UNIVERSITY OF
GOTHENBURG

Doctoral thesis for the Degree of Doctor of Philosophy, Faculty of Medicine

Study of the Colonic Mucus Layer by Mass Spectrometry

Sjoerd van der Post

Institute of Biomedicine
Department of Medical Biochemistry
Sahlgrenska Academy
University of Gothenburg
2014

A doctoral thesis at a University in Sweden is produced either as a monograph or as a collection of papers. In the latter case, the introductory part constitutes the formal thesis, which summarizes the accompanying papers. These have already been published or are manuscripts at various stages (in press, submitted, or manuscript).

ISBN 978-91-628-9246-3
http://hdl.handle.net/2077/36909

Sjoerd.van.der.Post@medkem.gu.se
University of Gothenburg
Institute of Biomedicine
Sahlgrenska Academy
SWEDEN

Printed by Ale Tryckteam
Bohus, Sweden 2014

Voor mijn familie

ABSTRACT

Sjoerd van der Post
Department of Medical Biochemistry, Institute of Biomedicine
Sahlgrenska Academy at the University of Gothenburg

The mucus covering our internal mucosal surfaces is a part of the innate immune system, and the first line of defense against microbial challenges. The need of an efficient defense system is especially important in the lower parts of the digestive tract where the microbiota reaches its highest density. In the colon, the mucus forms a dense layer that prevents bacteria from accessing the epithelial surface. The gel-forming mucin 2 (MUC2) is the major structural component of the colonic mucus layer, forming large net-like structures by oligomerization in the N- and C-terminal regions. A dysfunctional mucus layer that allows bacteria to pass through and access the underlying epithelium has been associated with inflammatory bowl diseases such as ulcerative colitis. However, detailed understanding of the molecular mechanisms behind the defective mucus layer is lacking. This lack of knowledge can largely be explained by the limited information regarding the composition and processing of the mucus during normal conditions. This thesis aims to broaden the knowledge regarding the protein composition of the human colonic mucus, and the molecular properties of the heavily glycosylated MUC2 mucin.

Proteomic and mass spectrometry approaches were used to characterize the composition of the human colonic mucus layer in health an disease, and to determine how alterations in protein abundance and modification of the MUC2 mucin affect the function of the mucus gel. Our results showed that the human colonic mucus is comprised of approximately 50 proteins. The protein composition of the mucus layer was shown to be unaffected in patients with ulcerative colitis, though the relative abundance of 13 mucus proteins including the structural components MUC2 and FCGBP were shown to be decreased during active disease.

The mucin protein family is characterized by a heavily *O*-glycosylated core that is resistant against proteolytic degradation. However, our results showed that the C-terminal part of the protein is also modified by *N*- and *O*-glycans, and that site specific *O*-glycosylation plays an important role in protecting the protein from proteolytic degradation by bacterial proteases. In addition, we could correlate the relative abundance of various glycosyltransferases required for *O*-glycosylation in the different parts of the colon, to the previously characterized segmental pattern of terminating glycans on the MUC2.

Taken together, the results from this thesis show that the human colonic mucus is composed of a relatively small number of proteins that are organized around the heavily *O*-glycosylated MUC2 mucin, and suggests that decreased amounts of the core mucus proteins in combination with impaired *O*-glycosylation of the MUC2 renders the mucus layer more permeable to bacteria and susceptible to proteolytic degradation.

Key words: MUC2, mucin, intestine, proteomics, mass spectrometry
ISBN: 978-91-628-9246-3

LIST OF PAPERS

This thesis is based on the following papers, referred to in the text by their Roman numerals.

I. van der Post, S., Jabbar, K.S., Sjövall, H., Johansson, M. E. V., and Hansson G.C. **The protein composition of the human colonic mucus: reduced levels of core structural components in active ulcerative colitis.** *Manuscript*

II. van der Post, S*., Subramani, D. B*., Bäckström, M., Johansson, M. E. V., Vester-Christensen, M. B., Mandel, U., Bennett, E. P., Clausen, H., Dahlén, G., Sroka, A., Potempa, J., and Hansson, G. C. (2013) **Site-specific *O*-glycosylation on the MUC2 mucin protein inhibits cleavage by the *Porphyromonas gingivalis* secreted cysteine protease (RgpB).** *Journal of Biological Chemistry* **288**, 14636–14646. *Equal contribution

III. van der Post S., Thomsson K. A., and Hansson G. C. **A multiple enzyme approach for the characterization of glycan modifications on the C-terminus of the intestinal MUC2 mucin.** *Journal of Proteome Research in press.*

IV. Ambort, D., van der Post, S., Johansson, M. E. V., Mackenzie, J., Thomsson, E., Krengel, U., and Hansson, G. C. (2011) **Function of the CysD domain of the gel-forming MUC2 mucin.** *Biochemical Journal* **436**, 61–70

V. van der Post S., and Hansson G. C. (2014) **Membrane protein profiling of human colon reveals distinct regional differences.** *Molecular & Cellular Proteomics,* **13**, 2277-2287

CONTENTS

ABBREVIATIONS

CID	Collision-induced dissociation
CLCA1	Calcium-activated chloride channel regulator 1
DDA	Data dependent acquisition
DIA	Data independent acquisition
ECD	Electron-capture dissociation
ER	Endoplasmic reticulum
ESI	Electro spray ionization
ETD	Electron transfer dissociation
FASP	Filter-aided sample preparation
FCGBP	IgG Fc-gamma binding protein
FDR	False discovery rate
GalNAc	*N*-acetylgalactosamine
GalNAc-T	*N*-acetylgalactosamine-transferase
GI	Gastrointestinal
GuHCl	Guanidine hydrochloride
HCD	Higher-energy collisional dissociation
HILIC	Hydrophilic interaction-liquid chromatography
PNGase F	Peptide *N*-glycosidase F
PTS-domain	Proline, threonine and serine rich domains
PTM	Post-translational modification
RELMβ	Resistin-like molecule beta
RgpB	Arg-gingipain B
RP	Reverse phase
SDS-PAGE	Sodium dodecyl sulfate-polyacrylamide gel electrophoresis
SRM	Single reaction monitoring
TFF3	Trefoil factor 3
LC	Liquid chromatography
MALDI	Matrix-assisted laser desorption ionization
MS	Mass spectrometry
MS/MS	Tandem mass spectrometry
MUC	Mucin
TOF	Time-of-flight
UC	Ulcerative colitis
vWF	von Willebrand factor
VWD	von Willebrand D-domain
XIC	Extracted ion-chromatogram
ZG16	Zymogen granule protein 16

BACKGROUND

On a daily basis we are continuously exposed to infectious and toxic substances that can be harmful and cause disease if not handled in the right way. To protect ourselves from these imminent threats our immune system has developed various strategies to prevent development of disease. The first line of defense is the actual prevention of contact and uptake of any pathogens, which is established by physically separating the inside of our body from the outside world. The way this separation occurs varies depending on the exposed organ, for example the skin is covered by multiple layers of dead keratinized cells, protecting it from physical damage and pathogens. On surfaces where active transport of nutrients and gasses occur a different strategy is applied. These epithelial surfaces are instead covered by a viscous layer of proteins referred to as mucus. This layer can be found covering the epithelial cells lining the digestive tract, respiratory system, reproductive organs and the urinary tract. In the respiratory tract mucus is involved in trapping pathogens and particles while still allowing active exchange of gasses and preventing the underlying tissue from drying out. The mucus in the gastrointestinal (GI) tract is produced and secreted by specialized secretory cells called goblet cells. The secreted mucus is adapted to the function of the respective organ and varies in composition and thickness along the length of the GI tract. In the stomach the mucus protects the epithelial cells from the acidic environment, whereas in the small and large intestine the mucus mainly functions as a protective barrier limiting the interaction between the commensal flora and the epithelium. The secretion and formation of the various types of mucus is tightly regulated and when abnormalities occur, commensals and pathogens can breach the mucus barrier, which facilitates invasion of the underlying epithelium. Increased epithelial bacterial interactions will trigger a response from immune cells in the *lamina propria*, resulting in development of acute or chronic inflammation depending on the underlying mechanisms behind loss of barrier function. In this model the mucus functions as the first line of defense in prevention of infection and inflammation.

The intestines and the mucus barrier

The main functions of the small and large intestines are to aid in food digestion, allow efficient absorption of nutrients, ions and fluids and function as a protective barrier against all the potentially harmful substances and microorganisms that pass through our digestive tract. The human intestine is covered by a single layer of fast renewing cells in both the small and large intestine aiding in these processes. The epithelium is composed of proliferative crypts, which contain intestinal stem cells that differentiate into specialized cell types (Figure 1). In addition to the crypts the small intestine has protruding villi to increase the absorptive surface area of the epithelium. Stem cells are found at the base of each crypt, proliferating while migrating along the crypt and/or villi and differentiating into four different cell types, renewing the complete epithelial cell layer every 4 – 5 days (van der Flier & Clevers, 2009). The majority of the cells in the intestine are absorptive enterocytes required for the absorption of nutrients, ions and water. In addition three types of secretory cells are found, Paneth cells, enteroendocrine and goblet cells. Paneth cells are exclusively found in the small intestine and remain along the base of the crypt producing antimicrobial proteins to keep the lower crypt sterile. The enteroendocrine cells

produce various hormones secreted at the basolateral side involved in signaling via the bloodstream and nervous system. Mucus components are solely produced by the goblet cells which increase in number along the length of the intestine (Karam, 1999).

In addition to the exposure to potential pathogens, our intestines are also harbor the resident commensal microbiota that are found in numbers exceeding trillions of over more than a 1,000 different bacterial species. These bacteria assist in the final stage of digestion, synthesize essential vitamins, and promote good host physiology. The highest bacterial density is found in the large intestine, and the complexity and diversity of the gut microbiota has only recently been fully resolved (Arumugam *et al.*, 2011). Although the resident microbiota plays an important role in promoting host physiology and health, these microorganisms are potentially harmful, and need to be handled in a correct way. The increasing bacterial load along the proximal to distal axis is reflected in the increased thickness and density of the mucus layer along the length of the GI tract (Luckey, 1972; Ermund *et al.*, 2013). In the small intestine, where the majority of nutrient absorption takes place, the mucus forms a loose and permeable structure that allows for efficient uptake. In the large intestine the mucus is composed of two distinct layers; an outer loose and permeable layer that harbors the commensal flora, and a thinner inner layer that is adherent to the epithelium and devoid of bacteria (Johansson *et al.*, 2008).

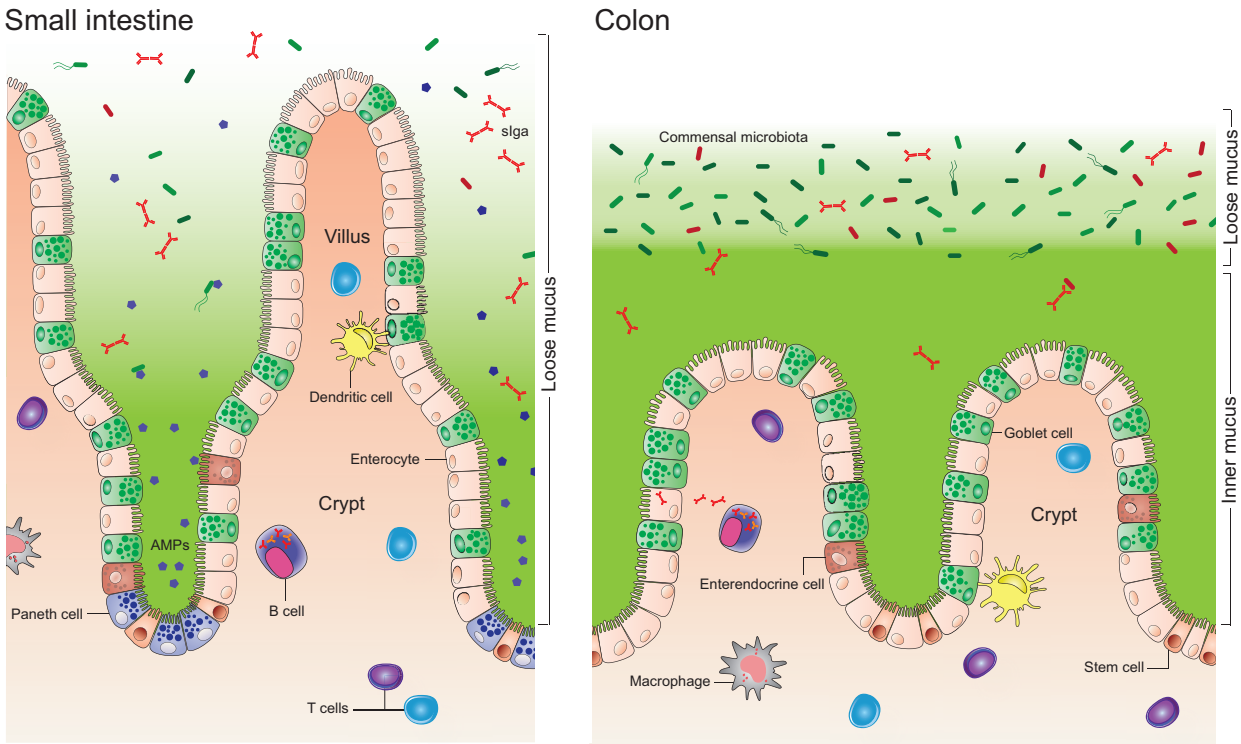

Figure 1. Overview of the intestinal mucosa in small intestine and colon. The mucus barrier in the small intestine is composed of single loose layer accessible for the bacteria, while there is as two layer system in the colon with an inner layer that is dense and devote of bacteria and a loose layer that harbors the commensal flora.

The colonic mucus is secreted from the goblet cells as highly organized stratified layers, which are impenetrable for bacteria. Over time conformational changes occur in the polymers that loosens their structure resulting in a less organized matrix with larger pore sizes (Johansson *et al.*, 2014). This structure allows bacteria to enter and is referred to as the loose layer. Since mucus is

constitutively secreted, the dense inner layer is constantly renewed ensuring sufficient protection of the underlying epithelium. The protein responsible for the core structure of the intestinal mucus gel is the MUC2 protein, a heavily *O*-glycosylated and extensively disulfide linked protein that is highly resistant to the harsh environment in the intestinal lumen. The importance of this protein in epithelial defense became evident when it was shown that MUC2 deficient mice that lack intestinal mucus develop spontaneous colitis around the time of weaning (Van der Sluis *et al.*, 2006). In addition, mice that lack the core 1-type glycosyltransferase which results in limited oligosaccharide extensions on the Muc2 protein, and mice with mutations in the Muc2 gene also develop spontaneous colitis, which further support the importance of the mucus layer in maintenance of intestinal homeostasis (Heazlewood *et al.*, 2008; Fu *et al.*, 2011).

Mucins

Proteins from the mucin glycoprotein family are selectively found on epithelial cells in all vertebrates and can be separated into two categories; secreted mucins and transmembrane mucins. The secreted mucins are involved in the formation of the mucus layers that covers the epithelial surfaces, and the membrane bound mucins protect the apical epithelial surface by forming the glycocalyx, and potentially act as sensors for the luminal milieu (Hattrup & Gendler, 2008; Johansson *et al.*, 2011). The main feature that distinguishes proteins of this family is the potential to become heavily *O*-glycosylated, typically contributing to over >80% of the glycoproteins molecular mass. All members of the mucin family have large repeated sequences of the amino acids serine, threonine and proline, so called PTS-domains, which are highly modified by *O*-glycosylation. The number of tandem repeats varies between the different proteins and contributes to the individual protein properties. The high density of *O*-glycans limits formation of secondary structures, resulting in long linear protein stretches that extend perpendicular from the cell membrane in the case of transmembrane mucins, or form large sheets in the case of secreted mucins.

Secreted mucins

The human gel-forming mucin family encompasses MUC2, MUC5AC, MUC5B, MUC6, MUC7 and possibly MUC19 that all lack transmembrane spanning domains, forming large oligomeric complexes with the exception of MUC7 that is secreted as a monomer in saliva. MUC19 is suggested to be expressed in human although has only been identified at protein level in mice, pigs and horses (Rousseau *et al.*, 2008). These proteins all have similar domain structures and their protein core is composed of a PTS-domain(s) which are highly *O*-glycosylated and distinguish the mucin protein family (Perez-Vilar & Hill, 1999). *O*-glycosylation of the central protein domain has a double role, firstly; negatively charged sugars bind water, which is essential for the gel forming properties of the mucus gel and secondly; the *O*-glycans protect the protein backbone from proteolytic degradation (Loomes *et al.*, 1999). The *O*-glycans are estimated to contribute to 50 – 90% of the proteins mass, which highlights the extensiveness of the glycosylation. Both protein termini are composed of multiple von Willebrand domains, which are involved in intermolecular oligomer formation (Vischer & Wagner, 1994). In the case of MUC2

which is found in the intestines, trimers are formed between N-termini, while the C-termini forms dimers, generating sheets of ring-like structures (Asker *et al.*, 1998; Lidell *et al.*, 2003b). Other secreted mucins such as MUC5B found in the respiratory tract form linear polymers by dimerization at both termini (Ridley *et al.*, 2014). Additional intramolecular disulfide bonds are formed at both termini between the highly number of cysteine residues, which add to the rigid structure giving further resistance to proteolytic degradation. All features combined results in a highly organized oligomer, which serves both as a lubricant and as a protective layer that is highly resistant to both endogenous and bacterial proteases.

