Glycopeptidomics Analysis of a Cell Line Model Revealing Pathogenesis and Potential Marker Molecules for the Early Diagnosis of Gastric MALT Lymphoma

Di Xiao^{a*}, Le Meng^b, Yanli Xu, Huifang Zhang^a, Fanliang Meng^a, Lihua He^a, Jianzhong Zhang^{a*}

^a State Key Laboratory of Infectious Disease Prevention and Control, Collaborative Innovation Center for Diagnosis and Treatment of Infectious Diseases, National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing 102206, China.

^b Center for Biomedical Mass Spectrometry, Department of Biochemistry, Boston University, Boston, MA 02118, USA.

^c Hebei University of Engineering, Affiliated Hospital, College of Medicine, Handan, China

* To whom correspondence should be addressed: Di Xiao, Professor, PhD, Tel/Fax: +86-10-58900754, E-mail: xiaodi@icdc.cn; Jianzhong Zhang, Professor, PhD, Tel/Fax: 86-10-58900707, E-mail: zhangjianzhong@icdc.cn

ABSTRACT

Background: Gastric mucosa-associated lymphoma (GML) is a mature B cell tumor related to *Helicobacter pylori* (*H.pylori*) infection. The clinical manifestations of GML are not specific, so GML commonly escapes diagnosis or is misdiagnosed, leading to excessive treatment. The pathogenesis of *H.pylori*-induced GML is not well understood and there are no molecular markers for early GML diagnosis.

Methods: Glycopeptidomics analyses of host cell lines (a BCG823 cell line, C823) and C823 cells infected by *H. pylori* isolated from patients with GML (GMALT823), gastritis (GAT823), gastric ulcer (GAU823) and gastric cancer (GAC823) were carried out to clarify the host reaction mechanism against GML and identify potential molecular criteria for the early diagnosis of GML.

Findings: Thirty-three samples were analyzed and approximately 2000 proteins, 200 glycoproteins and 500 glycopeptides were detected in each sample. O-glycans were the dominant glycoforms in GMALT823 cells only. Four specific glycoforms in GMALT823 cells and 2 specific glycoforms in C823 and GMALT823 cells were identified. Eight specific glycopeptides of from 7 glycoproteins were found in GMALT823 cells; of these glycopeptides, 6 and 3 specific glycopeptides had high affinity for T cell epitopes and have conformational B cell epitopes, respectively.

Interpretation: The relationship between the predominant glycoforms of host cells and the development of host disease was determined, and the glycoproteins, glycosylation sites and glycoforms might be closely related to the formation of GML, which provides new insight into the pathogenic mechanisms of *H. pylori* infection and suggests molecular indicators for the early diagnosis of GML.

Keywords: Gastric MALT lymphoma; *Helicobacter pylori*; Cell line model; Glycopeptidomics

1. Introduction

Helicobacter pylori (*H. pylori*) is a gram-negative, microaerophilic bacterium adapted for survival in the human stomach, where it can cause chronic gastritis, peptic ulcer disease, gastric mucosa-associated lymphoma (GML) and gastric adenocarcinoma. *H. pylori* has been officially confirmed as a class I carcinogen and remains the only bacterial pathogen established as capable of causing human cancer [1]. Fifty percent of the world's total population is estimated to be infected with *H. pylori* [2]. Mucosa-associated lymphoid tissue (MALT) lymphoma is a low-grade malignant lymphoma. It usually occurs in the stomach, lungs and lacrimal glands, which have no lymphoid tissue. MALT lymphoma is derived from GML [3], a mature B cell tumor related to *H. pylori* infection that accounts for 30%-50% of all extranodal lymphomas and 2-8% of all gastric cancers [4, 5]. *H. pylori* infection is the main factor leading to GML. When GML cells were isolated and cultured by standard methods, all the cells died within 5 days, but following the addition of heat-inactivated whole-cell *H. pylori*, the lymphocytes proliferated [6]. The phenotype of *H. pylori* is conservative, but its gene mutation rate is high. The genetic polymorphism of *H. pylori* is associated with the clinical outcome of *H. pylori* infection [7]. As a form of chronic antigen stimulation, persistent *H. pylori* colonization recruits immune lymphocytes that migrate to and infiltrate the site of *H. pylori* infection in the stomach, leading to the loss of regulation of B-lymphocyte proliferation and differentiation [8]. The clinical manifestations of GML are not specific. Many pathologists are unfamiliar with MALT lymphoma; hence, it is easy for MALT lymphoma to escape diagnosis or be misdiagnosed, and excessive treatment is common [5]. The pathogenesis of *H. pylori*-induced GML is not well understood, although some immunologic mechanisms are thought to be involved [8, 9]. Studies to identify virulence factors or genetic markers of *H. pylori* strains associated with GML have also been performed. Whole-genome comparisons showed that *H. pylori* strains isolated from patients with GML are different from those isolated from patients with gastritis and ulcers [10]. At present, there is no molecular marker for GML diagnosis. The diagnosis of GML mainly depends on pathological diagnosis. Therefore, it is important and of theoretical, clinical, and practical value to clarify the host reaction mechanism and establish molecular criteria for the early diagnosis of GML. Because *H. pylori* strains are difficult to colonize in the stomach of mice, research on *H. pylori*

cannot be completed in animal models. Therefore, with the help of a host cell model, the molecular mechanism of *H. pylori* alone can be examined. More than 90% of known clinical diagnostic markers are glycoproteins. Protein glycosylation is one of the most common and functionally important forms of posttranslational modifications and plays a central role in many biological processes and pathways [11, 12, 13, 14, 15]. In this study, glycopeptidomics analyses of host cell models infected with *H. pylori* from different sources were carried out, and this study aimed to reveal the biological effects of GML-related isolates on host cells and identify potential early diagnostic markers among these glycopeptides.

2. Methods

2.1. Bacterial strains and phenotypic identification

A total of 32 *H. pylori* strains (2 isolated from GML patients, 10 each isolated from gastritis patients, gastric ulcer patients and gastric cancer patients) from the *H. pylori* strain library of the China CDC were used in this study. The *H. pylori* strains were identified by microscopic examination and urease, catalase, and oxidase activity tests. All strains were grown on Columbia blood agar (CM0331, OXOID) plates and inoculated at 37 °C for 24 to 48 h under microaerophilic conditions.

2.2. Host cell line selection

The GES1 human gastric epithelial cell line and the BCG823 human gastric cancer cell line were used as host cell lines. Transcriptomics analysis showed that BCG823 cell lines infected by GML isolates were clustered together, and trends in their upregulated and downregulated differentially expressed genes were similar. However, differential gene expression among GES1 cell lines infected with various *H. pylori* strains was heterogeneous (Supplementary Figure 1). Therefore, the BCG823 cell line was used as the target host cell line and infected with *H. pylori*.

2.3. Sample preparation

2.3.1. Cell line infection

BCG823 cells were cultured in 100 ml/L FBS1640 medium in a 50 ml/L CO₂ gas environment for 48 h at 37 °C. *H. pylori* isolates were inoculated on Columbia agar plates containing 50 ml/L sheep blood and cultured at 37 °C for 72 h in a microaerobic environment with 50 ml/L O₂, 100 ml/L CO₂, and 850 ml/L N₂. Then, the colonies were scraped and collected. Bacteria were added to the cell culture dish at a bacteria/BCG823 cell ratio of 100:1 and allowed to infect BCG823 cells for 4 h at 37 °C. The cell culture media were discarded, and the cells were washed with PBS 3 times at 4 °C

and then scraped into a centrifuge tube. For convenience, the BCG823 cell line and BCG823 cell lines infected with *H. pylori* isolates from the gastric mucosa of patients with gastritis, gastric ulcer, gastric cancer and GML were named C823, GAT823, GAU823, GAC823 and GMALT823 cells, respectively.

