

Proof Delivery Form

Journal and Article number: BCJ-2018-0405

Number of pages (not including this page): 18

Biochemical Journal

Please check your proof carefully to ensure (a) accuracy of the content and (b) that no errors have been introduced during the production process.

- You are responsible for ensuring that any errors contained in this proof are marked for correction before publication. Errors not marked may appear in the published paper.
- Corrections should only be for typographical errors in the text or errors in the artwork; substantial revision of the substance of the text is not permitted.
- Please answer any queries listed below.
- If corrections are required to a figure, please supply a new copy.

Your proof corrections and query answers should be returned as soon as possible (ideally within 48 hours of receipt). Please upload your corrected proof and any additional files (e.g. artwork) via the online proof review page from which you downloaded this file. You can also provide any specific instructions or comments in the 'Response Comments' box on the online page.

Notes:

1. Please provide the paper's reference number in any correspondence about your article
2. If you have any queries, please contact the publisher by email (production@portlandpress.com) or by telephone +44 (0)20 7685 2410

Supplementary Material:



This proof does not contain any supplementary material. If supplementary content is associated with this article it will be published, in the format supplied by the authors, with the online version of the article.

Queries for author:

- Q1: Please review the highlighting of the author surnames. Bearing in mind that this will direct how the authors are indexed by the journal and PubMed, please confirm if this is correct or indicate any changes that are needed.
- Q2: References are not in sequential order. Please arrange the same in sequence.
- Q3: The abbreviations which are not used frequently in the text have been deleted from the Abbreviations list. Please check.
- Q4: Please confirm that all sources of funding (including all relevant grant numbers) have been acknowledged in Funding section.
- Q5: Please confirm that this statement is an accurate reflection of any competing interests (or lack thereof) of the author(s).
- Q6: Please provide the main title for Figure 6.
-

Research Article

The N-terminal domain of unknown function (DUF959) in collagen XVIII is intrinsically disordered and highly O-glycosylated

Inderjeet Kaur^{1,2}, Salla Ruskamo², Jarkko Koivunen^{1,2}, Ritva Heljasvaara^{1,2,3,4}, Jarkko J. Lackman⁵, Valerio Izzi^{1,2}, Ulla E. Petäjä-Repo⁵,  Petri Kursula^{2,3} and  Taina Pihlajaniemi^{1,2}

¹Oulu Center for Cell-Matrix Research, University of Oulu, Oulu FIN-90014, Finland; ²Biocenter Oulu and Faculty of Biochemistry and Molecular Medicine, University of Oulu, Oulu FIN-90014, Finland; ³Department of Biomedicine, University of Bergen, Bergen N-5020, Norway; ⁴Centre for Cancer Biomarkers (CCBIO), University of Bergen, Bergen N-5020, Norway; ⁵Medical Research Center Oulu and Research Unit of Biomedicine, Faculty of Medicine, University of Oulu, Oulu FIN-90014, Finland

Correspondence: Taina Pihlajaniemi (taina.pihlajaniemi@oulu.fi)

Collagen XVIII (ColXVIII) is a non-fibrillar collagen and proteoglycan that exists in three isoforms: short, medium and long. The medium and long isoforms contain a unique N-terminal domain of unknown function, DUF959, and our sequence-based secondary structure predictions indicated that DUF959 could be an intrinsically disordered domain. Recombinant DUF959 produced in mammalian cells consisted of ~50% glycans and had a molecular mass of 63 kDa. Circular dichroism spectroscopy confirmed the disordered character of DUF959, and static light scattering indicated a monomeric state for glycosylated DUF959 in solution. Small-angle X-ray scattering showed DUF959 to be a highly extended, flexible molecule with a maximum dimension of ~23 nm. Glycosidase treatment demonstrated considerable amounts of O-glycosylation, and expression of DUF959 in HEK293 *SimpleCells* capable of synthesizing only truncated O-glycans confirmed the presence of *N*-acetylgalactosamine-type O-glycans. The DUF959 sequence is characterized by numerous Ser and Thr residues, and this accounts for the finding that half of the recombinant protein consists of glycans. Thus, the medium and long ColXVIII isoforms contain at their extreme N-terminus a disordered, elongated and highly O-glycosylated mucin-like domain that is not found in other collagens, and we suggest naming it the Mucin-like domain in ColXVIII (MUCL-C18). As intrinsically disordered regions and their post-translational modifications are often involved in protein interactions, our findings may point towards a role of the flexible mucin-like domain of ColXVIII as an interaction hub affecting cell signaling. Moreover, the MUCL-C18 may also serve as a lubricant at cell–extracellular matrix interfaces.

Introduction

Collagens form a diverse family of extracellular matrix (ECM) molecules which, in addition to their characteristic collagenous sequences, have a variety of non-collagenous domains and can be subject to complex post-translational modifications (PTMs), such as hydroxylation, glycosylation and proteolytic processing [1]. Collagen XVIII (ColXVIII) is a non-fibrillar collagen and proteoglycan which belongs to a small subgroup of basement membrane-associated multiplexins (multiple triple-helix domains with interruptions) within the collagen superfamily (reviewed in ref. [2]). Mammalian ColXVIII is produced as three variant $\alpha 1$ (XVIII) collagen chains, which differ from each other in terms of their N-terminal domain structure and tissue distribution [3–7]. The short ColXVIII isoform is the predominant form in epithelial and endothelial basement membranes, while the medium and long forms are the major isoforms expressed by hepatocytes and occur at low levels in many other tissues [8–11].

Received: 4 June 2018
Revised: 28 September 2018
Accepted: 15 October 2018

Accepted Manuscript online:
16 October 2018
Version of Record published:
0 Month 2018

The three ColXVIII isoforms share a C-terminal non-collagenous domain 1 (NC1), including what is known as the endostatin domain [12], an interrupted central collagenous portion which gives the large ColXVIII molecule its flexibility, and an N-terminal NC11 sequence which includes a laminin-G-like/thrombospondin-1 (LAM-G/TSP-1, or shortly TSP-1) motif. The proteolytically released ~20 kDa monomeric endostatin is a potent anti-angiogenic and anti-tumorigenic peptide that has been reported to exhibit additional biological activities, e.g. in promoting apoptosis and autophagy and regulating tissue fibrosis (reviewed in refs [2,13,14]).

A conspicuous feature of the N-termini of the medium and long ColXVIII isoforms is a unique Domain of Unknown Function (DUF959), which in the medium isoform precedes the TSP-1 domain. In the long ColXVIII, a cysteine-rich frizzled-like motif (FZ18) homologous to the extracellular ligand-binding segment of frizzled receptors for Wnt/Wingless signaling molecules is located between the DUF959 and TSP-1 domains [3–7]. At present, the properties and functions of the differing N-terminal NC11 sequences of the three ColXVIII isoforms are only partly understood. Studies with genetically modified mice have shown that the short ColXVIII plays key roles in retinal angiogenesis [15,16], and overexpression of the TSP-1 domain in the mouse eye interferes with the normal functions of the full-length ColXVIII, leading to various defects in the cornea, lens and retina [15]. The FZ18 module of the long ColXVIII interacts with Wnt signaling molecules and antagonizes their functions [17–19]. This inhibition has been shown to lead to reduced tumor cell proliferation and cell cycle arrest in human liver and colon cancer cell lines whose growth is driven by Wnt/ β -catenin signaling [18,19]. In human hepatocellular carcinoma, low amounts of medium/long ColXVIII are associated with large tumors, tumor progression and a high recurrence rate, and the FZ18 domain is negatively associated with Wnt/ β -catenin activity [10,19–21]. The medium and long ColXVIII are crucial for hepatocyte adhesion to liver ECM and for hepatocyte survival during toxin-induced liver injury [22]. In addition, the N-terminal sequences of the medium/long ColXVIII isoforms regulate adipocyte differentiation and the maintenance and function of fat deposits [8].

ColXVIII is a highly glycosylated molecule, which increases the functional complexity of this collagen. All isoforms of vertebrate ColXVIII have three conserved Ser-Gly consensus attachment sites for glycosaminoglycans (GAGs), in the N-terminal NC11 and the central NC8 and NC9 domains [23–25]. In chicks, all three Ser-Gly sites have been shown to carry GAGs, either heparan sulfate (HS) or mixed heparan and chondroitin sulfate (CS) chains, depending on the cell type [26]. In humans and mice, short ColXVIII is mainly a HS proteoglycan [11,27], and the HS chains have been reported to interact with the cell adhesion protein L-selectin and chemokines to regulate renal inflammation [27–30] and with apolipoprotein E, possibly affecting the lipoprotein-trapping function [8].

In view of the unique nature of DUF959, we set out to characterize mammalian ColXVIII DUF959 biochemically and biophysically. We show here that this domain situated at the extreme N-terminus of both the medium and long ColXVIII chains is intrinsically disordered and contains an abundance of mucin-like *N*-acetylgalactosamine (GalNAc) *O*-glycan chains. Understanding these features will help in deciphering the diverse functions of the ColXVIII isoforms.

Methods

Protein expression and purification

A stable HEK293 cell line stably expressing murine DUF959 (ColXVIII's N-terminal amino acids 1–265) fused with a C-terminal His-tag was generated [8]. DUF959 was purified from conditioned cell culture medium (CM) by immobilized metal ion affinity chromatography using the ProBond™ Purification System (Invitrogen) (10 ml of resin for 300–500 ml of CM). The His-tagged protein was eluted with a buffer containing 250 mM imidazole, 150 mM NaCl and 50 mM sodium phosphate, pH 7.2. After imidazole removal and buffer exchange by ultrafiltration (Amicon Ultra Centrifugal Filter, Millipore), DUF959 was further purified by size-exclusion chromatography (SEC) on a HiLoad 16/60 Superdex 200 pg column (GE Healthcare), using 50 mM sodium phosphate, 150 mM NaCl, pH 7.2, as a running buffer. Fractions containing DUF959 were pooled, concentrated and frozen in liquid nitrogen. A typical yield from this purification scheme was 0.8–1.0 mg of DUF959 from 500 ml of CM. The authenticity of the DUF959 was verified by Western blotting using a specific polyclonal antibody generated against residues 34–237 within the DUF959 portion of murine medium/long ColXVIII [8] and a Penta-His antibody (Qiagen), and its purity was assessed by Coomassie-stained SDS-PAGE. Further validation of DUF959 was performed by MALDI-TOF mass spectrometry.

Bioinformatics

Predictions of the sequence-based secondary structure and intrinsically disordered region (IDR) of murine DUF959 were acquired from Jpred4 [31], GlobPlot2 [32], IUPRED [33], DisoPred3 [34] and PONDR-FIT [35]. GlobPlot2 uses a running sum of residue propensities for being in an ordered or disordered state and is the optimal tool for finding domain boundaries in folded proteins. DisoPred3 identifies IDRs based on the protein sequence and recognizes putative ligand-binding motifs within them. PONDR-FIT, on the other hand, combines many different disorder and domain boundary prediction methods. We also used ANCHOR [36] to predict possible intrinsically disordered ligand-binding regions in DUF959. The *NetOGlyc 4.0* (<http://www.cbs.dtu.dk/services/NetOGlyc>) [37] and *NetNGlyc 1.0* (<http://www.cbs.dtu.dk/services/NetNGlyc>) prediction servers were used to assess whether DUF959 has putative GalNAc-type O-glycosylation and N-glycosylation sites, respectively.

Relevant single-nucleotide polymorphisms (SNPs) in DUF959 were inferred from the HaploReg v4.1 database [88] via a two-step procedure: first, the DUF959 region was mapped into the Human Genome assembly GRCh38.p2 using BLAST and used to select the corresponding area on the COL18A1 NM_030582.3 transcript, and only the SNPs within this region were retained. Next, two groups of SNPs, corresponding to two types of amino acid change, were prioritized: the first group included SNPs which changed any residue into Ser or Thr and vice versa, potentially affecting O-glycosylation, while the second included SNPs which changed any residue into Arg and vice versa, potentially affecting proteolytic cleavage [89]. Functional annotation of SNPs, if present, was also reported.

Molecular weight determination by static light scattering

The oligomeric state and absolute molecular weight (MW) of DUF959 in solution were determined using analytical SEC with static light scattering (SLS). Approximately 28 µg of purified DUF959 protein in 50 mM sodium phosphate, 150 mM NaCl, pH 7.2, was loaded into a Superdex 200 10/300 GL gel filtration column connected to an ÄKTA Purifier (GE Healthcare). The system also included a mini-DAWN TREOS multi-angle static light scattering detector (Wyatt Technology) and a Shimadzu refractive index detector for protein concentration and absolute MW determination. The MW of DUF959 was determined based on the measured light scattering and refractive index and/or UV absorbance using ASTRA 5.3 software (Wyatt Technology).

