

PROTEASE AND BINDING POLYPEPTIDE FOR O-GLYCOPROTEINS

Field of the Invention

The present invention relates to a novel endoprotease, mutants thereof having
5 binding but lacking or having reduced hydrolyzing activity, and use in methods of studying
and isolating O-linked glycoproteins.

Background of the Invention

Recently, the impact of glycosylation on biological functions has gathered
10 increased attention, in particular in regards to O-linked glycans. However, while the
interest for these important protein modifications have been renewed, the tools to
efficiently study the glycans, and the glycoproteins, have been lacking.

Several exo- and endoglycosidases which are highly useful for both O-linked
glycan removal from native proteins and for glycan sequencing have been developed.
15 Both of these approaches can be used individually to reduce the heterogeneity of
glycoproteins, thus facilitating the analysis of the protein and its fragmented peptides in
mass spectrometry. A more efficient analysis of the biological effect of the glycans by
downstream analysis of the functions affected by the hydrolysis can also be carried out.
However, such tools are not efficient, for example, for facilitating the identification of O-
20 linked glycoproteins, determination of the site of glycosylation and purification of O-
linked glycopeptides.

The first O-glycoprotein-specific endoprotease, binding to O-glycans and mainly
hydrolyzing R-N-bonds close to the glycan, was reported in 1991/1992 (Abdullah *et al.*, J
Bacteriol 173, 5597–5603 (1991); Abdullah *et al.*, Infect Immun 60, 56–62 (1992).
25 However, this enzyme is of limited usefulness for medicine and biotechnology because it is
specific only for O-glycans comprising sialic acids (most but far from all O-linked
glycans) and has specific amino acid demands, resulting in low levels of hydrolysis in
general. There is a need for better tools for studying O-linked glycoproteins.

Summary of the Invention

The present inventors have identified, purified and characterised a novel polypeptide from *Akkermansia muciniphila*, referred to herein as LS. This polypeptide acts as an endoprotease, specifically cleaving/hydrolysing amino acid bonds N terminal to and in proximity of an O-linked glycan, without showing any specificity or limitation to a particular amino acid sequence.

The inventors have also modified the sequence of LS and have identified mutants that are able to bind to O-linked glycans but lack or have a reduced ability to hydrolyze the glycoproteins. These mutants can be used for selective removal, enrichment or purification of free O-glycans, O-glycopeptides and/or O-glycoproteins.

Accordingly, in a first aspect of the invention, there is provided a polypeptide having endoprotease activity specific for O-glycosylated proteins which comprises:

- (a) an amino acid sequence of SEQ ID NO: 1;
- (b) an amino acid sequence which is at least 85% identical to the amino acid sequence of SEQ ID NO: 1 or
- (c) an amino acid sequence which is a fragment of the sequence of SEQ ID NO: 1 or a fragment of an amino acid sequence which is 85% identical to the amino acid sequence of SEQ ID NO: 1.

The invention also provides a method of hydrolysing an O-glycoprotein, wherein the method comprises contacting a sample comprising the protein with a polypeptide of the invention and optionally further comprising the detection or analysis of the hydrolysis products.

Additionally there is provided a method for assessing the glycosylation status of a protein, comprising contacting a sample comprising the protein with a polypeptide of the invention and detecting and/or analysing the products produced, optionally wherein the presence or absence of cleavage products is used to determine the presence or absence of an O-glycoprotein in the sample, and/or wherein said analysis is conducted to identify the type of a O-glycan chain and/or its position of attachment to an O-glycoprotein.

In a second aspect of the invention, there is provided a polypeptide which is capable of binding to an O-glycan, O-glycopeptide and/or O-glycoprotein and which lacks or has reduced endoprotease activity specific for O-glycosylated proteins comprising:

- (a) an amino acid sequence of SEQ ID NO: 5;

(b) an amino acid sequence which is at least 85% identical to the amino acid sequence of SEQ ID NO: 5 or SEQ ID NO: 20; or

(c) an amino acid sequence which is a fragment of the sequence of SEQ ID NO: 5 or SEQ ID NO: 20, or a fragment of an amino acid sequence which is 85%

5 identical to the amino acid sequence of SEQ ID NO: 5 or SEQ ID NO: 20.

The invention also provides a method of binding to an O-glycan, O-glycopeptide and/or O-glycoprotein, wherein the method comprises contacting a sample comprising the O-glycan, O-glycopeptide and/or O-glycoprotein with a polypeptide of the invention, and optionally determining whether or not an O-glycan, O-glycopeptide or O-glycoprotein has
10 been bound and/or separating the O-glycan and any linked glycoprotein, the O-glycopeptide or the O-glycoprotein from the resulting mixture.

Additionally there is provided a method for assessing the glycosylation status of a protein, comprising contacting a sample comprising the protein with a polypeptide of the invention and determining whether or not the protein is bound by the said polypeptide.

15 There is also provided a method for detecting O-glycopeptides and/or O-glycoproteins in a sample, wherein the method comprises:

(a) contacting said sample with a polypeptide of the invention to thereby allow formation of a complex between the polypeptide of the invention and the O-linked glycopeptide and/or O-glycoprotein (an O-linked glycopeptide/protein-polypeptide
20 complex);

(b) optionally separating said polypeptide from the contacted sample; and

(c) determining whether the separated polypeptide is bound to an O-linked glycopeptide or glycoprotein, thereby determining the presence or absence of O-linked glycopeptides or glycoproteins in the sample.

25

Brief Description of the Figures

Figure 1: LS expression and purification. LS was expressed as a fusion protein with a C-terminal His-tag in the pET21(a)+ vector. After transformation into BL21(DE3) Star, four individual clones were expressed and purified on His GravityFlow columns to homogeneity. Based on total quantity of protein in the purified samples, as well as purity
30 based on SDS-PAGE, all four investigated clones expressed equally well.

Figure 2: LS specifically acts upon proteins containing O-glycans – figure shows the products analysed by SDS-PAGE. Incubation of LS with IgG or IgA resulted in a specific degradation of IgA, but no visible activity against IgG (Herceptin/trastuzumab). All incubations took place o/n at 37°C in PBS. Addition of sialidase (Am0707) was not
5 necessary for the activity of LS during these conditions.

Figure 3: Optimal enzymatic conditions. LS is active in a broad pH range (A), tolerates NaCl well (B), but is highly sensitive against EDTA (C, D), and is partly inhibited by Zn²⁺ (D). All experiments (except pH assay) were conducted in PBS o/n, at 37°C. For pH optimum determination, the enzyme was incubated in 20 mM Tris-HCl (pH 6.8-8.8) or
10 50 mM acetic acid (pH 5.6).

Figure 4: Activity of LS is regulated by glycan composition. (A) Sequential removal of specific glycans before hydrolysis with LS for 30 minutes resulted in very low activity in a sialylated protein, high activity in an asialylated protein, and no activity in a sample with removed galactoses. (S) sialidase, (SG) sialidase and galactosidase, (LS) LS.
15 (B) Prolonged incubation (o/n) of fully glycosylated (Enbrel) or sialidase treated (Enbrel(S)) glycoproteins resulted in full hydrolysis in both samples Enbrel may also be referred to herein as etanercept. (C) The TNF α binding part of etanercept (TNF α R) was pretreated with sialidase (“sialidase”), O-glycosidase/sialidase (“O-glyc”), or with PNGaseF (“N-glyc”), to remove sialic acids, O-glycans, and N-glycans, respectively. LS
20 was added to the samples, and incubation was allowed to continue o/n before analysis. LS had activity in all samples except those treated with O-glycosidase.

Figure 5: Search results showing that LS hydrolyses the glycoprotein N-terminal of O-glycans. Etanercept hydrolyzed to fragments with LS, and subsequently deglycosylated with O-glycosidase treatment were subjected to mass spec analysis (LC/MS
25 and MS/MS). Identified peptides (white and hatched boxes) were fitted to the sequence of Etanercept based on m/z values and MS/MS data, with y' and b' ions marked as small grey boxes. All gwhite and hatched boxes (e.g. peptides) start directly at a T or an S, where O-glycans have been attached. The preceding amino acid varies (P, S, H, T, G), and seems likely to not influence the hydrolysis. (A) Analysis using a biased approach, specifically
30 searching for peptides generated with an S/T-peptidase. (B) Analysis using an unbiased approach.

Figure 6: Inactivated LS specifically binds to O-linked glycoproteins. The metalloprotease active site was mutated to remove the catalytic capacity while not affecting substrate affinity or interaction. Specifically this was done by exchanging an E to an A, thus creating clone “LS_{mut}” (also referred to as LS_{E206A}). (A) While LS was able to hydrolyze Enbrel in the presence of sialidase, the inactivated Lsmut could not hydrolyze Enbrel under the conditions tested. Loss of activity was verified on SDS-PAGE. (B) Despite having lost hydrolytic activity, LS_{mut} was still able to bind to O-glycoprotein. Specific binding was verified on spin columns with immobilized LS_{E206A}, demonstrating a specific affinity for O-linked glycoproteins. By immobilizing LS_{mut} on sepharose we were able to affinity purify IgA. Herceptin (Trastuzumab), lacking O-glycans, as well as O-glycosidase treated IgA, did not bind to the column, but could be detected in the flowthrough (FT). Neur = Neuraminidase/Sialidase 0707.

Figure 7: α 2-3 bonded sialic acids limit the efficiency of LS. Concurrent incubation of LS with a set of diverse sialidases for 30 min – 20 h, using Enbrel as a glycoprotein substrate, revealed the higher efficiency in the presence of the α 2-3 specific sialdiase 1757, or with the Mix (0707 + 1757), while the broad spectrum sialidase 0707 was not necessary for seemingly full activity of LS, thus suggesting that α 2-6 (and α 2-8) bonds are not a concern for LS activity.

Figure 8: Schematic presentation of LS activity. LS preferentially binds terminal galactoses attached to O-linked GalNAc, resulting in an N-terminal hydrolysis of the serine or threonine upon which the glycan is attached. Presence of sialic acids will reduce the efficiency of LS, but not inhibit it. No activity can be seen upon N-linked glycans.

Figure 9: Results of an experiment in which erythropoietin is cleaved with different combinations of LS, PNGaseF to remove N-glycans, Sialidases to remove sialic acid, and with an O-glycosidase to remove O-glycans. Reaction products were analysed by SDS-PAGE, RPLC and ESI mass spectrometry (A) Results of SDS-PAGE analysis: Lane 1 = EPO treated with PNGaseF and sialidase; Lane 2 = EPO treated with PNGaseF and sialidase + LS; Lane 3 = EPO treated with PNGaseF + LS; Lane 4 = EPO treated with PNGaseF, sialidase, O-glycosidase before LS; Lane 5 = enzyme control. Band X = uncleaved EPO; Band Y = N terminal fragment of EPO digested by LS; Band Z = C terminal fragment of EPO digested by LS. Lanes 2 and 3 show that LS cleaves EPO where the sialic acids have been removed as well as where they are intact. Lanes 2 – 3 show that

LS also cleaves EPO where N-glycans have been removed with PNGaseF. Lane 4 shows that LS does not cleave EPO where the O-glycans have been removed. (B) UV chromatogram shows results of RPLC separation for EPO treated with PNGaseF and sialidase + LS. Two major peaks were identified as shown. Peak 1 is the C terminal
 5 fragment of EPO digested by LS; Peak 2 is the N terminal fragment of EPO digested by LS; (C,D) show results of mass spectrometry analysis. Fig C shows masses of the C terminal fragment of EPO with O-glycan still attached to the (now N terminal) serine (Square = GlcNAc, Circle = Galactose). Differences in mass are due to differences in degradation of the O-glycan (loss of a terminal Galactose) in some parts of the sample,
 10 likely caused by the ionisation energy in the MS instrument; Fig D shows the N terminal fragment of EPO lacking glycan, plus undigested EPO with glycan still attached.

Figure 10: Results of experiments which show that LS_{E206A} retains some activity whereas LS_{H205A/E206A} (may also be referred to as LS_{HE206AA}) is completely inactive. The activity of the LS mutants LS_{E206A} (A) and LS_{H205A/E206A} (B) was evaluated against
 15 asialylated O-glycosylated substrates including the TNF α binding part of etanercept (TNFaR2; may also be referred to herein as TNFaR) and etanercept itself (Etanercept), as compared to the wild type LS enzyme. Different concentrations of the LS mutants were added to 1 μ g substrate (1:1 – 15:1, enzyme:substrate), incubating in PBS at 37°C overnight before analysis on SDS-PAGE.

20 A) Lane 1: Asialylated substrate only; Lane 2: LS only, lane 3: 0.5 μ g LS_{E206A}, lane 4: 5 μ g LS_{E206A}, lane 5: TNFaR2 + LS (1:1 ratio), lane 6: TNFaR2 + LS_{E206A} (1:1 ratio).
 B) Lane 1: Asialylated substrate only; Lane 2: LS_{H205A/E206A} + Etanercept (15:1 ratio), lane 3: LS_{H205A/E206A} + Etanercept (5:1 ratio), lane 4: LS_{H205A/E206A} + Etanercept (1:1 ratio), lane 5: LS + Etanercept (1:1 ratio), lane 6: LS_{H205A/E206A}

25 **Figure 11:** Results of experiments which show that LS_{H205A/E206A} immobilised on resin specifically binds to O-glycan containing proteins. Figures show SDS-PAGE analyses for starting/loading material, flowthrough (FT) and Eluate (E) in each case (A) Samples included BSA (bovine serum albumin), Etanercept, IgA, and IgG, native or pretreated with Sialidase mixtures +/- O-glycosidase as shown. (B) Sample included a
 30 mix of O-glycosylated proteins (TNFaR and ApoE), N-glycosylated proteins (aflibercept, AGP (alpha-1-acid glycoprotein), IgG Fc (Fc domain of IgG) and non-glycosylated

proteins (BSA), pre-treated with Sialidase mixture. (C) Sample included a mix of N-glycosylated proteins (cetuximab, aflibercept, AGP) and non-glycosylated proteins (BSA, carbonic anhydrase), pre-treated with Sialidase mixture.

Figure 12: Results of experiment showing that immobilised LS_{H205A/E206A} has a concentration dependent capacity for binding of O-glycoproteins. Asialylated Etanercept (50-250 µg; in 100 µl PBS) was added to 50 µl PBS-equilibrated LS_{H205A/E206A} resin with different immobilization conditions of LS_{H205A/E206A} (5-15 mg/mL). Proteins were allowed to bind to the resin for 2 hours at room temperature with end-over-end rotation. The resin was washed three times with PBS (350 µl) and then eluted with the addition of 8 M Urea (50 µl, 5 min incubation; 2 repeats). To study the effect of urea and guanidine hydrochloride (GHC) in the binding, it was included in the binding buffer together with 50 µg asialylated Etanercept, but otherwise handled identically. (A) All samples were separated on SDS-PAGE, and band intensities determined through densitometry using GelDoc EZ and the software ImageLab. Stated percentages indicate the band intensity as compared to the control. (B) Protein binding capacity, as determined by the band intensity, was plotted versus the amount of immobilized LS_{H205A/E206A}.

Figure 13: Results of experiment showing that LS_{H205A/E206A} can affinity purify around 3 mg of etanercept /mL of resin. Asialylated etanercept (10-200 µg; 100 µl in PBS) was added to 50 µl PBS-equilibrated LS_{H205A/E206A} resin. Proteins were allowed to bind to the resin for 2 hours at room temperature with end-over-end rotation. The resin was washed three times with PBS (350 µl) and then eluted with the addition of 8 M Urea (50 µl, 5 min incubation; 2 repeats). All samples were separated on SDS-PAGE, and band intensities determined through densitometry using GelDoc EZ and the software ImageLab. Stated percentages indicate the band intensity as compared to the control.

Figure 14: Results of experiment showing that LS_{H205A/E206A}-substrate interaction is insensitive to high ionic strengths and differences in buffer volume/type, and works over a broad pH range. (A) Asialylated etanercept, 50 µg; 100 µl in PBS with added NaCl at concentrations as shown (B) Asialylated etanercept, 50 µg; 100-300 µl in PBS as shown; (C) Flow through from samples of asialylated etanercept (50 µg) and BSA (50 µg) in different buffers at different pHs as shown; (D) Eluates from the samples of C.

Figure 15: Results of experiment showing that denaturation or addition of detergents elutes O-glycoproteins bound to LS_{H205A/E206A}. Asialylated Etanercept (50 µg; 100 µl in PBS) was added to 50 µl PBS-equilibrated LS_{H205A/E206A} resin. Proteins were allowed to bind to the resin for 2 hours at room temperature with end-over-end rotation.

5 The resin was washed three times with PBS (350 µl) and then eluted with the addition of (A) 1-8 M Urea or (B) 1.25-10% SDS (50 µl, 5 min incubation; 2 repeats). All samples were separated on SDS-PAGE for analysis.

Figure 16A: Results of experiment showing enzymatic elution of LS_{H205A/E206A}-bound O-glycoproteins using LS. Asialylated Abatacept (10 µg, 100 µl in PBS) and

10 Etanercept (50 µg; 100 µl in PBS) were added to 50 µl PBS-equilibrated LS_{H205A/E206A} resin. Proteins were allowed to bind to the resin for 2 hours at room temperature with end-over-end rotation. The resin was washed three times with binding buffers (350 µl) before addition of 50 Units LS in a total volume of 100 µl PBS. The samples were incubated for an additional 6-24 hours at 37°C with shaking (450 rpm). LS released O-

15 glycoproteins/glycopeptides were collected through centrifugation (1000 g, 1 min) before the column was finally eluted with the addition of 8 M Urea (50 µl, 5 min incubation; 2 repeats). All samples were separated on SDS-PAGE for analysis.

Figure 16B: Results of mass spec analysis (LC/MS and MS/MS) of Etanercept eluted with LS. Identified peptides (Fig 16B.1) were consistent with those generated in a LS digestion of etanercept (Fig 16B.2). Identified peptides (white boxes) were fitted to the

20 sequence of Etanercept based on m/z values and MS/MS data, with y' and b' ions marked as small grey boxes. All white boxes (e.g. peptides) start directly at a T or an S, where O-glycans are attached.

Figure 17: Results demonstrating the affinity purification and enrichment of O-glycosylated serum proteins. (A) Asialylated serum (20 µl; 100 µl in PBS) was added to 50 µl PBS equilibrated LS_{H205A/E206A} resin. Proteins were allowed to bind to the resin for 2

25 hours at room temperature with end-over-end rotation. The resin was washed three times with binding buffer (350 µl) and then eluted with the addition of 8 M Urea. (B) To investigate the impact of glycans for the interaction, samples were pre-treated with a sialidase mixture +/- O-glycosidase. The downstream purification was performed as

30 described above. (C) Serum (40 µl) was mixed with PBS (up to 100 µl) and sialidase

mixture (50-500 Units) and added to a PBS equilibrated column, incubating for 2 hours at room temperature with end-over-end rotation, after which the samples were washed and eluted as described above. All samples were separated on SDS-PAGE for analysis.

Figure 18: Results showing enrichment of O-glycoproteins from human serum.

- 5 Human serum diluted 2.5x in PBS to 100 µl was applied on 50 µl PBS equilibrated LS_{H205A/E206A} resin in a spin column. 50-500 Units Sialidase mixture, was added and co-incubated on the resin at RT for 2h. Flow through was collected, and the resin was washed 5-10 times with PBS. The bound proteins were eluted in 8 M urea followed by denaturation and reduction with the addition of 5 mM DTT and a 60-minute incubation at
- 10 37°C. The reduced cysteines were alkylated with 15 mM iodoacetamide at RT in dark for 30 minutes. Samples were buffer exchanged to 50 mM Tris pH 8.0 on a Spin Desalting Column. Trypsin (2.5 µg) was added to the solution and digestion was overnight at 37°C. Peptides were separated and analysed using RP-LC MS/MS on a C18 column in a 0.1%FA in MQ : 0.1%FA in 95% ACN gradient at 45°C and a flow of 0.2 ml/min. Detection was
- 15 on an ESI-Q-TOF. Data were converted to mgf format files and searched against the Swiss Prot database (A) Identified peptides coming from proteins annotated as O-glycosylated proteins or non-O-glycosylated. Only proteins with >6 matching peptides and a MASCOT score >200 were included. (B) Different washing steps resulted in changes in identified peptides, as well as changed ratio of O-glycosylated to non-O-glycosylated proteins (C).
- 20 Sia = Sialidase treated; Sia Pre = Sialidase Pre-treated.

- Figure 19:** Results of experiments that show that the immobilized double-mutant also binds to shorter O-glycopeptides. **A** shows representative results for LC/MS analysis of binding to a prepared mix of an O-glycosylated peptide (glycodrosocin (GD)) and several non-glycosylated peptides (H2686, H4062 H8390 and insulin oxidized beta chain (IOB)). **B** shows a schematic diagram of IgA illustrating that tryptic digest will product a single O-glycosylated peptide. **C** shows representative results for LC/MS analysis of binding to the tryptic digest of IgA.
- 25

- Figure 20:** Results of experiments that show that the immobilized double-mutant compares favorably to other commercially available O-glycoprotein binding matrixes. **A**
- 30 shows representative SDS-PAGE gels comparing the presence of etanercept or asialylated etanercept (etanercept(S)) in flow through (FT) or Eluate (E) following incubation with

different immobilized lectins or LS double mutant as shown. **B** shows densitometry analysis of the gels, relative to positive control of 1.5 µg directly-loaded substrate.

Brief Description of the Sequences

5 SEQ ID NO: 1 is the amino acid sequence of a polypeptide having O-glycoprotein-specific endoprotease activity.

 SEQ ID NO: 2 is an amino acid sequence of an exemplary polypeptide of the invention having O-glycoprotein-specific endoprotease activity. Relative to SEQ ID NO: 1 it includes an additional N terminal Methionine and a C-terminal linker + His₆ tag. The
10 polypeptide consisting of this sequence may be referred to herein as LS.

 SEQ ID NO: 3 is a nucleotide sequence encoding a polypeptide having the sequence of SEQ ID NO: 2.

 SEQ ID NO: 4 is the wild-type amino acid sequence of a polypeptide isolated from *Akkermansia muciniphila* having O-glycoprotein-specific endoprotease activity. Relative to
15 SEQ ID NO: 1 it includes a signal motif at the N terminus.

 SEQ ID NO: 5 is the amino acid sequence of a polypeptide that is capable of binding to O-glycans but lacks or has reduced O-glycoprotein-specific endoprotease activity.

 SEQ ID NO: 6 is an amino acid sequence of an exemplary polypeptide of the invention that is capable of binding to O-glycans but lacks or has reduced O-glycoprotein-specific endoprotease activity. Relative to SEQ ID NO: 5 it includes an additional N terminal
20 Methionine and a C-terminal linker + His₆ tag. The polypeptide consisting of this sequence may be referred to herein as LSE_{206A}.

 SEQ ID NO: 7 is a nucleotide sequence encoding a polypeptide having the sequence of SEQ ID NO: 6.

25 SEQ ID NO: 8 is the metalloprotease domain motif of an exemplary polypeptide of the invention that has O-glycoprotein-specific endoprotease activity.

 SEQ ID NO: 9 is the wild-type amino acid sequence of a sialidase, Am1757, isolated from *Akkermansia muciniphila*. It includes a signal motif at the N terminus.

 SEQ ID NO: 10 is the wild-type amino acid sequence of a sialidase, Am1757, lacking
30 a signal motif at the N terminus relative to SEQ ID NO: 9.

SEQ ID NO: 11 is an amino acid sequence of an exemplary sialidase, Am1757. Relative to SEQ ID NO: 10 it includes an additional N terminal Methionine and a C-terminal linker + His₆ tag.

5 SEQ ID NO: 12 is wild-type amino acid sequence of a sialidase, Am0707, isolated from *Akkermansia muciniphila*. It includes a signal motif at the N terminus.

SEQ ID NO: 13 is the wild-type amino acid sequence of a sialidase, Am0707, lacking a signal motif at the N terminus relative to SEQ ID NO: 12.

10 SEQ ID NO: 14 is an amino acid sequence of an exemplary sialidase, Am0707. Relative to SEQ ID NO: 13 it includes an additional N terminal Methionine and a C-terminal linker + His₆ tag.

SEQ ID NO: 15 is the amino acid sequence of an O-glycosidase isolated from *S. oralis*.

SEQ ID NOs: 16 and 17 are primer sequences.

SEQ ID NO: 18 shows the amino acid sequence of EPO.

15 SEQ ID NO: 20 is the amino acid sequence of a polypeptide that is capable of binding to O-glycans but lacks or has reduced O-glycoprotein-specific endoprotease activity.

