GlycoForum - Technical Note

O-GlcNAcAtlas: A database of experimentally identified O-GlcNAc sites and proteins

Junfeng Ma1,2, Yaoxiang Li2, Chunyan Hou3, and Ci Wu2

1Department of Oncology, Lombardi Comprehensive Cancer Center, Georgetown University Medical Center, Washington, DC 20057, USA, and 2Dalian Institute of Chemical Physics, Chinese Academy of Sciences, Dalian, Liaoning 116023, China

Received 24 November 2020; Revised 31 December 2020; Editorial Decision 1 January 2021; Accepted 1 January 2021

Abstract

O-linked β-N-acetylglucosamine (O-GlcNAc) is a post-translational modification (i.e., O-GlcNAcylation) on the serine/threonine residues of proteins. As a unique intracellular monosaccharide modification, protein O-GlcNAcylation plays important roles in almost all biochemical processes examined. Aberrant O-GlcNAcylation underlies the etiologies of a number of chronic diseases. With the tremendous improvement of techniques, thousands of proteins along with their O-GlcNAc sites have been reported. However, until now, there are few databases dedicated to accommodate the rapid accumulation of such information. Thus, O-GlcNAcAtlas is created to integrate all experimentally identified O-GlcNAc sites and proteins. O-GlcNAcAtlas consists of two datasets (Dataset-I and Dataset-II, for unambiguously identified sites and ambiguously identified sites, respectively), representing a total number of 4571 O-GlcNAc modified proteins from all species studied from 1984 to 31 Dec 2019. For each protein, comprehensive information (including species, sample type, gene symbol, modified peptides and/or modification sites, site mapping methods and literature references) is provided. To solve the heterogeneity among the data collected from different sources, the sequence identity of these reported O-GlcNAc peptides are mapped to the UniProtKB protein entries. To our knowledge, O-GlcNAcAtlas is a highly comprehensive and rigorously curated database encapsulating all O-GlcNAc sites and proteins identified in the past 35 years. We expect that O-GlcNAcAtlas will be a useful resource to facilitate O-GlcNAc studies and computational analyses of protein O-GlcNAcylation. The public version of the web interface to the O-GlcNAcAtlas can be found at http://oglcnac.org/.

Key words: database, O-GlcNAc, proteomics

Introduction

O-linked β-N-acetylglucosamine (O-GlcNAc), which was discovered in early 1980s, is a post-translational modification (i.e., O-GlcNAcylation) on the serine/threonine residues of proteins (Torres and Hart 1984; Holt and Hart 1986). Distinct from the traditional glycosylation (i.e., N-glycosylation, O-glycosylation and glycosylphosphatidylinositol-anchored glycosylation), O-GlcNAcylation is a unique intracellular monosaccharide modification without being further elongated into complex sugar structures (Wells et al. 2001; Hart et al. 2007). After several decades’ endeavor, it has been revealed that O-GlcNAcylation exists in all metazoans (including animals, insects and plants), some bacteria, fungi and virus. By modulating various aspects of target proteins (e.g., activity, localization and stability), O-GlcNAcylation exerts diverse functional roles in many biochemical processes (Hart et al. 2011; Bond and Hanover 2013; Yang and Qian 2017; Hart 2019). Mounting evidence has demonstrated that deregulated protein O-GlcNAcylation underlies multiple human diseases, especially in
A number of bioinformatics platforms and databases have been developed for (glyco)proteins and glycans (Abrahams et al. 2020; Li et al. 2020), including PhosphoSite Plus (Hornbeck et al. 2019), dbPTM (Huang et al. 2019), MS-viewer (Baker and Chalkley 2014), UniCarbKB (Campbell et al. 2014), N-GlycositeAtlas (Sun et al. 2019), GlyGen (York et al. 2020), Glycosciences.DB (Böhm et al. 2019), GlyTouCan (Tiemeyer et al. 2017) and GlyConnect (Aloci et al. 2019). Unfortunately, these databases cover limited information of O-GlcNAc sites and proteins. Until now, few databases have been created to specifically accommodate the rapid accumulation of O-GlcNAc information on proteins. The database of O-GlcNAcylated proteins and sites (dbOGAP) which was constructed in 2011 contains ∼400 O-GlcNAcylation sites and has not been updated (Wang et al. 2011). Undoubtedly, there is an urgent need to create a comprehensive and curated O-GlcNAc-specific database. A human O-GlcNAc protein database has been recently introduced (Wulff-Fuentes et al. In review). Herein, we describe O-GlcNAcAtlas, a manually curated database of all experimentally identified O-GlcNAc sites and proteins from all species studied in the past 35 years (from 1984 to 31 Dec 2019). By enabling users to search and retrieve data easily, O-GlcNAcAtlas is proposed to facilitate O-GlcNAc studies (e.g., interrogation of functions of O-GlcNAcylation and of specific modification sites) on proteins in different biomedical settings.