2.3.2. Protein extraction

The BCG823 cell line and BCG823 cell lines infected with *H. pylori* were lysed with 8 M urea in 50 mM triethyl ammonium bicarbonate (TEAB) by sonication for 20 s. The protein samples were collected via centrifugation at 16,000 ×g for 10 minutes at 4 °C. The protein concentration of the supernatant was determined using a BCA Protein Assay Kit (Thermo Fisher Scientific).

2.3.3. Trypsin digestion and peptide purification

One-hundred micrograms of protein per condition was transferred into a new tube, and the volume was adjusted to a final volume of 100 µL with 100 mM TEAB. The proteins were reduced by incubation with 200 mM TCEP at 55 °C for 1 h and alkylated by incubation with 375 mM iodoacetamide (IAA) (Thermo Scientific) for 30 minutes while protected from light at room temperature. TEAB (100 mM) was used to adjust the urea concentration to lower than 1 M in all the protein samples, and then the proteins were digested to peptides using trypsin (Promega) at a trypsin/protein ratio of 1:50 (w/w) overnight at 37 °C. The resulting tryptic peptides were dried by speed vacuum at 4 °C and desalted with a C18 spin column (Thermo Scientific) according to the manufacturer's protocol.

2.3.4. Intact glycopeptide enrichment via hydrophilic interaction liquid chromatography (HILIC)

Glycopeptides in the samples were enriched by HILIC (The Nest Group, Inc.). Briefly, the tryptic and desalted peptides were resuspended in 80% ACN. The appropriate amounts of HILIC particle in 80% ACN were placed in Pierce spin columns (Thermo Scientific) and equilibrated three times using 80% ACN, which was followed by sample loading three times and washing two times with 80% ACN. Then, glycopeptides bound to the HILIC column were eluted three times with 100 µL of 0.1% TFA. The samples were dried by a SpeedVac and stored at -80 °C until analysis.

2.3.5. Nano-HPLC-MS/MS analysis

The dried and desalted glycopeptides were reconstituted in 0.1% formic acid (FA) and separated on a nanoAcquity ultraperformance liquid chromatography (UPLC) system and an EASY-nLC 1000 system (Thermo Scientific) fitted with a nanoAcquity Symmetry C18 trap column (100 µm×2 cm, nanoViper C18, 5 µm, 100 Å) and an analytical column (75 µm×15 cm, nanoViper C18, 3 µm, 100 Å). Mobile phase A was 100:0.1 HPLC-grade water/FA, and mobile phase B was 80:20:0.1 ACN/HPLC-grade water/FA. Each sample was loaded on the trapping column at a 2.0 µL/min flow rate and then separated on the analytical column using a 100-minute 3-35% mobile phase B linear

gradient at a 0.8 $\mu\text{L}/\text{min}$ flow rate. A retention time calibration mixture (Thermo Scientific) was used to optimize liquid chromatography (LC) and mass spectrometry (MS) parameters and monitor the stability of the system.

The analytical column was coupled to a high-resolution Q-Exactive Plus mass spectrometer (Thermo Fisher Scientific, San Jose, CA) with a nanoelectrospray ion source operated in positive ion mode. The source was operated at 2.0 kV with the transfer capillary temperature maintained at 250 °C and the S-lens RF level set at 60. MS spectra were obtained by scanning over an m/z range of 350–2000. Mass spectra in both MS and MS/MS were acquired in an Orbitrap mass analyzer with 1 microscan per spectrum. The resolving power for MS and MS/MS was set at 70,000 and 17,500, respectively. Tandem MS data on the top 20 most abundant multiply charged precursors were acquired in parallel with MS data, with higher energy collisional dissociation (HCD) at a normalized collision energy of 30 V. Precursors were isolated using a 2.0 m/z window, and dynamic exclusion of 60 s was enabled during precursor selection.

2.4. Data analysis

2.4.1. Protein and glycopeptide identification

Database searches were performed using Byonic software (v2.13.17, Protein Metrics, Inc.). The following parameters were set for the search: cleavage sites, RK; cleavage side, C-terminal; digestion specificity, fully specific; missed cleavages, 2; precursor mass tolerance, 10 ppm; fragmentation type, QTOF/HCD; fragment mass tolerance, 0.02 Da; modifications, carbamidomethyl and oxidation; glycans, 182 human N-glycans with no multiple fucose moieties in the N-glycan database and 70 human O-glycans in the O-glycan database; and protein false discovery rate (FDR), 1% FDR (or 20 reverse counts). All the other settings were set at their default values.

2.4.2. Glycopeptide, glycoform and glycosylation analysis

Byonic scores reflect the absolute quality of the peptide-spectrum match and not the relative quality compared to other candidate peptides. The Byonic score ranges from 0 to approximately 1000, with 300 being a good score, 400 a very good score, and peptide-spectrum matches with scores over 500 almost certainly correct. The DeltaMod value indicates whether modifications are confidently localized; DeltaMod values over 10 indicate a high likelihood that all modification placements are correct. Therefore, a score over 300, a DeltaMod value over 10, a $q\text{-value} < 0.05$, and an $\text{FDR} < 0.1\%$ were set as thresholds in this study. Systematic and comprehensive analyses of specific glycopeptides, glycoforms and glycosylation sites related to GMALT823 from all the proteins identified by Byonic were carried out.

2.4.3. Ingenuity pathway analysis (IPA)

The commercial software IPA was used to analyze proteins related to the specific glycopeptides. The diseases and functions along with the upstream regulatory factors and downstream regulatory factors of these proteins were analyzed by the BUILD function of IPA with a species selection of human. The OVERLAY tool was used to analyze the canonical pathways and biomarker information for the proteins.

2.4.3. T cell epitope and conformational B cell epitope prediction

T cell epitope prediction

T cell epitopes were identified using prediction tools at the Immune Epitope Database and Analysis Resource (IEDB-AR), a database of experimentally characterized immune epitopes (<http://tools.immuneepitope.org>). T cell epitopes were classified based on their binding affinity for human major histocompatibility complex (MHC) alleles using the half-maximal inhibitory concentration of a biological substance (IC_{50}) as the unit of measure. The threshold was set as follows: peptides with IC_{50} values <50 nM were considered high-affinity peptides, those with IC_{50} values <500 nM were considered intermediate-affinity peptides and those with IC_{50} values <5000 nM were considered low-affinity peptides. MHC class I T cell epitope prediction was performed using the artificial neural network (ANN) method [16], and MHC class II T cell epitope prediction was performed using the consensus method, a combination of the average relative binding matrix method and the stabilization matrix alignment method (SMM-align) [17].

Conformational B cell epitope prediction

SWISS-MODEL Workspace and ElliPro were used for three-dimensional (3D) structure template modeling for selected proteins [18, 19]. A template model of each protein was obtained by submitting protein sequences in FASTA format and modeling those sequences. ElliPro was used to predict conformational B cell epitopes from selected proteins using a modeled 3D structure template for each protein. The default values (a minimum score of 0.5 and a maximum distance of 6 angstroms) were selected for 3D structure prediction. Each predicted epitope formed by a group of amino acid residues was viewed with a Java viewer for chemical structures in 3D to illustrate the epitope 3D structure and relative orientation to the protein molecule.

2.4.4. Other bioinformatic methods

Signal peptides and transmembrane helices were predicted using the online tool Phobius (<http://phobius.sbc.su.se/index.html>). Cartoon representations of proteins and protein features, which were mapped with all the peptides identified from all MS/MS experiments, were generated using the online software tool Protter v.1.0 [20]. The N- and O-glycosylation sites of specific proteins were

predicted using the online software NetNGlyc 1.0 Server and NetOGlyc 4.0 Server [21], respectively.

3. Results

3.1. Protein and glycopeptide identification

The number of total proteins, the number of glycoproteins, the kinds of glycoforms, and the numbers of total peptides and glycopeptides in each sample (33 samples in total) were determined by Byonic analysis and independent comprehensive analysis (Table 1).

