**Open database searching enables the identification and comparison of bacterial glycoproteomes without defining glycan compositions prior to searching**

Ameera Raudah Ahmad Izaham[1] and Nichollas E. Scott[1#]

[1]Department of Microbiology and Immunology, University of Melbourne at the Peter Doherty Institute for Infection and Immunity, Melbourne 3000, Australia

[#]To whom correspondence and requests for materials should be addressed N.E.S (Nichollas.scott@unimelb.edu.au).

**Running title:** glycopeptide identification independent of glycan databases

**ABBREVIATIONS**

mass spectrometry (MS)

Luria Bertani (LB)

phosphate-buffered saline (PBS)

sodium dodecyl sulfate (SDS)

normalized collisional energy (NCE)

Zwitterionic Hydrophilic Interaction Liquid Chromatography (ZIC-HILIC)

Tris(hydroxymethyl)aminomethane (Tris)

Trifluoroacetic acid (TFA)

Dithiothreitol (DTT)

Electron-transfer/higher-energy collision dissociation (EThcD)

higher-energy collision dissociation (HCD)

Collision-induced dissociation (CID)

Automatic Gain Control (AGC)

Dimethyl sulfoxide (DMSO)

2,4-diacetamido-2,4,6 trideoxyglucopyranose (diNAcBac)

peptide spectrum matches (PSMs)

Glucose (Glc)

galactose (Gal)

N-Acetylgalactosamine (GalNAc)

N-Acetylglucosamine (GlcNAc)

N-acetylhexoseamine (HexNAc)

Hexose (Hex)

GlcNAc3NAcA4OAc (2,3-diacetamido-2,3-dideoxy-α-D-glucuronic acid)

**ABSTRACT**

Mass spectrometry has become an indispensable tool for the characterisation 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. While 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 prior to database searching. Utilising this approach, we demonstrate how wide tolerance searching can be used to compare glycan utilisation across bacterial species by examining the glycoproteomes of eight Burkholderia species (*B. pseudomallei; B. multivorans; B. dolosa; B. humptydooensis; 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 peptides sequences. Combined, these results show that open searching is a robust computational approach for the determination of glycan diversity within bacterial proteomes.

## INTRODUCTION

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 proteome 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 utilised 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 recognised as a common modification [16-19]. While 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 utilise 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 non-trivial analytical undertaking. Yet, through advancements in mass spectrometry (MS) [28, 30, 33, 34], these once obscure modifications are increasingly recognisable 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 characterise novel glycosylation systems [23-32]. This dependency 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 characterising 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 datasets [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 utilised to examine chemical modifications such as formylation [44] and mis-alkylation [45] as

well as large modifications such as DNA-peptide crosslinks [43] and eukaryotic glycosylation [46, 47]. Although this approach is effective, it is not without trade-offs being computationally more expensive than traditional searches leading to longer search times [48]. To date, these searches have typically been undertaken using **±**500 Da tolerances [39-43] yet large delta mass windows of ±1000Da [43, 46] and even +3000Da [46, 47] have been reported. Despite the growing application of open database searching in eukaryotic proteomics, few bacterial studies have utilised 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-rhammnosylation [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 prior to database searching. Using Byonic™, which enables both glycopeptide and open database searching [14, 51], we benchmark wide mass open searching on three previously characterised 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 characterised 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% $CO_2$, 5% $O_2$, 85% $N_2$) 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 1.

**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 phosphate-buffered saline (PBS) and bacterial cells collected by scraping. Cells were washed 3 times in PBS to remove media contaminates, then collected by centrifugation at 10,000 x *g* at 4˚C for 10 min and then snap frozen. Snap frozen cell samples were resuspended in 4% SDS, 100mM Tris pH 8.0,

20mM Dithiothreitol (DTT) and boiled at 95˚C with shaking at 2000rpm for 10 min. Samples were clarified by centrifugation at 17,000 x *g* for 10 min, the supernatants then collected, and protein concentration determined by a bicinchoninic acid assay (Thermo Scientific). 1mg 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 x g for 10 min at 0˚C. The precipitated protein pellets were resuspended in 80% ice-cold acetone and precipitated for an additional 4 hours at -20˚C. Samples were centrifuged at 17,000 x *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 $NH_4HCO_3$ then reduced for 1 hour with 20mM DTT followed by being alkylated with 40mM chloroacetamide for 1 hour in the dark. Samples were then digested with Lys-C (1/200 w/w) for 3 hours before being diluted with 5 volumes of 40 mM $NH_4HCO_3$ 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, USA) according to the manufacturer's instructions. tC18 Sep-Pak columns were conditioned with 10 bed volumes of Buffer B (0.1% formic acid, 80% 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 ˚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 0.5cm of 10 µm ZIC-HILIC resin (Millipore, Massachusetts, United States) into p200 tips containing a frit of C8 Empore™ (Sigma) material. Prior to use, the columns were washed with ultra-pure 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 µg/µL (a total of 300 µg 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 non-glycosylated peptides and bound peptides eluted with 10 bed volumes of ultra-pure water. Eluted peptides were dried by vacuum centrifugation and stored at -20 ˚C.

**Reverse phase LC-MS:** ZIC-HILIC enriched samples were re-suspended in Buffer A* and separated using a two-column chromatography set up composed of a PepMap100 C18 20 mm x 75 µm trap and a PepMap C18 500 mm x 75 µm analytical column (Thermo Fisher Scientific). Samples were concentrated onto the trap column at 5 µL/min for 5 minutes with Buffer A (0.1% formic acid, 2% DMSO) and then infused into an Orbitrap Fusion™ Lumos™ Tribrid™ Mass Spectrometer (Thermo Fisher Scientific) at 300 nl/minute via the analytical column using a Dionex Ultimate 3000 UPLC (Thermo Fisher Scientific). 185-minute 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 minutes, then from 28% B to 40% B over 10 minutes, then from 40% B to 100% B over 2 minutes. The composition was held at 100% B for 3 minutes, and then dropped to 2% B over 5

minutes before being held at 2% B for another 15 minutes. The Lumos™ 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 seconds and Orbitrap MS/MS HCD scans of precursors (NCE 30%, maximal injection time of 80 ms, AGC $1*10^5$ with a resolution of 15000). HexNAc oxonium ion (204.087 m/z) product-dependent MS/MS analysis [56] was used to trigger three additional scans of potential glycopeptides; a Orbitrap EThcD scan (NCE 15%, maximal injection time of 250 ms, AGC $2*10^5$ with a resolution of 30000); a ion trap CID scan (NCE 35%, maximal injection time of 40 ms, AGC $5*10^4$) 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 $2*10^5$ with a resolution of 30000). 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 batched processed using Byonic v3.5.3 (Protein Metrics Inc. [14]) with the proteome databases denoted within Table 1. Data was searched on a desktop with two 3.00GHz Intel Xeon Gold 6148 processors, a 2TB SDD 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 while oxidation of methionine was included as a variable modification. A maximum mass precursor tolerance of 5 ppm was allowed while 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 datasets from the same biological samples were combined using R and only glycopeptides with a Byonic score >300 were used for further analysis. This score cut-off 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 datasets. 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 datasets. All mass spectrometry 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 dataset identifier: PXD018587. Data can be accessed using the **username:** reviewer97600@ebi.ac.uk, **Password:** 7dcY3KOm

**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 characterisation 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 prior to 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 utilized 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 β-GlcNAc-α1,3-[GlcNAc1,6-]GlcNAc-α1,4-GalNAc-α1,4-GalNAc-α1,3-diNAcBac (1243.507Da) and β-GlcNAc-α1,3-[Glc1,6-]GlcNAc-α1,4-GalNAc-α1,4-GalNAc-α1,4-diNAcBac (1202.481Da) where diNAcBac is the bacterial specific sugar 2,4-diacetamido-2,4,6 trideoxyglucopyranose [25].

We searched ZIC-HILIC glycopeptide enrichments of *C. fetus fetus* allowing a wildcard mass of 200Da to 1600 Da on asparagine enabling the processing of 3-hour LC-MS runs by open searching in less than 2 hours (Supplementary figure 1A). Examination of the detected modifications by binning the observed delta masses in 0.001Da increments demonstrated a clear cluster of modifications with masses >1000Da (Figure 1A). Within these masses 1242.501Da and 1201.475Da were the most numerous delta masses observed (Figure 1A, Supplementary Table 1) yet these are one Dalton 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 Dalton (Supplementary Figure 2A). 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 (Supplementary Figure 2B). 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 due to errors in mono-isotope assignments.

