




OPEN  
ANALYSIS

# The human *O*-GlcNAcome database and meta-analysis

Eugenia Wulff-Fuentes <sup>1,3</sup>, Rex R. Berendt<sup>1,3</sup>, Logan Massman <sup>1</sup>, Laura Danner<sup>1</sup>, Florian Malard<sup>1</sup>, Jeet Vora<sup>2</sup>, Robel Kahsay<sup>2</sup> & Stephanie Olivier-Van Stichelen<sup>1</sup> 

Over the past 35 years, ~1700 articles have characterized protein *O*-GlcNAcylation. Found in almost all living organisms, this post-translational modification of serine and threonine residues is highly conserved and key to biological processes. With half of the primary research articles using human models, the *O*-GlcNAcome recently reached a milestone of 5000 human proteins identified. Herein, we provide an extensive inventory of human *O*-GlcNAcylated proteins, their *O*-GlcNAc sites, identification methods, and corresponding references ([www.oglcnac.mcw.edu](http://www.oglcnac.mcw.edu)). In the absence of a comprehensive online resource for *O*-GlcNAcylated proteins, this list serves as the only database of *O*-GlcNAcylated proteins. Based on the thorough analysis of the amino acid sequence surrounding 7002 *O*-GlcNAc sites, we progress toward a more robust semi-consensus sequence for *O*-GlcNAcylation. Moreover, we offer a comprehensive meta-analysis of human *O*-GlcNAcylated proteins for protein domains, cellular and tissue distribution, and pathways in health and diseases, reinforcing that *O*-GlcNAcylation is a master regulator of cell signaling, equal to the widely studied phosphorylation.

## Introduction

The current trend of sugar-rich food intake poses a significant concern for public health worldwide. In the United States, the average US citizen consumes 17 teaspoons of added sugar every day<sup>1</sup>, whereas the American Heart Association (AHA) recommended upper limit is at most half of that. Increased sugar consumption correlates with increased diet-related diseases like obesity, diabetes, cardiovascular diseases, and other metabolic syndromes<sup>2</sup>. Therefore, understanding the molecular mechanisms through which the excess intake of sugar impacts metabolic regulation is critical to prevent and treat these diseases. Amongst the affected mechanisms is the modification of proteins by *O*-GlcNAcylation<sup>3</sup>.

The *O*-GlcNAc modification is a nutrient rheostat that transiently regulates functions, localization, and stability of proteins in response to fluctuations in nutrient intake (Fig. 1a,b)<sup>4</sup>. Indeed, the nucleotide sugar donor for this modification, UDP-GlcNAc, is the final product of the Hexosamine Biosynthetic Pathway (HBP). This pathway integrates carbohydrate, amino acid, nucleotide, and fatty acid metabolisms to maintain a suitable pool of UDP-GlcNAc. UDP-GlcNAc is then used for glycan synthesis, including *O*-GlcNAcylation catalyzed by the *O*-GlcNAc Transferase (OGT). On the other hand, the *O*-GlcNAcase (OGA) dynamically hydrolyzes *O*-GlcNAc, releasing the protein's modification<sup>4</sup>.

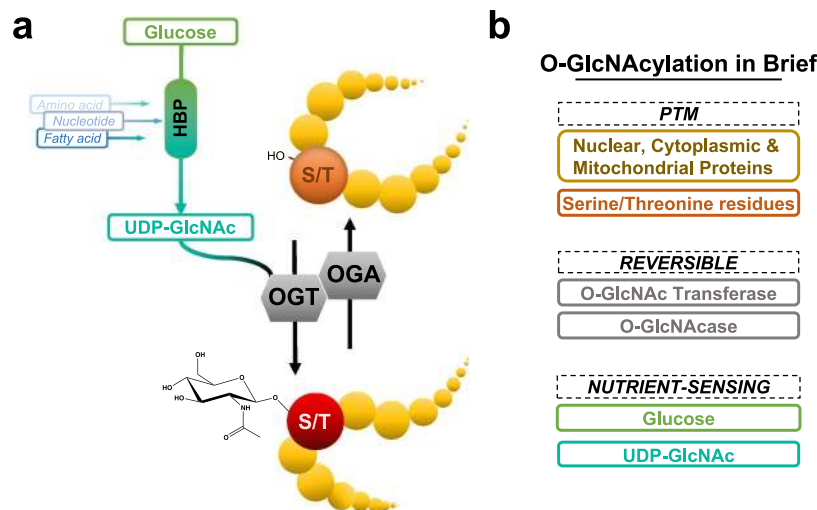
Since its discovery in 1984, *O*-GlcNAc research has been thriving. On average, one article every two days has been published over the past decade (Figure S1a)<sup>5</sup>. However, the absence of an *O*-GlcNAc database has been narrowing our view of the field. Therefore, we combined the results from all *O*-GlcNAc articles published to date and objectively provide a comprehensive picture of this modification. We have created an extensive database of 5072 human *O*-GlcNAcylated proteins (available at [www.oglcnac.mcw.edu](http://www.oglcnac.mcw.edu)) and performed a meta-analysis of surrounding amino acid sequences, protein domains, and molecular functions of these proteins in health and diseases.

## Results

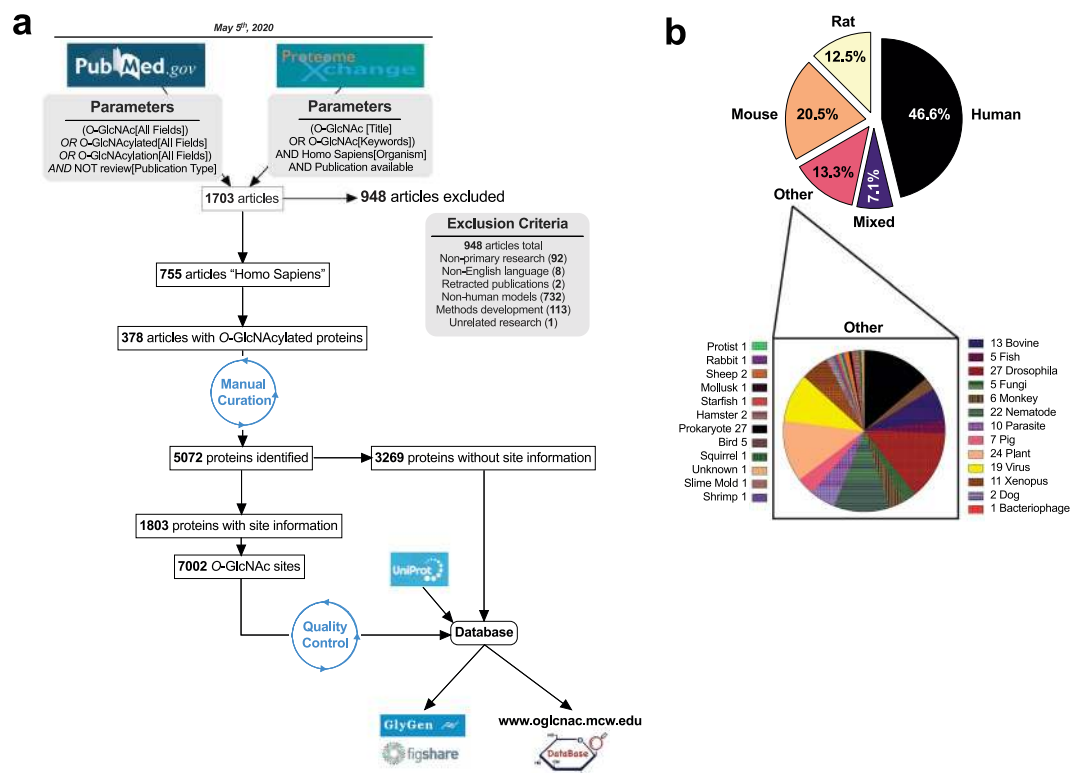
**Over 5000 proteins and 7000 sites compose the human *O*-GlcNAcome.** About 1703 articles were extracted from a combined search on PubMed and Proteome Exchange (Fig. 2a, Figshare file 1<sup>6</sup>). A total of 948 articles were excluded from our list using the following criteria: non-primary research articles, non-English languages, retracted publications, and articles studying non-human proteins. About 85% of *O*-GlcNAc articles

<sup>1</sup>Department of Biochemistry, Medical College of Wisconsin, Milwaukee, USA. <sup>2</sup>Department of Biochemistry & Molecular Medicine, The George Washington School of Medicine and Health Sciences, Washington, DC, 20052, USA.