Circular dichroism spectroscopy

Conventional circular dichroism (CD) spectroscopy was carried out using a Chirascan Plus instrument (Applied Photophysics). The cuvette path length was 1 mm. Thermal denaturation was carried out from +22°C to +90°C using a continued temperature ramp scan of 1°C/min for DUF959 protein in 20 mM phosphate buffer, pH 7.2. The wavelength range of 200–260 nm was used for data analysis. T_m values were calculated with GLOBAL3 (Applied Photophysics) by curve fitting the spectral changes as a function of temperature.

Synchrotron radiation CD spectroscopy

Synchrotron radiation CD (SRCD) data were collected on the beamline UV-CD12 at ANKA (Karlsruhe Institute of Technology, Karlsruhe, Germany). The measurements were carried out using 0.75–1.5 mg/ml DUF959 at +20°C with a 200-µm path length quartz cuvette. SRCD spectra were recorded in the 180–280 nm wavelength range at 0.5-nm intervals. All scans were repeated three times and averaged. The buffer contribution was then subtracted. Spectra for DUF959 were also recorded with 25% and 50% trifluoroethanol (TFE) and in the presence of dodecyl phosphocholine micelles.

Small-angle X-ray scattering

Small-angle X-ray scattering (SAXS) data were collected on the SAXS beamline P12 in the PETRA III synchrotron storage ring at EMBL/DESY (Hamburg, Germany). Scattering patterns for DUF959 were measured in 50 mM sodium phosphate, 150 mM NaCl, pH 7.2, with protein concentrations of 2.3, 4.6 and 9.1 mg/ml. No signs of protein aggregation were observed, but concentration-dependent repulsion was visible at higher protein concentrations. To overcome this, the scattering curves with the lowest and highest protein concentrations were merged and the merged curve was used for further analysis and modeling. Data processing and analysis were carried out using the ATSAS software package [38]. The distance distribution function was calculated using GNOM [39] and R_g and D_{max} distributions using the Ensemble Optimization Method 2.0 (EOM) [40,41] and

Guinier analysis. *Ab initio* dummy residue models were built using DAMMIF [42]. Ten independent DAMMIF models were averaged and filtered with DAMAVER [43]. Chain-compatible *ab initio* modeling was performed using GASBOR [44]. MW was estimated by comparing the forward scattering intensity (I_0) of DUF959 with that of glucose isomerase (173 kDa). The dissimilar partial specific volume and higher contrast difference ($\Delta\rho$) of glycan residues relative to protein was taken into account as described previously [45], a partial specific volume of 0.672 being used for DUF959, based on its estimated glycan content and the nature of the glycans [46]. The estimated $\Delta\rho$ of DUF959 (4.79×10^{23} e/cm³) was taken to be the average of the protein (4.47×10^{23} e/cm³) and glycan (5.10×10^{23} e/cm³) values.

Cell culture

The HEK293 SimpleCell line (HEK293-SC) with a knockout of COSMC [37] and the wild-type (WT) HEK293 cell line were maintained in 1× DMEM supplemented with 10% FBS, 100 units/ml penicillin and 0.1 mg/ml streptomycin at 37°C in a humidified atmosphere of 5% CO₂. For transient DUF959 transfections, 9×10^5 cells were seeded on six-well plates and cultured for 24 h before transfection. The Fugene6 transfection reagent (Invitrogen) was used according to the manufacturer's instructions. After 48 h, CM was collected for Western blot analysis, enzymatic deglycosylation and *Vicia villosa* agglutinin (VVA) purification analyses. Cells were harvested and lysed in PBS, pH 7.4, containing a protease inhibitor cocktail (Complete Protease Inhibitor Tablet, EDTA-free, Roche) for Western blotting.

Enzymatic deglycosylation

Purified DUF959 in 50 mM sodium phosphate, 4 mM calcium phosphate, pH 5.5, was used for enzymatic deglycosylation with O-glycosidase and neuraminidase (exo- α -sialidase) (both from Roche). The enzymes were added to a final concentration of 50 mU/ml and samples were incubated at +30°C for 16 h before terminating the reaction by adding 20 μ l of SDS sample buffer.

Heparitinase I and chondroitinase ABC digestions were performed with 20 mU chondroitinase ABC (Sigma) in 15 μ l of 100 mM Tris, 120 mM sodium acetate, pH 8.0, and with 5 mU heparitinase I (Seikagaku) in 15 μ l of PBS, 3 mM CaCl₂, pH 7.0, for 16 h at +30°C. Endoglycosidase H (Merck Millipore, 100 mU/ml) and peptide:N-glycosidase F (PNGase F) (100 U/ml) (Sigma) digestions of DUF959 were performed in 50 mM sodium phosphate, pH 5.5 and 7.5, respectively.

Lectin affinity chromatography

Purification with VVA affinity resin was performed as described recently [47]. Briefly, the CM of transiently transfected HEK293-WT cells and HEK293-SCs was collected after 48 h of DUF959 expression, their protein concentration was measured with the DC protein assay (Bio-Rad), and 500 μ g of protein was added to 20 μ l of agarose-bound VVA (Vector Laboratories) in 50 mM sodium phosphate buffer, 150 mM NaCl, pH 7.2, and 0.1% bovine serum albumin. The samples were incubated for 16 h in a mixer at +4°C, and after extensive washing of the resin, the bound O-GalNAc-containing proteins were eluted three times with 50 μ l of Glycoprotein Elution solution (Vector Laboratories). The buffer was changed to 50 mM sodium phosphate, 150 mM NaCl, pH 7.2, by concentrating the eluates to 20 μ l with centrifugal filter units (30 kDa, Millipore) and diluting to 60 μ l. The samples were analyzed by SDS-PAGE and Western blotting.

Results

Purification of recombinant DUF959

The N-terminal non-collagenous NC11 regions of the medium isoform of murine ColXVIII and the recombinant DUF959 are depicted schematically in Figure 1A. The 265-residue expression construct for murine DUF959 includes the authentic ColXVIII N-terminal signal sequence (residues 1–26), DUF959 (residues 27–265) and a C-terminal hexahistidine tag (Figure 1A). After cleavage of the signal sequence, the theoretical MW of the DUF959 core protein is ~26 kDa. Recombinant DUF959 was produced in stable HEK293 cells and purified from the CM using metal ion and SEC, the DUF959 being eluted from the latter as a single, sharp symmetrical peak (Figure 1B). SDS-PAGE of the pooled peak fractions indicated the presence of a smear-like protein band of size 72–95 kDa recognized by ColXVIII-specific anti-medium/long-18 and penta-His antibodies (Figure 1C).

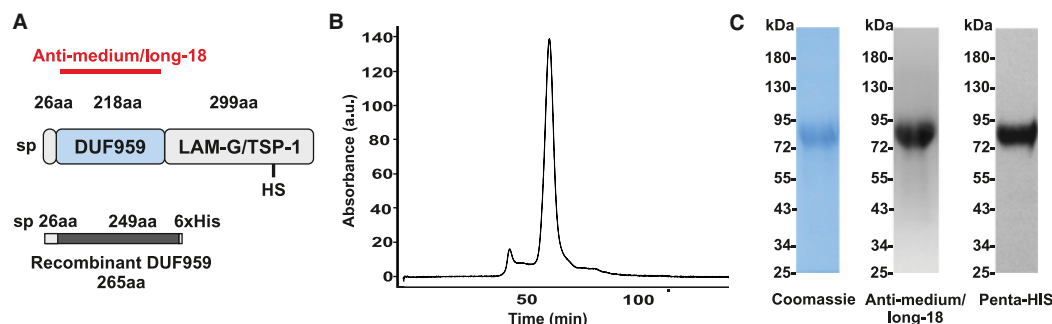


Figure 1. Expression, purification and determination of the oligomeric status of the DUF959 of ColXVIII.

(A) Schematic picture of the domain organization of the N-terminal non-collagenous portion (NC11) of the murine medium ColXVIII isoform. Its features include a signal peptide (sp) of 26 residues, a domain of unknown function (DUF959) of 218 residues and an LAM-G/TSP-1 domain of 299 residues. The 265-residue recombinant fragment, consisting of the signal peptide and the first 249 residues of NC11 and including a C-terminal hexahistidine tag (6xHis), is indicated. The red horizontal line denotes the epitope region for the polyclonal mouse anti-medium/long-18 antibody. HS, previously identified heparan sulfate attachment site [26]. (B) Elution profile of HEK293-secreted recombinant DUF959 from a HiLoad 16/60 Superdex 200 pg column (a.u., arbitrary units). (C) The purity of the recombinant DUF959 was assessed by SDS-PAGE and its authenticity confirmed by Western blotting with anti-medium/long ColXVIII and penta-His antibodies. At least three independent replicates showed a similar size-exclusion elution profile, and Coomassie and Western blot signals.

DUF959 is predicted to be mainly intrinsically disordered

Sequence analyses with JPREP4 suggested that DUF959 has no secondary structure (red dashed line, Figure 2A), while DisoPred3 (disordered residues marked with green x, Figure 2A), PONDR-FIT (Figure 2B) and IUPRED (Figure 2C) predicted one folded region in the middle of the DUF959. Altogether, the secondary structure prediction analyses pointed to the presence of one short (eight residues) and two long (84 and 107 residues) disordered regions in DUF959. Moreover, ANCHOR, which identifies potential ligand-binding regions in IDRs, indicated that DUF959 has multiple putative binding segments (Figure 2C) that may fold upon binding to an interaction partner.

DUF959 is non-globular, highly glycosylated but monomeric in solution

The oligomeric state and glycan content of DUF959 were determined using SEC-SLS. In SEC, DUF959 eluted as a sharp single peak with a relatively small elution volume (Figure 2D), while conjugate analysis of the SLS data revealed that the MW of DUF959 is 63 kDa, of which 49% is protein, a figure corresponding well with the predicted ~26-kDa MW, the rest being composed of glycans. This MW was used for concentration measurements in subsequent assays.

CD spectroscopy indicates a disordered conformation

To characterize the secondary structure content of the DUF959, SRCD spectra were recorded. In this case, the spectrum was typical of disordered proteins, showing a prominent negative minimum at 198 nm and a relatively low ellipticity near 210 nm (Figure 3A). No remarkable changes in the secondary structure content were discovered when the temperature was increased to 90°C (Figure 3B). SRCD spectra were also recorded for DUF959 in the presence of TFE. This solvent lowers the dielectric constant of the solution and is generally used for stabilizing protein secondary structures. In 50% (v/v) TFE, the CD spectrum of DUF959 showed only minor changes towards helical characteristics, and no significant folding was observed (Figure 3C). We did not detect any conformational changes in DUF959 in the presence of dodecyl phosphocholine micelles, which are commonly used to mimic lipid membrane-binding conditions (data not shown).

SAXS confirms the flexible and disordered nature of DUF959

To further characterize the structure of DUF959 in solution, we measured SAXS for the fully glycosylated protein. Based on the relative forward scattering (I_0) values of DUF959 and the reference protein, the MW of DUF959 was estimated to be 70 kDa, which corresponds well to the mass determined by SEC-SLS (Figure 2D).