20 SEQ ID NO: 21 is an amino acid sequence of an exemplary polypeptide of the invention that is capable of binding to O-glycans but lacks or has reduced O-glycoprotein-specific endoprotease activity. Relative to SEQ ID NO: 20 it includes an additional N terminal Methionine and a C-terminal linker + His₆ tag. The polypeptide consisting of this sequence may be referred to herein as LS_{HE206AA} or LS_{H205A/E206A}.

SEQ ID NO: 22 is a nucleotide sequence encoding a polypeptide having the sequence of SEQ ID NO: 21.

25 SEQ ID NOs: 23, 24 and 25 are sequences of disrupted metalloprotease domain motifs, each from a polypeptide of the invention that is capable of binding to O-glycans but lacks or has reduced O-glycoprotein-specific endoprotease activity.

SEQ ID NOs: 26, 27 and 28 are the amino acid sequences of polypeptides each having O-glycoprotein-specific endoprotease activity.

30 SEQ ID NO: 29, 30 and 31 are the amino acid sequences of exemplary polypeptides having O-glycoprotein-specific endoprotease activity. Relative to SEQ ID NOs: 26, 27 and 28, respectively, SEQ ID NOs: 29, 30 and 31 each include an additional N terminal Methionine and a C-terminal linker + His₆ tag.

SEQ ID NOs: 32, 33 and 34 are wild-type amino acid sequences of polypeptides having O-glycoprotein-specific endoprotease activity, which were isolated from *Pseudomonas aeruginosa* PAO1, *Bacteroides thetaiotaomicron* VPI-5482, and *Clostridium perfringens*, respectively. Relative to SEQ ID NOs: 26, 27 and 28, respectively, each includes a signal motif at the N terminus.

SEQ ID NOs: 35, 36 and 37 are the amino acid sequences of polypeptides that are each capable of binding to O-glycans but lack or have reduced O-glycoprotein-specific endoprotease activity.

SEQ ID NO: 38, 39, and 40 are the amino acid sequences of exemplary polypeptides of the invention that are each capable of binding to O-glycans but lack or have reduced O-glycoprotein-specific endoprotease activity. Relative to SEQ ID NOs: 35, 36 and 37, respectively, SEQ ID NOs: 38, 39 and 40 each include an additional N terminal Methionine and a C-terminal linker + His₆ tag.

SEQ ID NOs: 41 – 43 are the amino acid sequences of representative metalloprotease motifs of O-glycoprotein-specific endoproteases.

SEQ ID NOs: 44 – 46 are the amino acid sequences of representative disrupted metalloprotease motifs of polypeptides that are capable of binding to O-glycans but lack or have reduced O-glycoprotein-specific endoprotease activity.

SEQ ID NO: 47 is the amino acid sequence of the glycodrosocin peptide. There is a O-glycosylation site on the T residue.

SEQ ID NOs: 48 to 50 are the amino acid sequences of peptides which are not O-glycosylated.

Detailed Description of the Invention

It is to be understood that different applications of the disclosed products and methods may be tailored to the specific needs in the art. It is also to be understood that the terminology used herein is for the purpose of describing particular embodiments of the invention only, and is not intended to be limiting. All publications, patents and patent applications cited herein, whether supra or infra, are hereby incorporated by reference in their entirety.

As used in this specification and the appended claims, the singular forms “a”, “an”, and “the” include plural referents unless the content clearly dictates otherwise. Thus, for example, reference to “a polypeptide” includes “polypeptides”, and the like.

5 *General polypeptide features*

A “polypeptide” is used herein in its broadest sense to refer to a compound of two or more subunit amino acids, amino acid analogs, or other peptidomimetics. The term “polypeptide” thus includes short peptide sequences and also longer polypeptides and proteins. The terms “protein”, “peptide” and “polypeptide” may be used interchangeably.

10 As used herein, the term “amino acid” refers to either natural and/or unnatural or synthetic amino acids, including both D or L optical isomers, and amino acid analogs and peptidomimetics.

A polypeptide may be produced by suitable method, including recombinant or synthetic methods. For example, the polypeptide may be synthesised directly using standard techniques known in the art, such as Fmoc solid phase chemistry, Boc solid phase chemistry or by solution phase peptide synthesis. Alternatively, a polypeptide may be produced by transforming a cell, typically a bacterial cell, with a nucleic acid molecule or vector which encodes said polypeptide. Production of polypeptides by expression in bacterial host cells is described below and is exemplified in the Examples. The invention provides nucleic acid molecules and vectors which encode a polypeptide of the invention. The invention also provides a host cell comprising such a nucleic acid or vector. Exemplary polynucleotide molecules encoding polypeptides disclosed herein are provided as SEQ ID NOs: 3 and 7. Each of these sequences includes at the 5' end a codon for the N terminal methionine (ATG) and, prior to the stop codon (TAA) at the 3' end, codons for a Gly-Ser-Gly linker and a 6x His tag, which may optionally be excluded. The optional inclusion of an additional methionine and a tag are discussed in more detail below.

The terms “nucleic acid molecule” and “polynucleotide” are used interchangeably herein and refer to a polymeric form of nucleotides of any length, either deoxyribonucleotides or ribonucleotides, or analogs thereof. Non-limiting examples of polynucleotides include a gene, a gene fragment, messenger RNA (mRNA), cDNA, recombinant polynucleotides, plasmids, vectors, isolated DNA of any sequence, isolated RNA of any sequence, nucleic acid probes, and primers. A polynucleotide of the invention

encodes a polypeptide of the invention and may be provided in isolated or substantially isolated form. By substantially isolated, it is meant that there may be substantial, but not total, isolation of the polypeptide from any surrounding medium. The polynucleotides may be mixed with carriers or diluents which will not interfere with their intended use and still
5 be regarded as substantially isolated. A nucleic acid sequence which “encodes” a selected polypeptide is a nucleic acid molecule which is transcribed (in the case of DNA) and translated (in the case of mRNA) into a polypeptide in vivo when placed under the control of appropriate regulatory sequences, for example in an expression vector. The boundaries of the coding sequence are determined by a start codon at the 5' (amino) terminus and a
10 translation stop codon at the 3' (carboxy) terminus. For the purposes of the invention, such nucleic acid sequences can include, but are not limited to, cDNA from viral, prokaryotic or eukaryotic mRNA, genomic sequences from viral or prokaryotic DNA or RNA, and even synthetic DNA sequences. A transcription termination sequence may be located 3' to the coding sequence.

15 Polynucleotides can be synthesised according to methods well known in the art, as described by way of example in Sambrook *et al* (1989, Molecular Cloning - a laboratory manual; Cold Spring Harbor Press). The nucleic acid molecules of the present invention may be provided in the form of an expression cassette which includes control sequences operably linked to the inserted sequence, thus allowing for expression of the polypeptide of
20 the invention in vivo (e.g. in prokaryotic or eukaryotic expression systems). These expression cassettes, in turn, are typically provided within vectors (e.g., plasmids or recombinant viral vectors). Such an expression cassette may be administered directly to a host subject. Alternatively, a vector comprising a polynucleotide of the invention may be administered to a host subject. Preferably the polynucleotide is prepared and/or
25 administered using a genetic vector. A suitable vector may be any vector which is capable of carrying a sufficient amount of genetic information, and allowing expression of a polypeptide of the invention.

The present invention thus includes expression vectors that comprise such polynucleotide sequences. Such expression vectors are routinely constructed in the art of
30 molecular biology and may for example involve the use of plasmid DNA and appropriate initiators, promoters, enhancers and other elements, such as for example polyadenylation signals which may be necessary, and which are positioned in the correct orientation, in

order to allow for expression of a peptide of the invention. Other suitable vectors would be apparent to persons skilled in the art. By way of further example in this regard we refer to Sambrook *et al.*

The invention also includes cells that have been modified to express a polypeptide
5 of the invention. Such cells typically include prokaryotic cells such as bacterial cells, for example *E. coli*. Such cells may be cultured using routine methods to produce a polypeptide of the invention.

A polypeptide may be derivatised or modified to assist with their production, isolation or purification. For example, where a polypeptide of the invention is produced by
10 recombinant expression in a bacterial host cell, the sequence of the polypeptide may include an additional methionine (M) residue at the N terminus to improve expression. As another example, the polypeptide of the invention may be derivatised or modified by addition of a ligand which is capable of binding directly and specifically to a separation means. Alternatively, the polypeptide may be derivatised or modified by addition of one
15 member of a binding pair and the separation means comprises a reagent that is derivatised or modified by addition of the other member of a binding pair. Any suitable binding pair can be used. In a preferred embodiment where the polypeptide for use in the invention is derivatised or modified by addition of one member of a binding pair, the polypeptide is preferably histidine-tagged or biotin-tagged. Typically the amino acid coding sequence of
20 the histidine or biotin tag is included at the gene level and the polypeptide is expressed recombinantly in *E. coli*. The histidine or biotin tag is typically present at either end of the polypeptide, preferably at the C-terminus. It may be joined directly to the polypeptide or joined indirectly by any suitable linker sequence, such as 3, 4 or 5 glycine residues, or a mixture of glycine and serine residues. The histidine tag typically consists of six histidine
25 residues, although it can be longer than this, typically up to 7, 8, 9, 10 or 20 amino acids or shorter, for example 5, 4, 3, 2 or 1 amino acids.

A polypeptide may be provided in a substantially isolated or purified form. That is, isolated from the majority of the other components present in a cellular extract from a cell in which the polypeptide was expressed. By substantially purified, it will be understood
30 that the polypeptide is purified to at least 50%, 60%, 70%, 80% or preferably at least 90% homogeneity. Purity level may be assessed by any suitable means, but typically involves SDS-PAGE analysis of a sample, followed by Coomassie Blue detection. A polypeptide

may be mixed with carriers, diluents or preservatives which will not interfere with the intended purpose of the polypeptide and still be regarded as substantially isolated or purified. Where a polypeptide is provided in a composition with an additional active component, such as another polypeptide, each said polypeptide will individually be
5 purified to a high level of homogeneity prior to mixing in an appropriate ratio for the intended purpose of each. For example, two polypeptides may be each be purified to at least 90% homogeneity prior to combining in a 1:1 ratio.

A polypeptide (or mixture thereof) may be provided in lyophilised form, suitable for reconstitution in aqueous solution prior to use. The lyophilised composition has
10 improved stability enabling longer storage of the polypeptide. A method of preparing a polypeptide (or mixture thereof) in lyophilised form, comprising freeze-drying said polypeptide (or mixture) in a suitable buffer, such as Tris-buffered saline (TBS), is provided herein. A polypeptide is typically substantially purified prior to freeze-drying. The resulting polypeptide (or mixture) in lyophilised form is also provided. A method of
15 preparing a solution of a polypeptide (or mixture), comprising providing the polypeptide (or mixture) in lyophilised form and reconstituting with a suitable carrier or diluent, such as water, is also provided.

A polypeptide may be immobilised using methods known in the art, for example as described in Datta S *et al.*, Enzyme immobilization: an overview on techniques and
20 support materials, 3 *Biotech*, 3(1):1-9 (2013). For example, the polypeptide may be immobilised by adsorption, covalent binding, affinity immobilization or entrapment. Materials that can be used as supports include but are not limited to for example, natural supports such as agarose, collagen, gelatin, cellulose, pectin, sepharose, inorganic materials such as ceramics, silica, glass, activated carbon or charcoal, or synthetic polymers. For
25 example, the polypeptide may be immobilised on sepharose or agarose, optionally provided as a resin.

Polypeptides having endoprotease activity

30 Functional features of a polypeptide having endoprotease activity

In one embodiment, the present invention relates to a polypeptide having endoprotease activity specific for O-glycosylated proteins. In other words, the polypeptide

has O-glycoprotein-specific endoprotease activity. The polypeptide cleaves any O-linked glycoprotein, preferably any human O-linked glycoprotein. Examples of O-linked glycoproteins include any protein that comprises or consists all or part of an immunoglobulin, including full length antibodies, Fc fragments and Fc fusion proteins, particularly those of the IgA, IgD and IgG3 isotypes. Another example of an O-linked glycoprotein is Etanercept, which is a fusion protein of the ligand binding domain of human TNF α receptor 2 linked to the Fc portion of IgG1, with numerous O-glycosylation sites. Other examples of O-linked glycoproteins include erythropoietin (EPO), TNF α Receptors, fetuin, and plasminogen.

The hydrolysis (*i.e.* cleavage) of the substrate glycoprotein typically occurs with high specificity at a peptide bond N-terminal to and in close proximity to an O-glycosylated serine or threonine, and is O-glycan dependent. A polypeptide of the invention is preferably capable of cleaving such a peptide bond in close proximity to every O-glycosylated site in the substrate glycoprotein. The reaction preferably does not show any amino acid specificity or limitation, and in particular does not require any specific amino acid(s) to be present N-terminal to the O-glycosylated serine or threonine. When assessed using standard mass spectrometry parameters, the cleavage site is generally observed to be at the peptide bond immediately N terminal to each O-glycosylated residue.

Endoprotease activity and specificity of a given polypeptide may be determined by means of a suitable assay. For example, a standard O-glycoprotein substrate, such as an IgA molecule or erythropoietin (EPO), may be incubated with a test polypeptide. The starting materials and the reaction products may then be analysed by SDS-PAGE and/or mass spectrometry to determine the presence of cleavage products (if any) and if required also to further characterise those products. A glycoprotein substrate which is not O-glycosylated, such as an IgG1 molecule, may be used as a negative control. The results may be compared to those obtained in the same assay when the substrate is contacted with an exemplary polypeptide of the invention, such as a polypeptide consisting of the amino acid sequence of SEQ ID NO: 2. One unit of the polypeptide of SEQ ID NO: 2 is defined as the amount required to digest > 90% of 1 μ g of Erythropoietin (EPO) in combination with one unit of a sialidase mixture in 20 mM Tris buffer pH 6.8, overnight at 37°C as monitored by SDS-PAGE (preferred sialidase mixtures are described further below). A

test polypeptide preferably achieves a similar level of activity when present in the same amount. Exemplary assays are also described in the Examples.

Structural features of a polypeptide having endoprotease activity

5 This section sets out the structural features of a polypeptide in accordance with this embodiment, which apply in addition to the functional features outlined in the preceding section.

 The polypeptide is typically at least 150, 200, 250, 275, 280, 290, 300, 310, 320, 330, 340, 350 or 360 amino acids in length. The polypeptide is typically no larger than
10 400, 395, 390, 385, 380, 375, 370 or 365 amino acids in length. It will be appreciated that any of the above listed lower limits may be combined with any of the above listed upper limits to provide a range for the length the polypeptide. For example, the polypeptide may be 150 to 400 amino acids in length, or 280 to 380 amino acids in length. The polypeptide is preferably 340 to 380 amino acids in length, most preferably 360 to 375 amino acids in
15 length.

 The primary structure (amino acid sequence) of the polypeptide is based on the primary structure of the polypeptide encoded by the *Amuc1119* gene of *Akkermansia muciniphila*. The full sequence of this polypeptide is shown in SEQ ID NO: 4, which includes a signal motif at positions 1-24. The sequence with the signal motif removed is
20 shown in SEQ ID NO: 1.

 The polypeptide of the invention may comprise, consist essentially, or consist of the sequence of SEQ ID NO: 1.

 Alternatively, the polypeptide of the invention may comprise, consist essentially, or consist of a variant of the amino acid sequence of SEQ ID NO: 1 which is at least 50%
25 identical to the amino acid sequence of SEQ ID NO: 1. The variant sequence may be at least 60%, at least 70%, at least 80%, at least 85%, at least 90%, at least 95%, at least 98% or at least 99% identical to the sequence of SEQ ID NO:1. The identity level is preferably at least 85% or higher. Identity relative to the sequence of SEQ ID NO: 1 can be measured over a region of at least 100, at least 200, at least 300 or at least 350 or more contiguous
30 amino acids of the sequence shown in SEQ ID NO: 1, or more preferably over the full length of SEQ ID NO: 1. A variant is typically of a length which is no more than 50 amino

acids longer or shorter than the reference sequence, and is preferably of approximately (or exactly) the same length as the reference sequence.

Amino acid identity may be calculated using any suitable algorithm. For example the PILEUP and BLAST algorithms can be used to calculate identity or line up sequences
5 (such as identifying equivalent or corresponding sequences (typically on their default settings), for example as described in Altschul S. F. (1993) *J Mol Evol* 36:290-300; Altschul, S, F *et al* (1990) *J Mol Biol* 215:403-10. Software for performing BLAST analyses is publicly available through the National Center for Biotechnology Information (<http://www.ncbi.nlm.nih.gov/>). This algorithm involves first identifying high scoring
10 sequence pair (HSPs) by identifying short words of length W in the query sequence that either match or satisfy some positive-valued threshold score T when aligned with a word of the same length in a database sequence. T is referred to as the neighbourhood word score threshold (Altschul *et al*, supra). These initial neighbourhood word hits act as seeds for initiating searches to find HSPs containing them. The word hits are extended in both
15 directions along each sequence for as far as the cumulative alignment score can be increased. Extensions for the word hits in each direction are halted when: the cumulative alignment score falls off by the quantity X from its maximum achieved value; the cumulative score goes to zero or below, due to the accumulation of one or more negative-scoring residue alignments; or the end of either sequence is reached. The BLAST
20 algorithm parameters W, T and X determine the sensitivity and speed of the alignment. The BLAST program uses as defaults a word length (W) of 11, the BLOSUM62 scoring matrix (see Henikoff and Henikoff (1992) *Proc. Natl. Acad. Sci. USA* 89: 10915-10919) alignments (B) of 50, expectation (E) of 10, M=5, N=4, and a comparison of both strands.

The BLAST algorithm performs a statistical analysis of the similarity between two
25 sequences; see e.g., Karlin and Altschul (1993) *Proc. Natl. Acad. Sci. USA* 90: 5873-5787. One measure of similarity provided by the BLAST algorithm is the smallest sum probability (P(N)), which provides an indication of the probability by which a match between two polynucleotide or amino acid sequences would occur by chance. For example, a sequence is considered similar to another sequence if the smallest sum
30 probability in comparison of the first sequence to the second sequence is less than about 1, preferably less than about 0.1, more preferably less than about 0.01, and most preferably less than about 0.001. Alternatively, the UWGCG Package provides the BESTFIT program

which can be used to calculate identity (for example used on its default settings) (Devereux *et al* (1984) *Nucleic Acids Research* 12, 387-395).

The sequence of a polypeptide of the invention may comprise a variant of the amino acid sequence of SEQ ID NO: 1 in which modifications, such as amino acid
 5 additions, deletions or substitutions are made relative to the sequence of SEQ ID NO: 1. Unless otherwise specified, the modifications are preferably conservative amino acid substitutions. Conservative substitutions replace amino acids with other amino acids of similar chemical structure, similar chemical properties or similar side-chain volume. The amino acids introduced may have similar polarity, hydrophilicity, hydrophobicity, basicity,
 10 acidity, neutrality or charge to the amino acids they replace. Alternatively, the conservative substitution may introduce another amino acid that is aromatic or aliphatic in the place of a pre-existing aromatic or aliphatic amino acid. Conservative amino acid changes are well-known in the art and may be selected in accordance with the properties of the 20 main amino acids as defined in Table A1 below. Where amino acids have similar
 15 polarity, this can be determined by reference to the hydropathy scale for amino acid side chains in Table A2. A sequence of a polypeptide of the invention may comprise a variant of the amino acid sequence of SEQ ID NO: 1 in which up to 10, 20, 30, 40, 50 or 60 conservative substitutions are made.

Table A1 - Chemical properties of amino acids

Ala (A)	aliphatic, hydrophobic, neutral	Met (M)	hydrophobic, neutral
Cys (C)	polar, hydrophobic, neutral	Asn (N)	polar, hydrophilic, neutral
Asp (D)	polar, hydrophilic, charged (-)	Pro (P)	hydrophobic, neutral
Glu (E)	polar, hydrophilic, charged (-)	Gln (Q)	polar, hydrophilic, neutral
Phe (F)	aromatic, hydrophobic, neutral	Arg (R)	polar, hydrophilic, charged (+)
Gly (G)	aliphatic, neutral	Ser (S)	polar, hydrophilic, neutral
His (H)	aromatic, polar, hydrophilic, charged (+)	Thr (T)	polar, hydrophilic, neutral
Ile (I)	aliphatic, hydrophobic, neutral	Val (V)	aliphatic, hydrophobic, neutral
Lys (K)	polar, hydrophilic, charged(+)	Trp (W)	aromatic, hydrophobic, neutral
Leu (L)	aliphatic, hydrophobic, neutral	Tyr (Y)	aromatic, polar, hydrophobic

Table A2 - Hydropathy scale

<u>Side Chain</u>	<u>Hydropathy</u>
Ile	4.5
Val	4.2
Leu	3.8
Phe	2.8
Cys	2.5
Met	1.9
Ala	1.8

	Gly	-0.4
	Thr	-0.7
	Ser	-0.8
5	Trp	-0.9
	Tyr	-1.3
	Pro	-1.6
	His	-3.2
	Glu	-3.5
10	Gln	-3.5
	Asp	-3.5
	Asn	-3.5
	Lys	-3.9
	Arg	-4.5

15 The amino acid sequence of a polypeptide of the invention may comprises a variant of the amino acid sequence of SEQ ID NO: 1 as described above. However, certain residues in the amino acid sequence of SEQ ID NO: 1 are preferably retained within the said variant sequence. For example, the said variant sequence typically retains certain residues which are known to be required for endoprotease activity. Thus, the glutamate at
20 position 182 of SEQ ID NO: 1 (corresponds to position 206 of SEQ ID NO: 4) is preferably retained in the amino acid sequence of a polypeptide of the invention. This residue is thought to be required for electron transfer in the active site. Thus, a polypeptide of the invention typically comprises a variant of the amino acid sequence of SEQ ID NO: 1 which has a glutamate (E) at the position in said variant sequence which corresponds to
25 position 182 of SEQ ID NO: 1. Similarly, the histidine at position 181 of SEQ ID NO: 1 (corresponds to position 205 of SEQ ID NO: 4) is preferably retained in the amino acid sequence of a polypeptide of the invention. This residue is thought to be required for binding to a zinc ion co-factor.

The said glutamate and said histidine residue are both typically comprised within a
30 metalloprotease domain having the motif HEbbH, where b is an amino acid with no charge, such as amino acid A, C, F, G, I, L, M, N, P, Q, S, T, V or W. A preferred example of such a domain has the sequence HELGH (SEQ ID NO: 41), which corresponds to positions 181 to 185 of SEQ ID NO: 1 (positions 205 to 209 in SEQ ID NO: 4). Thus, a polypeptide of the invention typically comprises a variant of the amino acid sequence of
35 SEQ ID NO: 1 which comprises the motif HEbbH (such as HEIGH (SEQ ID NO: 42) or HELGH, preferably HELGH), at positions corresponding to positions 181 to 185 of SEQ

ID NO: 1. A polypeptide of the invention typically comprises an O-glycan specific binding domain located C-terminal to the metalloprotease domain.

The motif HEBbH may be compromised within a larger metalloprotease domain having the motif abxHEbbHbc, where a is amino acid V, T or G, b is an amino acid with
 5 no charge, such as amino acid A, C, F, G, I, L, M, N, P, Q, S, T, V or W, x is any amino acid, and c is a hydrophobic amino acid such as A, C, F, I, L, M, P, V, W or Y. A preferred example of such a domain has the sequence GMAHELGHGL (SEQ ID NO: 8), which corresponds to positions 178 to 187 of SEQ ID NO: 1 (positions 202 to 211 in SEQ ID NO: 4). Other examples include GVAHELGHNF (SEQ ID NO: 43). Thus, a
 10 polypeptide of the invention preferably comprises a variant of the amino acid sequence of SEQ ID NO: 1 which comprises the motif abxHEbbHbc, (such as GMAHELGHGL or GVAHELGHNF, preferably GMAHELGHGL), at positions corresponding to positions 178 to 187 of SEQ ID NO: 1. A polypeptide of the invention typically comprises an O-glycan specific binding domain located C-terminal to the metalloprotease domain.

15 Alternatively, a polypeptide of the invention may comprise, consist essentially, or consist of a shorter fragment of SEQ ID NO: 1 or of a variant thereof as described above. The fragments may be described as a truncated form of SEQ ID NO: 1 which retains O-glycoprotein specific-endoprotease activity. Such fragments are shorter than SEQ ID NO: 1 and are typically at least 100, 150 or 200 amino acids in length. The fragments typically
 20 comprise a metalloprotease domain at positions corresponding to positions 178 to 187 of SEQ ID NO: 1, including a glutamic acid residue (E) at a position which corresponds to position 182 of SEQ ID NO: 1 and a histidine residue (H) at a position which corresponds to position 181 of SEQ ID NO: 1, and an O-glycan specific binding domain located C-terminal to the metalloprotease domain.

