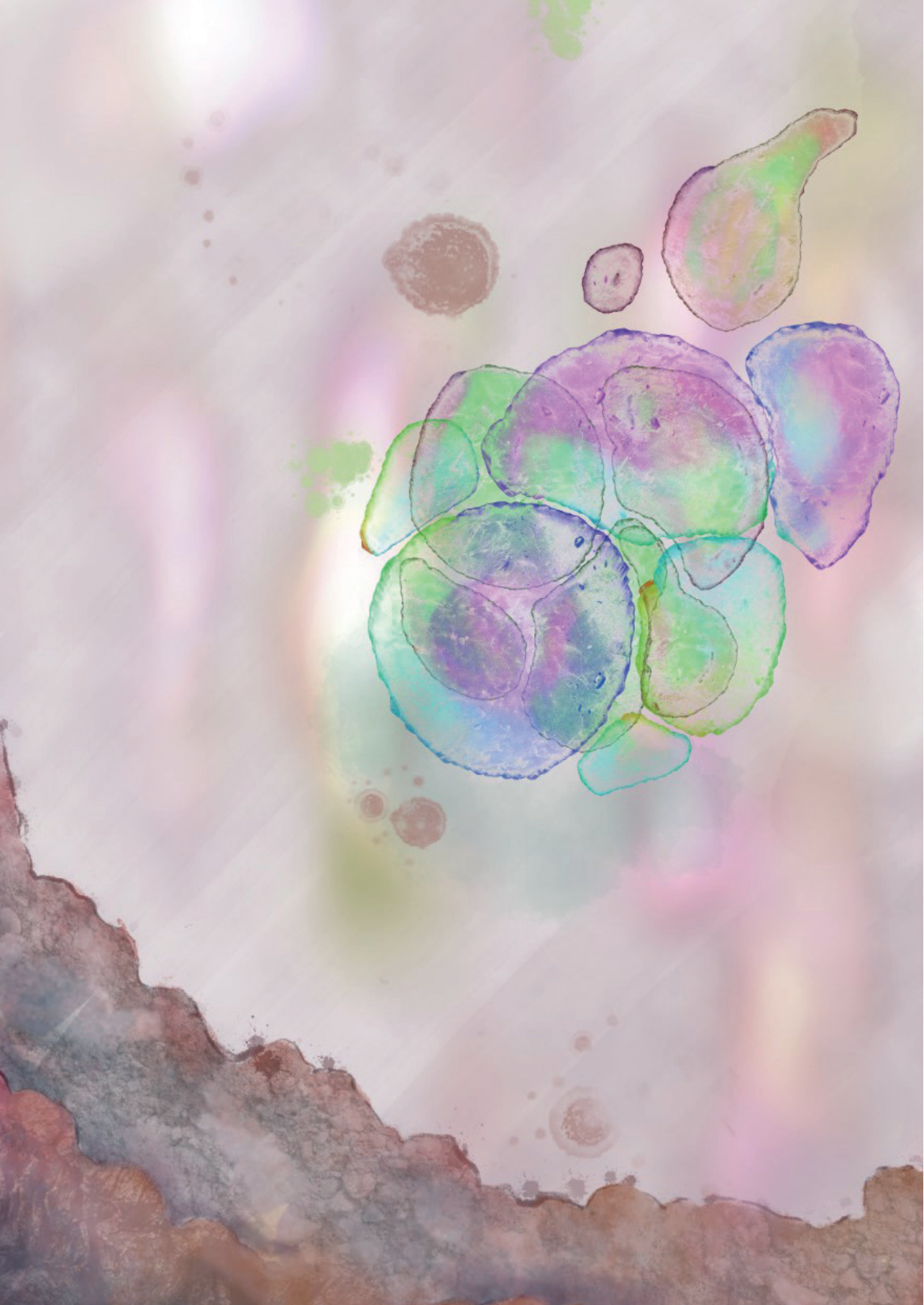
Glycoproteomics-based signatures of cancer cell lines

Pirro, M.