<sup>3</sup>These authors contributed equally: Eugenia Wulff-Fuentes, Rex R. Berendt. ✉e-mail: [solivier@mcw.edu](mailto:solivier@mcw.edu)



**Fig. 1** The O-GlcNAcylation cycling. **(a)** The O-GlcNAcylation cycling in humans. **(b)** An O-GlcNAcylation in Brief portion highlights the features and actors of O-GlcNAcylation. *HBP*: Hexosamine Biosynthesis Pathway; *OGT*: O-GlcNAc Transferase; *OGA*: O-GlcNAcase; *PTM*: Post-Translational Modification; *S/T*: Serine/Threonine.



**Fig. 2** Literature search for O-GlcNAcylation identified over 5000 human O-GlcNAcylation proteins. **(a)** Method flowchart and exclusion criteria. **(b)** Study models found in the O-GlcNAcylation literature.

were either human-, mouse- or rat-based studies (Fig. 2b). Moreover, we observed O-GlcNAcylation in all seven biological kingdoms, highlighting the ubiquitous and conserved nature of this modification (Fig. 2b).

Of the 755 remaining articles, half of them identified human O-GlcNAcylation proteins (Fig. 2a, Figshare file 1<sup>6</sup>). Each O-GlcNAcylation protein was cataloged using their UniProtKB entry number as the primary identifier. References, techniques used, and O-GlcNAcylation sites were also documented. The list was further manually curated for duplicate entries and merged if identical. When possible, unreviewed UniProtKB entries were merged or replaced with reviewed entries. In the end, 5072 O-GlcNAcylation proteins were inventoried from the O-GlcNAcylation literature (Figshare file 2<sup>6</sup>). A subset of 1803 proteins contained 7002 distinct O-GlcNAcylation sites were identified through varied techniques. Therefore, this inventory is the most extensive library of

O-GlcNAcylated proteins and site information made to date<sup>7</sup>. This list was first deposited to Figshare (#12443495) and GlyGen (#GLY\_000517). In addition, we created an online platform to search the database in a more direct manner ([www.oglcnac.mcw.edu](http://www.oglcnac.mcw.edu)).

For each entry, an O-GlcNAc score between 0 and 100, was attributed to each O-GlcNAcylated protein, reflecting the exhaustiveness of the O-GlcNAcylation description for each protein. The score combined the following parameters: (1) the timespan between the first and the last identification. Having a protein identified 20 years ago and still confirmed in recent analysis is a strong indication of reliability for the O-GlcNAc status; (2) the total number of references showing O-GlcNAcylation for a given protein; (3) the number of independent investigators (first and last authors) that reported the protein as O-GlcNAcylated, which limits the probability for errors; (4) the number of yearly citations for each paper was also added as an estimation of peers-validation of the publication in the field; (5) a modifier that underweighted papers with lots of proteins identified but few citation and overweighted publications that carefully characterized one protein and was well cited. By providing this O-GlcNAc score, we propose an unbiased meta-analysis to score the confidence in the identification of the O-GlcNAcylation on a protein. In our list, proteins with low score (<10) needs to be considered with caution and verified for their O-GlcNAcylation status.

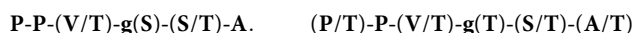
From our inventory, most of the O-GlcNAcylated proteins were identified by MS-based studies, almost always combined with labeling strategies such as click-chemistry (4032/5072 proteins) (Figshare file 2<sup>6</sup>). This confirmed that the majority of the O-GlcNAcylated proteins has been discovered by mass spectrometry (Figure S1b). Indeed, due to technical limitations, O-GlcNAcylation was overlooked for a long time. However, reliable mass-spectrometry protocols lead to proper and efficient mapping of cellular O-GlcNAcylation. Besides, the sub-stoichiometric nature of the modification still necessitates some enrichment before identification, such as lectin or enzymatic labeling and pull down. As a result, of the 1803 proteins with site information, only 15 used non-MS strategies, often opting for *in silico* analysis, pull-down and mutagenesis (Figshare file 2<sup>6</sup>).

### Analysis of human O-GlcNAc sites refined the O-GlcNAcylation semi-consensus sequence.

Over the past decades, 7002 O-GlcNAcylation sites were characterized in 1803 proteins (Fig. 2a, Figshare file 2<sup>6</sup>). Of those, 58.4% were found on serine as opposed to 41.6% threonine residues. However, this follows the relative abundance of serine/threonine in the human proteome (S:60.9%/T:39.1%) (<https://www.uniprot.org/uniprot/?query=proteome:UP000005640>). It suggests that the OGT has little to no preferences toward one residue over the other.

Heavily modified proteins (at least 50 sites) were identified as (1) involved in the formation of stress granules, including UBP2L (88 sites) and PRC2C (68 sites); (2) part of the pre-synaptic interface, including Bassoon (73 sites) and PCLO (62 sites); and (3) part of the nuclear transport complexes, including Host Cell Factor 1 (HCF1) (169 sites), Nucleoporin 214 (Nup214) (177 sites), Nup153 (131 sites), Nup98 (91 sites), Nup62 (87 sites) and Nup58 (53 sites) (Figshare file 2<sup>6</sup>). So far, the latter are the most O-GlcNAcylated proteins described to date.

O-GlcNAc is a signaling regulator more akin to phosphorylation, despite some distinct characteristics. Whereas protein phosphorylation is governed by about 518 specific kinases and about 200 phosphatases<sup>8,9</sup>, only two enzymes regulate the addition and removal of O-GlcNAc from thousands of protein targets. Therefore, while phosphorylation sites can be well predicted based on consensus sequences, O-GlcNAc sites do not have a strict consensus. To refine the preferred sequence for O-GlcNAcylation, we performed a sequence alignment on the 7002 O-GlcNAc sites in this study (Fig. 3/Figshare file 2<sup>6</sup>). An abundance of serine or threonine residues around the O-GlcNAc sites was overall advantageous. On the contrary, the presence of glutamic acid or cysteine was mostly unfavorable to O-GlcNAcylation. The following sequences were extracted as the semi-consensus sequence for both serine and threonine O-GlcNAc modification in human proteins (Fig. 3):

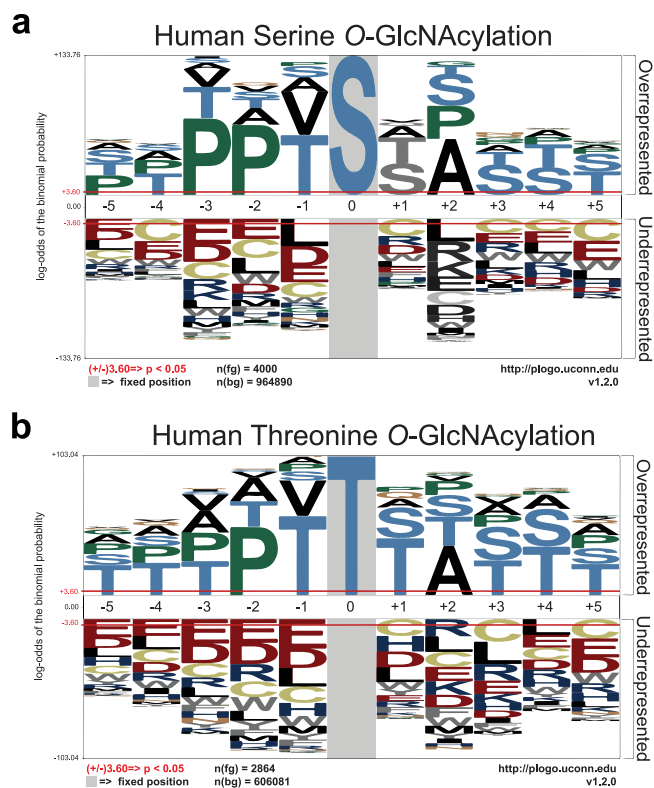


### Cellular distribution analysis confirmed O-GlcNAcylated proteins are concentrated in nuclear and cytoplasmic compartments.