Table 1. Summary of glycosylation modifications

Description	Name of cell line or isolate	Number		Number of glycosylated peptides*	Kind of glycoform		
		Protein/Peptide	Glycoprotein (O/N**)/Glycopeptide		N-glycan	O-glycan	Total
Cell line	BCG 823	2323/13899	229 (155/74)/539	134	14	26	40
	MALT	1825/10347	191 (131/60)/470	126	12	29	41
	H879	2571/16322	313 (185/128)/767	297	30	54	84
	liuwanhua	915/3259	125 (64/61)/345	78	15	4	19
	zhanghaiying	835/3045	80 (46/34)/227	37	9	1	10
Gastritis-related <i>H. pylori</i> isolates	HLJ208	1266/4831	176 (93/83)/436	86	17	4	21
	HZ00020	1297/5164	151 (86/65)/312	63	15	4	19
	YN3-77	1137/4346	164 (99/65)/403	98	20	4	24
	HZ00073	1169/4324	144 (80/64)/314	31	9	2	11
	YN3-35	1229/4536	151 (84/67)/367	126	22	12	34
	1-192	1440/8608	57 (42/15)/115	42	13	8	21
	26695	1868/11204	131 (81/50)/322	73	12	8	20
	XA91	576/1777	91 (49/42)/228	32	9	2	11
	XA203	929/3469	92 (48/44)/215	45	9	4	13
	XA137	855/2198	72 (41/31)/205	39	8	2	10
Gastric ulcer-related <i>H. pylori</i> isolates	HZ00124	1119/4589	142 (78/64)/372	89	16	8	24
	YN4-102	1156/4621	136 (81/55)/292	48	10	3	13
	HZ00058	1287/5206	135 (75/60)/381	89	23	5	28
	HLJ253	1387/5791	148 (81/67)/389	107	22	11	33
	wumingqun	1098/3893	137 (69/68)/330	57	14	1	15
	gaosonggui	1138/4355	160 (86/74)/388	74	13	4	17
	YN1-122	1253/4413	173 (96/77)/348	64	12	3	15
	4	1470/6888	168 (88/80)/339	92	16	5	21
	HLJ011	931/3517	102 (52/50)/260	45	11	5	16
	HLJ021	525/1585	62 (34/28)/169	33	9	3	12
Gastric cancer-related <i>H. pylori</i> isolates	Zhou183	1239/4923	165 (85/80)/440	88	21	5	26
	YTC-6	931/3256	134 (69/65)/377	78	13	4	17
	YTC-3	632/3121	71 (40/31)/225	40	11	2	13
	YTC-9	853/2769	120 (67/53)/270	49	11	3	14
	BJWJ-18	884/3174	143 (76/67)/315	61	15	2	17
	BJWJ-11	998/3404	114 (58/56)/272	62	11	5	16
	HZ00066	895/2616	108 (62/46)/247	41	9	4	13
	HLJ005	1180/6551	44 (29/15)/90	34	13	1	14

* Data that did not satisfy the cutoff values in this study (score over 300, DeltaMod value over 10, q-value<0.05, FDR <0.1%) were removed. Some peptides and glycosylation sites appeared more than two times due to the collection of different scans. These peptides were incorporated into one entry. ** The number of glycoproteins, O-glycoproteins and N-glycoproteins.

The enrichment efficiency of glycoproteins and glycopeptides in the C823 cell line infected by *H. pylori* was approximately 12% and 7%, respectively, and the distribution trend of glycoproteins and

glycopeptides was consistent (Figure 1). For C823, GMALT823, GAT823, GAU823 and GAC823 cells, O-glycans were the dominant glycoprotein, and there were more O-glycans than N-glycans. However, the distributions of O- and N-glycans in the samples were different. There were more kinds of glycoforms with O-glycans than those with N-glycans in C823 and GMALT823 cells, and the number of O-glycoforms was 1.9 times and 2.1 times (on average) higher than that of N-glycoforms in C823 and GMALT823 cells, respectively. In contrast, there were more kinds of glycoforms with N-glycans than O-glycans in GAT823, GAU823, and GAC823 cells, and the number of N-glycoforms was 4.3 times (on average) higher than that of O-glycoforms, with the number of total glycoforms significantly lower than that in C823 cells (Table 1, Figure 2).

Figure 1 The enrichment efficiency of glycoproteins and glycopeptides of the BCG823 and BCG823 cell lines infected with *H. pylori*.

Figure 2 Summary of the number of glycoproteins and glycoforms of all the samples used in this study. A: The distribution of the numbers of glycoproteins; there are more O-glycans than N-glycans, and O-glycans are the dominant type of glycoprotein. B: The distribution of the numbers of glycoforms; there are more O-glycans than N-glycans in C823 and GMALT823 cells, but in GAT823, GAU823 and GAC823 cells, the opposite is true.

3.2 Analysis of specific glycopeptides, glycoforms and glycosylation sites

Two new glycosylation sites in GMALT823 cells were observed (these proteins and peptide sequences were both found in C823, GAT823, GAU823, and GAC823 cells but were not glycosylated); these glycosylation sites were the 497T site of protein disulfide isomerase A4 (PDIA4), in which the glycoform was HexNAc(2), and the 71S site of heterogeneous nuclear ribonucleoproteins A2/B1 (HNRNPA2B1), in which the glycoform was HexNAc(1). Two glycoforms of two glycosylation sites were absent in GMALT823 cells (these sites had these glycoforms in C823, GAT823, GAU823, and GAC823 cells, but the glycoforms of these sites were different in GMALT823 cells.); these were the 221N site of cathepsin L1 (CTSL), in which the glycoform was HexNAc(2)Hex(6), and the 69N site of galectin-3-binding protein (LGALS3), in which the glycoform was HexNAc(2)Hex(9) (Table 2, Figures 3, 4).

Compared with the glycosylation sites in C823 cells, four glycosylation sites (O-glycan) in four proteins were consistently found in GMALT823 cells. Protein microtubule-associated tumor

suppressor 1 (MTUS1) was not expressed in GAU823 and GAC823 cells, but in C823 and GMALT823 cells, two glycoforms HexNAc(2)Hex(1)NeuAc(2) and HexNAc(1)Hex(1)NeuAc(1) at the 734S site were stable. The 259S site of the HNRNPA2B1 protein, which was deglycosylated in GAT823, GAU823 and GAC823 cells, was the only glycosylation site with the glycoform HexNAc(2)Hex(2)Fuc(2)NeuAc(1) in C823 and GMALT823 cells. The glycoform of the 45S site of protein calnexin (CANX) was different or deglycosylated in GAT823, GAU823 and GAC823 cells, but the glycoform HexNAc(1)Hex(1)NeuAc(2) at this site was stable in C823 and GMALT823 cells. The glycoform of the 271T site of nuclease-sensitive element-binding protein 1 (YBX1) was different in GAT823, GAU823 and GAC823 cells, but the glycoform HexNAc(1)Hex(2)Fuc(1) was stable in C823 and GMALT823 cells (Table 2, Figures 3, 4). The glycosylation sites 497T of PDIA4 and the 71S and 259S sites of HNRNPA2B1 were previously unreported glycosylation sites found in this study (Table 2).