CHAPTER 7

GENERAL DISCUSSION

General Discussion

The interaction between tumor and immune cells in the microenvironment plays a key role in oncogenesis. This recognition can be mediated by changes in glycosylation on tumor cells which are sensed by lectins, such as the Macrophage Galactose-type Lectin (MGL), expressed on immune cells, resulting in immunosuppressive responses [30]. MGL binding to the Tn-epitope on MUC1 triggers DCs to stimulate T-regulatory responses and tolerance against the tumor, while suppressing T-effector cells responsible for tumor eradication. This undesirable consequence in cancer led to interest in the characterization of targets of MGL. This thesis dealt with the analysis of glycoproteins binding to MGL expressed by different tumor cell models. For this purpose, a robust method to enrich MGL binding proteins, in combination with their mass spectrometry-based identification was established. Because of the high level of the Tn-antigen, due to a *Cosmc* mutation, the Jurkat cell line was initially used for this purpose (Chapter 2) but a slightly adjusted method was later applied to high- and low-MGL binding CRC cell lines as well (Chapter 3 and 5). These analyses led to the identification of hitherto unknown MGL binders. In Chapter 4, we focused on to the contribution of *N*-glycan MGL binding glycotopes in CRC cell lines, most probably corresponding to LacdiNAc structures, by implementing overall *N*-glycan release in our workflows. With this, we were able to show that *N*-glycoproteins represent a hitherto underestimated group of MGL binding protein in these cell lines. Also several secreted proteins from the CRC cell lines could bind to MGL (Chapter 5), indicating that the interaction with immune cells can also be mediated by this group of proteins. The results from previously published transcriptomics and *N*-/*O*-glycomic analyses could not explain the different expression of MGL binding proteins on the CRC cell lines used. For this reason, in Chapter 4, we extended our research with full comparative quantitative proteomics analyses, in an attempt to explain the differences in MGL binding, for example by different levels of MGL binding proteins or proteins involved in glycosylation pathways. Additionally, in Chapter 6, we used such a quantitative proteomics dataset also to test the suitability of a previously suggested mass spectrometry-based method to discriminate O-GalNAc (Tn) *versus* O-GlcNAc, which led to the first site-specific identification of *O*-glycosylation of both intracellular and secreted anterior gradient protein 2 (AGR2).

Notwithstanding these accomplishments, several research challenges have been encountered and these will be discussed further below.

Methodological challenges

Although glycoproteomics is nowadays more widely used, the characterization of specific cellular glycoprotein changes remains a challenging task. The detailed characterization of the MGL binding proteins as presented in this thesis provides the first step in order to initiate further experiments to understand the mechanisms behind tumor evasion from the immune system and identify potential antitumoral targets. One of the tasks of the work presented here was the development and application of suitable methods to specifically enrich and analyze MGL binding glycoproteins, especially in combination with the annotation of their subcellular localization, and identification of glycopeptides that could explain their binding to MGL.

Our ultimate method, using lectin pull-down combined with LC-MS/MS analyses followed by stringent selection criteria after data processing, represents a robust approach for the identification of specific MGL ligands from complex mixture of proteins obtained from a total cell lysate. Because such approaches always suffer from proteins that non-specifically bind to the materials that are used, the choice of proper negative controls was pivotal. Initially, we used pull-downs in the presence of EDTA for this purpose, later we applied a more elegant approach by using a mutated MGL, lacking the carbohydrate binding domain. In the case of MGL, other options are the specific elution with synthetic free *N*-acetylgalactosamine and EGTA. In addition to the results presented in this thesis, we have also applied these methodologies with the Jurkat cell model, using detection of the specific MGL ligand CD43, as readout for elution efficiency. The results of these experiments showed that, although they were effective, both GalNAc and EGTA were less efficient in CD43 elution compared to EDTA.

The use of a total cellular protein extract has the disadvantage of losing information about the cellular location of glycoproteins. This limitation also applies for glycosylation analysis, where interpretation of results is complicated by the presence of intracellular glycan precursors in the total cellular homogenate. To tackle this issue, previous *N*-glycosylation profiling of CRC cell lines, *N*-glycans are enzymatically “shaved” from intact cells [53]. Even though, the analysis of such samples showed a lower percentage of high mannose glycans

compared to the levels identified in a total lysate, the cell surface glycan analysis resulted in a less high throughput method, was more time consuming and required a higher number of cells. Since we assume that only cell surface and secreted epitopes are in contact with MGL on immune cells *in vivo*, we determined the subcellular location for our identified proteins by mining the Uniprot database. Although this is generally accepted as an appropriate method, it is complicated by the fact that many proteins do have not a single, but multiple cellular locations, and interconnection between different organelles exists [157]. Hence, the *in silico* prediction of protein location would benefit from supporting data. For this purpose, different methods are available. For example, enrichment of plasma membrane proteins prior to lectin pull-down is one option. The enrichment is usually achieved by differential centrifugation involving a density gradient or ultracentrifugation. However, proteins coming from other compartments (e.g. the cytoskeleton) or attached to the membrane, usually heavily contaminate such membrane preparations [158]. Furthermore, the amount of starting material needed to obtain a suitable amount of plasma membrane proteins to proceed with the second step of (lectin) enrichment that would still allow detection by mass spectrometry is drastically increased compared to the simple workflow with whole cell lysates.

