



## Towards a standardized bioinformatics infrastructure for N- and O-glycomics

Rojas-Macias, Miguel A.; Mariethoz, Julien; Andersson, Peter; Jin, Chunsheng; Venkatakrisnan, Vignesh; Aoki, Nobuyuki P.; Shinmachi, Daisuke; Ashwood, Christopher; Madunic, Katarina; Zhang, Tao; Miller, Rebecca L.; Horlacher, Oliver; Struwe, Weston B.; Watanabe, Yu; Okuda, Shujiro; Levander, Fredrik; Kolarich, Daniel; Rudd, Pauline M.; Wuhrer, Manfred; Kettner, Carsten; Packer, Nicolle H.; Aoki-Kinoshita, Kiyoko F.; Lisacek, Frédérique; Karlsson, Niclas G.

*Published in:*  
Nature Communications

*DOI:*  
[10.1038/s41467-019-11131-x](https://doi.org/10.1038/s41467-019-11131-x)

*Publication date:*  
2019

*Document version*  
Publisher's PDF, also known as Version of record

*Document license:*  
[CC BY](https://creativecommons.org/licenses/by/4.0/)

*Citation for published version (APA):*  
Rojas-Macias, M. A., Mariethoz, J., Andersson, P., Jin, C., Venkatakrisnan, V., Aoki, N. P., ... Karlsson, N. G. (2019). Towards a standardized bioinformatics infrastructure for N- and O-glycomics. *Nature Communications*, 10(1), [3275]. <https://doi.org/10.1038/s41467-019-11131-x>

REVIEW ARTICLE

<https://doi.org/10.1038/s41467-019-11131-x>

OPEN

# Towards a standardized bioinformatics infrastructure for *N*- and *O*-glycomics

Miguel A. Rojas-Macias<sup>1</sup>, Julien Mariethoz <sup>2,3</sup>, Peter Andersson<sup>1</sup>,  
Chunsheng Jin<sup>1</sup>, Vignesh Venkatakrisnan <sup>1</sup>, Nobuyuki P. Aoki<sup>4,5</sup>,  
Daisuke Shinmachi <sup>4,5</sup>, Christopher Ashwood <sup>6,7</sup>, Katarina Madunic <sup>8</sup>,  
Tao Zhang<sup>8</sup>, Rebecca L. Miller <sup>9</sup>, Oliver Horlacher<sup>2</sup>, Weston B. Struwe<sup>10</sup>,  
Yu Watanabe<sup>11</sup>, Shujiro Okuda <sup>11</sup>, Fredrik Levander <sup>12</sup>, Daniel Kolarich <sup>13,14</sup>,  
Pauline M. Rudd<sup>15</sup>, Manfred Wuhrer<sup>8</sup>, Carsten Kettner <sup>16</sup>,  
Nicolle H. Packer <sup>6,13,14</sup>, Kiyoko F. Aoki-Kinoshita <sup>4</sup>,  
Frédérique Lisacek <sup>2,3,17</sup> & Niclas G. Karlsson <sup>1</sup>

The mass spectrometry (MS)-based analysis of free polysaccharides and glycans released from proteins, lipids and proteoglycans increasingly relies on databases and software. Here, we review progress in the bioinformatics analysis of protein-released *N*- and *O*-linked glycans (*N*- and *O*-glycomics) and propose an e-infrastructure to overcome current deficits in data and experimental transparency. This workflow enables the standardized submission of MS-based glycomics information into the public repository UniCarb-DR. It implements the MIRAGE (Minimum Requirement for A Glycomics Experiment) reporting guidelines, storage of unprocessed MS data in the GlycoPOST repository and glycan structure registration using the GlyTouCan registry, thereby supporting the development and extension of a glycan structure knowledgebase.

**P**osttranslational modifications of proteins play an essential role in modifying amino acids in proteins, thereby extending their functions and regulating their activities. A census of all possible protein forms, now commonly called proteoforms, was recently estimated<sup>1</sup>. In this renewed view of protein diversity, glycoforms are increasingly being shown to play a major role

<sup>1</sup> Department of Medical Biochemistry and Cell Biology, Institute of Biomedicine, Sahlgrenska Academy, University of Gothenburg, Gothenburg 40530, Sweden. <sup>2</sup> Proteome Informatics Group, SIB Swiss Institute of Bioinformatics, Geneva 1211, Switzerland. <sup>3</sup> Computer Science Department, University of Geneva, Geneva 1227, Switzerland. <sup>4</sup> Soka University, Hachioji, 192-8577 Tokyo, Japan. <sup>5</sup> SparqLite LLC., Hachioji, 192-0032 Tokyo, Japan. <sup>6</sup> Department of Molecular Sciences, Macquarie University, Sydney 2109, Australia. <sup>7</sup> Department of Biochemistry, Medical College of Wisconsin, Milwaukee, WI 53226, USA. <sup>8</sup> Leiden University Medical Center, Leiden 2333ZA, Netherlands. <sup>9</sup> Copenhagen Centre for Glycomics, Department of Cellular and Molecular Medicine, University of Copenhagen, København DK-2200, Denmark. <sup>10</sup> Department of Chemistry, Chemistry Research Laboratory, University of Oxford, Oxford OX1 3TA, UK. <sup>11</sup> Graduate School of Medical and Dental Sciences, Niigata University, 950-2181 Niigata, Japan. <sup>12</sup> National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Department of Immunotechnology, Lund University, Lund 22387, Sweden. <sup>13</sup> Institute for Glycomics, Gold Coast Campus, Griffith University, Gold Coast, QLD QLD 4222, Australia. <sup>14</sup> ARC Centre for Nanoscale BioPhotonics, Macquarie University and Griffith University, North Ryde and Gold Coast, NSW and QLD NSW 2109 and QLD 4222, Australia. <sup>15</sup> Bioprocessing Technology Institute, AStar, Singapore 138668, Singapore. <sup>16</sup> Beilstein-Institut, Frankfurt am Main 60487, Germany. <sup>17</sup> Section of Biology, University of Geneva, Geneva 1211, Switzerland. Correspondence and requests for materials should be addressed to N.G.K. (email: [niclas.karlsson@medkem.gu.se](mailto:niclas.karlsson@medkem.gu.se))

in both health and disease<sup>2–5</sup>. In fact, glycosylation has been shown to be involved in the vast majority of cellular interactions and complex networks. Glycosylation is the basis of most biological events, including protein structural stability, recognition, immunological responses, cancer metastasis and the attachment of pathogens to host cells as the first step in the process of infection<sup>6,7</sup>. Furthermore, the importance of glycosylation is highlighted by the extreme consequences of genetic defects in the glycosylation machinery<sup>8</sup>. Congenital Disorders of Glycosylation are a result of the loss of function of different enzymes involved in *N*-linked and *O*-linked oligosaccharide biosynthesis<sup>9</sup>, resulting in severe illness, organ failure and premature death. The importance of protein glycosylation demands that technologies used for structural determination and function are accurate, robust, and information-rich.

Here, we review the latest mass spectrometry (MS) technology for analysing released *N*- and *O*-linked glycans. We also describe the progress that has been made in glycoinformatics software as well as structural/experimental databases and repositories. As researchers are requested to submit an increasing amount of analytical data into public repositories, we propose a standardized workflow for MS glycomics data recording based on community reporting guidelines and uploading of structural and experimental data to tailor-made databases and repositories.