25 Any polypeptide of the invention which comprises SEQ ID NO:1 or a variant thereof, or a fragment of either thereof, may optionally include an additional methionine at the N terminus and/or a histidine or other tag at the C terminus. Such additional sequences may aid with expression and/or purification. A histidine tag preferably consists of six histidine residues. The histidine tag is preferably linked to the C terminus by a linker,
 30 which is typically a short sequence of amino acids, such as 3 – 5 amino acids. The linker typically consists predominantly of glycine and serine residues, and may preferably include the sequence GSG. For example GSG and GSGLE are suitable linkers.

In summary therefore, a polypeptide of the invention is a polypeptide having O-glycoprotein-specific endoprotease activity which comprises:

- (a) an amino acid sequence of SEQ ID NO: 1;
 - (b) an amino acid sequence which is at least 85% identical to the amino acid sequence of SEQ ID NO: 1 or
 - (c) an amino acid sequence which is a fragment of the sequence of SEQ ID NO: 1 or a fragment of an amino acid which is 85% identical to the amino acid sequence of SEQ ID NO: 1;
- optionally wherein said polypeptide includes an additional methionine at the N terminus and/or a histidine tag at the C terminus, which tag may be joined to the C terminus by a linker.

The sequence of an exemplary polypeptide of the invention is provided as SEQ ID NO: 2.

The polypeptide may comprise or consist of the amino acid sequence of SEQ ID NO: 2.

An exemplary polynucleotide sequence encoding this polypeptide is shown in SEQ ID NO:

3.

Alternative polypeptides which have O-glycoprotein-specific endoprotease activity have been identified in *Pseudomonas aeruginosa* PAO1, *Bacteroides thetaiotaomicron* VPI-5482, and *Clostridium perfringens* (see three peptidases described in Noach *et al*; PNAS 2017, pE679-688 and supporting appendices, specifically Materials and Methods for Cloning, Protein Expression and Purification). The full length sequences of these polypeptides are provided as SEQ ID NOs: 32, 33 and 34. Each of these sequences includes a metalloprotease domain having the motif HEBbH as described above. The *Clostridium perfringens* sequence also has the longer metalloprotease domain having the motif abxHEbbHbc as described above. Each of these sequences can be optionally modified to remove any signal sequence or pro-enzyme sequences that may be present and/or to include an additional methionine at the N terminus and/or a histidine or other tag at the C terminus. Such additional sequences may aid with expression (e.g. in *E. coli*) and/or purification. Corresponding sequences with signal and other immature sequences removed are provided as SEQ ID NOs: 26, 27 and 28. Versions of these sequences optimised for expression in *E. coli* and subsequent purification (by inclusion of an additional methionine at the N terminus and a histidine tag at the C terminus) are provided as SEQ ID NOs: 29, 30 and 31. In methods described herein for use of a polypeptide of

the invention having O-glycoprotein-specific endoprotease activity, the polypeptide of the invention may optionally be replaced with one of these polypeptides. Preferred polypeptides for use in such methods therefore comprise, consist essentially, or consist of any one of SEQ ID NOs: 26 to 31.

5

Methods using the endoprotease activity of the polypeptide

The present invention also provides a method of hydrolysing an O-glycoprotein, wherein the method comprises contacting a sample of said protein with a polypeptide of the invention having O-glycoprotein-specific endoprotease activity and optionally further comprising detection of the hydrolysis products.

10

The present invention may also include a method for assessing the glycosylation status of a protein, comprising contacting a sample of said with a polypeptide of the invention having O-glycoprotein-specific endoprotease activity and analysing the products produced. The presence of cleavage products indicates that the protein in said sample is O-glycosylated, and thus the method may also be used for the detection of O-glycoproteins. The cleavage products may optionally be further analysed to identify the glycan chain and its position of attachment to the protein.

15

In such methods, a sample is contacted with a polypeptide of the invention under conditions suitable for polypeptide to interact with any proteins in the sample and for hydrolysis/cleavage reactions (endoprotease activity) to occur. Suitable conditions include incubation with a polypeptide of the invention for at least 20 minutes, 30 minutes, 40 minutes, 50 minutes, 60 minutes, 70 minutes, 80 minutes, 90 minutes or 120 minutes, 3 hours, 5 hours, 10 hours, or overnight. Incubation preferably takes place at room temperature, more preferably at approximately 20°C, 25°C, 30°C, 35°C, 40°C or 45°C, and most preferably at approximately 37°C. The methods described above may be carried out under any suitable pH. Suitable pH values include, for example, a pH of around 3.0, 3.5, 4, 4.5, 5, 5.5, 6, 6.5, 7, 7.5, 8, 8.5, 9 or 9.5. Preferred pH for the activity of a polypeptide of the invention is in the range 5.6 to 6.8. The method may be conducted in any suitable buffer, such as tris buffered saline (TBS) or phosphate buffered saline (PBS). The approximate ratio of the polypeptide of the invention to the protein content of the sample (enzyme:substrate) may be 1:1, 2:1, 4:1, 6:1, 10:1, 15:1, 20:1, 1:2, 1:4, or 1:6, 1:10, 1:15, 1:20, 1:40, 1:100, 1: 200 or 1:400. A preferred ratio is 1:20. Higher proportions of

20

25

30

enzyme to substrate may be beneficial if a shorter reaction time is required, or if the O-glycoprotein is heavily sialyated. Alternatively an earlier or simultaneous sialidase incubation step may be used to reduce sialic acid content, as is discussed in more detail below. The substrate is typically present at a concentration of 0.1mg/ml to 10mg/ml,
5 preferably around 0.1 to 2mg/ml.

The detection or analysis of the products produced may be assessed by any suitable analytical method, such as but not limited to mass spectrometry, HPLC, affinity chromatography, gel electrophoresis, SDS-PAGE, ELISA, lectin blotting, spectrometry, capillary electrophoresis and other standard laboratory techniques for the analysis of
10 proteins.

The sample in any of the above methods may be a sample taken from a patient, preferably a human patient. The results obtained may be used for a diagnostic purpose, for example to detect the presence of cancers which involve O-linked glycosylation. Such a use may involve comparison of the results obtained from the patient sample to those
15 obtained using a sample obtained from a healthy control.

In any method of the invention, the polypeptide can be used in combination with another enzyme such as a protease or glycosidase. The additional protease or glycosidase will typically further digest the substrate proteins, which may enhance the activity of the polypeptide of the invention and/or permit easier or more detailed analysis of the products.

20 For example, the present inventors have determined that a polypeptide of the invention demonstrates improved endoprotease activity if the O-glycans of a substrate protein are first modified to remove sialic acid. Thus, in a preferred method of the invention, the sample is contacted with an agent to remove sialic acid. Said agent may preferably be a sialidase enzyme or a mixture of such enzymes, which may be present in a
25 suitable buffer such as TBS or PBS. The buffer preferably comprises a low concentration of NaCl, typically up to 300mM, 250mM, 200mM, or 150mM. The NaCl concentration is preferably around 150mM, such as between 125mM and 175mM. Sialidases (or neuraminidases) catalyse the cleavage of terminal sialic acids from complex carbohydrates on glycoproteins and show a high degree of specificity. These enzymes target three
30 distinct sialic acid bonds which are commonly found within O-glycoproteins, namely α 2-3, α 2-6 and α 2-8 bonds. Sialidases that are suitable for use in the methods described include broad spectrum sialidases that target all of the α 2-3, α 2-6, or α 2-8 bonds, as well as narrow

spectrum sialidases that typically target only one type of bond. The α 2-3 bond is the most common in human glycoproteins, so if a narrow spectrum sialidase is used it is preferable that it targets this bond. Suitable sialidases may include viral or mammalian sialidases but are preferably sialidases isolated from bacteria, including but are not limited to strains of

5 *Clostridium perfringens*, *Arthrobacter ureafaciens*, *Streptococcus pneumoniae*, *Vibrio cholera* and *Akkermansia muciniphila*.

A preferred narrow spectrum sialidase is Am1757 isolated from *Akkermansia muciniphila*. Am1757 has specific activity against α 2-3 bonds. The wild-type sequence of Am1757 is provided as SEQ ID NO: 9, which includes a signal sequence. The wild-type

10 sequence of Am1757 lacking the signal sequence is provided as SEQ ID NO: 10. These sequences can be optionally modified to include an additional methionine at the N terminus and/or a histidine or other tag at the C terminus. Such additional sequences may aid with expression (e.g. in *E. coli*) and/or purification. A histidine tag preferably consists of six histidine residues. The histidine tag is preferably linked to the C terminus by a

15 linker, which is typically a short sequence of amino acids, such as 3 – 5 amino acids. The linker typically consists predominantly of glycine and serine residues, and may preferably include the sequence GSG. For example GSG and GSGLE are suitable linkers. An exemplary Am1757 sequence having an additional methionine at the N terminus and a GSGLE linker and His₆ tag at the C terminus is provided as SEQ ID NO: 11. Any

20 reference to Am1757 in the present disclosure may mean any of SEQ ID NOs: 9, 10 or 11, but preferably refers to a polypeptide which comprises or consists of the amino acid sequence of SEQ ID NO: 10. Most preferred is a polypeptide which consists of the amino acid sequence of SEQ ID NO: 11.

A preferred broad spectrum sialidase is Am0707 isolated from *Akkermansia*

25 *muciniphila*. Am0707 has activity against α 2-3, α 2-6 and α 2-8 bonds. The wild-type sequence of Am0707 is provided as SEQ ID NO: 12, which includes a signal sequence. The wild-type sequence of Am0707 lacking the signal sequence is provided as SEQ ID NO: 13. These sequences can be optionally modified to include an additional methionine at the N terminus and/or a histidine or other tag at the C terminus. Such additional

30 sequences may aid with expression and/or purification. A histidine tag preferably consists of six histidine residues. The histidine tag is preferably linked to the C terminus by a linker, which is typically a short sequence of amino acids, such as 3 – 5 amino acids. The

linker typically consists predominantly of glycine and serine residues, and may preferably include the sequence GSG. For example GSG and GSGLE are suitable linkers. An exemplary Am0707 sequence having an additional methionine at the N terminus and a GSGLE linker and His₆ tag at the C terminus is provided as SEQ ID NO: 14. Any
5 reference to Am0707 in the present disclosure may mean any of SEQ ID NOs: 12, 13 or 14, but preferably refers to a polypeptide which comprises or consists of the amino acid sequence of SEQ ID NO: 13. Most preferred is a polypeptide which consists of the amino acid sequence of SEQ ID NO: 14.

A preferred sialidase mixture capable of hydrolyzing all sialic acid bonds
10 comprises Am1757 and Am0707 isolated from *Akkermansia muciniphila*. The mixture of Am1757 and Am0707 is typically in a 1:1 ratio. A particularly preferred mixture may comprise a polypeptide consisting of the amino acid sequence of SEQ ID NO: 11 and a polypeptide consisting of the amino acid sequence of SEQ ID NO: 14.

Methods of the invention may preferably comprise incubating a sample with
15 Am1757 or with a mixture of Am1757 and Am0707 prior to or simultaneously with the polypeptide of invention, under conditions suitable for activity of the sialidases. The invention also provides a composition (in lyophilised or solution form) comprising a polypeptide of the invention and Am1757 or a mixture of Am1757 and Am0707. Such a composition may preferably be lyophilised in tris buffered saline, which may be at around
20 pH7.6. In such a composition, Am1757 and Am0707 will preferably be present in a 1:1 ratio to each other, with the total sialidase content (Am1757+Am0707) also being present at a 1:1 ratio relative to the polypeptide of the invention. For example, if a composition includes 2000 units of the polypeptide of the invention, it will also include 2000 units of sialidase, in which said 2000 units of sialidase comprise 1000 units Am1757 and 1000
25 units Am0707. A unit of sialidase mixture is typically the amount required to hydrolyse sialic acids from $\geq 90\%$ of 1 μg glycoprotein (fetuin) when incubated in 20 mM Tris pH 6.8 at 37 °C for 2 h at 37°C as monitored by SDS-PAGE. A unit of the polypeptide of the invention is typically the amount required to digest $> 90\%$ of 1 μg of Erythropoietin (EPO) when incubated in 20 mM Tris buffer pH 6.8, overnight with one unit of sialidase mixture
30 at 37°C as monitored by SDS-PAGE.

The invention also provides a kit comprising a polypeptide of the invention in a separate container from Am1757 or a mixture of Am1757 and Am0707, with instructions for the combined use of the different enzymes.

As another example, in any of the methods described herein, the sample may be
5 incubated with an N-glycosidase prior to, at the same time as, or after contacting the sample with the polypeptide of the invention, to remove N-glycans from target proteins. An exemplary N-glycosidase is PNGaseF. Other N-glycosidases that may be used when the sample includes immunoglobulins are EndoS (see SEQ ID NO: 1 of WO2008071418) or EndoS2 (may be referred to as EndoS49 - see SEQ ID NO: 1 of WO2013037824). Each
10 of these enzymes removes the N-linked glycoprotein from Asn-297 of IgG1. In a particular embodiment, the sample may be contacted with an N-glycosidase and a sialidase (or mixture thereof) in addition to the polypeptide of the invention. In such a method, the sialidase (or mixture) may be applied first, prior to simultaneous addition of the N-glycosidase and the polypeptide of the invention. Such a method is particularly suitable
15 for the subsequent assessment of O-glycosylation sites, typically achieved by separation of the products e.g. using RPLC and subsequent analysis of the different fractions e.g. using mass spectrometry.

As another example, in any of the methods described herein, the sample may be incubated with a protease prior to, at the same time as, or after contacting the sample with
20 the polypeptide of the invention, to further digest the target protein. Suitable general proteases include trypsin, chymotrypsin, Lys-C, Asp-N, Glu-C, Arg-C or similar endoproteases, or Arg-gingipain (RgpB) of *Porphyromonas gingivalis*.

If the sample includes immunoglobulins, immunoglobulin proteases may be used such as SpeB (see sequence in WO2015040125), Immunoglobulin G-degrading enzyme of
25 *S. pyogenes* (IdeS - see sequence in WO2015040125), Immunoglobulin G-degrading enzyme of *S. equi* subspecies *zooepidemicus* (IdeZ), Lys-gingipain (Kgp) of *Porphyromonas gingivalis*, and Immunoglobulin G-degrading Enzyme of *S. agalactiae* (IgdE_{agalactiae} - see SEQ ID NO: 3 of PCT/EP2017/052463). Use of any combination of these proteases in a method of the present invention may assist with determination of
30 glycosylation sites on monoclonal antibodies and subunits thereof, for example using mass spectrometry (middle down approach).

As another example, in any of the methods described herein, the sample may be incubated with an O-glycosidase after contacting the sample with a polypeptide of the invention. For instance, to simplify the analysis of the products produced, the products are subjected to digestion by an O-glycosidase for removal of the O-glycans prior to further analysis by any suitable method. Suitable O-glycosidases may be obtained from a strain of *Enterococcus faecalis*, *Streptococcus oralis*, or *Bifidobacterium bifidum*, preferably *Enterococcus faecalis* or *Streptococcus oralis*, most preferably *Streptococcus oralis*. The sequence of an exemplary O-glycosidase from *Streptococcus oralis* is provided as SEQ ID NO: 15.

10

Polypeptides which bind to O-linked glycoproteins but lack or have a reduced endoprotease activity

Functional features of a polypeptide lacking endoprotease activity

15 In one embodiment, the present invention relates to a polypeptide lacking or having a reduced endoprotease activity, whilst retaining the ability to bind O-glycans. In other words, the polypeptide may be described as an O-glycan-specific binding agent which does not significantly hydrolyse a glycoprotein to which said glycan is attached.

O-glycoprotein endoprotease activity may be determined using any suitable method, but may typically employ the same assay as described above for polypeptides of the invention which possess such activity. A lack of activity in a test polypeptide will be indicated by the absence of cleavage products following incubation with an O-glycoprotein substrate. Cleavage of the same substrate by a polypeptide comprising the amino acid sequence of SEQ ID NO: 1 may be used as a positive control. A reduction in activity in the test polypeptide may be determined by comparison to the same control. The polypeptide of the invention typically has O-glycoprotein endoprotease activity which is reduced relative to the activity of a polypeptide comprising the amino acid sequence of SEQ ID NO: 1. The polypeptide of the invention typically has O-glycoprotein endoprotease activity which is less than 95%, 90%, 85%, 80%, 75%, 70%, 65%, 60%, 55%, 50%, 45%, 40%, 35%, 30%, 25%, 20%, 15%, 10% or 5% compared to the activity of a polypeptide comprising the amino acid sequence of SEQ ID NO: 1.

The ability of a polypeptide to bind to O-glycan or O-glycoprotein may also be assessed by any suitable method. One such method involves immobilising a test polypeptide, e.g. on sepharose in a spin column, followed by incubation with a sample containing O-glycoproteins and/or O-glycans. If the test polypeptide has O-glycan and/or
5 O-glycoprotein binding ability, the O-glycoproteins and/or O-glycans will be detectable bound to the column or in a subsequent eluent. Preferably, the polypeptide is able to bind all O-glycoproteins which are hydrolysable by a polypeptide comprising the amino acid sequence of SEQ ID NO: 1.

Exemplary assays of this type are described in the examples.

10

Structural features of a polypeptide lacking endoprotease activity

This section sets out the structural features of a polypeptide in accordance with this embodiment, which apply in addition to the functional features outlined in the preceding section. A polypeptide in accordance with this embodiment of the invention may possess
15 the same structural features as described above in connection with a polypeptide having endoprotease activity, with the exception that the amino acid sequence is modified by one or more amino acid additions, deletions or substitutions such that said activity is reduced or eliminated. Typically, a polypeptide in accordance with this embodiment of the invention will not include an intact metalloprotease motif of HEbbH or abxHEbbHbc. Said motif
20 may be disrupted by addition, deletion or substitution, but is preferably disrupted by at least one amino acid substitution. Preferably, the substitution involves replacement of the glutamic acid (E) residue in the said motif with an alternative amino acid and/or the replacement of the histidine (H) residue in the position corresponding to the 1st position of the shorter motif (the 4th position of the longer motif) and/or replacement of the histidine
25 (H) residue in the position corresponding to the 5th position of the shorter motif (the 8th position of the longer motif). Preferably, either or both or all three said substitutions is non-conservative. The substitution of the E residue should reduce or eliminate electron transfer. The substitution of either of the H residues should reduce or eliminate Zinc ion co-factor binding. The E residue is therefore preferably substituted with a non-polar or
30 uncharged amino acid, such as A, C, F, G, I, L, M, N, P, Q, S, T, V or W, but is most preferably substituted with Alanine (A) or Glycine (G). The H residues may each

individually be substituted with any non-H amino acid, but non-polar amino acids such as A and G are again preferred.

Thus, a polypeptide of the invention may comprise, consist essentially, or consist of the amino acid sequence of SEQ ID NO: 1 in which the metalloprotease motif of HEbbH
5 or abxHEbbHbc is disrupted, preferably by the replacement of the glutamic acid residue in the position corresponding to position 182 of SEQ ID NO: 1 and/or the replacement of the histidine residue corresponding to position 181 of SEQ ID NO: 1 with an alternative amino acid and/or the replacement of the histidine residue corresponding to position 185 of SEQ ID NO: 1 with an alternative amino acid. In other words, the polypeptide may be
10 described as not comprising the metalloprotease motif HEbbH and preferably comprising a disrupted version of said motif, such that:

- (a) H in the first position is replaced with an alternative amino acid, preferably A or G; and/or
- (b) E in the second position is replaced with an uncharged amino acid,
15 optionally A, C, F, G, I, L, M, N, P, Q, S, T, V or W, preferably A or G; and/or
- (c) H in the fifth position is replaced with an alternative amino acid, preferably A or G

wherein b in the said motif is an uncharged amino acid, optionally A, C, F, G, I, L, M, N,
20 P, Q, S, T, V or W.

The said polypeptide may therefore be described as comprising the motif xbbbx, wherein:

- (a) x is preferably any amino acid except H, and is preferably A or G; and/or
- (b) b is an uncharged amino acid, optionally A, C, F, G, I, L, M, N, P, Q, S, T,
25 V or W, preferably A or G;

optionally wherein said motif is present in said polypeptide at positions corresponding to positions 181 to 185 of SEQ ID NO: 1.

The said polypeptide may therefore comprise a disrupted metalloprotease motif, for example with any one of the following sequences: HALGH (SEQ ID NO: 44), AELGH
30 (SEQ ID NO: 45) or most preferably AALGH (SEQ ID NO: 46). Sequences comprising this type of specific change to SEQ ID NO: 1 are shown in SEQ ID NO: 5 and SEQ ID NO: 20. In other words therefore, a polypeptide of this embodiment of the invention may

comprise, consist essentially, or consist of the amino acid sequence of SEQ ID NO: 5 or SEQ ID NO: 20.

The said polypeptide may alternatively be described as comprising the motif abxxbbbxbc, wherein:

- 5 (a) a is amino acid V, T or G;
 - (b) b is an uncharged amino acid, optionally A, C, F, G, I, L, M, N, P, Q, S, T, V or W, preferably A or G;
 - (c) x is any amino acid except that the amino acid in the 4th and/or 8th position of the motif is preferably not H, and is preferably A or G; and
 - 10 (d) c is a hydrophobic amino acid, optionally A, C, F, I, L, M, P, V, W or Y;
- optionally wherein said motif is present in said polypeptide at positions corresponding to positions 178 to 187 of SEQ ID NO: 1.

The said polypeptide may therefore comprise a disrupted metalloprotease motif, for example with any one of the following sequences: GMAHALGHGL (SEQ ID NO: 23),
 15 GMAAELGHGL (SEQ ID NO: 24) or most preferably GMAAALGHGL (SEQ ID NO: 25). Sequences comprising this type of specific change to SEQ ID NO: 1 are shown in SEQ ID NO: 5 and SEQ ID NO: 20. In other words therefore, a polypeptide of this embodiment of the invention may comprise, consist essentially, or consist of the amino acid sequence of SEQ ID NO: 5 or SEQ ID NO: 20.

20 Alternatively, the polypeptide of the invention may comprise, consist essentially, or consist of a variant of the amino acid sequence of SEQ ID NO: 5 or SEQ ID NO: 20 which is at least 50% identical to the amino acid sequence of SEQ ID NO: 5, provided that a glutamic acid residue is not introduced in the position corresponding to position 182 of SEQ ID NO: 1 and/or a histidine residue is not introduced in the position corresponding to
 25 position 181 of SEQ ID NO: 1 and/or a histidine residue is not introduced in the position corresponding to position 185 of SEQ ID NO: 1.

The variant sequence may be at least 60%, at least 70%, at least 80%, at least, 85%, at least 90%, at least 95%, at least 98% or at least 99% identical to the sequence of SEQ ID NO: 5. The identity level is preferably at least 85% or higher. Identity relative to the
 30 sequence of SEQ ID NO: 5 or SEQ ID NO: 20 can be measured over a region of at least 100, at least 200, at least 300 or at least 350 or more contiguous amino acids of the sequence shown in SEQ ID NO: 5 or SEQ ID NO: 20, or more preferably over the full

length of SEQ ID NO: 5 or SEQ ID NO: 20. A sequence of a polypeptide of the invention may comprise a variant of the amino acid sequence of SEQ ID NO: 5 or SEQ ID NO: 20 in which up to 10, 20, 30, 40, 50 or 60 conservative substitutions are made. Determinations of sequence identity and an explanation of conservative and non-conservative substitutions are providing in the section relating to polypeptides having endoprotease activity and apply
5 equally here.

Alternatively, a polypeptide of the invention may comprise, consist essentially, or consist of a shorter fragment of SEQ ID NO: 5 or SEQ ID NO: 20, or of a variant thereof as described above. The fragments may be described as a truncated form of SEQ ID NO: 5
10 or SEQ ID NO: 20 which retains O-glycoprotein binding activity. Such fragments are shorter than SEQ ID NO: 1 and are typically at least 100, 150 or 200 amino acids in length.

Any polypeptide of the invention which comprises SEQ ID NO: 5 or SEQ ID NO :20, or a variant thereof, or a fragment of either thereof, may optionally include an additional methionine at the N terminus and/or a histidine or other tag at the C terminus.
15 Such additional sequences may aid with expression and/or purification. A histidine tag preferably consists of six histidine residues. The histidine tag is preferably linked to the C terminus by a linker, which is typically a short sequence of amino acids, such as 3 – 5 amino acids. The linker typically consists predominantly of glycine and serine residues, and may preferably include the sequence GSG. For example GSG and GSGLE are
20 suitable linkers.