Results and discussion

By following the workflow shown in Figure 1A, we assembled all experimentally identified O-GlcNAc sites and proteins for a comprehensive database O-GlcNAcAtlas. Literature mining from PubMed yielded a total of 2236 O-GlcNAc-relevant articles (Supplementary Figure S1A). Among them, 225 articles contain O-GlcNAc sites on proteins (Supplementary Figure S1B). O-GlcNAc related information in each publication was manually retrieved, curated and compiled. Of special note, to minimize and avoid misleading and confusion, stringent selection criteria were applied to select O-GlcNAc sites and proteins. For large-scale proteomics studies, proteins without O-GlcNAc peptides/sites identified were not included. Each entry from low-throughput studies was also carefully curated. Moreover, to maintain scientific rigorous as described by the original authors of these studies, both unambiguously identified O-GlcNAc sites and ambiguously identified sites were recorded and categorized. Last but not the least, we fully respect the original authors’ discoveries, but by adding curators’ comments, we hope viewers can be aware of what happened to specific entries (e.g., mistakenly labeled modification residues or position in a peptide sequence identified) during curation.

Basically, O-GlcNAcAtlas consists of two datasets, depending on the ambiguity of O-GlcNAc sites mapped. Dataset-I contains unambiguously assigned O-GlcNAc sites, while Dataset-II is for O-GlcNAc sites ambiguously identified (mainly due to the low localization scores by software tools especially for peptides with clustered serine/threonine residues). Despite the ambiguity of specific modification sites, the corresponding peptides can be positively identified, so do the O-GlcNAc proteins. Thus, Dataset-II is also an important part of O-GlcNAcAtlas as it provides useful information for the confirmation of O-GlcNAc status for some proteins. Overall, 9348 O-GlcNAc sites were unambiguously identified, corresponding to 8151 peptides and 3918 proteins (Supplementary Figure S1C). In addition, 3028 peptides on 1507 proteins were found to be O-GlcNAcylated, corresponding to 6520 ambiguous sites (Supplementary Figure S1D).

To our knowledge, in contrast to all currently existing databases, O-GlcNAcAtlas distinguishes the ambiguity of O-GlcNAc sites as reported by the original authors. Moreover, even assuming that all O-GlcNAc sites presented in other databases are unambiguous, O-GlcNAcAtlas contains a substantially higher number of unambiguous O-GlcNAc sites experimentally identified from all species studied in the past 35 years. Undoubtedly, this enriched and rigorously curated O-GlcNAc site information can be beneficial in multiple ways. For example, (1) it will help with the development of O-GlcNAc site-specific antibodies for proteins; (2) it will enable the investigation of site-specific functional roles of many proteins and (3) by combining information from other publicly available resources (e.g., UniProtKB, dbPTM, PhosphoSite Plus and Protein Data Bank (Burley et al. 2019), it will facilitate the exploration of potential cross-talk between O-GlcNAc and other post-translational modifications for specific proteins of interest.