## Methods and reporting standards for MS-based glycomics

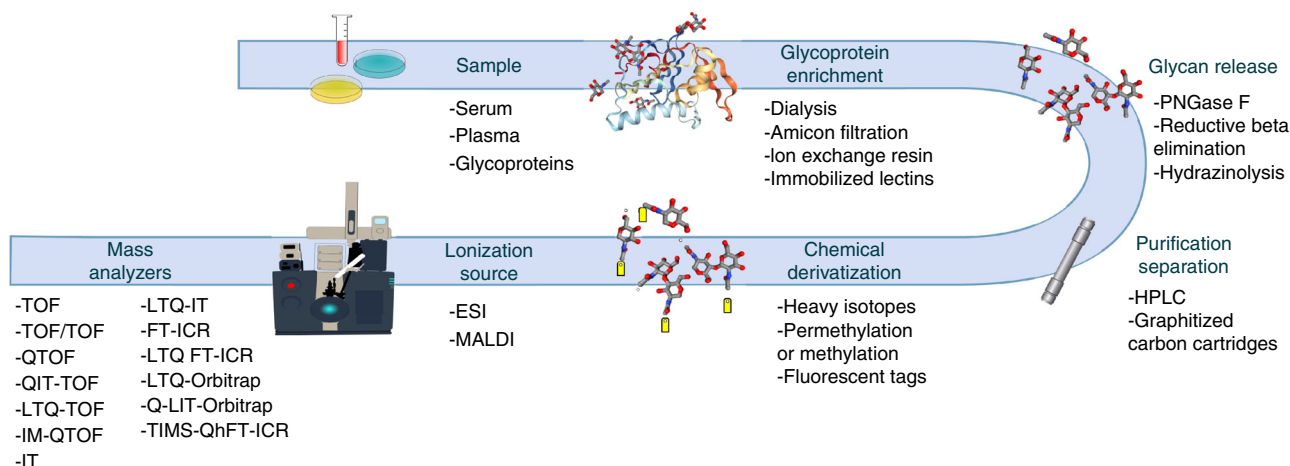
**Structural characterisation of glycans by MS.** At a first glance, MS is not the ideal choice for structural characterisation of glycans. While a precursor mass is sufficient to assign a composition (e.g. the number of constituting hexoses, *N*-acetylhexosamines etc.), it will not allow distinguishing between different isomeric structures, which is one of the major obstacles in the characterisation of glycans. A single mass measurement cannot resolve different isomeric monosaccharides such as glucose, mannose or galactose, nor does MS allow the assignment of pyranose, furanose or linear forms, or differentiation of enantiomers (D or L form).

More detailed insights into glycan structures can be obtained through MS/MS experiments, whereby the glycans are fragmented in the mass spectrometer. MS/MS CID (collision induced

dissociation) and HCD (higher-energy collisional dissociation) fragmentation can help to determine the primary sequence of a glycan, including branching points and elongation. However, additional input is required to identify the glycan linkage position (e.g., 1, 2, 3, 4 and 6) and configuration ( $\alpha$  or  $\beta$ ). Knowledge about the principles and rules of glycan biosynthesis—gained from structural glycobiology work using for instance NMR and studies characterising glycosyltransferase specificity—can reduce the number of conceivable MS assignments. In addition, specific cross-ring fragmentation in CID and HCD can sometimes enable linkage position assignment<sup>10,11</sup>. However, to fully assign a novel structure, a combination of MS, biosynthetic rules, chemical and enzymatic treatment, monosaccharide analysis, retention time and/or NMR is necessary. Multistage MS<sup>n</sup> fragmentation<sup>12</sup>, ion mobility MS<sup>13</sup> and ion spectroscopy<sup>14,15</sup> can also be used to reduce the number of conceivable MS assignments. Furthermore, electron activation fragmentation techniques (referred to as ExD techniques)<sup>16</sup> such as electron capture dissociation (ECD), electron transfer dissociation (ETD), electronic excitation dissociation (EED) and electron detachment dissociation (EDD) have been shown to provide extensive cross ring fragmentation allowing more detailed structural characterisation of glycans.

Despite its limitations, MS has become the central tool for the study of protein glycosylation, largely due to its speed, high sensitivity, partial structural identification and capacity to deal with mixtures, and has been used extensively for glycomics screening/profiling<sup>17,18</sup>. The glycomics profiling of free and/or released glycans by MS has involved the use of a considerable variety of upfront dedicated isolation, derivatization and characterisation techniques that, together with increasingly sophisticated MS instrumentation, has been used to increase speed, depth and efficiency of analysis. A generic glycomics workflow for *N*- and *O*-glycans released from proteins has been described before<sup>19</sup> and is summarized in Fig. 1.

It is important to point out that this review is focused on protein-based glycomics - which is different from glycoproteomics. While glycomics can generate detailed information about glycan structure(s), the methods used to release the glycans from the protein inevitably obliterate the localisation of the glycosylation site within a protein/peptide sequence. Glycoproteomics on



**Fig. 1** General Overview of a glycomics MS workflow. The analytical options at each step are shown below. A similar workflow has been described in ref. <sup>19</sup>. TOF time of flight mass analyzer, TOF/TOF TOF tandem MS, Q quadrupole mass analyzer, QTOF tandem MS combining Q and TOF, IT ion trap mass analyzer, QIT dual operational Q and IT MS, QIT-TOF QIT combined with TOF tandem MS; LTQ(-IT), linear trap quadrupole mass analyser (a linear ion trap), LTQ-TOF tandem MS combining LTQ and TOF, IM ion mobility, IM-QTOF ion mobility cell/QTOF mass spectrometer, FT-ICR Fourier-transform ion cyclotron resonance mass analyzer, LTQ FT-ICR tandem MS combining LTQ and FT-ICR, LTQ-Orbitrap tandem MS combining LTQ and Orbitrap, LIT linear ion trap, Qh quadrupole/hexapole interface, TIMS trapped ion mobility spectrometry

the other hand tries to address glycosylation by analysing intact glyco-peptides/proteins. The caveat is that traditional fragmentation used on (glyco)peptides such as CID and HCD only provide limited information about oligosaccharide structure, and even site localisation can sometimes be difficult due to loss of the entire glycan<sup>20</sup>. In this context, ECD and later on ETD showed to be the fragmentation methods of choice for site localisation of glycans, since the fragmentation occurs primarily in the peptide chain<sup>20,21</sup>. Glycoproteomics analysis, at this stage, therefore can identify the site and mass (composition) of the glycan(s) on a particular site, but provides little details about glycan sequence, branching or linkage.

**Comparative glycomics.** The goal of a glycomic experiment is not always to fully characterise all glycan structures in a sample. Instead, glycomic profiling is often applied to compare samples and focuses on the identification of abnormalities and differences. Several approaches such as MS, HPLC, LC-MS and capillary electrophoresis are used for glycomic profiling. They provide different levels of glycan characterisation ranging from mass profiling, to partial sequence and full structural assignments (based on complementary information), to absolute or relative quantification of individual structures in a biological sample.

In label-free MS-based analyses, the abundances of pseudo-molecular ions (e.g.  $[M - nH]^{n-}$  or  $[M + nNa]^{n+}$  ions) are used to identify differences between samples. Furthermore, several quantitative glycomics methods are based on derivatization approaches and heavy labelled isotopes (reviewed in refs. 22–24). Quantification using stable isotope standards and MS has been shown to provide excellent precision in glycomics<sup>25–27</sup> and glycoproteomics<sup>28</sup>. However, the low number of freely available stable isotope standards is currently the limiting factor for implementing absolute quantitation in MS-based glycomics for a wider range of glycans available from single cells or tissues.

Relative quantification using fluorescent tagging in connection with HPLC or capillary electrophoresis provides the benefit of stoichiometric response from individual glyco components and is the gold standard for glycomic relative quantification<sup>29</sup>. Cross-laboratory comparisons have shown that MS can provide similar quantification results<sup>30,31</sup>, where differences between laboratories can mainly be attributed to differences in sample preparation and data accumulation protocols. This illustrates the need to accurately record protocols for structural assignment of a well-defined sample data and sample handling protocols, and quantitative aspect.

**Adopting omics reporting guidelines for glycomics.** MS based omics entails analysing a multitude of samples generating large amounts of data, and using software to transform these data into biological information. To make this process transparent and reproducible, there is a need for consistent reporting of experimental methods and procedures in publications. Many omics fields have addressed these concerns by developing guidelines for the reporting, collecting and distributing of data and information. This started with MIAME launched for the handling of micro-arrays<sup>32</sup>, followed by the MIAPE guidelines for proteomics<sup>33</sup>, STRENDa in enzymology<sup>34,35</sup>, CIMR in metabolomics<sup>36,37</sup> among others. There are currently more than 150 reporting guidelines published and registered in the FAIRSharing portal<sup>38</sup>.