Surprisingly, our open search also revealed additional glycoforms corresponding to formylated glycans (+27.99Da) as well as a modification corresponding to the loss of a HexNAc (-203.079Da) or Hex (-162.053Da) from the 1243Da or 1202Da glycans respectively (Figure 1B). MS/MS analysis supports these delta masses as unexpected but bona fide glycoforms (Supplementary Figure 3A to J). Formylated glycans have been previously observed [25, 28] during ZIC-HILIC enrichment

and are most likely artefacts due to the high concentrations of formic acid [44] used during enrichment. These formylated glycans represented a significant proportion of all potential glycopeptide PSMs (Figure 1B, Supplementary Table 1). Consistent with glycan formylation being artifactual, it is not observed on C*. fetus fetus* glycopeptides within unenriched samples (Supplementary Figure 4). 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] (Figure 1C and D). We find the open search defined mass of the 1243Da and 1202Da 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 (1040.423Da, 1068.419Da, 1202.475Da, 1230.469Da, 1243.501Da, 1271.497Da, 1299.492Da, Figure 1B) as a focused search [64]. Consistent with previous studies focused searches outperformed open database searches [39, 46, 47] improving the identification of unique glycopeptides by 35% and glycoproteins by 28% (Figure 1D, Supplementary Table 2). 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 (Supplementary figure 5). 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, Supplementary Table 2). Combined, these results demonstrate open searching allows the detection of heterogeneous bacterial *N*-linked glycopeptides without the need to define glycans prior to searching.

**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-(β-GlcNAc-6-)-α-Gal-6-β-Glc-3-β-GalNAc where GlcNAc3NAcA4OAc corresponds to the bacterial sugar 2,3-diacetamido-2,3-dideoxy-α-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 respectively [26, 52]). *A. baumannii* glycopeptide enrichments were searched allowing a wildcard mass of 200Da 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 hours (Supplementary figure 1B). 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 (Figure 2A, Supplementary Table 3). These novel glycan masses are consistent with formylation (+27.99Da) as well as the loss of HexNAc (-203.079Da) from the 1030Da glycan with MS/MS analysis supporting these assignments (Supplementary Figure 6).

Examination of these masses revealed the most numerous delta masses (1029.362 Da, 987.355 Da and 1043.378 Da) were one Dalton less than the expected *A. baumannii* glycan masses (Figure

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 Dalton (Supplementary figure 7). Using the masses 1030.368 Da, 988.357 Da, 1044.383 Da, 827.281 Da and 1058.358 Da, we researched these *A. baumannii* datasets 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 (Figure 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 (Figure 2D, Supplementary table 4). 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 (Figure 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 (Supplementary Figure 8). Similar to *C. fetus fetus*, a large proportion of glycopeptide PSMs were identified with formylated glycans (>309 out of the total 2282 glycopeptide PSMs, Supplementary Table 4). It is important to note that the delta masses of multiply glycosylated peptides fall outside the 2000 Da window used for open 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 β-Gal-(1,3)-α-GalNAc-(1,3)-β-GalNAc and Suc-β-Gal-(1,3)-α-GalNAc-(1,3)-β-GalNAc where Suc is Succinyl with these glycans corresponding to the masses 568.211Da and 668.228Da 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 ~20hours to process (Supplementary figure 1C). These open searches revealed the presence of the expected glycoforms of *B. cenocepacia* (568.207Da and 668.223Da) as well as additional formylated variants (596.202Da, 624.197, and 696.218Da) leading to the identification of five unique glycoforms (Figure 3A, Supplementary Table 5). 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 (Figure 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 to open searches (Figure 3B, Supplementary Table 6). Although the improvement in 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 to assignments from open searches (Supplementary Figure 9). 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, Supplementary figure 10). In contrast, focused searches identified ~40% of all PSMs (Figure 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 (Figure 3C, Supplementary table 6). This 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 utilisation 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 utilise 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. humptydooensis 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 568Da and 668Da glycoforms as well as their formylated variants are present in all strains (Figure 4A and Supplementary Figure 11, Supplementary Tables 7 to 14). Having generated "delta mass fingerprints" for each species, we assessed if these profiles could enable the comparison of glycan utilisation using Pearson correlation and hierarchical clustering (Figure 4B and Supplementary Figure 12). Consistent with the similarities in the delta mass fingerprints Pearson correlation and hierarchical clustering resulted in the grouping of all Burkholderia species compared to the delta mass fingerprints of *C. fetus fetus* and *A. baumannii* (Figure 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 datasets, focused searches significantly improved the identification of glycopeptides and glycoproteins in all Burkholderia species (Supplementary figure 13, Supplementary Table 15 to 22).

**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 utilization 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 568/668Da 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 568/668Da glycans (Figure 5A). Examination of these delta masses readily revealed the presence of PSMs matching the modification of peptides with single (203.077Da) or double (406.158Da) HexNAc residues, two 568Da glycans (1136.422Da) and an unexpected mass at 990.390Da (Figure 5A). Examination of PSMs assigned to this 990Da 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 (Figure 5B). Incorporation of this unexpected glycan mass into a focused search with the known Burkholderia glycans demonstrated that the 990.390Da glycan is observed on multiple peptide substrates yet less than 6% of all glycopeptide PSMs correspond to this novel glycan (Figure 5C). Thus, this demonstrates that open searching provides an effective means to detect unexpected glycoforms which could be overlooked due to the low frequency of their occurrence in glycoproteomic datasets.

## 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 utilize 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 prior to 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 artefacts 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 non-similar glycan profiles such as those seen in *C. fetus fetus* or *A. baumannii*.

Within this work we utilized 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 a number of 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 to open searches for most datasets (Supplementary figure 5, 8 and 9). 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 non-optimized glycopeptide analysis can lead to a reduction in glycoproteome coverage [64]. Although Byonic was used within this study it should be noted alternative non-commercial 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 (Supplementary Figure 14A to F). Yet, as with our open Byonic searches MSfragger did not identify as many unique glycopeptides / glycoproteins as focused Byonic searches (Supplementary Figure 15A to F). These results demonstrate that open searching can be used to identify glycopeptides, yet due to the unique challenges associated with glycopeptide identification open searches can be less sensitive than focused searches.

Although we have utilised 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 utilizes 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 (Supplementary Figure 16). 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 while infrequently observed glycans, such as the 990Da 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 datasets. Although we utilized 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 artefacts on a substantial number of glycopeptide PSMs (Supplementary Table 2, 4, 6, 15 to 22) 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. While 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], while 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 datasets 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* datasets 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 to focused searches discussed above, supports that open searching is a useful discovery tool yet typically under reports unique glycopeptides and glycoproteins within datasets. The simplest solution to this issue is to use open searching as a means to identify glycans which can then be included as variables modification 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.

**ACKNOWLEDGEMENTS**

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

**DATA AVAILABILITY**

All raw data is available through the PRIDE repository, PRIDE accession: PXD018587.

**username:** reviewer97600@ebi.ac.uk, **Password:** 7dcY3KOm

**REFERENCES:**

1. Struwe WB, Robinson CV. Relating glycoprotein structural heterogeneity to function - insights from native mass spectrometry. Curr Opin Struct Biol. 2019;58:241-8. Epub 2019/07/22. doi: 10.1016/j.sbi.2019.05.019. PubMed PMID: 31326232; PubMed Central PMCID: PMCPMC7104348.
2. Brockhausen I, Stanley P. O-GalNAc Glycans. In: rd, Varki A, Cummings RD, Esko JD, Stanley P, Hart GW, et al., editors. Essentials of Glycobiology. Cold Spring Harbor (NY)2015. p. 113-23.
3. Stanley P, Taniguchi N, Aebi M. N-Glycans. In: rd, Varki A, Cummings RD, Esko JD, Stanley P, Hart GW, et al., editors. Essentials of Glycobiology. Cold Spring Harbor (NY)2015. p. 99-111.
4. Khoury GA, Baliban RC, Floudas CA. Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. Sci Rep. 2011;1. Epub 2011/10/29. doi: 10.1038/srep00090. PubMed PMID: 22034591; PubMed Central PMCID: PMCPMC3201773.
5. Moremen KW, Tiemeyer M, Nairn AV. Vertebrate protein glycosylation: diversity, synthesis and function. Nat Rev Mol Cell Biol. 2012;13(7):448-62. Epub 2012/06/23. doi: 10.1038/nrm3383. PubMed PMID: 22722607; PubMed Central PMCID: PMCPMC3934011.
6. Xu C, Ng DT. Glycosylation-directed quality control of protein folding. Nat Rev Mol Cell Biol. 2015;16(12):742-52. Epub 2015/10/16. doi: 10.1038/nrm4073. PubMed PMID: 26465718.
7. Freeze HH, Hart GW, Schnaar RL. Glycosylation Precursors. In: rd, Varki A, Cummings RD, Esko JD, Stanley P, Hart GW, et al., editors. Essentials of Glycobiology. Cold Spring Harbor (NY)2015. p. 51-63.
8. Kawano S, Hashimoto K, Miyama T, Goto S, Kanehisa M. Prediction of glycan structures from gene expression data based on glycosyltransferase reactions. Bioinformatics.