Another approach that can be used to study exposed plasma membrane (glyco)proteins is the labeling of these proteins prior to the subsequent affinity capture. Most labels target protein primary amines (N-termini and lysine side chains) or make use of hydrazide chemistry which targets periodate-oxidized sugar moieties of surface glycoproteins. When coupled with biotinylation, such methods can be combined with affinity purification through streptavidin-coated beads and in case the sugars are directly labeled, glycopeptides could be enriched directly. However, in general these workflows lead to a relatively low cell membrane enrichment and have other major limitations: living cells have to be used to prevent the penetration of labels inside the cells and the method is relatively laborious and time consuming.

During our studies, we have tried two plasma membrane protein enrichments methods. One method is an available commercial kit based on characteristic physical properties of plasma membranes compared to intracellular membranes, which allow for their separation in a two-phase system. The second one made use of biotin labeling of plasma membrane proteins from intact cultured cells for subsequent streptavidin affinity pull-downs of cellular extracts. Although we could achieve a decent plasma membrane enrichment (approx. 40%) and we

could show that cell surface biotinylation of Jurkat cells in combination with affinity purification lead to efficient capturing of CD43, many intracellular proteins were still present in such samples. Moreover, the limited benefits were outweighed by the increase of starting material required for plasma membrane enrichment, affecting both costs and time. Hence, in the end we decided to not include these sample preparation steps in our final workflow to identify MGL binding proteins, as presented in this thesis. In our hands, applying incorporation and capturing of glycoproteins with clickable sugars [159, 160] was unsuccessful.

Finally, the last essential part of our method is the mass spectrometry-based identification of glycopeptides. For several MGL binding proteins we were able to identify a glycopeptide that could explain the binding to MGL, while for others, such as one of our prime candidate in the high-MGL binding CRC cell lines, c-Met, we have not found a glycopeptide that carried the specific glycotope. This could be due to several reasons.

First of all, for the mapping of post-translational modifications a high sequence coverage is commonly required. For this purpose, optimal solubilization in combination with a digestion method that results in peptides with a suitable size and hydrophobicity compatible with the LC-MS/MS analysis is pivotal. In our work, we have used the gold-standard protease in proteomics, trypsin. However, for certain glycoproteins, trypsin may not be the best choice because of the location of its cleavage sites (Lys/Arg-Xaa, Xaa≠Pro) or their poor accessibility due to extensive glycosylation. Although the latter could be overcome by a deglycosylation step prior to proteolytic digestion, this will result in the loss of glycan-specific information for each peptide. On the other hand, such an approach helped us in Chapter 4 to demonstrate a major contribution of *N*-glycoproteins to the binding to MGL in the CRC cell lines.

The use of alternative proteases can be introduced in the attempt of increasing protein coverage [158]. Among those, chymotrypsin is frequently used since it generally results in smaller peptides because it cleaves after multiple, primarily hydrophobic, residues, e.g. Phe, Tyr, Trp and Leu. However, in Chapter 3, the use of chymotrypsin did not improve our results. With the aim to get a better understanding of the contribution of the Tn antigen and LacdiNAc motifs on our MGL binding proteins, we also tested the enrichment of glycopeptides through HILIC of tryptic peptides from MGL pull-downs or using tryptic peptides as input of MGL pull-downs. However, none of those methods increased the number of glycopeptides that we could identify, leaving an open question regarding the epitopes responsible for each MGL binding protein. For c-Met specifically, we additionally tested a set of antibodies for

immunoprecipitation but none of these were successful in our hands. For our studies, we were interested in the endogenous MGL binding proteins in the CRC cell lines. Recently, others have used a glyco-engineered high Tn expressing ovarian cancer cell line to study MGL binding proteins in these cells [130]. Instead of using pull-down assays and protein extracts as input material, they made use of MGL-columns for direct glycopeptides capturing after neuroaminidase treatment. Although this allowed them to gain broader insight in the importance of multiple Tn-epitopes on individual peptides for the binding to MGL, the underlying proteins themselves have less immunological relevance.