As was discussed in the previous paragraph, structural characterisation of glycans using only MS is difficult. Multiple guidelines are required for the multiple techniques that are used to convert the analytical data into detailed structural glycomic information. To acknowledge the complexity of glycan structural characterisation, the glycomics community launched the

MIRAGE (Minimum Information Required for A Glycomics Experiment) initiative in 2011. The MIRAGE initiative is formed by experts from the diverse areas of glycomics research and supported by the Beilstein-Institut<sup>39</sup>. Up to now this has resulted in guidelines for glycomics sample preparation<sup>40</sup> (<https://doi.org/10.3762/mirage.1>), MS analysis<sup>41</sup> (<https://doi.org/10.3762/mirage.2>), glycan microarray analysis<sup>42</sup> (<https://doi.org/10.3762/mirage.3>), and liquid chromatography analysis<sup>43</sup> (<https://doi.org/10.3762/mirage.4>). The MS guidelines require not only reporting of experimental conditions, but also disclosure of raw MS data and annotated spectra. Making these guidelines widely applicable require the development of workflows that describe what is to be reported and how to record glycomics MS data. In addition, adoptability of the guidelines requires a web-based software pipeline that facilitates the flow from MS data acquisition to public disclosure of raw data and reporting of data structural interpretation/annotation.

### The current landscape of glycomic e-infrastructures

**Glycan structure repositories.** In 2008, the NIH work group “Frontiers in glycomics” emphasised the need for a curated, sustainably funded glyco-structure database<sup>44</sup>. Accounting for the variety of analytical methods used to assign glycan structures, the proposed structural database was expected to contain associated information about experimental and biosynthetic data. Pioneering attempts to create a comprehensive glycomic database were made in the 1980s with Carbank<sup>45</sup>. The Carbank institutions also implemented regular updates with information from new publications. Unfortunately, a funding crisis stopped this effort in the 1990s and the project was discontinued, but the assembled data lived on in the next-generation databases including SWEET-DB<sup>46</sup> and GlycosuiteDB<sup>47</sup> (later incorporated into UniCarbKB<sup>48</sup> and GlyConnect<sup>49</sup>, both having the same agenda as their ancestor. Then, integrative initiatives arose with the goal of centralising scattered data (e.g. GlycomeDB<sup>50</sup>), as well as combining it with in silico analytical tools such as GLYCOSCIENCES.de<sup>51</sup>, KEGG GLYCAN<sup>52,53</sup> and repositories provided by the Consortium for Functional Glycomics (CFG, <http://www.functionalglycomics.org/fg/>). The progress in the field was somewhat chaotic at the turn of the century, but has significantly evolved lately through a new generation of centralised and integrative resources, each one located on a different continent and being developed in mutual recognition. These are GlyGen in the US (<http://www.glygen.org/>), Glycomics@ExPASy<sup>54</sup> (<https://www.expasy.org/glycomics>) in Europe and GlyCosmos in Japan (<https://glycosmos.org/>). This recent trend may finally provide a long-term solution for stable and financially supported resources for glycobiology. For instance, GlyCosmos includes Gly-TouCan (<https://glytoucan.org/>)<sup>55</sup>, a registry that provides glycan structures with unique identifiers. GlyTouCan provides a foundation for developing complementary repositories, where each unique glycan recorded can be associated with additional experimental information, such as MS data, HPLC retention times and NMR spectra.

**Bioinformatic resources for MS-based glycomics.** To capture information contained in glycomics MS/MS data, UniCarb-DB was launched in 2011<sup>56,57</sup>. Since its introduction, several versions of UniCarb-DB have been released, mainly to improve the glycomics data quality, to increase the number of entries and to advance the usability of the application. UniCarb-DB is currently integrated in Glycomics@ExPASy and provides the framework for accessing experimental MS data that comprise fragmentation spectra, associated structures and metadata about biological origin. Currently, UniCarb-DB contains structural and fragmentation data of O-glycans and N-glycans obtained in positive and

negative MS ion modes. Additional MS fragmentation spectra of glycans are provided in the NIST Glycan Mass Spectral Reference Library (<https://chemdata.nist.gov/glycan/spectra>)<sup>58</sup>.

In parallel to the expansion of glycan structure databases, there has been slow but steady progress in the development of software for glycomics data analysis. The early GlycosidIQ automated the comparison of observed fragments with theoretical glyco-fragments derived from a structural database<sup>59</sup>. This approach has been adopted in commercial software<sup>60</sup>. GlycoReSoft was developed to aid glycan detection from LC-MS runs to compare different samples<sup>61</sup>. Other approaches convert mass spectra into structures relying on spectral libraries<sup>57,62</sup>. More advanced tools for glycomics analysis use partial de novo sequencing<sup>63</sup> including GlycoDeNovo<sup>64</sup> and the recently published Glycoforest<sup>65</sup>. High-throughput glycomics MS annotation tools (GRITS Toolbox<sup>66</sup>, [www.grits-toolbox.org/](http://www.grits-toolbox.org/)) and quantitation tools<sup>67</sup> are now available and increase the need for a common data exchange format. Providing data in an agreed format will help to make data publicly accessible, so that they can be scrutinized by others and used for the validation and curation of glycan structures, for instance those deposited in the GlyTouCan registry.

### A mirage-compatible e-infrastructure for MS-based glycomics

In order to implement the MIRAGE guidelines<sup>39</sup> into an MS-based glycomics e-infrastructure, two existing guidelines were used; (1) glycomics sample preparation<sup>40</sup> and (2) defined MS conditions<sup>41</sup>. The curators also proposed a HPLC experimental module, expanding on the guidelines to enable recording of LC-MS parameters. This section is planned to be expanded since the MIRAGE LC-guidelines recently were published<sup>43</sup>.

This first version of a data recording workflow will focus efforts on the most essential implementation of qualitative, structural information. Quantification guidelines will only be addressed at a superficial level, with expected expansion in subsequent versions. This can be justified considering that workflows for quantitative glycomics are still evolving and that the basic level of methods and software tools is yet to become common practice. Due to the lack of a long-term global public repository for MS glycomics raw data, the requirement to provide this quantitative information as part of the submission has not yet set to be compulsory.

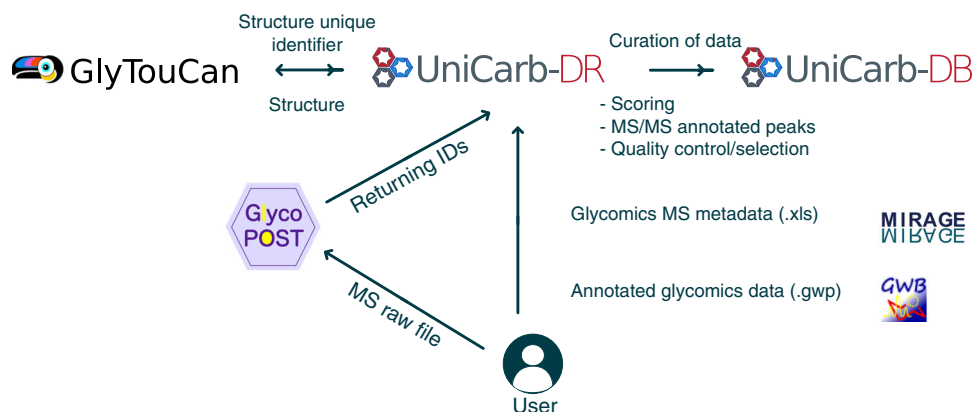
To support an e-workflow, we created the data repository UniCarb-DR (<http://unicarb-dr.biomedicine.gu.se/>) to facilitate submission of glycomics MS<sup>n</sup> data in compliance with the MIRAGE guidelines as part of a publication submission process (Fig. 2). This repository will serve as the interim storage of experimental MS fragment data and structures before data curation and annotation and subsequent transition into the UniCarb-DB database. An author can browse and re-enter submitted data before it is uploaded to the UniCarb-DB repository. We assume that in the near future journals will require data submission to be compulsory prior to publication as for other omics data. Hence, the user can submit the data, referring to it as a “manuscript”. For data uploaded after publication, PubMed ID (PMID), available from <https://www.ncbi.nlm.nih.gov/pubmed/> can be included.

Data deposition in the repository first requires user registration and login at <http://unicarb-dr.biomedicine.gu.se/signup>. Next, the user must provide a number of files and information (Fig. 2) including:

- (1) A compiled file with MIRAGE data (see example spreadsheet in Supplementary Data 1).
- (2) Compiled information about structures (proposed format is GlycoWorkbench)<sup>68</sup>.
- (3) Location of publicly accessible unprocessed MS files.
- (4) Unique structure identifier (this information is automatically generated by communication between UniCarb-DR and the GlyTouCan structural repository<sup>55</sup>).

Further information about these four steps is provided below. A detailed protocol for how to fill in spreadsheets, GlycoWorkbench files and how submit data to UniCarb-DR is available in Supplementary Note 1.