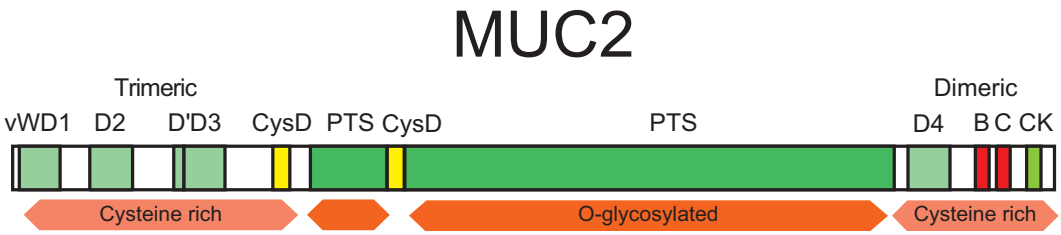

Figure 2. Domain organization of the MUC2 mucin and the specific features of the different regions.

MUC2

The MUC2 mucin is highly expressed and secreted in the small and large intestine, and is considered to be the main structural contributor of the intestinal mucus gel (Gum *et al.*, 1989; Carlstedt *et al.*, 1993). MUC2 was the first human gel-forming mucin to be partly sequenced, and is composed of an estimated 5,179 amino acids organized in an N-terminal region, two PTS-domains and a C-terminal region (Figure 2). The N-terminal region spans 1,400 amino acids with three complete and one truncated von Willebrand D domains (VWD). The N-terminus is followed by one small and one large PTS-domain of which the larger one is composed of approximately 100 tandem repeats of the consensus sequence PTTTPITTTTTVTPTPTPTGTQT, giving it a total length of around 2,300 amino acids (Toribara *et al.*, 1991). The C-terminal region is comprised of 840 amino acids spanning one VWD domain, two shorter von Willebrand B and C domains and a cystine-knot. In addition, two CysD domains are found on both sides of the small PTS-domain. The CysD domains are almost exclusively found in secreted mucins. One additional prominent feature of the terminal regions of the MUC2 is the high frequency of cysteine residues (1 out of 8 amino acids) that are responsible for formation of intra- and intermolecular disulfide bonds. The VWD domain in the proteins' C-terminus contain a GDPH motif which undergoes autocatalytic cleavage between the aspartic acid and proline under acidic conditions, resulting in a reactive C-terminus potentially cross linking the mucin (Lidell *et al.*, 2003a). However, it is not known how and when the GDPH cleavage is triggered and if the attachment sites are random or if there is specificity. In the case of heavy chain 3 (ITIH3) autocatalytic cleavage of the GDPH motif resulted in the formation of covalent bond with *N*-acetylgalactosamine (GalNAc), which is as well a potential candidate in the mucus layer rich in glycoproteins (Kaczmarczyk *et al.*, 2002).

The MUC2 N- and C-terminal regions show large sequence similarity with the blood glycoprotein von Willebrand factor (vWF) (Sadler, 1998). The vWF is involved in hemostasis by

mediating platelet adhesion to connective tissue, and by binding blood clotting factor VIII. The absence or a dysfunctional vWF lead to bleeding disorders, and the protein has therefore been much more intensively studied compared to gel-forming mucins. Hence, the majority of the structural knowledge regarding oligomerization of the MUC2 and the role of the various domains is based on its homology with the vWF protein (Huang *et al.*, 2008; Dang *et al.*, 2011).

MUC2 biosynthesis

The main role of the intestinal goblet cells is production and secretion of the MUC2 mucin, which is the main component of the intestinal mucus layer. Secreted mucins can be considered among the most complex proteins synthesized by human cells due to their extensive glycosylation, high number of disulfide bonds, intracellular oligomerization, and long-term storage in secretory granules. This requires cells with a specialized secretory machinery and is the reason why most cell lines cannot be used for the production of recombinant MUC2 (Bäckström *et al.*, 2013). In the goblet cell, the protein is directed to the ER by its signal peptide, where it becomes *N*-glycosylated (high-mannose type), and forms homo-dimers via its C-terminal (Figure 3). The protein holds 30 potential *N*-glycosylation consensuses sequences that are likely involved in protein folding and are required for dimerization. Inhibition of the *N*-glycosylation pathway results in accumulation of the protein in the ER, and mutations of selected aspartic acids in the cystine-knot has also been shown to prevent dimer formation (Asker *et al.*, 1998; Bell *et al.*, 2003).

Following dimerization the protein enters the Golgi were the *N*-glycans are further processed, and the PTS domains become *O*-glycosylated mucin domains. *O*-glycosylation is initiated by addition of GalNAc to serines and threonines on the protein backbone by members of the UDP-*N*-acetylgalactosamine-polypeptide *N*-acetylgalactosaminyl-transferases enzyme family (GalNAc-T's). The GalNAc-T enzyme family contains twenty different members, all described to be involved in initiating *O*-glycan synthesis. These transferases have different substrate specificity, and have been shown to be expressed in a cell and developmental stage specific manner (Bennett *et al.*, 2012). After addition of the first GalNAc the protein passes through the Golgi compartments where additional monosaccharide residues are added. The first step in elongation of the *O*-glycan is the core formation, followed by chain extension and finally addition of terminal monosaccharaides (Jensen *et al.*, 2010). The majority of the core extensions found in the human colon are based on core-3 and core-4 structures (Robbe *et al.*, 2004; Holmen Larsson *et al.*, 2009). The oligosaccharide chain is then further extended with galactoses and *N*-acetylglucosamines, varying in length from 2 up to 12 residues (Holmen Larsson *et al.*, 2009). In the *trans*-Golgi network the polysaccharide extension is terminated by addition of sialic acid or GalNAc. Additionally, specific residues can also be sulfated, fucosylated or acetylated. The resulting oligosaccharides show large heterogeneity in chain-length, composition and terminal epitopes, the profile of which can change in time, upon infection or in inflammatory bowel diseases (Larsson *et al.*, 2011). More than 100 different *O*-glycan structures have been identified on MUC2 isolated from the human small and large intestine (Robbe *et al.*, 2003). The large diversity in glycan epitopes has been suggested to serve as targets for microbial adhesins, allowing selection of beneficial microbial species and thereby preventing pathogens from adhering to the mucus gel (Hooper & Gordon, 2001; Staubach *et al.*, 2012). As the epitopes vary along the GI

tract the host creates niches for selected bacterial species to adhere. The main variation occurs in the terminal epitopes, increasing the acidity of the glycans towards the distal colon by increasing levels of sialylation, an opposing gradient of fucosylation and sulfation appears towards the small intestine in human (Robbe *et al.*, 2003). When completely glycosylated each MUC2 monomer has a mass of ~2.5 MDa where 80% of the mass is due to the added glycans, occupying over 70% of the serines and threonines in the PTS-domain (Carlstedt *et al.*, 1993). In the late Golgi the MUC2 N-terminal forms disulfide linked homo-trimers in the VWD3 domain resulting in large oligomers (Godl *et al.*, 2002). The VWD1 and 2 domains are further responsible for directing the protein to storage granules were it is densely packed on a ring like oligomeric platform in a high calcium and low pH dependent manner (Ambort *et al.*, 2012). Mucins are stored in secretory vesicles for extended periods of time before secreted into the intestinal lumen, occupying most of the apical cytoplasm. The exact mechanisms by which mucin exocytosis is triggered are only partly resolved, however, the process is driven by increased levels of intracellular calcium resulting in fusion of mucin vesicle to the plasma membrane and release of the stored protein. Recent studies by our group have shown that the densely packed mucins expand in a pH- and calcium-dependent matter into the lumen as large net-like sheets (Ambort *et al.*, 2012; Gustafsson *et al.*, 2012b). Upon exocytosis the densely packed MUC2 expands in volume approximately a 1,000 times to form the mucus layers (Verdugo *et al.*, 1991).

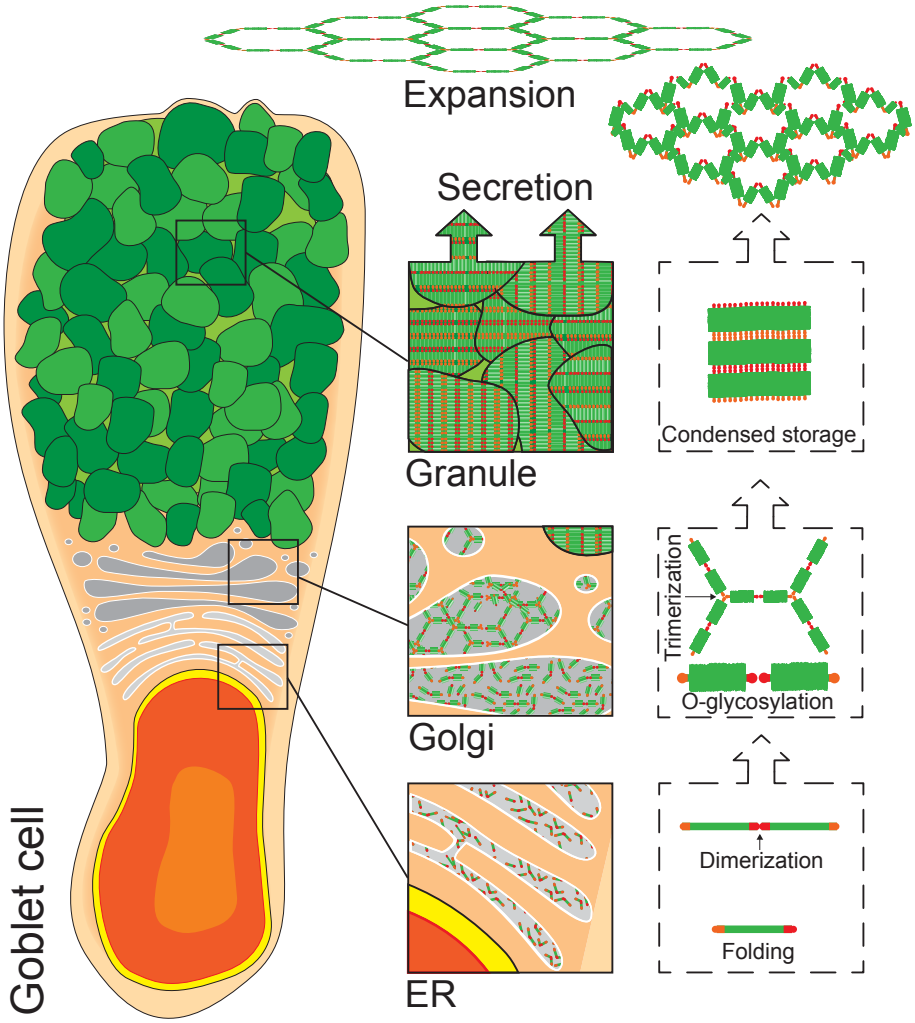

Figure 3. The goblet cell is responsible for the biosynthesis of MUC2. Highlighted are the various steps of the oligomerization process.

Mucus protein composition

Structural and granule specific proteins

Mucus is a heterogeneous mixture of molecules composed of approximately 95% water, while electrolytes, carbohydrates, proteins, amino acids and lipids make up the remaining part. The main structural component forming the intestinal mucus gel is the MUC2 mucin, however, immunohistochemistry and proteomics studies have shown that the intestinal mucus contains several hundred proteins. Not all of the identified proteins are considered to be an intrinsic part of the mucus layer since mucus retains exfoliated cells, and traps materials that passes through the digestive tract (Johansson *et al.*, 2009). This results in a complex mixture of intracellular, food derived, bacterial and actual mucus associated proteins, which has complicated the study of the protein composition and only limited information is available on the proteins that are required for a functional mucus barrier. The proteins that make up the actual mucus gel can be grouped based on their function into three categories, structural components, antimicrobial proteins, and proteins with regulatory functions. In addition to MUC2, the only other protein that is suggested to be a structural component of the mucus is the IgG Fc-gamma binding protein (FCGBP). This large protein is expressed in most mucin expressing cells and was initially reported to selectively bind IgG antibodies at the Fc region (Kobayashi *et al.*, 2002). However, the protein sequence contains 13 VWD domains, which are mainly found in proteins forming oligomeric structures suggesting that it has additional roles. Most of the VWD domains include an autocatalytic GDPH motif, where studies have shown that extensive washing of collected mucus with chaotropic agents did not result in loss of FCGBP which indicates that the protein is covalently linked to MUC2 (Johansson *et al.*, 2009). As FCGBP is found in the mucus granules and is secreted simultaneously to MUC2, it is hypothesized to form heteromers with MUC2 via the reactive anhydrides formed after GDPH cleavages in FCGBP. Only a few proteins are known to be localized to the mucin granules, trefoil factor 3 (TFF3) a protein disulfide linked to FCGBP that is required to maintain the integrity of the mucosal barrier after epithelial damage (Albert *et al.*, 2010), calcium-activated chloride channel regulator 1 (CLCA1) (Komiya *et al.*, 1999), the recently identified resistin-like molecule beta (RELMβ) and zymogen granule membrane protein 16 (ZG16) which will be discussed in the next section. CLCA1 was initially suggested to form an ion channel but is now believed to regulate the secretory capacity of other channels (Yurtsever *et al.*, 2012). Studies have also shown that CLCA1 drives mucus secretion in mice and horses, although the mechanism by which it regulates mucus secretion is unclear. RELMβ is secreted into the mucus as hexamers and trimers protecting against worm infections by limiting their motility (Patel *et al.*, 2004; Herbert *et al.*, 2009). Limiting the movement of the parasitic worms will trap them in the mucus while the peristalsis in the colon will move the parasite in the distal direction.

Antimicrobial components

The dense mucus gel in the colon limits the ability of the microbiota to reach the epithelium (Johansson *et al.*, 2008). However, in the small intestine where the mucus is permeable and non-adherent, the epithelial cells secrete proteins with antimicrobial properties to prevent bacterial invasion. The best characterized family of antimicrobial proteins found in the intestines are the

defensins, a family of small cationic proteins able to disrupt the cell membrane of bacteria and fungi (Ayabe *et al.*, 2000). Defensins and other antimicrobial proteins (*e.g.* lysozyme) are secreted by specialized Paneth cells at the bottom of the intestinal crypts, or by infiltrated neutrophils. As Paneth cells are selectively found in the small intestine and not in the colon, the presence of antimicrobial proteins and peptides is higher in the small intestine compared to the colon. The looser non-adherent mucus in the small intestine requires more active defense measures than for the dense layer found in the large intestine. Here the only protein that plays an active role in preventing bacteria from reaching the epithelium is ZG16 a small lectin-like protein secreted from the goblet cell granule (Tateno *et al.*, 2012). Recent work by our group has shown that ZG16 binds to Gram-positive bacteria, not actively killing them, but forming aggregates which limit further movement in the mucus (Bergstrom *et al.* unpublished).

Plasma cells in the *lamina propria* underneath the intestinal epithelium are responsible for production of large amounts of immunoglobulin A (IgA) found in the mucus (Johansen & Kaetzel, 2011). sIgA is transported through the enterocytes via pIgR and into the mucus layer and the intestinal lumen, forming the first line of antigen-specific immune defense recognizing both pathogens and commensals. Studies have suggested that the expression of pIgR is directly regulated by the commensal flora which thereby controls the IgA level in the mucus, since every transcytosis consumes one pIgR molecule (Hapfelmeier *et al.*, 2010).

In addition to the above described proteins there are membrane proteins that are cleaved from the epithelium or shed into the mucus such as the transmembrane mucins, most of which contain a sea urchin sperm domain (SEA) that breaks upon mechanical force (Pelaseyed *et al.*, 2013). With the increased interest in mucus and development of proteomics techniques it is expected that more components will be identified in the coming years.