Finally, notwithstanding the developments and accomplishments in mass spectrometric based glycoproteomics, the MS-based analysis of glycopeptides remains a difficult task, due to typical technical and intrinsic limitations e.g. low abundance, glycosylation site and glycan structure heterogeneity, low ionization efficiency [18]. Interpretation of glycopeptides by informatic tools pose additional complications in glycoproteomics. The laborious process of manual glycopeptide annotation is now facilitated by several search algorithms which perform automated glycopeptide assignment. We have used Byonic™, a search engine able to annotate both *N*- and *O*-glycopeptides from MS/MS spectra obtained by one or multiple fragmentation methods but in recent years many other algorithms have been presented (such as GlycoMAster DB, GlycoPeptideSearch and SugarQb) [69, 140], but knowledge on the pro's and con's for each of these strategies in relation to the type of data is lacking. For the most important glycopeptides, however, we always validated the assignment with manual interpretation of the spectra. In the coming years, we expect that further developments in search algorithms for glycopeptides will aid the glycoproteomics researcher with this rate limiting step in their workflows. Some developments for the assignment of glycopeptides also make use of the ratio of HexNAc specific diagnostic ions to discriminate different classes of glycopeptides [140, 147]. As an example, the O-Pair method was recently presented, which uses such information to aid with the HexNAc interpretation, supporting the annotation of GalNAc-Gal rather than Man-GlcNAc for H1N1 [142]. We have applied a similar strategy to infer and annotate single HexNAc containing peptides from a complex quantitative proteomics dataset, based on the ratio of the HexNAc oxonium ions at *m/z* 138.05496 and 144.08665. This allowed the discrimination between *O*-GlcNAcylation and *O*-GalNAcylation on a set of cellular proteins. As a result, we were the first to unequivocally assign a single *O*-GalNAc structures on intracellular AGR2, whereas more complex *O*-glycans were found on

secreted AGR2. Notwithstanding the role that secreted AGR2 plays as regulator of epithelial morphogenesis and tumorigenesis [153] and the hint towards a role of *O*-glycosylation in modulating the extracellular function of AGR2 [152], our approach clarified the nature of *O*-linked glycosylation that were previously suggested only based on band shifts in SDS-PAGE and bioinformatic prediction, respectively [152]. Moreover, it supports other recent data that showed that translocation of GalNAc transferases from the Golgi to the ER results in *O*-glycosylation of oxidoreductases in tumor cells, resulting in their translocation to the cell surface where they stimulate matrix degradation [154]. Hence, our data may aid the study of AGR2-mediated cell adhesion, and the role of *O*-linked glycosylation therein.

Molecular mechanisms and biological functions involved in protein glycosylation

Notwithstanding the importance of studying the changes in glycosylation, understanding the underlying mechanism is key for a better comprehension of the cellular alterations. In comparison to the amount of studies demonstrating changes in glycosylation, unraveling the underlying mechanisms or biological effect is lacking behind. In order to study this, the rapid advances in high throughput omics approaches play an important role. Multi-omics studies are used to analyze biological samples at different levels, such as genomics, transcriptomics, proteomics and metabolomics. However, biological contextualization, integration and interpretation of multi-dimensional data remains challenging.

In Chapter 3, we sought to correlate the higher expression of MGL binding glycan epitopes with available transcriptomics data on glycosyltransferases and overall *N-/O*-glycosylation [52-54]. However, gene expression of relevant glycosyltransferases involved in terminal GalNAc epitopes formation, such as β4GalNAc-T3 for LacdiNAc expression and T-synthase/Cosmc for Tn antigen formation did not appear to be the mechanisms driving higher MGL binding. Also, the glycomics data did not provide evidence for differences in LacdiNAc expression between the cell lines [53], whereas Tn- and STn- antigen analyses were inconclusive due to technical challenges related to their retention during chromatography, as part of the *O*-glycan analysis [54]. The structure of cellular glycan moieties is the results of a highly regulated and dynamic crosstalk between a big repertoire of glycosyltransferases and glycosidases [161]. In the *Cosmc* mutated cell model Jurkat, the inactivity of the chaperon

required for correct T-synthase activity to form T antigen is sufficient to induce a higher expression of MGL epitopes on the cell surface [39]. However, this mutation is not found either in breast cancer [42] or CRC cell lines [83]. On the other hand, CRC cell lines carrying the oncogenic mutation BRAFV600E mutation display increased levels of MGL ligands associated with the higher expression of enzyme N-acetylgalactosaminyltransferase 3 (GALNT3) [83]. A cell model carrying a GALNT3 mutation, resulting in no functional GALNT3 expression, was used in order to confirm this finding and indeed showed lower MGL binding in flow cytometry [83]. However, these models have to be used with caution knowing that the glycosylation machinery is highly redundant and characterized by competitions of enzymes for the same substrates. In fact the absence of one enzyme could result in a new glycan epitope expression as consequence of the competing enzymatic role of glycosyltransferases in the cell [162].