**Step 1: Recording MIRAGE data using webform.** Experimental data needs to be provided in a spreadsheet with data fields reflecting the general structure of the MIRAGE guidelines. Pre-filling and downloading of the MIRAGE compliant spreadsheets are possible in the web form (<http://unicarb-dr.biomedicine.gu.se/generate>). Three different spreadsheets are available: (1) sample preparation, (2) LC and (3) MS guidelines. These can be generated individually or combined into one file containing several sheets (see Supplementary Data 1). These spreadsheets can be



**Fig. 2** Workflow for MIRAGE data submission to UniCarb-DR. The workflow was designed to support the MIRAGE guidelines for sample preparation, including HPLC, MS and annotation. Metadata that are relevant to the entire experiment can be manually recorded in the on-line form accessible on the UniCarb-DR website (<http://unicarb-dr.biomedicine.gu.se/>) that automatically generates spreadsheets (.xls). GlycoWorkBench files (.gwp) are proposed to be used for data annotation of individual structures. MS raw files can be submitted to GlycoPOST and returned IDs can be included in the MIRAGE report. UniCarb-DR also communicates with the global glycostructure repository GlyTouCan to generate unique identifiers for structures submitted to UniCarb-DR. UniCarb-DR will become one source of curated data in UniCarb-DB

modified off-line using common software packages such as Excel. The templates use controlled vocabularies (e.g. for tissue, taxonomy or instrument description) to facilitate user input and simplify data exchange. They can be extended for harmonisation with other standard initiatives such as the HUPO Proteomics Standards Initiative (<http://www.psidev.info>). Additional glyco-related ontologies are proposed (Supplementary Data 2) and will be expanded in line with existing ontologies proposed by the MIRAGE commission and subsequently included in the input form.

**Step 2: Recording structures and MS fragmentation.** The open source software GlycoWorkbench developed within the EuroCarb project to assist manual annotation of MS/MS data<sup>68</sup> can be used to record glycan structures. GlycoWorkbench provides a straightforward interface to draw glycan structures in cartoon formats using the embedded GlycanBuilder module<sup>68</sup>. Glycan structures are stored in a linear format (.gws) for easy parsing and recording into databases. All recorded data can be stored in an XML-type Glycoworkbench file (.gwp file extension) (Supplementary Fig. 1). A Glycoworkbench template is available at <https://unicarb-dr.biomedicine.gu.se/generate> and an example of a filled in file is available in Supplementary Data 3. GlycoWorkbench allows the recording of individual structures as a “Scan” with associated fragment data (fragment list is imported from MS software as centroided data). We suggest utilizing the ability of GlycoWorkbench to record ion trees, using Glyco-workbench “Scans” to record MS<sup>2</sup> (i.e. MS/MS) for each structure, and sub-“Scans” to record MS<sup>3</sup>, MS<sup>4</sup> etc. A GlycoWorkbench file that includes several structures and MS<sup>n</sup> data, for each structure “Scans” and sub-“Scans” needs to be defined directly under the “Workspace” item.

Supplementary Fig. 1 shows the sections that are typically included in a.gwp file. A tag is represented by the “<” and “>” symbols and defines the different elements in a file. These elements are delimited by a start tag e.g. <scan> and an end tag, e.g. </scan>. The example shown in Supplementary Fig. 1 belongs to a single structure somewhat simplified, highlighting important MIRAGE tags. In order to be MIRAGE-compatible, we introduced a “Notes” section for recording orthogonal assignment methods, scoring and validation (see “MIRAGE parameters in UniCarb-DR” below). The format of the “Notes” section needs to be respected in order to upload its content to UniCarb-DR (see Supplementary Methods for more details on the proposed “Notes” format).

**Step 3: Depositing MS raw files.** To host the vast volume of glycomics MS raw data, we propose a model of data sharing similarly to the one implemented in proteomics by the ProteomeXchange consortium<sup>69</sup> that includes PRIDE<sup>70</sup> and JPOST (Japan Proteome STandard Repository/Database), among others. To provide open data access that complies with the MIRAGE requirements, we engaged with JPOST. We developed a pipeline enabling permanent data storage in GlycoPOST (<http://glycopost.glycosmos.org/>), a dedicated repository for MS-based glycomics data. The current model requires submission of MS raw data to GlycoPOST, whilst MIRAGE and GlycoWorkbench files are uploaded and read by the UniCarb-DR submission workflow. To further simplify this process, we are working on integrating submission of raw data, annotated spectra and MIRAGE meta-data in a seamless workflow using both UniCarb-DR and GlycoPOST (Fig. 2). Meanwhile, the applications are streamlined since the same MIRAGE spreadsheets are used in both applications and the user has the option of including GlycoPOST generated URLs of raw data into MIRAGE spreadsheets before

submission to UniCarb-DR. GlycoPOST also accepts and stores other types of files, allowing users to upload information about experimental design, sample log files including quality control samples and blank runs, as well as additional information that is potentially useful for checking the quality and reproducibility of glycomics experiments containing multiple samples. Hence, data repositories and the MIRAGE commission will depend on each other when the next version of guidelines is to be developed. Alternatives to GlycoPost for the long-term storage of MS raw data can potentially be included in the workflow. For instance, we have also used an MS Laboratory Information Management System called Proteios Software Environment (<http://www.proteios.org>)<sup>71</sup> to upload data to Swestore (<http://www.snrc.se/allocations/swestore/>), where Swestore generated Uniform Resource Identifiers (URIs) that have been included as part of scientific publications<sup>72</sup>. In this case, MS raw data are provided both as vendors’ preferred format and as files converted into the open source mzML format<sup>73</sup>. The mzML format is not only describing spectral data but also contains information requested in the MIRAGE guidelines. In the short term, we strongly recommend uploading mzML data separately to GlycoPOST or other repositories for non-vendor software dependent data access. In the long term, this submission of raw MS data can be integrated in the UniCarb-DR upload. We note that the MIRAGE bioinformatics subgroup has considered mzIdentML<sup>74</sup> and mzTab<sup>75</sup> potential formats that could be augmented with glycomics data, and this will also be implemented in the long term.

**Step 4: Registration of submitted glycan structures.** GlyTouCan<sup>55</sup> is a glycan structure repository promoted by the glyco-community as the prime location for generating unique identifiers for individually reported glycan structures and compositions. Glycan structures should be submitted to this repository as part of the publication of glycomics data. To avoid duplicate submissions to both UniCarb-DR and GlyTouCan we have developed a tool that assesses whether the structures submitted to UniCarb-DR are already deposited in GlyTouCan. In this case, the GlyTouCan ID provides a link to UniCarb-DR. If a UniCarb-DR submitted structure is not available in GlyTouCan, a new ID will be generated and communicated to UniCarb-DR. This process will commence after the submission of data to UniCarb-DR.