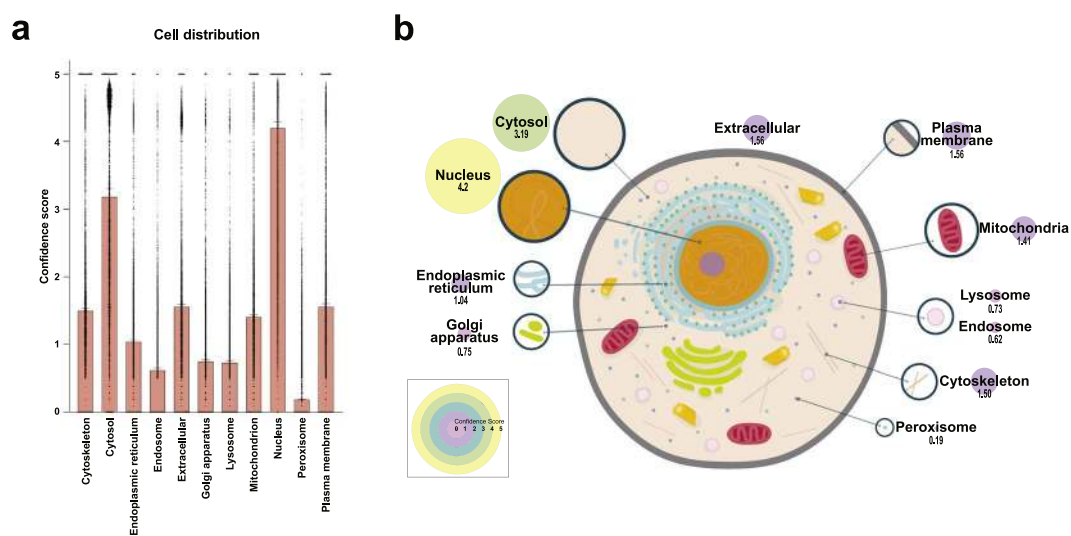
The 5072 O-GlcNAcylated proteins were submitted to the STRING database, and cellular distribution information was extracted, including a confidence score for each protein and compartment ranging from 0 (not present) to 5 (highly significant)<sup>10</sup>. As previously highlighted in the literature, O-GlcNAcylated proteins were mainly nuclear and cytoplasmic, with a median confidence score of 4.20 and 3.19, respectively (Fig. 4, Figshare file 3a<sup>6</sup>). Although some mitochondrial proteins were O-GlcNAcylated<sup>11</sup>, the median confidence score for that compartment was only 1.41. Therefore, only a small percentage of human O-GlcNAcylated proteins are mitochondrial compared to the nucleus or cytoplasm.

### Tissue distribution underlined the strong characterization of brain and liver O-GlcNAcylation.

To understand where O-GlcNAc deregulation would have the most impact, we studied the organ distribution of O-GlcNAcylated proteins<sup>12</sup>. Like cell compartment distribution, a confidence score was attributed for each O-GlcNAcylated protein, and the median was represented for each organ (Fig. 5/Figshare file 3b<sup>6</sup>). O-GlcNAcylation was found in all major organs. However, tissue distribution interestingly correlated with high interest organs in the field as well as the availability of those tissues to a vast number of researchers. Therefore, the most O-GlcNAcylated proteins were identified in the brain and liver, with respective scores of 4.47 and 3.86. This distribution correlated with the tissue distribution of OGT<sup>13,14</sup>. Most organs showed some level O-GlcNAcylated proteins. The least O-GlcNAc abundant sites in the human body were bones (0.78), saliva (1.10), gall bladder (1.12), and urine (1.20). In addition to being rarer and harder to process, those sites are also composed of generally fewer cells, which therefore would be overall limited in intracellular proteins. This analysis emphasized some understudied tissues that could be of high significance such as the pancreas. As new studies emerge, this distribution will be refined through the web interface.



**Fig. 3** O-GlcNAcylation semi-consensus sequence. (a) Serine and (b) Threonine O-GlcNAcylation semi-consensus sequences based on the 7002 human O-GlcNAcylation sites surrounding sequences.

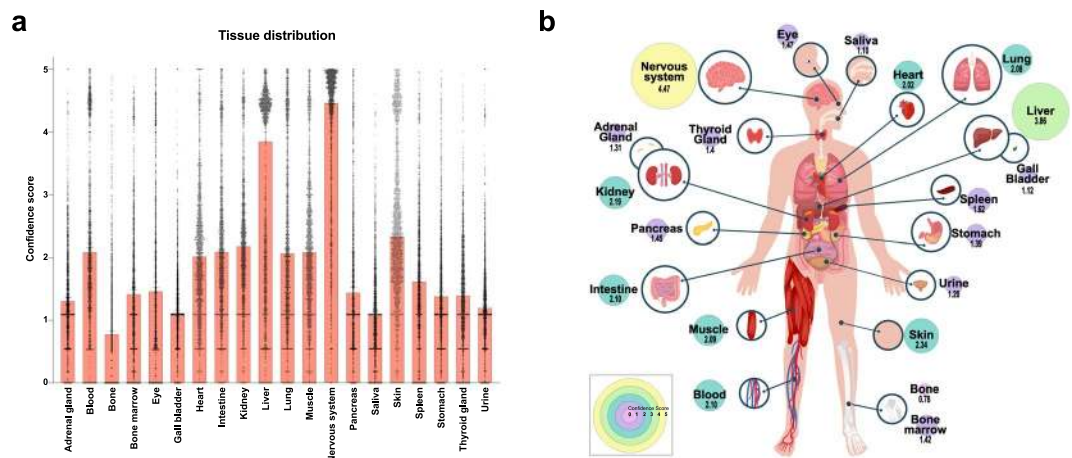


**Fig. 4** Cellular distribution of the O-GlcNAcylated proteins. (a,b) Cellular localization of the human O-GlcNAcylation sites (n = 4969). (a) The median and 95% Confidence intervals are represented. (b) Circle sizes and colors represent the median confidence score.

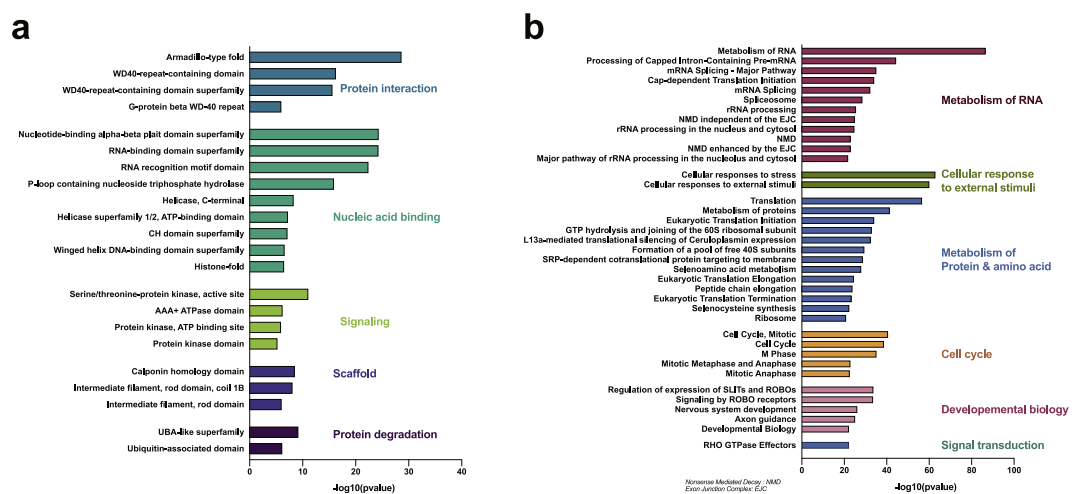
**Enrichment analyses highlighted the role of O-GlcNAcylation in regulating genetic information and cell signaling.** We then performed an enrichment analysis for protein domains (InterPro) and signaling pathways (KEGG and Reactome) on the human O-GlcNAcylated proteins (Fig. 6/Figshare file 4a/b<sup>6</sup>). For both protein domains and pathways, the enrichments highlighted the importance of O-GlcNAcylation in controlling genetic information and regulating protein signaling.

