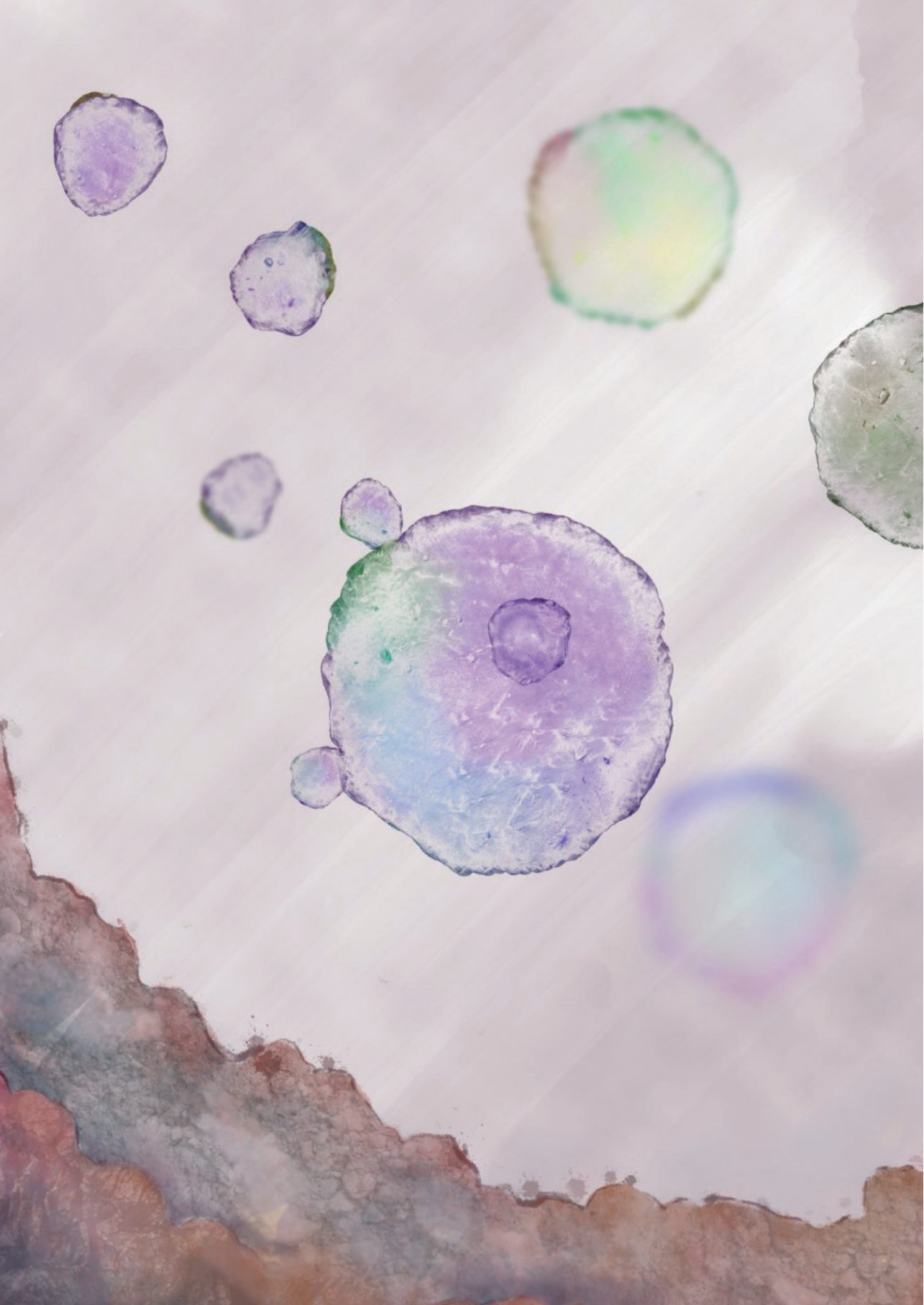
Glycoproteomics-based signatures of cancer cell lines

Pirro, M.



CHAPTER 1
INTRODUCTION

INTRODUCTION
Glycosylation

Protein glycosylation is one of the most common post translational modifications. The presence of carbohydrates on proteins, collectively known as glycans, plays a key role in the correct folding, stability, localization, protection from proteases and function of proteins [1]. In fact, glycosylation is found on most secreted and cell surface proteins, where it presents a key factor for solubility and stability and where it can promote cell-cell interaction, adhesion, migration, proliferation and activation of a large variety of intracellular signals [2, 3]. Of note, glycosylation can also occur on intracellular proteins.

The high complexity of mammalian glycans is the result of intricate biosynthetic pathways and the presence of more than 200 enzymes involved in synthesis and extension of sugar moieties. Some glycosyltransferases share the same glycan acceptor and can compete influencing glycosylated features in a cell or tissue [4]. Although the glycan diversity is high, the number of monosaccharides that are involved in human protein glycosylation is rather constrained. The 10 monosaccharides are fucose (Fuc), galactose (Gal), glucose (Glc), *N*-acetylgalactosamine (GalNAc), *N*-acetylglucosamine (GlcNAc), glucuronic acid (GlcA), iduronic acid (IdoA), mannose (Man), sialic acid (SA) and xylose (Xyl). The polymerization of these monosaccharides and linkage to specific amino acids of the peptide backbone make up most of the human *N*-glycans and *O*-glycans [5]. The presence or absence of glycans contribute to the macroheterogeneity of glycoproteins, while the actual size, structure and site of glycosylation contribute to their microheterogeneity.