2005;21(21):3976-82. Epub 2005/09/15. doi: 10.1093/bioinformatics/bti666. PubMed PMID: 16159923.
9. Campbell MP, Royle L, Radcliffe CM, Dwek RA, Rudd PM. GlycoBase and autoGU: tools for HPLC-based glycan analysis. Bioinformatics. 2008;24(9):1214-6. Epub 2008/03/18. doi: 10.1093/bioinformatics/btn090. PubMed PMID: 18344517.
10. Bohm M, Bohne-Lang A, Frank M, Loss A, Rojas-Macias MA, Lutteke T. Glycosciences.DB: an annotated data collection linking glycomics and proteomics data (2018 update). Nucleic Acids Res. 2019;47(D1):D1195-D201. Epub 2018/10/26. doi: 10.1093/nar/gky994. PubMed PMID: 30357361; PubMed Central PMCID: PMCPMC6323918.
11. McDonald AG, Tipton KF, Davey GP. A Knowledge-Based System for Display and Prediction of O-Glycosylation Network Behaviour in Response to Enzyme Knockouts. PLoS Comput Biol. 2016;12(4):e1004844. Epub 2016/04/08. doi: 10.1371/journal.pcbi.1004844. PubMed PMID: 27054587; PubMed Central PMCID: PMCPMC4824424.
12. Akune Y, Lin CH, Abrahams JL, Zhang J, Packer NH, Aoki-Kinoshita KF, et al. Comprehensive analysis of the N-glycan biosynthetic pathway using bioinformatics to generate UniCorn: A theoretical N-glycan structure database. Carbohydr Res. 2016;431:56-63. Epub 2016/06/19. doi: 10.1016/j.carres.2016.05.012. PubMed PMID: 27318307.
13. Hu H, Khatri K, Klein J, Leymarie N, Zaia J. A review of methods for interpretation of glycopeptide tandem mass spectral data. Glycoconj J. 2016;33(3):285-96. Epub 2015/11/28. doi: 10.1007/s10719-015-9633-3. PubMed PMID: 26612686; PubMed Central PMCID: PMCPMC4882288.
14. Bern M, Kil YJ, Becker C. Byonic: advanced peptide and protein identification software. Curr Protoc Bioinformatics. 2012;Chapter 13:Unit13 20. Epub 2012/12/21. doi: 10.1002/0471250953.bi1320s40. PubMed PMID: 23255153; PubMed Central PMCID: PMCPMC3545648.
15. Liu MQ, Zeng WF, Fang P, Cao WQ, Liu C, Yan GQ, et al. pGlyco 2.0 enables precision N-glycoproteomics with comprehensive quality control and one-step mass spectrometry for intact glycopeptide identification. Nat Commun. 2017;8(1):438. Epub 2017/09/07. doi: 10.1038/s41467-017-00535-2. PubMed PMID: 28874712; PubMed Central PMCID: PMCPMC5585273.
16. Nothaft H, Szymanski CM. New discoveries in bacterial N-glycosylation to expand the synthetic biology toolbox. Curr Opin Chem Biol. 2019;53:16-24. Epub 2019/07/10. doi: 10.1016/j.cbpa.2019.05.032. PubMed PMID: 31288139.
17. Szymanski CM, Wren BW. Protein glycosylation in bacterial mucosal pathogens. Nature reviews Microbiology. 2005;3(3):225-37. doi: 10.1038/nrmicro1100. PubMed PMID: 15738950.
18. Koomey M. O-linked protein glycosylation in bacteria: snapshots and current perspectives. Curr Opin Struct Biol. 2019;56:198-203. Epub 2019/05/13. doi: 10.1016/j.sbi.2019.03.020. PubMed PMID: 31078896.
19. Joshi HJ, Narimatsu Y, Schjoldager KT, Tytgat HLP, Aebi M, Clausen H, et al. SnapShot: O-Glycosylation Pathways across Kingdoms. Cell. 2018;172(3):632- e2. Epub 2018/01/27. doi: 10.1016/j.cell.2018.01.016. PubMed PMID: 29373833.
20. Schaffer C, Messner P. Emerging facets of prokaryotic glycosylation. FEMS microbiology reviews. 2017;41(1):49-91. Epub 2016/08/28. doi: 10.1093/femsre/fuw036. PubMed PMID: 27566466; PubMed Central PMCID: PMCPMC5266552.
21. Macek B, Forchhammer K, Hardouin J, Weber-Ban E, Grangeasse C, Mijakovic I. Protein post-translational modifications in bacteria. Nature reviews Microbiology.

2019;17(11):651-64. Epub 2019/09/06. doi: 10.1038/s41579-019-0243-0. PubMed PMID: 31485032.
22. Imperiali B. Bacterial carbohydrate diversity - a Brave New World. Curr Opin Chem Biol. 2019;53:1-8. Epub 2019/06/09. doi: 10.1016/j.cbpa.2019.04.026. PubMed PMID: 31176085; PubMed Central PMCID: PMCPMC6893104.
23. Fathy Mohamed Y, Scott NE, Molinaro A, Creuzenet C, Ortega X, Lertmemongkolchai G, et al. A general protein O-glycosylation machinery conserved in Burkholderia species improves bacterial fitness and elicits glycan immunogenicity in humans. The Journal of biological chemistry. 2019. Epub 2019/07/28. doi: 10.1074/jbc.RA119.009671. PubMed PMID: 31350337.
24. Harding CM, Nasr MA, Kinsella RL, Scott NE, Foster LJ, Weber BS, et al. Acinetobacter strains carry two functional oligosaccharyltransferases, one devoted exclusively to type IV pilin, and the other one dedicated to O-glycosylation of multiple proteins. Molecular microbiology. 2015;96(5):1023-41. Epub 2015/03/03. doi: 10.1111/mmi.12986. PubMed PMID: 25727908.
25. Nothaft H, Scott NE, Vinogradov E, Liu X, Hu R, Beadle B, et al. Diversity in the protein N-glycosylation pathways within the Campylobacter genus. Molecular & cellular proteomics : MCP. 2012;11(11):1203-19. doi: 10.1074/mcp.M112.021519. PubMed PMID: 22859570; PubMed Central PMCID: PMC3494190.
26. Scott NE, Kinsella RL, Edwards AV, Larsen MR, Dutta S, Saba J, et al. Diversity within the O-linked protein glycosylation systems of acinetobacter species. Molecular & cellular proteomics : MCP. 2014;13(9):2354-70. doi: 10.1074/mcp.M114.038315. PubMed PMID: 24917611; PubMed Central PMCID: PMC4159654.
27. Scott NE, Nothaft H, Edwards AV, Labbate M, Djordjevic SP, Larsen MR, et al. Modification of the Campylobacter jejuni N-linked glycan by EptC protein-mediated addition of phosphoethanolamine. The Journal of biological chemistry. 2012;287(35):29384-96. Epub 2012/07/05. doi: 10.1074/jbc.M112.380212. PubMed PMID: 22761430; PubMed Central PMCID: PMCPMC3436159.
28. Scott NE, Parker BL, Connolly AM, Paulech J, Edwards AV, Crossett B, et al. Simultaneous glycan-peptide characterization using hydrophilic interaction chromatography and parallel fragmentation by CID, higher energy collisional dissociation, and electron transfer dissociation MS applied to the N-linked glycoproteome of Campylobacter jejuni. Molecular & cellular proteomics : MCP. 2011;10(2):M000031-MCP201. doi: 10.1074/mcp.M000031-MCP201. PubMed PMID: 20360033; PubMed Central PMCID: PMCPMC3033663.
29. Hadjineophytou C, Anonsen JH, Wang N, Ma KC, Viburiene R, Vik A, et al. Genetic determinants of genus-Level glycan diversity in a bacterial protein glycosylation system. PLoS Genet. 2019;15(12):e1008532. Epub 2019/12/24. doi: 10.1371/journal.pgen.1008532. PubMed PMID: 31869330.
30. Ulasi GN, Creese AJ, Hui SX, Penn CW, Cooper HJ. Comprehensive mapping of O-glycosylation in flagellin from Campylobacter jejuni 11168: A multienzyme differential ion mobility mass spectrometry approach. Proteomics. 2015;15(16):2733-45. Epub 2015/04/18. doi: 10.1002/pmic.201400533. PubMed PMID: 25884275; PubMed Central PMCID: PMCPMC4975691.
31. Jervis AJ, Wood AG, Cain JA, Butler JA, Frost H, Lord E, et al. Functional analysis of the Helicobacter pullorum N-linked protein glycosylation system. Glycobiology.