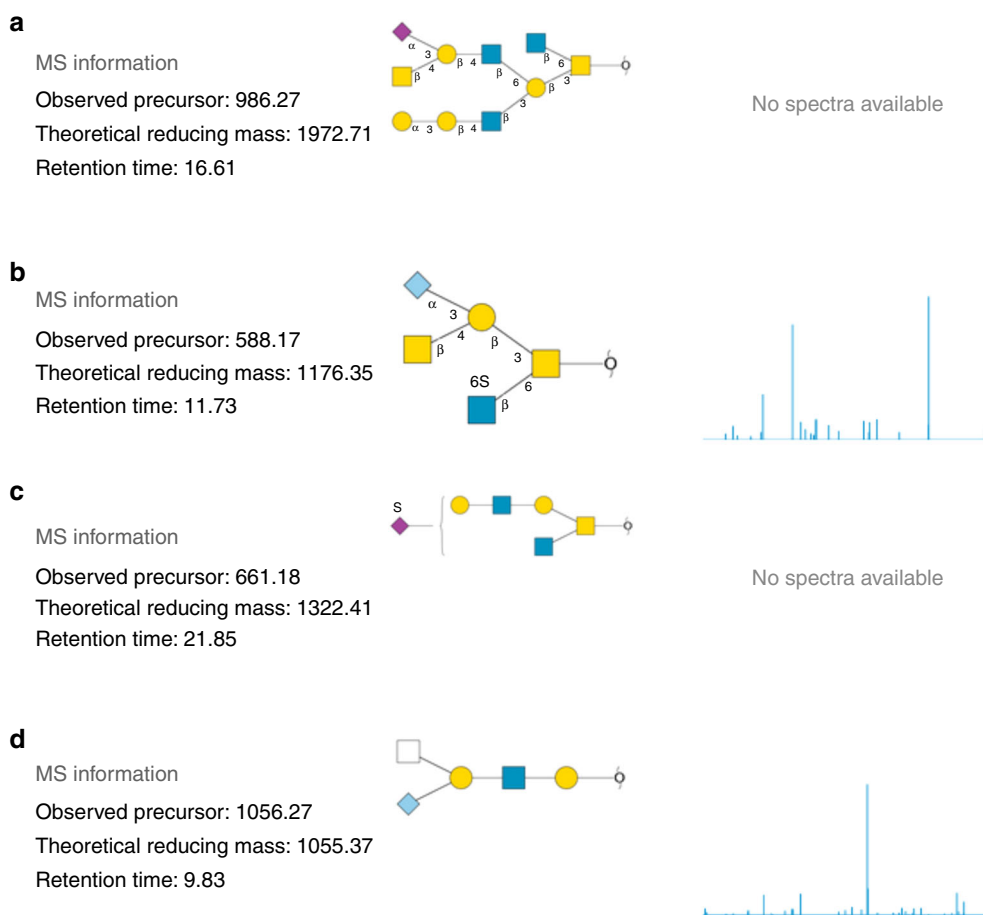
**MIRAGE parameters in UniCarb-DR.** The MIRAGE guidelines are generic and flexible in order to collect information from different types of experiments studying glycoconjugates. However, the use of commonly defined vocabularies is required to compare data within UniCarb-DR and to share data with other glycomics and life science databases. To preserve the flexibility of the MIRAGE guidelines in the reporting process we propose free text fields to describe experiments, whilst a rigorous reporting language is implemented only for key MIRAGE parameters (e.g., tissue, MS device). Inspired by the organization of PRIDE<sup>76</sup>, four different types of formats of the MIRAGE parameters were encoded in UniCarb-DR (Table 1) and outlined in the Supplementary Methods and Supplementary Data 2. To comply with the controlled vocabulary but still enable glossary update, new terms can be suggested by sending a request to administrators of UniCarb-DR at <http://unicarb-dr.biomedicine.gu.se/about>.

**Upload of MIRAGE compatible MS/MS spectra to UniCarb-DR.** MIRAGE-compliant data sets along with data stored in.gwp files of both individual intact structures and fragmentation spectra can be submitted as supplementary material associated with a publication. We also propose uploading these collected and structured glycomics information (spreadsheets and.gwp files) to

**Table 1** Examples of parameters identified and included in the MIRAGE guidelines

Defined formats	Existing ontologies	New parameters in glycoanalytics	Free text
Unique identifiers for structure and entry Dates	Cell lines <a href="http://www.clo-ontology.org/">www.clo-ontology.org/</a> Species <a href="http://www.ncbi.nlm.nih.gov/taxonomy">www.ncbi.nlm.nih.gov/taxonomy</a>	Treatments and orthogonal methods for isolation/characterisation Software used for data processing and structure assignment MS scoring method	Growth/harvest conditions for recombinantly produced material Treatments and/or storage conditions for material isolated from tissues Synthesis steps for chemically derived material Purification steps
Oligosaccharide structure (.gws, glyco-CT <sup>84</sup> and WURCS <sup>85</sup> )	Tissues <a href="http://meshb.nlm.nih.gov/record/ui?ui=D014024">meshb.nlm.nih.gov/record/ui?ui=D014024</a> MS devices <a href="http://www.ebi.ac.uk/ols/ontologies/ms">www.ebi.ac.uk/ols/ontologies/ms</a> Software <a href="http://www.ebi.ac.uk/ols/ontologies/ms">www.ebi.ac.uk/ols/ontologies/ms</a> Attached proteins <a href="http://www.uniprot.org/">www.uniprot.org/</a>	Validation of assigned structure	Sprayer features  Voltages and other parameters relevant for ESI experiments CID/HCD settings

The columns describe (from left to right): defined formats to control the input and presentation of certain data, ontologies or existing vocabulary that can be adopted from various sources, parameters that need to be expanded into ontologies or controlled vocabulary to use in glycomics, free unrestricted text



**Fig. 3** Different types of entries managed by the UniCarb-DR submission tool. The glycan structures are displayed in SNFG cartoon notation<sup>83</sup>. **a** Structure proposed from LC-MS without an LC-MS/MS spectrum. **b** Fully assigned structure with an associated LC-MS/MS spectrum. **c** Partially assigned structure proposed from LC-MS without an LC-MS/MS spectrum. **d** Partially assigned structure with an associated LC-MS/MS spectrum

<http://unicarb-dr.biomedicine.gu.se/uploadData>. Before uploading, the user is required to register at <http://unicarb-dr.biomedicine.gu.se/signup>. The database allows structures of full and partial assignment to be uploaded (Fig. 3). The reporting of orthogonal methods (i.e. NMR, HPLC retention time mapping, and chemical/enzymatic treatment) is also possible and justifies

the fact that UniCarb-DR can be used to accept structures where MS, but not MS<sup>n</sup> data, has been collected. Figure 3, displays examples of assigned structures in UniCarb-DR (<http://unicarb-dr.biomedicine.gu.se/references/1>), both with and without associated fragmentation data. In the latter case, structures were assigned based on retention time (RT) and biosynthetic

knowledge about the constituting monosaccharides, linkage position and configuration. Observed RT is of course of limited use outside a particular experiment. Hence, we promote recording relative RT using external RT markers that generate, for instance, a nominal size corresponding to number of glucose units (GU)<sup>77</sup> or relative to an internal common landmark oligosaccharide<sup>78</sup>.

We have assembled an expandable list of treatments and orthogonal methods commonly used for isolation/characterisation in glycomics experiments (Table 1 and Supplementary Data 2). The current records in UniCarb-DR have been uploaded using data generated in various laboratories by researchers in the author list. During this process we found that the requirement to record the full information about individual structures (e.g. scoring and orthogonal method validation) is time consuming due to lack of software, and is often not feasible. Hence, UniCarb-DR is also accepting data with partial MIRAGE records for an individual structure, requiring only the record of the precursor ion mass but no information about scoring and validation. Many authors of this paper are part of the MIRAGE committee. In light of these limitations, we have proposed to only require partial MIRAGE compliance for submitted records of an individual structure, with the commitment to move towards full MIRAGE compliance as glycomics software develops further.

### Discussion and future perspectives

The lack of an established formalised description of glycomics experiments may stall progress in the glycobiology field. The research community has always relied on sharing scientific results. Here, we are proposing a solution for sharing glycomics structures with associated MS experimental conditions and data. For this we are using spreadsheets in combination with GlycoWorkbench files. This format is a step towards enforcing MIRAGE-compliant scientific publications in glycomics. Past experience in introducing guidelines for glycomics studies as part of publications<sup>79</sup> has shown that, if there is a clear pathway and format, researchers will conform to get their manuscripts published in quality journals. With the tools presented in this report, glycomics MS/MS<sup>n</sup> reporting standards can be adopted at an early stage of a project. The spreadsheet can be completed and modified as the project unfolds and the use of GlycoWorkbench files for saving glycomics structural interpretations can be implemented for data housekeeping. Both the spreadsheet and the.gwp formats are flexible enough to support a variety of glycomics MS, requiring only limited modifications of templates provided. Hence, journal editors will be in the position to ensure MIRAGE compliance by requesting that authors provide these template files as supplementary data. The MIRAGE committee is continuously corresponding with relevant journal editors on novel MIRAGE protocols, software tools and repositories to implement MIRAGE guidelines as part of the publication requirements. With an increasing awareness of MIRAGE formats and associated enabling technologies, we expect that an increasing number of researchers and reviewers will insist that, not only their own, but also other groups' data are MIRAGE-compliant in scientific publications.

The use of spreadsheets (generated from web form, see Supplementary Data 1), deposition of raw data (e.g. in GlycoPost) and the GlycoWorkbench.gwp format are supporting the upload of glycomics MS metadata to UniCarb-DR. This workflow has been tested for datasets of intact as well as reducing-end-derivatized glycans, which were analysed by MS, LC-MS and -MS/MS in negative and positive ion modes using CID and HCD fragmentation. All the data can be assembled manually, allowing the workflow to be used by both beginners in glycomics as well as

advanced glycomics MS institutes. However, another purpose of defining the upload format is to provide a template for the output from software-aided glycomics discovery pipelines. The GlycoWorkbench structure format has already been adopted in other glycomics commercial (GlycoQuest, Bruker, Bremen Germany) and academic (GRITS Toolbox (<http://www.grits-toolbox.org/>)) software projects<sup>66</sup>. Hence, automated submission to UniCarb-DR is likely to be easily implemented for these tools.