Role of the colonic mucus in ulcerative colitis

Ulcerative colitis (UC) is one of the two principal types of inflammatory bowel diseases affecting the large intestine. The disease involves chronic relapsing inflammation of the colonic mucosa that originates in the distal colon and progresses in the proximal direction. The underlying etiology is unknown, but the disease is increasing in frequency in developing countries and suggested to be caused by a combination of genetic and environmental factors. Genome wide association studies have not identified specific genetic factors underlying UC, although certain loci are associated with an increased susceptibility for UC (Danese & Fiocchi, 2011; Khor *et al.*, 2011). The general hypothesis is that a genetically predisposed individual in combination with external factors will develop inappropriate immune responses towards the commensal flora. This hypothesis is supported by studies of monozygotic twins with UC showing that only in 10% of the cases both individuals develop UC, highlighting the importance of external factors such as diet, smoking habits and the use of antibiotics (Tysk *et al.*, 1988). Furthermore, all genetically engineered mouse models of UC do not develop colitis when raised under germ free conditions, suggesting that the commensal microbiota is responsible for driving the inflammation (Sartor, 2008). In UC patients alterations have been observed in microbial composition, although no strain was specifically linked to development of disease (Qin *et al.*, 2010). As the microbiota resides in the outer mucus layer, and can stimulate mucus secretion there is an potential

relationship between UC and the mucus layer, as a defect in this synergistic system will result in increased immune responses from the underlying *lamina propria*. Studies have shown that UC patients with active disease have a thinner mucus layer that is *O*-glycosylated with shorter glycan chains with less sulfated epitopes (Pullan *et al.*, 1994; Corfield *et al.*, 1996; Larsson *et al.*, 2011). The microbiota uses the mucin glycans as an energy source by secreting glycosidases that slowly degrade the mucus gel. Shorter glycan chains will lead to faster exposure of the protein core and more rapid degradation of the protein backbone. This hypothesis is supported by the recent observation that the mucus gel in UC patient is more permeable to bacteria sized beads when compared to control patients (Johansson *et al.*, 2014). The percentage of the mucus layer that was accessible by the microbiota was significantly increased in active UC, and this discontinuity appeared to increase with severity of UC. Overall, these studies suggest that the mucus layer is an important factor in development of UC, by preventing interaction between the host and the commensal microbiota. However, little is known concerning the underlying mechanisms, and whether an altered mucus layer is causing the disease or is secondary to the inflammatory process. One potential reason for a less structured mucus layer is alterations in the protein composition, a question that we addressed in this thesis work by studying the mucus protein composition in various stages of UC by the use of mass spectrometry.

PROTEOMICS

Proteomics is the generic term coined for the large-scale study of proteins, which includes the determination of their identity, quantity, modifications and interactions. This potentially allows the study of all proteins expressed by an organism at any given time point, commonly referred to as the proteome. The proteome is unlike the genome dynamic and can rapidly change depending on cell specific requirements, and is thus far more challenging to study especially in complex organisms. Only this year a draft of the complete human proteome was presented which aimed to characterize and identify the proteins in all tissue types and biological fluids (Kim *et al.*, 2014; Wilhelm *et al.*, 2014). The results gave a valuable insight into variations in biological processes in different tissue types, and can potentially be used for selection of specific biomarkers. Global proteomics studies generally result in large datasets that require elaborate data mining using various bioinformatics tools similar to other -omics fields (Kumar & Mann, 2009). Functional proteomics focuses more on protein complexes, individual proteins or even a single modified amino acid residue. Protein function is highly regulated by various modifications on individual amino acids, such as phosphorylated, ubiquitylated or glycosylated residues. Analyses of these modified sites are referred to as post-translational modifications (PTM) analyses (Mann & Jensen, 2003). These types of analyses often require enrichment techniques developed for the specific modification, combined with targeted mass spectrometry analyses. The techniques used in proteomic experiments vary widely from protein purification to gel electrophoresis, and mass spectrometry is most often used at the final stage for identification and characterization of the proteins of interest. Developments in mass spectrometry have been the main driving force in the field of proteomics over the last decade, rapidly becoming the most essential technique for large-scale protein identification and PTM analyses.

Mass spectrometry

A mass spectrometer is an analytical instrument used to determine the mass-to-charge ratio (m/z) of a charged molecule, in which m is the mass and z is the charge state of the ion. The technique is based on three basic steps, ionization of molecules in an ionization source, followed by gas-phase separation in the mass analyzer and finally detection to record the m/z value of the molecule (Figure 4). To achieve this, various types of instruments have been developed based on numerous principles for these three basic steps (de Hoffmann & Stroobant, 2013). In proteomics applications there are two methods commonly used to generate gas-phase ions; electro-spray ionization (ESI) and matrix-assisted laser desorption ionization (MALDI). The ionization event is followed by an ion separation method such as time-of-flight (TOF), quadrupole, ion trap or orbitrap mass analyzers. The ions that pass through the mass analyzer are then converted into a signal that can be read by a detector. The type of detector used depends on the design of the instrument, and can be based on conversion dynode, microchannel plate electron multipliers or image current detection. The work described in this thesis is based on electrospray ionization coupled to a linear ion trap-orbitrap tandem mass spectrometer (Hu *et al.*, 2005). The principles behind this instrument as well as other commonly applied MS techniques within biological mass spectrometry will be discussed in more detail.

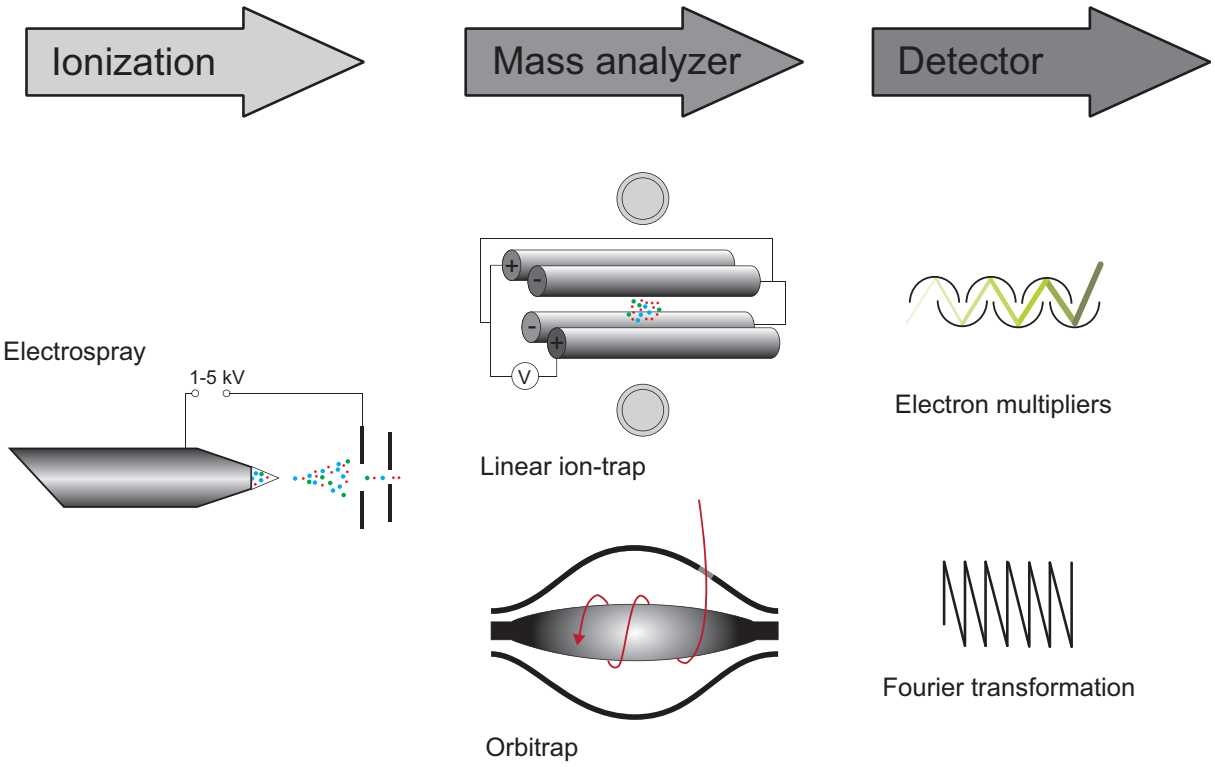

Figure 4. A mass spectrometer always contains the following elements, an ionization source, one or multiple mass analyzers for separation and a detector to "count' the ions. In the presented work electrospray ionization was used to produce ions, combined with two types of mass analyzers. The linear iontrap, in which ions are trapped in an alternating electric field and excited based on their *m/z*, as ions are excited out of the trap they will hit the detectors. The method of detection is based on electron multiplier, amplifying the signal of each ion in a cascade of secondary ions. The orbitrap is using electrostatics and DC voltage to trap ions, which will oscillate around the detector. Based on the detected image current the *m/z* can be determined using Fourier transform.

Ionization

In the ion source, the analyzed sample is ionized prior to analysis in the mass spectrometer, this involves the addition or removal of a charge. The ionized molecule can then be manipulated in an electric field and guided through the mass spectrometer and finally detected. The process of ionization occurs at the front end of the mass spectrometer as the first step of analysis. The two most commonly used ionization techniques in biological mass spectrometry are, as previously mentioned ESI (Fenn *et al.*, 1989) and MALDI (Karas & Hillenkamp, 1988) that are characterized by the stable formation of ions and absence of fragments. Introduction of these two ionization techniques has been driving the field of biological mass spectrometry. In ESI the ionization process occurs between the tip of the LC column and the inlet of the instrument. In positive mode a high potential difference is applied (1 – 3 kV for nanospray) which forces formation of a small liquid cone; referred to as a Taylor cone. The sample is vaporized into small droplets that are sprayed towards the heated inlet of the instrument, resulting in evaporation of the volatile mobile phase. Evaporation of the mobile phase reduces the droplet size and forces the molecules closer to one another until they become too close and fission occurs. This process continues until the droplets only contain a single ion that is then guided into the high vacuum region of the mass spectrometer (Wilm, 2011). An alternative theory suggests that when droplets reach a certain size charged gas-phase ions are directly formed from the droplets surface (Kebarle, 2000). ESI allows

for continuous formation of multiply charged ions by direct coupling of the analytical liquid chromatography column to the mass spectrometer (Quenzer *et al.*, 2001).

ESI of tryptic peptides is preferably performed under acidic conditions, resulting in mainly doubly protonated peptides $(M+2H)^{2+}$. The number of obtained charges is depending on the number of basic amino acids. In the case of tryptic peptides there is always one basic amino acid (K or R) at the C-terminus due to the enzyme specificity, and in addition the primary amine at the N-terminus is protonated. The addition of ≥2 charges makes it possible to select only peptides for fragmentation analyses, as most other ionized compounds will only carry a single charge. Additionally, the fixed charge on each side of the peptide is beneficial for peptide sequencing, as discussed in more detail below (Steen & Mann, 2004).

For MALDI ionization the analyte is embedded in an excess of a matrix molecules and excited using a laser. The matrix is generally consists of an acidic low molecular mass compound with strong absorption in the range of the selected laser (Mank *et al.*, 2004). The co-crystallized spot of matrix and analyte is irradiated using a laser pulse, inducing rapid heating of the crystals resulting in a small gaseous cloud of matrix and analyte. The exact mechanism of ion transfer is not fully understood, however, one theory is that the charged sublimated matrix collides with the analyte and transfers its charge resulting in predominantly single charged ions (Karas & Hillenkamp, 1988).

Mass analyzers and detection

After ionization the analyte enters the mass spectrometer, which functions under high vacuum. This is required to prevent ions from undergoing collisions with other gaseous molecules before they reach the detector. The ionized analyte will first be guided into a stable ion-current before detection, this is done by a sets of 4, 6 or 8 rods on which an oscillating potential is applied focusing the ions into its center trajectory. The instrument (ion-trap – orbitrap) used in this thesis work is composed of two mass analyzers based on two different principles for ion separation and detection. This type of instruments is referred to as "hybrid" tandem mass spectrometry allowing parallel data acquisition by combining the benefits of both a fast and a highly accurate analyzer. The two mass analysers are coupled linear to each other with octapoles for efficient ion transfer in between the two mass analysers.

The linear ion trap mass analyzer is highly sensitive, and has a fast duty cycle. Ions are accumulated in the ion trap between a set of four perfectly parallel hyperbolic rods (quadrupole) on which an oscillating electric potential is applied. A fixed potential on the back plate of the trap is forming a comb in which ions accumulate. When sufficient ions are gathered the potential on the front-plate is raised and ions are physically trapped. When trapped ions will have a stable trajectory in the oscillating electric field, in which their resonance frequency is depending on the *m/z* value. Detectors are placed on both sides of the quadrupole, and when a ramped radio frequency voltage is applied ions will increase their natural waveform and hit the detectors. This increase in waveform depends on the *m/z*, which makes it possible to separate and/or isolate ions of different mass. The detectors used for recording the mass spectra are electron multipliers, which register and amplify the impact of an ion into a cascade of secondary electrons producing a small electric current. The number of secondary electrons generated depends on the total number

of ions with a specific *m/z* hitting the detector at the same time. By ramping the radio frequency voltage it is possible to mass measure all ions in a specific mass window results in a mass spectrum of all trapped ions.

The orbitrap analyzer distinguishes itself by its high mass accuracy and resolution, although the scan rate is slower compared to ion traps. The principle of the orbitrap is based on trapping ions in an electrostatic field, where they cycle around an axial electrode in the center of a barrel shaped outer electrode (Makarov, 2000). Ions are orbiting around the electrode with a frequency proportional to their *m/z*. The frequency of the harmonically oscillating ions can be recorded using image current detection, which can then be transformed into mass spectra by Fourier transformation.

The preferred method of data acquisition in a proteomics experiment using the described instrumentation is parallel data collection, where spectra with high mass accuracy are collected in the orbitrap of all ions entering the instrument. Simultaneously in the ion trap multiply charged ions are isolated based on the information in the full precursor scan, and fragmentized to obtain sequence information. This acquisition process is data-dependent (DDA) where ions are selected based on their intensity in the orbitrap, allowing for efficient unsupervised data collection. The ion selection is based on a minimum signal threshold set to acquire high quality fragment spectra, after which they are excluded for further analyzes so spectra can be obtained on all other ions entering the analyzer. The typical duty cycle for mass analysis and simultaneous fragmentation is around one second, although this may vary based on the required resolution and signal intensity.

Spectral data can as well be collected independent of the acquired data in a method that is referred to as data independent analysis (DIA, *i.e*, AIF, MS^E and SWATH.) (Venable *et al.*, 2004). The method is based on the sequential fragmentation of a fixed precursor windows (10 – 100 m/z) covering the complete mass range, in principle allowing identification of all peptides entering the instrument. The main difference with DDA data interpretation is the loss of the relationship between peptide precursor and fragment mass, therefor multiplexed fragmentation spectra are searched against spectral libraries or using modified database search engines (Geiger *et al.*, 2010a; Gillet *et al.*, 2012).

Single reaction monitoring (SRM) is targeted data acquisition method relying on prior knowledge of the proteins in a sample. The mass spectrometer is set to record only the product ions from the fragmentation of a single peptide over a defined retention time window (Picotti *et al.*, 2009). These analyses are mainly performed on triple quadrupole instruments (QqQ) were the first quadrupole is set to isolate the precursor, followed by collision in the second and detection of the fragment ions in the last quadrupole. SRM is the standard method for targeted quantification as it allows for consistent recordings of the intensities of predefined target fragment ions across the analysis. However, SRM is limited to measurements of a few thousands transmissions per analysis, limiting the number of proteins quantified per analysis (Costenoble *et al.*, 2011).

Peptide sequencing by mass spectrometry

Isolated ions can be fragmentized into smaller fragments to obtain more detailed information on their structure or peptide sequence. There are various techniques available to fragmentize ions applied in biological mass spectrometry including collision-induced dissociation (CID), electron

transfer dissociation (ETD) (Syka *et al.*, 2004), high energy collision dissociation (HCD) (Olsen *et al.*, 2007) and electron capture dissociation (ECD) (Zubarev *et al.*, 1998). The principle of peptide fragmentation is based on controlled cleavage of peptide bonds or the lateral amino acid side chain. The site of cleavage depends on the fragmentation technique that is used and will result in one or two ion series (a, b, c from the N-terminal side or x, y, z derived from the C-terminal), according to the nomenclature by Roepstorff-Fohlmann-Biemann (Figure 5) (Roepstorff & Fohlman, 1984; Johnson *et al.*, 1988).