Table 2 Summary of specific glycosylation modifications in the BCG823 cell line infected with MALT lymphoma-related *H. pylori* isolates

Protein Name	Gene Name	Sequence	Glycoform	Description	Score	Delta Score
>sp Q9ULD2 Microtubule-associated tumor suppressor 1	MTUS1	K.VGPPVS*[+1150.4024]C[+57.0214]LR.R	HexNAc(2)Hex(1)N euAc(2)	This protein was not expressed in GAU823 and GAC823 cells. This site and glycoform were only found in C823 and GMALT823 cells.	434.5	35.6
>sp P22626 Heterogeneous nuclear ribonucleoproteins A2/B1	HNRNPA2B1	K.VGPPVS*[+656.2276]C[+57.0214]LR.R	HexNAc(1)Hex(1)N euAc(1)	This site was deglycosylated in GAT823, GAU823 and GAC823 cells.	464	39.7
>sp P27824 Calnexin	CANX	G.GNFGGS*[+1313.47563]PGYGGGR.G	HexNAc(2)Hex(2)F uc(2)NeuAc(1)	The glycoform of this site was changed or this site was deglycosylated in GAT823, GAU823 and GAC823 cells.	466.4	61.2
>sp P67809 Nuclease-sensitive element-binding protein 1	YBX1	V.IDIEDLDLVIEVEDS*[+947.3230]KPDTTAPPSP K.V	HexNAc(1)Hex(1)N euAc(2)	The glycoform of this site was different in GAT823, GAU823 and GAC823 cells.	635.5	61.3
>sp P13667 Protein disulfide isomerase A4	PDIA4	E.DKENQGET*[+673.2429]QGQPPQRR.Y	HexNAc(1)Hex(2)F uc(1)	This protein and sequence were found in all samples, but glycosylation of this site was found in only GMALT823 cells.	612.2	474.5
>sp P22626 Heterogeneous nuclear ribonucleoproteins A2/B1	HNRNPA2B1	M.EPEEFSDT**[+406.15875]LREFVTAFK.K	HexNAc(2)	All samples had this protein and sequence, but only GMALT823 cells lacked this glycoform on this glycosylation site.	309.3	195.9
>sp P07711 Cathepsin L1	CTSL	F.GFVTFSS*[+203.07937]MAEVDAAAMARPHSIDGR .V	HexNAc(1)		556.1	389.0
>sp Q08380 Galectin-3-binding protein	LGALS3	K.YSVAN*[+1378.47569]DTGFVDIPKQEK.A	HexNAc(2)Hex(6)		543.8	355.0
		R.ALGFEN*[+1864.63416]ATQALGR.A	HexNAc(2)Hex(9)		524.5	257.6

* indicates that this site is a glycosylation site by online prediction. ** indicates that this site is not a glycosylation site by online prediction. The modifications with a gray background are stable glycosylation modifications in C823 and GMALT823 cells. The

modifications with a white background are specific modifications in GMALT823 cells.

Figure 3 Cartoon schematics of glycoproteins containing specific glycosylated peptides related to gastric MALT lymphoma. The glycosylation sites 497T of PDIA4 and 71S of HNRNPA2B1 were present in only GMALT823 cells. The glycoforms HexNAc(2)Hex(6) and HexNAc(9) were absent at the glycosylation sites 221N of CTSL and 69N of LGALS3 in MALT823 cells, respectively. The glycosylation sites 734S of MTUS1, 45S of CANX, 271T of YBX1, and 259S of HNRNPA2B1 were stable in C823 and GMALT823 cells. The MTUS1 protein could not be expressed in GAT823, GAU823 and GAC823 cells. Glycosylation sites and glycoforms in the blue, yellow and green frames are present in only GMALT823 cells, absent in GMALI823 cells and steady in GAMALT823 and C823 cells, respectively.

Figure 4 Linear glycan sequence shown by HCD MS/MS fragmentation of the ions on a QE-HF. The correspondence of ion, glycopeptide sequences in the mass spectrum: A1: m/z 642.5522 [z=4], M.EPEEFDSDT[+406.15875]LREFVTFK. K, protein PDIA4; A2: m/z 707.8376 [z=4], F.GFVTFSS[+203.07937]MAEVDAAAMAARPHSIDGR. V, protein HNRNPA2B1; A3: m/z 1249.0179 [z=2], K.YSVAN[+1378.47569]DTGFVDIPKQEK. A, protein CTSL; A4: m/z 1071.7864 [z=3], R.ALGFEN[+1864.63416]ATQALGR. A, protein LGALS3BP; B1: m/z 712.3184 [z=3], K.VGPPVS[+1150.4024]C[+57.0214]LR. R, protein MTUS1; B2: 820.8885 [z=2], K.VGPPVS[+656.2276]C[+57.0214]LR. R, protein MTUS1; B3: m/z 1249.0179 [z=1], G.GNFGGS[+1313.47563]PGYGGGR. G, protein HNRNPA2B1; B4: m/z 1030.4617 [z=4], V.IDIEDDLDDVIEEVEDS[+947.3230]KPDTTAPPSSPK. V, protein CANX; B5: m/z 697.0580 [z=4], E.DKENQGDET[+673.2429]QGQPPQRR. Y, protein YBX1. * represents an oxonium ion

3.3. IPA analysis

The 7 proteins related to specific glycopeptides were analyzed using IPA. The related diseases and functions, upstream regulatory factors and downstream regulatory factors, canonical pathways, biomarkers, roles in the cell, Entrez Gene summary and molecular function of these proteins are listed in Supplementary Table 1.

3.4. T cell epitope and conformational B cell epitope prediction

For T cell epitope prediction, the predicted output was given in units of IC₅₀ nM for combinatorial library and SMM_align. Therefore, a lower number indicated a higher affinity. Here, peptides with IC₅₀ values <50 nM (high affinity) and <500 nM (intermediate affinity) are listed in Table 3. The glycopeptide MEPEEFDSDTLREFVTFKK from protein PDIA4 was predicted to have high-affinity and intermediate-affinity MHC class I epitopes and intermediate-affinity MHC class II epitopes. The glycopeptide FGFVTFSSMAEVDAAAMAARPHSIDGRV from the HNRNPA2B1 protein was predicted to have high-affinity and intermediate-affinity MHC class I and MHC class II epitopes. The glycopeptide GGNFGGSPGYGGGRG from the HNRNPA2B1 protein was predicted to have high-affinity and intermediate-affinity MHC class II epitopes. The specific glycopeptide sequences from the CTSL, LGALS3, and MTUS1 proteins were predicted to have intermediate-affinity MHC class II epitopes.

Table 3 High-affinity and intermediate-affinity MHC class I and II T cell epitopes of specific glycopeptides

Protein	Gene	Glycopeptide Sequence	MHC Class I T cell epitope		MHC Class II T cell epitope	
			Sequence (Positions) / Affinity	Alleles	Sequence (Positions) / Affinity	Alleles
P13667	PDIA4	MEPEEFDSDTLREFVTFKK	EEFDSDTLR (481-489) / H EEFDSDTLR (481-489) / I	HLA-A*68:01 HLA-A*33:01	EPEEFDSDTLREFVT(489-503) / I EFDSDTLREFVTAFK(492-506) / I	HLA-DRB3*01:01 HLA-DRB5*01:01, HLA-DRB3*01:01, HLA-DRB1*04:05 HLA-DRB1*01:01
P22626	HNRNP A2B1	FGFVTFSSMAEVDAAAMAARPHSIDGRV	SMAEVDAAAMA(71-80) / H SMAEVDAAAMA(71-80) / I	HLA-B*35:01, HLA-A*02:03 , HLA-B*15:01 HLA-B*15:01 HLA-A*02:03 HLA-A*02:06 HLA-A*02:01 HLA-A*68:02	GFGEVTFSSMAEVDAA(63-77) / H GFGEVTFSSMAEVDAA(63-77) / I FVTFSSMAEVDAAAMA(66-80) / I GFVTFSSMAEVDAAAM (65-79) / I FGFVTFSSMAEVDAA(64-78) / I GFVTFSSMAEVDAAAM(65-79) / I GPGGGNFGGSPGYGG(250-264) / H GNFGGSPGYGGGRG(254-268) / I GPGGGNFGGSPGYGG(250-264) / I GNFGGSPGYGGGRG(253-267) / I	HLA-DRB1*04:05, HLA-DRB1*01:01, HLA-DRB1*04:01, HLA-DRB1*09:01, HLA-DQA1*05:01/ DQB1*03:01 HLA-DRB1*01:01, HLA-DRB1*09:01 HLA-DQA1*05:01/ DQB1*03:01, HLA-DQA1*01:02/ DQB1*06:02 HLA-DQA1*05:01/ DQB1*03:01 HLA-DQA1*01:02/ DQB1*06:02 HLA-DQA1*05:01/ DQB1*03:01 HLA-DRB1*01:01, HLA-DRB1*01:01 HLA-DQA1*05:01/
		GGNFGGSPGYGGGRG				

