Journal of Proteomics xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Journal of Proteomics

journal homepage: www.elsevier.com/locate/jprot

In silico approaches for unveiling novel glycobiomarkers in cancer

Rita Azevedo [a,b], André M.N. Silva [c], Celso A. Reis [b,d,e], Lúcio Lara Santos [a], José Alexandre Ferreira [a,b,d,e,f,g,*]

[a] Experimental Pathology and Therapeutics Group, Research Centre, Portuguese Institute of Oncology of Porto (IPO-Porto), Porto, Portugal
[b] Institute of Biomedical Sciences Abel Salazar, University of Porto, Porto, Portugal
[c] UCIBIO-REQUIMTE, Department of Chemistry and Biochemistry, Faculty of Sciences, University of Porto, Porto, Portugal
[d] Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), Porto, Portugal
[e] Institute for Research and Innovation in Health (i3S), University of Porto, Porto, Portugal
[f] Porto Comprehensive Cancer Center (P.ccc), Porto, Portugal
[g] International Iberian Nanotechnology Laboratory (INL), Braga, Portugal

ARTICLE INFO

Article history:
Received 26 February 2017
Received in revised form 1 August 2017
Accepted 2 August 2017
Available online xxxx

Keywords:
Glycoproteomics
Bioinformatics
Cancer biomarkers
Glycosylation
Biomarker discovery
Glycobiology

ABSTRACT

Glycosylation is one of the most common and dynamic post-translational modification of cell surface and secreted proteins. Cancer cells display unique glycosylation patterns that decisively contribute to drive oncogenic behavior, including disease progression and dissemination. Moreover, alterations in glycosylation are often responsible for the creation of protein signatures holding significant biomarker value and potential for targeted therapeutics. Accordingly, many analytical protocols have been outlined for the identification of abnormally glycosylated proteins by mass spectrometry. Nevertheless, very few studies undergo a comprehensive mining of the generated data. Herein, we build on bladder cancer *O*-glycoproteomics datasets resulting from a hyphenated technique comprising enrichment by *Vicia villosa* agglutinin (VVA) lectin and nanoLC-ESI-MS/MS to propose an *in silico* step-by-step tutorial (Panther, UniProtKB, NetOGlyc, NetNGlyc, Oncomine, Cytoscape) for biomarker discovery in cancer. We envisage that this approach may be generalized to other mass spectrometry-based analytical approaches, including *N*-glycoproteomics studies, and different types of cancers.
Significance: The glycoproteome is an important source of cancer biomarkers holding tremendous potential for targeted therapeutics. We now present an *in silico* roadmap for comprehensive interpretation of big data generated by mass spectrometry-based glycoproteomics envisaging the identification of clinically relevant glycobiomarkers.

1. Introduction

Glycosylation is the most common post-translational modification of cell-surface and secreted proteins and comprises two main classes of modifications: *O*-GalNAc glycans occurring in Ser/Thr residues and *N*-glycans in Asn residues. Nevertheless, other less abundant forms of glycosylation may also be encountered at the cell membrane, including *O*-fucosylation, *O*-glucosylation, *O*- and *C*-manosylation [1]. Glycosylation is determinant for correct protein trafficking, folding, stability and resistance to proteolysis [2]. Moreover, glycans at the cell-surface play a key role in several biological processes including cell recognition, cell-cell and cell-matrix adhesion, mediation of intracellular signaling and interactions with the immune system [2–4]. Changes in glycosylation patterns of membrane proteins have been described as a common feature in cancer, being detected even in pre-malignant lesions [1,5]. Deregulations in glycosylation pathways that may stem from mutations in glycogenes, altered transcription and/or mislocalization of glycosyltransferases throughout the secretory pathway, metabolic shifts leading to unbalanced biosynthesis of sugar donors and several other factors. These events may ultimately influence the lengths and branching of glycan structures [6], promote changes in terminal glycan epitopes [1,7], overfucosylation [8] and oversialylation [9,10]. Particularly, the overexpression of ST6GalNAc-I [9,11] and mutations in *Cosmc* leading to inactivation of T-synthase activity [12], lead to a premature stop in protein *O*-glycosylation, causing the accumulation of short-chain *O*-glycans at tumour cells surface, including the Tn antigen and its sialylated counterpart, sialyl-Tn (STn; (Neu5Acα2-6GalNAc-*O*-Ser/Thr residues in membrane and secreted glycoproteins) [4,9,13–15]. These alterations present a pancarcinoma nature and can be found in the majority of late stage human solid tumours [4,9,16]. The expression of short-chain *O*-glycans affects the adhesive properties of tumour cells favoring motility, invasion and metastatic potential [4,9,17]. Furthermore, this pattern of abnormal *O*-glycosylation has been linked to the activation of key intracellular oncogenic pathways [18], chemoresistance [19], immune tolerance [3,18], having been shown to negatively influence prognosis [20,21]. Moreover, many tumours also (co)overexpress other short-chain