Among the 9348 unambiguous O-GlcNAc sites, >98% were identified during 2010–2019 (Supplementary Figure S2A). Moreover, >98% of all sites were unambiguously assigned by MS (Supplementary Figure S2B). While ∼15% of all sites were identified by two or more publications, the majority (85%) were found only once (Supplementary Figure S2C). And it turns out that the distribution of serine and threonine residues is 62%:38% (slightly less than a ratio of 2:1) (Supplementary Figure S2D).

Besides 3918 proteins with unambiguous O-GlcNAc sites, 1507 proteins were matched with ambiguous O-GlcNAc sites. Providing 854 proteins were overlapped between the two sets, and in total, 4571 O-GlcNAc proteins were identified (Supplementary Figure S3A). Among the O-GlcNAc proteins, ~77% (3535 out of 4571 proteins) were identified by one study (Supplementary Figure S3B), with 27 proteins identified by at least 10 times (Supplementary Table S1). About 62% of proteins are derived from human and ~38% (1728 out of 4571 proteins) are from other species (mainly common model systems, such as mouse, rat, Caenorhabditis elegans, Drosophila, Arabidopsis and wheat) (Supplementary Figure S3C). The details of O-GlcNAc proteins/sites information from different species are shown in Table 1. Regarding human proteins, most O-GlcNAc proteins were identified from model cell lines (e.g., HeLa cells and HEK293
cells) (Supplementary Figure S3D). Moreover, hundreds of O-GlcNAc proteins were identified from tissues/cells of special research interest (e.g., primary T cells and brain).

To facilitate the use of the O-GlcNAcAtlas resource, a web interface has been developed for users to browse and search efficiently for their O-GlcNACylated proteins of interest. O-GlcNAcAtlas can be searched using UniProtKB accession, protein name, gene symbol and peptide sequence as keywords, and the results can be filtered further. The search output includes the basic annotations for all the matched entries (as exemplified in Figure 1B). The accession number of each entry is linked to the detailed annotation for the site information of specific proteins (as exemplified in Figure 1B). So far, O-GlcNAcAtlas supports several functions including data searching, browsing and retrieving. Moreover, search results can be directly downloaded and saved from the O-GlcNAcAtlas webpage.

Concluding remarks

To appreciate the tremendous efforts in O-GlcNAc research in the past 35 years, we aimed to create a comprehensive and rigorously curated database of O-GlcNAc sites and proteins. O-GlcNAcAtlas not only includes data from case-by-case studies but also integrates high-throughput data from proteomics studies. For either low-throughput or high-throughput studies, we tried our best to carefully curate each entry, with the curators’ comments added. With O-GlcNAcAtlas, we aim to provide a one-stop portal for biomedical investigators to search O-GlcNAcylated proteins and sites. We anticipate it will largely facilitate O-GlcNAc-targeting basic and translational research in multiple aspects (e.g., elucidation of O-GlcNAc site-specific functional roles of proteins).

Methods

The system flow of the construction of the O-GlcNAcAtlas is presented in Figure 1A. Specifically, O-GlcNAcAtlas was compiled through a manual curation of the literature published between 1984 and 31 Dec 2019. The following search items: “O-linked β-N-acetylglucosamine,” “O-GlcNAc” or “O-GlcNAcylation” were used to extract publications from PubMed. O-GlcNAc sites information in each publication was retrieved and evaluated by at least two curators. Besides O-GlcNAc sites, related information (including species, sample type, peptide sequence, protein name and site-mapping methods used) was also extracted. To determine the positions of O-GlcNAcylated serine/threonine residues, the experimentally identified peptides were then mapped to the UniProtKB protein entries based on the database identifier or sequence similarity. The O-GlcNAcylated peptides/sites that could not align exactly to a protein sequence were annotated with curators’ comments. Finally,
each mapped O-GlcNAc site was attributed to the corresponding literature (PubMed ID).

A user-friendly, web-based graphical user interface was created with HTML, CSS and Bootstrap. The backend server was running on a collection of services developed using Python programming language (version 3.8.1) and was coupled with the MySQL database. All entries, given a unique O-GlcNAcAtlas identification number, were organized in the MySQL database. Accession, protein name, gene name and peptide sequence were set as keywords input to retrieve the O-GlcNAc information of proteins of interest. The results could be downloaded and saved in multiple formats (including CSV, Excel and PDF).