The focus of this report is on *N*- and *O*-linked glycans released from proteins and their spectra generated by CID and HCD. This is because the vast majority of data currently available are *N*- and *O*-linked CID fragmentation data using positive and negative ion modes. UniCarb-DR will also be able to host CID/HCD data of free glycans, glycans released from glycolipids and glycosaminoglycans. In principle, the repository could also host data of intact glycolipids, but the formalisation of the aglycon is currently missing in GlycoWorkbench. The lack of detailed structural information about the glycan moiety in global glycoproteomics data complicates recording of these data with the workflow presented here. However, the main obstacle for including glycoproteomics data is that UniCarb-DR currently is not capturing peptide sequences and glycosite information, and is not associating data to a protein. For the time being, glycopeptide MS data is being collected in databases such as MS-Viewer<sup>80</sup> and, as partially curated data, in GlyConnect<sup>49</sup> where they are integrated with multiple related sources of information on the recorded attached glycan composition.

Other workflows utilized in glycomics, such as permethylation and other type of derivatisation recordable in GlycoWorkbench followed by MS with or without coupled LC separation, are easily implemented if the spectra are from single isomers. With several isomers present in one spectrum, these data can still be recorded in GlycoWorkbench (several structures recorded in one "Scan"), but the UniCarb-DR format will need to be modified in future versions to enable easy upload of this mixed-structure data. Similar concerns apply to workflows involving multi-stage MS<sup>n</sup>, despite the flexibility of recording sub-"Scans" in GlycoWorkbench. Glycomics MS workflows including ion mobility MS will require updating the spreadsheets and GlycoWorkbench files with information about collisional cross sections and ion-mobility parameters currently not considered in MIRAGE or will need to link to databases that contain cross section information from carbohydrates such as GlycoMob<sup>81</sup>. Fragmentation data generated by ExD or other type of fragmentation techniques producing non-standard fragments cannot be recorded in GlycoWorkbench and will require further adjustments of the UniCarb-DR upload procedure. This would involve expanding the type and associated metadata for individual fragments. Currently, non-standard fragment peaklist can be uploaded, without annotation of non-standard fragments. We also request help from the community to identify additional major glycomics workflows for us to adapt the data submission accordingly.

In addition to adapting the submission process to a broader range of experimental workflows, we are also aiming to automate submission to UniCarb-DR. To this end, we plan to accept direct submissions from glycan structure assignment tools such as Glycoforest<sup>65</sup>. In 2020, a web version of Glycoforest will manage the automation of structure assignment to MS/MS spectra. Glycoforest first generates consensus spectra from the MS/MS data, and then assigns structures to the consensus spectra. The resulting assignments can be manually checked and if necessary corrected by the user. The direct submission of spectra and their associated assignments will contribute to the expansion of UniCarb-DR and help reduce human error in the submission process.



The commitment to store glycomics MS datasets is essential. The annotation of glycomics LC-MS data is currently based only on the knowledge of the interpreters<sup>65</sup>, be it a software or a human researcher or both. From this it can be concluded that it is highly unlikely that all information from a glycomics raw data set will be extracted in a single analysis. Hence, glycomics raw MS data should be considered as libraries that will be re-analysed to harvest new knowledge and to ask new questions. This is even more important if glycomics evolves similarly to proteomics and will increasingly rely on data independent acquisition<sup>82</sup> in addition to data depend acquisition as a means to generate data from clinical or other reference samples. These glycomics libraries will provide essential information for hypothesis-driven glycomics. Similar to PRIDE and ProteomeXchange initiatives, the glycomics community needs to voice the unanimous opinion that this is needed, and target both national and international life science e-infrastructure organizations and journals. The MIRAGE committee already identified this requirement by introducing the recommendation for raw data deposition in the guidelines.

A pipeline for curation of experimental data from data repositories into databases is changing how curated structural databases will be generated. The previous top-down approach of a database generator and curator searching literature for information will shift to researchers submitting and managing their own data. Researchers and curators will need software tools to help in the curation process. The metadata in the reporting guidelines and evaluation of the accompanying publication, complemented with present and future biosynthetic knowledge, will aid the curation process. This process must remain objective and transparent in that information can only be added, but not deleted or altered (unless permitted by the data supplier). The MIRAGE guidelines can only be strengthened by such an approach that supports the unbiased assessment of data quality. The mission of UniCarb-DR and Unicarb-DB is to support the development of a knowledgebase of glycan structures by providing the pipeline for storage and curation of glycomics experimental MS data.