In Chapter 4, we aimed to dissect the mechanisms behind the higher MGL ligands expression in HCT116 and HT29 compared to LS174T by comparative quantitative proteomics. Although based on these data we could rule out differential expression of the MGL ligands in general, and confirmed a correlation with higher GALNT3 expression in HT29 cells, the overall proteomics data covered only 30% of the 245 glycosylation-related proteins, leaving a role of other relevant glycosyltransferases elusive. Hence, we could not attribute a specific mechanism responsible for the malignant phenotype binding to MGL, which appears rather unique per cell line. Based on the GALNT3 results, we used the GALNT3 KO cell model (in HT29) [83], to perform MGL pull-downs in the attempt of shedding light on ligands binding to MGL due to aberrant *O*-glycosylation. However, within the top 25 MGL ligands shown in Chapter 3, we did not observe significant differences in MGL binding between mock and deficient GALNT3 cells (data not shown). Moreover, when studying overall proteomics changes in two independent clones of HT29 GALNT3 knockout cells, we were regrettably confronted with large proteomics changes related to interferon responses (data not shown), and we deemed this most likely to be due to side effects of the CRISPR-Cas technology that was used to generate these cells, but this warrants further investigation. In the future, targeted MS approaches could be used to get more information on the differences in levels of the enzymes involved in the glycosylation machinery in the CRC cell lines that could explain their differential binding to MGL. However, the final shaping of glycoconjugates depends not only on the expression levels of enzymes involved in glycosylation but also on their activity,

location and substrate availability. For this reason, the overall characterization of genes, transcripts, and proteins expression is often not sufficient to explain glycosylation changes in cancer.

Our quantitative proteomics analyses showed lower levels of CDX2 in the two high MGL binding cell lines (HCT116 and HT29) compared to the lower MGL binding cell line, LS174T. CDX2 is a transcription factor known to be involved, together with CDX1, Hepatocyte Nuclear Factor (HNF) 1α and 4α, in the regulation of intestinal differentiation genes, MUC2, FUT2 and ST6GalNAc-I [117]. Given this knowledge, we postulated a hypothesis on the role of CDX2 in the modulation of downstream targets such as genes encoding proteins involved in MGL binding motifs. Multiple studies have shown the association between glycosylation changes and transcriptional factors [53, 163, 164]. However, aberrant glycosylation and altered signal transduction have a bidirectional relationship. In fact, on the one hand alteration of signal transduction pathways may directly influence the levels of glycosylation enzymes ultimately leading to altered glycosylation. On the other hand, tumor associated carbohydrate antigens on cell surface receptors can regulate cell signaling resulting in the alteration of cancer cell behavior [163]. This is the case of epidermal growth factor receptor (EGFR), part of the ERBB receptor family. This receptor homo- or heterodimerizes upon ligand binding with consequent activation of intracellular signaling, mainly through mitogen-activated protein kinase (MAPK), Janus kinase (JAK)/signal transducer and activator of transcription (STAT) and phosphatidylinositol 3-kinase (PI3K)/protein kinase B (Akt) pathways. For example, α2,6-sialylation of EGFR decrease ligand binding, tyrosine phosphorylation and intracellular signaling in a colon cancer cell model [164]. Similarly, modulation of glycosylation of the hepatocyte growth factor receptor (c-Met) has been reported to regulate the activity of the receptor. C-Met is a transmembrane tyrosine receptor, whose activation is mediated by HGF binding and subsequent Tyr-1234/1235 phosphorylation [165]. Deficiency of ST6GAL1 reduces α2,6 sialylation on c-Met and abolishes motility of HCT116 cells, independent of hepatocyte growth factor (HGF) engagement [166]. On the other hand, higher expression of C1GALT1, the enzyme elongating Tn antigen, enhances HGF-induced phosphorylation and activation of c-Met, thereby promoting proliferation of hepatocellular carcinoma cells [167]. In Chapter 3, we found c-Met as one of the major cell surface MGL binding proteins and, in Chapter 4, we demonstrated that its binding is due to differential *N*-glycosylation carried by high MGL binding cell lines, which, at the same time, are characterized by increased c-Met