- (1) *Protein interaction and cell signaling*: Protein domains and pathway enrichment emphasized a central role for O-GlcNAcylation in signaling. Indeed, the most enriched domains in human O-GlcNAcylated





**Fig. 5** Tissues distribution of the human *O*-GlcNAcylated proteins (n = 2052). **(a)** The median and 95% Confidence intervals are represented. **(b)** Circle sizes and colors represent the median confidence score.



**Fig. 6** Protein domains and pathways enrichment of the human *O*-GlcNAcylated proteins (n = 5059). **(a)** Protein domain enrichment analysis of the human *O*-GlcNAcylated proteins. **(b)** Pathway enrichment analysis of the human *O*-GlcNAcylated proteins.

proteins were “Protein Interaction” and “Signaling” and included domains such as the Armadillo, WD40 (propeller-like arrangements of 4–8  $\beta$ -folds) and kinase domains (Fig. 6a/Figshare file 3a<sup>6</sup>). Since these domains present great interaction surfaces, proteins with these domains are usually vital for signal transduction. Therefore, *O*-GlcNAcylated proteins in this category were central to signaling pathways and included  $\beta$ -catenin, the regulatory-associated protein of mTOR (RPTOR), Receptor of activated protein C kinase 1 (RACK1), and the histone-binding protein RBBP4 (Figshare file 2<sup>6</sup>). In this category, essential protein transporters like importins were also enriched. In the pathway analysis, cellular response to stimuli or stress, cell cycle, and development were significantly enriched in *O*-GlcNAcylated proteins (Fig. 6b/Figshare file 4b<sup>6</sup>). Indeed, *O*-GlcNAc-modified proteins were involved in cellular responses to various external stimuli, including hypoxia and heat stress or response to amino acid deficiency<sup>15</sup>. Heat-shock proteins (HSPs) constituted part of this category. Finally, many actors of cell cycle and development were also enriched, including Cyclin D, Retinoblastoma protein (RB), Minichromosome Maintenance proteins (MCM3,6,7), Vimentin, Enhancer of Zeste 2 Polycomb Repressive Complex 2 Subunit (EZH2), and Transcriptional repressor protein Yin and yang 1 (YY1) (Figshare file 2<sup>6</sup>).

- (2) *Nucleic acid-binding and metabolism of RNA*: Both protein domain and pathway enrichments emphasized the predominant role of *O*-GlcNAcylation in the control of genetic materials. *O*-GlcNAcylated proteins often presented nucleic acid-binding domains, including *DNA-binding* (histone-fold, TATA-box binding protein (TBP) domain) and *RNA-binding* (nucleotide-binding  $\alpha/\beta$  plait superfamily, RNA recognition motif domain, K homology domain, ribosome, translation protein beta-barrel domain, and the ribosomal protein S5 domain 2-type) (Fig. 6a/Figshare file 4a<sup>6</sup>). Therefore, a large portion of the human *O*-GlcNAcylated proteins contributed to RNA metabolism, as demonstrated by pathway enrichment analysis (Fig. 6b/Figshare

file 4b<sup>6</sup>). This included mRNA processing and splicing, nonsense-mediated decay (NMD), and ribosomal RNA (rRNA) processing. Amongst the *O*-GlcNAcylated proteins identified in those categories were the transcriptional repressor EWS, Enhancer of mRNA-decapping protein 3 (EDC3), GTP-binding nuclear protein Ran, Heterogeneous nuclear ribonucleoproteins (hnRNPs), Nuclear RNA export factor 1 (NXF1), Histones (2A/2B/3/4), TBP, and RNA polymerase II (Figshare file 2<sup>6</sup>).

- (3) *Metabolism of Protein*: The human *O*-GlcNAcome was enriched for proteins participating in the processing or degradation of proteins, including ribosomal and proteasomal subunits (Fig. 6a/Figshare file 4a<sup>6</sup>). Pathway enrichment analysis also emphasized various aspects of protein metabolism, including the translational machinery (initiation, elongation, termination), co-translational protein targeting, and translational silencing of stress response gene expression (Fig. 6b/Figshare file 4b<sup>6</sup>). Examples in this category were regulatory subunits of the proteasome (non-ATPase, 26S and 11S), 40S ribosomal proteins S2 and S16, RACK1, translation initiator eIF4 A, F, G, and translational co-factor p67 (Figshare file 2<sup>6</sup>).
- (4) *Cellular structure*: Scaffold protein domains, including the calponin homology domain and actin-binding domain, were also found enriched in the human *O*-GlcNAcome. This group contained *O*-GlcNAcylated plastinins and actinins (Figshare file 2<sup>6</sup>), which correlated with the partial localization of *O*-GlcNAcylated proteins in the cytoskeletal fraction (Fig. 4/Figshare file 3a<sup>6</sup>), pointing toward a more structural role for *O*-GlcNAcylation.
- (5) *Disordered domains*: Using the DisProt database, we also observed that *O*-GlcNAcylation was prevalent in intrinsically disordered proteins (Figshare file 4c<sup>6</sup>)<sup>16</sup>. Out of the 591 manually curated human proteins available in the database, 50% were *O*-GlcNAcylated proteins, including the isoform Tau-F,  $\alpha$ -synuclein, and CDK1 (Figshare file 4c<sup>6</sup>). The average disorder content of those *O*-GlcNAc proteins was 26%.

## Discussion

This article generated an extensive inventory serving as a standing database for human *O*-GlcNAc-modified proteins with 5000+ proteins and 7000+ *O*-GlcNAc sites available in the literature. In addition to being deposited to Figshare and GlyGen, we developed a web interface to interrogate the database ([www.oglcnac.mcw.edu](http://www.oglcnac.mcw.edu)). Consequently, we encourage feedback and requests for integration of new proteins using the website's submission tool. For each protein, an *O*-GlcNAc score describes the extent of the identification and is, therefore, an estimator of the confidence. There is no strict threshold to validate a protein's *O*-GlcNAcylation. However, caution is advised for low score proteins (<10) for which extra-validation should be performed. Still, site mapping is the most decisive proof of *O*-GlcNAcylation in the *O*-GlcNAc field, which was only found in 35.5% of the human *O*-GlcNAcylated proteins.

Based on 7002 sites, the sequence analysis suggested that although there is no consensus sequence for *O*-GlcNAcylation, some amino acids are found more frequently. With more than twice the input sequences, our findings corroborate previously published *O*-GlcNAcylation semi-consensus sequences<sup>17–21</sup>. They can be used to refine existing or create new *in silico* prediction tools for *O*-GlcNAcylation.

Previously, gene ontology already suggested the enrichment of *O*-GlcNAcylated protein on RNA binding/processing proteins in activated human T cells and *Arabidopsis thaliana*<sup>22,23</sup>. Based on our extensive inventory of human *O*-GlcNAcylated proteins, we confirmed that a predominant role for *O*-GlcNAcylation lies in controlling genomic information, likely in response to nutrient input. This function is accomplished by the RNA/DNA-binding ability of many *O*-GlcNAcylated proteins and their RNA metabolism action. Predominantly, *O*-GlcNAcylation regulates the activity of numerous transcription factors, modifies the histones, and regulates RNA polymerase II activity<sup>24–26</sup>. Interestingly, *O*-GlcNAcylation also controls intron retention on various mRNAs, including *OGT* and *OGA* themselves, allowing the enzymes' rapid production in response to nutrient input<sup>27,28</sup>. Consequently, acute *OGT* inhibition affects the spliceosome/RNA-splicing pathway, and particularly the ability to retain introns<sup>28</sup>.