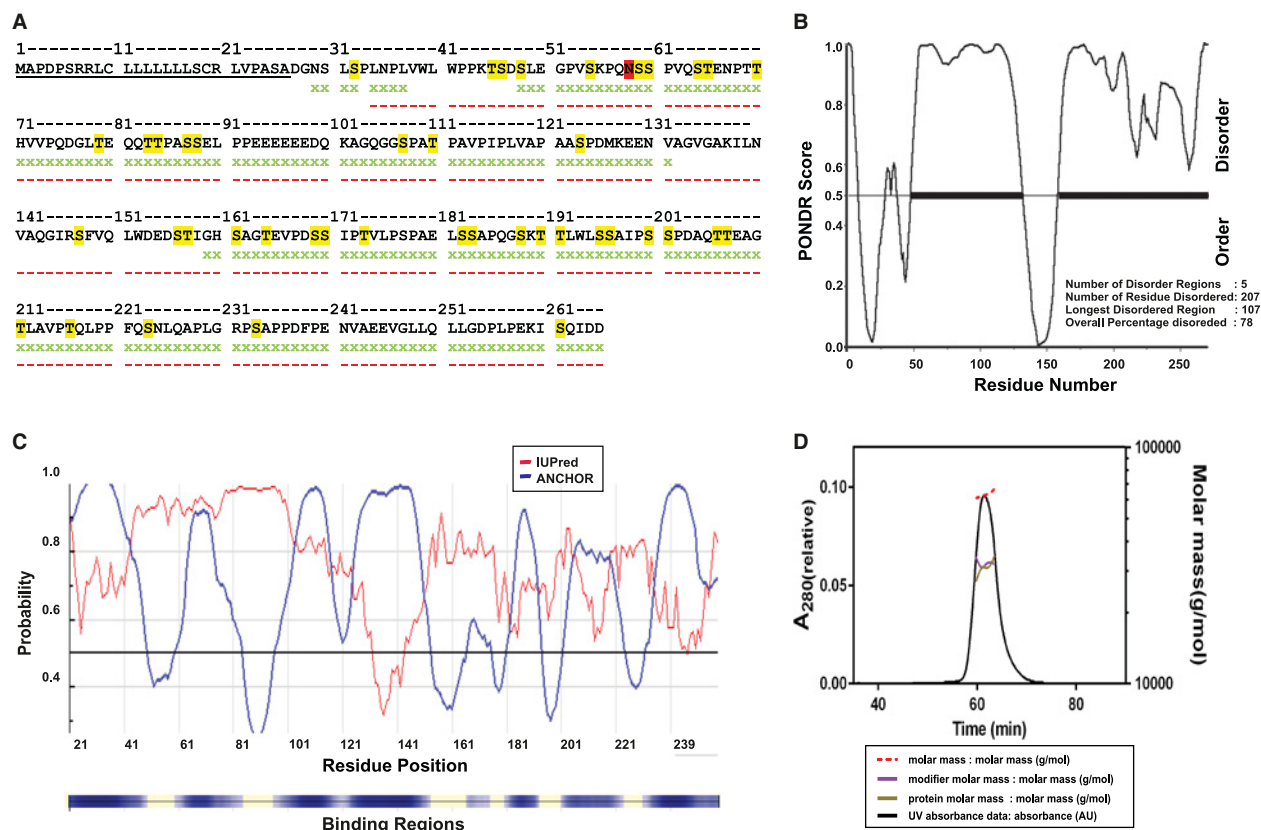


Figure 2. The DUF959 is predicted to be largely disordered with potential ligand-binding regions.

(A) The JPRED4 (red dashed lines indicate no secondary structure) and DisoPred3 (disordered residues marked with green x) predicted secondary structure shown under the amino acid sequence of DUF959. The underlined residues encompass the signal peptide. Potential O-glycosylation sites Ser (S) and Thr (T), as predicted with *NetOGlyc 4.0*, are highlighted in yellow, and one potential N-glycosylation Asn (N) site in red. (B) PONDR-FIT suggests the presence of an ordered segment in the middle of DUF959, while 78% of the domain is predicted to be disordered. (C) IUPRED (red line) predicted that most of DUF959 is disordered (tendency >0.5, threshold shown as the black horizontal line). ANCHOR identified multiple potential ligand-binding regions (shown in blue on the bar below the graph). (D) The MW and glycan content of the SEC peak of DUF959 were determined by SLS. The MW of the whole particle is 63 kDa (red dashed line), of which 31 kDa (green line) is composed of protein and 32 kDa (purple line) of glycans.

The featureless shape of the scattering pattern (Figure 4A), as well as the Kratky plot (Figure 4B) with its lack of any clear maximum, demonstrates the non-globular, flexible and mainly disordered nature of DUF959. The Kratky plot is typically used to identify highly flexible and disordered particles in SAXS [48]. Furthermore, the relatively large radius of gyration (R_g) and maximum intramolecular dimension (D_{max}) of DUF959 (Table 1) illustrate a highly extended shape. Since these are clear indications of the random coil-like behavior of DUF959, we employed the EOM method, optimized for disordered proteins, to study the heterogeneity of DUF959. This showed that there was a pool of slightly more compact DUF959 particles in the solution, and that the R_g and D_{max} distributions of the DUF959 ensemble did not follow a Gaussian distribution (Figure 4C,D). The average values for R_g and D_{max} as calculated using EOM 2.0 are slightly larger than those obtained from a Guinier plot and GNOM (Table 1). It is common for these values to be underestimated when using standard protocols for an IDR. The flexibility analysis incorporated into EOM 2.0, where $R_{flex} = 100\%$ for a fully flexible system and 0% for a fully rigid system, gives the R_{flex} value of 79% for DUF959 and 85% for the random pool, indicating that DUF959 is highly flexible but slightly more rigid than a random pool.

To obtain structural information on DUF959, we built three-dimensional *ab initio* envelopes using both dummy residues and chain-like reconstructions. All the models obtained showed an elongated shape 22-nm in length surrounded by more bulky glycan residues (Figure 4E,F). All the models (Figure 4G,H) fit well with the raw data (χ^2 values 0.86–0.95 for DAMMIF models and 1.30 for GASBOR model) (Table 1).

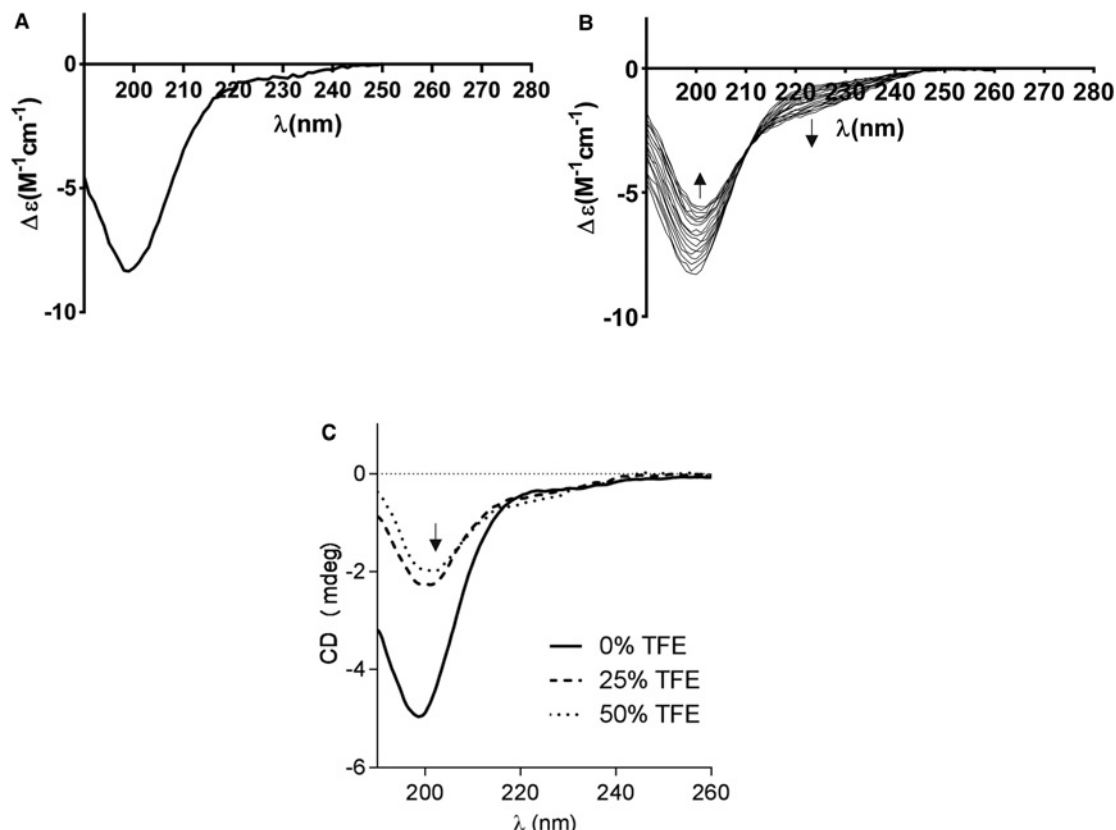


Figure 3. Secondary structure analysis of DUF959 by CD and SRCD spectroscopic analysis.

(A) Conventional CD analysis indicates the disordered/unfolded nature of DUF959. (B) The temperature scan (22–90°C) shows no peak shift in the CD spectra, indicating that protein unfolding does not occur with rising temperature. The changes seen in the spectra were caused by protein aggregation during heating. The direction of spectral change is indicated by the arrows. (C) SRCD spectra for DUF959 recorded in the presence of 0, 25 and 50% TFE. Only minor changes were observed, but TFE did not induce any significant folding of the DUF959. Each SRCD spectrum is an average of three measurements.

DUF959 glycosylation

Mammalian ColXVIII is typically modified by HS GAG chains, while chick ColXVIII also occasionally contains CS side chains [11,26,27]. The multiple Ser residues and acidic amino acids in DUF959 could serve as attachment sites for GAGs [49]. Therefore, heparitinase I and chondroitinase ABC digestions were performed with the recombinant protein purified from HEK293 cells, but there was no visible change in the 72–95 kDa MW of DUF959 after heparitinase I (Figure 5A) or chondroitinase (data not shown) treatment, suggesting that DUF959 does not contain HS or CS side chains. *NetNGlyc 1.0* predicted that DUF959 is not N-glycosylated, but there is one putative N-glycosylation site based on the consensus sequence (Asn-X-Ser/Thr/Cys, where X can be any amino acid except Pro) [50]. PNGase F and endoglycosidase H (which remove either all types, or high mannose and some hybrid N-linked carbohydrates, respectively) digestions did not affect the protein size, confirming the *NetNGlyc 1.0* result and demonstrating that DUF959 is not N-glycosylated (Figure 5B).

The *NetOGlyc 4.0* prediction suggested the multiple Ser and Thr residues of DUF959 as potential O-glycosylation sites (Figure 2A). To test for the presence of O-glycans, *in vitro* enzymatic deglycosylation of recombinant DUF959 protein was performed with O-glycosidase and neuraminidase, which are commonly used to identify proteins modified with GalNAc-type O-glycans. Neuraminidase removes the terminal sialic acids from the glycans and O-glycosidase removes the desialylated Gal β 1–3GalNAc O-glycans (T antigen) from the protein structure. Treatment with neuraminidase alone did not result in any visible change in the DUF959 MW (data not shown), but treatment with both O-glycosidase and neuraminidase resulted in two smeared proteins with MW of ~55–70 and ~32–36 kDa (Figure 5C), so that the difference in MW between the non-treated