N-glycosylation structure and biosynthesis

N-glycans are linked via an *N*-glycosidic linkage to an asparagine (**N**) of a polypeptide chain within the *N*-glycosylation consensus motif **N**-X-S/T (X= any amino acid except proline, S= serine, T= threonine). *N*-glycans synthesis starts in the endoplasmic reticulum (ER) as a lipid linked oligosaccharide (LLO) consisting of 14 sugars [6]. This precursor is initially transferred to a nascent polypeptide chain [7]. The processing of the *N*-glycan continues in the lumen of the ER and in the Golgi, thanks to the action of a series of glycosyltransferases and glycosidases [6]. At the end of the process, most *N*-glycans share a common pentasaccharide

core, consisting of two *N*-acetylglucosamines and three mannoses. Depending on the saccharides that elongate the core, three classes of *N*-glycans can be distinguished: high mannose, complex or hybrid (Figure 1A) [8].

Mucin type O-glycan structure and biosynthesis

In *O*-glycosylation, we distinguish different types based on the first oligosaccharide and linkage. The most abundant type is represented by the so-called mucin type *O*-GalNAc glycosylation. Differently from *N*-glycosylation, where a lipid-linked precursor is transferred to the protein, this type of glycosylation is characterized by a GalNAc linked to the hydroxyl group of a serine or threonine, and it highly decorates mucins and many other common glycoproteins [9]. The initial *O*-GalNAc monosaccharide also known as Tn antigen (GalNAcα1-Ser/Thr), can be further elongated with sialic acid forming the STn antigen (NeuAcα2,6-GalNAcα1-Ser/Thr) or to form core 1 to 8 *O*-GalNAc glycans depicted in figure 1B [9].