P07711	CTSL	KYSVA N DTGFVDI PKQEKA	<u>VA</u> N DTGFVDIPKQEK(219-233) /I	DQB1*03:01 HLA-DRB1*11:01, HLA-DRB5*01:01
			NPKYSVA N DTGFVDI(214-228) /I	HLA-DRB1*13:02
			KYNPKYSVA N DTGFV(212-226) /I	HLA-DRB1*13:02
Q08380	LGALS 3BP	RALGFENATQALG RA	<u>LG</u> FENATQALGRAAF(65-79) /I	HLA-DRB1*01:01, HLA-DQA1*05:01/ DQB1*03:01, HLA-DQA1*01:02/ DQB1*06:02
			VCRALGFENATQALG(61-75) /I	HLA-DQA1*05:01/ DQB1*03:01
			VCRALGFENATQALG(61-75) /I	HLA-DRB1*04:01
			VCRALGFENATQALG(61-75) /I	HLA-DRB1*01:01
Q9ULD2	MTUS1	KVGPPV S CLRR	IAAPKAKVGPPVSCL(722-736) /I	HLA-DQA1*05:01/ DQB1*03:01
			VGPPV S CLRRNSDNR(729-743) /I	HLA-DRB1*11:01
			GPPV S CLRRNSDNRN(730-744) /I	HLA-DRB1*13:02
			PPV S CLRRNSDNRNP(731-745) /I	HLA-DRB1*11:01

H: high-affinity binding, IC₅₀ of <50 nM; I: intermediate-affinity binding, IC₅₀ of <500 nM; L: low-affinity binding, IC₅₀ of <5000 Mm. A lower IC₅₀ is indicative of a higher binding affinity to host MHC alleles. The underlined residues are the smm_align_core sequence. The gray amino acids are glycosylation sites.

For conformational B cell epitope prediction, neither program found a suitable model for the specific glycopeptide sequences from the HNRNPA2B1, MTUS1, CANX and YBX1 proteins. The specific glycopeptide sequences MEPEEFDSDTLREFVTAFKK, KYSVANDTGFVDIPKQEKA and RALGFENATQALGRA from the PDIA4, CTSL, and LGALS3 proteins, respectively, were predicted to have conformational B cell epitopes (Table 4). The 3D structures of B cell epitopes in specific glycopeptide sequences and model-template alignments of the PDIA4, CTSL, and LGALS3 proteins are shown in Figure 5.

Table 4 Conformational B cell epitopes of specific glycopeptides

Protein	Gene	Glycopeptide Sequence	Conformational B Cell Epitopes		
			Peptides	Score	PDB ID
P13667	PDIA4	MEPEEFDSDTLREFVTAFKK	PEEFDSDTLRE (490-500)	0.606	5wwf.2
P07711	CTSL	KYSVA N DTGFVDIPKQEKA	QYVQDNGGLDSEESYPYEAT EESCKYNPKYSVANDTG (188-224)	0.681	1cs8.1
Q08380	LGALS3BP	RALGFENATQALGRA	LGFE N A (65-70)	0.557	1by2.1

The gray amino acids are glycosylation sites.

Figure 5 Conformational B cell epitopes of specific glycosylated peptides. In the ball-and-stick models (A), yellow balls are the residues of the predicted epitopes, and white sticks are the structures of nonepitopes and core residues. In the model-template alignment (B), each epitope is shown with predicted residues and residue positions: P13667, PEEFDSDTLRE (490-500); P07711, QYVQDNGGLDSEESYPYEATEESCKYNPKYSVANDTG (188-224); and Q08380, LGFENA (65-70). The amino acid in the green frame is the glycosylation site.

4. Discussion

Histological examination, immunohistochemistry and clonal analysis of B cells can be used for the pathological diagnosis of GML [22]. To diagnose small lymphomas, which cannot be detected by the naked eye under endoscopy, the "observe and wait" method is usually adopted, after which a diagnosis can be made when the lymphoma is observed [23]. Unlike other gastric cancers, early GML is a reversible tumor, and the eradication of *H. pylori* at the early stage of GML causes 60-80% of MALT lymphomas to subside [5, 24]. Early detection and treatment of GML can greatly reduce the burden of this disease. Therefore, identifying molecular markers for the early diagnosis of GML has important clinical and social value.

Glycoforms are related to specific biological properties. Certain glycoforms are mass-produced in a disease-specific manner [25]. Aberrant protein glycosylation has been observed during cancer development and progression. Given the functional role of aberrant glycosylation in cancer, many researchers have attempted to utilize specific glycoforms as cancer biomarkers for various clinical purposes [25, 26]. This study showed that the dominant glycoforms in the BCG823 cell line both before and after *H. pylori* infection are O-glycans. However, the glycoform type was different before and after infection with *H. pylori* from patients with different diseases. Before and after infection with MALT lymphoma isolates, the BCG823 cell line expressed more O-glycan glycoforms than N-glycan glycoforms; however, after infection with isolates from patients with gastritis, gastric ulcer and gastric cancer, the dominant glycoforms in BCG823 cell lines were N-glycan glycoforms. These results suggest that a predominance of N-glycans in host cells may lead to the development of gastritis, gastric ulcers and gastric cancer, while a predominance of O-glycans in host cells may lead to the development of MALT lymphoma. In addition, the differentially and stably expressed glycoforms in GMALT823 cells suggest that if the HexNAc(2)Hex(6) glycoform on the 221N site of the CTSL protein or/and the HexNAc(2)Hex(9) glycoform on the 69N site of the LGALS3 protein are detected in host cells, gastric disease will not develop into MALT lymphoma. If the HexNAc(2) glycoform at the 497T site of the PDIA4 protein and/or the HexNAc(1) glycoform at the 71S site of HNRNPA2B1 are detected, the host cell tends to develop MALT lymphoma. If the HexNAc(2)Hex(2)Fuc(2)NeuAc(1) glycoform at the 259S site of HNRNPA2B1, the HexNAc(1)Hex(1)NeuAc(2) glycoform at the 45S site of CANX, the HexNAc(1)Hex(2)Fuc(1) glycoform at the 271T site of YBX1, and the two HexNAc(2)Hex(1)NeuAc(2) and HexNAc(1)Hex(1)NeuAc(1) glycoforms at the 734S site of MTUS1 are detected, the host cell may not develop or trend towards developing MALT lymphoma. Therefore, the predominant type of glycoforms and the 9 glycoforms of 8 specific glycopeptides in host cells identified in this study are potential molecular markers to predict the developmental course of disease from *H. pylori* infection. The pathogenesis of GML is not yet clear [22]. Previous studies have elucidated some immunological and virulence factors and gene marker-related mechanisms involved in GML [4, 27, 28, 29]. Among the eight specific glycopeptides in seven proteins found in this study, six T cell receptor epitopes (Table 3) and three conformational B cell epitopes (Table 4) mediate cellular and humoral immunity, respectively. Changes in specific glycosylation sites and glycoforms caused by

infection with MALT lymphoma-related *H. pylori* isolates may affect the binding of T/B cell epitopes, thus positively or negatively affecting cellular or humoral immunity.

In this study, seven specific glycoproteins (MTUS1, HNRNPA2B1, CANX, YBX1, PDIA4, CTSL and LGALS3) were found in GMALT823 cells; these proteins can all be detected in bodily fluids (Supplementary Table 1).

MTUS1 is a tumor suppressor in lung cancer that promotes cell proliferation and migration [30]. MTUS1 has a significant impact on the proliferation and metastatic potential of gastric cancer cell lines, as it has shown a potential anticancer effect in gastric cancer cell lines [31]. In this study, stable glycopeptides of the MTUS1 protein were found in host cells and host cells infected by MALT lymphoma isolates, but these glycopeptides did not express or exhibited specific glycopeptides in host cells infected by other *H. pylori* isolates. The specific glycopeptides and glycosylation sites of MTUS1 may play an important role in the development of MALT lymphoma.