Expected of post-translational modifications, *O*-GlcNAcylation is also involved in regulating protein production and processing, including the translational machinery itself. Both *OGT* and *OGA* are tightly associated with ribosomes. Overall, *O*-GlcNAcylation enhances translation complex formation and stability by modifying ribosomal proteins, co-factors such as RACK1<sup>29</sup> or p67, and the translation initiation complex eIF2<sup>30–32</sup>. Finally, *O*-GlcNAcylation also prevents protein degradation by modifying subunits of the proteasome<sup>33</sup>. Altogether, *O*-GlcNAcylation is involved in every step of expressing our genetic information, from the activation or repression of genes, the production of mRNA and protein, and their processing, which overall defines cellular functions.

The responsiveness of *O*-GlcNAcylation to nutrient input is a critical characteristic of this modification. It allows for regulation of various cell signaling cascades in response to environmental changes. For example, the ability to thrive in favorable conditions engages *O*-GlcNAc cycling in many aspects. *O*-GlcNAcylation levels vary during the cell cycle, and its alteration disturbs mitotic phosphorylation, cyclin expression patterns, and cytokinesis, ultimately resulting in cell cycle failure and cell death<sup>34</sup>. Additionally, many cell cycle actors (cyclins, Cyclin-dependent Kinases) and pluripotency factors (OCT4 and SOX2) are modified in embryonic stem cells (ESCs), where *O*-GlcNAcylation regulates their transcriptional activity<sup>35</sup>. Finally, in *Drosophila*, *Ogt* is encoded by the essential Polycomb Group (PcG) gene *super sex comb (sxc)*<sup>36</sup>, crucial for proper patterns of development. *O*-GlcNAcylated proteins are also enriched on Polycomb response elements (PREs)<sup>37</sup>. *OGT* interacts with the Polycomb repressive complex 2 (PRC2) in human cells, one of the two PcG protein complexes in mammals<sup>38</sup>. The tight control of *OGT* expression during development by X-inactivation<sup>39</sup> also emphasizes the importance of *OGT* regulation during development. Altogether, those studies collectively support the vital role that *O*-GlcNAc plays in regulating the cell cycle and embryonic development.

Furthermore, the ability to quickly respond to stress is also a feature in which *O*-GlcNAcylation is involved. A global increase of *O*-GlcNAc levels is shown in response to several models of cellular stress induction<sup>40–42</sup>. Those

studies led to the identification of HSP-70 family members as *O*-GlcNAc-binding lectins<sup>43</sup>. *In vivo* studies in trauma-hemorrhage and ischemia/reperfusion injury models showed decreased *O*-GlcNAc levels following stress injury with subsequent necrosis and apoptosis<sup>44,45</sup>. Taken together, these data suggest that *O*-GlcNAcylation plays a protective role in response to cellular stress.

Across the years, the essential nature of *O*-GlcNAcylation has been observed in human and animal models. In humans, OGA or OGT deletions are not viable. Indeed, *OGT* is necessary for ESC viability, and conditional *Ogt* knockout is lethal in mice<sup>46</sup>. On the other hand, conditional deletion of *Oga* in mice is perinatal lethal, leading to severe phenotypes including altered glucose homeostasis, metabolic defects, and deregulation of genes related to growth and metabolism<sup>47</sup>. OGA inhibition leads to impaired differentiation of mESCs and activates transcription of genes usually epigenetically repressed in mESCs<sup>48</sup>. Recently, OGT deregulation has been identified in patients with X-linked Intellectual Disability (XLID) patients, accentuating once more the importance of *O*-GlcNAcylation in the brain. In those patients, OGT's single nucleotide polymorphism generates improper *O*-GlcNAc and OGA levels<sup>49–51</sup>, but does not completely abolish *O*-GlcNAcylation.

It is now widely accepted that balanced levels of *O*-GlcNAcylation are necessary for normal cell physiological function and that deregulation of those levels leads to diseases. The brain is the organ concentrating the most *O*-GlcNAcylated proteins in humans so far and is also one of the most glucogenic organs. Not surprisingly, the brain consumes about 120 g of glucose daily<sup>52</sup>. In a resting state, up to 60% of the whole-body glucose is used by the brain. The central nervous system is home to many *O*-GlcNAcylated proteins, including Microtubule-associated protein Tau,  $\alpha$ -Synuclein,  $\beta$ -Amyloid precursor protein (APP), and Huntingtin (Figshare file 2<sup>6</sup>). These proteins are linked to various neurodegenerative pathologies, including Alzheimer's, Parkinson's, and Huntington's diseases. Alteration of *O*-GlcNAcylation levels has been extensively linked to neurodegeneration<sup>53</sup>. Increased *O*-GlcNAcylation is typically protective in patients with neurodegenerative diseases. The human Tau protein carries *O*-GlcNAc residues that compete with key phosphorylation residues and prevents its aggregation<sup>54</sup>.

Similarly, *O*-GlcNAc protects neurons from Parkinson's disease by modifying  $\alpha$ -Synuclein and preventing its aggregation. Finally, *O*-GlcNAcylation on APP decreases the production of  $\beta$ -Amyloid peptides and reduces the formation of amyloid plaques<sup>55</sup>. *O*-GlcNAcylation in neurodegenerative diseases is the hotspot of *O*-GlcNAc research with promising treatment options emerging<sup>56,57</sup>.

The liver is a central platform for glucose metabolism in the human body and, as a result, shows a high concentration of *O*-GlcNAcylated proteins. The physiological role of *O*-GlcNAcylation in glucose metabolism includes the modification and regulation of the Insulin receptor, IRS1, AKT, GSK3 $\beta$ , FOXO1, and PGC1 $\alpha$  (Figshare file 2<sup>6</sup>). In mice, alteration of *O*-GlcNAc enzymes in the liver causes insulin resistance and dyslipidemia<sup>58</sup>, which is also the main feature of Non-Alcoholic Fatty Liver Disease (NAFLD). Increasing *O*-GlcNAcylation is protective against hepatic steatosis, supporting the critical role of *O*-GlcNAc in the liver<sup>59</sup>. Finally, liver fibrosis, one of the symptoms of Non-Alcoholic Steatohepatitis (NASH), is prominent in mice depleted of *O*-GlcNAcylation in the liver<sup>60</sup>. Similar to the brain, these studies suggest a protective role for *O*-GlcNAcylation in the liver.

In light of the significant advances in the field and the recent milestone of 5000 human proteins identified, *O*-GlcNAcylation is a highly prevalent area for biomedical research. Furthermore, it is even studied as a treatment option, for example, in neurodegenerative diseases<sup>56,57,61,62</sup>, increasing its significance for the medical community. With more technical innovations and growing interest in *O*-GlcNAc research, the role of *O*-GlcNAcylation in common pathologies is expected to be the focus of more intensive research in the near future.

## Methods

**Literature search.** To identify *O*-GlcNAcylated proteins, a systematic literature search was conducted (Fig. 2/Figshare file 1<sup>6</sup>). PUBMED search was finalized on May 5<sup>th</sup>, 2020, with the following search terms: “*O*-GlcNAc,” “*O*-GlcNAcylated,” and “*O*-GlcNAcylation” to englobe the entire *O*-GlcNAc literature. Previous reviews were excluded by adding the terms “NOT review” in the search criteria (Fig. 2). No restriction was imposed on the publication date.