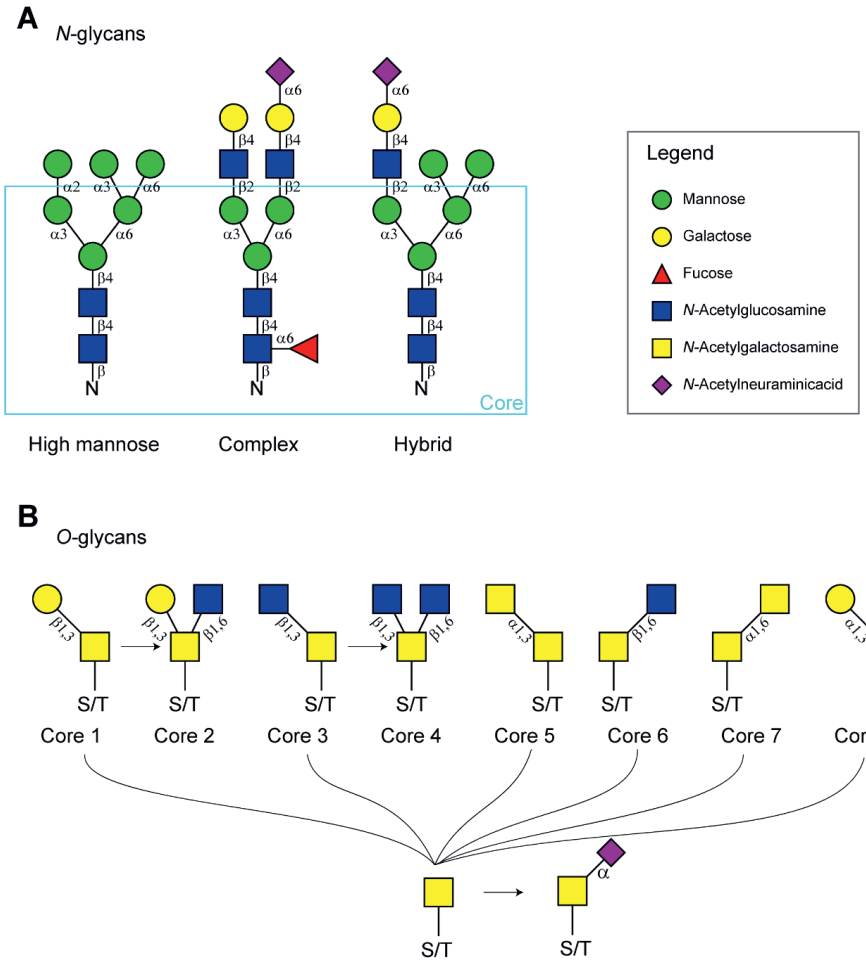


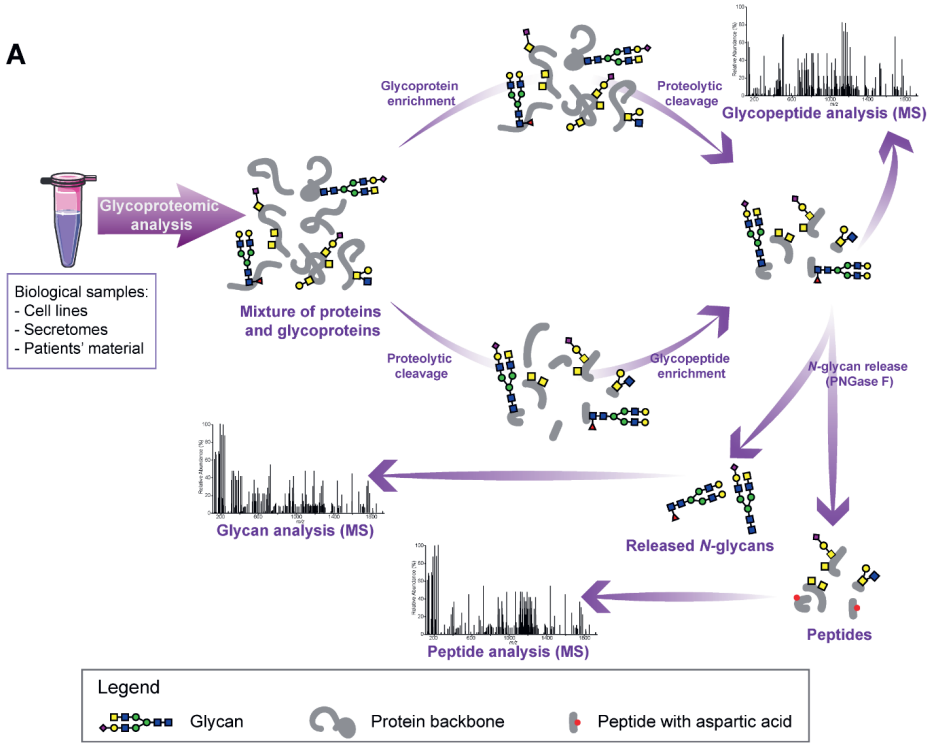
Figure 1: Common N- and O-glycan structures. A) *Three N-glycans types: high mannose, complex and hybrid. Blue square highlights the N-glycan core structure. N: asparagine residue of the polypeptide chain.* ***B)*** *O-GalNAc glycans core 1-8 attached to S (serine) or T(threonine) of polypeptide backbone. The addition of N-acetylneuraminic acid on O-GalNAc-S/T prevents its further elongation to extended O-glycans structures. Linkages types (α or β) are depicted in the figure.*

Aberrant glycosylation in cancer

Glycan profiles of proteins do not follow a predefined template since they are regulated by multiple factors, depending on cell type and tissue. Moreover, glycosylation patterns respond

to both physiological and pathological changes. Due to their important role, aberrant glycosylation in human diseases such as cancer, can contribute to a more malignant phenotype, characterized by cell-cell adhesion impairment, increased migration and invasion resulting in tumor metastasis [10]. Being a highly regulated post-translational modification the mechanisms that can generate abnormal glycosylation are multiple: type, function, location and expression level of specific glycosyltransferases and their chaperones, expression level of glycosidases, availability of protein substrates or sugar donors, activity of nucleotide sugar transporters, pH of ER and Golgi and competing reactions between different glycosyltransferases for similar substrates [1].

Cancer-associated alterations in protein glycosylation can be divided into two groups, truncated or *de novo* glycan structures. Both are commonly known as tumor-associated carbohydrate antigens (TACA), because they are lower abundant or even absent on normal mammalian cells [11]. The major types of glycosylation changes include (overexpression of) truncated *O*-glycans, such as the above mentioned Tn antigen, its sialylated form STn (NeuAcα2,6-GalNAcα1-Ser/Thr), the disaccharide Thomsen–Friedenreich antigen or T antigen (Galβ1,3-GalNAcα1-O-Ser/Thr) [5]. Within the specific *N*-glycans changes, increased sialylation, core/antenna fucosylation [5], high-mannose type and tri- or tetra-antennary glycans are commonly seen in different type of cancer tissues and cell lines [1, 12]. *N*- and *O*-glycans can be further extended by a GalNAcβ1,4-GlcNAcβ1- (LacdiNAc or LDN) group at the non-reducing termini [13, 14]. Fucosylated glycans associated with cancer are specific Lewis blood group antigens such as LeX, LeY, LeA and LeB [15]. Moreover, LeA and LeX can be expressed by cancer cells in their sialylated species SLeA and SLeX, respectively (Figure 2) [5].