In summary therefore, a polypeptide of the invention is a polypeptide having O-glycoprotein binding activity but lacking or having reduced O-glycoprotein-specific endoprotease activity which comprises:

- (a) an amino acid sequence of SEQ ID NO: 5 or SEQ ID NO: 20;
- 25 (b) an amino acid sequence which is at least 85% identical to the amino acid sequence of SEQ ID NO: 5 or SEQ ID NO: 20;
- (c) an amino acid sequence which is a fragment of the sequence of SEQ ID NO: 5 or SEQ ID NO: 20, or a fragment of an amino acid which is 85% identical to the amino acid sequence of SEQ ID NO: 5 or SEQ ID NO: 20;
- 30 optionally wherein said polypeptide includes an additional methionine at the N terminus and/or a histidine tag at the C terminus, which tag may be joined to the C terminus by a linker.

The sequence of an exemplary polypeptide of the invention is provided as SEQ ID NO: 6.
 The polypeptide may comprise or consist of the amino acid sequence of SEQ ID NO: 6.
 An exemplary polynucleotide sequence encoding this polypeptide is shown in SEQ ID NO:
 7. The sequence of another exemplary polypeptide of the invention is provided as SEQ ID
 5 NO: 21. The polypeptide may comprise or consist of the amino acid sequence of SEQ ID
 NO: 21. An exemplary polynucleotide sequence encoding this polypeptide is shown in
 SEQ ID NO: 22.

The said polypeptide is preferably provided in immobilised form, such as on
 agarose or sepharose, optionally provided as a resin.

10 Additional polypeptides of the invention having O-glycoprotein binding activity
 but lacking or having reduced O-glycoprotein-specific endoprotease activity may be
 produced by disrupting the metalloprotease domain motif HEbbH or abxHEbbHbc in any
 other polypeptide having O-glycoprotein-specific endoprotease activity which comprises
 such a motif. In methods described herein for use of a polypeptide of the invention lacking
 15 or having reduced endoprotease activity, references to the polypeptide of the invention
 include such polypeptides. Disruption of the said motif is preferably achieved as described
 above such that:

- (a) H in the first position is replaced with an alternative amino acid, preferably
 A or G; and/or
- 20 (b) E in the second position is replaced with an uncharged amino acid,
 optionally A, C, F, G, I, L, M, N, P, Q, S, T, V or W, preferably A or G;
 and/or
- (c) H in the fifth position is replaced with an alternative amino acid, preferably
 A or G
- 25 wherein b in the said motif is an uncharged amino acid, optionally A, C, F, G, I, L, M, N,
 P, Q, S, T, V or W.

The said polypeptide may therefore be described as comprising the motif xbbbx,
 wherein:

- (a) x is preferably any amino acid except H, and is preferably A or G; and/or
- 30 (b) b is an uncharged amino acid, optionally A, C, F, G, I, L, M, N, P, Q, S, T,
 V or W, preferably A or G;

Other polypeptides having O-glycoprotein-specific endoprotease activity and which may be disrupted in this way are described above as having been identified in *Pseudomonas aeruginosa* PAO1, *Bacteroides thetaiotaomicron* VPI-5482, and *Clostridium perfringens* (see three peptidases described in Noach *et al*; PNAS 2017, pE679-688 and supporting
5 appendices, specifically Materials and Methods for Cloning, Protein Expression and Purification). The full length sequences of these polypeptides are provided as SEQ ID NOs: 32, 33 and 34. Corresponding mature sequences (e.g. with signal and other sequences removed) are provided as SEQ ID NOs: 26, 27 and 28. Versions of these sequences optimised for expression in *E. coli* and subsequent purification (by inclusion of an additional
10 methionine at the N terminus and a histidine tag at the C terminus) are provided as SEQ ID NO: 29, 30 and 31. Each of SEQ ID NOs: 26 to 34 therefore includes a metalloprotease domain having the motif HEbbH which may be disrupted to produce a motif xbbbx as described above, to product an additional polypeptide of the invention. Versions of SEQ ID NOs: 26, 27 and 28 in which the HEbbH motif has been so disrupted are provided as SEQ
15 ID NOs: 35, 36 and 37. Versions of these sequences optimised for expression in *E. coli* and subsequent purification (by inclusion of an additional methionine at the N terminus and a histidine tag at the C terminus) are provided as SEQ ID NO: 38, 39, and 40. Polypeptides of the invention which lack or have reduced O-glycoprotein-specific endoprotease activity may comprises, consist essentially, or consist of any one of SEQ ID NOs: 35, 36, 37, 38, 39
20 or 40.

Methods using LS mutant lacking or having reduced endoprotease activity

The present invention also provides a method of binding to an O-glycan, wherein the method comprises contacting a sample comprising the O-glycan with a polypeptide of
25 the invention capable of binding to an O-glycan and which lacks or has reduced endoprotease activity specific for O-glycosylated proteins. The method optionally further includes determining whether or not an O-glycan has been bound and/or separating the O-glycan and any linked glycoprotein from the resulting mixture.

The present invention may also include a method for assessing the glycosylation
30 status of a protein, comprising contacting a sample of said with a polypeptide of the invention capable of binding to an O-glycan and which lacks or has reduced endoprotease

activity specific for O-glycosylated proteins, and determining whether or not the protein is bound by the said polypeptide.

The present invention may also include a method for detecting O-linked glycoproteins in a sample, wherein the method comprises contacting said sample with a polypeptide of the invention capable of binding to an O-glycan and which lacks or has reduced endoprotease activity specific for O-glycosylated proteins, to thereby allow formation of an O-linked glycoprotein-polypeptide complex. The method may optionally include separating said polypeptide from the contacted sample and determining whether the separated polypeptide is bound to the O-linked glycoproteins, thereby determining the presence or absence of O-linked glycoproteins in the sample can thereby be determined. The method may also be used for isolating an O-glycan or O-linked glycoprotein from a sample containing O-glycans or O-linked glycoproteins.

In such methods, a sample is contacted with a polypeptide of the invention under conditions suitable for the polypeptide to interact with any O-glycan or proteins in the sample and for binding to occur. Suitable conditions include incubation with a polypeptide of the invention for at least 20 minutes, 30 minutes, 40 minutes, 50 minutes, 60 minutes, 70 minutes, 80 minutes, 90 minutes or 120 minutes, 3 hours, 5 hours, 10 hours, or overnight, typically with mixing e.g. end-over-end mixing. Incubation preferably takes place at room temperature, more preferably at approximately 20°C, 25°C, 30°C, 35°C, 40°C or 45°C, and most preferably at approximately 37°C. The methods described above may be carried out under any suitable pH. Suitable pH values include, for example, a pH of around 3.0, 3.5, 4, 4.5, 5, 5.5, 6, 6.5, 7, 7.5, 8, 8.5, 9 or 9.5. Preferred pH for the activity of a polypeptide of the invention is in the range 5.6 to 6.8. The method may be conducted in any suitable buffer, such as tris buffered saline (TBS) or phosphate buffered saline (PBS). The approximate ratio of the polypeptide of the invention to the protein content of the sample may be 1:1, 2:1, 4:1, 6:1, 10:1, 15:1, 20:1, 1:2, 1:4, or 1:6, 1:10, 1:15, 1:20, 1:40, 1:100, 1: 200 or 1:400 (wt:wt). A preferred ratio is 1:1 (wt:wt). Higher proportions of polypeptide to substrate may be beneficial if a shorter reaction time is required, or if the O-glycoprotein is heavily sialylated. Alternatively an earlier or simultaneous sialidase incubation step may be used to reduce sialic acid content, as is discussed in more detail below. The substrate is typically present at a concentration of around 0.01mg/ml to

10mg/ml, preferably around 0.1mg/ml to 10mg/ml, around 0.01mg/ml to 2mg/ml, or around 0.1mg/ml to 2mg/ml.

The detection or analysis of the sample to determine whether an O-glycan or O-linked glycoprotein has been bound may be assessed by any suitable analytical method, such as but not limited to mass spectrometry, HPLC, affinity chromatography, gel
5 electrophoresis, SDS-PAGE, ELISA, lectin blotting, spectrometry, capillary electrophoresis and other standard laboratory techniques for the analysis of proteins. For example, the molecular weight of the polypeptide may be analysed. The polypeptide of the invention bound to an O-glycan or O-linked glycoprotein will have a higher molecular
10 weight than a polypeptide not bound to an O-glycan or O-linked glycoprotein.

Separation of the bound O-glycan or O-linked glycoprotein and the polypeptide of the invention may be carried out by any suitable separation means. For example, the separation means may comprise a population of magnetic nanoparticles. These may be separated from a sample using magnetic field separation, preferably high-gradient
15 magnetic field separation. Examples of reagents or separating means are populations of magnetic particles capable of binding to the polypeptide of the invention. For example, where the polypeptide is derivatised with a histidine tag, the magnetic particles contain on their surface chelating groups which carry a nickel, copper or zinc ion. Alternatively, where the polypeptide is derivatised with a biotin tag, the magnetic particles contain on
20 their surface streptavidin.

The separation means may also comprise a solid support to which the polypeptide of the invention is immobilised. Examples of solid supports include those described in previous sections, and may include agarose or sepharose resins, cross-linked agarose beads, or similar. The support may be used as the matrix in an affinity chromatography
25 column. Alternatively the solid support may comprise a suitable silica-based material or polystyrene, or a plastic container such as a microtiter plate or equivalent, to which the polypeptide of the invention can be directly adsorbed.

Alternative separation means include reagents comprising antibodies specific to the polypeptide of the invention, which may be generated by methods standard in the art.
30 Antibodies in this sense include a monoclonal antibody, a polyclonal antibody, a single chain antibody, a chimeric antibody, a CDR-grafted antibody or a humanized antibody. The antibody may be an intact immunoglobulin molecule or a fragment thereof such as a

Fab, F(ab')₂ or Fv fragment. If more than one antibody is present, the antibodies preferably have different non-overlapping determinants such that they may bind to the polypeptide of the invention simultaneously. The antibody may be bound to a solid support or may be labeled or conjugated to another chemical group or molecule to assist with their separation or isolation. For example, typical chemical groups include
5 fluorescent labels such as Fluorescein (FITC) or Phycoerythrin (PE), or tags such as biotin.

Other suitable means of separation include elution of the protein from the (typically immobilised) polypeptide by contacting the polypeptide from the contacted sample with a suitable elution buffer. The choice of elution buffer may depend on the acid-sensitivity of
10 the protein. Preferred elution buffers may comprise high molar concentrations of urea (typically at least 5, 6, 7 or most preferably at least 8M) or high concentrations of a detergent (typically at least around 1%, 5% or 10%). Suitable detergents include Nonidet P40, Triton X-100, Tween 20, CHAPS, sodium deoxycholate, and RapiGest SF surfactant, but Sodium dodecyl sulphate (SDS) is preferred. High molar urea is preferred over
15 detergent since downstream procedures are more likely to be sensitive to the presence of detergent.

Another preferred elution buffer comprises a suitable concentration of a polypeptide of the invention which has O-glycoprotein endoprotease activity, e.g. a polypeptide of SEQ ID NO: 1. Cleavage at the O-glycan by this polypeptide will release
20 bound O-glycoproteins, thus obviating the need for Urea or detergent-based elution.

Preferred methods of elution of O-glycoproteins from immobilised polypeptides of the invention are demonstrated in the Examples.

The sample in any of the above methods may be a sample taken from a patient, preferably a human patient. The results obtained may be used for a diagnostic purpose, for
25 example to detect the presence of cancers which involve O-linked glycosylation. Such a use may involve comparison of the results obtained from the patient sample to those obtained using a sample obtained from a healthy control.

In any method of the invention, the polypeptide can be used in combination with another enzyme such as a protease or glycosidase. The additional protease or glycosidase
30 will typically further digest the substrate proteins or glycans, which may permit easier or more detailed analysis of the products.

For example, the polypeptide of the invention may be used in combination with an agent to remove sialic acid. Said agent may preferably be a sialidase enzyme or a mixture of such enzymes as described in the section above. The invention also provides a composition (in lyophilised or solution form) comprising a polypeptide of the invention and Am1757 or a mixture of Am1757 and Am0707. The invention also provides a kit comprising a polypeptide of the invention in a separate container from Am1757 or a mixture of Am1757 and Am0707, with instructions for the combined use of the different enzymes.

As another example, in any of the methods described herein, the sample may be incubated with an N-glycosidase prior to, at the same time as, or after contacting the sample with the polypeptide of the invention, to remove N-glycans from target proteins. An exemplary N-glycosidase is PNGaseF. Other N-glycosidases that may be used when the sample includes immunoglobulins are EndoS (see SEQ ID NO: 1 of WO2008071418) or EndoS2 (may be referred to as EndoS49 - see SEQ ID NO: 1 of WO2013037824). Each of these enzymes removes the N-linked glycoprotein from Asn-297 of IgG1. The sample may be contacted with an N-glycosidase and a sialidase (or mixture thereof) in addition to the polypeptide of the invention. In such a method, the sialidase (or mixture) may be applied first, prior to simultaneous addition of the N-glycosidase and the polypeptide of the invention.

As another example, in any of the methods described herein, the sample may be incubated with a protease prior to, at the same time as, or after contacting the sample with the polypeptide of the invention, to further digest the target protein. Suitable general proteases include trypsin, chymotrypsin, Lys-C, Asp-N, Glu-C, Arg-C or similar endoproteases, or Arg-gingipain (RgpB) of *Porphyromonas gingivalis*.

If the sample includes immunoglobulins, immunoglobulin proteases may be used such as SpeB (see sequence in WO2015040125), Immunoglobulin G-degrading enzyme of *S. pyogenes* (IdeS - see sequence in WO2015040125), Immunoglobulin G-degrading enzyme of *S. equi* subspecies *zooepidemicus* (IdeZ), Lys-gingipain (Kgp) of *Porphyromonas gingivalis*, and Immunoglobulin G-degrading Enzyme of *S. agalactiae* (IgdE_{agalactiae} - see SEQ ID NO: 3 of PCT/EP2017/052463). Use of any combination of these proteases in a method of the present invention may assist with analysis of the substrate protein or glycan, for example using mass spectrometry.

As another example, in any of the methods described herein, the isolated O-linked glycoprotein may be incubated with an O-glycosidase to remove the O-glycans prior to further analysis by any suitable method. Suitable O-glycosidases may be obtained from a strain of *Enterococcus faecalis*, *Streptococcus oralis*, or *Bifidobacterium bifidum*,
5 preferably *Enterococcus faecalis* or *Streptococcus oralis*, most preferably *Streptococcus oralis*. The sequence of an exemplary O-glycosidase from *Streptococcus oralis* is provided as SEQ ID NO: 15.

The following Examples illustrate the invention:

10 **Example 1**

MATERIALS AND METHODS

Mutagenesis of LS

Site-directed mutagenesis using Q5 (NEB) was performed according to manufacturer's instructions (annealing temperature 68°C, 3 min elongation) using primers
15 E206A_fwd 5'-ATGGCGCACGC GCTGGGCCACG-3' and 5'-GCCACCGTAC CATTTCGTC-3' (rev); thus changing a glutamic acid to an alanine in an *Amuc1119* gene from *Akkermansia muciniphila* to create the mutant, Amuc1119_{E206A} (LS_{E206A}). The construct was transformed into DH5α *E. coli*, isolated and verified using sequencing (GATC Biotech).

20

Recombinant expression of LS and LS_{E206A}

The gene *Amuc1119* from *Akkermansia muciniphila* ATCC BAA-835, and a mutant, Amuc1119_{E206A} (Amuc1119 – LS; Amuc1119_{E206A} – LS_{E206A}), were codon optimized for expression in *E. coli* (DNA 2.0) and cloned into an expression vector with a
25 C-terminal 6xHis-tag as part of the fusion protein.

The codon-optimized genes were transformed into BL21(DE3) Star cells. *E. coli* was routinely cultured in LB at 37°C, 180 rpm. In the presence of the plasmid, 50 µg/mL kanamycin was added. After overnight incubation, cultures were diluted 1:20 in fresh LB(kana), and grown until OD₆₂₀ ~ 0.7-0.8, after which recombinant protein expression was
30 induced by addition of 1 mM IPTG, and the expression continued for 6 hours before the cells were collected and frozen. Frozen cells were thawed and resolved in His binding buffer (20 mM NaP pH 7.4, 500 mM NaCl, 20 mM imidazole), and sonicated for release of intracellular

proteins. Cell debris was removed by centrifugation. Sterile filtered supernatant was affinity purified on a nickel column, and re-buffered to 20 mM Tris-HCl pH 8.0 on a PD-25 column. Concentration of the proteins was determined using the Nanodrop, and purity estimated through SDS-PAGE.

5

Activity assessment using protein substrate

TNF α R was mixed together with LS at a 2:1 ratio and incubated for 15-60 minutes at 37°C after which the proteins were separated on a 4-20% Novex gradient SDS-PAGE. The impact of NaCl (0-1 M), divalent cations, EDTA, and pH on LS activity was investigated, and differences in generated hydrolytic fragments measured through densitometric analysis using Gel Doc EZ (BioRad).

10

Time and dose dependency for activity

TNF α R (0.5 μ g) was incubated with varying doses of LS for 15 or 60 minutes at 37°C in PBS after which the proteins were separated on a 4-20% Novex gradient SDS-PAGE. The intensity of the generated fragments (densitometry) was used for determining optimal dose and time for efficient incubation conditions.

15

Substrate specificity

LS was incubated with a variety of N- and O-linked substrates overnight at 37°C at a ratio of 2:1 (substrate:enzyme). LS was incubated with EPO (0.3mg/ml) at a ratio of 50:1 (substrate:enzyme). The proteins were separated and analyzed on 4-20% Novex gradient SDS-PAGE gels.

20

25 Immobilization of LS_{E206A}

LS_{E206A} was resuspended into a coupling buffer (0.2 M NaHCO₃, 0.5 M NaCl pH 8.3), and concentrated to 20 mg/ml. NHS-activated sepharose 4 Fast Flow (GE Healthcare) was prepared for coupling according to manufacturer's instructions (e.g. HCl wash and equilibration in coupling buffer). LS_{E206A} was immobilized by overnight incubation with the sepharose at 4°C, slowly rocking for constant mixing. The sepharose was blocked by addition of 0.1 M Tris pH 8.5, washed with 3 repetitions of 0.1 M Tris pH 8.5 / 0.1 M NaAc, 0.5 M NaCl pH 5.0, and stored in EtOH until usage.

30

Binding affinity of LS_{E206A}

Spin columns with 50 µl immobilized LS_{E206A} (e.g. ca 50 µg protein) equilibrated in PBS were incubated with 10 µg glycoprotein pretreated with either a sialidase mix (Am0707:Am1757), or with a combination of sialidases and a *Streptococcus oralis* Endo-α-N-acetyl-galactosaminidase (e.g. an O-glycosidase). The samples were allowed to incubate for 2 h at 37°C after which the columns were washed with PBS (10 volumes; 100 g, 30 s) and eluted with 0.1 M glycine pH 3.0. Fractions were analyzed on SDS-PAGE.

10 Mass spectrometry analysis

Etanercept (Enbrel®) is clinically approved Fc-fusion protein that binds to TNFα. etanercept contains several O-glycans. In order to test enzymatic cleavage specificities the endoprotease was incubated together with etanercept overnight at 37°C. To simplify mass spectrometric analysis, a second round of enzymatic treatment was done to remove the remaining O-glycans using sialidase and O-glycosidase (overnight, in PBS, 1:40 ratio of all single enzymes). The generated peptides were analyzed by MS/MS after separating the peptides by C18 reversed phase liquid chromatography.

RESULTS

20 LS is a putative metalloprotease

Based on sequence and domains similarity, LS shares homology with several metalloproteases, containing the putative active site sequence GMAHELGHGL, sharing similarity to the general metalloprotease sequence abxHEbbHbc (a = V/T, b = no charge, c = hydrophobic). The histidines are generally involved in substrate binding and Zn²⁺ affinity, while the glutamic acid together with the histidines mediate the electron transfer, and thus the hydrolytic effect. To be able to further characterize the enzyme, we constructed an LS_{E206A} mutant, capable of binding the substrates, but lacking or having reduced hydrolytic capabilities by altering the E to an A. Further modifications (e.g. altering H to A) may be necessary for full inactivity. Both constructs expressed well, and were readily purified using affinity chromatography based on their His-tags (Fig. 1).

LS specifically hydrolyses glycoproteins with O-glycans

In order to investigate the substrate specificity of LS, the protease was incubated with a diversity of proteins. As shown in Fig 2, LS was incubated with IgA and Herceptin (trastuzumab). LS was only able to act upon proteins having O-linked glycans, such as IgA.

- 5 While the presence of terminal sialic acids seemingly partly inhibits the activity of LS, the absence of sialic acids is not a prerequisite for hydrolysis (Fig. 4).

LS can act upon O-linked glycoproteins under diverse conditions

Densitometric analyses of SDS-PAGE gels were carried out to evaluate the enzymatic properties of LS. LS is active under most conditions, with a preference for a slightly acidic pH and a low NaCl concentration (Fig. 3A-B). While both Mg^{2+} and Ca^{2+} ions positively affected the hydrolytic activity of LS, the presence of Zn^{2+} significantly lowered the activity, and EDTA completely abolished it (Fig. 3C-D).

15 O-linked galactosidase residues are critical for activity of LS

While having an increased activity in the absence of terminal sialic acids, the importance of the other carbohydrates in the O-glycans for the activity of LS was not fully understood. While the activity of LS is significantly increased in the absence of terminal sialic acids, the removal of galactoses completely inhibits the activity of LS (Fig. 4A).

- 20 Further, the lower activity of LS on sialylated proteins is not due to an inability to hydrolyze the bond in the presence of sialic acids, as demonstrated by the full hydrolysis after overnight incubation (Fig. 4B). The activity of LS fully relies on O-glycans, since removal of N-glycans did not affect the hydrolysis by LS (Fig. 4C).

25 O-linked glycans direct the cleavage site of LS

Having demonstrated that the O-glycan is critical for activity, we next sought to investigate the specific cleavage site of LS. Using mass spectrometry, we were able to demonstrate that LS hydrolyzes the amino bond between the O-glycosylated Ser/Thr and its N-terminal amino acid, regardless of its type (*e.g.* proline does not seem to inhibit the hydrolysis) (Fig. 5).

Using etanercept as a model protein due to its high abundance of O-linked glycans, the glycoprotein was treated with LS, after which it was subsequently treated with O-

glycosidases to facilitate the mass spec analysis. m/z values generated from the mass spec analysis, in combination with MS/MS data, were fitted to etanercept. All identified peptides had an N-terminal serine or threonine, consistent with LS cleaving just N-terminal of the O-glycans (Fig. 5). The analysis identified peptides both in a directed search
5 (defining S/T hydrolysis in the parameters; Fig. 5A), and in an unbiased approach (Fig. 5B).

A hydrolytic inactive variant of LS specifically bind to O-glycan containing proteins

With the ability of LS to bind to O-glycans and specifically hydrolyze the amino acid
10 bond next to the glycan (*e.g.* next to Ser/Thr), we hypothesized that an E₂₀₆A mutant of LS would lack hydrolytic activity, but retain binding ability. Such a tool would be valuable among others for a) identifying O-linked glycoproteins, b) affinity-purify O-linked glycopeptides for removal or for study, and c) affinity-purify O-glycans.

Fig. 6A shows that the mutant LS did not have any detectable hydrolytic activity.
15 While LS was able to hydrolyze etanercept in the presence of sialidase, the LS_{mut} could not hydrolyze etanercept, confirming that the genetic alteration indeed inactivated O-glycoprotease under the conditions tested.

LS_{E206A} was immobilised onto sepharose and added onto spin columns for easier handling. Importantly, the binding of LS_{E206A} to different substrates perfectly correlated
20 with the hydrolytic activity of LS (Fig. 6B). LS_{E206A} (labelled as LS_{mut}) demonstrated a specific affinity for O-linked glycoproteins. By immobilizing LS_{mut} on sepharose we were able to affinity purify IgA. However, we were not able to elute the protein, likely due to a strong affinity. Herceptin (trastuzumab), lacking O-glycans, as well as O-glycosidase treated IgA, did not bind to the column, but could be detected in the flowthrough (FT).

25

2-3 sialic bonds are important to remove for full LS activity

We recently determined that endoprotease activity relied on specific sialic acid bonds, necessitating the removal of both 2-3 and 2-6 linked sialic acids for full effect. To determine the individual role of specific sialic acid bonds for the activity of LS, we
30 incubated Enbrel with different sialidases in combination with LS for 30 min – 20 h. Removal of 2-3 bonds seemed sufficient for hydrolysis by LS (Fig. 7).