phosphorylation. However, the direct influence of *N*-glycosylation on c-Met activation remains to be addressed, having as limitation factor the lack of knowledge of the localization of the specific c-Met glycosylation responsible for MGL binding. In fact, manipulation of site-specific glycosylation in cell models could be used to study c-Met signaling not only relative to the mere glycosylated epitope but also in combination with lectin recognition and binding. Although current literature focuses extensively on the immune response mediated by MGL-ligand binding, such studies could show potential effects of MGL binding in the tumor cells instead. Recent evidences of ligand-specific conformational changes in the MGL CRD [35] and different binding efficiencies to the secondary binding site [36] support the intriguing possibility that each individual MGL ligand may activate different signaling cascades in MGL expressing DCs and promote the establishment of glycoproteomic analyses in cancer research. Finally, future research on the biological functions MGL ligands and their validation in CRC tissues and/or body fluids of CRC patients is expected to shed more light on the immune suppressive role of MGL during tumor progression and cancer immunotherapy.

Conclusion and future perspectives

The results presented in this thesis contribute to a better understanding of glycoproteins aberrantly glycosylated in tumor cells, providing novelty in technology as well as in glycobiology and glycoproteomics. Despite the recent advances in cancer biomarker discovery, personalized medicine and population screening approaches, CRC remains a major cause of death worldwide. Characterization of MGL ligands in CRC cell lines can offer valuable insights into understanding the mechanisms triggered by MGL recognition in this tumor. However, the validation of given candidates and investigation of MGL-mediated intracellular signaling and their functional effects, both in tumor and immune cells, are yet needed. As an example, one approach that was not applied in our work is represented by the proximity ligation assay (PLA). This method is based on the use of two specific probes (antibodies or lectins), one targeting the peptide backbone and one specific for a glycan epitope. If binding to both targets in close molecular proximity, the two probes will hybridize with the help of bridging oligonucleotides and a ligase. A polymerase amplifies the forming closed nucleotide in combination with labeled nucleotides which leads to a fluorescent or chromogenic signal allowing the *in situ* detection of the co-expression [168]. The PLA technique can be used as

tool to validate suspected protein glycosylation not only in cell lines but also CRC patient tissues.

The clinical relevance of MGL as prognostic marker for stage III CRC tumors and as target of antitumoral treatments sets challenges towards the translation of our *in vitro* studies into clinical ones. In fact, MGL ligands discovery using patients' cancer tissues as material sources could bring new highlights on the molecular mechanisms responsible of immune suppression and set advances for personalized immunotherapeutic treatment. Stage III tumors are characterized by local lymph-node invasion that allows cancer cells to come into contact with the peripheral immune system [45]. DCs subsets in lymph-nodes, suppress immune responses after MGL ligands recognition on cancer cells resulting in immune evasion of tumor cells [39]. The reduced glycolytic activity of MGL-stimulated DCs was recently suggested as mechanism associated to the anti-inflammatory and pro-tumoral activity [169], which can contribute to the worse prognosis of CRC patients [45]. When MGL recognizes and binds terminal GalNAc carrying structures, oligomerization of the lectin occurs with subsequent ligand internalization, processing, presentation in MHC class II (called HLA II in human) and T cell response [34]. Notwithstanding the crystal structures of the human MGL CRD in complex with GalNAc [170] and the fact that ligand affinity and avidity are crucial for the clustering of MGL and its endocytic activity [41], the study of MGL ligands is essential for the design of immunotherapies based on targeting MGL-expressing DCs. In fact, using different glycan epitopes (LacdiNAc/Tn), modifying Tn-density and/or steric structure of peptide backbone can activate different intracellular signaling and immune responses. On the other hand, targeted therapies could aim the inhibition mechanisms behind MGL ligands expression in stage III CRC patients, which remains still unclear. For instance, while BRAF mutation seems to have a correlation with MGL binding, only 10 % of CRC patients carries this variation, suggesting that various oncogenic alterations may modulate the expression of tumor associated antigens. An exhaustive characterization with different omics-platforms of cancer tissues from patients with poor prognosis could unravel not only the interactions between MGL and cancer cells highly expressing MGL ligands but also the mechanisms driving this process. Finally, the in depth and thorough investigations of three-dimensional (3D) cultures, such as organoids [171], and ultimately CRC cancer tissues will guarantee the correct interpretation of *in vitro* research for clinical applications.