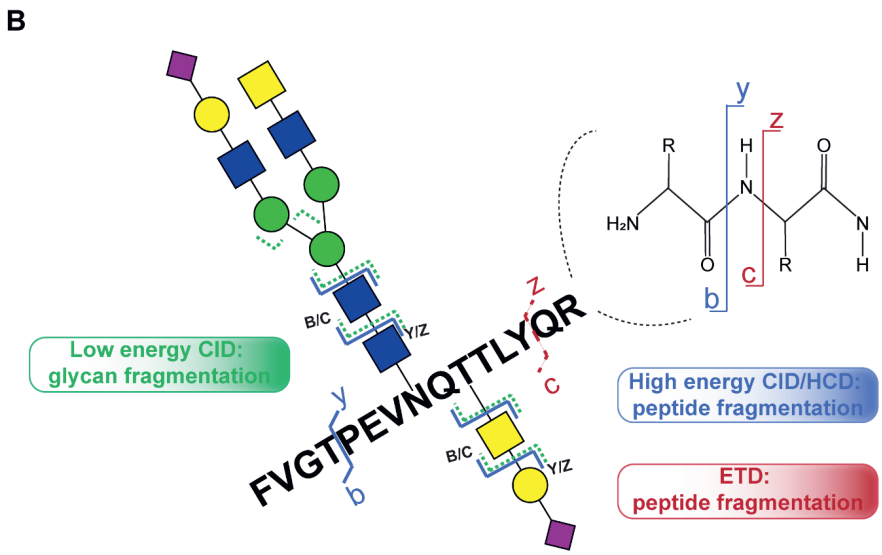


Figure 2: Strategies for glycoproteomic analysis. A) *Glycoproteomic workflow: Glycoproteins from multiple sources are enzymatically digested prior or after a glycoprotein/glycopeptide enrichment step. Enriched sample are then either directly analyzed as intact glycopeptides or after a step of N-glycans release (PNGase F treatment). Both glycan and glycopeptides undergo to Mass Spectrometry*

(MS) analysis. ***B)*** *Glycopeptide MS/MS fragmentation methods. The depicted glycopeptide sequence belongs to that of a tryptic N-glycopeptide from TIMP1, while the O-glycan was added for illustrative purposes. Location of fragmentation and glycosidic cleavages are exemplary.*

Study of glycoproteins

Glycosylation changes have been associated with the acquisition of hallmark capabilities (survival, proliferation and dissemination), playing an important role in neoplastic transformation and tumor progression [16, 17]. To study such changes, it became important to develop strategies to not only comprehensively analyze glycan diversity, using glycomics tools, but at the same time identify the proteins bearing these specific glycans (glycoproteomics). Recent technical advances in the field of glycoproteomics, using advanced mass spectrometry (MS) methods, coupled to analytical separation methods, such as (ultra) high performance liquid chromatography (U)HPLC have significantly improved the characterization of protein glycosylation in a protein- and glycosylation site-specific manner [18]. In addition, many complementary separation and detection methods, such as the use of glycan specific lectins, can, together, give an in-depth characterization of the glycoproteome of a complex sample.

Mass spectrometry-based (glyco)proteomics

The most common glycoproteomic studies are based on the characterization of glycopeptides obtained by the digestion of proteins with site-specific proteases (most commonly trypsin) of a large variety of protein mixtures, defining the so called bottom-up approach (Figure 3A) [19]. These samples can be individual purified glycoproteins, sets of co-migrating proteins on an SDS-PAGE gel, secreted proteins or even highly complex protein extracts (e.g. cell lines proteomes, patients' material). The MS methods used for the study of glycoproteins and glycopeptides vary depending on the research question and the complexity of the starting material [20]. Glycopeptides are usually hard to analyze due to i) glycosylation microheterogeneity which results in different glycoforms of the same peptide and its low abundance in the total peptide pool, ii) the complexity of the glycan structure and iii) the low ionization efficiency compared to the unmodified peptide [18, 20]. To reduce the technical challenges and the complexity of information obtained in a spectrum due to the presence of

the carbohydrate moiety, glycopeptides can be analyzed after release of all or just a selected class of glycans. For example, peptide-N-glycosidase F (PNGase F) is an effective endoglycosidase widely used for the removal of *N*-linked glycans from glycoproteins. The reaction can be performed both in-gel and in-solution and the enzyme cleaves the amide bond of the *N*-glycosylated Asn residue of proteins [21]. This reaction results in a change from asparagine to aspartic acid, which can be used in proteomic data analysis for the assignment of the original site of glycosylation [22]. For *O*-glycans, although specific *O*-glycosidase exist, chemical release remains the most commonly applied approach for complete *O*-glycan release [22]. The most used reaction is β-elimination, consisting in the cleavage of the *O*-glycosydic bond on Ser or Thr. As for PNGase F, the reaction changes the amino acid at the site of elimination marking the *O*-linked glycosylated locations [22], however side reactions and partial degradation of the peptide portion reduces its application for further MS-based proteomics [23]. Moreover, obviously, any method that releases the glycans will result in the loss of information regarding which glycan is attached to a particular amino acid in a certain protein [24]. On the other hand, MS-based glycoproteomics analysis of intact glycopeptides can provide information about the peptide backbone and glycan structure in a single measurement. Since the comprehensive analysis of intact glycopeptides is technically challenging, many glycopeptides/glycoproteins enrichment techniques have been optimized. Among those, we recognize hydrophilic interaction-liquid chromatography (HILIC), which enriches glycopeptides based on their hydrophilicity and size of the glycan moiety, chemical and metabolic labelling of glycans, and affinity purification [20].

In addition to enrichment strategies, specific mass spectrometry approaches for the analysis of glycopeptides have been developed. A mass spectrometer comprises two major units: the ion source for producing gas-phase analyte ions, and the mass analyzer, which separates the ions based on their mass/charge ratio (*m/z*) and detects each resolved ionic specie with its relative abundance [19]. These components exist in different flavors and can be used in different combinations.