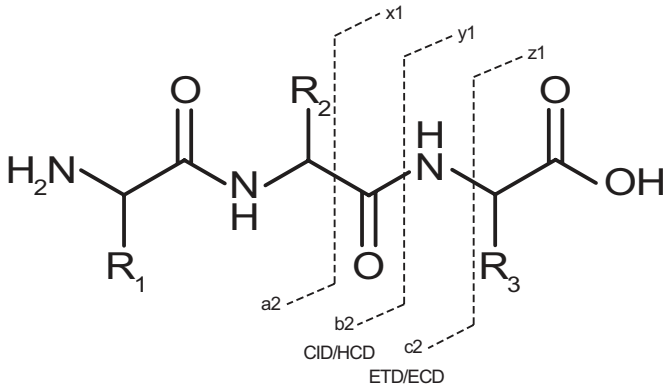

Figure 5. Fragment nomenclature of N- and C-terminal derived ions after protein backbone fragmentation. The observed fragment ions based on CID/HCD and ETD/ECD fragmentation are annotated in the figure

When performing CID fragmentation in the ion trap the isolated peptide is trapped and accelerated to reach a higher kinetic energy, followed by collision with an inert gas. During collision the kinetic energy is transformed into internal energy resulting in cleavage of the peptide bonds. CID fragmentation is performed under controlled conditions and generates random sized peptide fragments. The resulting b- and y-ion series can be used to resolve the peptide sequence, based on the mass differences between the ions in both series representing the sequential loss of the amino acids from the N- or C-terminal end (Eng *et al.*, 1994). The choice of fragmentation technique depends on both the type of instrument that is available and the experimental question. CID and HCD are the preferred methods when obtaining spectra for peptide sequencing, whereas ETD and ECD are often used for analyzes of post-translational modifications (PTM's) to determine the modified sites and for longer peptides (Wiesner *et al.*, 2008). During CID and HCD fragmentation the labile PTMs dissociate from their attachment site due to the lower energy barrier compared to the peptide backbone, which prevents accurate site localization. In ETD fragmentation, singly charged radical anions are collided with the cationic peptide (Syka *et al.*, 2004; Mikesh *et al.*, 2006) inducing general peptide backbone cleavage while the modification is retained on its amino acid residue. ECD is based on a similar principle introducing low energy electrons into the collision cell to induce fragmentation (Zubarev *et al.*, 1998). Both fragmentation techniques are therefore frequently used for site localization of modifications such as phosphorylation and glycosylation (Chi *et al.*, 2007; Steentoft *et al.*, 2011).

Peptide identification by mass spectrometry

The general strategy for identification of proteins is enzymatic digestion into peptide fragments, which are then subjected to mass spectrometry analysis. This strategy is referred to as bottom-up

proteomics, which is in contrast to top-down proteomics were the complete protein is analyzed and identified by fragmentation induced in the instrument. Complete proteins can be analyzed by mass spectrometry although cumbersome, due to limited solubility, lower sensitivity in the higher mass range, and unpredictable masses due to PTMs. However, top-down proteomics often leads to overall higher sequence coverage allowing identification of isoforms and more accurate protein quantification compared to the analyses on peptides level (Waanders *et al.*, 2007; Tran *et al.*, 2011). The recent introduction of a new high-sensitive orbitrap allows the analyses of mega Dalton structures previously limited to TOF instruments, and the improvement of separation techniques allowing the more routine analysis of multiple proteins per analysis (Ahlf *et al.*, 2012; Rose *et al.*, 2012).

The standard approach for protein identification is still based on one-dimensional gel electrophoresis for separation of a protein mixture after which stained protein bands of interest are excised, washed and digested (Shevchenko *et al.*, 2006). The enzymes used for digestion are selected based on their high specificity and activity, such as trypsin and Lys-C. Less specific enzymes should be avoided, as they will generate small overlapping sequences that complicate the analysis. Extracted peptides can be directly analyzed or separated into multiple fractions by liquid chromatography to increase the number of peptide identifications. Peptide separation by HPLC can be performed either directly coupled to the mass spectrometer as in the case of ESI (Martin *et al.*, 2000; Shen *et al.*, 2001), or offline when using MALDI as ionization source (Marcus *et al.*, 2007). Proteomics often entails identification of a large number of peptides in a complex mixture, and the duty cycle of the mass spectrometer is limiting the identification when all peptides are introduced directly (Thakur *et al.*, 2011). Therefore the peptides are normally separated by chromatography prior to introduction to the mass spectrometer by chromatography. The preferred chromatographic method is reverse phase (RP) chromatography using C18 material; separating peptides based on their hydrophobicity, which can be directly coupled to the ionization source. Complete peak separation is not required since the mass spectrometer can record multiple ions at once. On average a peptide elutes in a 10 - 60 seconds time window depending on the slope of the gradient giving the instrument enough time to collect a fragmentation spectra for each individual peptide. The signal intensity of an ion is directly proportional to the volume in which the peptide elutes. Downscaling the columns to the nanoscale range (inner diameter 75 μm or less), and decreasing the flow rate has greatly improved the sensitivity and the sample quantities required for analysis (Liu *et al.*, 2007). Continuous developments of the chromatography interface has made it possible to identify thousands of proteins in a single run (Thakur *et al.*, 2011). A more global strategy to study the proteome of a cell population requires a different approach then when only a subset is analyzed, such as a protein band. For a more global approach, peptides are often separated in multiple fractions prior analyzes by LC-MS/MS. This approach is referred to as 2D LC-MS/MS (Washburn *et al.*, 2001) and combines two orthogonal chromatographic phases dramatically increasing the depth of the proteomic analysis (*i.e* normal-phase – reverse phase or ion-exchange – reverse phase) and can be preformed offline or online. The extensive fractionation is required to overcome the large dynamic range in protein abundance in a cell (Picotti *et al.*, 2009).

Protein identification

High throughput identification of proteins from mass spectrometry data is a fully automated process. Spectral data for every fragmentized peptide is reduced into a list of all detected ion masses and combined with the mass of the peptides submitted to a search engine. The principle of identification is based on the comparison of *in silico* digested protein sequences with peptide masses observed by the mass spectrometer, followed by comparison of the fragmentation spectra (Figure 6) (Taylor & Johnson, 1997; Perkins *et al.*, 1999). Therefore the use enzymes with high specificity is of great importance for the success of the identification process (Olsen *et al.*, 2004), and the identification requires a protein sequence database available for the analyzed species. Most modern mass spectrometers measure the mass of an ion with very high mass accuracy allowing only a few parts per million error tolerance between the observed and the theoretical peptide mass (Nesvizhskii & Aebersold, 2004). The high mass accuracy reduces the potential peptide candidates to a relative low number. The theoretical fragmentation spectra of the candidate peptides are then compared to the observed fragmentation spectra. Calculation of theoretical spectra is based on the type of fragmentation applied during mass spectrometry analysis. The quality of the peptide match is presented by a score calculated based on the number of fragment ions matched. The score threshold for positive peptide identifications is probability based typically depending on the number of entries in a protein database. The more entries in a protein database, the higher the required score is due to the increased chance for random false positive identifications when large numbers of spectra a searched (Keller *et al.*, 2002; MacCoss *et al.*, 2002). In the final step, peptide hits are assigned to one or multiple proteins if homology occurs. In some cases protein identifications will only be based on a single peptide, this is typically for large-scale studies were a significant number of the identified proteins would be based on only a single peptide. In a small-scale study these hits are often not considered as true positives. Stringent filtering by removing all single peptide identifications will together with the false positive identification remove important biological information (Peng *et al.*, 2003). To circumvent this problem methods are developed to give a better estimation of the required score assigned to a spectra by determining the false discovery rate (Choi & Nesvizhskii, 2008; Käll *et al.*, 2008). Spectra can be searched towards a decoy database, composed of identical protein sequences in reversed order combined with the actual protein database. The results will be a combination of reversed and forward hits, and the frequency and peptide score of both is used to set the threshold. The peptide score threshold for a positive identification can be set to the intercept of the two frequency plots where <1% is a false positive. This method will allow for inclusion of normally omitted single peptide identifications, although one still have to take into consideration that the analyzed data set will include false positive protein identifications.

Quantitative mass spectrometry based proteomics

Mass spectrometry provides a perfect platform for large-scale quantification and comparisons of proteins under different sample conditions. However, data obtained by mass spectrometry analysis of peptides is not directly quantitative, as the signal recorded is no measure of its abundances. The intensity depends on the ionization efficiency, which in turn depends on the chemical properties of the combined amino acids (Eyers *et al.*, 2011). Therefore the signal

observed for the various peptides derived from the same protein will be highly variable. To overcome this issue various methods have been developed over the years to allow for both absolute and relative protein quantification. The applied methods can be separated in two groups based on either stable isotopic labelling or without the addition of labels ("label free"). The use of isotopic labelling is considered more accurate compared to label free methods, however, label free methods can be applied to virtually any sample whereas the introduction of isotopic labels has limitations.

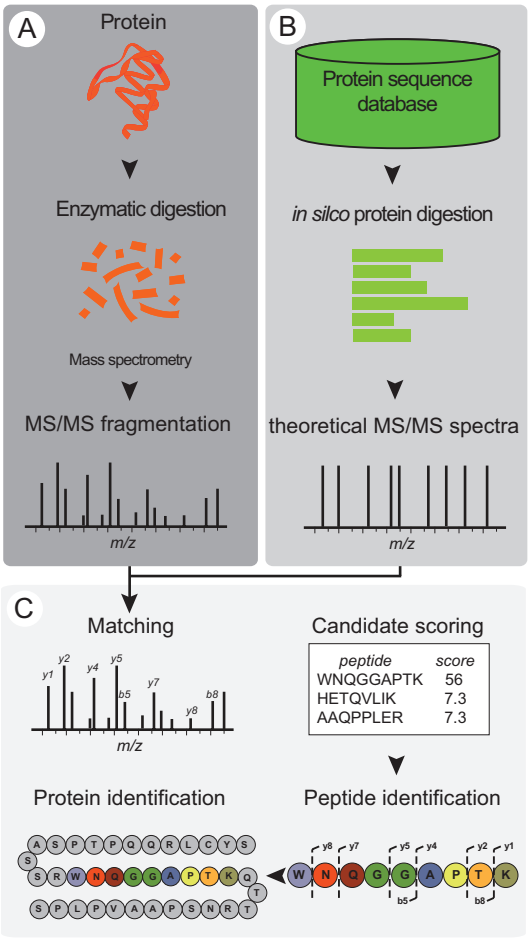

Figure 6. Overview of the protein identification process. (A) Proteins are digested into peptides, which are then subjected to mass spectrometry analysis collecting MS/MS spectra of the peptides. (B) A protein database containing large numbers of protein sequences is digested theoretically with the same enzyme used in the experiment. Followed by the calculation of the theoretical fragment ions. (C) The fragment ions of peptides that fit the mass of the observed peptide are then compared to the observed fragmentation spectra. The potential hits are then ranked by score and the peptide is assigned to the protein of origin.

Stable isotope labelling

Stable isotopes can be introduced at various stages during a proteomics experiments both by *in vivo* and *in vitro* methods incorporated enzymatically, chemically or metabolically. Isotopic reagents are typically used to generate a heavy and light state without changing the chemical properties of the peptides, allowing simulations analysis and differentiation of the isotopic pairs. One of the first methods that was used to introduce stable isotopes was enzymatic labelling by performing protein digestion in heavy water ($H_2{}^{18}O$) resulting in incorporation of ^{18}O at the C-terminal generating a 4 Dalton mass shift as compared to digestion in non deuterated water. The

method allowed the relative quantification between conditions although limited to the use of MALDI. The major drawback with deuterated peptides is the altered chromatographic behavior and the overlapping isotopic envelop at higher charge states.

Chemical labelling can be subdivided into two groups; isotopic and isobaric. The first category includes methods such as isotope-coded affinity tags (ICAT) (Gygi *et al.*, 1999), isotope-coded protein labels (ICPL) (Schmidt *et al.*, 2005) and dimethyl labelling (Hsu *et al.*, 2005) were quantification is performed at MS level for up to three channels based on the extracted ion chromatograms. Dimethyl labelling has emerged as a cost effective, easy and flexible method recently extended to include 5 different labelling channels (Wu *et al.*, 2014). Isobaric labelling involves chemical derivatization of the peptides primary amine groups adding a tag with an identical mass to each peptide. Samples for comparisons are differentially labelled after tryptic digestion and mixed together. The generated fragmentation spectra will reveal distinct reporter ions in the low mass region for the different tags that then are used for peptide and protein quantification. The most commonly used isobaric labelling methods are tandem mass tags (TMT) (Thompson *et al.*, 2003) and isobaric tags for relative and absolute quantitation (iTRAQ) (Ross *et al.*, 2004) allowing the multiplexing of up to 8 samples.

Metabolic incorporation of stable isotope ($^{15}N/^{13}C$) labeled nutrients in growth media of cultured cells is a method for global coding of proteomes. Incorporation of ^{15}N has the drawback that the introduced mass shift depends on the number of nitrogen's in the peptides sequence and only after peptide identification the isotopic pairs can be calculated. Therefore stable isotope labeling by amino acids in cell culture (SILAC) is a more frequently used method introducing one or more isotopic variants of essential amino acids, mainly lysine and arginine since at least one of these is included in every tryptic peptide (Ong *et al.*, 2002). Culture media is depleted of arginine and or lysine and supplemented with isotopic variants and after multiple cell cycles full incorporation is achieved and samples can be directly mixed with normal unlabeled cells for comparison. The method was originally developed for the comparison of cell lines, however when a cell line is labelled that resembles the cell type of a tissue of interest is can also be used as a standard for the relative quantification of for example clinical samples (Geiger *et al.*, 2010b). In addition, SILAC can also be used to label model organisms such as *Mus Musculus*, *Drosophila Melanogaster* and *Caenorhabditis Elegans* by feeding the animals a SILAC diet (Krüger *et al.*, 2008; Sury *et al.*, 2010). Although SILAC is the regarded as the most accurate method for quantitative comparisons of two or more samples the method does not give information on the absolute protein concentrations. To obtain absolute measurements of protein levels synthesized standard isotopic peptides in a known concentration are added to the samples prior analysis and the concentration of the peptides in the samples is calculated by comparison to the synthetic peptides (Gerber *et al.*, 2003). All the described isotopic labelling methods have their respective advantages and drawbacks and the selected method depends on costs, available instrumentation and applicability. One general rule is that the earlier in an experiment that samples are combined the less variation is observed due to sample handling favoring the use of metabolic labelling (Ong & Mann, 2005).

Label free quantification

Label free protein quantification is based on comparison of the combined intensities of the identified peptides or the frequency of which each peptide is identified (Neilson *et al.*, 2011). The major difference to isotopic labelling is that the compared samples are analyzed by separate LC-MS/MS runs. Initially the identification score from the database search was used as a measure of protein abundance. However, the summed peptide score was shown to be more depending on the peptide sequences than the actual protein abundance (Sadygov *et al.*, 2004). The number of peptides observed for each protein is an easy and more accurate method for comparing protein abundance between samples. The principle is based on the assumption that more abundant proteins result in peptides that will be more frequently selected for fragmentation in a DDA experiment. Over the years the method has been improved applying different normalization strategies correcting for variation in protein length and number of theoretically observable peptides (Lundgren *et al.*, 2010).

The intensity of the analyzed peptides can also be used as measure of abundance using either the intensity or the area underneath an eluting peak (Zoetendal *et al.*, 2002; Liu *et al.*, 2004; Eckburg *et al.*, 2005). Label-free quantification has been broadly applied although as samples are analyzed in parallel it is sensitive to variability in sample preparation, LC reproducibility, and ionization efficiency all affecting the accuracy of the quantification. Therefore parallel sample preparation, chromatographic peak alignment, data normalization and statistical analysis are required to avoid inaccuracies in quantification (Derrien *et al.*, 2004; Khan *et al.*, 2009).

The overall strength of label-free quantification is the potential to apply it to any type of sample, although considerations are needed, as it is prone to the introduction of errors. Various studies have shown that most relative quantification methods give similar results (Asara *et al.*, 2008; Bevins & Salzman, 2011). Label-free methods have as benefit of deeper proteome coverage and the ability to overcome a wider dynamic range.

Mass spectrometry based glycoproteomics

The majority of secreted and extracellular proteins are modified by *O*-glycans and/or *N*-glycans, which are required for proper protein folding, protein-protein interactions, and increase the overall stability of a protein. The study of both the attachment sites on the protein and the attached glycan(s) is referred to as glycoproteomics. The large heterogeneity in glycan moieties, and the size and complexity of the added glycans have hampered the comprehensive analysis and localization of these modifications (Haslam *et al.*, 2006). For mucin type *O*-glycosylation (GalNAc), no consensus motif exists making the prediction and identifications of *O*-glycoproteins difficult. *N*-glycosylation can be predicted based on the N-X-S/T motif (in which X can be any amino acid except proline) which in general has a high site occupancy (Apweiler *et al.*, 1999). Different approaches can be considered when studying site-specific protein glycosylation by mass spectrometry. One method is by releasing the glycans from the protein backbone, which can be achieved by the use of specific glycosidases that are able to hydrolyze glycosidic bonds, or by chemical release (Jensen *et al.*, 2012).