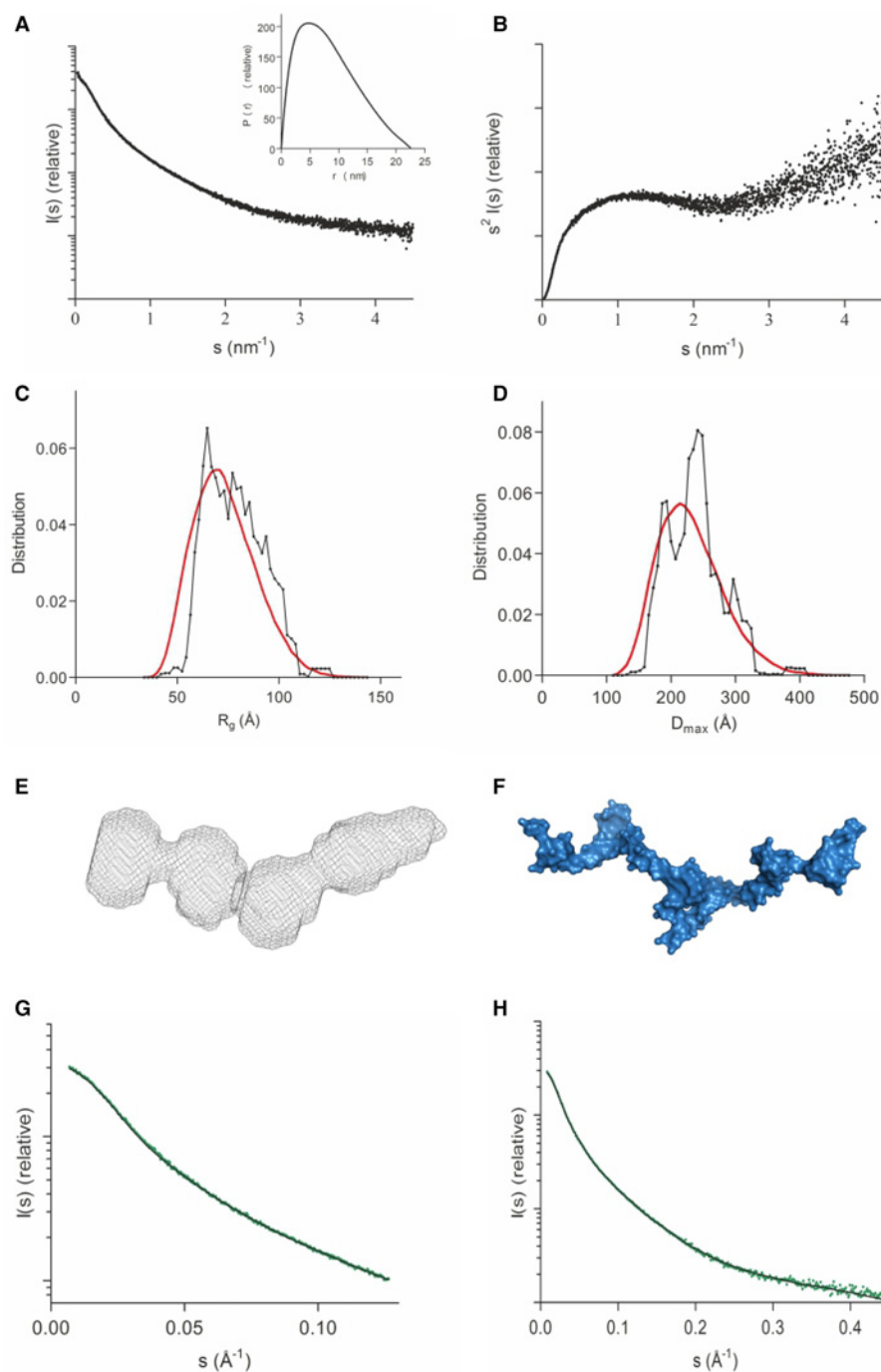


Figure 4. SAXS analysis of DUF959.

(A) Small-angle scattering profile and distance distribution function (inset) of DUF959. The maximum distance of DUF959 in solution is 22 nm. (B) The Kratky plot for DUF959 is typical of flexible particles: no clear maxima exist. (C) Radius of gyration (R_g) and (D) maximum distance (D_{max}) distributions of the EOM ensemble (black) and those of a pool of random polypeptides (red). EOM analysis revealed that a small, more compact DUF959 population also existed in the solution. (E) *Ab initio* model of DUF959, built using DAMMIF. The fit of the model (black) vs. the experimental data (green) is shown in (G). (F) Chain-compatible *ab initio* model of DUF959 generated by GASBOR. The fit of the GASBOR model (black) vs. the experimental data (green) is shown in (H). The SAXS data were collected with three protein concentrations (2.3, 4.6 and 9.1 mg/ml).

Table 1 SAXS parameters

R_g and D_{max} obtained from SAXS data using the Guinier plot, GNOM or EOM χ^2 values of the *ab initio* models of DUF-959 are also been listed.

SAXS data analysis			
R_g (Å) Guinier	R_g (Å) EOM	D_{max} (Å) GNOM	D_{max} (Å) EOM
63	79	226	241
χ^2 values (model vs. data)			
DAMMIF	GASBOR	EOM	
0.855–0.949	1.303	1.345	

and digested proteins must be primarily due to the removal of O-glycans. The presence of two different deglycosylated species could be due to incomplete deglycosylation of the protein by O-glycosidase; the ~55–70 kDa species representing desialylated glycans where only part of the glycans are removed while the ~32–36 kDa species may represent unglycosylated DUF959 which due to its disordered nature migrates slightly slower in the SDS–PAGE than expected for the ~26 kDa DUF959 (Figure 5C). The result nevertheless clearly demonstrates the presence of O-glycans in DUF959.

DUF959 contains GalNAc-type O-linked glycans

The possible presence of GalNAc-type O-glycosylation in the domain of unknown function (DUF) domain of ColXVIII was assessed by means of *SimpleCell* technology, employing HEK293-SCs with a knockout of COSMC [37] followed by VVA affinity chromatography. In HEK293-SCs, O-glycosylation is hindered after the initial addition of a GalNAc residue, due to deletion of the COSMC chaperone. COSMC is required for the activity of the core 1 synthase C1GalT1 that controls further elongation of the O-GalNAc glycan structure [51,52]. O-glycoproteins expressed in the HEK293-SC can be purified with agarose-bound VVA that binds to terminal β - and α -linked GalNAc residues, with a preference for a single α -GalNAc bound to a Ser or Thr residue (Tn antigen).

To confirm that DUF959 is modified by GalNAc-type O-glycans, the DUF959 construct was transfected into HEK293-WT cells and HEK293-SCs. Western blotting of the CM displayed a difference in expressed DUF959 between these two cell lines. More specifically, DUF959 purified from transiently transfected HEK293-WT cells was expressed as a protein with a MW of ~72–95 kDa (Figure 5D), corresponding to the smeared band observed in the stably transfected HEK293 cell line (Figure 5A,B). In contrast, DUF959 purified from HEK293-SCs showed a faster migrating smeared band of ~43–55 kDa (Figure 5D). This is an indication of the presence of truncated O-glycans on the protein when expressed in the HEK293-SCs. There are some differences in the apparent MW of the enzymatically deglycosylated and HEK293-SC-purified unmodified DUF959 (Figure 5C,D), most likely on account of the initial addition of GalNAc residues to the HEK293-SCs, which would have increased the size of the protein. The difference in the apparent MW indicates that many of the potential sites shown in Figure 2A are modified by the ~0.2 kDa GalNAc moieties, resulting in a 43–55 kDa species which could well represent an aberrantly migrating disordered DUF959 decorated with single GalNAc residues (Figure 5D).

To confirm the presence of GalNAc residues, aliquots from the same CM samples were subjected to purification with agarose-bound VVA, and the elutes were analyzed by Western blotting (Figure 5E). The DUF959 expressed in the HEK293-SCs bound to VVA and was detected as a species of MW ~50–55 kDa. As expected, the DUF959 expressed in HEK293-WT cells was unable to bind to VVA and thus could not be detected (Figure 5E). Unexpectedly, a band of ~32 kDa was visible in all VVA elutes, probably representing an endogenous GalNAc-modified protein that binds to VVA column and cross-reacts with the polyclonal anti-medium/long-18 antibody (Figure 5E). Collectively, the results from *SimpleCells* and lectin affinity chromatography indicate that the DUF959 is post-translationally modified by GalNAc-type O-glycans.

Discussion

Of the three N-terminally distinct isoforms of ColXVIII, DUF959 is present in the long and medium variants, while the short variant lacks this domain. The short form results from the use of an upstream promoter of the

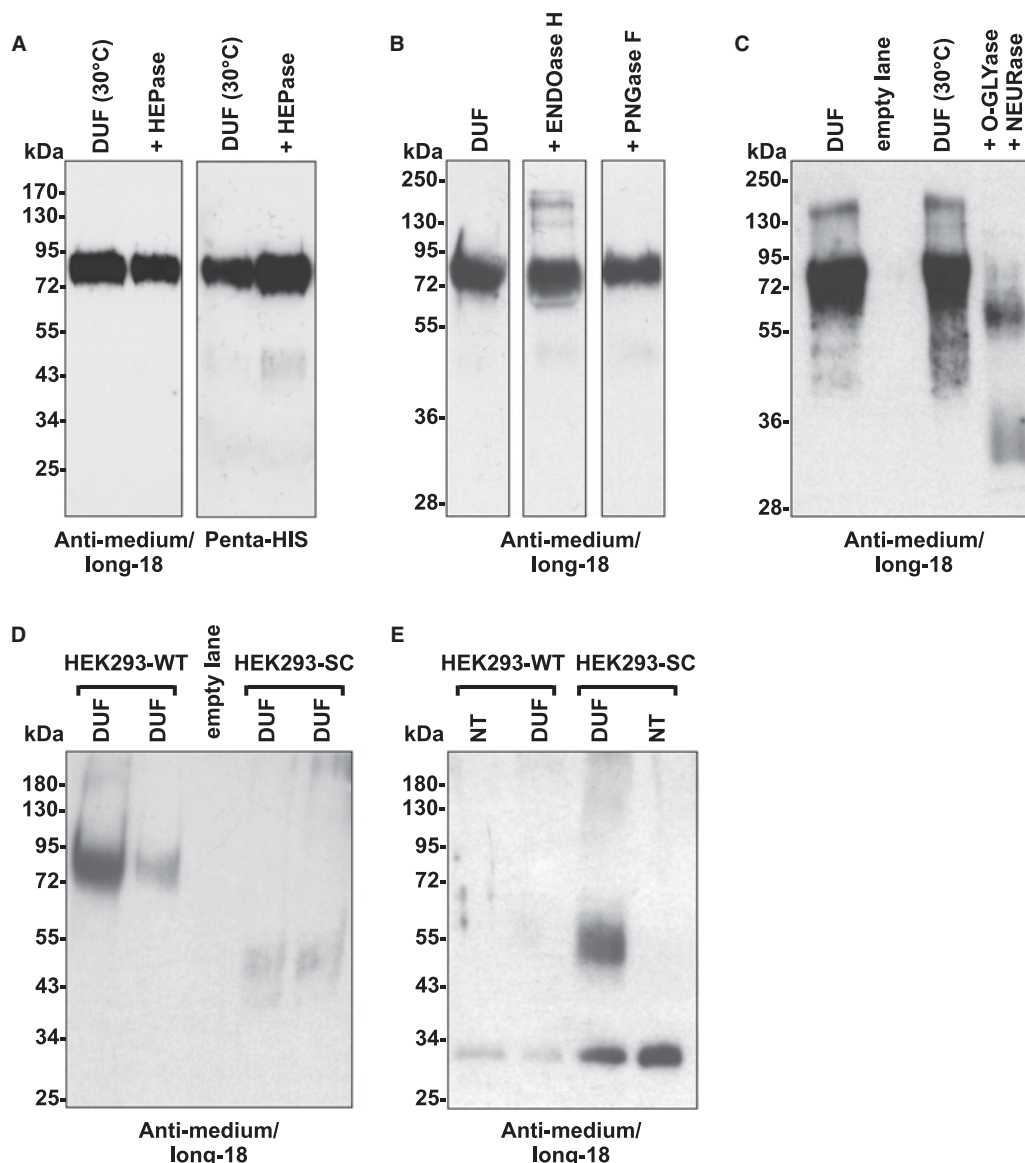


Figure 5. The DUF959 of ColXVIII is rich in mucin-like O-glycosylation.

(A) In a Western blot with anti-medium/long-18 or penta-His antibody, the purified recombinant DUF959 (DUF) from the conditioned CM of the stable HEK293 cells appeared as a smear at ~72–95 kDa. Before loading onto SDS–PAGE, the purified DUF959 samples were incubated at 30°C with or without heparitinase I (HEPase) for 16 h. There was no change in protein size after heparitinase I digestion. (B) Treatment of DUF959 with endoglycosidase H (ENDOase H) or PNGase F at 30°C for 16 h did not affect the size of DUF959. (C) Treatment with a mixture of O-glycosidase (O-GLYase) and neuraminidase (NEURase) at 30°C for 16 h resulted in the appearance of two smaller smeared bands of ~55–70 and ~32–36 kDa band, while untreated DUF959 or incubation of DUF959 at 30°C without enzymes showed the most intense broad signal between 60 and 95. Anti-medium/long-18 antibody was used to visualize DUF959 in Western blotting. (D) Purification of DUF959 from the transiently transfected HEK293-WT CM shows a smeared band between 72 and 95 kDa, while a ~43–55 kDa band is seen in the HEK293-SC CM. Two separate experiments are shown. (E) VVA lectin affinity chromatography of DUF959 expressed either in HEK293-WT cells or in HEK293-SCs. The CM was subjected to VVA chromatography, and the bound and eluted material was analyzed with Western blotting using the anti-medium/long-18 antibody. In contrast with the DUF959 expressed in HEK293-WT cells, DUF959 expressed in HEK293-SCs binds to VVA lectin and has a MW of ~50–55 kDa. CM samples from non-transfected (NT) HEK293-WT cells and HEK293-SCs are shown as controls. An additional unidentified VVA-binding ~32 kDa protein that cross-reacts with anti-medium/long-18 antibody is seen in all elutes. VVA affinity chromatography was performed with CMs collected from two transient DUF959 transfections in HEK293-WT or HEK293-SC cells. Other data shown in this figure are from at least three independent replicates.