PDIA4 is also an important glycoprotein that was examined in our study. Studies have shown that PDIA4 facilitates tumor cell growth via inhibiting the degradation and activation of procaspases 3 and 7 via their mutual interaction in a CGHC-dependent manner [32]. PDIA4 regulates MTPP (Supplementary Table1), a major and an essential lipid transfer protein that transfers phospholipids and triacylglycerols to nascent ApoB for the assembly of lipoproteins [33]. Clinical studies have shown that the level of ApoB/ApoA1 in gastric cancer patients is negatively correlated with the survival rate, while PDIA4 is positively correlated with the survival rate of gastric cancer patients^[34]. In this study, a new specific 497T glycosylation site with the glycoform HexNAc(2) was observed in host cells infected by MALT lymphoma isolates, indicating that this glycosylation site may be an important factor in the development of MALT lymphoma in host cells.

HNRNPA2B1 plays a direct role in cancer development, cancer progression, gene expression, and signal transduction. Knockdown of HNRNPA2B1 reduced breast cancer cell proliferation, induced apoptosis, and prolonged the S phase of the cell cycle in vitro. In addition, HNRNPA2B1 knockdown suppressed subcutaneous tumorigenicity in vivo [35]. HNRNPA2B1 can specifically bind to the COX-2 core promoter and regulate the growth of non-small-cell tumors [36]. In this study, a new specific 71S glycosylation site with the glycoform HexNAc(1) was observed in the HNRNPA2B1 protein from host cells infected by MALT lymphoma isolates, and the 259S site with the glycoform HexNAc(2)Hex(2)Fuc(2)NeuAc(1) was expressed steadily in only host cells and host cells infected

by MALT lymphoma isolates, suggesting that these two specific glycosylation sites play an important role in the pathogenesis of MALT lymphoma caused by *H. pylori* infection.

Our study showed the deletion of the glycoform HexNAc(2)Hex(6) on the 221N site of CTSL and the glycoform HexNAc(2)Hex(9) on the 69N site of LGALS3 in host cells after infection with MALT lymphoma isolates. These changes were specific for MALT lymphoma *H. pylori* isolates and not observed for *H. pylori* isolates from patients with gastritis, gastric ulcer and gastric cancer. CTSL is involved in autophagy- and phage maturation-related pathways and associated with many cancers, such as renal cancer, skin cancer, soft tissue sarcoma, and colon cancer, and immune diseases (Supplementary Table 1). CTSL expression may be linked to cancer grade, invasion and stage [37]. LGALS3 binds Galectin-3 (Gal3), which regulates its function. Gal3 plays an important role in innate immunity against infection and the colonization of *H. pylori*. Large lymphoid clusters consisting of mostly B cells have frequently been observed in the gastric submucosa of Gal3-deficient mice [38], which is consistent with our results. Changes in the glycoforms on the glycosylation site of LGALS3 may lead to changes in its biological function, which may decrease the expression of Gal3 or cause a loss of biological function, resulting in B lymphocyte aggregation. This result suggests that deletion of the 221N glycosylation site of CTSL and the 69N glycosylation site of LGALS3 may be important factors in the development of MALT lymphoma.

CANX is a prognostic marker and potential therapeutic target in colorectal cancer [39]. In melanoma models, CANX knockout enhanced the infiltration and effector functions of T cells in the tumor microenvironment and inhibited tumor growth [40]. YBX1 is involved in the antigen presentation pathway and phage maturation, which is related to gastric cancer and gastric epithelial cancer (Supplementary Table 1). YBX1 binds to HOXC-AS3 to mediate the tumorigenesis of gastric cancer [41]. The YBX1 protein is associated with cancer proliferation in numerous tissues, and its gene may be a prognostic marker for poor outcome and drug resistance in certain cancers (Supplementary Table 1). In this study, the glycoforms of the 45S site of CANX and the 271T site of YBX1 were stable in host cells infected with *H. pylori* isolates from patients with MALT lymphoma; however, glycoforms in these glycosylation sites changed or were deglycosylated in host cells infected with *H. pylori* isolates from patients with gastritis, gastric ulcers and gastric cancer. Changes in the glycosylation sites of these two glycoproteins may be related to the development of MALT lymphoma in cells.

Missed diagnosis and misdiagnosis of GML are common in the clinic. The overtreatment of early GML, such as treatment with systemic chemotherapy, occurs in China and many developing countries. Therefore, it is of great theoretical and clinical value to clarify the host response

mechanism and identify and establish molecular markers for the early diagnosis of GML. In this study, the relationship between the predominant glycoform of host cells and the development of host disease was determined by glycopeptidomics analysis of a cell line model. Seven glycoproteins, eight glycosylation sites and 9 glycoforms might be closely related to the formation of GML, which provides new insight into the pathogenic mechanisms of *H. pylori* infection and suggests molecular indicators for the early diagnosis of GML.

Acknowledgements

We thank Joseph Zaia, Cheng Lin and Catherine E. Costello of Center for Biomedical Mass Spectrometry, Department of Biochemistry, Boston University for their assistance with the glycomics technical guidance.

Funding sources

This work was supported by Major Infectious Diseases Such as AIDS and Viral Hepatitis Prevention and Control Technology Major Projects (Grant No. 2018ZX10712-001, and Grant No. 2018ZX10733-402).

Author contributions

Di Xiao, Jianzhong Zhang conceived the study. Di Xiao, Le Meng, Yanli Xu, Huifang Zhang, Fanliang Meng and Lihua He completed the bacterial and cell cultures, sample preparation and data acquisition. Di Xiao, Le Meng did data analysis and bioinformatics. Di Xiao wrote the manuscript. All authors reviewed the final manuscript.

Declaration of Competing Interests

The authors have declared that no competing interests exist.

REFERENCES

- [1] Deguchi R, Takagi A, Kawata H, Inoko H, Miwa T. Association between CagA+ *Helicobacter pylori* infection and p53, bax and transforming growth factor-beta-II gene mutations in gastric cancer patients. *Int J Cancer* 2001;91:481-5.
- [2] Deng B, Li Y, Zhang Y, Bai L, Yang P. *Helicobacter pylori* infection and lung cancer: a review of an emerging hypothesis. *Carcinogenesis* 2013;34:1189-95.
- [3] Dogan A, Du M, Koulis A, Briskin MJ, Isaacson PG. Expression of lymphocyte homing receptors and vascular addressins in low-grade gastric B-cell lymphomas of mucosa-associated lymphoid tissue. *Am J Pathol* 1997;151:1361-9.
- [4] Zullo A, Hassan C, Ridola L, Repici A, Manta R, Andriani A. Gastric MALT lymphoma: old and new insights. *Ann Gastroenterol* 2014;27:27-33.
- [5] Moleiro J, Ferreira S, Lage P, Dias Pereira A. Gastric malt lymphoma: Analysis of a series of consecutive patients over 20 years. *United European Gastroenterol J* 2016;4:395-402.
- [6] Raderer M, Kiesewetter B, Ferreri AJ. Clinicopathologic characteristics and treatment of marginal zone lymphoma of mucosa-associated lymphoid tissue (MALT lymphoma). *CA Cancer J Clin* 2016;66:153-71.
- [7] Saribasak H, Salih BA, Yamaoka Y, Sander E. Analysis of *Helicobacter pylori* genotypes and correlation with clinical outcome in Turkey. *J Clin Microbiol* 2004;42:1648-51.
- [8] de Jong D, Aleman BM, Taal BG, Boot H. Controversies and consensus in the diagnosis, work-up and treatment of gastric lymphoma: an international survey. *Ann Oncol* 1999;10:275-80.
- [9] Du MQ, Isaacson PG. Gastric MALT lymphoma: from aetiology to treatment. *Lancet Oncol* 2002;3:97-104.
- [10] Wang HC, Cheng FC, Wu MS, Shu HY, Sun HS, Wang YC, et al. Genome Sequences of Three *Helicobacter pylori* Strains from Patients with Gastric Mucosa-Associated Lymphoid Tissue Lymphoma. *Genome Announc* 2015;3.
- [11] Hoseki J, Ushioda R, Nagata K. Mechanism and components of endoplasmic reticulum-associated degradation. *J Biochem* 2010;147:19-25.
- [12] Liu YC, Yen HY, Chen CY, Chen CH, Cheng PF, Juan YH, et al. Sialylation and fucosylation of epidermal growth