### Supplementary data

Supplementary data for this article are available online at http://glycob.oxfordjournals.org/.

### Acknowledgements

We are indebted to Dr. Gerald Hart for his insight during the initiation of this project several years ago. We wish to thank Drs. Zhangzhi Hu and Leslie Arminski for helpful discussions. We would like to acknowledge the researchers who, by generously answering our curators’ questions regarding O-GlcNAc sites in specific publications, have contributed to improve the database. We would appreciate if investigators can report any potentially missing sites and/or send the datasets in their new publications to us so that we can update this database in a timely manner.

### Funding

The authors are partially supported by National Institutes of Health/ National Cancer Institute (P30-CA-51008).

### Conflict of interest statement

None declared.

### Abbreviations

- O-GlcNAc
- O-linked β-N-acetylglucosamine
- MS, mass spectrometry

### References

Abrahams JL, Taherzadeh G, Jarvas G, Gutman A, Zhou Y, Campbell MP. 2020. Recent advances in glycoinformatic platforms for glycomics and glycoproteomics. *Caer Open Struct Biol.* 62:56–69.

Alfaro JF, Gong CX, Montoro ME, Aldrich JT, Clauss TR, Purvine SO, Wang Z, Camp DG 2nd, Shabanowitz J, Stanley P, et al. 2012. Tandem mass spectrometry identifies many mouse brain O-GlcNAcylated proteins including EGF domain-specific O-GlcNAc transferase targets. *Proc Natl Acad Sci U S A.* 109(19):7280–7285.

Aloci D, Mariethoz J, Gastaldello A, Gasteiger E, Karlsson NG, Kolarich D, Packer NH, Lisacek F. 2019. GlyConnect: Glycoproteomics goes visual, interactive, and analytical. *J Proteome Res.* 18(2):664–677.

Baker PR, Chalkley RJ. 2014. MS-viewer: A web-based spectral viewer for proteomics results. *Mol Cell Proteomics.* 13(5):1392–1396.

Bohm M, Bohne-Lang A, Frank M, Loss A, Rojas-Macias MA, Lütteke T. 2019. Glycosciences. DB: An annotated data collection linking glycomics and proteomics data (2018 update). *Nucleic Acids Res.* 47(D1):D1195–D1201.

Bond MR, Hanover JA. 2013. O-GlcNAc cycling: A link between metabolism and chronic disease. *Am Rev Nutr.* 33:205–229.

Burley SK, Berman HM, Bi C, Chen L, Costanzo LD, Christie C, Dalenberg K, Duarte JM, Dutta S, Feng Z, et al. 2019. RCSB protein data Bank: Biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. *Nucleic Acids Res.* 47(D1):D446–D474.

Campbell MP, Peterson R, Mariethoz J, Gasteiger E, Akune Y, Aoki-Kinoshita KF, Lisacek F, Packer NH. 2014. UniCarbKB: building a knowledge platform for glycoproteomics. *Nucleic Acids Res.* 42(D1):D215–D221.

Ferrer CM, Sodi VL, Reginao MJ. 2016. O-GlcNAcylation in cancer biology: Linking metabolism and signaling. *J Mol Biol.* 428(16):3282–3294.

Hanover JA, Chen WP, Bond MR. 2018. O-GlcNAc in cancer: An oncometabolism-fueled vicious cycle. *J Bienerg Biomembr.* 50(3):153–173.

Hart GW, Housley MP, Slawson C. 2007. Cycling of O-linked beta-N-acetylglucosamine on nucleocytoplasmic proteins. *Nature.* 446(7139):1017–1022.

Hart GW, Slawson C, Ramirez-Correa G, Lagerlof O. 2011. Cross talk between O-GlcNAcylation and phosphorylation: Roles in signaling, transcription, and chronic disease. *Annu Rev Biochem.* 80(1):825–858.