ColXVIII gene, whereas the medium and long forms result from the use of a downstream promoter and alternative splicing of the ensuing RNA transcript [3–6,24,53,54]. Recent studies have indicated specific roles for the isoforms, the short form being essential for retinal development and the other two forms affecting liver function and adipose tissue formation [8,15]. These *in vivo* data render it necessary to investigate the properties of the N-terminal non-collagenous domains that distinguish the ColXVIII isoforms. While the FZ18 domain, specific to the longest isoform, and the TSP-1 domain, found in all three isoforms, can be envisioned as sharing properties with other proteins having similar domains [17,18,55–59], the structural and functional properties of DUF959 are not known.

Secondary structure prediction tools identified DUF959 as an IDR interrupted by short ordered regions. We produced DUF959 as a 265-residue His-tagged recombinant protein in mammalian cells and characterized its properties after purification. CD spectroscopy showed that DUF959 indeed has an intrinsically disordered structure. Since TFE did not induce any notable increase in the α -helical content of DUF959, it is probable that DUF959 does not have a strong propensity for folding into an α -helical conformation. This might be due to the large amount of bulky glycan residues contained in it, which hinder the effect of TFE on the polypeptide backbone and physically block the formation of secondary structures. On the other hand, it is rather common for IDRs to form short segments of secondary structure upon binding to ligand proteins or membranes. Once the binding partners for the DUF959 have been identified, higher-resolution studies will be required to observe such structures.

Estimation of the MW of glycoproteins by conventional means is difficult, but SLS can be used as an accurate method for this [45]. SLS shows the MW of DUF959 to be 63 kDa, and approximately one half of this MW is made up of a polypeptide chain and the other half of glycans. Since the sequence-based MW of DUF959 is ~26 kDa, SLS shows a monomeric state for the glycosylated DUF959 in solution. The MW of the molecule in solution can also be estimated based on SAXS, and we derived it using the forward scattering intensity of the sample relative to that of a reference protein. In the case of glycoproteins, differences in $\Delta\rho$ and the partial specific volumes of the reference protein and a glycoprotein need to be considered [45]. Based on SAXS, the MW of fully glycosylated DUF959 is 70 kDa, which is in good agreement with the MW derived from SLS (63 kDa).

SAXS is a widely used method for studying intrinsically disordered proteins, and we used it here to characterize the glycosylated structurally disordered DUF959. The scattering pattern and Kratky plot confirm that DUF959 is a highly extended, flexible molecule, and that its maximum dimension is 226–241 Å (Table 1). Although no *ab initio* modeling programs have yet been optimized for glycoproteins, these can successfully be used [45]. Both modeling methods produced similar models, an elongated, disordered polypeptide chain with bulky volumes attached, presumably representing the linked glycans. The precise positions of the glycan residues cannot be determined by SAXS, however. EOM analysis shows that a fraction of the DUF959 particles adopt a conformation that is more compact than a random polypeptide chain, a form of behavior that indicates that under certain conditions, such as ligand binding, the DUF959 may partly fold.

IDRs in proteins are frequently subjected to various PTMs, including glycosylation [60,61]. Actually, glycosylation is the most common type of PTM in proteins, playing a vital role in cellular differentiation and in pathological disorders [62–64]. This may include changes in the glycosylation pattern of the cell surface and secreted glycoproteins and the facilitation of malignant transformation and cancer progression. Protein O-glycosylation, for example, is implicated in various physiological processes, such as cell adhesion, trafficking, cell signaling and pathogen–host interactions [63,65].

Collagens undergo a wide range of PTMs, including extensive hydroxylation of proline and lysyl residues as well as N- and O-glycosylation. The latter include O-glycosylation of some hydroxylysine residues within the triple-helical collagenous domain by the addition of galactose, which can be further modified by adding of glucose [1,66]. The importance of these glycosylations for collagen stability and function has been demonstrated in several studies [66–69]. Our current results point to a previously unrecognized PTM among the collagens, namely extensive mucin-like O-glycosylation of the DUF959 of ColXVIII.

Sequence analyses suggest that DUF959 can be expected to contain multiple putative sites for Ser/Thr O-linked glycosylation and one potential site for N-glycosylation. We investigated the glycosylation patterns of recombinant DUF959 and found that its digestion with O-glycosidase and neuraminidase resulted in a decrease in the MW of the protein, suggesting the presence of O-glycans (Figure 5C). This was confirmed by expressing DUF959 in HEK293-SCs, which do not support the elongation of O-glycans beyond the initial GalNAc-Ser/Thr linkage [37,52]. As expected, the expressed DUF959 was substantially smaller in size in these cells than in WT HEK293 cells (Figure 5D). VVA has been successfully used to enrich GalNAc-Ser/Thr-modified proteins

from various sources [47,51,70,71], and here DUF959 produced in HEK293-SCs became bound to the GalNAc-Ser/Thr-binding VVA lectin, confirming the presence of GalNAc-type O-glycans (Figure 5E).

The abundance of Pro/Thr/Ser (PTS) residues and the cell-based analyses together imply that DUF959 resembles mucin proteins in terms of its O-glycosylation. Closer examination revealed that DUF959 sequence does not perfectly fulfill the criteria of a mucin-type protein, i.e. the presence of a typical PTS domain containing at least 5% of Pro and over 25% of Thr and Ser residues in a tandem repeat organization [72]. In mucins, this domain is extensively O-glycosylated through GalNAc linkages at the Thr and Ser residues, resulting in a 'bottle brush' glycan configuration around the protein core that typically constitutes over 50% of the glycoprotein mass [73,74]. It should be noted here that lack of tandem repeats is seen in some mucins and mucin-like glycoproteins, e.g. in endomucin/MUC14 and MUC15 [75], and thus DUF959 with its high degree of O-glycosylation could be considered a novel mucin-like glycoprotein.

Previous biophysical analyses of mucin glycopeptides indicate that the flexibility of the mucins is remarkably reduced upon O-glycosylation [76–78]. The stiffness of the fully glycosylated mucin glycopeptides resembles the stability of folded globular protein [77,79]. By CD spectroscopy (Figure 3A), we, however, showed that the conformation of DUF959 is similar to a disordered random coil. Also, Kratky plot (Figure 4B) as well as EOM analysis from SAXS data clearly confirm the flexible nature of DUF959. Despite DUF959 being highly glycosylated, the glycan content is still low enough to retain the flexibility and disordered nature of the polypeptide chain.

In addition to mucin-type or mucin-like protein glycosylation, O-glycans can be distributed to isolated sites of proteins without mucin-like features, referred to site-specific O-glycosylation [80]. Formation of GalNAc-Ser/Thr linkages are controlled by ~20 differentially expressed GalNAc-transferases (GalNAc-T), which have essentially distinct peptide substrate specificities and show specialized functions in the cells [80–82]. Of a particular interest for us is a recent finding that loss of GalNAc-T2 leads to dysregulated lipid metabolism in humans and rodents due to abnormal site-specific O-glycosylation of various targets, including angiotensin-like protein 3 and ApoC-III which both are liver-secreted regulators of triglycerides and high-density lipoprotein cholesterol [83]. In this study, glycoproteomic analysis revealed that Ser107 and Thr110 sites of DUF959 are O-glycosylated by GalNAc-T2 in the mouse liver, but the role of this modification was not addressed further. An important mechanism maintaining lipid homeostasis seems to be the GalNAc-T2 O-glycosylation-regulated substrate proteolysis by proprotein convertases and matrix metalloproteases [84,85]. The DUF959-containing medium/long ColXVIII variants are enzymatically processed with currently unknown proteases [3,19], and it remains to be studied whether the reported site-specific O-glycosylation of DUF959 by GalNAc-T2 [83] is involved in the regulation of ColXVIII processing and, subsequently, lipid metabolism.

We propose on the basis of the present data that DUF959 is a mucin-like domain with a brush-like structure (Figure 6A). The mucin-like domain of the medium and long isoforms of ColXVIII may provide functional properties not present in the other classical basement membrane components, namely collagen IV, laminins, nidogens and perlecan [86]. It should be mentioned here that our *NetOGlyc* and other sequence analyses of collagen XV, structurally related to ColXVIII and the second member of the multiplexin subfamily of collagens, revealed potential mucin features also in this collagen and warrant biochemical and biophysical studies for it as well. We speculate that the mucin-like properties of DUF959 may have a role in cell adhesion and survival in tissues where the medium and long ColXVIII isoforms are expressed, e.g. in the liver [22]. As also reported for mucins [75,87], the O-glycans in DUF959 may affect cells by forming a signaling platform for various growth factors and other signaling molecules. O-glycans may also specifically regulate processing of N-terminus of ColXVIII, thereby affecting the activity of its functional domains as described for FZC18 [3,19], or protect ColXVIII from cell surface proteolysis and thereby shield the cells from attack by other cells [63,73,88,89]. A further possibility is that DUF959 may act as a lubricant at cell–matrix interfaces, because of the potential gel-like properties endowed by high glycosylation.

Altogether 11 260 SNPs, multiple nucleotide polymorphisms and small insertion/deletions (indels) have been identified in the *COL18A1* gene to date, and out of these, 361 SNPs are located in the region that encodes the DUF959 (Figure 6B). It was recently demonstrated that one of the SNPs in the DUF959 region, rs114139997; pGly111Arg, is associated with circulating lipid content and coronary artery disease [90], and that in line with this, genetic polymorphism in the FZ18 region of *COL18A1* is associated with fat deposition and obesity [91]. Moreover, lack of the *Col18a1* medium and long variants in mice results in markedly reduced adiposity and imbalance in circulating triglycerides [8], and lack of all three variants causes hypertriglyceridemia and increases atherosclerosis [92,93], further supporting the role of ColXVIII in regulating lipid metabolism. As mentioned above, site-specific O-glycosylation by GalNAc-T2 is a critical regulator of lipid metabolism, and

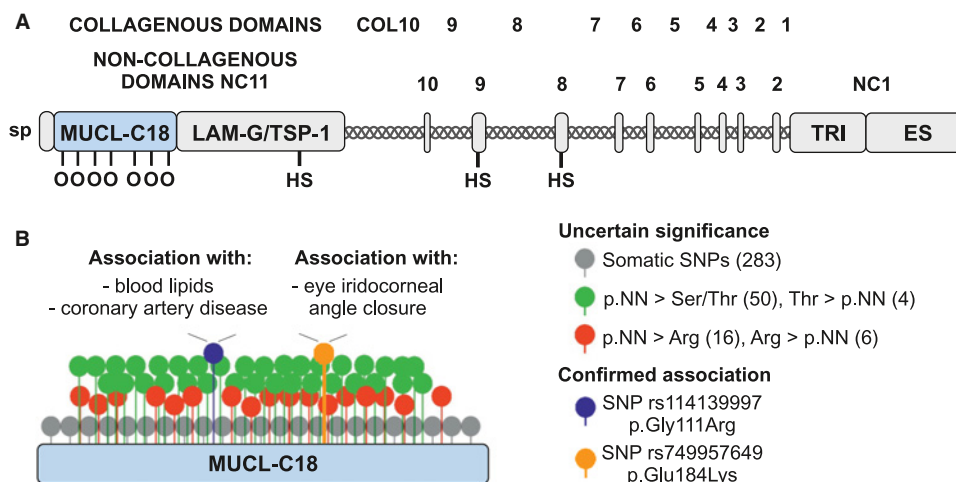


Figure 6. (A) Schematic representation of the full-length medium ColXVIII isoform, including the mucin-like features of DUF959, which we now propose to name MUCL-C18. This ColXVIII isoform consists of a sp, the N-terminal non-collagenous NC11 region, a MUCL18 domain, an LAM-G/TSP-1 homology region, 10 triple-helical collagenous domains (COL1–10), nine short non-collagenous sequences (NC2–NC10) and a C-terminal NC1 with a trimerization (TRI) domain and an endostatin (ES) domain. O-linked glycosylation (O) in the MUCL-C18 domain is indicated. Sites reported to be occupied by HS chains [26] are indicated. (B) SNPs in MUCL-C18 according to the HaploReg v4.1 database [99] and a recent publication by Suri et al. [94]. Gray pins: total number ($n = 283$) of somatic SNPs identified; green: amino acid variation to Ser or Thr ($n = 50$), or from Thr to any other amino acid ($n = 4$) potentially affecting O-glycosylation (note that no variation was observed in Ser); red: amino acid variation to Arg ($n = 16$) or from Arg to any other amino acid ($n = 6$) potentially involved in proteolytic cleavage [100]; blue: an SNP within the MUCL18 with confirmed physiopathological association with lipid metabolism [90]; orange: an SNP within the MUCL-C18 with confirmed pathophysiological association with primary open angle glaucoma [94].