**Proteome exchange search.** To identify more *O*-GlcNAcylated proteins, we have also included a search from Proteome Exchange (<http://www.proteomexchange.org/>), repository for mass spectrometry data. Our search criteria were “*O*-GlcNAc” in either the “Title” or “Keywords” field and “Homo Sapiens” in the “Species” field. Only datasets with a publication available were retained in our initial literature list.

**Literature curation.** The search results were then compiled and screened manually by the authors. Papers were rejected if they met any part of the exclusion criteria (Fig. 2). In short, non-primary research, non-English language, and retracted articles were excluded. Similarly, a study in which the source of the protein used could not be identified was also excluded. Publications detailing novel techniques and methods or computational studies were excluded unless they showed direct protein *O*-GlcNAcylation. Finally, articles were excluded when no direct evidence of protein *O*-GlcNAcylation was shown.

**Human research curation.** Only *O*-GlcNAcylated human proteins were retained for the final *O*-GlcNAcome list. The research was considered human if human proteins were used to study the *O*-GlcNAc status, whether endogenous, purified, or recombinant. Studies using non-human proteins in a human-based cell system were excluded and considered non-human models. Publications showing the modification of human origin proteins and non-human proteins were included and classified as “human mixed.”

***O*-GlcNAcylated protein curation.** A total of 5072 *O*-GlcNAcylated human proteins were extracted and cross-referenced, when available, to UniProtKB Entries. For each protein, the UniProt accession number (AC), Entry name, protein name, gene names, curation status (“Reviewed” vs. “unreviewed”), modification sites, and

identification techniques used were documented. Duplicate entries were merged and updated with any new information discovered throughout curation. Identified proteins from experiments that included protein IDs were also cross-referenced with UniProtKB. The list was then manually screened a second time to further improve accuracy by increasing the number of curated UniProtKB entries. “Unreviewed” UniProtKB entries were replaced or merged with corresponding “reviewed” entries when available. When multiple UniProtKB entries were available, “reviewed” entries were preferred to “unreviewed.” “Unreviewed” entries sharing gene names with “reviewed” entries were merged into the appropriate “reviewed” UniProt KB entry. Finally, any proteins identified by articles as fragments, and therefore “unreviewed,” were replaced with the full-length protein entry when available. The final table constitutes the human *O*-GlcNAcome database (Figshare file 2<sup>6</sup>).

**O-GlcNAc sites quality control.** A python script in the GlyGen<sup>63,64</sup> backend pipeline was used to process each protein entry of the database table and perform quality control (QC). In the QC checks, the reported UniProtKB accessions in the database were matched with the GlyGen’s human protein master list of UniProtKB canonical accessions to ensure all of the proteins belong to the human species. Next, the *O*-GlcNAc site data that contained the amino acid residue and its position were mapped to the GlyGen’s FASTA sequence for the corresponding human UniProtKB canonical proteins. The amino acid residues and positions that did not match with the FASTA sequence of the canonical protein were flagged and excluded from the table. Similarly, protein entries that did not have any site information were also excluded. The entries that passed all the quality control steps were exported as a final CSV dataset file. In contrast, the excluded entries were exported in a log file with the reasons for exclusion for further manual verification.

When site information differed from the original publication to align the UniprotKB canonical sequence, a disclaimer was added, and the changes were indicated in the note section.

**O-GlcNAc Score.** The *O*-GlcNAc score was assigned for each protein identified (Figshare file 2<sup>6</sup>). The score was calculated based on the following equation, in which all components were normalized by their maximal value in the dataset of 5072 entries:

$$S(x) = R(x)^{norm} + C(x)^{norm} + T(x)^{norm} + fA(x)^{norm} + lA(x)^{norm} + B(x)^{norm}$$

Briefly, given  $x$  the list of all protein entries and  $x$  a single entry, and considering the list of references  $Nx$  and  $P$  the number of protein entries documented in an index  $i$  of  $Nx$ :

- $R$  is the length of the list of references  $Nx$ .
- $C$  is the sum of per-year citations for each index  $i$  of  $Nx$ .
- $T$  is the time span between the first and last reference publication.
- $fA$  and  $lA$  are the number of distinct first and last author, respectively, within  $Nx$ .
- $B$  is a bonus term computed for each index  $i$  of  $Nx$  and averaged over  $R$ . Higher  $P_i$  negatively impacts  $B_i$  whereas higher  $C_i$  positively impacts  $B_i$ .

This absolute score is theoretically contained in the [0,6] range. We converted in a relative scale by normalizing score values over the top score protein entry.

**O-GlcNAc semi-consensus sequences.** *O*-GlcNAcylation semi-consensus sequence was generated after quality control using the  $-5$  to  $+5$  amino-acid sequence surrounding the 7002 *O*-GlcNAc sites. A frequency graph was generated using pLogo<sup>65</sup>.

**String network mapping.** Cytoscape was used to analyze the list of proteins using the STRING database<sup>66</sup>. Briefly, the UniProtKB ID list was imported using the “Import Network from public database” function and selecting the “STRING: protein query” option. *Homo sapiens* was chosen as the target species, and the following parameters were applied to create the STRING network: “confidence (score) cutoff”: 0.40; “Max additional interactors”: 0. A total of 4975 proteins matched to the STRING database and were used for the distribution and enrichment analysis. Due to the large number of proteins to analyze, four-node tables were generated and compiled together.

**Cellular and tissue distribution.** From the final node table (Cytoscape), UniProtKB ID and confidence score for each “Compartment” and “Tissue” column were extracted (Figshare file 3a/b<sup>6</sup>). Empty cells were assigned a 0 score, the lowest confidence score. A score of 5 corresponded to the highest confidence<sup>10,12</sup>. A total of 2923 or 6 entries did not have tissues or compartment information, respectively, and were therefore excluded from the analysis. For each compartment or tissue, the median confidence score was used for the analysis and graphic representation.

**Enrichment analysis.** To perform enrichment analysis of the human *O*-GlcNAcome, the HumanMine server<sup>67</sup> was used. Similarly, 5059 protein/gene couples matched the various databases used to perform enrichment analysis (Figshare file 4a/b<sup>6</sup>). Protein domain enrichment was performed using the following parameters: Test correction: Holm-Bonferroni; Max p-value: 0.05; Background population: Default. Gene ontology enrichment was also performed for molecular function using similar parameters. Finally, pathway enrichment for KEGG and Reactome datasets was performed using identical parameters. For all analyses, p-values were extracted and plotted.



**Intrinsic disorder protein domain.** The total 5072 human O-GlcNAcylated protein entries were mapped to the DisProt database<sup>16</sup> (Figshare file 4c<sup>6</sup>). The human database contains 591 manually curated intrinsically disordered proteins, and 299 DisProt entries were mapped as O-GlcNAcylated.

**Website.** A website interface was created to browse through the database and submit comments and request for integration ([www.oglcnac.mcw.edu](http://www.oglcnac.mcw.edu)).

It relies on the non-relational database management system MongoDB and is based on the Django web framework for rendering (<https://www.djangoproject.com/>). Backend processes were all developed using the Python programming language (v3.7) and the pymongo library (<https://pypi.org/project/pymongo/>) for database server-client interactions. GNU/Linux Debian-based systems with gunicorn (Python http) (<https://gunicorn.org/>) and NginX (SSL/reverse proxy) (<https://www.nginx.com/resources/wiki/>) were used for development and production of the O-GlcNAc Database.

### Data availability

The database can be accessed through the web platform [www.oglcnac.mcw.edu](http://www.oglcnac.mcw.edu). The O-GlcNAc database repository was also deposited on Figshare under the <https://doi.org/10.6084/m9.figshare.12443495.v14><sup>6</sup>. Finally, the full curated repository was also deposited and integrated into the GlyGen database under the URL: [https://data.glygen.org/GLY\\_000517](https://data.glygen.org/GLY_000517), under Creative Commons Attribution (CC BY 4.0) license. The URL is linked to the dataset's entry page. It contains the download option to download the dataset, BioCompute Object, and README that provides the bioinformatics workflow and metadata details.