The release of *N*-glycans requires only a single glycosidase PNGase F/A cleaving between the innermost GlcNAc and asparagine. When the resulting peptides are analyzed the previously

occupied asparagines are identified by a mass increment of ΔM = 0.98 due to deamidation after the release of the *N*-glycan. This methodology allows for the identification of occupied sites in efficient and a systematic way (Zielinska *et al.*, 2010). By the removal of the glycan the actual composition remains elusive, loosing important biological information. The analysis of *N*-glycosylated peptides requires enrichment using lectins, size exclusion chromatography or hydrophilic interaction-liquid chromatography (HILIC) (Wuhrer *et al.*, 2007). These methods are limited to the analysis of peptide digests of simple protein mixtures or purified proteins, due to the complexity of the resulting samples and the lack of software tools for the automated the interpretation of the resulting spectra.

Protein *O*-glycosylation is potentially even more complex and challenging to study as these are often found in mucin-domains lacking enzymatic cleavage sites, and glycosidases able to release the full oligosaccharide are absent. Exoglycosidases do exists that are able to hydrolyze most glycosidic bonds to trim down the structures stepwise, although often completely released by reductive beta elimination hydrolyzing the protein backbone (Fukuda, 2001). The introduction of the genetically engineered "SimpleCells" system by Clausen and co-workers allowed for the first time in-depth analysis of the *O*-glycoproteome based on *O*-glycopeptides (Steentoft *et al.*, 2011). In these cells the Cosmc gene is knocked-out that functions as a dedicated chaperone for the glycosyltransferase required for extension, terminating the glycosylaton after the initial addition of GalNAc to the protein backbone. These short *O*-glycosylated peptides can be enriched using lectin columns and identified by standards peptide identification algorithm.

AIM OF THESIS

The overall aim of this thesis work was to adapt mass spectrometry methods for studies of the protective mucus layer in the colon and to utilize these methods to obtain a better understanding of its composition, modifications and function.

Specific aims

I. To characterize the protein composition of the human colonic mucus, and to determine whether the mucus protein composition is altered in patients with ulcerative colitis.

II. To define the proteolytic effect of the secreted proteases RgpB from the bacterium *Porphyromonas gingivalis* on the MUC2 mucin.

III. To combine various proteomics methods to obtain full sequence coverage of a protein enabling complete characterization of its post-translational modifications.

IV. To study the role of the CysD domain in the MUC2 mucin on the formation of oligomeric structures.

V. To study the segmental variation in membrane protein expression along the axis of the normal human colon.

METHODS

Sample preparation prior to mass spectrometry (I, II, III, IV and V)

Before a protein sample can be analyzed by mass spectrometry, the sample has to be subjected to various preparation steps that make the sample suitable for analysis. The methods of choice depend on the nature of the sample and the biological/research question addressed. The work presented in this thesis is based on analysis of proteins digested by proteases, after which the peptide fragments are analyzed by mass spectrometry, identified and matched to proteins for identification. This method is commonly referred to as "bottom-up" proteomics, which is in contrast to "top-down" were the full protein is analyzed and subjected to fragmentation in the mass analyzer and the identification is based on the resulting protein fragments.

Factors to consider when choosing sample preparation methods include the expected protein complexity, if all proteins in a sample have to be identified, or if the analysis only requires identification or characterization of one particular protein. Independently of the research question each approach requires the protein sample to be solubilized, and stabilizing bonds have to be reduced to allow the protease used for digestion to access the protein backbone. In the presented work, two different approaches were used, in-gel digestion and in-solution digestion on molecular mass cut-off filters.

In-gel digestion (II, III and IV)

In this thesis all single protein identifications were performed using in-gel digestion. Proteins were separated by SDS-PAGE electrophoresis after which the proteins were stained, and the bands of interest were excised and treated with proteases (Shevchenko *et al.*, 1996). The main advantage of this is method is the large reduction in sample complexity introduced by the molecular mass separation on the gel. In addition, interfering buffer components and detergents are removed from the sample while the proteins are retained in the gel-matrix. On the contrary, in-gel digestion tends to result in lower peptide recovery, and higher variation in quantification of peptides derived from the same protein as compared to in-solution digestion (Havlis & Shevchenko, 2004). This is likely due to steric hindrance, and the inability of the protease to diffuse into the gel matrix. As the method was solely used for identification, and in most cases sufficient amount of recombinant protein was available, these factors were not expected to alter the outcome of the experiments. The work in paper III and IV required the extraction of glycosylated and cross-linked peptides after digestion; these peptides are expected to be large and more difficult to extract. High concentrations of organic solvents dehydrate the gel limiting the pore size and reducing the extraction efficiency. The first phase of peptide extraction was therefore performed in a buffer not containing any organic solvents (Kolarich *et al.*, 2012).

In-solution digestion (I, II and V)

Protein digestion in-solution is considered the method of choice for identification and quantification of large numbers of proteins in complex mixtures. However, the method is

hampered by reduced trypsin efficiency in the buffers and detergents used during sample preparation and present in the initial sample. To overcome this problem, protein digestion can be performed on molecular mass cut-off membranes in a spin-filter format. The method was initially described by Manza *et al*. and later adapted and coined filter-aided sample preparation (FASP) by Wiśniewski *et al.* (Manza *et al.*, 2005; Wiśniewski *et al.*, 2009). The overall benefit of this sample preparation method is the ability to perform efficient buffer exchange via centrifugation while the proteins remain on the filter, allowing the use of high concentrations of SDS for initial protein solubilization, which is then completely removed by subsequent washes with chaotropic agents like urea. When trypsin is introduced to the proteins in the spin-columns the resulting peptides whose size is below the molecular mass cut-off of the filter can be collected by centrifugation. In paper II and V the FASP method was applied with serial digestions, first by the endoproteinase Lys-C followed by trypsin. The purpose of using two proteases is that Lys-C is active in high concentrations of urea, when the proteins are maximally denatured. Trypsin is then added after dilution with the purpose of increasing the numbers of peptides used for identification and improving quantification (Wiśniewski & Mann, 2012). This is especially imported for the analysis of membrane proteins, which are in general more difficult to solubilize.

In paper I, preparation of the mucus samples was performed using guanidine hydrochloride (GuHCl) as a chaotropic agent. GuHCl was preferred over urea due to its higher efficiency in denaturing the highly charged mucins (Carlstedt *et al.*, 1983). Both solubilization and reduction of the disulfide bonds were performed in GuHCl, before the sample was transferred to the filters units. Another advantage of GuHCl is that disulfide bond reduction can be performed at elevated temperatures without the risk of introducing unwanted protein modifications as occurs with urea. Similar protein identification rates are observed when comparing urea to GuHCl when used in combination with Lys-C (Poulsen *et al.*, 2013). GuHCl is seldom used in proteomics studies as trypsin only tolerates minimal concentrations (Proc *et al.*, 2010). However, since the FASP method allows for effective buffer exchange, GuHCl can be used during the sample preparation as long as it is exchanged prior to addition of trypsin. Both methods were evaluated and the results showed more efficient solubilization of the mucus in GuHCl and an increase in peptide identifications of mucus associated proteins as compared to SDS/urea (data not presented).

Enrichment of membrane proteins (II and V)

Analysis of membrane spanning proteins was used to identify the GalNAc transferases expressed in the colonic epithelium in paper II, and to characterize and profile all membrane spanning proteins along the human colon in paper V. In standard proteomics workflows, membrane proteins are often underrepresented because of their amphiphilic properties. Various methods for enriching membrane proteins have been developed over the years, varying from extraction of the total membrane pool to specific isolation of the plasma membrane (Vuckovic *et al.*, 2013). The traditional methodology relies on mechanical or osmotic disruption of the cells followed by density gradient centrifugation. This type of approach requires relatively large amounts of starting material and is therefore not ideal for the small tissue biopsies that were used in both paper II and V. Instead we used an established method for enrichment of membrane proteins based on sodium carbonate washes at high pH in combination with ultracentrifugation (Fujiki *et*

al., 1982; Wu *et al.*, 2003). This method is less specific compared to other described methods but can be used to efficiently enrich both the plasma membrane and intracellular membranes from small sample quantities. Using this method around 50% of the proteins identified in paper I and II contained membrane spanning regions, including the GalNAc-transferases found in the Golgi membrane.

Chromatography (I, II, III, IV and V)

Chromatography is an essential tool in proteomics, used to reduce the protein and peptide complexity in a sample prior to mass spectrometry analysis (Neverova & Van Eyk, 2005). The limiting factor in a typical experiment is the analysis speed and sensitivity of the mass spectrometer. Separation of a tryptically digested protein sample into multiple fractions will increase the depth of the analysis and improve sequence coverage, resulting in an increased number of identified proteins and more accurate protein quantification. The separation of proteins by size using SDS-PAGE is the simplest method used to reduce sample complexity. However, a wide range of fractionation strategies for both proteins and peptides are available. These methods often combine multiple separation techniques, as single dimension separation often is inadequate. Various types of chromatography can be used prior to protein digestion such as size exclusion, ion exchange, and reverse phase. At the peptide level, reverse phase, ion exchange and isoelectric focusing are the most common separation strategies. The benefit of performing chromatography at peptide level is the smaller variety in physical and chemical properties between peptides compared to proteins, and the higher resolving power. In the case of analysis of mucus, reduction of sample complexity was performed at the peptide level due to the biochemical properties of the mucus. The MUC2 mucin forms large oligomers that upon secretion are a few million Dalton in mass, and even after reduction of the disulfide bonds the protein size far exceeds that of all other proteins in the sample. In addition, the heterogeneity in glycan modifications will vary the biophysical properties of each molecule, which hinders selective fractionation. When digested into peptides the majority of the MUC2 peptides will have similar properties as peptides derived from proteins found in the mucus gel, allowing for simultaneous analysis of the different mucus components.

Peptide fractionation by offline chromatography (II and V)

The separation of a peptide mixture in one dimension is often not sufficient to adequately identify all of its components by mass spectrometry. Therefore, an additional separation step can be added off-line to separate the peptide mixture into multiple fractionations, or even in an additional dimension. In principle, any type of chromatography can be used to separate the peptides, while the final online separation is typically reverse phase chromatography (Wang & Hanash, 2003). Various methods are commonly applied such as: strong-anion exchange, strong-cation exchange, high pH reverse phase and off-gel isoelectric focusing (Lau *et al.*, 2011; Branca *et al.*, 2013). When selecting a separation method one needs to consider that the final stage is based on peptide hydrophobicity under acidic conditions and the first dimension should be orthogonal to this method. Under ideal conditions this will result in fractions of equal complexity that elute

in a broad window from the LC-RP column. In this thesis work, off-line peptide separation was performed using HILIC, which allowed for a more in-depth analysis of the complex samples in paper II and V, and improved separation of the different glycopeptide isoforms in paper II. HILIC was introduced in the 1990s and is based on a hydrophilic stationary phase, and a hydrophobic mobile phase, and is mainly used for analysis of small molecules (Alpert, 1990). Peptides elute in a reversed order as compared to standard reversed phase with the hydrophilic peptides eluting at the end of the gradient. Although HILIC is mainly used to separate small molecules, recent studies have shown that it also performs well in a variety of proteomics applications (Di Palma *et al.*, 2011; Engholm-Keller *et al.*, 2011). The sole use of water and organic solvents as the mobile phase allows for direct analysis of the collected fractions after lyophilizing and reconstitution without further sample cleanup. In glycoproteomics HILIC is used to separate glycans and glycosylated peptides, and to differentiate between different isomers (Wuhrer *et al.*, 2009). Glycopeptides have a longer retention time compared to non-glycosylated peptides due to the hydrophilic character of the glycan, which makes it possible to separate glycopeptides from non-modified peptides (Hägglund *et al.*, 2004). In paper II a standard *O*-glycosylated peptide could efficiently be separated based on the number of GalNAc modified residues, supporting the use of HILIC in improving the resolution of glycopeptide analyses.

Peptide separation by online chromatography (I, II, III, IV and V)

Chromatography directly coupled to a mass spectrometer is referred to as LC-MS. The chromatography step concentrates the peptides on the column prior to elution into the mass spectrometer, which increases the sensitivity. The method allows for direct analysis of the eluting peptides, but is limited to ESI ionization. The number of peptides that can be analyzed by LC-MS depends on the column length, the column inner diameter, the packing material, the temperature and the length of the gradient (Shen *et al.*, 2001; Luo *et al.*, 2005; Xu *et al.*, 2009). One requirement for direct analysis of the eluting peptides is that the solvents used during chromatography are compatible with the mass spectrometry analysis and do not interfere with the ionization process. This excludes the use of most salts, and limits the chromatography mainly to reverse phase for the analysis of peptides. In the presented work we used reverse phase chromatography on C18 material that retains the peptides in aqueous solutions, that will then elute over an organic solvent gradient depending on hydrophobicity. The columns were packed in-house with 3μm C18 material in fused silica capillaries with an inner diameter of 75 – 100 μm. Each chromatographic setup was composed of two separate columns, one so called pre-column, a short column used to trap peptides during sample loading (5 cm long, 100 μm inner diameter) and an analytical column for the actual chromatographic separation (15 – 20 cm long, 75 μm inner diameter). The pre-column was connected between two micro-tees connected to a divert valve allowing sample loading at high flow rates, while using a split to reduce the flow during the actual elution of the peptides over the chromatographic column (Meiring *et al.*, 2002). This configuration was chosen to be able to efficiently load peptides onto the setup, while the split allowed the use of reduced flow rates in the nl/min range without using a nano-LC system. The length of the gradient was adapted to the expected complexity of the analyzed samples and ranged between 30 and 90 minutes.

Mass spectrometry (I, II, III, IV and V)

The majority of the presented work in this thesis is based on mass spectrometry analysis, an essential tool in the field of proteomics. We used a linear ion trap-orbitrap (LTQ-Orbitrap) for all MS analysis, composed of an ESI interface coupled to an ion trap mass analyzer allowing the transfer of ions via a C-trap into an orbitrap mass analyzer. The instrument is equipped with three different types of fragmentation techniques, ETD, CID and HCD. The first two are performed in the ion trap and for HCD a dedicated cell is coupled to the back of the C-trap. Both ion trap and orbitrap can be used for fragment ion detection, and the preferred fragmentation and detection method is selected depending on the preferred experimental outcome. In the following sections the use of various mass spectrometry strategies and their applications will be discussed.

Characterization of O- and N-glycopeptide modifications (II and III)

Mass spectrometry is the primary tool used for the study of site-specific protein glycosylation. Characterization of glycoproteins is mainly performed at the peptide level after enzymatic digestion, identifying both the amino acid sequence and the glycan structure attached. The information obtained from glycopeptide spectra is dependent on the fragmentation method applied. When applying CID fragmentation, the main fragments observed will be from the cleavage of the glycosidic bonds characterized by abundant Y-type ions and diagnostic B-type oxonium ions (Huddleston *et al.*, 1993). Fragments from the peptide backbone are normally not observed. On the contrary ETD or ECD results in selective fragmentation of the peptide backbone (Wuhrer *et al.*, 2007). The combination of the two complementary fragmentation techniques results in spectra from which both peptide and glycan composition can be elucidated. Identification requires the manual validation of the spectra as no algorithms are available to efficiently analyze the spectra. Low mass oxonium ions typical for glycan backbone fragmentation can be used as indicator that a spectra is derived from a glycopeptide.

The characterization of *N*- and *O*-glycosylation in this thesis was performed on the MUC2 C-terminal in paper III, and on a synthetic peptide in paper II. Mass spectrometry analyses were performed on glycopeptides using HCD fragmentation with fragment detection in the orbitrap. The high mass accuracy of the orbitrap makes the selection of potential glycopeptide derived spectra based on the oxonium ions more straightforward. The peptide sequences were identified based on manual interpretation of ETD spectra.