Q6

DUF959 was found among the GalNAc-T2 targets in liver [83], implicating that O-glycosylation of ColXVIII may regulate its ability to control lipid homeostasis.

In addition to the lipid phenotype, DUF959 of ColXVIII may have a role in glaucoma pathogenesis. Recently, a new SNP (rs749957649, p.Glu184Lys) in human DUF959 was identified and associated with iridocorneal primary angle closure glaucoma (PACG) [94]. PACG is an anatomical disorder of the anterior segment of the eye characterized by blocking the aqueous humor drainage as a result of iris apposition the trabecular meshwork and subsequent increased intraocular pressure [95]. These findings are in line with previous reports, showing that specifically medium/long ColXVIII is expressed in ciliary and iris epithelia in mice [15], that mice lacking all three ColXVIII isoforms show ciliary body atrophy, adherence of iris to cornea and higher intraocular pressure in young knockout mice [96], and that some Knobloch patients with *COL18A1* mutations manifest a glaucoma phenotype [97,98]. Finally, even if the few identified SNPs in DUF959 do not directly change O-glycosylation sites, it is possible to envisage a role for *COL18A1* SNPs in changing the patterns of glycosylation within the DUF959, something that might have unexpected biological consequences.

Concluding remarks

We have shown that the DUF959, situated at the extreme N-terminal end of the long and medium variants of ColXVIII, is intrinsically disordered and highly O-glycosylated at its Ser/Thr residues. Such properties have often been found in other proteins to mediate important biological events, for example by recruiting/storing ligands and assembling signaling complexes via their PTMs or regulating proteolysis [60]. These findings can help to illuminate the functional properties of the DUF959 in future studies. Moreover, due to its high degree of GalNAc-type O-glycosylation, we suggest that the DUF959 should be renamed the Mucin-like domain in ColXVIII (MUCL-C18).

Q3 Abbreviations

CD, circular dichroism; CM, culture medium; ColXVIII, collagen XVIII; CS, chondroitin sulfate; DUF, domain of unknown function; ECM, extracellular matrix; EOM, ensemble optimization method; FZ18, frizzled-like motif in

ColXVIII; GAG, glycosaminoglycan; Gal, galactose; GalNAc, *N*-acetylgalactosamine; HEK293-SC, HEK293 SimpleCell line; HS, heparan sulfate; IDR, intrinsically disordered region; LAM-G/TSP-1, laminin-G-like/thrombospondin-1; MUCL-C18, mucin-like domain in ColXVIII; MW, molecular weight; NC, non-collagenous; PACG, primary open angle glaucoma; PNGase F, peptide : *N*-glycosidase F; PTM, post-translational modification; PTS, Pro/Thr/Ser; SAXS, small-angle X-ray scattering; SEC, size-exclusion chromatography; SLS, static light scattering; SNP, single-nucleotide polymorphism; Sp, signal peptide; SRCD, synchrotron radiation circular dichroism; TFE, trifluoroethanol; VVA, *Vicia villosa* agglutinin; WT, wild type.

Author Contribution

I.K. planned the studies, expressed and purified DUF959, performed deglycosylation experiments, CD spectroscopy, bioinformatic analyses and SLS studies, and wrote the manuscript. S.R. performed SAXS, SRCD, and SLS studies, and bioinformatics analyses, and wrote the manuscript. J.L.L. performed the lectin affinity chromatography and wrote parts of the manuscript. V.I. performed the SNP analysis. U.E.P.-R. supervised the O-glycosylation experiments and wrote the manuscript. J.K., R.H., P.K. and T.P. planned and supervised the studies and wrote the manuscript.

Funding

This work was supported by grants from Biocenter Oulu, the Sigrid Jusélius Foundation and the Academy of Finland [grants nos 294617, 275225 and 284605].

Acknowledgements

We thank Jaana Peters, Maija Seppänen and Aila White for their technical assistance. We gratefully acknowledge Prof. Henrik Clausen, University of Copenhagen, for the HEK293-SCs, and beam time and excellent beamline support at the ANKA and EMBL/DESY synchrotron beamlines. The use of the facilities and expertise of the Biocenter Oulu protein crystallography core facility, a member of Biocenter Finland and Instruct-FI, and the Biocenter Oulu core facilities for Proteomics and protein analysis, are gratefully acknowledged.

Competing Interests

The Authors declare that there are no competing interests associated with the manuscript.

References

- Myllyharju, J. and Kivirikko, K.I. (2004) Collagens, modifying enzymes and their mutations in humans, flies and worms. *Trends Genet.* **20**, 33–43 [https://doi.org/S0168-9525\(03\)00319-6](https://doi.org/S0168-9525(03)00319-6)
- Heljasvaara, R., Aikio, M., Ruotsalainen, H. and Pihlajaniemi, T. (2017) Collagen XVIII in tissue homeostasis and dysregulation — lessons learned from model organisms and human patients. *Matrix Biol.* **57–58**, 55–75 [https://doi.org/S0945-053X\(16\)30120-2](https://doi.org/S0945-053X(16)30120-2)
- Elamaa, H., Snellman, A., Rehn, M., Autio-Harmanen, H. and Pihlajaniemi, T. (2003) Characterization of the human type XVIII collagen gene and proteolytic processing and tissue location of the variant containing a frizzled motif. *Matrix Biol.* **22**, 427–442 <https://doi.org/S0945053X03000738>
- Muragaki, Y., Timmons, S., Griffith, C.M., Oh, S.P., Fadel, B., Quertermous, T. et al. (1995) Mouse Col18a1 is expressed in a tissue-specific manner as three alternative variants and is localized in basement membrane zones. *Proc. Natl Acad. Sci. U.S.A.* **92**, 8763–8767 <https://doi.org/10.1073/pnas.92.19.8763>
- Rehn, M. and Pihlajaniemi, T. (1995) Identification of three N-terminal ends of type XVIII collagen chains and tissue-specific differences in the expression of the corresponding transcripts: the longest form contains a novel motif homologous to rat and drosophila frizzled proteins. *J. Biol. Chem.* **270**, 4705–4711 <https://doi.org/10.1074/jbc.270.9.4705>
- Saarela, J., Ylikarppa, R., Rehn, M., Purmonen, S. and Pihlajaniemi, T. (1998) Complete primary structure of two variant forms of human type XVIII collagen and tissue-specific differences in the expression of the corresponding transcripts. *Matrix Biol.* **16**, 319–328 [https://doi.org/S0945-053X\(98\)90003-8](https://doi.org/S0945-053X(98)90003-8)
- Suzuki, O.T., Sertie, A.L., Der Kaloustian, V.M., Kok, F., Carpenter, M., Murray, J. et al. (2002) Molecular analysis of collagen XVIII reveals novel mutations, presence of a third isoform, and possible genetic heterogeneity in knobloch syndrome. *Am. J. Hum. Genet.* **71**, 1320–1329 [https://doi.org/S0002-9297\(07\)60854-4](https://doi.org/S0002-9297(07)60854-4)
- Aikio, M., Elamaa, H., Vicente, D., Izzi, V., Kaur, I., Seppinen, L. et al. (2014) Specific collagen XVIII isoforms promote adipose tissue accrual via mechanisms determining adipocyte number and affect fat deposition. *Proc. Natl Acad. Sci. U.S.A.* **111**, 3043 <https://doi.org/10.1073/pnas.1405879111>
- Kinnunen, A.I., Sormunen, R., Elamaa, H., Seppinen, L., Miller, R.T., Ninomiya, Y. et al. (2011) Lack of collagen XVIII long isoforms affects kidney podocytes, whereas the short form is needed in the proximal tubular basement membrane. *J. Biol. Chem.* **286**, 7755–7764 <https://doi.org/10.1074/jbc.M110.166132>
- Musso, O., Theret, N., Heljasvaara, R., Rehn, M., Turlin, B., Campion, J.P. et al. (2001) Tumor hepatocytes and basement membrane-producing cells specifically express two different forms of the endostatin precursor, collagen XVIII, in human liver cancers. *Hepatology* **33**, 868–876 [https://doi.org/S0270-9139\(01\)11897-5](https://doi.org/S0270-9139(01)11897-5)