### Code availability

Source code for the GlyGen QC and integration can be found in the Github repository: <https://github.com/glygener/glygen-backend-integration/blob/master/pipeline/integrator/make-proteiform-dataset.py>

Received: 2 October 2020; Accepted: 5 January 2021;

Published: 21 January 2021

### References

1. Johnson, R. K. *et al.* Dietary Sugars Intake and Cardiovascular Health: A Scientific Statement From the American Heart Association. *Circulation* **120**, 1011–1020 (2009).
2. Rippe, J. M. & Angelopoulos, T. J. Relationship between Added Sugars Consumption and Chronic Disease Risk Factors: Current Understanding. *Nutrients* **8** (2016).
3. Liu, K., Paterson, A. J., Chin, E. & Kudlow, J. E. Glucose stimulates protein modification by O-linked GlcNAc in pancreatic beta cells: linkage of O-linked GlcNAc to beta cell death. *Proc. Natl. Acad. Sci. USA* **97**, 2820–2825 (2000).
4. Olivier-Van Stichelen, S. & Hanover, J. A. You are what you eat: O-linked N-acetylglucosamine in disease, development and epigenetics. *Curr. Opin. Clin. Nutr. Metab. Care* **18**, 339–345 (2015).
5. Torres, C. R. & Hart, G. W. Topography and polypeptide distribution of terminal N-acetylglucosamine residues on the surfaces of intact lymphocytes. Evidence for O-linked GlcNAc. *J. Biol. Chem.* **259**, 3308–3317 (1984).
6. Olivier-Van Stichelen, S. The human O-GlcNAc database. *figshare* <https://doi.org/10.6084/m9.figshare.12443495.v14> (2020).
7. Ma, J. & Hart, G. W. O-GlcNAc profiling: from proteins to proteomes. *Clin. Proteomics* **11**, 8 (2014).
8. Manning, G., Whyte, D. B., Martinez, R., Hunter, T. & Sudarsanam, S. The protein kinase complement of the human genome. *Science* **298**, 1912–1934 (2002).
9. Sacco, F., Peretto, L., Castagnoli, L. & Cesareni, G. The human phosphatase interactome: An intricate family portrait. *FEBS Lett.* **586**, 2732–2739 (2012).
10. Binder, J. X. *et al.* COMPARTMENTS: unification and visualization of protein subcellular localization evidence. *Database J. Biol. Databases Curation* **2014**, bau012 (2014).
11. Zhao, L., Feng, Z., Yang, X. & Liu, J. The regulatory roles of O-GlcNAcylation in mitochondrial homeostasis and metabolic syndrome. *Free Radic. Res.* **50**, 1080–1088 (2016).
12. Palasca, O., Santos, A., Stolte, C., Gorodkin, J. & Jensen, L. J. TISSUES 2.0: an integrative web resource on mammalian tissue expression. *Database* 2018 (2018).
13. Kreppel, L. K., Blomberg, M. A. & Hart, G. W. Dynamic glycosylation of nuclear and cytosolic proteins. Cloning and characterization of a unique O-GlcNAc transferase with multiple tetratricopeptide repeats. *J. Biol. Chem.* **272**, 9308–9315 (1997).
14. Lubas, W. A., Frank, D. W. & Krause, M. & Hanover, J. A. O-Linked GlcNAc Transferase Is a Conserved Nucleocytoplasmic Protein Containing Tetratricopeptide Repeats. *J. Biol. Chem.* **272**, 9316–9324 (1997).
15. Martinez, M. R., Dias, T. B., Natov, P. S. & Zachara, N. E. Stress-induced O-GlcNAcylation: an adaptive process of injured cells. *Biochem. Soc. Trans.* **45**, 237–249 (2017).
16. Hatos, A. *et al.* DisProt: intrinsic protein disorder annotation in 2020. *Nucleic Acids Res.* **48**, D269–D276 (2020).
17. Woo, C. M. *et al.* Mapping and Quantification of Over 2000 O-linked Glycopeptides in Activated Human T Cells with Isotope-Targeted Glycoproteomics (Isotag). *Mol. Cell. Proteomics MCP* **17**, 764–775 (2018).
18. Kao, H.-J. *et al.* A two-layered machine learning method to identify protein O-GlcNAcylation sites with O-GlcNAc transferase substrate motifs. *BMC Bioinformatics* **16**(Suppl 18), S10 (2015).
19. Pathak, S. *et al.* The active site of O-GlcNAc transferase imposes constraints on substrate sequence. *Nat. Struct. Mol. Biol.* **22**, 744–750 (2015).
20. Wang, S. *et al.* Quantitative proteomics identifies altered O-GlcNAcylation of structural, synaptic and memory-associated proteins in Alzheimer's disease. *J. Pathol.* **243**, 78–88 (2017).
21. Leney, A. C., El Atmioui, D., Wu, W., Ovaa, H. & Heck, A. J. R. Elucidating crosstalk mechanisms between phosphorylation and O-GlcNAcylation. *Proc. Natl. Acad. Sci. USA* **114**, E7255–E7261 (2017).
22. Lund, P. J., Elias, J. E. & Davis, M. M. Global Analysis of O-GlcNAc Glycoproteins in Activated Human T Cells. *J. Immunol. Baltim. Md* **1950**(197), 3086–3098 (2016).
23. Xu, S.-L. *et al.* Proteomic analysis reveals O-GlcNAc modification on proteins with key regulatory functions in Arabidopsis. *Proc. Natl. Acad. Sci. USA* **114**, E1536–E1543 (2017).
24. Fardini, Y., Dehennaut, V., Lefebvre, T. & Issad, T. O-GlcNAcylation: A New Cancer Hallmark? *Front. Endocrinol.* **4**, 99 (2013).
25. Hanover, J. A. Epigenetics gets sweeter: O-GlcNAc joins the 'histone code'. *Chem. Biol.* **17**, 1272–1274 (2010).
26. Ranuncolo, S. M., Ghosh, S., Hanover, J. A., Hart, G. W. & Lewis, B. A. Evidence of the involvement of O-GlcNAc-modified human RNA polymerase II CTD in transcription *in vitro* and *in vivo*. *J. Biol. Chem.* **287**, 23549–23561 (2012).

