouelhadid, S., North, S. J., Hitchen, P., Vohra, P., Chintoan-Uta, C., Stevens, M., Dell, A., Cuccui, J., and Wren, B. W. (2019) Quantitative analyses reveal novel roles for N-glycosylation in a major enteric bacterial pathogen. *MBio* **10**
39. Chick, J. M., Kolippakkam, D., Nusinow, D. P., Zhai, B., Rad, R., Huttlin, E. L., and Gygi, S. P. (2015) A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. *Nat. Biotechnol.* **33**, 743–749
40. Na, S., Bandeira, N., and Paek, E. (2012) Fast multi-blind modification search through tandem mass spectrometry. *Mol. Cell. Proteomics* **11**, M111010199
41. Devabhaktuni, A., Lin, S., Zhang, L., Swaminathan, K., Gonzalez, C. G., Olsson, N., Pearlman, S. M., Rawson, K., and Elias, J. E. (2019) TagGraph reveals vast protein modification landscapes from large tandem mass spectrometry datasets. *Nat. Biotechnol.* **37**, 469–479
42. Solntsev, S. K., Shortreed, M. R., Frey, B. L., and Smith, L. M. (2018) Enhanced global post-translational modification discovery with MetaMorpheus. *J. Proteome Res.* **17**, 1844–1851
43. Kong, A. T., Leprevost, F. V., Avtonomov, D. M., Mellacheruvu, D., and Nesvizhskii, A. I. (2017) MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. *Nat. Methods.* **14**, 513–520
44. Lenco, J., Khalikova, M. A., and Svec, F. (2020) Dissolving peptides in 0.1% formic acid brings risk of artificial formylation. *J. Proteome Res.* **19**, 993–999
45. Muller, T., and Winter, D. (2017) Systematic evaluation of protein reduction and alkylation reveals massive unspecific side effects by iodine-containing reagents. *Mol. Cell. Proteomics* **16**, 1173–1187
46. Trinidad, J. C., Schoepfer, R., Burlingame, A. L., and Medzihradsky, K. F. (2013) N- and O-glycosylation in the murine synaptosome. *Mol. Cell. Proteomics* **12**, 3474–3488
47. Medzihradsky, K. F., Kaasik, K., and Chalkley, R. J. (2015) Tissue-specific glycosylation at the glycopeptide level. *Mol. Cell. Proteomics* **14**, 2103–2110
48. Li, Q., Shortreed, M. R., Wenger, C. D., Frey, B. L., Schaffer, L. V., Scalf, M., and Smith, L. M. (2017) Global post-translational modification discovery. *J. Proteome Res.* **16**, 1383–1390
49. Cveticic, N., Semanjski, M., Soufi, B., Krug, K., Gruic-Sovolj, I., and Macek, B. (2016) Proteome-wide measurement of non-canonical bacterial mistranslation by quantitative mass spectrometry of protein modifications. *Sci. Rep.* **6**, 28631
50. Lassak, J., Keilhauer, E. C., Fürst, M., Wuichet, K., Gödeke, J., Starosta, A. L., Chen, J.-M., Søgaard-Andersen, L., Rohr, J., Wilson, D. N., Häussler, S., Mann, M., and Jung, K. (2015) Arginine-rhamnosylation as new strategy to activate translation elongation factor P. *Nat. Chem. Biol.* **11**, 266–270
51. Bern, M., Cai, Y., and Goldberg, D. (2007) Lookup peaks: a hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry. *Anal. Chem.* **79**, 1393–1400
52. Iwashkiw, J. A., Seper, A., Weber, B. S., Scott, N. E., Vinogradov, E., Stratillo, C., Reiz, B., Cordwell, S. J., Whittall, R., Schild, S., and Feldman, M. F. (2012) Identification of a general O-linked protein glycosylation system in *Acinetobacter baumannii* and its role in virulence and biofilm formation. *PLoS Pathog.* **8**, e1002758