2018;28(4):233-44. Epub 2018/01/18. doi: 10.1093/glycob/cwx110. PubMed PMID: 29340583; PubMed Central PMCID: PMCPMC6025236.
32. Jervis AJ, Butler JA, Lawson AJ, Langdon R, Wren BW, Linton D. Characterization of the structurally diverse N-linked glycans of Campylobacter species. Journal of bacteriology. 2012;194(9):2355-62. Epub 2012/03/06. doi: 10.1128/JB.00042-12. PubMed PMID: 22389484; PubMed Central PMCID: PMCPMC3347071.
33. Madsen JA, Ko BJ, Xu H, Iwashkiw JA, Robotham SA, Shaw JB, et al. Concurrent automated sequencing of the glycan and peptide portions of O-linked glycopeptide anions by ultraviolet photodissociation mass spectrometry. Anal Chem. 2013;85(19):9253-61. doi: 10.1021/ac4021177. PubMed PMID: 24006841; PubMed Central PMCID: PMCPMC3816934.
34. Zampronio CG, Blackwell G, Penn CW, Cooper HJ. Novel glycosylation sites localized in Campylobacter jejuni flagellin FlaA by liquid chromatography electron capture dissociation tandem mass spectrometry. Journal of proteome research. 2011;10(3):1238-45. Epub 2010/12/17. doi: 10.1021/pr101021c. PubMed PMID: 21158479; PubMed Central PMCID: PMCPMC3049644.
35. Cain JA, Dale AL, Niewold P, Klare WP, Man L, White MY, et al. Proteomics reveals multiple phenotypes associated with N-linked glycosylation in Campylobacter jejuni. Molecular & cellular proteomics : MCP. 2019. Epub 2019/01/09. doi: 10.1074/mcp.RA118.001199. PubMed PMID: 30617158.
36. Elhenawy W, Scott NE, Tondo ML, Orellano EG, Foster LJ, Feldman MF. Protein O-linked glycosylation in the plant pathogen Ralstonia solanacearum. Glycobiology. 2016;26(3):301-11. doi: 10.1093/glycob/cwv098. PubMed PMID: 26531228; PubMed Central PMCID: PMCPMC4736539.
37. Lithgow KV, Scott NE, Iwashkiw JA, Thomson EL, Foster LJ, Feldman MF, et al. A general protein O-glycosylation system within the Burkholderia cepacia complex is involved in motility and virulence. Molecular microbiology. 2014;92(1):116-37. doi: 10.1111/mmi.12540. PubMed PMID: 24673753.
38. Abouelhadid S, North SJ, Hitchen P, Vohra P, Chintoan-Uta C, Stevens M, et al. Quantitative Analyses Reveal Novel Roles for N-Glycosylation in a Major Enteric Bacterial Pathogen. MBio. 2019;10(2). Epub 2019/04/25. doi: 10.1128/mBio.00297-19. PubMed PMID: 31015322; PubMed Central PMCID: PMCPMC6478998.
39. Chick JM, Kolippakkam D, Nusinow DP, Zhai B, Rad R, Huttlin EL, et al. A mass-tolerant database search identifies a large proportion of unassigned spectra in shotgun proteomics as modified peptides. Nat Biotechnol. 2015;33(7):743-9. doi: 10.1038/nbt.3267. PubMed PMID: 26076430; PubMed Central PMCID: PMCPMC4515955.
40. Na S, Bandeira N, Paek E. Fast multi-blind modification search through tandem mass spectrometry. Molecular & cellular proteomics : MCP. 2012;11(4):M111 010199. Epub 2011/12/22. doi: 10.1074/mcp.M111.010199. PubMed PMID: 22186716; PubMed Central PMCID: PMCPMC3322561.
41. Devabhaktuni A, Lin S, Zhang L, Swaminathan K, Gonzalez CG, Olsson N, et al. TagGraph reveals vast protein modification landscapes from large tandem mass spectrometry datasets. Nat Biotechnol. 2019;37(4):469-79. Epub 2019/04/03. doi: 10.1038/s41587-019-0067-5. PubMed PMID: 30936560; PubMed Central PMCID: PMCPMC6447449.
42. Solntsev SK, Shortreed MR, Frey BL, Smith LM. Enhanced Global Post-translational Modification Discovery with MetaMorpheus. Journal of proteome research. 2018;17(5):1844-51. Epub 2018/03/27. doi: 10.1021/acs.jproteome.7b00873. PubMed PMID: 29578715.

43. Kong AT, Leprevost FV, Avtonomov DM, Mellacheruvu D, Nesvizhskii AI. MSFragger: ultrafast and comprehensive peptide identification in mass spectrometry-based proteomics. Nature methods. 2017;14(5):513-20. Epub 2017/04/11. doi: 10.1038/nmeth.4256. PubMed PMID: 28394336; PubMed Central PMCID: PMCPMC5409104.
44. Lenco J, Khalikova MA, Svec F. Dissolving Peptides in 0.1% Formic Acid Brings Risk of Artificial Formylation. Journal of proteome research. 2020;19(3):993-9. Epub 2020/02/19. doi: 10.1021/acs.jproteome.9b00823. PubMed PMID: 32068400.
45. Muller T, Winter D. Systematic Evaluation of Protein Reduction and Alkylation Reveals Massive Unspecific Side Effects by Iodine-containing Reagents. Molecular & cellular proteomics : MCP. 2017;16(7):1173-87. Epub 2017/05/26. doi: 10.1074/mcp.M116.064048. PubMed PMID: 28539326; PubMed Central PMCID: PMCPMC5500753.
46. Trinidad JC, Schoepfer R, Burlingame AL, Medzihradszky KF. N- and O-glycosylation in the murine synaptosome. Molecular & cellular proteomics : MCP. 2013;12(12):3474-88. Epub 2013/07/03. doi: 10.1074/mcp.M113.030007. PubMed PMID: 23816992; PubMed Central PMCID: PMCPMC3861701.
47. Medzihradszky KF, Kaasik K, Chalkley RJ. Tissue-Specific Glycosylation at the Glycopeptide Level. Molecular & cellular proteomics : MCP. 2015;14(8):2103-10. Epub 2015/05/23. doi: 10.1074/mcp.M115.050393. PubMed PMID: 25995273; PubMed Central PMCID: PMCPMC4528240.
48. Li Q, Shortreed MR, Wenger CD, Frey BL, Schaffer LV, Scalf M, et al. Global Post-Translational Modification Discovery. Journal of proteome research. 2017;16(4):1383-90. Epub 2017/03/02. doi: 10.1021/acs.jproteome.6b00034. PubMed PMID: 28248113; PubMed Central PMCID: PMCPMC5387672.
49. Cvetesic N, Semanjski M, Soufi B, Krug K, Gruic-Sovulj I, Macek B. Proteome-wide measurement of non-canonical bacterial mistranslation by quantitative mass spectrometry of protein modifications. Sci Rep. 2016;6:28631. doi: 10.1038/srep28631. PubMed PMID: 27377007.
50. Lassak J, Keilhauer EC, Furst M, Wuichet K, Godeke J, Starosta AL, et al. Arginine-rhamnosylation as new strategy to activate translation elongation factor P. Nat Chem Biol. 2015;11(4):266-70. doi: 10.1038/nchembio.1751. PubMed PMID: 25686373; PubMed Central PMCID: PMCPMC4451828.
51. Bern M, Cai Y, Goldberg D. Lookup peaks: a hybrid of de novo sequencing and database search for protein identification by tandem mass spectrometry. Anal Chem. 2007;79(4):1393-400. Epub 2007/01/25. doi: 10.1021/ac0617013. PubMed PMID: 17243770.
52. Iwashkiw JA, Seper A, Weber BS, Scott NE, Vinogradov E, Stratilo C, et al. Identification of a general O-linked protein glycosylation system in Acinetobacter baumannii and its role in virulence and biofilm formation. PLoS pathogens. 2012;8(6):e1002758. doi: 10.1371/journal.ppat.1002758. PubMed PMID: 22685409; PubMed Central PMCID: PMC3369928.
53. Lees-Miller RG, Iwashkiw JA, Scott NE, Seper A, Vinogradov E, Schild S, et al. A common pathway for O-linked protein-glycosylation and synthesis of capsule in Acinetobacter baumannii. Molecular microbiology. 2013;89(5):816-30. Epub 2013/06/21. doi: 10.1111/mmi.12300. PubMed PMID: 23782391.
54. Bzdyl NM, Scott NE, Norville IH, Scott AE, Atkins T, Pang S, et al. Peptidyl-Prolyl Isomerase ppiB Is Essential for Proteome Homeostasis and Virulence in Burkholderia