MS-based glycoproteomic studies use different levels of information for the identification not only of the microheterogeneity of glycans present but also to identify the peptide sequence (and, if unique, the protein that it is derived from), and the glycosylation site. Knowing the mass of a peptide and of the individual monosaccharides, MS analysis allows the identification of the glycan composition of a glycopeptide. Additional information on the structure of

specific glycan composition can be obtained by using tandem mass spectrometry (MS/MS, MS^2) on selected ions. The choice of the fragmentation method will result in either glycan or peptide fragmentation, giving rise to specific ions depending on the type of fragments that are formed (Figure 3B) [19]. We recognize three different main fragmentation modes: collision-induced dissociation (CID), higher-energy collisional dissociation (HCD) and electron-transfer dissociation (ETD). Low energy CID predominantly results in glycosidic bond cleavages and are therefore suitable to obtain additional information about the glycan structure. In HCD, also peptide bonds are fragmented, generally resulting in characteristic **b** and **y** ions, corresponding to N- and C-terminal fragments, respectively. With ETD, gas-phase electron-transfer reactions of singly charged anions and multiply charged protonated peptides generate peptide backbone fragmentation into **c-** and **z-** type ions, due to **N-C_α** bond cleavage, but leave the glycosidic bond intact. Hence, peptide sequence information can be obtained by ETD and HCD, where ETD is preferred for site-specific analysis. Often, a combination of fragmentation methods is necessary for full characterization of a glycopeptide [20].

Modern-type hybrid mass spectrometers, such as those using Orbitrap technology, often enable the combination of HCD, CID and ETD, as well as the possibility of routing ions to the different types of mass analyzers depending on the desired application. In addition to glycopeptide identifications MS-based analysis of glycopeptides can be combined with relative quantification approaches, such as Tandem Mass Tag (TMT), where of up to 16 samples can be compared. With this method, the quantification of the labeled glycopeptides is based on the detection and relative quantification of the reporter ions, corresponding to different labels/samples, in MS^2 or even MS^3 [18, 25-27].

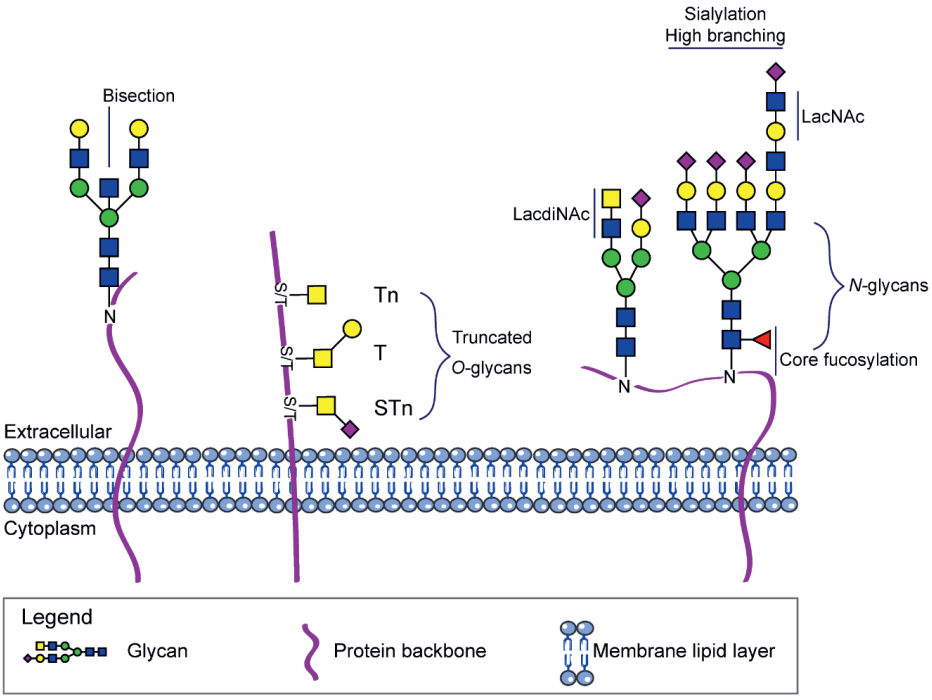


Figure 3: Examples of aberrant N- and O-glycosylation in cancer. *N (asparagine), S (serine) and T (threonine) residues of the polypeptide chain.*

Lectin-affinity purification

Lectin-affinity purification uses glycan-binding proteins to purify a specific set of glycoproteins (or more in general glycoconjugates) from a heterogeneous complex mixture of proteins [28]. Lectin-affinity purification methodologies are often used as starting steps for MS glycomics and glycoproteomics, working as preconcentration and enrichment of certain glycoproteins/glycopeptides [28]. In fact, lectins are proteins capable of recognizing more or less restricted classes of sugars or glycan structures. They can derive from multiple sources, ranging from viruses, bacteria, plant, animals or humans [29]. Immobilized lectins on sepharose, agarose, magnetic beads or other resins, either in suspension or immobilized in a column, bind to glycoproteins in a non-covalent and reversible fashion, and therefore the

captured ligands can be selectively released by competitive elution (e.g. using a corresponding free sugar or analog) [28].

Lectin-binding to aberrant glycosylation

In addition to their use as tools to enrich for certain glycoproteins, endogenous lectins which are part of the innate immune system against non-self, can also recognize TACA expressed on plasma membrane or secreted proteins. This interaction can lead to tumor progression and metastasis through evasion of anti-tumor immune responses. Animal lectins are categorized into several families, but in human cancer biology we recognize three main groups, C-, S- and I-type [11]. C-type lectins are the largest family of lectins and share a carbohydrate recognition domain (CRD) which requires calcium in order to bind the specific glycosylated ligands [30]. Among those, the Macrophage galactose-type lectin (MGL) is the best studied of this family [31]. Within the S-type, the most studied are Galectins, which are associated to cancer initiation and progression as well as tumor escape [32]. Siglecs, belonging to I-type lectins recognize different sialylated glycoconjugates, including STn, and can activate or inhibit the immune system [11].