53. Lees-Miller, R. G., Iwashkiw, J. A., Scott, N. E., Seper, A., Vinogradov, E., Schild, S., and Feldman, M. F. (2013) A common pathway for O-linked protein-glycosylation and synthesis of capsule in *Acinetobacter baumannii*. *Mol. Microbiol.* **89**, 816–830
54. Bzdyl, N. M., Scott, N. E., Norville, I. H., Scott, A. E., Atkins, T., Pang, S., Sarovich, D. S., Coombs, G., Inglis, T. J. J., Kahler, C. M., and Sarkar-Tyson, M. (2019) Peptidyl-prolyl isomerase ppiB is essential for proteome homeostasis and virulence in *Burkholderia pseudomallei*. *Infect. Immun.* **87**
55. Rappsilber, J., Mann, M., and Ishihama, Y. (2007) Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. *Nat. Protoc.* **2**, 1896–1906

56. Saba, J., Dutta, S., Hemenway, E., and Viner, R. (2012) Increasing the productivity of glycopeptides analysis by using higher-energy collision dissociation-accurate mass-product-dependent electron transfer dissociation. *Int. J. Proteomics*. **2012**, 560391
57. Lee, L. Y., Moh, E. S., Parker, B. L., Bern, M., Packer, N. H., and Thaysen-Andersen, M. (2016) Toward automated N-glycopeptide identification in glycoproteomics. *J. Proteome Res.* **15**, 3904–3915
58. Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T., Mann, M., and Cox, J. (2016) The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods*. **13**, 731–740
59. Perez-Riverol, Y., Csordas, A., Bai, J., Bernal-Llinares, M., Hewapathirana, S., Kundu, D. J., Inuganti, A., Griss, J., Mayer, G., Eisenacher, M., Pérez, E., Uszkoreit, J., Pfeuffer, J., Sachsenberg, T., Yilmaz, S., Tiwary, S., Cox, J., Audain, E., Walzer, M., Jamuczak, A. F., Ternent, T., Brazma, A., and Vizcaino, J. A. (2019) The PRIDE database and related tools and resources in 2019: improving support for quantification data. *Nucleic Acids Res.* **47**, D442–D50
60. Vizcaino, J. A., Csordas, A., del-Toro, N., Dianas, J. A., Griss, J., Lavidas, I., Mayer, G., Perez-Riverol, Y., Reisinger, F., Ternent, T., Xu, Q.-W., Wang, R., and Hermjakob, H. (2016) 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44**, D447–D456
61. Oliveira, L. M., Resende, D. M., Dorneles, E. M. S., Horácio, E. C. A., Alves, F. L., Gonçalves, L. O., Tavares, G. S., Stynen, A. P. R., Lage, A. P., and Ruiz, J. C. (2016) Complete genome sequence of type strain *Campylobacter fetus* subsp. *fetus* ATCC 27374. *Genome Announc* **4**
62. Shin, B., Jung, H.-J., Hyung, S.-W., Kim, H., Lee, D., Lee, C., Yu, M.-H., and Lee, S.-W. (2008) Postexperiment monoisotopic mass filtering and refinement (PE-MMR) of tandem mass spectrometric data increases accuracy of peptide identification in LC/MS/MS. *Mol. Cell. Proteomics* **7**, 1124–1134
63. Avtonomov, D. M., Kong, A., and Nesvizhskii, A. I. (2019) DeltaMass: automated detection and visualization of mass shifts in proteomic open-search results. *J. Proteome Res.* **18**, 715–720
64. Khatri, K., Klein, J. A., and Zaia, J. (2017) Use of an informed search space maximizes confidence of site-specific assignment of glycoprotein glycosylation. *Anal. Bioanal. Chem.* **409**, 607–618
65. Smith, M. G., Gianoulis, T. A., Pukatzki, S., Mekalanos, J. J., Ornston, L. N., Gerstein, M., and Snyder, M. (2007) New insights into *Acinetobacter baumannii* pathogenesis revealed by high-density pyrosequencing and transposon mutagenesis. *Genes Dev.* **21**, 601–614