pseudomallei. Infection and immunity. 2019;87(10). Epub 2019/07/25. doi: 10.1128/IAI.00528-19. PubMed PMID: 31331957; PubMed Central PMCID: PMCPMC6759293.
55. Rappsilber J, Mann M, Ishihama Y. Protocol for micro-purification, enrichment, pre-fractionation and storage of peptides for proteomics using StageTips. Nature protocols. 2007;2(8):1896-906. doi: 10.1038/nprot.2007.261. PubMed PMID: 17703201.
56. Saba J, Dutta S, Hemenway E, Viner R. Increasing the productivity of glycopeptides analysis by using higher-energy collision dissociation-accurate mass-product-dependent electron transfer dissociation. Int J Proteomics. 2012;2012:560391. doi: 10.1155/2012/560391. PubMed PMID: 22701174; PubMed Central PMCID: PMCPMC3369405.
57. Lee LY, Moh ES, Parker BL, Bern M, Packer NH, Thaysen-Andersen M. Toward Automated N-Glycopeptide Identification in Glycoproteomics. Journal of proteome research. 2016;15(10):3904-15. Epub 2016/08/16. doi: 10.1021/acs.jproteome.6b00438. PubMed PMID: 27519006.
58. Tyanova S, Temu T, Sinitcyn P, Carlson A, Hein MY, Geiger T, et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. Nature methods. 2016;13(9):731-40. Epub 2016/06/28. doi: 10.1038/nmeth.3901. PubMed PMID: 27348712.
59. Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, Kundu DJ, et al. The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic Acids Res. 2019;47(D1):D442-D50. Epub 2018/11/06. doi: 10.1093/nar/gky1106. PubMed PMID: 30395289; PubMed Central PMCID: PMCPMC6323896.
60. Vizcaino JA, Csordas A, del-Toro N, Dianes JA, Griss J, Lavidas I, et al. 2016 update of the PRIDE database and its related tools. Nucleic Acids Res. 2016;44(D1):D447-56. doi: 10.1093/nar/gkv1145. PubMed PMID: 26527722; PubMed Central PMCID: PMCPMC4702828.
61. Oliveira LM, Resende DM, Dorneles EM, Horacio EC, Alves FL, Goncalves LO, et al. Complete Genome Sequence of Type Strain Campylobacter fetus subsp. fetus ATCC 27374. Genome Announc. 2016;4(6). Epub 2016/12/17. doi: 10.1128/genomeA.01344-16. PubMed PMID: 27979934; PubMed Central PMCID: PMCPMC5159567.
62. Shin B, Jung HJ, Hyung SW, Kim H, Lee D, Lee C, et al. Postexperiment monoisotopic mass filtering and refinement (PE-MMR) of tandem mass spectrometric data increases accuracy of peptide identification in LC/MS/MS. Molecular & cellular proteomics : MCP. 2008;7(6):1124-34. Epub 2008/02/28. doi: 10.1074/mcp.M700419-MCP200. PubMed PMID: 18303012.
63. Avtonomov DM, Kong A, Nesvizhskii AI. DeltaMass: Automated Detection and Visualization of Mass Shifts in Proteomic Open-Search Results. Journal of proteome research. 2019;18(2):715-20. Epub 2018/12/14. doi: 10.1021/acs.jproteome.8b00728. PubMed PMID: 30523686.
64. Khatri K, Klein JA, Zaia J. Use of an informed search space maximizes confidence of site-specific assignment of glycoprotein glycosylation. Anal Bioanal Chem. 2017;409(2):607-18. Epub 2016/10/14. doi: 10.1007/s00216-016-9970-5. PubMed PMID: 27734143; PubMed Central PMCID: PMCPMC5283608.
65. Smith MG, Gianoulis TA, Pukatzki S, Mekalanos JJ, Ornston LN, Gerstein M, et al. New insights into Acinetobacter baumannii pathogenesis revealed by high-density pyrosequencing and transposon mutagenesis. Genes Dev. 2007;21(5):601-14. Epub 2007/03/09. doi: 10.1101/gad.1510307. PubMed PMID: 17344419; PubMed Central PMCID: PMCPMC1820901.
66. Holden MT, Seth-Smith HM, Crossman LC, Sebaihia M, Bentley SD, Cerdeno-Tarraga AM, et al. The genome of Burkholderia cenocepacia J2315, an epidemic pathogen of cystic

fibrosis patients. Journal of bacteriology. 2009;191(1):261-77. Epub 2008/10/22. doi: 10.1128/JB.01230-08. PubMed PMID: 18931103; PubMed Central PMCID: PMCPMC2612433.
67. Parker BL, Thaysen-Andersen M, Solis N, Scott NE, Larsen MR, Graham ME, et al. Site-specific glycan-peptide analysis for determination of N-glycoproteome heterogeneity. Journal of proteome research. 2013;12(12):5791-800. Epub 2013/10/05. doi: 10.1021/pr400783j. PubMed PMID: 24090084.
68. Riley NM, Hebert AS, Westphall MS, Coon JJ. Capturing site-specific heterogeneity with large-scale N-glycoproteome analysis. Nat Commun. 2019;10(1):1311. Epub 2019/03/23. doi: 10.1038/s41467-019-09222-w. PubMed PMID: 30899004; PubMed Central PMCID: PMCPMC6428843.
69. Darula Z, Medzihradszky KF. Analysis of Mammalian O-Glycopeptides-We Have Made a Good Start, but There is a Long Way to Go. Molecular & cellular proteomics : MCP. 2018;17(1):2-17. Epub 2017/11/23. doi: 10.1074/mcp.MR117.000126. PubMed PMID: 29162637; PubMed Central PMCID: PMCPMC5750848.
70. Hinneburg H, Stavenhagen K, Schweiger-Hufnagel U, Pengelley S, Jabs W, Seeberger PH, et al. The Art of Destruction: Optimizing Collision Energies in Quadrupole-Time of Flight (Q-TOF) Instruments for Glycopeptide-Based Glycoproteomics. J Am Soc Mass Spectrom. 2016;27(3):507-19. Epub 2016/01/06. doi: 10.1007/s13361-015-1308-6. PubMed PMID: 26729457; PubMed Central PMCID: PMCPMC4756043.
71. Alagesan K, Hinneburg H, Seeberger PH, Silva DV, Kolarich D. Glycan size and attachment site location affect electron transfer dissociation (ETD) fragmentation and automated glycopeptide identification. Glycoconj J. 2019;36(6):487-93. Epub 2019/10/23. doi: 10.1007/s10719-019-09888-w. PubMed PMID: 31637569.
72. Scott NE, Marzook NB, Cain JA, Solis N, Thaysen-Andersen M, Djordjevic SP, et al. Comparative proteomics and glycoproteomics reveal increased N-linked glycosylation and relaxed sequon specificity in Campylobacter jejuni NCTC11168 O. Journal of proteome research. 2014;13(11):5136-50. Epub 2014/08/06. doi: 10.1021/pr5005554. PubMed PMID: 25093254.
73. Darula Z, Medzihradszky KF. Glycan side reaction may compromise ETD-based glycopeptide identification. J Am Soc Mass Spectrom. 2014;25(6):977-87. Epub 2014/03/26. doi: 10.1007/s13361-014-0852-9. PubMed PMID: 24664807; PubMed Central PMCID: PMCPMC4036456.
74. Yu G, Zhang Y, Zhang Z, Song L, Wang P, Chai W. 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. 2010;82(22):9534-42. Epub 2010/10/30. doi: 10.1021/ac102300r. PubMed PMID: 21028831.
75. Hagglund P, Matthiesen R, Elortza F, Hojrup P, Roepstorff P, Jensen ON, et al. An enzymatic deglycosylation scheme enabling identification of core fucosylated N-glycans and O-glycosylation site mapping of human plasma proteins. Journal of proteome research. 2007;6(8):3021-31. Epub 2007/07/20. doi: 10.1021/pr0700605. PubMed PMID: 17636988.
76. Mysling S, Palmisano G, Hojrup P, Thaysen-Andersen M. Utilizing ion-pairing hydrophilic interaction chromatography solid phase extraction for efficient glycopeptide enrichment in glycoproteomics. Anal Chem. 2010;82(13):5598-609. Epub 2010/06/12. doi: 10.1021/ac100530w. PubMed PMID: 20536156.
77. Ding W, Nothaft H, Szymanski CM, Kelly J. Identification and quantification of glycoproteins using ion-pairing normal-phase liquid chromatography and mass spectrometry.