Identification of proteolytic cleavage sites (paper II)

Proteolytic processing of a protein is considered a post-translational modification that irreversibly modifies its function. Studies of the specificity of this type of modification is challenging, as our genome encodes for several hundred different proteases which limits the majority of the research in this field to *in vitro* techniques (Puente *et al.*, 2003). More recently, proteomics approaches have emerged that specifically focus on capturing these events in cells and tissues by targeting the neo N-termini formed after proteolytic hydrolysis of the peptide bonds (Doucet & Overall, 2008). Identification of the neo N-termini as a result of proteolytic cleavage of

a protein is possible by specific labeling that distinguish these from other peptides during MS analysis. Labeling of primary amine groups (N-termini and lysines) can be performed prior to protein digestion using various amine reactive reagents; after tryptic digestion newly formed internal peptide N-terminal amine groups can be differentially labeled or negatively enriched before mass spectrometry analysis (Kleifeld *et al.*, 2010). For mass spectrometry analysis it is important to select a labeling reagent that adds a fixed charge to the N-termini to enable the detection of sequence ions upon fragmentation. In paper II we identified the neo N-termini after proteolytic cleavage of MUC2 by labeling with trimethoxyphenyl phosphonium (TMPP) a reagent originally introduced to enhance sequencing by MALDI based mass spectrometry (Huang *et al.*, 1997). The labeling reaction was performed at a controlled pH at which only the α-amines are labeled due to their lower pKa value as compared to the lysine side chains (Huang *et al.*, 1999). The protein was digested after the labeling reaction and the peptides were analyzed by LC-MS/MS. The derivatized peptides were identified based on their modified N-termini, in combination with their increased retention time. Addition of a TMPP label increases the peptide hydrophobicity, resulting in elution outside the window where the majority of the peptides elute (Gallien *et al.*, 2009).

Identification of disulfide linked peptides (IV)

Disulfide bonds formed between cysteines are important for protein folding and stabilization of secondary and tertiary structures. Therefore, identification and characterization of their location can help to understand how a protein is folded and is structured. Most of the current knowledge on protein disulfide linkages comes from detailed protein structures obtained by NMR or X-ray crystallography, which are time-consuming techniques that require large quantities of protein. Disulfide linked peptides can also be analyzed by mass spectrometry, however this approach is faced with various technical challenges (Gorman *et al.*, 2002). Firstly, the majority of algorithms developed for the identification of peptides by mass spectrometry are not designed to identify two peptides linked together and present in a single fragmentation spectrum. For this reason in a general proteomics experiment disulfide bonds are reduced, thus information regarding disulfide bonds is lost. Secondly, disulfide bonds do not dissociate during CID or HCD fragmentation, resulting in limited sequence information. Thirdly, sample handling can induce rearrangement of disulfide bonds when free cysteine is available and/or when protein digestion is performed at neutral or alkaline pH (Ryle & Sanger, 1955). However, recent advances have been made assisting in the analysis of linked peptides with the introduction of algorithms dedicated to deal with identification of cross-linked peptides in simple protein mixtures (Choi *et al.*, 2010; Yang *et al.*, 2012). Furthermore, the availability of ETD on orbitrap mass spectrometers that enables dissociation of disulfide bonds, in combination with CID/HCD allows for systematic analysis of disulfide bridges (Liu *et al.*, 2014). Disulfide scrambling can be limited by alkylation of free sulfhydryl groups, and protein digestion under slightly acidic conditions.

Identification of which specific cysteine residues are linked via disulfide bonds was of great interest in this thesis work for the characterization of the cysteine rich CysD domain in MUC2 responsible for oligomerization of the protein. We characterized the intramolecular disulfide bonds of the CysD domain using Asp-N protein digestion followed by CID fragmentation. This

approach allowed us to identify the potentially disulfide-linked peptides by database searches using Mascot against a specially designed database containing all possible inter- and intramolecular linked peptide combinations for CysD as single entries after Asp-N digestion (Singh *et al.*, 2010). Candidate spectra were manually curated, and all linked cysteines were identified. No disulfide rearrangement was observed although protein digestion was performed at basic pH.

Label free peptide quantification (I and V)

Mass spectrometry based proteomics is not directly quantitative as the response from the mass analyzer depends on the ionization efficiency of the peptide, which in turn depends on the chemical properties of the combined amino acids in the peptide sequence (*i.e.* length, charge and hydrophobicity) (Eyers *et al.*, 2011). Peptides derived from the same protein will therefore give different signal intensities when detected in the mass spectrometer. Comparison of the combined signal for all identified peptides assigned to a protein provides a good relative estimation of its abundance when comparing two samples. The relative protein abundance can be compared using either stable isotope-labeling methods, or label-free methods as described earlier (Ong & Mann, 2005). A label-free method based on the extracted ion-chromatograms (XIC) was chosen for the quantitative work presented in this thesis. The method was selected for its flexibility as it can be applied to basically any sample type, allowing comparison of a larger numbers of samples as in paper I. Another option would have been spectral counting, but this approach was not considered due to its limitations when using long exclusion times, and the poor performance for small proteins (Zhang *et al.*, 2009). The XIC method has its own limitations, as samples are prepared in parallel which could result in unexpected variations, and the method is hampered by the low sensitivity of detection of small changes in protein abundance (Asara *et al.*, 2008). To compensate for possible variations in sample preparation introduced during and after protein digestion, a titrated mixture of stable isotope-labeled standard peptides was added to each sample, and used to normalize the data after acquisition. This approach allowed us to compare the relative protein abundance between over one hundred different samples. A similar approach was applied in paper V to profile the relative expression of membrane proteins along the colon. The use of isotopic labeling was considered, with isobaric labeling as the only feasible option with the number of samples and origin. Quantification using isobaric labels is performed after peptide fragmentation by detection of different low-mass reporter ions that are specific for each individual sample (Ross *et al.*, 2004). These specific ions are not retained when performing CID fragmentation with ion-trap detection, however, this can be overcome by performing pulsed Q dissociation (PQD), although the efficiency is debated and will require instrument dependent optimization (Bantscheff *et al.*, 2008). An alternative method available on linear ion trap-orbitrap instruments is HCD fragmentation with detection of the fragment ions in the orbitrap. However, the optimal collision energy required for accurate peptide quantification is generally higher than what is required for optimal peptide identification (Dayon *et al.*, 2010). Therefore both CID and HCD spectra are often acquired for each precursor, performing quantification on the HCD spectra while using the CID spectra for peptide identification (Köcher *et al.*, 2009). This strategy results in a significant increase in duty cycle and in theory limits the number of unique spectra sampled

by 50%. We therefore decided to use label-free quantification based on XIC as presented in paper V.

Protein identification by mass spectrometry (I, II, III, IV and V)

Identification of peptides from fragmentation spectra is a key step in mass spectrometry based proteomics. The data acquisition speed of modern mass spectrometers enables sequencing of thousands of peptides per hour, demanding efficient algorithms for protein identification and quantification. Developments of search algorithms for interpretation of peptide fragmentation spectra started in the late 1990s with the introduction of Sequest and Mascot (Link *et al.*, 1999; Perkins *et al.*, 1999). Before that period, peptide identifications were solely performed on the basis of the peptide mass, or by manually interpreting a small stretch of amino acids in an MS/MS spectra, which combined with the size of the N- and C-terminal fragments was matched against a protein database (peptide sequence tag) (Mann & Wilm, 1994). The basic difference in the new algorithm introduced was the use of theoretical fragmentation of a peptide for comparison towards the observed fragments in the spectra. Today there are various additional database search engines available, such as OMSSA (Geer *et al.*, 2004), Xtandem! (Craig & Beavis, 2004) Andromeda (Cox *et al.*, 2011) and Crux (Park *et al.*, 2008). The main difference between these programs is that they use different methods to score the fragmentation spectra, which will result in slight variations in the proteins identified. Combining the results from multiple search engines will therefore increase the number of protein identifications (McIlwain *et al.*, 2014).

We used Mascot for its overall good performance, and Andromeda for all of our quantitative experiments as it is integrated into the MaxQuant environment. Andromeda has some additional features that are not available in other software. It corrects the data at the MS level for systematic errors by performing a first and secondary search on recalibrated data, as well as an additional search for co-isolated peptides by removing the identified fragments and re-searching the spectra once more (Kryuchkov *et al.*, 2013).

Biopsy collection (I, II and V)

All experiments involving human tissue were approved by the ethical committee of the Sahlgrenska University Hospital. Written informed consent was obtained from all study subjects. In the present thesis, a large number of subjects were included in the various studies. The control material consisted of patients that were referred to colonoscopy for reasons such as bleedings of unspecified origin, polyp surveillance, diverticulitis or altered bowel habit, in which the colonoscopy and the macroscopic appearance of the mucosa was normal. The control patients in paper V were specifically selected to have no intestinal disease history. In addition to the control subjects, UC patients were included in the different studies presented in paper I. The UC patients were either referred to colonoscopy as a part of their disease surveillance program or due to clinical reasons related to the disease. The disease activity was evaluated in two ways. First by the gastroenterologist performing the colonoscopy via the endoscopic Mayo score (Lewis *et al.*, 2008), and then by a pathologist that evaluated the clinical biopsies obtained from the respective segments: caecum, ascending colon, right colon, transverse colon, left colon, descending colon,

sigmoid colon, and rectum. The UC patients were divided into two groups; remission and acute inflammation. The remission patients had an endoscopic Mayo score 0 and patients with acute inflammation had Mayo scores 1 to 3 and the histological profile was characterized by at least the presence of cryptitis or crypt abscesses. All patients underwent colonoscopy due to their own clinical symptoms or disease and the only additional procedure that was related to the research studies was obtaining extra biopsies. Data obtained from patients were collected in such a way that the patients' identities were kept confidential. Since our control group consisted of patients that were referred to colonoscopy due to various medical reasons it is possible that this patient group diverge from a control group composed of healthy volunteers with no history of bowel disorders.

RESULTS AND DISCUSSION

Composition of the human colonic mucus in control and UC patients (Paper I)

The main findings in paper I were:

- The human colonic core mucus is composed of approximately 50 proteins that were consistently identified in over 100 patients.
- We identified novel proteins providing new insights into the function of the highly organized mucus layer, such as CHGA, RNASE3 and CTSZ.
- The abundance of the known mucus components MUC2, FCGBP, CLCA1 and ZG16 was decreased in active UC patients.
- The previously observed loss of barrier function of the mucus layer could be correlated to reduced abundance of the main structural components MUC2 and FCGBP.

Ulcerative colitis (UC) is one of the most predominant inflammatory bowel diseases, causing chronic reoccurring inflammation of the colonic mucosa (Danese & Fiocchi, 2011). The underlying pathogenesis is still unknown, but one common hypothesis is that a genetically predisposed individual in combination with external factors will develop inappropriate immune responses towards the commensal flora. This will most likely involve more direct interaction between the microbiota and the host, which is in general prevented by the inner mucus layer (Johansson *et al.*, 2008). The protective inner mucus layer of the human colon is a highly structured system composed of multiple proteins. Patients with UC have an inner mucus layer that is more penetrable to bacteria, something that could be caused by an altered protein composition (Johansson *et al.*, 2014). In the present study we explored the mucus composition of healthy human colon, and compared the relative protein abundance between control patients and UC patients in various stages of the disease.

Sigmoid colon biopsies were collected from 111 patients undergoing routine colonoscopy, including 47 controls and 64 patients with UC, 36 of whom had active disease, and 28 in remission. Mucus samples were collected *ex vivo* from two biopsies per patient after 1 hour of incubation in a horizontal Ussing-type chamber (Gustafsson *et al.*, 2012b). The collected mucus was on-filter digested (FASP) by trypsin, and the resulting peptides were analyzed by LC-MS/MS (Wiśniewski *et al.*, 2009). Both protein identification and quantification were performed using the MaxQuant software package (Cox & Mann, 2008). These analyses resulted in the identification of over 1,500 different proteins, with a median of 955 proteins for the control group and 910 proteins for the UC patient group. These numbers comprise both the true mucus proteins, and protein derived from shed cells. True mucus components were distinguished from the background by applying the criteria that mucus proteins should contains either a signal peptide, membrane spanning domains or being lipidated, and the protein should be present in 95% of the samples. By applying these criteria on the initial protein list, we selected a group of 48 proteins that we consider to be the core mucus proteome.

Protein-protein interaction network analysis on this subset of proteins suggested that most are part of networks with similar biological function. The majority of the identified proteins were associated with two distinct functional groups: the established mucus components, and a set of ER proteins responsible for protein biosynthesis. Proteins in the latter group all have a signal peptide sequence, and an ER retention motif. Despite the fact that the majority of these proteins will be retained in the ER, we did not exclude this group from the mucus proteome since they fulfilled the inclusion criteria. Furthermore, certain protein disulfide isomerases have been shown to have additional functions outside the ER (Turano *et al.*, 2002).

The selected mucus proteins plus an additional 7 proteins that were identified in 95% of the UC patients were then used for further comparative analysis between the different patient groups. Firstly we investigated potential protein-protein interaction partners by correlating the abundances of the respective proteins between all the patients. When proteins are part of a complex one would expect similar ratios in abundance in all the analyzed samples (Smits *et al.*, 2012). Cluster analysis was performed on the correlation matrix and confirmed the results showing that the core proteome is composed of two functional groups; this analysis grouped together the proteins with an ER retentions signal, and the group of previously identified mucus protein (Johansson *et al.*, 2009).

These analyses did not reveal any major differences in the composition of the mucus proteome when comparing control and UC patients, and we proceeded to investigate whether the protein abundance levels differed between the three patient groups. The results showed that 14 of the 55 mucus proteome candidates were significantly changed between the groups. The largest variation in abundance level was observed between the controls and active UC patients. All the established mucus components (*i.e.* MUC2, FCGBP, ZG16 and CLCA1) were less abundant in the active UC patient group, independently of the degree of inflammation as assessed by the endoscopic Mayo score (Lewis *et al.*, 2008). One protein, RNASE3, was found to be more abundant in the active UC patients compared to the controls. RNASE3 was the only protein in the curated dataset that is secreted from immune cells in the *lamina propria.* This small eosinophil cationic protein has been shown to have a bactericidal effect, and was shown earlier to be more abundant both during acute inflammation and in remission (Saitoh *et al.*, 1999). RNASE3 was also the only established antimicrobial protein identified in the mucus, indicating that the thickness and density of the colonic mucus layer is sufficient to keep the bacteria at a distance from the epithelium. This is in contrast with the small intestine where Paneth cells actively secrete defensins and other proteins to keep the crypt sterile.

A subset of the control and UC patients included in this study were also included in a previously published study by our group focusing on the quality of the secreted mucus by measuring the penetrability to beads the size of bacteria (Johansson *et al.*, 2014). This study showed that UC patients with active disease secreted mucus that was more penetrable compared to that from control patients. To assess whether an increase in mucus penetrability correlated with the observed decrease in abundance of the structural mucus components MUC2 and FCGBP, we compared mucus penetrability against protein abundance. The results showed that patients with low levels of MUC2 and FCGBP had a more penetrable mucus, suggesting that in UC patients the secreted mucus is composed of less structural molecules, resulting in greater penetrability.

The presented work represents the first systematic study into the composition and relative protein abundance of the human colonic mucus. No major changes in mucus protein composition were observed between the different groups. However, fourteen proteins were found to have significantly altered levels between the control group and the UC patients. Thirteen proteins where less abundant in the active UC patients, including the major structural components of the mucus layer MUC2 and FCGBP. These two proteins were present in the mucus in an equimolar ratio suggesting that these proteins are secreted as part of an oligomeric network (Johansson *et al.*, 2009). The abundance of these two proteins could also be correlated to previously published data on the quality of the secreted mucus, where low abundance of MUC2 and FCGBP was associated with increased mucus penetrability to beads the size of bacteria (Johansson *et al.*, 2014)

In summary, the inner colon mucus layer is a dynamic structure that in the healthy colon efficiently separates the luminal bacteria from the epithelium. During acute inflammation the inner mucus layer is transformed to a structure that is more penetrable to bacteria. This loss of barrier function must have a biochemical explanation, and the current work suggests that lower levels of the MUC2 mucin and other goblet cell products results in secretion of a less dense and more penetrable mucus gel that does not provide the same level of protection.

Mucin degradation by bacterial proteases (Paper II)

The main findings in paper II were:

- The bacterium *Porphyromonas gingivalis* secretes a cysteine protease (RgpB) that can degrade the MUC2 mucin.
- The cleavage sites were found in the C-terminal region similar to previous observations for a protease secreted by *Entamoeba histolytica.*
- We identified the GalNAc-transferases expressed in the human colon and identified T3 and T7 to be responsible for the *O*-glycosylation around the cleavage site, preventing degradation.
- Co-expression of the GalNAc-transferases T3 and the MUC2-C terminal made the protein resistant to proteolytic degradation by RgpB.