66. Holden, M. T. G., Seth-Smith, H. M. B., Crossman, L. C., Sebahia, M., Bentley, S. D., Cerdeño-Tárraga, A. M., Thomson, N. R., Bason, N., Quail, M. A., Sharp, S., Cherevach, I., Churcher, C., Goodhead, I., Hauser, H., Holroyd, N., Mungall, K., Scott, P., Walker, D., White, B., Rose, H., Iversen, P., Mil-Homens, D., Rocha, E. P. C., Fialho, A. M., Baldwin, A., Dowson, C., Barrell, B. G., Govan, J. R., Vandamme, P., Hart, C. A., Mahenthiralingam, E., and Parkhill, J. (2009) The genome of *Burkholderia cenocepacia* J2315, an epidemic pathogen of cystic fibrosis patients. *J. Bacteriol.* **191**, 261–277
67. Parker, B. L., Thaysen-Andersen, M., Solis, N., Scott, N. E., Larsen, M. R., Graham, M. E., Packer, N. H., and Cordwell, S. J. (2013) Site-specific glycan-peptide analysis for determination of N-glycoproteome heterogeneity. *J. Proteome Res.* **12**, 5791–5800
68. Riley, N. M., Hebert, A. S., Westphal, M. S., and Coon, J. J. (2019) Capturing site-specific heterogeneity with large-scale N-glycoproteome analysis. *Nat. Commun.* **10**, 1311
69. Darula, Z., and Medzhiradzky, K. F. (2018) 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
70. Hinneburg, H., Stavenhagen, K., Schweiger-Hufnagel, U., Pengelley, S., Jabs, W., Seeberger, P. H., Silva, D. V., Wuhrer, M., and Kolarich, D. (2016) The art of destruction: optimizing collision energies in quadrupole-time of flight (Q-TOF) instruments for glycopeptide-based glycoproteomics. *J. Am. Soc. Mass Spectrom.* **27**, 507–519
71. Alagesan, K., Hinneburg, H., Seeberger, P. H., Silva, D. V., and Kolarich, D. (2019) Glycan size and attachment site location affect electron transfer dissociation (ETD) fragmentation and automated glycopeptide identification. *Glycoconj. J.* **36**, 487–493
72. Scott, N. E., Marzook, N. B., Cain, J. A., Solis, N., Thaysen-Andersen, M., Djordjevic, S. P., Packer, N. H., Larsen, M. R., and Cordwell, S. J. (2014) Comparative proteomics and glycoproteomics reveal increased N-linked glycosylation and relaxed sequon specificity in *Campylobacter jejuni* NCTC11168 O. *J. Proteome Res.* **13**, 5136–5150
73. Darula, Z., and Medzhiradzky, K. F. (2014) Glycan side reaction may compromise ETD-based glycopeptide identification. *J. Am. Soc. Mass Spectrom.* **25**, 977–987
74. Yu, G., Zhang, Y., Zhang, Z., Song, L., Wang, P., and Chai, W. (2010) Effect and limitation of excess ammonium on the release of O-glycans in reducing forms from glycoproteins under mild alkaline conditions for glycomic and functional analysis. *Anal. Chem.* **82**, 9534–9542
75. Hägglund, P., Matthiesen, R., Elortza, F., Højrup, P., Roepstorff, P., Jensen, O. N., and Bunkenborg, J. (2007) An enzymatic deglycosylation scheme enabling identification of core fucosylated N-glycans and O-glycosylation site mapping of human plasma proteins. *J. Proteome Res.* **6**, 3021–3031
76. Mysling, S., Palmisano, G., Højrup, P., and Thaysen-Andersen, M. (2010) Utilizing ion-pairing hydrophilic interaction chromatography solid phase extraction for efficient glycopeptide enrichment in glycoproteomics. *Anal. Chem.* **82**, 5598–5609
77. Ding, W., Nothhaft, H., Szymanski, C. M., and Kelly, J. (2009) Identification and quantification of glycoproteins using ion-pairing normal-phase liquid chromatography and mass spectrometry. *Mol. Cell. Proteomics* **8**, 2170–2185