LS cleaves erythropoietin (EPO)

EPO was treated with PNGaseF, a sialidase (Smix, comprises Am0707 and Am1757) and/or an O-glycosidase and incubated with LS.

The resulting products were then analysed by SDS-PAGE and Coomassie blue staining, as well as RPLC and mass spectrometry. Results of SDS-PAGE are shown in Figure 9A, which shows that LS cleaves EPO both when sialic acids have been removed and when they are intact. Furthermore, LS also digests EPO where the N-glycans have been removed with PNGaseF, confirming that LS activity is not affected by N-glycan removal. However, LS did not cleave EPO when the O-glycans were removed with O-glycosidase, showing that O-glycan is needed for LS to cleave a protein. Equivalent results were observed at ratios of 10:1, 5:1 and 2:1 (substrate:enzyme) (data not shown).

The sample mixtures following incubation with PNGaseF, Smix and LS were separated by Reverse Phase Liquid Chromatography and analyzed by ESI mass spectrometry for identification of reaction products after enzymatic treatment.

Figure 9B shows an UV chromatogram from the RPLC. As expected, given that EPO has only one suggested O-glycan position (see predicted position in SEQ ID NO: 14 below), the chromatogram shows 2 peaks which correspond to the 2 fragments resulting from cleavage by LS.

These fragments were further analysed by MS (see Figures 9C and D) and identified as follows:

SAAPLRTITADTRKLFVRVYSNFLRGKCLKLYTGEACRTGD (Mass = 4900.5868Da – corresponds to the sequence C terminal to the cleavage point and thus includes the O-glycan still linked to the N terminal serine); and
APRLICDSRVLERYLLEAKEAEDITGCAEHCSLDENITVPDTKVDFYAWKRMEV
GQQAQVEVWQGLALLSEAVLRGQALLVNSSQPWEPLQLHVDKAVSGLRSLTTLR
ALGAQKEAISPPDAA (Mass = 13714.1199Da, corresponds to the sequence N terminal to the cleavage point).

Therefore the combined use of PNGaseF, sialidase mix and LS permitted the isolation and precise identification of the O-glycan-bearing serine in EPO. Methods of this type are applicable to any O-glycoprotein and permit the rapid identification of O-glycan attachment positions.

Example 2

Introduction

5

The LSE206A mutant described in Example 1 incorporates a site-directed mutation of the active site of LS (abxHEbbHbc to abxHAAbbHbc), removing the electron transfer capacity of the enzymatic cleft. As is explained further below, upon further stress testing it was found that although this change reduced O-glycoprotease activity relative to the wild-type sequence it did not completely eliminate it. Accordingly the inventors have developed and characterized another mutant incorporating an additional substitution in the enzymatic cleft. Specifically, a His residue important in the orientation of the co-factor zinc ion was replaced with an Ala. The resulting double-mutant is referred to as H205A/E206A (abxHEbbHbc to abxAAAbbHbc).

15

2.1 Production of the double-mutant

Site-directed mutagenesis using standard protocols (e.g. as in Example 1) was used to change both a histidine and a glutamic acid to alanine relative to the Amuc1119 gene of *Akkermansia muciniphila*, to create the double mutant, Amuc1119H205A/E206A (LS_{H205A/E206A}). The construct was transformed into *E. coli*, isolated and verified using sequencing as in Example 1. Expression in *E. coli* was conducted as described in Example 1. The sequence of the expressed protein is provided as SEQ ID NO: 21.

20

2.2 Characterization of the double-mutant

25

2.2.1 Double mutant fully inactivates the activity of LS

As shown in Example 1, the single mutant LSE206A was seen to be inactive given its inability to hydrolyze an O-glycoprotein in 2 hours. However, in a stress test it was found that O-glycoprotease activity was not completely abolished, but was rather only reduced in that some activity was observed at higher ratios of enzyme : O-glycoprotein and longer incubation times.

30

Incubation for 24 hours at a 1:1 (wt:wt) ratio for LSE206A : asialylated O-glycoprotein resulted in a significant hydrolysis of the substrate, though not to the same extent as wild type LS (Fig 10A). By contrast, the double-mutant LSH205A/E206A did not produce any evidence of hydrolysis even at 15:1 (wt:wt) ratios for the enzyme : O-glycoprotein after
5 overnight incubation (Fig 10B), suggesting that the enzyme was completely inactive with the addition of the second mutation.

2.2.2 Double mutant specifically binds O-glycoproteins

To assess binding to different proteins, immobilized LSH_{205A/E206A} (50 µl resin) (prepared
10 using the same protocol as Example 1) was equilibrated in PBS, after which 50 µg of different samples of proteins were added in a concentration of 0.5 mg/mL and incubated with end-over-end rotation for 2 hours at room temperature. The flowthrough was collected through centrifugation (200 g, 1 min) and the resin washed 3x with 350 µl PBS. Bound proteins were eluted by two sequential 5 minutes incubations with 50 µl 8 M Urea followed
15 by centrifugation (1000 g, 1 min). All samples were loaded in equal volumes. Starting / loading material, flowthrough, and eluate were assessed by SDS-PAGE.

In the first experiment (see Figure 11A), glycosylated or non-glycosylated proteins were pretreated with either a sialidase mix (Am0707:Am1757), or with a combination of the
20 sialidase mix and a *Streptococcus oralis* Endo- α -N-acetyl-galactosaminidase (e.g. an O-glycosidase) before being incubated with the resin, washed, and eluted. Pretreatment of the samples (Sialidase mixture +/- O-glycosidase) was done in accordance with the manufacturer's instructions. Only proteins with O-glycans bound to the resin, with an increased affinity in the absence of sialic acids. The presence of O-glycans was imperative
25 for any binding to take place, as shown by the lack of interaction after treatment with O-glycosidase.

In the second experiment (see Figure 11B), a mix of N-glycosylated, O-glycosylated and non-glycosylated proteins was incubated with LS double mutant resin. Only O-
30 glycosylated proteins (TNF α R and ApoE) were bound to the matrix and eluted with 8 M urea. N-glycosylated (afibercept, AGP (alpha-1-acid glycoprotein), Fc domain of IgG

(IgG Fc) and non-glycosylated (BSA) did not bind to the LS double mutant resin and were found in the flow through. Thus the double mutant resin specifically binds to only O-glycosylated proteins when a sample contains a mixture of N-, O- and non-glycosylated proteins.

5

In the third experiment (see Figure 11C), a mix of N-glycosylated and non-glycosylated proteins was incubated with LS double mutant resin. There was no non-specific binding even in the absence of possible competition from O-glycoproteins (none are present). No proteins were found in the eluate. Thus the double mutant resin does not bind to proteins lacking O-glycans.

10

2.2.3 Double mutant can be immobilized on resin at different concentrations to enhance capacity

To investigate the ability to improve the capacity of the immobilized double-mutant resin to bind more O-glycosylated proteins, different concentrations of double mutant (5-15 mg/mL) were used during immobilisation on resin. A representative gel is shown in Figure 12A. The % shown is the level of binding relative to positive control and was determined by densitometry analysis of the gel. The results are shown in the graph in Figure 12B. A dose-dependent increase in capacity with higher O-glycoprotein binding capacity was seen when using a higher concentration of double mutant during immobilization. Further experiments continued using 15 mg/mL of immobilized double mutant. Further, a high degree of O-glycoprotein binding was maintained even in the presence of 1 M Urea and 1 M GHCl, even though the latter significantly reduced binding efficiency.

15

20

2.2.4 The affinity purification ability of the double mutant is ~3 mg glycoprotein/mL resin

In order to specifically investigate the ability of the double-mutant resin to affinity purify O-glycoproteins, as well as the impact of sample concentration on this ability, different quantities and concentrations of asialylated etanercept were added to the resin. An individual column (containing 50µl of double mutant resin) had a capacity of binding around 150 µg O-glycoprotein, i.e. 3 mg O-glycoprotein / mL resin). Figure 13 shows a representative gel.

30

2.2.5 The binding of O-glycoproteins to double-mutant is not significantly affected by ionic strength or buffer volume/type, and works over a broad pH range

5 Sample proteins were allowed to bind to double-mutant resin for 2 hours at room temperature with end-over-end rotation under a range of different conditions to test the effect of ionic strength, buffer volume/type and pH on the binding ability of the resin. In each case, the resin was then washed three times with its respective binding buffer (350 µl) and then eluted with the addition of 8 M Urea (50 µl, 5 min incubation; 2 repeats). All
10 samples were then analysed by SDS-PAGE.

In a first experiment (see Figure 14A), to investigate the stability of the interaction in buffers with varying ionic strength, the sample consisted of asialylated etanercept, which was incubated with double-mutant-resin in 0-4 M NaCl, as well as performing all washing steps
15 with the respective concentration of NaCl. Addition of NaCl did not significantly affect the binding of asialylated etanercept.

In a second experiment (see Figure 14B), the sample consisted of asialylated etanercept in a range of different volumes of PBS. Wash steps used PBS. Varying the substrate volume
20 between 100-300 µl did not affect the efficiency significantly.

In a third experiment (see Figures 14C and D), the sample consisted of asialylated etanercept and BSA in different buffers (100 mM sodium acetate, 50 mM sodium phosphate and 50 mM Tris) at different pHs (pH 4-9). Wash steps used matching buffers. pH 6-8 was found
25 to work best whereas pH 4 did not work at all and pH 9 slightly less efficiently than pH 8. BSA, which does not contain any O-glycans, did not bind to the resin under any of the binding conditions.

2.2.6 Urea and SDS can elute affinity-bound O-glycoproteins

30 Based on the high affinity between the double mutant and its O-glycoprotein substrate, the inventors investigated different means to elute the bound protein from the resin, not based on ionic strength. Urea had a dose dependent elution, with close to 100% elution with the

usage of 8 M Urea (Fig. 15A). High concentrations of SDS (eg 5-10%) also eluted most of the bound protein (Fig. 15B). However, since many downstream applications are sensitive to the presence of detergents, use of high levels Urea is likely to be of more practical utility for non-enzymatic release of bound proteins/peptides.

5

2.2.7 Wildtype LS can be used to elute double-mutant-bound O-glycoproteins

The inventors speculated that the addition of LS to double-mutant-bound proteins may result in a release thereof, and thus not necessitate the addition of urea for elution. Both abatacept and etanercept could be hydrolyzed and eluted from the double-mutant-resin by LS in 6 h but had a slightly more complete elution after 24 h (Fig.16A). Addition of urea afterwards showed that very little O-glycoprotein remained attached to the affinity matrix, demonstrating that the LS elution strategy was highly efficient.

Etanercept eluted with LS was also subjected to mass spec analysis (LC/MS and MS/MS). Identified peptides (Fig 16B.1) were consistent with those generated in a LS digestion of etanercept (Fig 16B.2). Additional MS data from this experiment is shown in the following table:

Row	OK	Cmpd.	m/z meas.	Mr calc.	z	Δ m/z [ppm]	RMS90 [ppm]
1	TRUE	1901	392,18578	782,3545447	2	3,139355727	39,5628229
2	TRUE	1240	460,24964	459,2441464	1	-3,873732048	6,970140396
3	TRUE	2395	593,93513	1778,79071	3	-4,012614897	37,85883109
4	TRUE	1019	598,28693	2986,418959	5	-6,916951725	23,18052633
5	TRUE	1388	642,65448	1924,956428	3	-7,685331295	12,36799202
6	TRUE	2169	672,31437	1342,621401	2	-5,364932116	7,267732308
7	TRUE	2169	672,31437	1342,621401	2	-5,364932116	7,267732308
8	TRUE	1666	695,85597	1389,70747	2	-7,245075789	13,08730439
9	TRUE	1887	745,4213	744,4170254	1	-4,027105599	40,39538537
10	TRUE	2308	846,8813	1691,758682	2	-6,278826744	10,77131727
11	TRUE	1636	878,42107	1754,839667	2	-6,875687278	9,154478723
12	TRUE	2298	895,05953	2682,161633	3	-1,814552481	8,770132859
13	TRUE	2048	898,39499	1794,785625	2	-5,675814522	10,69492018
14	TRUE	2319	924,06736	2769,193661	3	-4,837599076	9,79089835
15	TRUE	2299	1072,96596	2143,922906	2	-2,58127791	10,42173435
16	TRUE	1315	1189,55981	2377,120652	2	-6,550738429	9,551870744
17	TRUE	1325	1189,55981	2377,120652	2	-6,550738429	9,851027886
18	TRUE	2377	1219,18407	3654,552682	3	-6,097374291	9,225746116
Row	Int.	Scores					Range
1	100030	20.3 (D.metascore:20.3,D.btScore:4.0,D.fragCov:50.0,D.intCov:8.2)					457 - 459
2	29654	23.0 (D.metascore:23.0,D.btScore:0.0,D.fragCov:12.5,D.intCov:42.3)					182 - 185
3	30476	31.2 (D.metascore:31.2,D.btScore:127.0,D.fragCov:40.0,D.intCov:24.3)					216 - 225
4	219206	26.5 (D.metascore:26.5,D.btScore:265.0,D.fragCov:33.3,D.intCov:21.1)					184 - 204

5	655550	23.9 (D.metascor:23.9,D.btScore:136.0,D.fragCov:46.7,D.intCov:12.3)		184 - 198	
6	374698	36.2 (D.metascor:36.2,D.btScore:61.0,D.fragCov:33.3,D.intCov:39.2)		217 - 225	
7	374698	36.2 (D.metascor:36.2,D.btScore:61.0,D.fragCov:33.3,D.intCov:39.2)		217 - 225	
8	39488	26.9 (D.metascor:26.9,D.btScore:12.0,D.fragCov:35.7,D.intCov:20.3)		186 - 199	
9	18900	66.8 (D.metascor:66.8,D.btScore:191.0,D.fragCov:92.9,D.intCov:48.1)		1 - 7	
10	108010	22.4 (D.metascor:22.4,D.btScore:13.0,D.fragCov:22.2,D.intCov:22.5)		217 - 225	
11	275854	29.6 (D.metascor:29.6,D.btScore:260.0,D.fragCov:39.3,D.intCov:22.3)		186 - 199	
12	30586	24.2 (D.metascor:24.2,D.btScore:509.0,D.fragCov:28.1,D.intCov:20.8)		216 - 231	
13	48978	35.0 (D.metascor:35.0,D.btScore:125.0,D.fragCov:35.0,D.intCov:35.0)		216 - 225	
14	38966	23.7 (D.metascor:23.7,D.btScore:253.0,D.fragCov:23.5,D.intCov:23.8)		216 - 232	
15	29176	20.1 (D.metascor:20.1,D.btScore:29.0,D.fragCov:25.0,D.intCov:16.2)		216 - 225	
16	2495538	23.1 (D.metascor:23.1,D.btScore:535.0,D.fragCov:59.4,D.intCov:9.0)		184 - 199	
17	182236	22.9 (D.metascor:22.9,D.btScore:4096.0,D.fragCov:43.8,D.intCov:12.0)		184 - 199	
18	360318	21.1 (D.metascor:21.1,D.btScore:33.0,D.fragCov:22.2,D.intCov:20.0)		208 - 225	
Row	#Cmpds.	P	Sequence	Modifications	Protein
1	1	0	Y.TQK.S	Acetyl: 1; Hex(1)HexNAc(1): 1	Etanercept
2	1	1	T.SPTR.S		Etanercept
3	5	2	P.STSFLLPMGP.S	Hex(1)HexNAc(1): 1, 2	Etanercept
4	13	4	P.TRSMAPGAVHLPQPVSTRSQH.T	Hex(1)HexNAc(1): 16, 17	Etanercept
5	6	1	P.TRSMAPGAVHLPQPV.S	Hex(1)HexNAc(1): 3	Etanercept
6	4	1	S.TSFLLPMGP.S	Oxidation: 7; Hex(1)HexNAc(1): 2	Etanercept
7	4	1	S.TSFLLPMGP.S	Oxidation: 7; Hex(1)HexNAc(1): 1	Etanercept
8	2	1	R.SMAPGAVHLPQPV.S.T		Etanercept
9	5	0	-.LPAQVAF.T		Etanercept
10	2	1	S.TSFLLPMGP.S	Hex(1)HexNAc(1): 1, 2	Etanercept
11	5	1	R.SMAPGAVHLPQPV.S.T	Hex(1)HexNAc(1): 14	Etanercept
12	2	3	P.STSFLLPMGSPSPAEG.S	Hex(1)HexNAc(1): 1, 2, 3	Etanercept
13	10	2	P.STSFLLPMGP.S	Oxidation: 8; Hex(1)HexNAc(1): 1, 2	Etanercept
14	3	4	P.STSFLLPMGSPSPAEGS.T	Hex(1)HexNAc(1): 1, 2, 3	Etanercept
15	5	2	P.STSFLLPMGP.S	Hex(1)HexNAc(1): 1, 2, 3	Etanercept
16	19	2	P.TRSMAPGAVHLPQPV.S.T	Hex(1)HexNAc(1): 3, 16	Etanercept
17	19	2	P.TRSMAPGAVHLPQPV.S.T	Hex(1)HexNAc(1): 1, 3	Etanercept
18	8	5	P.TPEPSTAPSTSFLLPMGP.S	Hex(1)HexNAc(1): 5, 6, 9, 10, 11	Etanercept

2.2.8 Double mutant can be used to affinity-purify O-glycoproteins from complex samples

- As a proof of concept that the system can function as a general affinity matrix for O-glycosylated proteins, not only in simplified systems but in complex media, the inventors investigated the ability of the double-mutant to purify O-glycoproteins from human serum. Human serum mainly consists of non-glycosylated (BSA) and N-glycosylated (IgG) proteins, with only a small fraction of the total serum proteome being O-glycosylated.
- Applying 20 µl sialidase treated serum (ca 1.2 mg protein) to a 50 µl immobilized double-mutant resin column allowed the removal of almost all of the non-glycosylated and N-glycosylated proteins, while eluting a few selected proteins (Fig. 17A). By adding higher

quantities of serum (e.g. 2.5 mg protein) with or without pre-treatment of sialidases and O-glycosidases it was demonstrated that the interaction is dependent on O-glycans and removal of terminal sialic acids (Fig. 17B). Further, it was concluded that pre-treatment with sialidases significantly increased the amount of bound O-glycoproteins as compared to non-sialidase-treated samples. Addition of 50 U sialidase mix (Am0707:Am1757) was sufficient to improve the amount of affinity purified O-glycoproteins (Fig. 17C).

By analysis in mass spectrometry the vast majority of the affinity purified serum proteins can be annotated as O-glycosylated proteins (see Fig. 18A, and the names in ***bold-italics*** in the following table). The number of identified O-glycoprotein peptides in relation to non-O-glycoprotein peptides could be affected by different stringencies in the washing steps, both in terms of total number of identified peptides (Fig. 18B) as well as on the ratio O-glycoprotein peptides vs non-O-glycoprotein peptides (Fig. 18C). Thus it is clear that the affinity resin is highly efficient in its ability to specifically and selectively affinity purify and enrich O-glycoproteins. Additional MS data from this experiment is shown in the following table.

Row	OK	Protein	MW [kDa]
1	TRUE	<i>Ig alpha-1 chain C region</i> OS=Homo sapiens GN=IGHA1 PE=1 SV=2	37,6
2	TRUE	<i>Apolipoprotein B-100</i> OS=Homo sapiens GN=APOB PE=1 SV=2	515,3
3	TRUE	<i>Kininogen-1</i> OS=Homo sapiens GN=KNG1 PE=1 SV=2	71,9
4	TRUE	<i>Complement C4-B</i> OS=Homo sapiens GN=C4B PE=1 SV=2	192,6
5	TRUE	<i>Complement C4-A</i> OS=Homo sapiens GN=C4A PE=1 SV=2	192,7
6	TRUE	Complement C3 OS=Homo sapiens GN=C3 PE=1 SV=2	187
7	TRUE	<i>Inter-alpha-trypsin inhibitor heavy chain H4</i> OS=Homo sapiens GN=ITIH4 PE=1 SV=4	103,3
8	TRUE	<i>Inter-alpha-trypsin inhibitor heavy chain H2</i> OS=Homo sapiens GN=ITIH2 PE=1 SV=2	106,4
9	TRUE	<i>Plasminogen</i> OS=Homo sapiens GN=PLG PE=1 SV=2	90,5
10	TRUE	<i>Inter-alpha-trypsin inhibitor heavy chain H1</i> OS=Homo sapiens GN=ITIH1 PE=1 SV=3	101,3
11	TRUE	Serum albumin OS=Homo sapiens GN=ALB PE=1 SV=2	69,3
12	TRUE	Alpha-2-macroglobulin OS=Homo sapiens GN=A2M PE=1 SV=3	163,2
13	TRUE	<i>Fibronectin</i> OS=Homo sapiens GN=FN1 PE=1 SV=4	262,5
14	TRUE	Complement C1s subcomponent OS=Homo sapiens GN=C1S PE=1 SV=1	76,6
15	TRUE	<i>N-acetylmuramoyl-L-alanine amidase</i> OS=Homo sapiens GN=PGLYRP2 PE=1 SV=1	62,2
16	TRUE	<i>Plasma protease C1 inhibitor</i> OS=Homo sapiens GN=SERPING1 PE=1 SV=2	55,1
17	TRUE	<i>Prothrombin</i> OS=Homo sapiens GN=F2 PE=1 SV=2	70
18	TRUE	C4b-binding protein alpha chain OS=Homo sapiens GN=C4BPA PE=1 SV=2	67
19	TRUE	<i>Coagulation factor XII</i> OS=Homo sapiens GN=F12 PE=1 SV=3	67,7
20	TRUE	Complement C1r subcomponent OS=Homo sapiens GN=C1R PE=1 SV=2	80,1
21	TRUE	<i>Ig delta chain C region</i> OS=Homo sapiens GN=IGHD PE=1 SV=2	42,2
22	TRUE	<i>Apolipoprotein F</i> OS=Homo sapiens GN=APOF PE=1 SV=2	35,4
23	TRUE	<i>Complement component C7</i> OS=Homo sapiens GN=C7 PE=1 SV=2	93,5
24	TRUE	Ig mu chain C region OS=Homo sapiens GN=IGHM PE=1 SV=3	49,3

25	TRUE	Inter-alpha-trypsin inhibitor heavy chain H3 OS=Homo sapiens GN=ITIH3 PE=1 SV=2	99,8
26	TRUE	<i>Histidine-rich glycoprotein</i> OS=Homo sapiens GN=HRG PE=1 SV=1	59,5
27	TRUE	Complement component C9 OS=Homo sapiens GN=C9 PE=1 SV=2	63,1
28	TRUE	<i>Alpha-2-HS-glycoprotein</i> OS=Homo sapiens GN=AHSG PE=1 SV=1	39,3
29	TRUE	<i>Protein AMBP</i> OS=Homo sapiens GN=AMBP PE=1 SV=1	39
30	TRUE	<i>Ig gamma-1 chain C region</i> OS=Homo sapiens GN=IGHG1 PE=1 SV=1	36,1
31	TRUE	Alpha-1-antitrypsin OS=Homo sapiens GN=SERPINA1 PE=1 SV=3	46,7
32	TRUE	<i>Plasma kallikrein</i> OS=Homo sapiens GN=KLKB1 PE=1 SV=1	71,3
33	TRUE	<i>Vitronectin</i> OS=Homo sapiens GN=VTN PE=1 SV=1	54,3
34	TRUE	<i>Apolipoprotein A-I</i> OS=Homo sapiens GN=APOA1 PE=1 SV=1	30,8
35	TRUE	Vitamin K-dependent protein S OS=Homo sapiens GN=PROS1 PE=1 SV=1	75,1
36	TRUE	<i>Serotransferrin</i> OS=Homo sapiens GN=TF PE=1 SV=3	77
37	TRUE	<i>Apolipoprotein E</i> OS=Homo sapiens GN=APOE PE=1 SV=1	36,1
38	TRUE	<i>von Willebrand factor</i> OS=Homo sapiens GN=VWF PE=1 SV=4	309,1
39	TRUE	<i>Proteoglycan 4</i> OS=Homo sapiens GN=PRG4 PE=1 SV=2	151