- factor receptor suppress its dimerization and activation in lung cancer cells. *Proc Natl Acad Sci U S A* 2011;108:11332-7.
- [13] Napoletano C, Rughetti A, Agervig Tarp MP, Coleman J, Bennett EP, Picco G, et al. Tumor-associated Tn-MUC1 glycoform is internalized through the macrophage galactose-type C-type lectin and delivered to the HLA class I and II compartments in dendritic cells. *Cancer Res* 2007;67:8358-67.
- [14] Dube DH, Bertozzi CR. Glycans in cancer and inflammation--potential for therapeutics and diagnostics. *Nat Rev Drug Discov* 2005;4:477-88.
- [15] Pinho SS, Reis CA. Glycosylation in cancer: mechanisms and clinical implications. *Nat Rev Cancer* 2015;15:540-55.
- [16] Tenzer S, Peters B, Bulik S, Schoor O, Lemmel C, Schatz MM, et al. Modeling the MHC class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class I binding. *Cell Mol Life Sci* 2005;62:1025-37.
- [17] Wang P, Sidney J, Dow C, Mothe B, Sette A, Peters B. A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. *PLoS Comput Biol* 2008;4:e1000048.
- [18] Arnold K, Bordoli L, Kopp J, Schwede T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* 2006;22:195-201.
- [19] Ponomarenko J, Bui HH, Li W, Fusseder N, Bourne PE, Sette A, et al. ElliPro: a new structure-based tool for the prediction of antibody epitopes. *BMC Bioinformatics* 2008;9:514.
- [20] Haserick JR, Klein JA, Costello CE, Samuelson J. Cryptosporidium parvum vaccine candidates are incompletely modified with O-linked-N-acetylgalactosamine or contain N-terminal N-myristate and S-palmitate. *PLoS One* 2017;12:e0182395.
- [21] Steentoft C, Vakhrushev SY, Joshi HJ, Kong Y, Vester-Christensen MB, Schjoldager KT, et al. Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. *EMBO J* 2013;32:1478-88.
- [22] Hu Q, Zhang Y, Zhang X, Fu K. Gastric mucosa-associated lymphoid tissue lymphoma and Helicobacter pylori infection: a review of current diagnosis and management. *Biomark Res* 2016;4:15.
- [23] Doglioni C, Ponzoni M, Ferreri AJ, Savio A, Gruppo Italiano Patologi Apparato D, Societa Italiana di Anatomia Patologica e Citopatologia Diagnostica/International Academy of Pathology Id. Gastric lymphoma: the histology report. *Dig Liver Dis* 2011;43 Suppl 4:S310-8.
- [24] Nobre-Leitao C, Lage P, Cravo M, Cabecadas J, Chaves P, Alberto-Santos A, et al. Treatment of gastric MALT lymphoma by Helicobacter pylori eradication: a study controlled by endoscopic ultrasonography. *Am J Gastroenterol* 1998;93:732-6.
- [25] Lee JH, Cho CH, Kim SH, Kang JG, Yoo JS, Chang CL, et al. Semi-quantitative measurement of a specific glycoform using a DNA-tagged antibody and lectin affinity chromatography for glyco-biomarker development. *Mol Cell Proteomics* 2015;14:782-95.
- [26] Kim YS, Hwang SY, Kang HY, Sohn H, Oh S, Kim JY, et al. Functional proteomics study reveals that N-Acetylglucosaminyltransferase V reinforces the invasive/metastatic potential of colon cancer through aberrant glycosylation on tissue inhibitor of metalloproteinase-1. *Mol Cell Proteomics* 2008;7:1-14.
- [27] Munari F, Lonardi S, Cassatella MA, Doglioni C, Cangi MG, Amedei A, et al. Tumor-associated macrophages as major source of APRIL in gastric MALT lymphoma. *Blood* 2011;117:6612-6.
- [28] Lin WC, Tsai HF, Kuo SH, Wu MS, Lin CW, Hsu PI, et al. Translocation of Helicobacter pylori CagA into Human B lymphocytes, the origin of mucosa-associated lymphoid tissue lymphoma. *Cancer Res* 2010;70:5740-8.
- [29] Saito Y, Suzuki H, Tsugawa H, Imaeda H, Matsuzaki J, Hirata K, et al. Overexpression of miR-142-5p and miR-155 in gastric mucosa-associated lymphoid tissue (MALT) lymphoma resistant to Helicobacter pylori eradication. *PLoS One* 2012;7:e47396.
- [30] Gu Y, Liu S, Zhang X, Chen G, Liang H, Yu M, et al. Oncogenic miR-19a and miR-19b co-regulate tumor

- suppressor MTUS1 to promote cell proliferation and migration in lung cancer. *Protein Cell* 2017;8:455-66.
- [31] Li X, Liu H, Yu T, Dong Z, Tang L, Sun X. Loss of MTUS1 in gastric cancer promotes tumor growth and metastasis. *Neoplasia* 2014;61:128-35.
- [32] Kuo TF, Chen TY, Jiang ST, Chen KW, Chiang YM, Hsu YJ, et al. Protein disulfide isomerase $\alpha 4$ acts as a novel regulator of cancer growth through the procaspase pathway. *Oncogene* 2017;36:5484-96.
- [33] Sirwi A, Hussain MM. Lipid transfer proteins in the assembly of apoB-containing lipoproteins. *J Lipid Res* 2018;59:1094-102.
- [34] Ma MZ, Yuan SQ, Chen YM, Zhou ZW. Preoperative apolipoprotein B/apolipoprotein A1 ratio: a novel prognostic factor for gastric cancer. *Onco Targets Ther* 2018;11:2169-76.
- [35] Hu Y, Sun Z, Deng J, Hu B, Yan W, Wei H, et al. Splicing factor hnRNPA2B1 contributes to tumorigenic potential of breast cancer cells through STAT3 and ERK1/2 signaling pathway. *Tumour Biol* 2017;39:1010428317694318.
- [36] Xuan Y, Wang J, Ban L, Lu JJ, Yi C, Li Z, et al. hnRNPA2/B1 activates cyclooxygenase-2 and promotes tumor growth in human lung cancers. *Mol Oncol* 2016;10:610-24.
- [37] Sudhan DR, Siemann DW. Cathepsin L targeting in cancer treatment. *Pharmacol Ther* 2015;155:105-16.
- [38] Park AM, Hagiwara S, Hsu DK, Liu FT, Yoshie O. Galectin-3 Plays an Important Role in Innate Immunity to Gastric Infection by *Helicobacter pylori*. *Infect Immun* 2016;84:1184-93.
- [39] Ryan D, Carberry S, Murphy AC, Lindner AU, Fay J, Hector S, et al. Calnexin, an ER stress-induced protein, is a prognostic marker and potential therapeutic target in colorectal cancer. *J Transl Med* 2016;14:196.
- [40] Chen Y, Ma D, Wang X, Fang J, Liu X, Song J, et al. Calnexin Impairs the Antitumor Immunity of CD4(+) and CD8(+) T Cells. *Cancer Immunol Res* 2019;7:123-35.
- [41] Zhang E, He X, Zhang C, Su J, Lu X, Si X, et al. A novel long noncoding RNA HOXC-AS3 mediates tumorigenesis of gastric cancer by binding to YBX1. *Genome Biol* 2018;19:154.