- 11 Saarela, J., Rehn, M., Oikarinen, A., Autio-Harminen, H. and Pihlajaniemi, T. (1998) The short and long forms of type XVIII collagen show clear tissue specificities in their expression and location in basement membrane zones in humans. *Am. J. Pathol.* **153**, 611–626 [https://doi.org/S0002-9440\(10\)65603-9](https://doi.org/S0002-9440(10)65603-9)
- 12 O'Reilly, M.S., Boehm, T., Shing, Y., Fukai, N., Vasios, G., Lane, W.S. et al. (1997) Endostatin: an endogenous inhibitor of angiogenesis and tumor growth. *Cell* **88**, 277–285 [https://doi.org/S0092-8674\(00\)81848-6](https://doi.org/S0092-8674(00)81848-6)
- 13 Folkman, J. (2006) Antiangiogenesis in cancer therapy — endostatin and its mechanisms of action. *Exp. Cell Res.* **312**, 594–607 [https://doi.org/S0014-4827\(05\)00545-8](https://doi.org/S0014-4827(05)00545-8)
- 14 Wallia, A., Yang, J.F., Huang, Y.H., Rosenblatt, M.I., Chang, J.H. and Azar, D.T. (2015) Endostatin's emerging roles in angiogenesis, lymphangiogenesis, disease, and clinical applications. *Biochim. Biophys. Acta* **1850**, 2422–2438 <https://doi.org/10.1016/j.bbagen.2015.09.007>
- 15 Aikio, M., Hurskainen, M., Brideau, G., Hagg, P., Sormunen, R., Heljasvaara, R. et al. (2013) Collagen XVIII short isoform is critical for retinal vascularization, and overexpression of the tsp-1 domain affects eye growth and cataract formation. *Invest. Ophthalmol. Vis. Sci.* **54**, 7450–7462 <https://doi.org/10.1167/iov.13-13039>
- 16 Fukai, N., Eklund, L., Marmaros, A.G., Oh, S.P., Keene, D.R., Tamarkin, L. et al. (2002) Lack of collagen XVIII/endostatin results in eye abnormalities. *EMBO J.* **21**, 1535–1544 <https://doi.org/10.1093/emboj/21.7.1535>
- 17 Hendaoui, I., Laverne, E., Lee, H.S., Hong, S.H., Kim, H.Z., Parent, C. et al. (2012) Inhibition of Wnt/beta-catenin signaling by a soluble collagen-derived frizzled domain interacting with Wnt3a and the receptors frizzled 1 and 8. *PLoS ONE* **7**, e30601 <https://doi.org/10.1371/journal.pone.0030601>
- 18 Laverne, E., Hendaoui, I., Coulouarn, C., Ribault, C., Leseur, J., Eliat, P.A. et al. (2011) Blocking Wnt signaling by SFRP-like molecules inhibits in vivo cell proliferation and tumor growth in cells carrying active beta-catenin. *Oncogene* **30**, 423–433 <https://doi.org/10.1038/onc.2010.432>
- 19 Quelard, D., Laverne, E., Hendaoui, I., Elamaa, H., Tirola, U., Heljasvaara, R. et al. (2008) A cryptic frizzled module in cell surface collagen 18 inhibits Wnt/beta-catenin signaling. *PLoS ONE* **3**, e1878 <https://doi.org/10.1371/journal.pone.0001878>
- 20 Musso, O., Rehn, M., Saarela, J., Th  ret, N., Li  tard, J., Hintikka et al. (1998) Collagen XVIII is localized in sinusoids and basement membrane zones and expressed by hepatocytes and activated stellate cells in fibrotic human liver. *Hepatology* **28**, 98–107 <https://doi.org/S0270913998002766>
- 21 Musso, O., Rehn, M., Theret, N., Turlin, B., Bioulac-Sage, P., Lotrian, D. et al. (2001) Tumor progression is associated with a significant decrease in the expression of the endostatin precursor collagen XVIII in human hepatocellular carcinomas. *Cancer Res.* **61**, 45–49 PMID:11196195
- 22 Duncan, M.B., Yang, C., Tanjore, H., Boyle, P.M., Keskin, D., Sugimoto, H. et al. (2013) Type XVIII collagen is essential for survival during acute liver injury in mice. *Dis. Model. Mech.* **6**, 942–951 <https://doi.org/10.1242/dmm.011577>
- 23 Halfter, W., Dong, S., Schurer, B. and Cole, G.J. (1998) Collagen XVIII is a basement membrane heparan sulfate proteoglycan. *J. Biol. Chem.* **273**, 25404–25412 <https://doi.org/10.1074/jbc.273.39.25404>
- 24 Oh, S.P., Kamagata, Y., Muragaki, Y., Timmons, S., Ooshima, A. and Olsen, B.R. (1994) Isolation and sequencing of cDNAs for proteins with multiple domains of gly-xaa-yaa repeats identify a distinct family of collagenous proteins. *Proc. Natl Acad. Sci. U.S.A.* **91**, 4229–4233 <https://doi.org/10.1073/pnas.91.10.4229>
- 25 Rehn, M. and Pihlajaniemi, T. (1994) Alpha 1(XVIII), a collagen chain with frequent interruptions in the collagenous sequence, a distinct tissue distribution, and homology with type XV collagen. *Proc. Natl Acad. Sci. U.S.A.* **91**, 4234–4238 <https://doi.org/10.1073/pnas.91.10.4234>
- 26 Dong, S., Cole, G.J. and Halfter, W. (2003) Expression of collagen XVIII and localization of its glycosaminoglycan attachment sites. *J. Biol. Chem.* **278**, 1700–1707 <https://doi.org/10.1074/jbc.M209276200>
- 27 Zaferani, A., Talsma, D.T., Yazdani, S., Celie, J.W., Aikio, M., Heljasvaara, R. et al. (2014) Basement membrane zone collagens XV and XVIII/proteoglycans mediate leukocyte influx in renal ischemia/reperfusion. *PLoS ONE* **9**, e106732 <https://doi.org/10.1371/journal.pone.0106732>
- 28 Celie, J.W., Keuning, E.D., Beelen, R.H., Drager, A.M., Zweegman, S., Kessler, F.L. et al. (2005) Identification of L-selectin binding heparan sulfates attached to collagen type XVIII. *J. Biol. Chem.* **280**, 26965–26973 <https://doi.org/M502188200>
- 29 Celie, J.W., Rutjes, N.W., Keuning, E.D., Soininen, R., Heljasvaara, R., Pihlajaniemi, T. et al. (2007) Subendothelial heparan sulfate proteoglycans become major L-selectin and monocyte chemoattractant protein-1 ligands upon renal ischemia/reperfusion. *Am. J. Pathol.* **170**, 1865–1878 [https://doi.org/S0002-9440\(10\)61396-X](https://doi.org/S0002-9440(10)61396-X)
- 30 Kawashima, H., Watanabe, N., Hirose, M., Sun, X., Atarashi, K., Kimura, T. et al. (2003) Collagen XVIII, a basement membrane heparan sulfate proteoglycan, interacts with L-selectin and monocyte chemoattractant protein-1. *J. Biol. Chem.* **278**, 13069–13076 <https://doi.org/10.1074/jbc.M212244200>
- 31 Drozdetskiy, A., Cole, C., Procter, J. and Barton, G.J. (2015) JPred4: a protein secondary structure prediction server. *Nucleic Acids Res.* **43**, 389 <https://doi.org/10.1093/nar/gkv332>
- 32 Linding, R., Russell, R.B., Neduva, V. and Gibson, T.J. (2003) Globplot: exploring protein sequences for globularity and disorder. *Nucleic Acids Res.* **31**, 3701–3708 <https://doi.org/10.1093/nar/gkg519>
- 33 Dosztanyi, Z., Csizmek, V., Tompa, P. and Simon, I. (2005) The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins. *J. Mol. Biol.* **347**, 827–839 [https://doi.org/S0022-2836\(05\)00129-4](https://doi.org/S0022-2836(05)00129-4)
- 34 Jones, D.T. and Cozzetto, D. (2015) DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics* **31**, 857–863 <https://doi.org/10.1093/bioinformatics/btu744>
- 35 Xue, B., Dunbrack, R.L., Williams, R.W., Dunker, A.K. and Uversky, V.N. (2010) PONDR-FIT: a meta-predictor of intrinsically disordered amino acids. *Biochim. Biophys. Acta* **1804**, 996–1010 <https://doi.org/10.1016/j.bbapap.2010.01.011>
- 36 Meszaros, B., Simon, I. and Dosztanyi, Z. (2009) Prediction of protein binding regions in disordered proteins. *PLoS Comput. Biol.* **5**, e1000376 <https://doi.org/10.1371/journal.pcbi.1000376>
- 37 Steentoft, C., Vakhrushev, S.Y., Joshi, H.J., Kong, Y., Vester-Christensen, M.B., Schjoldager, K.T. et al. (2013) Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. *EMBO J.* **32**, 1478–1488 <https://doi.org/10.1038/emboj.2013.79>
- 38 Petoukhov, M.V., Franke, D., Shkumatov, A.V., Tria, G., Kikhney, A.G., Gajda, M. et al. (2012) New developments in the ATSAS program package for small-angle scattering data analysis. *J. Appl. Crystallogr.* **45**, 342–350 <https://doi.org/10.1107/S0021889812007662>
- 39 Svergun, D.I. (1992) Determination of the regularization parameter in indirect-transform methods using perceptual criteria. *J. Appl. Crystallogr.* **25**, 495–503 <https://doi.org/10.1107/S0021889892001663>

- 40 Bernado, P., Mylonas, E., Petoukhov, M.V., Blackledge, M. and Svergun, D.I. (2007) Structural characterization of flexible proteins using small-angle X-ray scattering. *J. Am. Chem. Soc.* **129**, 5656–5664 <https://doi.org/10.1021/ja069124n>
- 41 Tria, G., Mertens, H.D., Kachala, M. and Svergun, D.I. (2015) Advanced ensemble modelling of flexible macromolecules using X-ray solution scattering. *IUCrJ* **2**, 207–217 <https://doi.org/10.1107/S205225251500202X>
- 42 Franke, D. and Svergun, D.I. (2009) DAMMIF, a program for rapid ab-initio shape determination in small-angle scattering. *J. Appl. Crystallogr.* **42**, 342–346 <https://doi.org/10.1107/S0021889809000338>
- 43 Volkov, V.V. and Svergun, D.I. (2003) Uniqueness of ab initio shape determination in small-angle scattering. *J. Appl. Crystallogr.* **36**, 860–864 <https://doi.org/10.1107/S0021889803000268>
- 44 Svergun, D.I., Petoukhov, M.V. and Koch, M.H. (2001) Determination of domain structure of proteins from X-ray solution scattering. *Biophys. J.* **80**, 2946–2953 [https://doi.org/S0006-3495\(01\)76260-1](https://doi.org/S0006-3495(01)76260-1)
- 45 Guttman, M., Weinkam, P., Sali, A. and Lee, K.K. (2013) All-atom ensemble modeling to analyze small-angle X-ray scattering of glycosylated proteins. *Structure* **21**, 321–331 <https://doi.org/10.1016/j.str.2013.02.004>
- 46 Lewis, M.S. and Junghans, R.P. (2000) Ultracentrifugal analysis of molecular mass of glycoproteins of unknown or ill-defined carbohydrate composition. *Methods Enzymol.* **321**, 136–149 [https://doi.org/S0076-6879\(00\)21191-9](https://doi.org/S0076-6879(00)21191-9)
- 47 Lackman, J.J., Goth, C.K., Halim, A., Vakhrushev, S.Y., Clausen, H. and Petaja-Repo, U.E. (2018) Site-specific O-glycosylation of N-terminal serine residues by polypeptide GalNAc-transferase 2 modulates human delta-opioid receptor turnover at the plasma membrane. *Cell Signal.* **42**, 184–193 [https://doi.org/S0898-6568\(17\)30285-1](https://doi.org/S0898-6568(17)30285-1)
- 48 Receveur-Brechot, V. and Durand, D. (2012) How random are intrinsically disordered proteins? A small angle scattering perspective. *Curr. Protein Pept. Sci.* **13**, 55–75 <https://doi.org/BSP/CPSP/E-Pub/166>
- 49 Lindahl, U., Couchman, J., Kimata, K. and Esko, J. D. (2015) Proteoglycans and sulfated glycosaminoglycans. In *Essentials of Glycobiology*, 3rd edn (Varki, A., Cummings, R.D., Esko, J.D., Stanley, P., Hart, G.W., Aebi, M. et al., eds), pp. 207–221, Cold Spring Harbor Laboratory Press, La Jolla, CA
- 50 Medzihradsky, K.F. (2008) Characterization of site-specific N-glycosylation. *Methods Mol. Biol.* **446**, 293–316 https://doi.org/10.1007/978-1-60327-084-7_21
- 51 Steentoft, C., Vakhrushev, S.Y., Vester-Christensen, M.B., Schjoldager, K.T., Kong, Y., Bennett, E.P. et al. (2011) Mining the O-glycoproteome using zinc-finger nuclease-glycoengineered SimpleCell lines. *Nat. Methods* **8**, 977–982 <https://doi.org/10.1038/nmeth.1731>
- 52 Wang, Y., Ju, T., Ding, X., Xia, B., Wang, W., Xia, L. et al. (2010) Cosmc is an essential chaperone for correct protein O-glycosylation. *Proc. Natl Acad. Sci. U.S.A.* **107**, 9228–9233 <https://doi.org/10.1073/pnas.0914004107>
- 53 Oh, S.P., Warman, M.L., Seldin, M.F., Cheng, S.D., Knoll, J.H., Timmons, S. et al. (1994) Cloning of cDNA and genomic DNA encoding human type XVIII collagen and localization of the alpha 1 (XVIII) collagen gene to mouse chromosome 10 and human chromosome 21. *Genomics* **19**, 494–499 [https://doi.org/S0888-7543\(84\)71098-6](https://doi.org/S0888-7543(84)71098-6)
- 54 Rehn, M., Hintikka, E. and Pihlajaniemi, T. (1996) Characterization of the mouse gene for the alpha 1 chain of type XVIII collagen (Col18a1) reveals that the three variant N-terminal polypeptide forms are transcribed from two widely separated promoters. *Genomics* **32**, 436–446 [https://doi.org/S0888-7543\(96\)90139-1](https://doi.org/S0888-7543(96)90139-1)
- 55 Adams, J.C. and Lawler, J. (2011) The thrombospondins. *Cold Spring Harb. Perspect. Biol.* **3**, a009712 <https://doi.org/10.1101/cshperspect.a009712>
- 56 Bovolenta, P., Esteve, P., Ruiz, J.M., Cisneros, E. and Lopez-Rios, J. (2008) Beyond Wnt inhibition: new functions of secreted frizzled-related proteins in development and disease. *J. Cell. Sci.* **121**, 737–746 <https://doi.org/10.1242/jcs.026096>
- 57 Cruciati, C.M. and Niehrs, C. (2013) Secreted and transmembrane Wnt inhibitors and activators. *Cold Spring Harb. Perspect. Biol.* **5**, a015081 <https://doi.org/10.1101/cshperspect.a015081>
- 58 Jones, S.E. and Jomary, C. (2002) Secreted frizzled-related proteins: searching for relationships and patterns. *BioEssays* **24**, 811–820 <https://doi.org/10.1002/bies.10136>
- 59 Resovi, A., Pinessi, D., Chiorino, G. and Tarabozetti, G. (2014) Current understanding of the thrombospondin-1 interactome. *Matrix Biol.* **37**, 83–91 <https://doi.org/10.1016/j.matbio.2014.01.012>
- 60 van der Lee, R., Buljan, M., Lang, B., Weatheritt, R.J., Daughdrill, G.W., Dunker, A.K. et al. (2014) Classification of intrinsically disordered regions and proteins. *Chem. Rev.* **114**, 6589–6631 <https://doi.org/10.1021/cr400525m>
- 61 Nishikawa, I., Nakajima, Y., Ito, M., Fukuchi, S., Homma, K. and Nishikawa, K. (2010) Computational prediction of O-linked glycosylation sites that preferentially map on intrinsically disordered regions of extracellular proteins. *Int. J. Mol. Sci.* **11**, 4991–5008 <https://doi.org/10.3390/ijms11124991>
- 62 Ohtsubo, K. and Marth, J.D. (2006) Glycosylation in cellular mechanisms of health and disease. *Cell* **126**, 855–867 [https://doi.org/S0092-8674\(06\)01086-5](https://doi.org/S0092-8674(06)01086-5)
- 63 Oliveira-Ferrer, L., Legler, K. and Milde-Langosch, K. (2017) Role of protein glycosylation in cancer metastasis. *Semin. Cancer Biol.* **44**, 141–152 [https://doi.org/S1044-579X\(17\)30040-8](https://doi.org/S1044-579X(17)30040-8)
- 64 Pinho, S.S. and Reis, C.A. (2015) Glycosylation in cancer: mechanisms and clinical implications. *Nat. Rev. Cancer* **15**, 540–555 <https://doi.org/10.1038/nrc3982>
- 65 Chia, J., Goh, G. and Bard, F. (2016) Short O-GalNAc glycans: regulation and role in tumor development and clinical perspectives. *Biochim. Biophys. Acta* **1860**, 1623–1639 <https://doi.org/10.1016/j.bbagen.2016.03.008>
- 66 Gjaltema, R.A. and Bank, R.A. (2017) Molecular insights into prolyl and lysyl hydroxylation of fibrillar collagens in health and disease. *Crit. Rev. Biochem. Mol. Biol.* **52**, 74–95 <https://doi.org/10.1080/10409238.2016.1269716>
- 67 Jurgensen, H.J., Madsen, D.H., Ingvarsen, S., Melander, M.C., Gardsvoll, H., Pathy, L. et al. (2011) A novel functional role of collagen glycosylation: interaction with the endocytic collagen receptor uparap/ENDO180. *J. Biol. Chem.* **286**, 32736–32748 <https://doi.org/10.1074/jbc.M111.266692>
- 68 Rautavuoma, K., Takaluoma, K., Sormunen, R., Myllyharju, J., Kivirikko, K.I. and Soininen, R. (2004) Premature aggregation of type IV collagen and early lethality in lysyl hydroxylase 3 null mice. *Proc. Natl Acad. Sci. U.S.A.* **101**, 14120–14125 <https://doi.org/10.1073/pnas.0404966101>
- 69 Ruotsalainen, H., Sipila, L., Vapola, M., Sormunen, R., Salo, A.M., Uitto, L. et al. (2006) Glycosylation catalyzed by lysyl hydroxylase 3 is essential for basement membranes. *J. Cell. Sci.* **119**, 625–635 <https://doi.org/119/4/625>
- 70 Steentoft, C., Bennett, E.P. and Clausen, H. (2013) Glycoengineering of human cell lines using zinc finger nuclease gene targeting: SimpleCells with homogeneous GalNAc O-glycosylation allow isolation of the O-glycoproteome by one-step lectin affinity chromatography. *Methods Mol. Biol.* **1022**, 387–402 https://doi.org/10.1007/978-1-62703-465-4_29