Row	pI	#Peptides	SC [%]	Scores	RMS90 [ppm]
1	6,1	149	75,1	1239.4 (M.expect:0.0,M.score:1239.4,M.siglimit:38.0)	5,46
2	6,6	137	41,2	7043.5 (M.expect:0.0,M.score:7043.5,M.siglimit:38.0)	6,02
3	6,3	86	45,7	1638.5 (M.expect:0.0,M.score:1638.5,M.siglimit:38.0)	5,49
4	6,9	78	63,3	4951.3 (M.expect:0.0,M.score:4951.3,M.siglimit:38.0)	6,46
5	6,7	76	61,3	4468.7 (M.expect:0.0,M.score:4468.7,M.siglimit:38.0)	6,12
6	6	75	64,1	4596.8 (M.expect:0.0,M.score:4596.8,M.siglimit:38.0)	6,13
7	6,5	52	56,9	2389.7 (M.expect:0.0,M.score:2389.7,M.siglimit:38.0)	6,29
8	6,4	49	50,4	2175.2 (M.expect:0.0,M.score:2175.2,M.siglimit:38.0)	7,02
9	7	45	67,3	2752.5 (M.expect:0.0,M.score:2752.5,M.siglimit:38.0)	5,54
10	6,3	39	43,8	1838.3 (M.expect:0.0,M.score:1838.3,M.siglimit:38.0)	5,16
11	5,9	37	63,7	2350.1 (M.expect:0.0,M.score:2350.1,M.siglimit:38.0)	5,59
12	6	36	38,1	1970.3 (M.expect:0.0,M.score:1970.3,M.siglimit:38.0)	5,67
13	5,5	33	25,3	1632.8 (M.expect:0.0,M.score:1632.8,M.siglimit:38.0)	5,58
14	4,8	23	46,1	1454.6 (M.expect:0.0,M.score:1454.6,M.siglimit:38.0)	6,01
15	7,3	22	56,4	1288.8 (M.expect:0.0,M.score:1288.8,M.siglimit:38.0)	5,72
16	6,1	22	46,2	1166.8 (M.expect:0.0,M.score:1166.8,M.siglimit:38.0)	6,15
17	5,6	20	43,4	1158.1 (M.expect:0.0,M.score:1158.1,M.siglimit:38.0)	6,13
18	7,2	20	51,4	1111.9 (M.expect:0.0,M.score:1111.9,M.siglimit:38.0)	5,13
19	8	20	39,7	691.9 (M.expect:0.0,M.score:691.9,M.siglimit:38.0)	5,03
20	5,8	19	39,6	1097.7 (M.expect:0.0,M.score:1097.7,M.siglimit:38.0)	6,07
21	8,1	19	46,9	687.5 (M.expect:0.0,M.score:687.5,M.siglimit:38.0)	4,87
22	5,4	19	43,3	344.0 (M.expect:0.0,M.score:344.0,M.siglimit:38.0)	7,71
23	6,1	18	38,3	1100.7 (M.expect:0.0,M.score:1100.7,M.siglimit:38.0)	5,73
24	6,3	17	50,9	1159.0 (M.expect:0.0,M.score:1159.0,M.siglimit:38.0)	5,25
25	5,5	16	28,3	747.2 (M.expect:0.0,M.score:747.2,M.siglimit:38.0)	5,8
26	7,1	15	35,4	964.5 (M.expect:0.0,M.score:964.5,M.siglimit:38.0)	5,33
27	5,4	15	34,3	876.5 (M.expect:0.0,M.score:876.5,M.siglimit:38.0)	5,86
28	5,4	14	46	941.9 (M.expect:0.0,M.score:941.9,M.siglimit:38.0)	5,78
29	5,9	14	55,4	886.3 (M.expect:0.0,M.score:886.3,M.siglimit:38.0)	6,06
30	8,5	14	60,3	825.9 (M.expect:0.0,M.score:825.9,M.siglimit:38.0)	6,04
31	5,4	14	39	740.9 (M.expect:0.0,M.score:740.9,M.siglimit:38.0)	6,02
32	8,6	14	26	647.7 (M.expect:0.0,M.score:647.7,M.siglimit:38.0)	5,62
33	5,6	13	37	752.2 (M.expect:0.0,M.score:752.2,M.siglimit:38.0)	5,1
34	5,6	13	46,8	723.0 (M.expect:0.0,M.score:723.0,M.siglimit:38.0)	5,57
35	5,5	13	27,7	676.0 (M.expect:0.0,M.score:676.0,M.siglimit:38.0)	6,18
36	6,8	13	22,9	544.9 (M.expect:0.0,M.score:544.9,M.siglimit:38.0)	6,02
37	5,6	13	45,1	535.3 (M.expect:0.0,M.score:535.3,M.siglimit:38.0)	5,73
38	5,3	13	6,3	505.5 (M.expect:0.0,M.score:505.5,M.siglimit:38.0)	5,52

39	9,5	13	5,6	214.2 (M.expect:0.0,M.score:214.2,M.siglimit:38.0)	4,7
----	-----	----	-----	--	-----

2.2.9 Immobilized double-mutant also binds to shorter O-glycopeptides

A series of experiments was performed to demonstrate the specificity of the LS double mutant also for O-glycopeptides. In the first experiment, a mix of an O-glycosylated peptide (glycodrosocin (GD) = GKPRPYSPRPTSHPRPIRV (SEQ ID NO: 47) with a core 1 O-glycan on the threonine) and several non-glycosylated peptides (H2686, H4062 H8390 and insulin oxidized beta chain (IOB)) was incubated with LS double mutant resin. (H2686 = YIYGSFK (SEQ ID NO: 48), H4062 = KKLVFFA (SEQ ID NO: 49), H8390 = FLPLILGKLVKGLL (SEQ ID NO: 50)).

The peptide mix was allowed to bind to 50 µl immobilized double-mutant resin for 2 hours at room temperature with end-over-end rotation. The resin was washed five times with binding buffer (300 µl) and then eluted with the addition of 8 M Urea. The peptides in load, flow through and eluate were analyzed with LC/MS. Separation was performed on a RP-LC C18 column (Advance BioPeptide Map 2.1x100 2.7µm from Agilent) and detected with ESI-Q-TOF Bruker Impact II. The results are shown in Figure 19A. Glycodrosocin, the only peptide in the mix containing an O-GalNAcGal, was predominantly found in the eluted fraction and the non-glycosylated peptides in the flow-through fraction

In the second experiment, it was investigated whether the LS double-mutant could enrich O-glycosylated peptides from a tryptic protein digest (e.g a different type of peptide mix). IgA was chosen as the target for digest. Based on the trypsin sites and reported O-glycosylation sites in IgA, a trypsin digest should result in only a single O-glycosylated peptide corresponding to positions 89-126 of IgA (see schematic diagram in Figure 19B). To create the tryptic digest, IgA was mixed with urea to 6M and DTT to 5 mM followed by incubation at 37°C for 1h. IAM was added to 15 mM followed by incubation at room temperature for 30 min in the dark. The sample was then buffer exchanged to 50 mM Tris, pH 8.0 on Zeba spin 7000 K column. Trypsin was then added at 1:20 followed by incubation at 37°C overnight. Trypsin inhibitor 1 mg/mg was added followed by incubation at room temperature for 20 min. Sialidase mix and NaCl were added to the resulting tryptic digest. The mix was allowed to bind to the resin for 2h at room temperature with

end-over-end rotation. The resin was washed ten times with PBS buffer (300 µl) and then eluted with the addition of 8 M Urea (50 µl, 2 min, 2 repeats).

Peptides of the load, flowthrough and eluate were separated and analysed using RP-LC MSMS on a C18 column (Advance BioPeptide Plus 2.1x150mm 2.7µm from Agilent Technologies) in a 0.1%FA in MQ: 0.1%FA in 95% ACN gradient at 45°C and a flow of 0.2ml/min. Detection was on an ESI-Q-TOF Bruker Impact II instrument. The results are shown in Figure 19C. O-glycosylated peptide 89-126 were significantly enriched in the eluate and the specific O-glycopeptide 89-126 was identified with intact mass.

10

2.2.10 Immobilized double-mutant compares favorably to other O-glycoprotein binding matrixes

The inventors evaluated the ability of the double-mutant to affinity purify O-glycoprotein as compared to other commercially available O-glycoprotein binding matrices, specifically the lectins Peanut agglutinin (PNA), and Vicia villosa lectin (VVA). Etanercept and asialylated etanercept were used as model substrates.

50 µg substrate was added in PBS (PNA and LS double mutant) or lectin binding buffer (VVA) to 50 µl volume of different immobilized lectin or LS double-mutant resins pre-equilibrated in the respective buffers (total 100 µl). (Lectin binding buffer is 20 mM Tris-HCl pH 7.4, 150 mM NaCl, 1 mM each of MgCl₂, CaCl₂, ZnCl₂, and MnCl₂). The substrate was allowed to interact with the resins with end-over-end-mixture for 2 h at room temperature. Non-bound proteins were washed away (100 g, 1 min; 3 x) with PBS or lectin binding buffer, respectively. The resins were dried by centrifugation (1000 g, 1 min). Bound proteins were eluted with the addition of 8 M urea (for PNA and LS double mutant resins) or VVA elution buffer according to manufacturer (for VVA resin) (50 µl, 5 min treatment before centrifugation 1000 g 1 min; 2 x) and both flow through (FT) and eluate (E) were analyzed on SDS-PAGE. 1.5 µg substrate was added to each gel (e.g. 3 µl) as a positive control and a densitometric analysis was performed to evaluate the efficiency of the resins, relative to the 1.5 µg loaded substrate which has assumed 100% efficiency. Representative gels for etanercept and asialylated etanercept (Etanercept S) are shown in Figure 20A. The

20
25
30

results of the densitometric analysis are shown in Figure 20B. LSH205A/E206A double mutant performs at least as well as the best performing commercial lectin for efficiency of purification of asialylated substrate.

5

Sequences

SEQ ID NO: 1

EVTVPDALKDRIALKKTARQLNIVYFLGSDTEPVPDYERRLSELLLYLQQFYGKEMQRHGYGARSFGLDIKSP
GRVNIIEYKAKNPAAHYPYENGSGWKAQELDEFFKAHPDRKKSQHTLIIMPTWNDEKNGPDNPGGVPFYGMG
10 RNCFALDYPAFDIKHLGQKTREGRLLTWKYGGMAHELGHGLNLPNHQTASDGKKYGTALMGSGNYTFGTSP
FLTPASCALLDACEVFSVTPSQQFYEGKPEVEVGDAISFKGDQILVSGNYKSPQTVKALNVYIQDPPYAVNQ
DYDAVSFSRRLGKKSGKFSMKIDKKELEGLNNNEFRISLMFILANGLHMQKHFTFHWDALQDYRDGSKS

SEQ ID NO: 2

MEVTVPDALKDRIALKKTARQLNIVYFLGSDTEPVPDYERRLSELLLYLQQFYGKEMQRHGYGARSFGLDIKSP
PGRVNIIEYKAKNPAAHYPYENGSGWKAQELDEFFKAHPDRKKSQHTLIIMPTWNDEKNGPDNPGGVPFYGMG
15 GRNCFALDYPAFDIKHLGQKTREGRLLTWKYGGMAHELGHGLNLPNHQTASDGKKYGTALMGSGNYTFGTSP
TFLTPASCALLDACEVFSVTPSQQFYEGKPEVEVGDAISFKGDQILVSGNYKSPQTVKALNVYIQDPPYAVNQ
QDYDAVSFSRRLGKKSGKFSMKIDKKELEGLNNNEFRISLMFILANGLHMQKHFTFHWDALQDYRDGSKSGSG
20 HHHHHH

SEQ ID NO: 3

ATGGAAGTCACGTGCGCGACGCCCTGAAAGATCGCATCGCGCTGAAGAAAACCGCTCGTCAGCTGAATATCG
TCTACTTCCTGGGTTCTGATACCGAACCGGTTCCGGACTACGAGCGCCGTCTGAGCGAGCTGCTGTTGTATCT
25 GCAGCAATTCTATGGTAAAGAAATGCAGCGCCATGGCTATGGCGCACGCAGCTTTGGTCTGGACATTAAGTCA
CCGGGTCTGTGAACATTATCGAGTACAAAGCGAAGAACCCGGCAGCGCATTACCCGTATGAGAATGGTGGCG
GCTGGAAAGCTGCACAAGAACTGGACGAATTTTTCAAGGCCCATCCAGACCGCAAGAAAAGCCAGCACACCCT
GATCATCATGCCTACCTGGAATGATGAGAAAAATGGTCTGACAATCCGGGTGGCGTTCCGTTCTATGGTATG
GGTCGTAATTGTTTGGCTTGGACTACCCGGCGTTTGATATCAAGCACCTGGGTGAGGTCAGAAAACGCGTGAGGGT
30 GTCTGCTGACGAAATGGTACGGTGGCATGGCGCACGAACTGGGCCACGGCCTGAATCTGCCGCACAATCACCA
GACCGCGAGCGATGGCAAGAAATATGGCACCGCCCTGATGGGTAGCGGCAACTACACGTTCCGGTACCAGCCCG
ACGTTCCCTGACCCCGCGAGCTGTGCGTCTGATGCTGCAAGTGTTCAGCGTTACCCCGAGCCAACAGT
TTTATGAGGGTAAGCCAGAGTCGAGGTGGTGTATGTTCAATTTCTTCAAGGGTGATCAAATCTTGGTCAG
CGGTAACATAAGAGCCCGCAAACCGTGAAAGCTCTGAACGTTTACATTACAGATCCGCCGTACGCCGTGAAC
35 CAAGACTACGATGCAGTGAGCTTTAGCCGTCTGCTGGGCAAAAAGTCCGGTAAGTTTAGCATGAAGATTGACA
AAAAAGAACTGGAAGGCCTGAATAACAACGAATTCCTGATTTCTTGGTATGTTTATTCGGCAACCGCTTACA
CATGCAGAAGCACTTTACGTTTCACTGGGATGCGCTGCAAGACTACCGTGACGGTAGCAAATCTGGTTCGGGT
CATCATCACCACCATCACTGA

40 SEQ ID NO: 4

MLKRLLSAFFSLFFLGAASGTSFAEVTVPDALKDRIALKKTARQLNIVYFLGSDTEPVPDYERRLSELLLYLQ
QFYGKEMQRHGYGARSFGLDIKSPGRVNIIEYKAKNPAAHYPYENGSGWKAQELDEFFKAHPDRKKSQHTLI
IMPTWNDEKNGPDNPGGVPFYGMGRNCFALDYPAFDIKHLGQKTREGRLLTWKYGGMAHELGHGLNLPNHQT
ASDGKKYGTALMGSGNYTFGTSPFLTPASCALLDACEVFSVTPSQQFYEGKPEVEVGDAISFKGDQILVSG
45 NYKSPQTVKALNVYIQDPPYAVNQDYDAVSFSRRLGKKSGKFSMKIDKKELEGLNNNEFRISLMFILANGLHM
QKHFTFHWDALQDYRDGSKS (signal sequence underlined)

SEQ ID NO: 5

EVTVPDALKDRIALKKTARQLNIVYFLGSDTEPVPDYERRLSELLLYLQQFYGKEMQRHGYGARSFGLDIKSP
GRVNIIEYKAKNPAAHYPYENGSGWKAQELDEFFKAHPDRKKSQHTLIIMPTWNDEKNGPDNPGGVPFYGMG
50 RNCFALDYPAFDIKHLGQKTREGRLLTWKYGGMAHALGHGLNLPNHQTASDGKKYGTALMGSGNYTFGTSP
FLTPASCALLDACEVFSVTPSQQFYEGKPEVEVGDAISFKGDQILVSGNYKSPQTVKALNVYIQDPPYAVNQ
DYDAVSFSRRLGKKSGKFSMKIDKKELEGLNNNEFRISLMFILANGLHMQKHFTFHWDALQDYRDGSKS

- SEQ ID NO: 6
 MEVTVPDALKDRIALKKTARQLNIVYFLGSDTEPVPDYERRLSELLLYLQQFYGKEMQRHGYGARSFGLDIKS
 PGRVNIIEYKAKNPAAHYPYENGGGWKAAQELDEFFKAHPDRKKSQHTLIIMPTWNDEKNPDPNPGGVFFYGM
 5 GRNCFALDYPAFDIKHLGQKTREGRLLTKWYGGMAHALGHGLNLPNHNQTASDGKKYGTALMGSGNYTFTGTSP
 TFLTASPASCALLDACEVFSVTPSQQFYEGKPEVEVGDAISFKGDQILVSGNYKSPQTVKALNVYIQDPPYAVN
 QDYDAVSFSRRLGKKSFKSMKIDKKEGLENNNEFRISLMFILANGLHMQKHFTFHWDALQDYRDGSKSGSG
 HHHHHH
- 10 SEQ ID NO: 7
 ATGGAAGTCACGTGTGCCGACGCCCTGAAAGATCGCATCGCGCTGAAGAAAACCGCTCGTCAGCTGAATATCG
 TCTACTTCCCTGGGTTCTGATACCGAACCGGTTCCGGACTACGAGCGCCGTCTGAGCGAGCTGCTGTTGTATCT
 GCAGCAATTCATGGTAAAGAAATGCAGCGCCATGGCTATGGCGCACGCAGCTTTGGTCTGGACATTAAGTCA
 CCGGGTCGTGTGAACATTATCGAGTACAAAGCGAAGAACCCGGCAGCGCATTACCCGTATGAGAATGGTGGCG
 15 GCTGGAAAGCTGCACAAGAACTGGACGAATTTTTCAAGGCCCATCCAGACCGCAAGAAAAGCCAGCACACCCCT
 GATCATCATGCCTACCTGGAATGATGAGAAAAATGGTCTGACAAATCCGGGTGGCGTTCGGTTCTATGGTATG
 GGTCGTAATTGTTTTGCGTTGGACTACCCGGCGTTTGATATCAAGCACCTGGGTGAGAAAACCGGTGAGGGTC
 GTCTGCTGACGAAATGGTACGGTGGCATGGCGCACGCGCTGGGCCACGGCCTGAATCTGCCGCACAATCACCA
 GACCGCGAGCGATGGCAAGAAATATGGCACCGCCCTGATGGGTAGCGGCAACTACACGTTCCGGTACCAGCCCG
 20 ACGTTCCCTGACCCCGCGAGCTGTGCGCTGCTGGATGCTGCGAAGTGTTTACGCGTTACCCCGAGCCAACAGT
 TTTATGAGGGTAAGCCAGAAGTCGAGGTTGGTGATGTTGCAATTTCTTCAAGGGTGATCAAAATCTTGGTCAG
 CGGTAACCTACAAGAGCCCGCAAACCGTGAAAGCTCTGAACGTTTACATTACGATCCGCCGTACGCCGTGAAC
 CAAGACTACGATGCAGTGAGCTTTAGCCGTCGTCTGGGCAAAAAGTCCGGTAAGTTTAGCATGAAGATTGACA
 AAAAAGAACTGGAAGGCCTGAATAACAACGAATTCGTTATTTCTTGTATGTTTCTTGGCAACCGGCTTACA
 25 CATGCAGAAGCACTTTACGTTTCACTGGGATGCGCTGCAAGACTACCGTGACGGTAGCAAATCTGGTTCGGGT
 CATCATCACCACCATCACTGA
- SEQ ID NO: 8
 GMAHELGHGL (metalloprotease motif)
- 30 SEQ ID NO: 9
 MKNLLFALLTGSFCCCYAQKKAAPVPEPEVVATPPADAGRGLIRVDSREIRHYSGRTRKEPDYLVSRDNGKTWE
 MKAAPAGYPPNYGGIPKESPAIVRNPLTREFIRVQPIGGFVFLSRGGLDGKWLAVTNDGKLEEDWKDPEKRKN
 LKKLGGIMRTPVFNKGRRVIVPFHNMGGGTKFHIISDDGGLTWHVSRNGVTSRHEARPPHQGVRFNNAVEA
 35 TVLEMKDGTWALARTSQDQAWQAFSKDYGETWSKPEPSRFFGTLTMTNLGRLLDDGTIVSLWNTMALPENAT
 AGNGTWEDVFTNRDSSHIIAMSGDEGKTWYGFREIILDEHRNHPGYATLDGPEDRGRKHQSEMVQLDKNRILISL
 GQHKNHRRLLVIVDRRWGAKTRATQTGKDLDSQWTIHTYIPQKKGHCYSYNRKPSAELVQDPSGGTKKVLQIKR
 LDDPELVNEKSNVDYRNGGATWNFPNGTGLVKFRFRVVDGEQADDSGLQVSLTDRLFNACDSTTKDYALFTF
 PIRLKPAPHLLLGMMKVPFTPGAWHEISLLWQGGQAVVSLDGKKAGTLKMANKSPNGASYIHFIISTGSQPDAG
 40 ILDDTVNARVK (signal sequence underlined)
- SEQ ID NO: 10
 QQKAAPVPEPEVVATPPADAGRGLIRVDSREIRHYSGRTRKEPDYLVSRDNGKTWEMKAAPAGYPPNYGGIPKE
 SPAIVRNPLTREFIRVQPIGGFVFLSRGGLDGKWLAVTNDGKLEEDWKDPEKRKNLKKLGGIMRTPVFNKGR
 45 RVIVPFHNMGGGTKFHIISDDGGLTWHVSRNGVTSRHEARPPHQGVRFNNAVEATVLEMKDGTWALARTSQ
 DQAWQAFSKDYGETWSKPEPSRFFGTLTMTNLGRLLDDGTIVSLWNTMALPENATAGNGTWEDVFTNRDSSHII
 AMSGDEGKTWYGFREIILDEHRNHPGYATLDGPEDRGRKHQSEMVQLDKNRILISLQHKNHRRLLVIVDRRWG
 AKTRATQTGKDLDSQWTIHTYIPQKKGHCYSYNRKPSAELVQDPSGGTKKVLQIKRLDDPELVNEKSNVDYRNG
 GATWNFPNGTGLVKFRFRVVDGEQADDSGLQVSLTDRLFNACDSTTKDYALFTFPIRLKPAPHLLLGMMKVP
 50 FTPGAWHEISLLWQGGQAVVSLDGKKAGTLKMANKSPNGASYIHFIISTGSQPDAGILDDTVNARVK
- SEQ ID NO: 11
 MQKAAPVPEPEVVATPPADAGRGLIRVDSREIRHYSGRTRKEPDYLVSRDNGKTWEMKAAPAGYPPNYGGIPK
 SPAIVRNPLTREFIRVQPIGGFVFLSRGGLDGKWLAVTNDGKLEEDWKDPEKRKNLKKLGGIMRTPVFNKGR
 55 RRVIVPFHNMGGGTKFHIISDDGGLTWHVSRNGVTSRHEARPPHQGVRFNNAVEATVLEMKDGTWALARTSQ
 DQAWQAFSKDYGETWSKPEPSRFFGTLTMTNLGRLLDDGTIVSLWNTMALPENATAGNGTWEDVFTNRDSSHII
 AMSGDEGKTWYGFREIILDEHRNHPGYATLDGPEDRGRKHQSEMVQLDKNRILISLQHKNHRRLLVIVDRRWG
 GAKTRATQTGKDLDSQWTIHTYIPQKKGHCYSYNRKPSAELVQDPSGGTKKVLQIKRLDDPELVNEKSNVDYRNG

GGATWNFNPGTTGLVKFRFRVVDGEQADDSGLQVSLTDRLFNACDSTTKDYALFTFP IRLKPAPHLLLGMMKV
PFTPGAWHEISLLWQGGQAVVSLDGKKAGTLKMANKSPNGASYIHFIISTGSQPDAGILLDTVNARVKGSGLEH
HHHHH

5 SEQ ID NO: 12

MTWLLCGRGKWNKVCRMNSVFKCLMSAVCAVALPAFGQEEKTGFPTDRAVTVFSAGEGNPYASIRIPALLSI
GKGQLLAFAGRYKNTDQGENDIIMSVSKNGGKTWSRPRAIAKAHGATFNNPCPVYDAKTRTVTVVFQRYPA
VKERQPNIPDGWDDEKCI RNFMIQSRNGGSSWTKPQEITKTTKRPSGVDIMASGPNAGTQLKSGAHKGR
LVIP MNEGPF GKWVISCIIYSDDGKSWKLGQPTANMKGMVNETSIAETDNGGVMMVARHWGAGNCRRI
AWSQDGGET 10 WGQVEDAPELFC DSTQNSLMTYSLSDQPAYGGKSRI LFSGPSAGRRIKGQVAMS
YDNGKTPVVKLLGEGGFAYSSLAMVEPGIVGVLYEENQEHIKKLKFVPITMEWLTDGEDTGLAPGKKAPVLK
(signal sequence underlined)