Molecular & cellular proteomics : MCP. 2009;8(9):2170-85. Epub 2009/06/16. doi: 10.1074/mcp.M900088-MCP200. PubMed PMID: 19525481; PubMed Central PMCID: PMCPMC2742440.
78. Lenz S, Giese SH, Fischer L, Rappsilber J. In-Search Assignment of Monoisotopic Peaks Improves the Identification of Cross-Linked Peptides. Journal of proteome research. 2018;17(11):3923-31. Epub 2018/10/09. doi: 10.1021/acs.jproteome.8b00600. PubMed PMID: 30293428; PubMed Central PMCID: PMCPMC6279313.
79. VÉRON MC, R. 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. INTERNATIONAL JOURNAL OF SYSTEMATIC BACTERIOLOGY. 1973;23(3):122-34.
80. Baumann P, Doudoroff M, Stanier RY. A study of the Moraxella group. II. Oxidative-negative species (genus Acinetobacter). Journal of bacteriology. 1968;95(5):1520-41. Epub 1968/05/01. PubMed PMID: 5650064; PubMed Central PMCID: PMCPMC252171.
81. Holden MT, Titball RW, Peacock SJ, Cerdeno-Tarraga AM, Atkins T, Crossman LC, et al. Genomic plasticity of the causative agent of melioidosis, Burkholderia pseudomallei. Proceedings of the National Academy of Sciences of the United States of America. 2004;101(39):14240-5. Epub 2004/09/21. doi: 10.1073/pnas.0403302101. PubMed PMID: 15377794; PubMed Central PMCID: PMCPMC521101.
82. Vandamme P, Holmes B, Vancanneyt M, Coenye T, Hoste B, Coopman R, et al. Occurrence of multiple genomovars of Burkholderia cepacia in cystic fibrosis patients and proposal of Burkholderia multivorans sp. nov. Int J Syst Bacteriol. 1997;47(4):1188-200. Epub 1997/10/23. doi: 10.1099/00207713-47-4-1188. PubMed PMID: 9336927.
83. Sahl JW, Vazquez AJ, Hall CM, Busch JD, Tuanyok A, Mayo M, et al. The Effects of Signal Erosion and Core Genome Reduction on the Identification of Diagnostic Markers. mBio. 2016;7(5). Epub 2016/09/22. doi: 10.1128/mBio.00846-16. PubMed PMID: 27651357; PubMed Central PMCID: PMCPMC5030356.
84. Winsor GL, Khaira B, Van Rossum T, Lo R, Whiteside MD, Brinkman FS. The Burkholderia Genome Database: facilitating flexible queries and comparative analyses. Bioinformatics. 2008;24(23):2803-4. Epub 2008/10/10. doi: 10.1093/bioinformatics/btn524. PubMed PMID: 18842600; PubMed Central PMCID: PMCPMC2639269.
85. Johnson SL, Bishop-Lilly KA, Ladner JT, Daligault HE, Davenport KW, Jaissle J, et al. Complete genome sequences for 59 burkholderia isolates, both pathogenic and near neighbor. Genome Announc. 2015;3(2). Epub 2015/05/02. doi: 10.1128/genomeA.00159-15. PubMed PMID: 25931592; PubMed Central PMCID: PMCPMC4417688.
86. Ginther JL, Mayo M, Warrington SD, Kaestli M, Mullins T, Wagner DM, et al. Identification of Burkholderia pseudomallei Near-Neighbor Species in the Northern Territory of Australia. PLoS Negl Trop Dis. 2015;9(6):e0003892. Epub 2015/06/30. doi: 10.1371/journal.pntd.0003892. PubMed PMID: 26121041; PubMed Central PMCID: PMCPMC4486726.

**Table 1. Strain list**

| Strains | Source (Description, Country, Year) | Reference | Proteome database |
|---|---|---|---|
| *C. fetus subsp. fetus NCTC 10842* | Brain of sheep fetus, France, 1952 | [79] | Uniprot: 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: UP000000605 |
| *Burkholderia cenocepacia* (LMG 16656 / J2315) | Human clinical specimen, United Kingdom, 1989 | [82] | Uniprot: UP000001035 |
| *Burkholderia multivorans* MSMB2008 | Soil isolate, Australia, 2012 | [83] | Burkholderia Genome Database [84], Strain number: 3016 |
| *Burkholderia dolosa* AU0158 | Human clinical specimen, USA unknown | [85] | Uniprot database: UP000032886 |
| *Burkholderia humptydooensis* 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 |

**FIGURE LEGENDS**

**Figure 1: Open searching analysis of *C. fetus fetus* NCTC 10842 glycopeptides. A)** *C. fetus fetus* glycopeptide delta mass plot of 0.001 Dalton 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 non-formylated glycans while black masses correspond to formylated glycans. **C)** Density and zoomed delta mass plot of the *C. fetus fetus* glycan masses 1243.507Da and 1202.481Da. **D)** Comparison of the unique glycopeptide sequences and glycoproteins observed between open and focused searches across *C. fetus fetus* datasets.

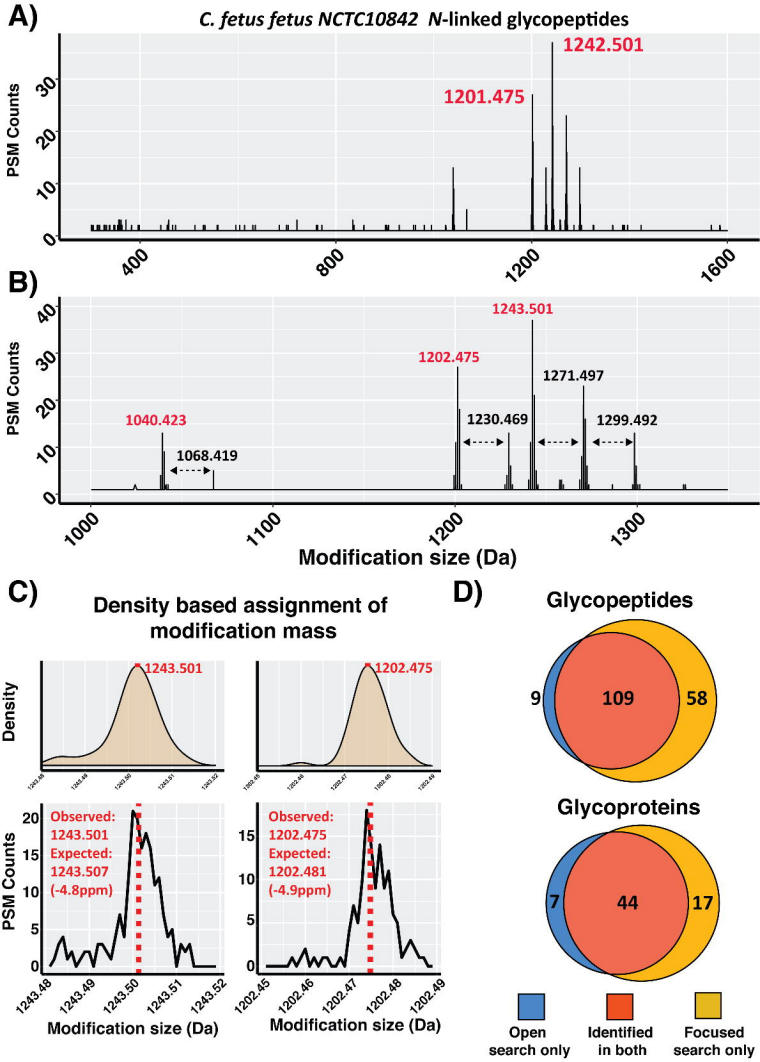
**Figure 2: Open searching analysis of *A. baumannii ATCC17978* glycopeptides. A)** *A. baumannii* glycopeptide delta mass plot of 0.001 Dalton increments showing the detection of PSMs modified with masses over 800 Da. **B)** Zoomed view of *A. baumannii* glycopeptide delta mass plot highlighting the most numerous observed modifications. **C)** Comparison of unique glycopeptide sequences and glycoproteins observed between open and focused searches across *A. baumannii* datasets **D)** Glycan mass plot showing the amount of glycan (in Da) observed on glycopeptides 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.