Hart GW. 2019. Nutrient regulation of signaling and transcription. *J Biol Chem.* 294(7):2211–2231.

Holt GD, Hart GW. 1986. The subcellular distribution of terminal N-acetylglucosamine moieties. Localization of a novel protein-saccharide linkage. *O-linked GlcNAc J Biol Chem.* 261(17):8049–8057.

Horbeck PV, Kornhauser JM, Latham V, Murray B, Nandhikonda V, Nord A, Skrzypek E, Wheeler T, Zhang B, Gnad F. 2019. 15 years of PhosphoSitePlus®: Integrating post-translationally modified sites, disease variants and isoforms. *Nucleic Acids Res.* 47(D1):D433–D441.
Huang KY, Lee TY, Kao HJ, Ma CT, Lee CC, Lin TH, Chang WC, Huang HD. 2019. dbPTM in 2019: Exploring disease association and cross-talk of post-translational modifications. Nucleic Acids Res. 47(D1):D298–D308.

Li J, Li Z, Duan X, Qin K, Dang L, Sun S, Cai L, Hsieh-Wilson LC, Wu L, Yi W. 2019. An isotope-coded photolabile probe for quantitative profiling of protein O-GlcNAcylation. ACS Chem Biol. 14(1):4–10.

Li X, Xu Z, Hong X, Zhang Y, Zou X. 2020. Databases and bioinformatic tools for glycochemistry and glycoproteomics. Int J Mol Sci. 21:6727.

Ma J, Hart GW. 2013. Protein O-GlcNAcylation in diabetes and diabetic complications. Expert Rev Proteomics. 10(4):365–380.

Ma J, Hart GW. 2014. O-GlcNAc profiling: From proteins to proteomes. Clin Proteomics. 11(1):8.

Ma J, Liu T, Wei A-C, Banerjee P, O’Rourke B, Hart GW. 2015. O-GlcNAcomics profiling identifies widespread O-linked β-N-acetylglucosamine modification (O-GlcNAcylation) in oxidative phosphorylation system regulating cardiac mitochondrial function. J Biol Chem. 290(49):29141–29153.

Ma J, Wu C, Hart GW. 2021. Analytical and biochemical perspectives of protein O-GlcNAcylation. Chem Rev. doi: 10.1021/acs.chemrev.0c00884.

Ma Z, Vosseller K. 2014. Cancer metabolism and elevated O-GlcNAc in oncogenic signaling. J Biol Chem. 289(50):34457–34465.

Maynard J, Chalkley RJ. 2021. Methods for enrichment and assignment of N-acetylglucosamine modification sites. Mol Cell Proteomics. doi: 10.1074/mcp.R120.002206.

Qin K, Zhu Y, Qin W, Gao J, Shao X, Wang YL, Zhou W, Wang C, Chen X. 2018. Quantitative profiling of protein O-GlcNAcylated sites by an isotope-tagged cleavable linker. ACS Chem Biol. 13(8):1983–1989.

Slawson C, Hart GW. 2011. O-GlcNAc signaling: Implications for cancer cell biology. Nat Rev Cancer. 11(9):678–684.

Sun S, Hu Y, Ao M, Shah P, Chen J, Yang W, Jia X, Tian Y, Thomas S, Zhang H. 2019. N-GlycositeAtlas: A database resource for mass spectrometry-based human N-linked glycoprotein and glycosylation site mapping. Clin Proteomics. 16:35.

Thompson JW, Sorum AW, Hsieh-Wilson LC. 2018. Deciphering the functions of O-GlcNAc glycosylation in the brain: The role of site-specific quantitative O-GlcNAcomics. Biochemistry. 57(27):4010–4018.

Tiemeyer M, Aoki K, Paulson J, Cummings RD, York WS, Karlsson NG, Lisacek F, Packer NH, Campbell MP, Aoki NP, et al. 2017. GlyToUCan: An accessible glycan structure repository. Glycobiology. 27(10):915–919.