The human gut hosts our commensal flora that is composed of over 1,000 bacterial species, reaching the highest density in the colon. The majority of these organisms belongs to the phyla *Bacteroidetes* and *Firmicutes* and are primarily found in the large intestine, where they live in the lumen and in the outer loose mucus layer (Qin *et al.*, 2010). The density of the inner mucus layer physically prevents bacteria from reaching the epithelium, thereby creating a bacteria free region between the sterile tissue and the lumen. Pathogenic bacteria such as *Citrobacter rodentium* have developed specific mechanisms to colonize the inner mucus layer in mice causing severe diarrhea (Bergstrom *et al.*, 2010). It is presumed that pathogenic bacteria express adhesins and proteases that allow them to overcome the mucus layer by binding and degrading the MUC2 mucin. This mechanism of colonization has been shown for the enterotoxigenic *Escherichia coli*. These bacteria are able to proteolytically degrade MUC2 by secretion of a serine protease (EatA), belonging to the serine protease autotransporters of the *Enterobacteriaceae* family found in a

broad range of pathogenic *E. coli* strains (Kumar *et al.*, 2014; Ruiz-Perez & Nataro, 2014). Previous research by members of our group have shown that the parasite *Entamoeba histolytica* has a similar mechanism to degrade the mucus layer by proteolytically cleaving MUC2 at two sites in the C-terminus (Lidell *et al.*, 2006). One of these cleavages occurred in a region just outside the PTS domain where the protein is not stabilized by disulfide bonds. Cleavage in this region will therefore dissolve the oligomeric structure of the mucus, and allow the pathogen to access the underlying epithelium. In this study we screened several bacterium for their potential to secrete proteases that are proteolytically active on the MUC2.

The study was performed using two recombinant proteins comprised of either parts of the N or C-terminus of MUC2. The recombinant N-terminal protein was composed of the first 1792 amino acids, and the C-terminal construct covered the last 981 amino acids (Godl *et al.*, 2002; Lidell *et al.*, 2003a). The recombinant proteins were incubated overnight with culture supernatants from various bacterial species. The initial screening step showed that the supernatant from the bacterium *Porphyromonas gingivalis* contained a protease that was able to cleave the MUC2 mucin. Analysis by SDS-PAGE was performed under reducing and non-reducing conditions and the results showed that secretions from *P. gingivalis* did not affect the MUC2 N-terminus, but degraded the MUC2-C dimer by cleaving the molecule at two positions. To identify the protease responsible for this proteolytic activity, the culture supernatant was fractionated by ion-exchange chromatography and the proteolytic activity of the different fractions was determined. Fractions that were able to cleave the MUC2-C protein were resolved by SDS-PAGE, and the protein content was characterized using mass spectrometry. These analysis identified three proteases, two arginine-gingipains (RgpA and B) and one lysine-gingipain (KgP), all associated with the pathogenicity of *P. gingivalis* (Guo *et al.*, 2010). Mutant bacterial strains deficient in the respective proteases were used to determine which of the three candidates was responsible for the cleavage. The results showed that the cysteine protease RgpB was responsible for the cleavage. This observation was further supported by pre-incubating the supernatant with the cysteine protease inhibitor E64, which abolished the proteolytic activity.

The exact cleavage sites were determined by protein N-terminal specific labeling and mass spectrometry analysis (Bertaccini *et al.*, 2013). One of the sites was shown to be in close proximity to the cleavage site previously observed for *E. histolytica* (Lidell *et al.*, 2006). This site allows disruption of the mucus gel, as it will not be stabilized by disulfide bonds any longer. However it is localized in a region of the protein that contains a high number of serine and threonine residues which have the potential to become *O*-glycosylated. The effect of *O*-glycosylation on the protein backbone on the protease activity was determined by *in vitro* glycosylation of a synthetic peptide (amino acid 4,306 – 4,335) covering the cleavage site. Proteomic analysis was conducted on isolated epithelial cells from human colonic tissue to determine which of the GalNAc-transferases are expressed in the tissue, as the 20 family members have different substrate activity (Bennett *et al.*, 2012). The GalNAc-transferases T1 - 5, 7 and 12 were identified, of which the first five are peptide specific, and T7 and T12 are glycopeptide specific, meaning they require initial GalNAc residues on the peptide for activity. The efficiency of the *in vitro* glycosylation of the synthetic peptide was determined by mass spectrometry using ETD fragmentation for the localization of the modified sites. Addition of one GalNAc to the threonine at the P2' position

from the cleavage site by GalNAc-transferase T3 abolished the proteolytic activity of RgpB. Interestingly, when the T3 glycosylated peptides were further processed by T7, nine out of ten potential *O*-glycosylation sites were found to be modified, highlighting the density of *O*-glycosylation of MUC2 that can be achieved. These results were confirmed in vivo by co-expressing the MUC2-C and GalNAc-T3 in CHO-K1 cells. The resulting recombinant protein was resistant to degradation by RgpB, and immunohistochemistry of colonic sections confirmed the localization of the GalNAc-T3 to the goblet cells.

These results highlight the importance of *O*-glycosylation on the MUC2 mucin in protecting the mucus from proteolytic degradation. Commensal bacteria express exoglycosidases for sequential release of carbohydrates and as the mucus is continuously secreted this degradation is in balance with the newly formed mucus. Pathogens such as the bacterium identified in this work express proteases that instead act on the protein backbone, resulting in rapid degradation of the mucus layer, potentially causing an inflammatory response. The majority of the unstructured PTS-domain is protected by *O*-glycans, and outside the PTS region the protein is protected from degradation by both *N*- and *O*-glycosylation in combination with extensive disulfide bonding. The high density of *O*-glycans is in general considered to be sufficient to protect the protein backbone, however, we show in this study that selected sites have to be modified to render the molecule resistant to proteases secreted by bacteria and the *E. histolytica* parasite.

Although the sequence homology of human MUC2 and mouse Muc2 is high, the arginine required for cleavage by RgpB is lacking in the mouse sequence, suggesting that mouse mucus is resistant to degradation by *P. gingivalis* as well as *E. histolytica*. This is supported by the fact that mice are naturally resistant to infection by *E. histolytica*. *P. gingivalis* is associated with periodontitis in the oral cavity, and the incidence of the bacterium in the gut is unclear, however members of the *Porphyromonadaceae* family are found in the human GI tract. The observation that both *E. histolytica* and *P. gingivalis* secrete proteases that act around the same site, suggests that this region is a susceptible region in the human MUC2 mucin that requires modification by *O*-glycosylation to protect it against degradation.

Characterization of the complex *N*- and *O*-glycosylation on the MUC2 C-terminus (Paper III)

The main findings in paper III were:

- Full characterization of a complex glycoprotein can be achieved by the use of complementary proteases in combination with ETD and HCD fragmentation, and accurate mass spectrometry.
- Almost all predicted *N*-glycosylation consensus sequences on MUC2 were found to be modified (17/18) and over 52 different glycopeptide forms were identified.
- *O*-glycosylation outside the PTS region was observed for the first time on MUC2, revealing 4 specific sites.

The extensive glycosylation on the MUC2 mucin is an important feature of the molecule and essential for formation of a functional mucus layer. The attached glycans account for an estimated 80% of the mass of the secreted protein and are mainly found on the two central PTS-domains (Carlstedt *et al.*, 1993). The PTS-domains are largely composed of repeats of the amino acids

serine, proline and threonine; length, sequence and mucin specific. The added *O*-glycans are responsible for binding water and a selective niche for bacteria (Hooper & Gordon, 2001). The two terminal regions of the protein contain 30 predicted *N*-glycosylation consensus sequences, 12 in the N-terminal, and 18 in the C-terminal part of the protein. The significance of *N*-glycosylation on mucin molecules has not received much attention. It has been suggested that *N*-glycosylation of the two terminal regions of MUC2 are responsible for folding of these regions in the ER and for further multimerization (Asker *et al.*, 1998). Mutation of the 9th *N*-glycosylation consensus sequence in the cystine-knot results in intracellular accumulation of MUC2, and significant loss of dimer formation (Bell *et al.*, 2003). The role of *O*-glycosylation outside the tandem repeats is unclear although it was recently shown that MUC1, MUC4, MUC5B, MUC16, MUC17, and MUC20 bear *O*-glycans outside the PTS region (Steentoft *et al.*, 2013). Studies of MUC2 *O*-glycosylation by mass spectrometry is routinely performed by releasing the glycans from the protein backbone and analyzing the released structures (Karlsson *et al.*, 1997). The majority of glycans detected by this method will be from the PTS domains, however all site-specific information is lost by releasing the glycan from the protein backbone, making it impossible to determine site occupancy. Here we explored the possibility of studying site specific *N*- and *O*-glycosylation on the MUC2 C-terminus by digesting the protein backbone into glycopeptides. These smaller glycopeptides can then be subjected to mass spectrometry analysis to characterize both site occupancy and glycan composition. To enable identification of all potential glycosylation sites we digested the protein with multiple enzymes, both specific and non-specific (Clowers *et al.*, 2007; Swaney *et al.*, 2010).

We used a recombinant protein (MUC2-C) composed of the last 981 amino acids of the human MUC2 (amino acids 4198 –5179) flanked at the N-terminal side by GFP and a Myc tag (Lidell *et al.*, 2003a). The purified, reduced and alkylated protein was analyzed by gel electrophoresis, and the fully glycosylated protein migrated at approximately 250 kD. The band was excised, subjected to in-gel digestion using trypsin, and analyzed by LC-MS/MS. This approach resulted in a sequence coverage of 36% based on the identified unmodified peptide sequences, and 39% when analyzing the de-*N*-glycosylated protein. To increase peptide coverage we evaluated other enzymes for protein digestion. Asp-N, subtilisin, and chymotrypsin were used to digest de-*N*-glycosylated recombinant and the peptides were analyzed by LC-MS/MS. Asp-N did not increase the sequence coverage, however subtilisin and chymotrypsin increased the combined coverage to 86%. The N-terminal part of the recombinant MUC2-C protein resembles the PTS regions and lacks potential cleavage sites for the enzymes used in this study. Due to this, the N-terminal part was not covered in our analysis, which is the main reason why full sequence coverage was not achieved.

Seventeen out of the eighteen predicted *N*-glycosylation sites were found deamidated after PGNase F treatment, indicating prior attachment of an *N*-glycan. When analyzing the *N*-glycosylated protein, 52 *N*-glycopeptides were identified including 38 glycoforms at 10 of the 18 potential *N*-glycosylation consensus sites. A broad heterogeneity was observed for the individual sites with up to seven forms per peptide. The spectra of 15 *O*-glycopeptides covering four different sites were elucidated in the different enzymatic digests. These results showed for the first time that *O*-glycosylation of MUC2 occurs outside the PTS domain. The characterized

structures were simple core-1 type structures with a maximum of 4 glycan residues, commonly observed in CHO-K1 cells that were used to produce the recombinant protein.

We could demonstrate that substantial information on site-specific glycosylation of a large purified glycoprotein can be obtained by combining standard proteomics methods without further preparative steps enriching for glycopeptides. This is an attractive approach to use when it is possible to obtain purified starting material. Most methodologies described for glycopeptide enrichment have the drawback of selecting for specific types of *N*-glycans or *O*-glycans, such as lectin based approaches or by chemical modification of the glycans (Nilsson *et al.*, 2009; Zielinska *et al.*, 2010). The identified *N*- and *O*-glycosylation sites add to information regarding the structure of the molecule, and which sites that are protected against proteolytic degradation in the intestine (Lidell *et al.*, 2006; van der Post *et al.*, 2013).

Function of the CysD domain in the MUC2 mucin (Paper IV)

The main findings in paper IV were:

- The second CysD domain in the MUC2 mucin forms non-covalent dimers.
- All cysteine residues are involved in formation of intramolecular disulfide bonds.
- The previously suggested *C*-mannosylated WXXW motif was unmodified in our analysis.

The major structural protein responsible for the formation of the mucus gel in the intestine is the MUC2 mucin. MUC2 is comprised of five distinct regions: an N-terminal part with von Willebrand D1-D2-D'-D3 domains and a CysD domain, a small PTS domain, a second CysD domain, a large PTS (tandem-repeated) domain, and a C-terminal part with von Willebrand D4-B-C domains and a cystine-knot (Perez-Vilar & Hill, 1999). The protein has been shown to form disulfide-linked dimers between the C-termini in the endoplasmic reticulum (Asker *et al.*, 1998), and disulfide-linked trimers via its N-terminal region in the trans-Golgi network (Godl *et al.*, 2002). This results in highly organized oligomeric net-like structures that upon secretion form the mucus layers that protect the underlying epithelium. Polymer formation is clearly depending on the von Willebrand domains and cystine-knot (Ambort *et al.*, 2012). However, the role of the CysD domain has remained elusive. This domain is almost exclusively found in the secreted mucins MUC2, MUC5B and MUC5AC, adjacent to or scattered within the heavily *O*-glycosylated PTS-domains. The domain spans 97 amino acids and contains 10 conserved cysteine residues, and a highly conserved WXXW motif of which the first tryptophan has the potential to become C-mannosylated (Löffler *et al.*, 1996). The CysD domain was recently adopted in various protein domain databases as referred to as WxxW domain since this as the most conserved part (IPR025155 and PF13330). To obtain a better understanding of the role of the CysD domain we studied the function and biochemical properties of the second CysD domain in the human MUC2.

The study was performed using a recombinant protein composed of a Myc tag, the second CysD in human MUC2 (residues 1782–1878), exons 1– 3 of the murine IgG-Fc region, and a C-terminal polyhistidine-tag. An enterokinase cleavage site was introduced between the CysD and the IgG part to enable separation of the two parts after protein-G purification. In addition, a

second protein construct was generated in which two PTS repeats (PTTTPITTTTTVTPTPTPTGTQT) were added after the CysD domain. This construct was used to study the effect of *O*-glycosylation on protein function. The first construct was analyzed by SDS-PAGE under reducing and non-reducing conditions, with or without enterokinase treatment. Under reducing conditions the CysD and IgG part of the protein were observed both before and after enterokinase treatment. However, when analyzed under non-reducing conditions the CysD part could not be observed after enterokinase treatment. This led to the conclusion that removal of the IgG part results in formation of insoluble aggregates. These aggregates are likely not formed in the full construct due to steric hindrance.

For further analysis, the recombinant protein was analyzed under native PAGE conditions after gel filtration. The analyzed intact protein formed four distinct bands representing tetra-, octa-, dodeca- and hexadeca-mers, with the first two as the predominate species. When the combined fractions were reduced, only the monomeric band representing the CysD-IgG construct was observed. Gel filtration following enterokinase treatment allowed for the separation of the CysD and IgG parts both eluting as dimers as observed by native PAGE. The second recombinant protein with the added PTS repeats showed similar results except for an increase in mass due to the added glycan moieties, indicating that the dimers are held together via non-covalent interactions, which could not be disrupted by analyses at varying pH and calcium concentrations observed by gel chromatography. To confirm that no intermolecular stabilizing disulfide bonds were formed between the domains the disulfide linkages were resolved by mass spectrometry. The protein was resolved by SDS-PAGE under non-reducing conditions followed by digestion with Asp-N, and the resulting peptides were analyzed. All ten cysteine residues of the CysD domain were determined to form five intramolecular disulfide bonds, and no indications of any intermolecular disulfide bonds could be found. In addition, the proposed *C*-mannosylation site in the WxxW domain was found to be unmodified, which contradicts previously published work on the CysD domains (Perez-Vilar *et al.*, 2004). However, these studies were performed on CysD domains from MUC5AC and MUC5B, suggesting potential heterogeneity in the glycosylation exists between CysD from different mucins.

This study showed that the CysD domain in the MUC2 mucin forms non-covalent dimers. The in this study used recombinant protein contained the Fc region of mouse IgG known to form 3 intermolecular disulfide links resulting in secretion of the produced protein as a dimer. However, after release of the IgG part the CysD still formed dimers, as confirmed by both native PAGE and gel filtration analysis. Dimerization of the CysD resulted in even higher oligomeric structure such as tetra-, octa-, dodeca- and hexadeca-mers as observed by native PAGE by heteromerization of the IgG dimers.

Based on these results we propose that the CysD domain is responsible for additional cross-linking of MUC2 in the mucus gel. As the CysD domains are found flanking the linear *O*-glycosylated PTS-domains, this type of further cross-linking could be responsible for regulating the pore size of the mucus gel. As the number of CysD domains and the distance between the domains varies among different gel-forming mucins, it is expected that the respective mucins produce a gels with a porosity that is adapted to the local requirements in the specific organ. Both MUC5AC and MUC5B contain a higher number of CysD domains compared to MUC2,

suggesting that these proteins form a denser mucus gels. However, additional functional and biochemical studies are required to confirm this hypothesis.

Profiling of the membrane protein expression along the human colon (Paper V)

The main findings in paper V were:

- The majority of the epithelial membrane proteins show stable expression along the length of the colon.
- The human proximal colon participates in nutrient digestion and absorption.
- The relative expression levels of glycosyltransferases responsible for *O*-glycosylation can be used as an indirect measurement of the terminating glycans found on the MUC2 mucin.
- We confirmed regional differences in the expression of a number of ion-transporters, and identified the segmental regulation of the goblet cell specific sodium bicarbonate co-transporter NBCN1.