◆ Enrichment efficiency glycoproteins ■ Enrichment efficiency of glycopeptides — Trendline

● specific glycosylation site (predicted site)
 ● specific glycosylation site(not prdicted site)
 ● predicted O-glycosylation site
 ● predicted N-glycosylation
 ● signal peptide
 ● specific glycopeptide
 ● coverage sequence
 — Tryps

P13667
PDIA4

497T-HexNAc(2)

P22626
HNRNPA2B1

718S-HexNAc(1)

259S-HexNAc(2)Hex(2)Fuc(2)NeuAc(1)

P07711
CTSL

221N-HexNAc(2)Hex(6)

69N-HexNAc(9)

Q08380
LGALS3

P27824
CANX

458S-HexNAc(1)Hex(1)NeuAc(2)

P67809
YBX1

271T-HexNAc(1)Hex(2)Fuc(1)

Q9ULD2
MTUS1

734S-HexNAc(2)Hex(1)NeuAc(2)

734S-HexNAc(1)Hex(1)NeuAc(1)

A

B

B5

A

B

bioRxiv preprint doi: <https://doi.org/10.1101/2020.06.02.126854>; this version posted June 2, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder. All rights reserved. No reuse allowed without permission.

```
Model_01 MRPRKAFLLLLLLGLVQLLAVAGAEGPDEDSSNRENAIEDEEEEEEDDDDEEDDLEVKEENGVLVLNDANFDNFVADKDTVLLEFYAPWCGHCKQFAPEYEKIA 105
3f8u.1.A -----
Model_01 NILKDKDPPIPVAKIDATSASVLASRFVDSGYPTIKILKKQAVDYEGSRTQEEIVAKVREVSQPDWTTPPEVTLVLTKENFDEVND ---ADIILVEFYAPWCG 207
3f8u.1.A -----VLELTDDNFESRISDTGSAGLMLVEFAPWCG 34
Model_01 HCKKLAPEYEKA AKELSKRSPPIPLAKVDATAETDLAKRFDVSGYPTLKIPRKGRPYD YNGPREKYGIVDYMIEQSGPPSKEILT LKQVQEF LKDGDDVIIIGV 311
3f8u.1.A HAKRLAPEYEAAATL LKG ---IVPLAKVD TAN(TNTCN)YGVSGYPTLKIPRKGRPYD YNGPREKYGIVDYMIEQSGPPSKEILT LKQVQEF LKDGDDVIIIGV 135
Model_01 FKGESDPAYQQYQDAANNLRDYKFHHTFSTEIAKFLKVSQGLVVMQPEKQSKYEPRSHMMDVQGSTODSAIKDFVLKYALPLVGHKVSND AKRYTR RPLV 415
3f8u.1.A DDDSFSAHSEFLKAASND RDNYRFAH DNVESLVNE DDDNGEGITLPSH DNVESLVNE DDDNGEGITLPSH DNVESLVNE DDDNGEGITLPSH 237
Model_01 VVYYSVDFSPFDYRAATQFWRSKVLEVAKDFPE ---YTPAIAD EEDYAGEVKDLGL SESGEDVNAAILDES GKKPFAMEPEEE DSDTLRE FVTAFKKGKLPV 513
3f8u.1.A TAYDQVDYEKNAGK SNYWRNRVMMVAKMFLDAGHKINFAVA RKTFSHELSDGLESTAGEIFVVAITAKGKF MQEEFSRDGKALERFLQDYFDGNLKRY 340
Model_01 IKSQPVPKNNKGPVKVVVGKTFDSIVMDPKKDVLI EFYAPWCGHCKQLEPVYN SLAKKYKGQKGLVI AKMDATANDVPSDRYKVEGFPTIYPAPSGDKKNPVKPE 618
3f8u.1.A LKSEPIPEPSNDGPVKVVAENFDEIVNNENKDVLI EFYAPWCGHCKNLEPKYKELGKELSKDPNIVIAKMDATANDVPSPEYVRGFFTIYFPANKKLNPKVE 444
Model_01 GGDRDLEHLSKFIEEHATKLSRTKEEL 645
3f8u.1.A GGRELSDFISYLQRBATNPPVIEE- 469
```

PDIA4

```
Model_01 MNPTLILAAFLGIASTLTDFHSLEAQWTKWKAMHNRLYGMNEEGWRRRAVWEKNMKMIELHNQEYREGKHSFTMAMNAFGDMTSEEFROVMNGFQNRKPRKGKV 105
1cs8.1.A -----SLTFDHSLEAQWTKWKAMHNRLYGMNEEGWRRRAVWEKNMKMIELHNQEYREGKHSFTMAMNAFGDMTSEEFROVMNGFQNRKPRKGKV 88
Model_01 FQEPLFYEAPRSVDWREKGYVTPVKNQGCQGCSCWAFSATGALEQGMFRKTGRRLISLSEQNLVDCSGPQGNEGCNGGLMDYAFQYVQDNGGLDSEESYPYEATEES 210
1cs8.1.A FQEPLFYEAPRSVDWREKGYVTPVKNQGCQGCSCWAFSATGALEQGMFRKTGRRLISLSEQNLVDCSGPQGNEGCNGGLMDYAFQYVQDNGGLDSEESYPYEATEES 193
Model_01 CKYNPKYSVANDTG FVDIPKQEKALMKAVATVGPISVAIDAGHESFLFYKEGIYFEPDCSS EMDHGVLVVGYGFESTESDNNKYWLKNSWGEEWGMGGYVKMA 315
1cs8.1.A CKYNPKYSVANDAGFVDIPKQEKALMKAVATVGPISVAIDAGHESFLFYKEGIYFEPDCSS EMDHGVLVVGYGFESTESDNNKYWLKNSWGEEWGMGGYVKMA 298
Model_01 KDRRNHCGIASAASYPTV 333
1cs8.1.A KDRRNHCGIASAASYPTV 316
```

CTSL

```
Model_02 MTPPRLFWVLLVAGTQGVNDGDMRLADGGATNQGRVEIFYRGQWGTVCNLDWLDLTASVVCRA LGFENATQALGRAAFGGSGPIMLDEVQCTGTEASL 100
1by2.1.A -----VNDGDMRLADGGATNQGRVEIFYRGQWGTVCNLDWLDLTASVVCRA LGFENATQALGRAAFGGSGPIMLDEVQCTGTEASL 86
Model_02 ADCKSLGWLKSNCRHERDAGVVC TNETRSTHTL DLSRELSEALGQIFDSQRGCDLSISVNVQGEDALGFCGHVTILTANLEAQALWKEPGSNVTMSVDAE 200
1by2.1.A ADCKSLGWLKSNCRHERDAGVVC TNETRSTHTL DLSRELSEALGQIFDSQRGCDLSISVNVQGEDALGFCGHVTILTANLEAQALWKEPGSNVTMSVDAE 119
Model_02 CVP MVRDLLRYFYSSRRIDITLSSVKCFHKLASAYGARQLQGYCASLFAILLPQDPSFQMPLDLAYAVATGDALLEKLCLOFLAWNFEALTQAEAWPSVP 300
1by2.1.A -----
Model_02 TDLLQLLLPRSDLAVPSELALLKAVDTWSWGERASHEEVEGLVEKIRFPMMPLPEELFELQFNLSLYWSHEALFQKKTLQALEFHTVPFQLLARYKGLNLT 400
1by2.1.A -----
Model_02 EDTYKPRIYTSPTWSAFVTDSSWSARKSQLVYQSRGRLVKYSSDYFQAPSDYRYYPYQSFQTPQHPSFLFQDKRVSWSLVYLPTIQSCWNYGFSCSSDE 500
1by2.1.A -----
Model_02 LPVLGLTKSGGSDRTIAYENKALMLCEGLFVADVTDFEGWKAAPSAIDTNSKSTSSFPAGHFNGFRTVIRPFYLTNSSGVD 585
1by2.1.A -----
```

LGALS3