SEQ ID NO: 13

15 QEEKTGFPDRAVTVFSAGEGNPYASIRIPALLSIGKGQLLAFAGRYKNTDQGENDIIMSVSKNGGKTWSRP
RAIAKAHGATFNNPCPVYDAKTRTVTVVFQRYPAVGERQPNIPDGWDDEKCI RNFMIQSRNGGSSWTKPQE
ITKTTKRPSGVDIMASGPNAGTQLKSGAHKGRLVIPMNEGPF GKWVISCIIYSDDGKSWKLGQPTANMKGMV
NETSIAETDNGGVMMVARHWGAGNCRRIAWSQDGGETWGQVEDAPELFC DSTQNSLMTYSLSDQPAYGGKSRI
LFSGPSAGRRIKGQVAMS YDNGKTPVVKLLGEGGFAYSSLAMVEPGIVGVLYEENQEHIKKLKFVPITMEWLTD
20 GEDTGLAPGKKAPVLK

SEQ ID NO: 14

MQEEKTGFPDRAVTVFSAGEGNPYASIRIPALLSIGKGQLLAFAGRYKNTDQGENDIIMSVSKNGGKTWSRP
PRAIAKAHGATFNNPCPVYDAKTRTVTVVFQRYPAVGERQPNIPDGWDDEKCI RNFMIQSRNGGSSWTKPQE
25 ITKTTKRPSGVDIMASGPNAGTQLKSGAHKGRLVIPMNEGPF GKWVISCIIYSDDGKSWKLGQPTANMKGMV
NETSIAETDNGGVMMVARHWGAGNCRRIAWSQDGGETWGQVEDAPELFC DSTQNSLMTYSLSDQPAYGGKSRI
LFSGPSAGRRIKGQVAMS YDNGKTPVVKLLGEGGFAYSSLAMVEPGIVGVLYEENQEHIKKLKFVPITMEWLTD
GEDTGLAPGKKAPVLKSGGLEHHHHHH

30 SEQ ID NO: 15 - O-glycosidase from *S. oralis*

MDKRFFEKRCKFSIRKFTLGVASVMIGATFFAASPLADQARVGSTDNLPSELADLDKKASDEGHDFDKEAAA
QNPESAETTEGPQTEEEELAQEKEKSEKPSNLPKELEDKLEKAEDNGREVDKQDLAQDTGKLVPE
DVAKTNG ELNYGATV KIKTPSGEGSGIVVAKDLVLT VSHNFIKDSQEGNIRKVVNDNQDGD
DIYSISYPGLPDVKFSKDD I IHWDRGYLKGFKNDLALVRLTVLENTPVEVTKKPVVKKIGDKLHV
FGYPEGKLNPIVNTTVDFAEPYGEG 35 VQIGIGYQGGKPGASGGGIFDTGKLVGVHQN
GVVGRSGGILFSPAQLKWIQDHMQGISSVKPADLEEKEKPA
EEKPKEDKPA AAKPETPKAVTPEWQTVANKEQQGTVTIREKGVRYNQLSSTAQN
DNDGKPALFEKQGLTVDA NGNATVDLTFKDDSEKGS RFGVFLFKFKDTKNNV
FVGIDQGGWFWEYKTPGNSTWYKGNRVAAPFGSVNRLS I TLKSDGQLNASNNDV
NLFDTVTLP GAVNENLNKNEKKILLKAGTYSNDRTVSVKTDNQEGVKADDT
PAQKET GPAVDDSKVTYDTIQSKVLKAVIDQAFPRVKEYTLNGHTLPGQVQFNQV
FINNHRTPEVTYKKINETTAEY 40 LMKLRDDAHLINAEMTVRLQVVDNQLHFDV
TKIVNHNQVTPGQKIDDERKLLSTISFLGNALVSVSSDQAGAK
FDGATMSNNTHVSGDDHIDVTNPMKDLAKGYMYGFVSTDKLAAGVWSNSQNSYGGG
SNDWTRLTAYKETVGNA NYVGIHSSEWQWEKAYKIVFPEYTKELPSAKVVIT
EDANADNKVDWQDGAIAYSIMNPNQGWKVKDITAY RIAMNFGSQAQNPFLMTLDGI
KKINLHTDGLGQGVLLKGYGSEGHDSGHLNYADIGKRIGGEDFKTLIEKAK KYGAHL
GIHVNASETYPE SKYFNENILRKNPDGSSYSGWNWLDQGINIDAAYDLAHGRLARWED
LKKKLGEG 45 LDFIYVDVWNGQSGDNGAWATHVLAKEINKQGWRF
AIEWGHGGEYDSTFQHWAAADLTGGYTNKGINSAITRF
IRNHQKDSWVG DYRSYGAANYPL LGGYSMKDFEGWQGRSDYNGYVTLNFAHDVMTKYF
QHFTVSKWENGTPV TMTDNGSTYKWTPEMKVELVDAAGNKVVTRKSNVNSPQYRERTV
TLNGRVIQDSAYLTPWNWDANGKKLP TEKEKMYFNTQAGATTWTLPSDWANSKVYLYKL
TDQGKTEEQELTVTDGKITLDDL ANQPYVLYRSKQTNPE MSWSEGMHIYDQGFNSGTL
KHWTISGDASKAEIVKSQGANEMLR IQGNKSKVSLTQKLTGLKPNKYAVYVGV 50
DNRSNAKASITVNTGEKEVTTYTNKSLALNYIKAYAHNNRRENATVDDTSYFQNM
YAFFTTGSDVSNVTLTLS REAGDEATYFDEIRTFENNSSMYGDKHDTGQGT
FKQDFENVAQGIFFPVVGGVEGVEDNRTHLSEKHDPYTOR GWNGKKVDDVIEGNW
SLKTNGLVSRRLVYQTI PQNFRFEAGKTYRVTFEYEAGSDNTYAFVVGKGEFQSGRR
GTQASNLEMHEL PNTWTDSSKAKKVTF LVTGAETGDTWVG IYSTGNASNTRGDAGG
NANFRGYNDFMMDNLQI EEITLTGKMLTENALKNYLPTVAMTNYTKESMDALKEAV
FNLSQADDDI SVEEARAEIAKIEALKNALVQKKT 55 ALVAEDFESLDAPAQPG
EGLENAFDGNVSSLWHTSWNGGDVGK PATMVLKEPTEITGLRYVPRASDSNGNLRD
VKLVVTDESGKEHTFNVDWPNNNKPKDIDFGKTIKAKKIVLTGTKTYGDGDKYQSA
AELIFTRPQVAETPL DLSGYEALAKAQKLTDKDNQEEVASVQASMKYATDNHLLTER
MVAYFADYLNQLKDSATKPDAPTSKGEEQ PPVLDVPEFKGGVNATEAAVHEVPEFKG
GVNAVQALVHELPEYKGGANAVLAAANEVPEYKGGANAVEALVNE

- KPAYTGVLATAGDQAAPTVEKPEYPLTPSPVADTKTPGAKDEEKL PATGEHSSEVALFLASVSIALSAAVLAT
KRKEEGSGLEHHHHH
- 5 SEQ ID NO: 16 - E206A_forward primer
ATGGCGCACGC GCTGGGCCACG
- SEQ ID NO: 17 - E206A_reverse primer
GCCACCGTAC CATTTTCGTC
- 10 SEQ ID NO: 18 - EPO
APRLICDSRVLERYLLEAKEAEDITTGCAEHCSLDENITVPDTKVDFYAWKRMEVGGQAVEVWQGLALLSEA
VLRGQALLVNSSQPWEPLQLHVDKAVSGLRSLTLLRALGAQKEAISPPDAASAAPLRTITADTRKLFVRVYS
NFLRGKLLKLYTGEACRTGDR
(Notes: predicted O-glycan bearing Serine is underlined; C terminal
15 Arginine is commonly truncated during expression)
- SEQ ID NO: 20
EVTVPDALKDRIALKKTARQLNIVYFLGSDTEPVPDYERRLSELLLYLQQFYGKEMQRHGYGARSFGLDIKSP
GRVNIIEYKAKNPAAHYPYENGSGWKAQELDEFFKAHPDRKKSQHTLIIMPTWNDEKNGPDNPGGVVPFYGMG
20 RNCFALDYPAFDIKHLGQKTREGRLLTWKYGGMAAALGHGLNLPNHNQTASDGKKYGTALMGSGNYTFGTSP
FLTPASCALLDACEVFSVTPSQQFYEGKPEVEVGDAISFKGDQILVSGNYKSPQTVKALNVYIQDPPYAVNQ
DYDAVSFSRRLGKKS GKFSMKIDKKELEGLNNNEFRISLMFILANGLHMQKHFTFHWDALQDYRDGSKS
- 25 SEQ ID NO: 21
MEVTVPDALKDRIALKKTARQLNIVYFLGSDTEPVPDYERRLSELLLYLQQFYGKEMQRHGYGARSFGLDIKS
PGRVNIIEYKAKNPAAHYPYENGSGWKAQELDEFFKAHPDRKKSQHTLIIMPTWNDEKNGPDNPGGVVPFYGM
GRNCFALDYPAFDIKHLGQKTREGRLLTWKYGGMAAALGHGLNLPNHNQTASDGKKYGTALMGSGNYTFGTSP
TFLTPASCALLDACEVFSVTPSQQFYEGKPEVEVGDAISFKGDQILVSGNYKSPQTVKALNVYIQDPPYAVN
30 QDYDAVSFSRRLGKKS GKFSMKIDKKELEGLNNNEFRISLMFILANGLHMQKHFTFHWDALQDYRDGSKSGSG
HHHHH
- SEQ ID NO: 22
- 35 ATGGAAGTCACTGTGCCGACGCCCTGAAAGATCGCATCGCGCTGAAGAAAACCGCTCGTCAGCTGAATATCG
TCTACTTCTGGGTTCTGATACCGAACCGGTTCCGGACTACGAGCGCCGTCTGAGCGAGCTGCTGTTGTATCT
GCAGCAATTCTATGGTAAAGAAATGCAGCGCCATGGCTATGGCGCACGCAGCTTTGGTCTGGACATTAAGTCA
CCGGGTCTGTGAACATTATCGAGTACAAAGCGAAGAACCCGGCAGCGCATTACCCGTATGAGAAATGGTGGCG
GCTGGAAAGCTGCACAAGAACTGGACGAATTTTCAAGGCCCATCCAGACCGCAAGAAAAGCCAGCACACCCCT
40 GATCATCATGCCTACCTGGAATGATGAGAAAAATGGTCCTGACAAATCCGGGTGGCGTTCGGTCTATGGTATG
GGTCGTAATTGTTTTCGTTGGACTACCCGGCGTTTGATATCAAGCACCTGGGTGAGAAACGGGTGAGGGTC
GTCTGCTGACGAAATGGTACGGTGGCATGGCGGCCGCGCTGGGCCACGGCCTGAATCTGCCGCACAATCACCA
GACCGCGAGCGATGGCAAGAAATATGGCACCGCCCTGATGGGTAGCGGCAACTACACGTTCCGTACCAGCCCG
ACGTTCTGACCCCGCGAGCTGTGCGCTGCTGGATGCCTGCGAAGTGTTACAGCGTTACCCCGAGCCAACAGT
45 TTTATGAGGGTAAGCCAGAAGTCGAGGTGGTGATGTTGCAATTTCTTCAAGGGTGATCAAATCTTGGTCAG
CGGTAACATAAGAGCCCGCAAACCGTGAAAGCTCTGAACGTTTACATTACAGATCCGCCGTACGCCGTGAAC
CAAGACTACGATGCAGTGAGCTTTAGCCGTCGTCTGGGCAGAAAAGTCCGGTAAGTTTAGCATGAAGATTGACA
AAAAAGAACTGGAAGGCCTGAATAACAACGAATTCCTGATTTCTTGGTATGTTTCTGGCAAACGGCTTACA
CATGCAGAAGCACTTTACGTTTCACTGGGATGCGCTGCAAGACTACCGTGACGGTAGCAAATCTGGTTCGGGT
50 CATCATCACCACCATCACTGA
- SEQ ID NO: 23
GMAHALGHGL (disrupted metalloprotease motif)
- 55 SEQ ID NO: 24
GMAAELGHGL (disrupted metalloprotease motif)

SEQ ID NO: 25
GMAAALGHGL (disrupted metalloprotease motif)

- 5 SEQ ID 26
Pseudomonas aeruginosa PAO1 (native sequence with removed signal sequence)

10 ATQEEILDAAVSGDSSQLTDSHLVALRLQQQVERIRQTRTQLLDGLYQNLSSQAYDPGAASMWVLPANPDNTL
PFLIGDKGRVLASLSLEAGGRGLAYGTNVLTQLSGTNAAHAPLLKRAVQWLVNGDPGAATAKDFKVSVVGVDK
TAALNGLKSAGLQPADAACNALTDASCASTSKLLVLGNGASAA¹⁵SLSATVRARLQAGLPILFVHTNGWNQSSTG
QQILAGLGLQEGPYGGNYWDKDRVPSSRTRRSVELGGAYGQDPALVQQIVDGSWRTDYDWSKCTSYVGRITC
DDVPGLSDFSKRVDVLKGALDAYNQKAQNLFALPGTTSRLWLLWADAVRQNIIRYPMDKAADTARFQETFFVAD
AIVGYVREAGAAQKELGSYAGQRQQSMPVSGSEETLTLTLP²⁰SAQGF¹⁵TAIGRMAAPGKRLSIRIEDAGQASLAV
GLNTQRIGSTRLWNTRQYDRPRFLKSPDIKLANQSV¹⁵ALVSPYGGLLQLVYSGATPGQTVTVKVTGAASQPFL
DIQPGEDSSQAIADFIQALDADKADWLEIRSGSVEVHAKVEKVRGSDKDYGGDVQRFIRELNEVFIDDAYTL
AGFAIPNQAKTPAIQQECAARGWDCDSE¹⁵TLHKLPGTQHINVDQYACGGGCSGNPYDQTWGLNPRGWGESHEL
GHN¹⁵LQVNRLKVYGGRSGEISNQIFPLHKDWRVLR¹⁵EFQNLDDTRVNYRNAYNLIVAGRAEADPLAGVYKRLWE
DPGT¹⁵YALNGERMAFYTQWVHYWADLKN¹⁵DPLQGW¹⁵DIWTL¹⁵LYLHQRQVDKSDWDANKAALGYGT¹⁵YQRPGNSGDA
SSTDGNDNLLGLSWLTQRDQRPTFALWGIRTSAAQAQVAAYGFAEQPAFFYANNRTNEYSTVKLLDMSQGS
PAWPPF

Metalloprotease motif underlined.

- 25 SEQID 27
Bacteroides thetaiotaomicron VPI-5482 (native sequence with removed signal sequence)

30 DKWEKEFRIRSYEPYSNIAEWADKLMTKKYSDLDNPTGISVKAGDDIIVLVGD¹⁵TYGQNISMQCIWETGTEYKQ
TASSGDVYMLNPGV¹⁵NKLT¹⁵MKGEGQLFV¹⁵MYNTELT¹⁵SN¹⁵TAKPIKIH¹⁵PLGSGTVNGFFDLKEHK¹⁵TDEKYAELLKK
STHKYFCIRGEKIMFYFHRNKLLEYVPNNILSAIHLWDNIVGWQQELMGIDDVRPSQVNNHLFAISPEGSYMW
ASDYQIGFVYTYLGNILLEDNVMAEDN¹⁵AWGPAHEIGHVHQA¹⁵INWASSTESSNNLFSNFI¹⁵IYKLGKYKSRGN
GLG¹⁵SVATARYANGQAWYNMGDATHQNEDETETHMRMNWQLWIYYHRCEYKTDFWQTLFKLMREVNMTGEDDPGK
KQLEFAKMASKAANQNLTDFFEMWGFFEPVNTTIEQYGT¹⁵YKYVSDAMIREAKEYMAQFPAPKHAFQYIEDRK
KSEFPSNDYRYS¹⁵AVGDVGYTQF¹⁵KENQKITKAITAELAGRKVSIQNGDEAVAFELRENDENGKLLYFSTFTTF
35 EIPSSILMVNAKLYAVQADGKRILL

Metalloprotease motif underlined.

- 40 SEQ ID NO: 28
Clostridium perfringens (native sequence with removed signal sequence)

45 VLELEMRGDSISEAKKRKVWNFQDWQITGLSARAGDKITVYVDVAEGDPTPTLLYKQSLTQHGGGATSFQLKPG
KNEITIPEINYESNGIPKDVIQGGDLFFTN¹⁵YKSDSQKRAPKVRIEGASKYPVFILGKSDENEVMKELEAYVEK
IKAEPKTT¹⁵PNIFAVSSNKSLEFVQATYALDWYKKNKTPKYTAEQWDQYIADAMGF¹⁵WGF¹⁵DN¹⁵SKDVNSDFNFRI
MPMVKNLSGGA¹⁵FMNAGNGVIGIRPGNQDAILAANKGWGAHELGHNFDTGGRTIVEVTNNMMPLFFESKYKTK
TRITDQNIWENNTYPKVGLDDYSNNELYNKADSTHLAQLAPLWQLYLYDNTFYGKFERQFRERDFGNKNREDI
YKSWVVAASDAMELDL¹⁵TEFFARHGIRVDDKV¹⁵KEDLAKYPKPDKKIYYLNDLAMNYKGDGTENAKVSVSTSGS
NGNIKLSFSVDENKDNILGYEIRRDGKYVGFTSND¹⁵SFVDTKSNLDEDGVYVVT¹⁵PYDRKLNTLNPIEVN

- 50 Metalloprotease motif underlined.

- SED ID NO: 29
Pseudomonas aeruginosa PAO1 (N-term Met, C-term linker/tag)

55 MATQEEILDAAVSGDSSQLTDSHLVALRLQQQVERIRQTRTQLLDGLYQNLSSQAYDPGAASMWVLPANPDNT
LPFLIGDKGRVLASLSLEAGGRGLAYGTNVLTQLSGTNAAHAPLLKRAVQWLVNGDPGAATAKDFKVSVVGVD

KTAALNGLKSAGLQPADAACNALTDASCASTSKLLVLNGGASAAASLSATVRARLQAGLPILFVHTNGWNQSSST
 GQQILAGLGLQEGPYGGNYWDKDRVPSSRTRTRSVELGGAYGQDPALVQQIVDGSWRDIDYDWSKCTSYVGRIT
 CDDVPGLSDFSQRVDVLKGALDAYNQKAQNLFPALPGTSLRLWLLWADAVRQNIIRYPMDKAADTARFQETFVA
 DAIVGYVREAGAAQKELGSYAGQRQQSMPVSGSEETLTTLTPSAQGFTAIGRMAAPGKRLSIRIEDAGQASLA
 5 VGLNTQRIGSTRLWNTRQYDRPRFLKSPDIKLQANQSVALVSPYGGLLQLVYSGATPGQTVTVKVTGAASQPF
 LDIQPGEDSSQAIADFIQALDADKADWLEIRSGSVEVHAKVEKVRGSIDKDYGGDVQRFFIRELNEVFIDDAYT
 LAGFAIPNQAKTPAIQQECAARGWDCDSETLHKLPQTQHINVDQYAQCGGGCSGNPYDQTWGLNPRGWGESHE
 LGHNQLQVNRLKVYGGRSGEISNQIFPLHKDWRVLRREFGQNLDDTRVNYRNAYNLIVAGRAEADPLAGVYKRLW
 EDPGTALNGERMAFYTQWVHYWADLKNLPLQGWDIWTLTLLYHQRQVQDKSDWDANKAALGYGTYAQRPGNSGD
 10 ASSTDGNDNLLGLSWLTQRDQRPTFALWGIRTSAAAQAQVAAYGFAEQPAFFYANNRTNEYSTVKLLDMSQG
 SPAWFFPGSGHHHHH

Metalloprotease motif underlined.

15 SEQ ID NO: 30
 Bacteroides thetaiotaomicron VPI-5482 (N-term Met, C-term linker/tag)

MDKWEKEFRIRSYEPYSNIAEWADKLMTKKYSDDLNPSTGISVKAGDDIIVLVGDTYQGNISMQCIWETGTEYK
 QTASSGDVYMLNPGVNKLTMKGGQLFVMYNTELTSTAKPIKIHPLGSGTVNGFFDLKEHKTDEKYAELLK
 20 KSTHKYFCIRGEKIMFYFHRNKLLEYVNNILSAIHLWDNIVGWQQLMGIDDVRPSQVNNHLFAISPEGSYM
 WASDYQIGFVYTYLGNILLEDNVMAAEDNAWGPAHEIGHVHQAINWASSTESSNNLFSNFIIYKLGKYKSRG
 NGLGSAVATARYANGQAWYNMGDATHQNEDETHMRMNWQLWIIYHRCEYKTDWFQTLFKLMREVNMTGEDPG
 KKQLEFAKMASKAANQNLTDFEMWGFEPVNTTIEQYGTYYVSDAMIREAKYMAQFPAPKHAQYIEDR
 KKSEFPSNDYRYSAVGDVGYTQFKENQKITKAITAELAGRKVSIQNGDEAVAFELRENDENGKLLYFSTFTT
 25 FEIPSSILMVNAKLYAVQADGKRILLGSGHHHHH

Metalloprotease motif underlined.

30 SEQ ID NO: 31
 Clostridium perfringens (N-term Met, C-term linker/tag)

MVLELEMRGDSISEAKKRKVWNFQDWQITGLSARAGDKITVYVDVAEGDPTPTLLYKQSLTQHGATSFQLKP
 GKNEITIPENYENSGIPKDVIOGGDLFFTNYSKSDSKRAPKVRIEGASKYPVFILGKSDENEVMKELEYVE
 KIKAEPTTPNIFAVSSNKSLEFVQATYALDWYKNNKTPKYTAEQWDQYIADAMGFWGFDSKDVNSDFNFR
 35 IMPMVKNLSGGAFMNGVIGIRPGNQDAILAANKGWGVAAHELGHNFDTGGRTIVEVTNNMPLFFESKYKT
 KTRITDQNIWENNTYPKVGLDDYSNNELYNKADSTHLAQLAPLWQLYLYDNTFYGKFERQFRERDFGKNKRED
 IYKSWVVAASDAMELDLTFEFARHGIRVDDKVKEDLAKYKPKDKKIYYLNDLAMNYKGDGFTENAKVSVSTSG
 SNGNIKLSFSVDENKDNILGYEIRRDGKYVGFTSNDSEFVDTKSNLDEGDVYVVTYPYDRKLNLTNPPIEVNGSG
 HHHHHH

40 Metalloprotease motif underlined.

SEQ ID NO: 32 Pseudomonas aeruginosa PA01 (full native sequence including
 signal sequence)
 45 MSLSTTAFFPSLQGENMSRSPIPRHRALLAGFCLAGALSAQAATQEEILDAALVSGDSSQLTDSHLVALRLQQQ
 VERIRQTRTQLLDGLYQNLQAYDPGAASMWLVPANPDNTLPFLIGDKGRVLASLSLEAGGRGLAYGTNVLTQ
 LSGTNAAHAPLLKRAVQWLNVGDPGAATAKDFKVSVGVGDKTAALNGLKSAGLQPADAACNALTDASCASTSK
 LLVLNGGASAAASLSATVRARLQAGLPILFVHTNGWNQSSSTGQQILAGLGLQEGPYGGNYWDKDRVPSSRTR
 SVELGGAYGQDPALVQQIVDGSWRDIDYDWSKCTSYVGRITCDDVPGLSDFSQRVDVLKGALDAYNQKAQNLFA
 LPGTSLRLWLLWADAVRQNIIRYPMDKAADTARFQETFVADAIVGYVREAGAAQKELGSYAGQRQQSMPVSGS
 50 EETLTTLTPSAQGFTAIGRMAAPGKRLSIRIEDAGQASLAVGLNTQRIGSTRLWNTRQYDRPRFLKSPDIKLQ
 ANQSVALVSPYGGLLQLVYSGATPGQTVTVKVTGAASQPFLLDIQPGEDSSQAIADFIQALDADKADWLEIRSG
 SVEVHAKVEKVRGSIDKDYGGDVQRFFIRELNEVFIDDAYTLAGFAIPNQAKTPAIQQECAARGWDCDSETLHK
 LPGTQHINVDQYAQCGGGCSGNPYDQTWGLNPRGWGESHELGHNLQVNRLKVYGGRSGEISNQIFPLHKDWRV
 LREFGQNLDDTRVNYRNAYNLIVAGRAEADPLAGVYKRLWEDPGTYALNGERMAFYTQWVHYWADLKNLPLQ
 55 WDIWTLTLLYHQRQVQDKSDWDANKAALGYGTYAQRPGNSGDASSTDGNDNLLGLSWLTQRDQRPTFALWGI
 RTSAAAQAQVAAYGFAEQPAFFYANNRTNEYSTVKLLDMSQSPAWPFP

Uniprot accession: Q9I5W4.1

Metalloprotease motif underlined.
Signal sequence bold, underlined.