* Corresponding author at: Experimental Pathology and Therapeutics Group, Research Centre, Portuguese Institute of Oncology of Porto (IPO-Porto), Porto, Portugal.
E-mail address: jose.a.ferreira@ipoporto.min-saude.pt (J.A. Ferreira).

http://dx.doi.org/10.1016/j.jprot.2017.08.004
1874-3919/

O-glycans, including sialylated forms of the T antigen (Gal-GalNAc-*O*-Ser/Thr) [22,23]. Notwithstanding, alterations in glycosylation may render cancer-specific protein glycoantigens holding tremendous potential for targeted therapy [24–27].

Given the importance of glycomarkers, an analytical workflow has been introduced for studying the *O*-glycoproteome of cancer cells, enabling the identification of Tn and STn expressing glycoproteins and glycosites. It includes the removal of sialic acids by neuraminidase digestion followed by a pre-enrichment of Tn-expressing glycoproteins based on *Vicia Villosa* agglutinin (VVA) affinity chromatography and subsequent protein identification by liquid chromatography-mass spectrometry (LC-MS). The strategy has been successfully applied to cancer cells of different origins [17,28,29], which has significantly broaden our understanding of the *O*-glycoproteome. However, difficulties persist in pinpointing relevant biomarkers arising from glycoproteomics experiments. This includes challenges related with high-confidence identification of glycopeptides and glycosylation sites using non-ETD fragmentation, which do constitute the routine of many proteomics laboratories. This may either be overcome by the combination of multiple LC-MS/MS runs and different fragmentation strategies [28,30] and/or big data curation using several freeware bioinformatics web-tools such as Panther [31], UniProtKB [32], NetOGlyc [29] and NetNGlyc [33]. The introduction of online tools such as Oncomine [34] and Cytoscape [35] further allows narrowing high number of identified species to glycobiomarkers with real clinical potential.

Recently, we have applied this strategy to the identification of abnormally *O*-glycosylated proteins in bladder cancer [23] and developed an *in silico* tutorial for big data curation and identification of relevant cancer glycobiomarkers. We now build on these procedures to provide a step-by-step practical approach for handling *O*-glycoproteomics information using bladder cancer as a model. We envisage that our protocol may be generalized to datasets generated by alternative methodologies, allowing more accurate assignments and the identification of relevant biomarkers for downstream clinical validation.

2. Methods

2.1. General overview on the generation of mass spectrometry data sets

The present work explores a pre-existing dataset generated by Cotton S, Azevedo R and co-workers [23]. Briefly, mass spectrometry data have been generated from a total protein pool from muscle invasive bladder tumours expressing high levels of the STn and also sialylated T antigens. The glycoproteomics workflow included the digestion with PNGase F followed by digestion with an α-neuraminidase. The samples were then pre-enriched for *O*-GalNAc expressing proteins by VVA affinity chromatography before protein identification. A bottom-up approach was adopted, with tryptic peptides being identified by a conventional high resolution nanoLC-MS experiment using an LTQ-Orbitrap mass spectrometer. Full experimental procedures and MS are detailed in our most recent publications [17,23]. Mass spectrometry proteomics data are available *via* ProteomeXchange [36], where they have been deposited *via* the PRIDE [37] partner repository with the dataset identifier PXD006813.