78. Lenz, S., Giese, S. H., Fischer, L., and Rappsilber, J. (2018) In-search assignment of monoisotopic peaks improves the identification of cross-linked peptides. *J. Proteome Res.* **17**, 3923–3931
79. Véron Mc, R. (1973) Taxonomic study of the genus *Campylobacter* Sebald and Véron and designation of the neotype strain for the type species, *Campylobacter fetus* (Smith and Taylor) Sebald and Véro. *Int. J. Syst. Bacteriol.* **23**, 122–134
80. Baumann, P., Doudoroff, M., and Stanier, R. Y. (1968) A study of the *Moraxella* group. II. Oxidative-negative species (genus *Acinetobacter*). *J. Bacteriol.* **95**, 1520–1541
81. Holden, M. T. G., Titball, R. W., Peacock, S. J., Cerdeño-Tárraga, A. M., Atkins, T., Crossman, L. C., Pitt, T., Churcher, C., Mungall, K., Bentley, S. D., Sebahia, M., Thomson, N. R., Bason, N., Beacham, I. R., Brooks, K., Brown, K. A., Brown, N. F., Challis, G. L., Cherevach, I., Chillingworth, T., Cronin, A., Crossett, B., Davis, P., DeShazer, D., Feltwell, T., Fraser, A., Hance, Z., Hauser, H., Holroyd, S., Jagels, K., Keith, K. E., Maddison, M., Moule, S., Price, C., Quail, M. A., Rabinowitz, E., Rutherford, K., Sanders, M., Simmonds, M., Songsvilai, S., Stevens, K., Tumapa, S., Vesaratchavest, M., Whitehead, S., Yeats, C., Barrell, B. G., Oyston, P. C. F., and Parkhill, J. (2004) Genomic plasticity of the causative agent of melioidosis, *Burkholderia pseudomallei*. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 14240–14245
82. Vandamme, P., Holmes, B., Vancanneyt, M., Coenye, T., Hoste, B., Coopman, R., Revets, H., Lauwers, S., Gillis, M., Kersters, K., and Govan, J. R. (1997) Occurrence of multiple genomovars of *Burkholderia cepacia* in cystic fibrosis patients and proposal of *Burkholderia multivorans* sp. nov. *Int. J. Syst. Bacteriol.* **47**, 1188–1200
83. Sahl, J. W., Vazquez, A. J., Hall, C. M., Busch, J. D., Tuanyok, A., Mayo, M., Schupp, J. M., Lummis, M., Pearson, T., Shippy, K., Colman, R. E., Allender, C. J., Theobald, V., Sarovich, D. S., Price, E. P., Hutcheson, A., Kortach, J., LiPuma, J. J., Ladner, J., Lovett, S., Koroleva, G., Palacios, G., Limmathurotsakul, D., Wuthiekanun, V., Wongsuwan, G., Currie, B. J., Keim, P., and Wagner, D. M. (2016) The effects of signal erosion and core genome reduction on the identification of diagnostic markers. *mBio* **7**
84. Winsor, G. L., Khaira, B., Van Rossum, T., Lo, R., Whiteside, M. D., and Brinkman, F. S. (2008) The *Burkholderia* Genome Database: facilitating flexible queries and comparative analyses. *Bioinformatics* **24**, 2803–2804
85. Johnson, S. L., Bishop-Lilly, K. A., Ladner, J. T., Daligault, H. E., Davenport, K. W., Jaisle, J., Frey, K. G., Koroleva, G. I., Bruce, D. C., Coyne, S. R., Broomall, S. M., Li, P.-E., Teshima, H., Gibbons, H. S., Palacios, G. F., Rosenzweig, C. N., Redden, C. L., Xu, Y., Minogue, T. D., and Chain, P. S. (2015) Complete genome sequences for 59 *Burkholderia* isolates, both pathogenic and near neighbor. *Genome Announc* **3**
86. Ginther, J. L., Mayo, M., Warrington, S. D., Kaestli, M., Mullins, T., Wagner, D. M., Currie, B. J., Tuanyok, A., and Keim, P. (2015) Identification of *Burkholderia pseudomallei* near-neighbor species in the Northern Territory of Australia. *PLoS Negl. Trop. Dis.* **9**, e0003892