MGL lectin: ligand specificity

MGL is a homotrimer cluster expressed on the surface of immature dendritic cells (DCs) and macrophages [33]. MGL specifically recognizes terminal GalNAc residues exposed C3- and C4-hydroxyl groups, having (S)Tn antigen and LacdiNAc as major ligands [34]. NMR analysis of the MGL-CRD, containing the three amino acid motif (Gln-Pro-Asp), revealed that it is highly dynamic and its conformation changes in a ligand-specific manner [35]. Its high plasticity might explain MGL ability of binding different TACAs with terminal GalNAc antigens. The type of glycan ligand can also have an effect on the signaling transmission of MGL to immune cells. In fact, activation of intracellular pathways is usually accompanied by structural changes of CRD-MGL upon ligand recognition [35]. In addition, a secondary binding site of MGL has been recently identified using a MGL mutant model carrying a threonine, rather than a histidine, at position 259 [36]. Binding affinity studies of this mutant showed reduced affinity for Tn antigen, LacdiNAc motifs and Tn-containing glycopeptides and failed to interact with STn

sugar. This loss of affinity suggests that MGL not only recognizes the sugar portion, but also the underlying protein backbone and most probably this recognition happens through the secondary binding site, where His259 plays a key role [36].

Human MGL in the Immune response

Immature DCs are highly efficient Antigen Presenting Cells (APC) [34]. When MGL expressed on their surface recognizes its ligands, it rapidly internalizes the antigen. The endocytosed ligands are transported along the endosomal-lysosomal pathway to be presented on MHC class II molecules [34]. MGL plays a significant role in the immune response to viruses (such as Ebola) [37] and parasites (e.g. *Schistosoma mansoni*) [38]. The first MGL counter receptor recognized in humans was the tyrosine phosphatase CD45 carrying GalNAc epitopes [39]. It is expressed by human T cells in five alternative isoforms (ABC, AB, BC, B and RO) with different glycosylation, which changes during T cell activation and differentiation [39]. MGL recognizes all isoforms except for CD45RO, which carries only two *O*-linked glycans. This evidence suggests that two *O*-linked glycans are not sufficient for MGL binding. On the other hand, MGL prefers binding to effector T cells, highly expressing CD45B isoforms, which suggests a role of MGL in various immunological pathways [34]. In fact, MGL-CD45 interaction results in a decreased activation of CD45 phosphatase and consequent reduced activation of its downstream target, tyrosine phosphatase Lck, and T-cell receptor (TCR)-mediated signaling. As a consequence, effector T cells decrease their proliferation and cytokine production, which could evolve into T cell death. This could be a self-protecting mechanism triggered to prevent chronic inflammation and consequent tissue damage [39]. This effect was first seen during prolonged stimulation of Acute T cell Leukemia cell model, Jurkat T cells, with recombinant MGL and concomitant TCR activation (with anti-CD3). Jurkat cells contain a mutation in the gene encoding for the T-synthase chaperone, Cosmc, responsible of the elongation of Tn antigen to T antigen. For this reason, Jurkat cells expose higher density of terminal *O*-GalNAc structures on glycoproteins, representing a suitable model for studying MGL binding. For example, CD43 expressed on Jurkat cells, represents another specific binder for the lectin [39] but other ligands on Jurkat cells are currently unknown.

MGL role in human cancer

Recent evidences have shown that MGL might be involved in immune responses to human adenocarcinomas. Tn and STn antigens are found in the vast majority of breast cancers [40]. For this reason, MGL was chosen as tool for glycoprofiling in this type of tumors in order to select patients who could benefit from MGL-based specific therapeutic approaches [41]. Tn and STn can be carried by MUC1, a glycoprotein bearing high density truncated *O*-glycans which is expressed on the surface of epithelial cells, and aberrantly overexpressed in many carcinomas [41]. Even though the overexpression of Tn and STn antigen cannot be explained by the inactivation or lack of expression of Cosmc in breast cancer cells, MUC1-Tn and MUC1-STn bind MGL lectin [42]. An accumulation of MGL positive glycan structures could partially be explained by upregulation of the carrier protein MUC1, together with the enhanced expression of GalNAc-transferases 6 (GALNT6) and translocation from cis- to trans-Golgi compartment of GalNAc-transferases 2 (GALNT2) and GALNT6, both responsible of Tn antigen formation [43]. Glycoproteins binding to MGL were characterized from total lysates of breast cancer samples, confirming that proteins with truncated *O*-glycans were primarily mucins (MUC1/5AC/16). Higher MGL staining of breast cancer patients' tissues was correlated with a better prognosis and overall survival [43]. In different cancer types, MGL has been linked to an immune evasive role because MGL binding was associated with a more aggressive tumor phenotype resulting in poor prognosis of the patients. This is the case of MGL studies on (adeno-)squamous cervical cancer [44] and colorectal cancer (CRC) [45]. In the first study, it was demonstrated that higher MGL ligand expression is associated with oncogenic PIK3CA mutations [44], whereas in the second study, to BRAF mutations [45]. MGL can be used as a tool to distinguish healthy tissues from CRC [46] and MGL binding may serve as a novel prognostic biomarker for stage III colon cancer patients, predicting lower survival and higher disease recurrence rate [45]. As found in breast carcinoma, MUC1 isolated from primary colon carcinoma tissues strongly binds MGL [46], even though this study was not focused on a more comprehensive analysis of other expressed MGL ligands.