- 71 Vakhrushev, S.Y., Steentoft, C., Vester-Christensen, M.B., Bennett, E.P., Clausen, H. and Lavery, S.B. (2013) Enhanced mass spectrometric mapping of the human GalNAc-type O-glycoproteome with SimpleCells. *Mol. Cell Proteomics* **12**, 932–944 <https://doi.org/10.1074/mcp.0112.021972>
- 72 Lang, T., Alexandersson, M., Hansson, G.C. and Samuelsson, T. (2004) Bioinformatic identification of polymerizing and transmembrane mucins in the puffer fish *fugu rubripes*. *Glycobiology* **14**, 521–527 <https://doi.org/10.1093/glycob/cwh066>
- 73 Bergstrom, K.S. and Xia, L. (2013) Mucin-type O-glycans and their roles in intestinal homeostasis. *Glycobiology* **23**, 1026–1037 <https://doi.org/10.1093/glycob/cwt045>
- 74 Dekker, J., Rossen, J.W., Buller, H.A. and Einerhand, A.W. (2002) The MUC family: an obituary. *Trends Biochem. Sci.* **27**, 126–131 [https://doi.org/S0968-0004\(01\)02052-7](https://doi.org/S0968-0004(01)02052-7)
- 75 Dhanisha, S.S., Guruvayoorappan, C., Drishya, S. and Abeesh, P. (2018) Mucins: structural diversity, biosynthesis, its role in pathogenesis and as possible therapeutic targets. *Crit. Rev. Oncol. Hematol.* **122**, 98–122 [https://doi.org/S1040-8428\(17\)30306-2](https://doi.org/S1040-8428(17)30306-2)
- 76 Coltart, D.M., Royyuru, A.K., Williams, L.J., Glunz, P.W., Sames, D., Kuduk, S.D. et al. (2002) Principles of mucin architecture: structural studies on synthetic glycopeptides bearing clustered mono-, di-, tri-, and hexasaccharide glycodomains. *J. Am. Chem. Soc.* **124**, 9833–9844 <https://doi.org/ja020208f>
- 77 Kramer, J.R., Onoa, B., Bustamante, C. and Bertozzi, C.R. (2015) Chemically tunable mucin chimeras assembled on living cells. *Proc. Natl Acad. Sci. U.S.A.* **112**, 12574–12579 <https://doi.org/10.1073/pnas.1516127112>
- 78 Shogren, R., Gerken, T.A. and Jentoft, N. (1989) Role of glycosylation on the conformation and chain dimensions of O-linked glycoproteins: light-scattering studies of ovine submaxillary mucin. *Biochemistry* **28**, 5525–5536 <https://doi.org/10.1021/bi00439a029>
- 79 Live, D.H., Williams, L.J., Kuduk, S.D., Schwarz, J.B., Glunz, P.W., Chen, X.T. et al. (1999) Probing cell-surface architecture through synthesis: an NMR-determined structural motif for tumor-associated mucins. *Proc. Natl Acad. Sci. U.S.A.* **96**, 3489–3493 <https://doi.org/10.1073/pnas.96.7.3489>
- 80 Schjoldager, K.T. and Clausen, H. (2012) Site-specific protein O-glycosylation modulates proprotein processing — deciphering specific functions of the large polypeptide GalNAc-transferase gene family. *Biochim. Biophys. Acta* **1820**, 2079–2094 <https://doi.org/10.1016/j.bbagen.2012.09.014>
- 81 Bard, F. and Chia, J. (2016) Cracking the glycome encoder: signaling, trafficking, and glycosylation. *Trends Cell Biol.* **26**, 379–388 [https://doi.org/S0962-8924\(15\)00251-2](https://doi.org/S0962-8924(15)00251-2)
- 82 Bennett, E.P., Mandel, U., Clausen, H., Gerken, T.A., Fritz, T.A. and Tabak, L.A. (2012) Control of mucin-type O-glycosylation: a classification of the polypeptide GalNAc-transferase gene family. *Glycobiology* **22**, 736–756 <https://doi.org/10.1093/glycob/cwr182>
- 83 Khetarpal, S.A., Schjoldager, K.T., Christoffersen, C., Raghavan, A., Edmondson, A.C., Reutter, H.M. et al. (2016) Loss of function of GALNT2 lowers high-density lipoproteins in humans, nonhuman primates, and rodents. *Cell. Metab.* **24**, 234–245 <https://doi.org/10.1016/j.cmet.2016.07.012>
- 84 King, S.L., Goth, C.K., Eckhard, U., Joshi, H.J., Haue, A.D., Vakhrushev, S.Y. et al. (2018) TAILS N-terminomics and proteomics reveal complex regulation of proteolytic cleavage by O-glycosylation. *J. Biol. Chem.* **293**, 7629–7644 <https://doi.org/10.1074/jbc.RA118.001978>
- 85 Schjoldager, K.T., Vester-Christensen, M.B., Bennett, E.P., Lavery, S.B., Schwientek, T., Yin, W. et al. (2010) O-glycosylation modulates proprotein convertase activation of angiotensin-like protein 3: possible role of polypeptide GalNAc-transferase-2 in regulation of concentrations of plasma lipids. *J. Biol. Chem.* **285**, 36293–36303 <https://doi.org/10.1074/jbc.M110.156950>
- 86 Pozzi, A., Yurchenco, P.D. and Iozzo, R.V. (2017) The nature and biology of basement membranes. *Matrix Biol.* **57–58**, 1–11 [https://doi.org/S0945-053X\(16\)30326-2](https://doi.org/S0945-053X(16)30326-2)
- 87 Singh, P.K. and Hollingsworth, M.A. (2006) Cell surface-associated mucins in signal transduction. *Trends Cell Biol.* **16**, 467–476 [https://doi.org/S0962-8924\(06\)00195-4](https://doi.org/S0962-8924(06)00195-4)
- 88 Corfield, A.P. (2015) Mucins: a biologically relevant glycan barrier in mucosal protection. *Biochim. Biophys. Acta* **1850**, 236–252 <https://doi.org/10.1016/j.bbagen.2014.05.003>
- 89 Hollingsworth, M.A. and Swanson, B.J. (2004) Mucins in cancer: protection and control of the cell surface. *Nat. Rev. Cancer* **4**, 45–60 <https://doi.org/10.1038/nrc1251>
- 90 Peloso, G.M., Auer, P.L., Bis, J.C., Voorman, A., Morrison, A.C., Stitzel, N.O. et al. (2014) Association of low-frequency and rare coding-sequence variants with blood lipids and coronary heart disease in 56,000 whites and blacks. *Am. J. Hum. Genet.* **94**, 223–232 <https://doi.org/10.1016/j.ajhg.2014.01.009>
- 91 Errera, F.I., Canani, L.H., Yeh, E., Kague, E., Armelin-Correa, L.M., Suzuki, O.T. et al. (2008) COL18A1 is highly expressed during human adipocyte differentiation and the SNP c.1136C>T in its 'frizzled' motif is associated with obesity in diabetes type 2 patients. *An. Acad. Bras. Cienc.* **80**, 167–177 <https://doi.org/S0001-37652008000100012>
- 92 Bishop, J.R., Passos-Bueno, M.R., Fong, L., Stanford, K.I., Gonzales, J.C., Yeh, E. et al. (2010) Deletion of the basement membrane heparan sulfate proteoglycan type XVIII collagen causes hypertriglyceridemia in mice and humans. *PLoS ONE* **5**, e13919 <https://doi.org/10.1371/journal.pone.0013919>
- 93 Moulton, K.S., Olsen, B.R., Sonn, S., Fukai, N., Zurakowski, D. and Zeng, X. (2004) Loss of collagen XVIII enhances neovascularization and vascular permeability in atherosclerosis. *Circulation* **110**, 1330–1336 <https://doi.org/10.1161/01.CIR.0000140720.79015.3C>
- 94 Suri, F., Yazdani, S., Chapi, M., Safari, I., Rasooli, P., Daftarian, N. et al. (2018) COL18A1 is a candidate eye iridocorneal angle closure gene in humans. *Hum. Mol. Genet.* **27**, 3772–3786 <https://doi.org/10.1093/hmg/ddy256>
- 95 Jonas, J.B., Aung, T., Bourne, R.R., Bron, A.M., Ritch, R. and Panda-Jonas, S. (2017) Glaucoma. *Lancet* **390**, 2183–2193 [https://doi.org/S0140-6736\(17\)31469-1](https://doi.org/S0140-6736(17)31469-1)
- 96 Ylikarppa, R., Eklund, L., Sormunen, R., Kontiola, A.I., Utriainen, A., Maatta, M. et al. (2003) Lack of type XVIII collagen results in anterior ocular defects. *FASEB J.* **17**, 2257–2259 <https://doi.org/10.1096/fj.02-1001fje>
- 97 Hull, S., Arno, G., Ku, C.A., Ge, Z., Waseem, N., Chandra, A. et al. (2016) Molecular and clinical findings in patients with knobloch syndrome. *JAMA Ophthalmol.* **134**, 753–762 <https://doi.org/10.1001/jamaophthalmol.2016.1073>
- 98 Paisan-Ruiz, C., Scopes, G., Lee, P. and Houlden, H. (2009) Homozygosity mapping through whole genome analysis identifies a COL18A1 mutation in an Indian family presenting with an autosomal recessive neurological disorder. *Am. J. Med. Genet. B Neuropsychiatr. Genet.* **150B**, 993–997 <https://doi.org/10.1002/ajmg.b.30929>
- 99 Ward, L.D. and Kellis, M. (2012) Haploreg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40**, 930 <https://doi.org/10.1093/nar/gkr917>
- 100 Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R.D. and Bairoch, A. (2003) ExPASy: the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* **31**, 3784–3788 <https://doi.org/10.1093/nar/gkg563>