- 5 SEQ ID NO: 33
Bacteroides thetaiotaomicron VPI-5482 (full native sequence including
signal and other sequences)

10 **MTIKRFITNLLALFTLFTVSLACKDTEKSIINSSFSISEEYLIQNLDKSSTSVQIPINTSMELAQWSVSYEAN**
WLQCSKQKTAAGTFLRITVNENTGETKRTANIKVTSTTATYTTITVNQYAKGEVIVEGDIKVTPTGGKASEHQ
EGQDIENTYDGFSTDGAAPFHTPWGQSAKFPVTLEYFYGKDEIDYLIYYTRSGNGNFGKVKVYTTTNPDRS
DYTLQGEYDFKEQNAPSKVSFSEGIKATGIKFEVLSGLGDFVSCDEMEFYKTNTDKTLQLLTVFTDITCTE
IKNNVTNEQIQALPDYFVRIAEAVRDNTYDKWEKEFRIRSYEPYSNIAEWADKLMTKKYSDDLNPGISVKAG
DDIIVLVGDTYQONISMQCIWETGTEYKQTASSGDVYMLNPGVNKLTMKGEGQLFVYNTELTSNTAKPIKIH
15 IPLSGTVNGFFDLKEHKTDEKYAELLKKSTHKYFCIRGEKIMFYFHRNKLLEYVPPNNILSAIHLWDNIVGWQ
QELMGIDDVRSQVNNHLFAISPEGSYMWASDYQIGFVYTYLGNILLEDNVMAAEDNAWGPAHEIGHVHQA
NWSSTESSNNLFSNFIIYKLGKYKSRGNGLGSVATARYANGQAWYNNMGDATHQNEDETETHMRMNWQLWIIYYH
RCEYKTDWFQTLFKLMREVNMTGEDPGKKQLEFAKMASKAANQNLTDFEMWGFEPVNTTIEQYGTYYV
SDAMIREAKYMAQFPAPKHAFOYIEDRKKSEFPSNDYRYSAGVDVGYYTQFKNQKITKAITAELAGRKVISI
20 QNGDEAVAFELRENDENGLLYFSTFTTTEIPSSILMVNAKLYAVQADGKRILL

Uniprot accession: Q89ZX7.1

Metalloprotease motif underlined.
Signal sequence bold, underlined.

- 25 Other sequences removed in mature protein bold, italic.

SEQ ID NO: 34

- 30 Clostridium perfringens (full native sequence including signal and other
sequences)

MNKRKIAAII LATMITNLSATTIDVLAQELNTKNNSKVEVSHDDESHQARVSKFDLYNSDKLDAYNQEFQVSR
SNIKSINNNGGKYNSSTIDKAIDGNLETHWETGKPNANFTNEVVVTFNEITNIDRIVYSARRDSARGKGFAK
EFETIYASLKDEGDDFNLVSSGEYTESTRDLVEIKFNPDTFKRLKFKFKKADQNWASAAEFMFYKEDKLNKFN
35 **GLFTDSSMNKVSEEFNTLEKLNAFENELKDHPIDYLYKEGLNNARAILTETSENPATLGQITYNLNDYNN**
QYRMPYKNIKAIKNNGRHYAAQNIKAIDNDVNTYWTGTNLNSSFNNEVEVEFNDLVTLDRIVYGSRSOLK
GFAEEVYIYASRTSKGDTYKLVATGAHEATKGLVEAKFEPTFEKRVKFKFKKSKQNSATLNELMFYKPDVYS
SIPKLFTDGTMSELSEEFNSLEKINAFKEKAKNHPLYNDFNETIELAESLISNPRKEDVLELEMRGDSISEAK
40 **KRKVWNFQDWQITGLSARAGDKITVYVDVAEGDPTPTLLYKQSLTQHGGATSFQLKPGKNEITIPEINYESNG**
IPKDVIGQGDLFFTNYKSDSQKRAPKVRIEGASKYPVFI LGKSDENEVMKELEAYVEKIKAEPKTTPNIFAVS
SNKSLEFVQATYALDWYKKNKTPKYTAEQWDQYIADAMGFWGF DNSKDVNSDFNFRIMPMVKNLSGGAFMNA
GNGVIGIRPGNQDAILAANKGWGAHELGHNFDTGGRTIVEVTNNMPLFFESKYTKTRITDQNIWENNTYP
KVGLDDYSNNELYNKADSTHLAQLAPLWQLYLYDNTFYGKFERQFRERDFGNKNREDIYKSWVVAASDAMELD
45 **LTEFFARHGIRVDDKVKEDLAKYKPKDKKIYYLNDLAMNYKGDGFTEAKVSVSTSGSNGNIKLSFSVDDENK**
DNILGYEIRRDGKYVGFTSNDSFVDTKSNLDEDDGVYVTPYDRKLNTLNP IEVNALQPTLSVNPVITLALGEE
FNEEEYIVAKDIKGNLSSESVKVSNNVT SKVGEYEVLYSLED SKGNEYTKTSKVNVS SRKEYMSDLTPKQS
SNGWGTVRKDKSISGGVIGLTRDGD FVDYKGLGLHSNAEYVYDLE GKDYDYFESYVGVDKAMSSRPASSVIF
KVLVDGEEKFNSGVMRSTTPQKYVKVDVKNAKELKLI VNDAGDGSSD HASFGDAKLATLSSKPIIKGENLAY
50 **SMDEKVDLMKGITATDIEDGNITSKVQIKSSDFVEGKSGIFTVVYSVTSDGLTSECSRTIAVTDKETQLSDL**
NWKSATIGSGSVRKDRAVSGNQIRLLNEDNSVETFAKGIGTHSYSEIVYNSEGYDIFDTWVGIDRHVADKKVS
SVKFKVYVDGELKAETDVMRIDTPKKRLVVDVRNSKEIKLVVDVADNGNWNWDHADWADAKFRNLAEYDASELN
KAIEEAKKLDLNNYTESSEALKNAISKGEEALLSKDKETINSAL EELNKEMNSLVKVDLNAVINIPDKYLLK
SIQNQLNKTGDTITLGDMYSLTTLTSGVEDLTGLENAKNLET LNMDYNEVKDLRPLSKLKKLNTLNAQE QFIA
55 **AGELKPSNGKVI GDSKVYNREGKNVAKTIRVVDKNGNTILEQDAKDEFTINTKDLSSGLYGVHVLFEDEGFSG**
VMFYLFNV

Uniprot accession: A0A0H2YN38.1

Metalloprotease motif underlined.

Signal sequence bold, underlined.
Other sequences removed in mature protein bold, italic.

SEQ ID NO: 35

5 Pseudomonas aeruginosa PAO1 (double mutant with removed signal sequence)

ATQEEILDAAVLVSGDSSQLTDSHLVALRLQQQVERIRQTRTQLLDGLYQNLQAYDPGAASMWVLPANPDNTL
PFLIGDKGRVLASLSLEAGGRGLAYGTNVLTQLSGTNAHAHAPLLKRAVQWLVNGDPGAATAKDFKVSVVGVDK
TAAINGLKSAGLQPADAAACNALTDASCASTSKLLVLGNGASAAASLTVRARLQAGLPILFVHTNGWNQSSSTG
10 QQILAGLGLQEGPYGGNYWDKDRVPSSRTRRSVELGGAYGQDPALVQQIVDGSWRTDYDWSKCTSYVGRITC
DDVPGLSDFSKRVVLDKALDAYNQKQNLFALPGTTSRLWLLWADAVRQNI RYPMDKAADTARFQET FVAD
AIVGVVREAGAAQKELGSYAGQRQQSMPVSGSEETLTLTLP SAQGFTAIGRMAAPGKRLSIRIEDAGQASLAV
GLNTQRIGSTRLWNTRQYDRPFLKSPDIKLQANQSVALVSPYGGLLQLVYSGATPGQTVTVKVTGAASQPFL
DIQPGEDSSQAIADFIQALDADKADWLEIRSGSVEVHAKVEKVRGSIDKDYGGDVQR FIRELNEVFIDDAYTL
15 AGFAIPNQAKTPAIQQECAARGWDCDSETHKLPGTQHINVDQYAQCGGCGSGNPYDQTWGLNPRGWGES SAAL
GHNLQVNRLKVYGGSGEISNQIFPLHKDWRVLRFGQNLDDTRVNYRNAYNLIVAGRAEADPLAGVYKRLWE
DPGTYALNGERMAFYTQVWHYWADLKN DPLQGWDIWTLLYLHQRQVDKSDWDANKAALGYGTYAQRPGNSGDA
SSTDGNDNLLLGLSWLTQRDQRPTFALWGIRTSAAAQQAQVAAYGFAEQPAFFYANNRTNEYSTVKLLDMSQGS
PAWFFP

20 Disrupted metalloprotease motif underlined

SEQ ID NO: 36

25 Bacteroides thetaiotaomicron VPI-5482 (double mutant with removed signal and other immature sequences)

DKWEKEFRIRSYEPYSNIAEWADKLMTKKYSDDLNP TGISVKAGDDIIVLVGDTYQGNISMQCIWETGTEYKQ
TASSGDVYMLNPGVKNLTMKGEGQLFVMYNTELTSTNAKPIKIHIPLGSGTVNGFFDLKEHKTDEKYAELLKK
STHKYFCIRGEKIMFYFHRNKLLEYVPPNNILSAIHLWNIWGWQQLMGIDDVRSQVNNHLFAISPEGSYMW
30 ASDYQIGFVYTYLGNILLEDNVMAAEDNAWGPA AAIGHVHQAAINWASSTESSNNLFSNFIIYKLGKYKSRGN
GLGSVATARYANGQAWYNMGDATHQNEDETETHMRMNWLWIYYHRCEYKTD FWQTLFKLMREVNMTGEDPGK
KQLEFAKMAKAAANQNLTDFEMWGFFEPVNTTIEQYGTYYVSDAMIREAKEYMAQFPAPKHAFQYIEDRK
KSEFPSNDYRYSAGVDVGYTYQFKENQKITKAITAELAGRKVSIQNGDEAVAFELRENDENGKLLYFSTFTTF
EIPSSILMVNAKLYAVQADGKRILL

35 Disrupted metalloprotease motif underlined

SEQ ID NO: 37

40 Clostridium perfringens (double mutant with removed signal and other immature sequences)

VLELEMRGDSISEAKKRKVWNFQDWQITGLSARAGDKITVYVDVAEGDPTPTLLYKQSLTQHGGATSFQLKPG
KNEITIPEINYESNGIPKDV IQGGDLFFTNKYKSDSQKRAPKVRIEGASKYPVFILGKSDENEVMKELEAYVEK
IKAEPKTPPNIFAVSSNKSLEFVQATYALDWYKKNKTPKYTAEQWDQYIADAMGFWGFDNSKDVNSDFNFRI
45 MPMVKNLSGGAFMNAGNVIGIRPGNQDAILAANKGW GVA AAALGHNFDTGGR TIVEVTNNMPLFFESKYKTK
TRITDQNIWENNTPKVGLDDYSNNELYNKADSTHLAQLAPLWQLYLYDNTFYGKFERQFRERDFGNKNREDI
YKSWVVAASDAMELDLTEFFARHGIRVDDKVKEDLAKYKPKDKKIYYLNDLAMNYKGDGF TENAKVSVSTSGS
NGNIKLSFSVDENKDNILGYEIRRDGKYVGFTSNDSFVDTKSNLDEDGVYVVT PYDRKLNTLNPIEVN

50 Disrupted metalloprotease motif underlined

SEQ ID NO: 38

55 Pseudomonas aeruginosa PAO1 (double mutant with removed signal sequence, with N-term Met, C-term linker/tag)

- MATQEEILDAAALVSGDSSQLTDSHLVALRLQQQVERIRQTRTQLLDGLYQNLSSQAYDPGAASMWVLPANPDNT
 LPFLIGDKGRVLASLSLEAGGRGLAYGTNVLTQLSGTNAAHAPLLKRAVQWLVNGDPGAATAKDFKVSVVGVVD
 KTAALNGLKSAGLQPADAACNALTDASCASTSKLLVLGNGASAASLSATVRARLQAGLPILFVHTNGWNQSST
 GQQILAGLGLQEGPYGGNYWDKDRVPSSRTRTRSVELGGAYGQDPALVQQIVDGSWRTDYDWSKCTSYVGRTT
 5 CDDVPGLSDFSKRVDVLKALDAYNQKAQNLFALPGTTSRLWLLWADAVRQNI RYPMDKAADTARFQET FVA
 DAIVGYVREAGAAQKELGSYAGQRQQSMPVSGSEETLTLTLPQAQGFATAIGRMAAPGKRLSIRIEDAGQASLA
 VGLNTQRIGSTRLWNTRQYDRPFLKSPDIKLANQSVLSPYGGLLQLVYSGATPGQTVTVKVTGAASQPF
 LDIQPGEDSSQAIADFIQALDADKADWLEIRSGSVEVHAKVEKVRGSDKDYGGDVQRFIRELNEVFIDDAYT
 LAGFAIPNQAKTPAIQQECAARGWDCDSETLHKLPGTQHINVDQYACGGGCSGNPYDQTWGLNPRGWGESAA
 10 LGHNLQVNRLKVYGGRSGEISNQIFPLHKDWRVLRFGQNLDDTRVNYRNAYNLIVAGRAEADPLAGVYKRLW
 EDPGTALNGERMAFYTQWVHYWADLKNDFLQGWDIWTLTYLHQRQVDKSDWDANKAALGYGTYAQRPGNSGD
 ASSTDGNDNLLGLSWLTQRDQRTFALWGIRTSAAQAQVAAYGFAEQPAFFYANNRTNEYSTVKLLDMSQG
 SPAWPFPGSGHHHHH
- 15 Disrupted metalloprotease motif underlined
- SEQ ID NO: 39
 Bacteroides thetaiotaomicron VPI-5482 (double mutant with removed signal
 and other sequences from immature protein, with N-term Met, C-term
 20 linker/tag)
- MDKWEKEFRIRSYEPYSNIAEWADKLMTKKYSDDLNP TGISVKAGDDIIVLVGDTYGQNISMQCIWETGTEYK
 QTASSGDVYMLNPGVKNLTKGEGQLFVMYNTELTSTAKPIKIHIPLGSGTVNGFFDLKEHKTKDEKYAELLK
 KSTHKYFCIRGEKIMFYFHRNKLLEYVPNNILSAIHLWDNIWGWQQLMGIDDVRPSQVNNHLFAISPEGSYM
 25 WASDYQIGFVYTYLGNILLEDNVMAAEDNAWGPA~~AA~~AIGHVHQAAINWASSTESSNNLFSNFIIYKLGKYKSRG
 NGLGSVATARYANGQAWYNMGDATHQNEDETETHMRMNWQLWIIYHRCEYKTD FWTFLFKLMREVNMTEGEDPG
 KKQLEFAKMAKAAQNLTDFEMWGFEPVNTTIEQYGTYYVSDAMIREAKEYMAQFPAPKHAFQYIEDR
 KKSEFPNDYRYSAGVDVGYYTQFKENQKITKAITAELAGRKVSIQNGDEAVAFELRENDENGKLLYFSTFTT
 FEIPSSILMVNAKLYAVQADGKRILLGSGHHHHH
- 30 Disrupted metalloprotease motif underlined
- SEQ ID NO: 40
 Clostridium perfringens (double mutant with removed signal and other
 35 sequences from immature protein, with N-term Met, C-term linker/tag)
- MVLELEMRGDSISEAKKRKVWNFQDWQITGLSARAGDKITVYVDVAEGDPTPTLLYKQSLTQHGATS FQLKP
 GKNEITIPEINYESNGIPKDVIOGGDLFFTNYKSDSQKRAPKVRIEGASKYPVFILGKSDENEVMKELEAYVE
 KIKAEPKTTPNIFAVSSNKSLEFVQATYALDWYKNNKTPKYTAEQWDQYIADAMGFWGFDNSKDVNSDFNFR
 40 IMPMVKNLSGGAFMNAGNVIGIRPGNQDAILAANKGWGVA~~AA~~ALGHNFDTGGRTIVEVTNNMPLFFESKYKT
 KTRITDQNIWENNTYPKVGLDDYSNNELYNKADSTHLAQLAPLWQLYLYDNTFYGKFERQFRERDFGNKNRED
 IYKSWVVAASDAMELDLTEFFARHGIRVDDKVKEDLAKYKPKDKKIYYLNDLAMNYKGDGFTENAKVSVSTSG
 SNGNIKLSFSVDENKDNILGYEIRRDGKYVGFTSNDSEFVDTKSNLDEDGVYVVTYPYDRKLNLTNPIEVNGSG
 HHHHHH
- 45 Disrupted metalloprotease motif underlined
- SEQ ID NO: 41
 50 HELGH (metalloprotease motif)
- SEQ ID NO: 42
 HEIGH (metalloprotease motif)
- 55 SEQ ID NO: 43
 GVAHELGHNF (metalloprotease motif)

- SEQ ID NO: 44
HALGH (disrupted metalloprotease motif)
- 5 SEQ ID NO: 45
AELGH (disrupted metalloprotease motif)
- SEQ ID NO: 46
AALGH (disrupted metalloprotease motif)
- 10 SEQ ID NO: 47
GKPRPYSPRPTSHPRPIRV (glycodrosocin peptide with O-gly site on the T)
- SEQ ID NO: 48
15 YIYGSFK (Non-O-glycosylated peptide)
- SEQ ID NO: 49
KKLVFFA (Non-O-glycosylated peptide)
- 20 SEQ ID NO: 50
FLPLILGKLVKGLL (Non-O-glycosylated peptide)

CLAIMS

1. A polypeptide having endoprotease activity specific for O-glycosylated proteins which comprises:
- 5 (a) an amino acid sequence of SEQ ID NO: 1;
- (b) an amino acid sequence which is at least 85% identical to the amino acid sequence of SEQ ID NO: 1 or
- (c) an amino acid sequence which is a fragment of the sequence of SEQ ID NO: 1 or a fragment of an amino acid sequence which is 85% identical to the amino acid sequence of
- 10 SEQ ID NO: 1.
2. The polypeptide according to claim 1, wherein the amino acid sequence which is at least 85% identical to the amino acid sequence of SEQ ID NO: 1, or the fragment thereof, comprises the motif HEbbH, wherein b is an uncharged amino acid, optionally A, C, F, G,
- 15 I, L, M, N, P, Q, S, T, V or W, and optionally wherein said motif is present in said polypeptide at positions corresponding to positions 181 to 185 of SEQ ID NO: 1.
3. The polypeptide according to claim 2, wherein said motif comprises the sequence HEIGH or HELGH, preferably HELGH.
- 20 4. The polypeptide according to any one of the preceding claims which comprises the motif abxHEbbHbc, wherein:
- (a) a is amino acid V, T or G;
- (b) b is an uncharged amino acid, optionally A, C, F, G, I, L, M, N, P, Q, S, T,
- 25 V or W;
- (c) x is any amino acid; and
- (d) c is a hydrophobic amino acid, optionally A, C, F, I, L, M, P, V, W or Y.
- optionally wherein said motif comprises the sequence GMAHELGHGL or GVAHELGHNF, preferably GMAHELGHGL.
- 30 5. The polypeptide according to any one of the preceding claims, which includes an additional methionine at the N terminus and/or a His tag at the C terminus, which tag may

be joined to the C terminus by a linker, optionally wherein said polypeptide comprises or consists of the amino acid sequence of SEQ ID NO: 2.

6. The polypeptide according to any one of claims 1 to 4, wherein the polypeptide is provided in solution, lyophilised, or immobilised, optionally wherein the polypeptide is provided together with a sialidase, preferably Am1757 or a mixture of Am1757 and Am0707, wherein Am1757 is a polypeptide consisting of SEQ ID NO: 11 and/or wherein Am0707 is a polypeptide consisting of SEQ ID NO: 14.
7. A method of hydrolysing an O-glycoprotein, wherein the method comprises contacting a sample comprising the protein with a polypeptide according to any one of claims 1 to 6 and optionally further comprising the detection or analysis of the hydrolysis products.
8. A method for assessing the glycosylation status of a protein, comprising contacting a sample comprising the protein with a polypeptide according to any one of claims 1 to 6 and detecting and/or analysing the products produced, optionally wherein the presence or absence of cleavage products is used to determine the presence or absence of an O-glycoprotein in the sample, and/or wherein said analysis is conducted to identify the type of a O-glycan chain and/or its position of attachment to an O-glycoprotein.
9. The method according to claim 7 or 8, wherein the analysis or detection is carried out by affinity chromatography, SDS-PAGE, HPLC or mass spectrometry.
10. The method according to any one of claims 7 to 9, wherein the sample is incubated with a sialidase prior to or at the same time as contacting the protein or sample with the polypeptide of claims 1 to 6, preferably wherein said sialidase is Am1757 or a mixture of Am1757 and Am0707.
11. The method according to claim 10, wherein Am1757 is a polypeptide consisting of SEQ ID NO: 11 and/or wherein Am0707 is a polypeptide consisting of SEQ ID NO: 14.

12. A polypeptide which is capable of binding to an O-glycan or O-glycoprotein and which lacks or has reduced endoprotease activity specific for O-glycosylated proteins comprising:

- 5 (a) an amino acid sequence of SEQ ID NO: 5 or SEQ ID NO: 20;
(b) an amino acid sequence which is at least 85% identical to the amino acid sequence of SEQ ID NO: 5 or SEQ ID NO: 20; or
(c) an amino acid sequence which is a fragment of the sequence of SEQ ID NO: 5 or SEQ ID NO: 20 or a fragment of an amino acid which is 85% identical to the
10 amino acid sequence of SEQ ID NO: 5 or SEQ ID NO: 20.

13. The polypeptide according to claim 12, wherein the amino acid sequence which is at least 85% identical to the amino acid sequence of SEQ ID NO: 5 or SEQ ID NO: 20, or the fragment thereof, does not comprise the metalloprotease motif HEBbH and preferably
15 comprises a disrupted version of said motif, such that:

- (a) H in the first position is replaced with an alternative amino acid, preferably A or G; and/or
(b) E in the second position is replaced with an uncharged amino acid, optionally A, C, F, G, I, L, M, N, P, Q, S, T, V or W, preferably A or G;
20 and/or
(c) H in the fifth position is replaced with an alternative amino acid, preferably A or G

wherein b in the said motif is an uncharged amino acid, optionally A, C, F, G, I, L, M, N, P, Q, S, T, V or W.

25 14. The polypeptide according to claim 12 or 13, which includes an additional methionine at the N terminus and/or a His tag at the C terminus, which tag may be joined to the C terminus by a linker, optionally wherein said polypeptide comprises or consists of the amino acid sequence of SEQ ID NO: 6 or 21.

30 15. The polypeptide according to any one of claims 12 to 14, wherein the polypeptide is provided in solution, lyophilised, or immobilised, optionally wherein the polypeptide is

provided together with a sialidase, preferably Am1757 or a mixture of Am1757 and Am0707.

16. A method of binding to an O-glycan, O-glycopeptide and/or O-glycoprotein
5 wherein the method comprises contacting a sample comprising the O-glycan, O-glycopeptide and/or O-glycoprotein with a polypeptide according to any one of claims 12 to 15, and optionally determining whether or not an O-glycan, O-glycopeptide and/or O-glycoprotein has been bound and/or separating the O-glycan and any linked glycoprotein, the O-glycopeptide or O-glycoprotein from the resulting mixture; optionally wherein the
10 method is for the purpose of isolating an O-glycan or linked glycoprotein, O-glycopeptide or O-glycoprotein from the sample.
17. A method for assessing the glycosylation status of a protein, comprising contacting
15 a sample comprising the protein with a polypeptide according to any one of claims 12 to 15 and determining whether or not the protein is bound by the said polypeptide.
18. A method for detecting O-glycopeptides and/or O-glycoproteins in a sample, wherein the method comprises:
- 20 (a) contacting said sample with a polypeptide according to any one of claims 12 to 15 to thereby allow formation of complex between an O-glycopeptide and/or O-glycoprotein and the said polypeptide;
- (b) optionally separating said polypeptide from the contacted sample; and
- (c) determining whether the separated polypeptide is bound to an O-linked
25 glycoprotein or glycopeptide, thereby determining the presence or absence of O-linked glycopeptides or glycoproteins in the sample.
19. The method according to any one of claims 16 to 18, wherein said determining and/or separating is carried out by affinity chromatography, SDS-PAGE, HPLC, lectin
30 blotting, ELISA or mass spectrometry.

20. The method according to any one of claims 16 to 19 which additionally comprises a step of eluting bound material from the polypeptide of any one of claims 12 to 15 with a buffer comprising:

- (a) High molar concentration Urea;
- 5 (b) High concentration detergent; or
- (c) A polypeptide according to any one of claims 1 to 5.