Here in, we have used the raw data provided to perform a new protein search. Data were analyzed using the SequestHT search engine with the Percolator algorithm for validation of protein identifications (Proteome Discoverer 1.4; Thermo Scientific). The four data files were combined in a single search using the Proteome Discoverer Daemon. Data were searched against the human proteome obtained from the SwissProt database on 22/11/2015, selecting trypsin as the enzyme and allowing for up to two missed cleavage sites, a precursor ion mass tolerance of 10 p.p.m. and 0.6 Da for product ions. Carbamidomethylcysteine was selected as a fixed modification, while oxidation of methionine (+15.994u) and modification of serine and threonine with HexNAc (Tn) (+203.08u), HexHexNAc (T) (+365.13u) were defined as variable modifications. Due to neutral loss or internal fragmentation of the glycan residue, glycated peptides have often poor identification scores under CID conditions. Thus for, Sequest results of low confidence peptides were also considered. Protein grouping filters were set to consider glycosylations with low confidence and ΔCn lower than 0.05. The strict maximum parsimony principle was applied. A protein filter counting peptides only on top-scored proteins was also set. Peptides were filtered for Xcorr ≥ 1.0. The identification of glycopeptides exhibiting the mentioned posttranslational modifications was a pre-requisite for inclusion of glycoproteins in the final list (Supplementary material 1 – sheet "Original dataset"). The generated protein list was curated for relevant bladder cancer glycobiomarkers using bioinformatics tools disclosed in the following sections.

2.2. Curation of glycoproteomics data

For curation of bladder cancer *O*-glycoproteomics dataset, we applied the following freeware bioinformatics web-tools: Panther version 11 (http://www.pantherdb.org/) [31], UniProtKB (http://www.uniprot.org/) [32], NetOGlyc version 4.0 (http://www.cbs.dtu.dk/services/NetOGlyc/) [29], Oncomine (https://www.oncomine.org/) [34] and the ClueGO and CluePedia apps for Cytoscape (http://www.cytoscape.org/) [35,38]. The NetNGlyc version 1.0 (http://www.cbs.dtu.dk/services/NetNGlyc/) [33] was also used as an example for researchers working with a putative *N*-glycosylated proteins list. The *in silico* roadmap for *O*-glycoproteomics curation and identification of relevant cancer glycobiomarkers is outlined in Fig. 1.

3. *In silico* strategies for unveiling novel glycobiomarkers in cancer

Herein, we present an *in silico* roadmap for mining *O*-glycoproteomics datasets for relevant cancer biomarkers holding potential for targeted therapeutics, using bladder cancer as an exemplificative model. Nevertheless, the above described protocol may be adopted to protein datasets generated using other samples or overall analytical approaches. The proposed strategy includes two major steps: i) identification of cell membrane proteins with potential glycosylation sites; and, ii) identification of glycoproteins associated with bladder cancer.

3.1. Identification of cell membrane proteins with glycosylation sites

Conventional mass spectrometry proteomics analytical strategies, such as the one described by us to address the bladder cancer *O*-glycoproteome [23], often use the more commonly available collision induced dissociation (CID) for peptide fragmentation. Under CID conditions, glycopeptides are known to generate a high number of low-confidence identifications. This is the result of preferential glycan fragmentation with only minimal fragmentation of the peptide backbone [39–42]. The main CID MS/MS spectra fragment ions are formed as a result of the cleavage of the glycosidic bonds and the formation of typical oxonium ions. These ions are not commonly taken into account by search engines retrieving missed identifications or poorly scored peptides. A possible analytical strategy is to consider these poorly identified peptides, but it warrants a careful validation of protein identification during data curation. In addition, we often note a high number of cytosolic and nuclear proteins after VVA affinity chromatography, suggesting possible isolation of glycoproteins with *O*-GlcNAc moieties, a post-translational modification commonly found in intracellular but not cell-surface proteins [43,44]. While *O*-GlcNAc class of glycans may also be a source of important cancer biomarkers [43,45], the present analytical strategy focus of glycans present at the cell-surface that may be explored for targeting cancer cells. Consequently, we propose to use Panther, UniProtKB and NetOGlyc for identifying cell surface glycoproteins yielding potentially *O*-GalNAc glycosylation extracellular domains. Accordingly, Panther is a classification system of proteins based in