Torres CR, Hart GW. 1984. Topography and polypeptide distribution of terminal N-acetylglucosamine residues on the surfaces of intact lymphocytes. Evidence for O-linked GlcNAc. J Biol Chem. 259(5):3308–3317.

Trinidad JC, Barkan DT, Gulledge BF, Thalhammer A, Sali A, Schoepfer R, Burlingame AL. 2012. Global identification and characterization of both O-GlcNAcylation and phosphorylation at the murine synapse. Mol Cell Proteomics. 11(8):215–229.

Vaidyanathan K, Wells L. 2014. Multiple tissue-specific roles for the O-GlcNAc post-translational modification in the induction of and complications arising from type II diabetes. J Biol Chem. 289(50):34466–34471.

Vasconcelos-Dos Santos A, Muniz de Queiroz R, Rodrigues BC, Todeschini AR, Dias WB. 2018. Hyperglycemia and aberrant O-GlcNAcylation: Contributions to tumor progression. J Bioenerg Biomembr. 50(3):175–187.

Wang J, Torii M, Liu H, Hart GW, Hu Z-Z. 2011. dbOAGP—An integrated bioinformatics resource for protein O-GlcNAcylation. BMC Bioinform. 12(1):91.

Wang S, Yang F, Petyuk VA, Shukla AK, Monroe ME, Gritsenko MA, Rodland KD, Smith RD, Qian WJ, Gong CX, et al. 2017. Quantitative proteomics identifies altered O-GlcNAcylation of structural, synaptic and memory-associated proteins in Alzheimer’s disease. J Pathol. 243(1):78–88.

Wang Z, Hart GW. 2008. Glycomic approaches to study GlcNAcylomicroarray profiling identifies widespread O-GlcNAcylation of structural, synaptic and memory-associated proteins in Alzheimer’s disease. J Pathol. 243(1):78–88.

Wani WY, Chatham JC, Darley-Usmar V, McMahon LL, Zhang J. 2017. O-GlcNAcylomicroarray and neurodegeneration. Brain Res Bull. 133:80–87.

Wells L, Vosseller K, Hart GW. 2001. Glycosylation of nucleocytoplasmic proteins: Signal transduction and O-GlcNAc. Science. 291(5512):2375–2378.

Woo CM, Lund PJ, Huang AC, Davis MM, Bertozzi CR, Pitteri SJ. 2018. Mapping and quantification of over 2000 O-linked glycopeptides in activated human T cells with isotope-targeted glycopeptide (Isotag). Mol Cell Proteomics. 17(4):764–775.

Xu SL, Chalkley RJ, Maynard JC, Wang W, Ni W, Jiang X, Shin K, Cheng L, Savage D, Huhmer AF, et al. 2017. Proteomic analysis reveals O-GlcNAc modification on proteins with key regulatory functions in Arabidopsis. Proc Natl Acad Sci U S A. 114(8):E1536–E1543.

Yang X, Qian K. 2017. Protein O-GlcNAcylation: Emerging mechanisms and functions. Nat Rev Mol Cell Biol. 18(7):452–465.

York WS, Mazumder R, Ranzinger R, Edwards N, Kalsbay R, Aoki-KinoshitaKF, Campbell MP, Cummings RD, FeiTi, Martin M, et al. 2020. Gly gen: Computational and informatics resources for glycoscience. Glycobiology. 30(2):72–73.

Yuzwa SA, Vocadlo DJ. 2014. O-GlcNAc and neurodegeneration: Biochemical mechanisms and potential roles in Alzheimer’s disease and beyond. Chem Soc Rev. 43(19):6839–6858.

Zhang P, Viner R, Teo CF, Boons G-J, Horn D, Wells L. 2011. Combining high-energy C-trap dissociation and electron transfer dissociation for protein O-GlcNAc modification site assignment. J Proteome Res. 10(9):4088–4104.

Zhu Y, Hart GW. 2021. Targeting O-GlcNAcylation to develop novel therapeutics. Mol Aspects Med. doi: 10.1016/j.mam.2020.100885.