Received: 24 January 2019 Accepted: 24 June 2019

Published online: 22 July 2019

## References

1. Aebersold, R. et al. How many human proteoforms are there? *Nat. Chem. Biol.* **14**, 206–214 (2018). **A viewpoint paper of how post translational modifications are one of the main contributors to generate the huge variety of proteoforms from a single gene.**
2. Everest-Dass, A. V., Moh, E. S. X., Ashwood, C., Shathili, A. M. M. & Packer, N. H. Human disease glycomics: technology advances enabling protein glycosylation analysis - part 1. *Expert Rev. Proteom.* **15**, 165–182 (2018).
3. Everest-Dass, A. V., Moh, E. S. X., Ashwood, C., Shathili, A. M. M. & Packer, N. H. Human disease glycomics: technology advances enabling protein glycosylation analysis - part 2. *Expert Rev. Proteom.* **15**, 341–352 (2018).
4. Kailemia, M. J., Park, D. & Lebrilla, C. B. Glycans and glycoproteins as specific biomarkers for cancer. *Anal. Bioanal. Chem.* **409**, 395–410 (2017).
5. Bennis, S. V. et al. Systems glycobiochemistry: integrating glycogenomics, glycoproteomics, glycomics, and other 'omics data sets to characterize cellular glycosylation processes. *J. Mol. Biol.* **428**, 3337–3352 (2016).
6. Varki, A. Glycan-based interactions involving vertebrate sialic-acid-recognizing proteins. *Nature* **446**, 1023–1029 (2007).
7. Rudd, P. M., Elliott, T., Cresswell, P., Wilson, I. A. & Dwek, R. A. Glycosylation and the immune system. *Science* **291**, 2370–2376 (2001).
8. Sparks, S. E. & Krasnewich, D. M. Congenital Disorders of N-Linked Glycosylation and Multiple Pathway Overview. in *GeneReviews*<sup>®</sup> (eds. Adam MP et al.) (University of Washington, Seattle, Washington, 1993).
9. Jaeken, J. & Matthijs, G. Congenital disorders of glycosylation. *Annu Rev. Genom. Hum. Genet.* **2**, 129–151 (2001).
10. Karlsson, N. G., Schulz, B. L. & Packer, N. H. Structural determination of neutral O-linked oligosaccharide alditols by negative ion LC-electrospray-MSn. *J. Am. Soc. Mass Spectrom.* **15**, 659–672 (2004).
11. Everest-Dass, A. V., Abrahams, J. L., Kolarich, D., Packer, N. H. & Campbell, M. P. Structural feature ions for distinguishing N- and O-linked glycan isomers by LC-ESI-IT MS/MS. *J. Am. Soc. Mass Spectrom.* **24**, 895–906 (2013).
12. Ashline, D. J., Zhang, H. & Reinhold, V. N. Isomeric complexity of glycosylation documented by MS(n). *Anal. Bioanal. Chem.* **409**, 439–451 (2017).
13. Harvey, D. J. & Struwe, W. B. Structural studies of fucosylated N-glycans by ion mobility mass spectrometry and collision-induced fragmentation of negative ions. *J. Am. Soc. Mass Spectrom.* **29**, 1179–1193 (2018).
14. Schindler, B. et al. Anomeric memory of the glycosidic bond upon fragmentation and its consequences for carbohydrate sequencing. *Nat. Commun.* **8**, 973 (2017).
15. Mucha, E. et al. Glycan fingerprinting via cold-ion infrared spectroscopy. *Angew. Chem. Int. Ed. Engl.* **56**, 11248–11251 (2017).
16. Pu, Y. et al. Separation and identification of isomeric glycans by selected accumulation-trapped ion mobility spectrometry-electron activated dissociation tandem mass spectrometry. *Anal. Chem.* **88**, 3440–3443 (2016).
17. Wuhler, M. Glycomics using mass spectrometry. *Glycoconj. J.* **30**, 11–22 (2013).
18. Zaia, J. Mass spectrometry and the emerging field of glycomics. *Chem. Biol.* **15**, 881–892 (2008).
19. An, H. J., Kronewitter, S. R., de Leoz, M. L. & Lebrilla, C. B. Glycomics and disease markers. *Curr. Opin. Chem. Biol.* **13**, 601–607 (2009).
20. Rudd, P. et al. Chapter 51 Glycomics and Glycoproteomics. In: *Essentials of Glycobiology* (eds Varki A. et al.) (Cold Spring Harbor (NY) (2015). **One of many basic chapters in THE textbook for glycobiology.**
21. Hakansson, K. et al. Electron capture dissociation and infrared multiphoton dissociation MS/MS of an N-glycosylated tryptic peptide to yield complementary sequence information. *Anal. Chem.* **73**, 4530–4536 (2001).
22. Veillon, L., Zhou, S. & Mechref, Y. Quantitative glycomics: a combined analytical and bioinformatics approach. *Methods Enzym.* **585**, 431–477 (2017).
23. Moh, E. S., Thaysen-Andersen, M. & Packer, N. H. Relative versus absolute quantitation in disease glycomics. *Proteom. Clin. Appl.* **9**, 368–382 (2015).
24. Orlando, R. Quantitative analysis of glycoprotein glycans. *Methods Mol. Biol.* **951**, 197–215 (2013).
25. Hecht, E. S., McCord, J. P. & Muddiman, D. C. Definitive screening design optimization of mass spectrometry parameters for sensitive comparison of filter and solid phase extraction purified, INLIGHT plasma N-glycans. *Anal. Chem.* **87**, 7305–7312 (2015).
26. Meitei, N. S., Apte, A., Snovida, S. I., Rogers, J. C. & Saba, J. Automating mass spectrometry-based quantitative glycomics using aminoxy tandem mass tag reagents with SimGlycan. *J. Proteom.* **127**, 211–222 (2015).
27. Reiding, K. R. et al. High-throughput serum N-glycomics: method comparison and application to study rheumatoid arthritis and pregnancy-associated changes. *Mol. Cell Proteom.* **18**, 3–15 (2019).
28. Chen, Z. et al. Site-specific characterization and quantitation of N-glycopeptides in PKM2 knockout breast cancer cells using DiLeu isobaric tags enabled by electron-transfer/higher-energy collision dissociation (ETHcD). *Analyst* **143**, 2508–2519 (2018).
29. Royle, L. et al. HPLC-based analysis of serum N-glycans on a 96-well plate platform with dedicated database software. *Anal. Biochem.* **376**, 1–12 (2008).
30. Wada, Y. et al. Comparison of the methods for profiling glycoprotein glycans—HUPO Human Disease Glycomics/Proteome Initiative multi-institutional study. *Glycobiology* **17**, 411–422 (2007).
31. Ito, H. et al. Comparison of analytical methods for profiling N- and O-linked glycans from cultured cell lines: HUPO Human Disease Glycomics/Proteome Initiative multi-institutional study. *Glycoconj. J.* **33**, 405–415 (2016).
32. Brazma, A. et al. Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat. Genet.* **29**, 365–371 (2001).
33. Taylor, C. F. et al. The minimum information about a proteomics experiment (MIAPE). *Nat. Biotechnol.* **25**, 887–893 (2007).
34. Tipton, K. F. et al. Standards for Reporting Enzyme Data: The STREND Consortium: What it aims to do and why it should be helpful. *Perspect. Sci.* **1**, 131–137 (2014).
35. Apweiler, R. et al. The importance of uniformity in reporting protein-function data. *Trends Biochem. Sci.* **30**, 11–12 (2005).
36. Jenkins, H. et al. A proposed framework for the description of plant metabolomics experiments and their results. *Nat. Biotechnol.* **22**, 1601–1606 (2004).
37. Goodacre, R. et al. Proposed minimum reporting standards for data analysis in metabolomics. *Metabolomics* **3**, 231–241 (2007).
38. McQuilton, P. et al. BioSharing: curated and crowd-sourced metadata standards, databases and data policies in the life sciences. *Database* **2016**, baw075–baw075 (2016).
39. York, W. S. et al. MIRAGE: the minimum information required for a glycomics experiment. *Glycobiology* **24**, 402–406 (2014).
40. Struwe, W. B. et al. The minimum information required for a glycomics experiment (MIRAGE) project: sample preparation guidelines for reliable reporting of glycomics datasets. *Glycobiology* **26**, 907–910 (2016).