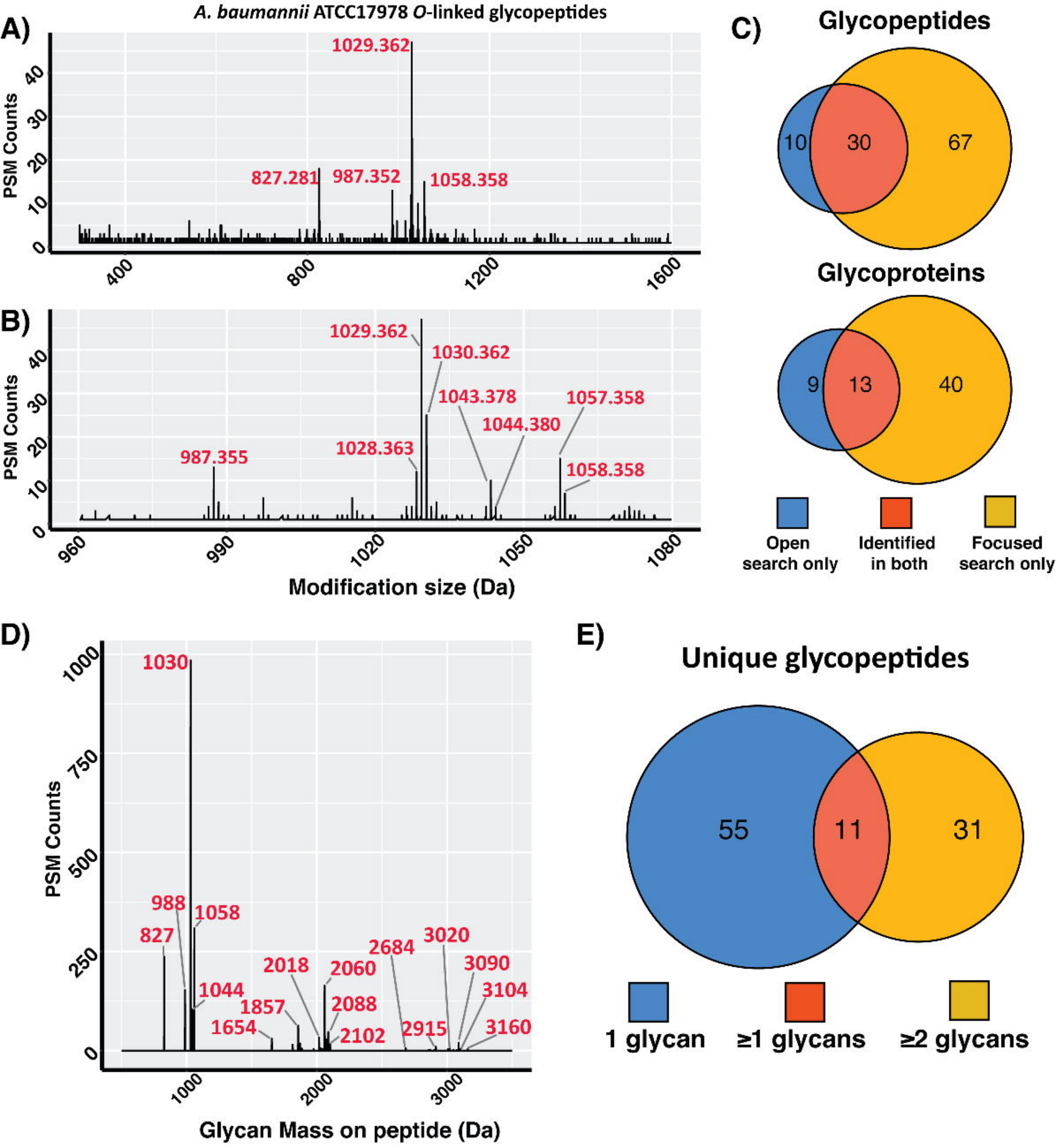
**Figure 3: Open searching analysis of *B. Cenocepacia* J2315 glycopeptides. A)** *B. Cenocepacia* glycopeptide delta mass plot of 0.001 Dalton increments showing the detection of PSMs modified with masses over 500 Da. Highlighted area shown in zoomed panel. **B)** Comparison of the unique glycopeptide sequences and glycoproteins observed between open and focused searches across *B. Cenocepacia* datasets **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.

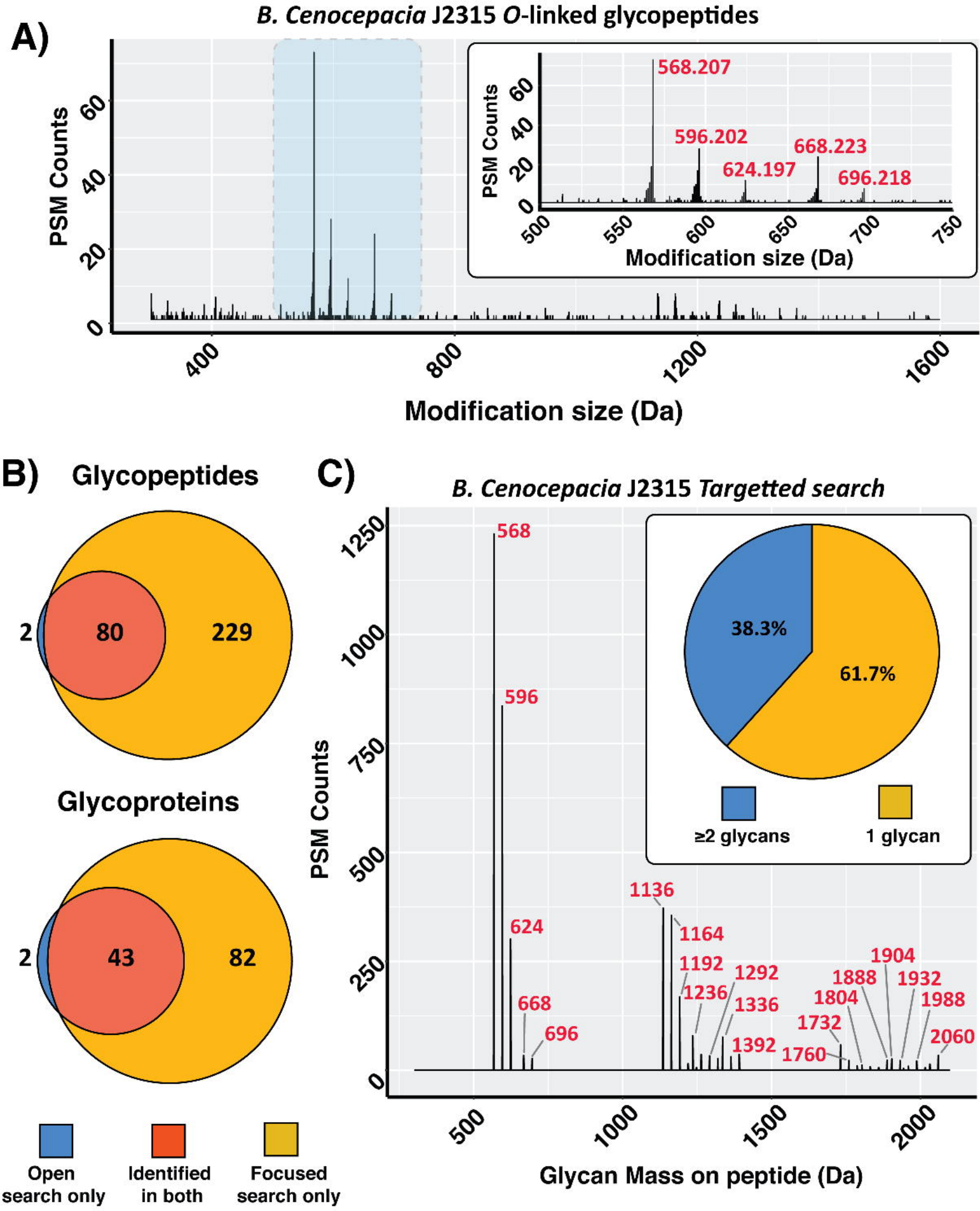
**Figure 4: Comparison of Burkholderia glycoproteomes using open searching. A)** Representative delta mass plots of four out of the eight Burkholderia strains examined demonstrating the 568Da and 668Da glycans are frequently identified delta masses in Burkholderia glycopeptide enrichments. Formylated glycans are denoted in black while Burkholderia O-linked glycans are in