Colorectal cancer

Colorectal cancer (CRC) is the third most common cause of death worldwide, both in women and men [47]. CRC arises from the mucosa of the bowel usually projecting on the lumen-side. It usually develops from a non-malignant lesion, called adenoma (polyp), progressing to carcinoma. The symptoms are generally not specific for CRC (e.g. abdominal pain, unexplained weight loss, iron deficiency or anemia) and for this reason it is hard to diagnose, especially at early stages [48]. About 20-25% of the patients are diagnosed at metastatic stages of the disease [49]. The only curative treatment for invasive CRC is surgical resection. For high-risk patients, adjuvant treatment is recommended after surgery, whereas resection accompanied by chemotherapy is necessary in case of distant metastasis [48]. In many western countries CRC -related death has declined, thanks to cancer screening programs (i.e. colonoscopy), the removal of adenomas and early detection of malignant lesions as well as the availability of targeted therapies [48]. However, cancer-related survival reduces with age, and the stage of the disease at the moment of diagnosis is the most important prognostic factor [48].

Risk factors and histopathological classification

The risk of developing CRC can be mainly found in lifestyle and behavioral factors, such as smoking, obesity, high red meat and alcohol consumption and physical inactivity [48] as well as genetic factors. In addition, age, male sex and diseases such as inflammatory bowel disease (IBD) and diabetes are associated with CRC onset [50].

Classification of CRC is based on the extension of primary tumor (T stage), lymph node infiltration from cancer cells (N stage) and occurrence of distant metastasis (M stage). These features determine the overall TNM classification system, rating from 1 to 4, which provides the basis for therapeutic decision [50]. In fact, risk assessment is necessary to decide adjuvant therapy administration after primary tumor resection, in order to reduce the risk of relapse and death [48]. The use of new predictors have been recently examined in order to help not only in the early diagnosis but also in the prediction of prognosis and therapy response in CRC patients.

Diagnostic screening and prognostic factors

CRC is diagnosed histologically by biopsies obtained during endoscopy [50], as sigmoidoscopy (more than 35% of CRC tumors are located in the rectosigmoid) or a more comprehensive colonoscopy. The screening procedures recommended in over-50 years-old women and men include also the fecal occult blood test (FOBT) and molecular non-invasive screening test. Detection of DNA, RNA or other molecules derived by tumor, found in stool or blood, would contribute to reduce CRC incidence and mortality [51]. In fact, molecular mechanisms responsible of CRC development are clinically important also for prognosis and treatment response of patients. The CRC carcinogenesis process is characterized by genetic alterations. Between those, the most prevalent involve mutations in KRAS, TP53, APC genes and markers for microsatellite instability (MSI) [51]. Among the investigated biomarkers also proteins play a big role. The Carcinoembryonic antigen (CEA), discovered in 1965, still remains the only specific biomarker for postoperative follow-up and therapy efficacy in CRC patients [48]. However, high levels of CEA in blood have been found also in other tumors and inflammatory diseases and it does not discriminate between benign and malignant polyps. With its relatively low sensitivity and specificity, CEA is not used for early diagnosis screening [51], and consequently there is still a need for reliable biomarkers for early detection.

Multi-omics of CRC cell lines for glycobiology studies

Cell lines have been widely used as pre-clinical model system. Their comprehensive characterization is important to select the optimal model(s) depending on the research question. In fact, cell lines show high molecular and glycosylation variability, suggesting that studying a single CRC cell line might not be representative and sufficient to generalize glycobiology findings and select relevant biomarkers. Altogether, multi-omics studies on CRC cell lines, not only give comprehensive information at multiple levels (gene/protein expression, single nucleotide mutations and glycosylation profiling) but also show the applicability of CRC cell line as representative molecular and glycomic models for primary carcinomas [52, 53].

CRC cell lines can be divided into two distinct clusters: colon-like, expressing gastro-intestinal specific markers of differentiation, and undifferentiated cell lines, characterized by

upregulated signaling typical of epithelial-mesenchymal transition such as TGFβ signaling [52]. For example, within the first group, HT29 cells, derived from stage 3 primary tumor, are microsatellite stable (MSS) and are characterized by mutations in TP53 and BRAF variation p.V600E. Although part of the same differentiated group, LS174T cells originates from stage 2 primary tumor, with MSI, wild-type for TP53 but carrying mutations in KRAS, PI3CA and a different variant of BRAF (p.D211G). On the other hand, in the undifferentiated group, HCT116 derives from stage 4 and presents MSI, KRAS and PI3CA mutations[52].