41. Kolarich, D. et al. The minimum information required for a glycomics experiment (MIRAGE) project: improving the standards for reporting mass-spectrometry-based glycoanalytic data. *Mol. Cell Proteom.* **12**, 991–995 (2013).
42. Liu, Y. et al. The minimum information required for a glycomics experiment (MIRAGE) project: improving the standards for reporting glycan microarray-based data. *Glycobiology* **27**, 280–284 (2017).
43. Campbell, M. P. et al. The minimum information required for a glycomics experiment (MIRAGE) project: LC guidelines. *Glycobiology* **29**, 349–354 (2019).
44. Packer, N. H. et al. Frontiers in glycomics: bioinformatics and biomarkers in disease. An NIH white paper prepared from discussions by the focus groups at a workshop on the NIH campus, Bethesda MD (September 11–13, 2006). *Proteomics* **8**, 8–20 (2008).
45. Doubet, S., Bock, K., Smith, D., Darvill, A. & Albersheim, P. The Complex Carbohydrate Structure Database. *Trends Biochem Sci.* **14**, 475–477 (1989). **Pioneering publication of a glyco database that evolved into CARBANK.**
46. Loss, A. et al. SWEET-DB: an attempt to create annotated data collections for carbohydrates. *Nucleic Acids Res.* **30**, 405–408 (2002).
47. Cooper, C. A., Harrison, M. J., Wilkins, M. R. & Packer, N. H. GlycoSuiteDB: a new curated relational database of glycoprotein glycan structures and their biological sources. *Nucleic Acids Res.* **29**, 332–335 (2001).
48. Campbell, M. P. et al. UniCarbKB: building a knowledge platform for glycoproteomics. *Nucleic Acids Res.* **42**, D215–D221 (2014).
49. Alocci, D. et al. GlyConnect: glycoproteomics goes visual, interactive and analytical. *J. Proteome Res.* **18**, 664–677 (2019).
50. Ranzinger, R., Herget, S., Wetter, T. & von der Lieth, C. W. GlycomeDB - integration of open-access carbohydrate structure databases. *BMC Bioinforma.* **9**, 384 (2008).
51. Lutteke, T. et al. GLYCOSCIENCES.de: an Internet portal to support glycomics and glycobiology research. *Glycobiology* **16**, 71R–81R (2006).
52. Hashimoto, K. et al. KEGG as a glycome informatics resource. *Glycobiology* **16**, 63R–70R (2006).
53. Kanehisa M. KEGG GLYCAN. in *A Practical Guide to Using Glycomics Databases*. (ed Aoki-Kinoshita, K. F.) 177–193 (Springer, Japan, 2017).
54. Mariethoz, J. et al. Glycomics@ExpASY: Bridging the Gap. *Mol. Cell Proteom.* **17**, 2164–2176 (2018).
55. Tiemeyer, M. et al. GlyYouCan: an accessible glycan structure repository. *Glycobiology* **27**, 915–919 (2017). **Development of a glycan repository that has been embraced by the glycomics community and provides a unique identifier for a recorded glycan structure.**
56. Hayes, C. A. et al. UniCarb-DB: a database resource for glycomics discovery. *Bioinformatics* **27**, 1343–1344 (2011). **Proposal of the UniCarb-DB database, the predecessor of the data repository UniCarb-DR proposed in this article.**
57. Campbell, M. P. et al. Validation of the curation pipeline of UniCarb-DB: building a global glycan reference MS/MS repository. *Biochim. Biophys. Acta* **1844**, 108–116 (2014).
58. Remoroza, C. A., Mak, T. D., De Leoz, M. L. A., Mirokhin, Y. A. & Stein, S. E. Creating a mass spectral reference library for oligosaccharides in human milk. *Anal. Chem.* **90**, 8977–8988 (2018).
59. Joshi, H. J. et al. Development of a mass fingerprinting tool for automated interpretation of oligosaccharide fragmentation data. *Proteomics* **4**, 1650–1664 (2004).
60. Apte, A. & Meitei, N. S. Bioinformatics in glycomics: glycan characterization with mass spectrometric data using SimGlycan. *Methods Mol. Biol.* **600**, 269–281 (2010).
61. Maxwell, E. et al. GlycReSoft: a software package for automated recognition of glycans from LC/MS data. *PLoS ONE* **7**, e45474 (2012).
62. Ashline, D. J., Hanneman, A. J., Zhang, H. & Reinhold, V. N. Structural documentation of glycan epitopes: sequential mass spectrometry and spectral matching. *J. Am. Soc. Mass Spectrom.* **25**, 444–453 (2014).
63. Sun, W., Lajoie, G.A., Ma, B. & Zhang, K. *Bioinformatics Research and Applications. ISBRA704. Lecture Notes in Computer Science*, Vol 9096 (eds Harrison R, et al.) (Springer, Cham, 2015).
64. Hong, P. et al. GlycoDeNovo - an Efficient Algorithm for Accurate de novo Glycan Topology Reconstruction from Tandem Mass Spectra. *J. Am. Soc. Mass Spectrom.* **28**, 2288–2301 (2017).
65. Horlacher O. et al. Glycoforest 1.0. *Anal. Chem.* **89**, 10932–10940 (2017).
66. Weatherly, D. B. et al. GRITS Toolbox—a freely available software for processing, annotating and archiving glycomics mass spectrometry data. *Glycobiology* **29**, 452–460 (2019).
67. Jansen, B. C. et al. MassyTools: A High-Throughput Targeted Data Processing Tool for Relative Quantitation and Quality Control Developed for Glycomics and Glycoproteomic MALDI-MS. *J. Proteome Res.* **14**, 5088–5098 (2015).
68. Damerell, D. et al. The GlycanBuilder and GlycoWorkbench glycomics informatics tools: updates and new developments. *Biol. Chem.* **393**, 1357–1362 (2012). **Development of GlycoWorkbench, which is instrumental for this publication.**
69. Vizcaino, J. A. et al. ProteomeXchange provides globally coordinated proteomics data submission and dissemination. *Nat. Biotechnol.* **32**, 223–226 (2014).
70. Vizcaino, J. A. et al. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **44**, D447–D456 (2016).
71. Hakkinen, J., Vincic, G., Mansson, O., Warell, K. & Levander, F. The proteios software environment: an extensible multiuser platform for management and analysis of proteomics data. *J. Proteome Res.* **8**, 3037–3043 (2009).
72. Jin, C. et al. Structural Diversity of Human Gastric Mucin Glycans. *Mol. Cell Proteom.* **16**, 743–758 (2017).
73. Martens, L. et al. mzML—a community standard for mass spectrometry data. *Mol. Cell Proteom.* **10**, R110 000133 (2011).
74. Jones, A. R. et al. The mzIdentML data standard for mass spectrometry-based proteomics results. *Mol. Cell Proteom.* **11**, M111 014381 (2012).
75. Griss, J. et al. The mzTab data exchange format: communicating mass-spectrometry-based proteomics and metabolomics experimental results to a wider audience. *Mol. Cell Proteom.* **13**, 2765–2775 (2014).
76. Jones, P. et al. PRIDE: a public repository of protein and peptide identifications for the proteomics community. *Nucleic Acids Res.* **34**, D659–D663 (2006).
77. Guile, G. R., Rudd, P. M., Wing, D. R., Prime, S. B. & Dwek, R. A. A rapid high-resolution high-performance liquid chromatographic method for separating glycan mixtures and analyzing oligosaccharide profiles. *Anal. Biochem.* **240**, 210–226 (1996).
78. Abrahams, J. L., Campbell, M. P. & Packer, N. H. Building a PGC-LC-MS N-glycan retention library and elution mapping resource. *Glycoconj. J.* **35**, 15–29 (2018).
79. Wells, L. & Hart, G. W. Athens Guidelines for the Publication of Glycomics D. Glycomics: building upon proteomics to advance glycosciences. *Mol. Cell Proteom.* **12**, 833–835 (2013).
80. Baker, P. R. & Chalkley, R. J. MS-viewer: a web-based spectral viewer for proteomics results. *Mol. Cell Proteom.* **13**, 1392–1396 (2014).
81. Struwe, W. B., Pagel, K., Benesch, J. L., Harvey, D. J. & Campbell, M. P. GlycoMob: an ion mobility-mass spectrometry collision cross section database for glycomics. *Glycoconj. J.* **33**, 399–404 (2016).
82. Gillet, L. C., Leitner, A. & Aebersold, R. Mass spectrometry applied to bottom-up proteomics: entering the high-throughput era for hypothesis testing. *Annu Rev. Anal. Chem.* **9**, 449–472 (2016).
83. Varki, A. et al. Symbol Nomenclature for Graphical Representations of Glycans. *Glycobiology* **25**, 1323–1324 (2015). **The proposal of cartoons to represent glycans, a popular strategy to communicate glycomics and glycoproteomic results to a broad audience.**
84. Herget, S., Ranzinger, R., Maass, K. & Lieth, C. W. GlycoCT—a unifying sequence format for carbohydrates. *Carbohydr. Res.* **343**, 2162–2171 (2008).
85. Tanaka, K. et al. WURCS: the Web3 unique representation of carbohydrate structures. *J. Chem. Inf. Model* **54**, 1558–1566 (2014).

## Acknowledgements

This work was financed by the European Union FP7 GastricGlycoExplorer ITN (No 316929), the Swedish Research Council (621–2013–5895), The Swedish Foundation for International Cooperation in Research and Higher Education (STINT) initiation grant (IB2015–5931) and institutional grant (IG2010–2050). The setup of UniCarb-DR was supported and promoted by the Swedish Infrastructure for Biological Mass Spectrometry (BioMS, [www.bioms.se](http://www.bioms.se)) supported by the Swedish Research Council. SIB is supported by the Swiss Federal Government through the State Secretariat for Education, Research and Innovation (SERI). ExpASY is maintained by and hosted at the SIB Swiss Institute of Bioinformatics. The MIRAGE project is supported by Beilstein-Institut. GlyYouCan and jPOST are supported by the Integrated Database Coordination Program of the National Bioscience Database Center (Japan) and Japan Science and Technology Agency.

## Author contributions

M.A.R.-M. designed and led the development of the UniCarb-DB database and UniCarb-DR repository; J.M. O.H. and P.A. contributed to the UniCarb-DR core implementation and the web interface; C.J. and N.G.K. contributed to the design of the upload forms, C.J., V.V., K.M., C.A., R.L.M. and T.Z. contributed with data for beta-testing of the data submission tool and feedback; C.J. wrote the manual for MIRAGE file generation and submission; N.P.A., D.S. and K.F.A.-K. contributed to the setting up the communication with GlyYouCan; F. Li. contributed with design of Proteios for glyco MS data; Y.W., S.O. and K.F.A.-K. established the harmonization with GlycoPost; W.B.S., D.K., P.M.R., M. W., C.K., N.H.P., K.F.A.-K., F. Li. and N.G.K. formed the advisory core that contributed with feedback on repository design, advised in implementation of glycomics MS/reporting guidelines and design of databases and coding; K.F.A.-K. and N.G.K. jointly received the funds for UniCarb-DR and GlyYouCan communication; F. Li. and N.G.K. jointly received funds for development of the UniCarb-DR infrastructure and structured the manuscript; N.G.K. proposed the UniCarb-DR concept and managed the project in collaboration with F. Li, M.A.R.-M., F.Li, K.F.A.-K., N.H.P. and N.G.K. jointly wrote the manuscript with input from the other authors. All authors signed off on the final version.

**Additional information**

**Supplementary Information** accompanies this paper at <https://doi.org/10.1038/s41467-019-11131-x>.

**Competing interests:** The authors declare no competing interests.

**Reprints and permission** information is available online at <http://npg.nature.com/reprintsandpermissions/>

**Peer review information:** *Nature Communications* thanks David C. Muddiman, Catherine Costello and other anonymous reviewer(s) for their contribution to the peer review of this work.

**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) 2019