27. Park, S.-K. *et al.* A Conserved Splicing Silencer Dynamically Regulates O-GlcNAc Transferase Intron Retention and O-GlcNAc Homeostasis. *Cell Rep.* **20**, 1088–1099 (2017).
28. Tan, Z.-W. *et al.* O-GlcNAc regulates gene expression by controlling detained intron splicing. *Nucleic Acids Res.* **48**, 5656–5669 (2020).
29. Zeidan, Q., Wang, Z., De Maio, A. & Hart, G. W. O-GlcNAc cycling enzymes associate with the translational machinery and modify core ribosomal proteins. *Mol. Biol. Cell* **21**, 1922–1936 (2010).
30. Li, X. *et al.* O-GlcNAcylation of core components of the translation initiation machinery regulates protein synthesis. *Proc. Natl. Acad. Sci. USA* **116**, 7857–7866 (2019).
31. Dierschke, S. K. *et al.* O-GlcNAcylation alters the selection of mRNAs for translation and promotes 4E-BP1-dependent mitochondrial dysfunction in the retina. *J. Biol. Chem.* **294**, 5508–5520 (2019).
32. Datta, B., Ray, M. K., Chakrabarti, D., Wylie, D. E. & Gupta, N. K. Glycosylation of eukaryotic peptide chain initiation factor 2 (eIF-2)-associated 67-kDa polypeptide (p67) and its possible role in the inhibition of eIF-2 kinase-catalyzed phosphorylation of the eIF-2 alpha-subunit. *J. Biol. Chem.* **264**, 20620–20624 (1989).
33. Zhang, F. *et al.* O-GlcNAc modification is an endogenous inhibitor of the proteasome. *Cell* **115**, 715–725 (2003).
34. Slawson, C. *et al.* Perturbations in O-linked beta-N-acetylglucosamine protein modification cause severe defects in mitotic progression and cytokinesis. *J. Biol. Chem.* **280**, 32944–32956 (2005).
35. Jang, H. *et al.* O-GlcNAc regulates pluripotency and reprogramming by directly acting on core components of the pluripotency network. *Cell Stem Cell* **11**, 62–74 (2012).
36. Sinclair, D. A. R. *et al.* Drosophila O-GlcNAc transferase (OGT) is encoded by the Polycomb group (PcG) gene, super sex combs (sxc). *Proc. Natl. Acad. Sci. USA* **106**, 13427–13432 (2009).
37. Akan, I., Love, D. C., Harwood, K. R. & Bond, M. R. & Hanover, J. A. Drosophila O-GlcNAcase Deletion Globally Perturbs Chromatin O-GlcNAcylation. *J. Biol. Chem.* **291**, 9906–9919 (2016).
38. Gao, J. *et al.* Proteomic analysis of the OGT interactome: novel links to epithelial-mesenchymal transition and metastasis of cervical cancer. *Carcinogenesis* **39**, 1222–1234 (2018).
39. Olivier-Van Stichelen, S. & Hanover, J. A. X-inactivation normalizes O-GlcNAc transferase levels and generates an O-GlcNAc-depleted Barr body. *Front. Genet.* **5**, 256 (2014).
40. Taylor, R. P. *et al.* Glucose deprivation stimulates O-GlcNAc modification of proteins through up-regulation of O-linked N-acetylglucosaminyltransferase. *J. Biol. Chem.* **283**, 6050–6057 (2008).
41. Káta, E. *et al.* Oxidative stress induces transient O-GlcNAc elevation and tau dephosphorylation in SH-SY5Y cells. *J. Cell. Mol. Med.* **20**, 2269–2277 (2016).
42. Zachara, N. E. *et al.* Dynamic O-GlcNAc modification of nucleocytoplasmic proteins in response to stress. A survival response of mammalian cells. *J. Biol. Chem.* **279**, 30133–30142 (2004).
43. Lefebvre, T. *et al.* Identification of N-acetyl-d-glucosamine-specific lectins from rat liver cytosolic and nuclear compartments as heat-shock proteins. *Biochem. J.* **360**, 179–188 (2001).
44. Yang, S. *et al.* Glucosamine administration during resuscitation improves organ function after trauma hemorrhage. *Shock Augusta Ga* **25**, 600–607 (2006).
45. Champattanachai, V. & Marchase, R. B. & Chatham, J. C. Glucosamine protects neonatal cardiomyocytes from ischemia-reperfusion injury via increased protein O-GlcNAc and increased mitochondrial Bcl-2. *Am. J. Physiol. Cell Physiol.* **294**, C1509–1520 (2008).
46. Shañ, R. *et al.* The O-GlcNAc transferase gene resides on the X chromosome and is essential for embryonic stem cell viability and mouse ontogeny. *Proc. Natl. Acad. Sci. USA* **97**, 5735–5739 (2000).
47. Keembiyehetty, C. *et al.* Conditional knock-out reveals a requirement for O-linked N-Acetylglucosaminase (O-GlcNAcase) in metabolic homeostasis. *J. Biol. Chem.* **290**, 7097–7113 (2015).
48. Speakman, C. M. *et al.* Elevated O-GlcNAc levels activate epigenetically repressed genes and delay mouse ESC differentiation without affecting naïve to primed cell transition. *Stem Cells Dayt. Ohio* **32**, 2605–2615 (2014).
49. Pravata, V. M. *et al.* A missense mutation in the catalytic domain of O-GlcNAc transferase links perturbations in protein O-GlcNAcylation to X-linked intellectual disability. *FEBS Lett.* **594**, 717–727 (2020).
50. Willems, A. P. *et al.* Mutations in N-acetylglucosamine (O-GlcNAc) transferase in patients with X-linked intellectual disability. *J. Biol. Chem.* **292**, 12621–12631 (2017).
51. Vaidyanathan, K. *et al.* Identification and characterization of a missense mutation in the O-linked β-N-acetylglucosamine (O-GlcNAc) transferase gene that segregates with X-linked intellectual disability. *J. Biol. Chem.* **292**, 8948–8963 (2017).
52. Berg, J. M., Tymoczko, J. L. & Stryer, L. Each Organ Has a Unique Metabolic Profile. (2002).
53. Akan, I., Olivier-Van Stichelen, S., Bond, M. R. & Hanover, J. A. Nutrient-driven O-GlcNAc in proteostasis and neurodegeneration. *J. Neurochem.* **144**, 7–34 (2018).
54. Ryan, P. *et al.* O-GlcNAc Modification Protects against Protein Misfolding and Aggregation in Neurodegenerative Disease. *ACS Chem. Neurosci.* **10**, 2209–2221 (2019).
55. Yuzwa, S. A. *et al.* Pharmacological inhibition of O-GlcNAcase (OGA) prevents cognitive decline and amyloid plaque formation in bigenic tau/APP mutant mice. *Mol. Neurodegener.* **9**, 42 (2014).
56. Wang, X. *et al.* MK-8719, a Novel and Selective O-GlcNAcase Inhibitor That Reduces the Formation of Pathological Tau and Ameliorates Neurodegeneration in a Mouse Model of Tauopathy. *J. Pharmacol. Exp. Ther.* **374**, 252–263 (2020).
57. Tavassoly, O., Yue, J. & Vocadlo, D. J. Pharmacological inhibition and knockdown of O-GlcNAcase reduces cellular internalization of α-synuclein preformed fibrils. *FEBS J.* <https://doi.org/10.1111/febs.15349> (2020).
58. Yang, X. *et al.* Phosphoinositide signalling links O-GlcNAc transferase to insulin resistance. *Nature* **451**, 964–969 (2008).
59. Guinez, C. *et al.* O-GlcNAcylation increases ChREBP protein content and transcriptional activity in the liver. *Diabetes* **60**, 1399–1413 (2011).
60. Zhang, B. *et al.* O-GlcNAc transferase suppresses necroptosis and liver fibrosis. *JCI Insight* **4** (2019).
61. Hastings, N. B. *et al.* Inhibition of O-GlcNAcase leads to elevation of O-GlcNAc tau and reduction of tauopathy and cerebrospinal fluid tau in rTg4510 mice. *Mol. Neurodegener.* **12**, 39 (2017).
62. Selnick, H. G. *et al.* Discovery of MK-8719, a Potent O-GlcNAcase Inhibitor as a Potential Treatment for Tauopathies. *J. Med. Chem.* **62**, 10062–10097 (2019).
63. York, W. S. *et al.* GlyGen: Computational and Informatics Resources for Glycoscience. *Glycobiology* **30**, 72–73 (2020).
64. Kahsay, R. *et al.* GlyGen data model and processing workflow. *Bioinformatics* **36**, 3941–3943 (2020).
65. O’Shea, J. P. *et al.* pLogo: a probabilistic approach to visualizing sequence motifs. *Nat. Methods* **10**, 1211–1212 (2013).
66. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
67. Smith, R. N. *et al.* InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. *Bioinforma. Oxf. Engl.* **28**, 3163–3165 (2012).

## Acknowledgements

We thank John Hanover, Ph.D., for his scientific insight in building this article. We thank the GlyGen project and Michael Tiemeyer, Ph.D., for their interest and support in finalizing the O-GlcNAc database.

### Author contributions

E.W.F., R.R.B., L.M., L.D. and S.O.V.S. performed the literature curation and built the initial O-GlcNAcome list. F.M. performed quality controls, created the O-GlcNAc score algorithm and the web interface. J.V. and R.K. performed quality controls and integration to the GlyGen platform. S.O.V.S. performed the meta-analysis. E.W.F., R.B., F.M. and S.O.V.S. wrote the paper.

### Competing interests

The authors declare no competing interests.

### Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41597-021-00810-4>.

**Correspondence** and requests for materials should be addressed to S.O.V.S.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021