Twenty six CRC cell lines have been comprehensively characterized for both their *N*- and *O*-glycosylation profile [53, 54]. Higher levels of fucosylation were found in various CRC cell lines, with highest expression on LS180 and its variant, LS174T, whereas the undifferentiated HCT116 and DLD1 cell lines are an exception, with very low fucosylation levels [53]. Fucosylated antigens such as LeX and LeY expressed by some CRC cell lines as well as tissue from CRC patients can be recognized by the C-type lectin DC-SIGN of immature DCs, which may induce immune-suppressive tumor microenvironment [55]. Multi-fucosylation is often accompanied by higher expression of α2,3-sialylation in some CRC cell lines, such as HT29, SW48, Lovo, indicative of sialyl Lewis antigen expression [53]. Sialylated species are sensed by sialic-acid binding immunoglobulin-type lectins (Siglecs) on tolerogenic DCs, resulting in increased anti-inflammatory response [56]. In colon, the major secreted glycoproteins are mucins, which are heavily *O*-glycosylated with core 1, 2, 3 and 4 glycans [12]. During CRC progression, mucins express low levels of core 3 and 4 [12], and these structures were in general found at lower levels in CRC cell lines [54]. In fact, mucins overexpress truncated *O*-glycans, such as T antigens (during early stages of the malignant transformation) and (S)Tn (hallmark of more advanced and poor differentiated colon cancers) [12]. (S)Tn antigens, together with *N*- or *O*-glycans carrying a LacdiNAc motif, are recognized by MGL, whose high ligand expression was found in CRC cell lines harboring BRAF mutation V600E, such as HT29 [45]. This *in vitro* finding was correlated with poor survival and disease recurrence in stage III CRC and MSI status patients, suggesting MGL potential to be used as prognostic marker for this subgroup of patients.

In conclusion, tumor-associated glycans drastically influence the immune response to tumors via glycan-binding receptors expressed on immune cells. This interaction may facilitate immune evasion or anti-tumor response in different conditions. Hence, well characterized cell

lines can be chosen and used for studying the interactions between the tumor microenvironment and cancer cells.

Scope of the thesis

The scope of this thesis is to increase our understanding of the (glyco)proteomic signatures of malignant cell lines (Acute T cell Leukemia and CRC). The discovery of glycoproteins binding to the immunological receptor MGL can be the starting point not only to learn the intracellular mechanisms activated in cancer by lectin recognition, but also to study their potential as biomarkers for improved prognosis and patient stratification.

A previous study identified CD45 and CD43 as major MGL binding proteins expressed by Jurkat, a highly expressing Tn antigens cell model. In **Chapter 2**, we expanded the existing study and optimized a method for capturing and MS-based identification of MGL binding proteins from these cells. It revealed a novel set of MGL binding proteins in this cell model, with glycopeptides carrying a variable number of Tn antigens, ranging from 1 to 11.

Triggered by the prognostic value of MGL binding in high stage CRC, we also explored colorectal cancer cell lines for proteins binding to MGL (**Chapter 3**). Adapting the protocol previously used for Jurkat, we noticed a high variability in MGL staining in different CRC cell lines, discriminating high binders, HCT116 and HT29, from low binder, LS174T. Interestingly, the characterization of several glycopeptides from the MGL binding proteins from the CRC cell lines indicated LacdiNAc on *N*-glycans, over Tn antigen, as MGL ligands on these cell lines. One of the main MGL binders in the two high-binding cell lines was the hepatocyte growth factor receptor c-Met.

In **Chapter 4**, we substantiated the role of *N*-glycans in the binding of proteins to MGL in CRC cell lines by a pre-treatment with PNGase F, prior to MGL pull-down and ligand characterization. Indeed, both on single protein level (c-Met) as well as overall, release of *N*-glycans drastically reduced the binding of CRC cell lines proteins to MGL. Hence, these studies underscored the relatively unrecognized importance of *N*-glycans in MGL binding to CRC cell lines. Moreover, in order to have more insight into potential mechanisms behind the high MGL ligands expression, we used comparative quantitative proteomics to analyze the differential expression of i) protein substrates carrying the MGL binding glycan epitopes, ii) glycosyltransferases and glycosidases involved in *N*- and *O*-glycosylation. This study not only ruled out different levels of protein substrates as a mechanism behind the differential MGL

binding, but also the role of mucins in binding MGL in CRC cell lines, contrary to what was seen in CRC tissues and other cancer types.

In **Chapter 5**, we showed for the first time that also secreted proteins from CRC cell lines bind MGL. Interestingly, also glycoproteins secreted by the low MGL binder cell line (LS174T) carry MGL-glycan ligands, suggesting that secreted proteins are not always the direct reflection of cell surface glycoprotein expression. This increase the importance of studying different type of samples.

Differently, **Chapter 6** addresses the implementation of an automated bioinformatic tool to overcome the challenging discrimination of the monosaccharide isomers *O*-GalNAc and *O*-GlcNAc. This is important for the characterization of the hallmark Tn antigen in MS/MS spectra through the ratio of specific oxonium ions obtained in the fragmentation patterns. Applying this method to CRC glycoprotein analysis, highlighted anterior gradient protein 2 (AGR2) as an *O*-GalNAc carrying protein differentially expressed in CRC cell lines, and demonstrated AGR2 *O*-glycosylation in a site-specific manner for the first time. Moreover, this study revealed different AGR2 *O*-glycosylation in secreted as compared to ER-resident AGR2.

Finally, **Chapter 7** gives a general discussion of our findings and suggests future perspectives to contribute to novel clinical applications. Also, current challenges in the field of CRC glycoproteomics are discussed.