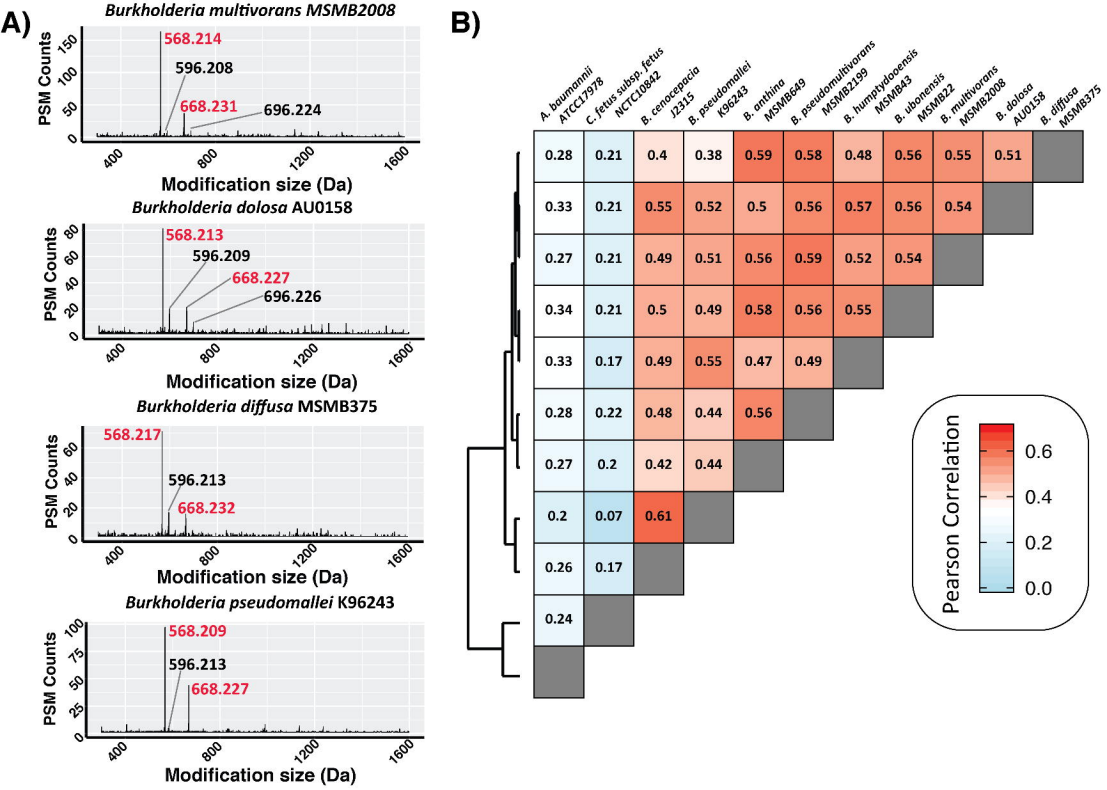
red. **B)** Pearson correlation and clustering analysis of delta mass plots enable the comparison and grouping of samples.

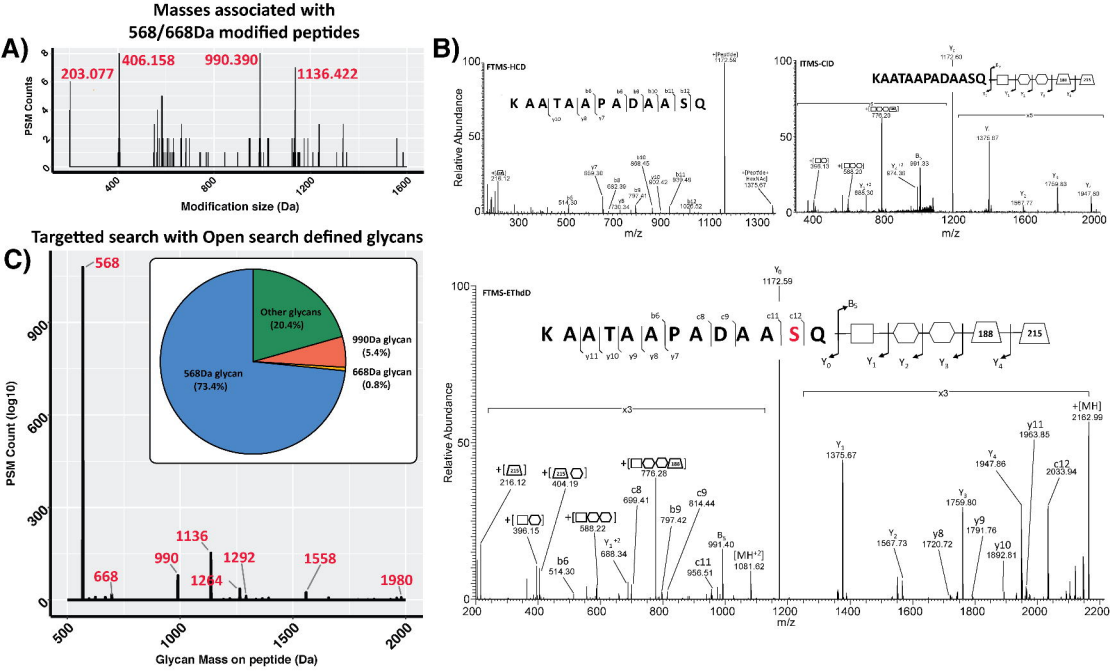
**Figure 5: Identification of minor glycoforms within *B. pseudomallei* K96243. A)** Delta mass plot, binned by 0.001 Dalton increments, showing delta masses observed for peptide sequences also modified with the 568 or 668D glycans. **B)** MS/MS analysis (FTMS-HCD, ITMS-CID and FTMS-EThcD) supporting 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.


A)
C. fetus fetus NCTC10842 N-linked glycopeptides
PSM Counts
1242.501
1201.475
B)
PSM Counts
1243.501
1202.475
1271.497
1040.423
1230.469
1299.492
1068.419
Modification size (Da)
C)
Density based assignment of modification mass
Density
1243.501
1202.475
PSM Counts
Observed: 1243.501 Expected: 1243.507 (-4.8ppm)
Observed: 1202.475 Expected: 1202.481 (-4.9ppm)
Modification size (Da)
D)
Glycopeptides
9
109
58
Glycoproteins
7
44
17
Open search only
Identified in both
Focused search only


A. baumannii ATCC17978 O-linked glycopeptides
A)
PSM Counts
1029.362
827.281
987.352
1058.358
B)
1029.362
1030.362
1043.378
1057.358
1044.380
1028.363
987.355
1058.358
Modification size (Da)
C)
Glycopeptides
10
30
67
Glycoproteins
9
13
40
Open search only
Identified in both
Focused search only
D)
1030
988
1058
827
1044
1857
1654
2018
2060
2088
2102
2684
2915
3020
3090
3104
3160
Glycan Mass on peptide (Da)
E)
Unique glycopeptides
55
11
31
1 glycan
≥1 glycans
≥2 glycans


A)
B. Cenocepacia J2315 O-linked glycopeptides
PSM Counts
Modification size (Da)
568.207
596.202
624.197
668.223
696.218
B)
Glycopeptides
2
80
229
Glycoproteins
2
43
82
Open search only
Identified in both
Focused search only
C)
B. Cenocepacia J2315 Targetted search
PSM Counts
Glycan Mass on peptide (Da)
568
596
624
668
696
1136
1164
1192
1236
1292
1336
1392
1732
1760
1804
1888
1904
1932
1988
2060
38.3%
61.7%
≥2 glycans
1 glycan


A)
Burkholderia multivorans MSMB2008
PSM Counts
568.214
596.208
668.231
696.224
Modification size (Da)
Burkholderia dolosa AU0158
568.213
596.209
668.227
696.226
Burkholderia diffusa MSMB375
568.217
596.213
668.232
Burkholderia pseudomallei K96243
568.209
596.213
668.227
B)
A. baumannii ATCC17978
C. fetus subsp. fetus NCTC10842
B. cenocepacia J2315
B. pseudomallei K96243
B. anthina MSMB649
B. pseudomultivorans MSMB2199
B. humptydoensis MSMB43
B. ubonensis MSMB22
B. multivorans MSMB2008
B. dolosa AU0158
B. diffusa MSMB375
Pearson Correlation



Masses associated with
568/668Da modified peptides
A)
203.077
406.158
990.390
1136.422
PSM Counts
Modification size (Da)
B)
FTMS-HCD
KAATAAPADAASQ
Relative Abundance
m/z
ITMS-CID
KAATAAPADAASQ
Relative Abundance
m/z
Targetted search with Open search defined glycans
C)
568
668
990
1136
1264
1292
1558
1980
Other glycans
(20.4%)
990Da glycan
(5.4%)
668Da glycan
(0.8%)
568Da glycan
(73.4%)
PSM Count (log10)
Glycan Mass on peptide (Da)
FTMS-EThdD
KAATAAPADAASQ
188
215
Relative Abundance
m/z