The colon is a highly dynamic organ in which a single layer of epithelial cells separates the densely colonized intestinal lumen from the largely sterile tissue, while still allowing active transport over the membrane. One of the main functions of the colon is to reabsorb water, ions and other key nutrients from waste material and recycle it back into the body, a processes known to gradually change along the length of the colon (Sandle, 1998). Gene expression data have shown that the expression level of various transporters and enzymes changes along the length of the human colon (LaPointe *et al.*, 2008). However, knowledge concerning the general protein expression along the axis of the healthy human colon is still limited. Segmental variation in protein expression is often ignored in studies of colonic function, although it is well known that diseases such as UC and Crohn's disease have clear segmental patterns. Improved understanding of region specific protein expression could be used to unravel some of the underlying mechanisms to why a disease manifests at a specific location in the colon. As all types of transport over the plasma membranes involve transmembrane spanning proteins the focus of this study was on this selected group of proteins.

Mass spectrometry analyses were performed on colonic biopsies obtained from the ascending, transverse, descending, and sigmoid colon, thereby covering almost the entire length of the colon. The protein composition was characterized and relative protein quantification was performed using a label-free approach. The method used for relative protein quantification was based on the sum of extracted ion-chromatograms for all peptides assigned to a protein (Chelius & Bondarenko, 2002). Colonic tissue was obtained from patients with no known colon disease history and with macroscopically normal mucosa. The epithelial cells were isolated from the tissue followed by extraction of membrane proteins. A total of four patients were included in this study, and the number of identified proteins ranged between 2,598 and 2,682 proteins per patient. Since we were specifically interested in the segmental distribution of the identified proteins, only the proteins that were identified in all four segments (1,729) were considered for further data analysis. Hierarchical clustering analyses were performed based on the relative protein intensities

for each protein in the different samples. The ascending and sigmoid segments from the different patients grouped together in separate clusters while the transverse and descending segments could not be differentiated. These results indicate that the two ends vary more in function compared to the middle region of the colon.

The efficiency of the membrane protein enrichment method was evaluated as these proteins tend to be underrepresented in proteomics studies due to their amphiphilic nature. Based on protein abundance estimations it was observed that membrane proteins were among the more abundant proteins (Schwanhäusser *et al.*, 2011). This trend was also observed on a global level when comparing the number of protein identifications in our study to the proteins that are expressed in colon epithelia based on antibody staining (Uhlén *et al.*, 2005). The percentage of protein identifications was higher among the more hydrophobic proteins and the ones with predicted transmembrane spanning domains, indicating that membrane proteins were efficiently enriched in our analysis.

A total of 261 proteins were significantly ($p < 0.1$) regulated between the ascending-transverse and descending-sigmoid colon. The biological functions of these proteins were used for enrichment analysis to determine which functions changed the most. In the ascending colon, protein synthesis was found to be the most enriched biological function. The more bioactive active epithelium in the proximal colon has a higher cell turnover, which requires an increase in protein biosynthesis. This observation is in line with the overall decrease in protein synthesis observed along the small intestine that is expected to continue to decline towards the distal colon (Nakshabendi *et al.*, 1999). Furthermore, metabolic processes were increased in the proximal part of the colon suggesting that the final stages of digestion are still ongoing in the proximal part of the colon. The resemblance with the ileum was also reflected in the increased expression of major histocompatibility complex (MHC) proteins responsible for presentation of foreign material by the intestinal epithelial cells to the immune cells. In the small intestine the mucus structure is non-adherent and more open allowing direct interactions between the epithelia cells and the luminal bacteria. Further distally in the digestive tract the mucus layer becomes denser preventing these types of interactions (Ermund *et al.*, 2013). In the distal colon, the proteins with the highest upregulation were the glycosyltransferases responsible for various stages of protein *O*-glycosylation. These proteins are selectively found in the Golgi membrane, and as the membrane enrichment method used in this study was not selectively enriching for the plasma membrane these proteins were also consistently identified. The major *O*-glycosylated protein in the intestine is the MUC2 mucin and the expression levels of the glycosyltransferases directly reflect the heterogeneity in glycan structures previously observed in glycomic studies (Robbe *et al.*, 2003; Holmen Larsson *et al.*, 2009). Segmental differences were mainly observed in the terminating glycans, showing a decrease in sulfation and an increase in sialylation in the proximal to distal direction.

When focusing on individual proteins, the significance threshold was lowered ($p < 0.05$) for the comparison of the two colon ends. This resulted in 144 proteins, of which 105 were predicted to contain transmembrane spanning domains. Among the regulated proteins various members of the solute carrier family (SLC) were identified; these are involved in transport over the plasma membrane. The reabsorption of ions is an important function of the colon and we identified the transporters at a protein level that are likely involved in ion-transport at the apical and basolateral

membrane. NHE3 (SLC9A3) and NBCn1 (SLC4A7) were significantly upregulated; NHE3 expression was previously shown to decrease towards the sigmoid colon (Farkas *et al.*, 2011). NBCn1 is a goblet cell specific sodium-bicarbonate co-transporter, and of particular interest in the context of regulation of mucus properties, as we and others have shown that bicarbonate secretion plays an important role in regulation of mucus formation (Gustafsson *et al.*, 2012a; Singh *et al.*, 2013).

This dataset is the first comprehensive profile of membrane protein levels along the length of the healthy human colon, emphasizing regional heterogeneity. This information can be used to obtain a better understanding of how various biological processes vary along the proximal-distal axis. Additionally, this dataset can be used as a reference when comparing between healthy and diseased tissue specimens, such as samples from patients with inflammatory bowel disease or colon cancer.

GENERAL CONCLUSIONS

The studies in this thesis were all aimed towards increasing the understanding of the functioning of the colonic mucus layer, with a special focus on the main component the MUC2 mucin and how its structure, modifications and abundance affect its function.

The MUC2 mucin is densely packed inside the goblet cells and upon secretion the protein expands into large oligomeric sheets that cover the surface of the intestinal epithelium with a dense layer of mucus. Under normal conditions the colonic mucus prevents direct contact with the billions of commensal bacteria that reside in the lumen and the intestinal epithelium. The properties of the intestinal mucus layer have been shown to be affected by various factors such as bacteria, the immune system and ion secretion. The underlying mechanisms as to how these factors alter the mucus barrier are not fully understood, although presumed to affect either the rate of synthesis and/or secretion and proper expansion of the protein into oligomeric sheets. Oligomerization of MUC2 occurs via di- and trimerization of the heavily disulfide stabilized termini. The N-terminus also contains CysD domains and our results showed that these domains are able to form additional cross-links by dimerization, allowing formation of even more elaborate complexes by linking the respective oligomeric sheets. Interestingly, this domain is almost exclusively found in secreted mucins, and the number of domains varies between the various secreted mucins, suggesting that the number of domains determines the degree of polymerization and thereby pore size. In between the two termini lies the highly *O*-glycosylated mucin domain. The high density of *O*-glycosylation on MUC2 is essential for the ability to bind water for gel formation and lubrication, and protecting the protein from proteolytic degradation by the commensal microbiota. Bacteria use the glycans as an energy source and produce exoglycosidases that release the oligosaccharides stepwise from the terminal end. Glycan epitopes have been shown to differ along the length of the colon, with increasing rates of sialylation towards the distal colon. In this part of the colon the bacterial load reaches its maximum. However, since only selected bacterial species are able to release terminal sialic acids from the *O*-glycans on MUC2, the mucus gel is protected against rapid degradation. By analyzing the membrane protein expression and relative abundance along the intestine we could confirm an increase in abundance of sialyltranferases in the distal colon compared to the proximal colon, together with other glycosyltransferases, which correlated with the observation made by glycomics studies of the MUC2 oligosaccharides.

The stepwise release of glycans from the terminal end is a relatively slow process and is presumed to be in concordance with the rate of mucus secretion, thereby preventing the bacteria from penetrating far into the mucus, and the slow rates of glycan degradation prevent proteases from acting on the peptide backbone.

In this model the length and density of the oligoaccharides is more important than the localization, as the possible modification sites in the PTS domain are abundant. However, outside the PTS region localization of the glycans is more important as the number of serines and threonines is limited. When we investigated the site-specific glycosylation of the C-terminus we identified one particular site that required modification in close proximity to prevent proteolytic degradation. Cleavage of the protein at this particular site would lead to disruption of the mucus polymers, thus proper glycosylation of this region plays an important role in the integrity of the

mucus barrier. The question that remains to be answered is how often this site remains unmodified and under what conditions. We hypothesize that when the tissue is under stress and forced to secrete more mucus the *O*-glycan density is reduced. This hypothesis is supported by studies showing that the rate of mucus secretion is increased in UC, and that the MUC2 is covered by shorter *O*-glycans. The shorter glycans will be removed more rapidly by bacterial exoglycosidases making the identified site accessible for proteases, and reducing the overall function of the molecule. Increased rates of mucus secretion in UC patients has been observed, highlighted by less filled goblet cells and excessive mucus in the stool. In addition to an altered glycosylation pattern we observed a reduction in protein abundance of MUC2 in UC patients, which suggests that the pool of mucus is more rapidly depleted, reflected in thinner mucus with a decreased barrier function as observed by Johansson *et al* 2014.

As with the majority of the –omics studies our work resulted in large amounts of data which biological importance was only partly understood. The results did however provided us with new insights on the MUC2 protein, mucus, and the underlying cells and complemented many ongoing projects in our laboratory, and the data is made publicly available for other researchers. Overall the work in this thesis has shown that intestinal mucus is a dynamic matrix, and mass spectrometry can be successfully used to study its protein composition, proteolytic degradation and modifications.

FUTURE PERSPECTIVES

The results presented in this thesis provide novel insights into the composition and processing of the protective mucus layer in the colon, and how mass spectrometry can be applied to study its various aspects. However, many questions and technical challenges still remain to be solved.

The methods established in this thesis work are now routinely used in our laboratory for the analysis of mucus samples from various parts of the gastrointestinal tract. All the quantitative work presented in this thesis is based on relative quantification between various conditions, and a logical continuation would be to establish methods to measure the absolute quantity of MUC2 and associated mucus proteins. This can be achieved by the addition of known concentrations of synthetic isotopically labeled peptides for the proteins of interest before mass spectrometry analysis. The outcome of these experiments will provide information on the molar amount of protein per volume of mucus that is required for a proper functioning mucus barrier.

The importance of MUC2 as the major structural component in the mucus gel is well established. However, the biological function of the majority of the additional mucus proteins remains elusive. In particular, the role of FCGBP, CLCA1 and ZG16 in the mucus is of great interest as these proteins are secreted from the goblet cells simultaneously with MUC2. The biochemical characterization of these proteins is therefore of importance to understand their specific role and their influence on the mucus properties.

All protein features of MUC2 are aimed to generate a molecule that is resistant to the harsh environment in the intestinal lumen. However, parts of the protein sequence were shown to be prone to proteolytic degradation by pathogens in cases where the sequence was not protected by *O*-glycosylation. As our experiments were performed on recombinant protein expressed in CHO cells, the question remains if the same sites are accessible *in vivo*. The GalNAc-transferases required for modification were identified in the epithelium along the complete axis of the colon suggesting that this site is modified in the human colon. Analysis of site-specific glycosylation of the protein *in vivo* is challenging but could be performed following the presented methodology. The identified cleavage site is lacking in mice, which limits studies regarding this particular site using mouse models. To determine whether proteolytic degradation of MUC2 is a general mechanism for pathogens to colonize the colon, additional bacterial species can be tested for secretion of proteases that are able to digest the recombinant MUC2-C, such as various pathogenic *E.coli* strains.

ADDITIONAL BIBLIOGRAPHY

1. Bell, A. W., Deutsch, E. W., Au, C. E., Kearney, R. E., Beavis, R., Sechi, S., Nilsson, T., Bergeron, J. J. M., Beardslee, T. A., Chappell, T., Meredith, G., Sheffield, P., Gray, P., Hajivandi, M., Pope, M., Predki, P., Kullolli, M., Hincapie, M., Hancock, W. S., Jia, W., Song, L., Li, L., Wei, J., Yang, B., Wang, J., Ying, W., Zhang, Y., Cai, Y., Qian, X., He, F., Meyer, H. E., Stephan, C., Eisenacher, M., Marcus, K., Langenfeld, E., May, C., Carr, S. A., Ahmad, R., Zhu, W., Smith, J. W., Hanash, S. M., Struthers, J. J., Wang, H., Zhang, Q., An, Y., Goldman, R., Carlsohn, E., **van der Post, S.**, Hung, K. E., Sarracino, D. A., Parker, K., Krastins, B., Kucherlapati, R., Bourassa, S., Poirier, G. G., Kapp, E., Patsiouras, H., Moritz, R., Simpson, R., Houle, B., LaBoissiere, S., Metalnikov, P., Nguyen, V., Pawson, T., Wong, C. C. L., Cociorva, D., Yates, J. R., III, Ellison, M. J., Lopez-Campistrous, A., Semchuk, P., Wang, Y., Ping, P., Elia, G., Dunn, M. J., Wynne, K., Walker, A. K., Strahler, J. R., Andrews, P. C., Hood, B. L., Bigbee, W. L., Conrads, T. P., Smith, D., Borchers, C. H., Lajoie, G. A., Bendall, S. C., Speicher, K. D., Speicher, D. W., Fujimoto, M., Nakamura, K., Paik, Y.-K., Cho, S. Y., Kwon, M.-S., Lee, H.-J., Jeong, S.-K., Chung, A. S., Miller, C. A., Grimm, R., Williams, K., Dorschel, C., Falkner, J. A., Martens, L., and Vizcaíno, J. A. (2009) **A HUPO test sample study reveals common problems in mass spectrometry–based proteomics.** *Nat Methods* **6**, 423–430
2. Johansson, M. E. V., Ambort, D., Pelaseyed, T., Schütte, A., Gustafsson, J. K., Ermund, A., Subramani, D. B., Holmén-Larsson, J. M., Thomsson, K. A., Bergström, J. H., **van der Post, S.**, Rodriguez-Piñeiro, A. M., Sjövall, H., Bäckström, M., and Hansson, G. C. (2011) **Composition and functional role of the mucus layers in the intestine.** *Cell Mol Life Sci* **68**, 3635–3641
3. Rodriguez-Piñeiro, A. M., **van der Post, S.**, Johansson, M. E. V., Thomsson, K. A., Nesvizhskii, A. I., and Hansson, G. C. (2012) **Proteomic study of the mucin granulae in an intestinal goblet cell model.** *J Proteome Res* **11**, 1879–1890
4. Wahlgren, W. Y., Omran, H., Stetten, Von, D., Royant, A., **van der Post, S.**, and Katona, G. (2012) **Structural characterization of bacterioferritin from blastochloris viridis.** *PLoS ONE* **7**, e46992
5. Skogberg, G., Gudmundsdottir, J., **van der Post, S.**, Sandström, K., Bruhn, S., Benson, M., Mincheva-Nilsson, L., Baranov, V., Telemo, E., and Ekwall, O. (2013) **Characterization of human thymic exosomes.** *PLoS ONE* **8**, e67554
6. Pelaseyed, T., Bergström, J. H., Gustafsson, J. K., Ermund, A., Birchenough, G. M. H., Schütte, A., **van der Post, S.**, Svensson, F., Rodriguez-Piñeiro, A. M., Nyström, E. E. L., Wising, C., Johansson, M. E. V., and Hansson, G. C. (2014) **The mucus and mucins of the goblet cells and enterocytes provide the first defense line of the gastrointestinal tract and interact with the immune system.** *Immunol. Rev.* **260**, 8–20
7. Skogberg, G., Lundberg, V., Lindgren, S., Gudmundsdottir, J., Sandström, K., Kämpe, O., Annerén, G., Gustafsson, J., Sunnegårdh, J., **van der Post, S.**, Telemo, E., Berglund, M., and Ekwall, O. (2014) **Altered expression of autoimmune regulator in infant down syndrome thymus, a possible contributor to an autoimmune phenotype.** *J. Immunol.* **193**, 2187–2195

8. Hannan, T. J., Roberts, P. L., Riehl, T. E., **van der Post, S.**, Binkley, J. M., Schwartz, D. J., Miyoshi, H., Mack, M., Schwendener, R. A., Hooton, T. M., Stappenbeck, T. S., Hansson, G. C., Stenson, W. F., Colonna, M., Stapleton, A. E., Hultgren, S. J. (2014) **Inhibition of cyclooxygenase-2 prevents chronic and recurrent cystitis,** *EBioMedicine***,** online

ACKNOWLEDGEMENTS

This work could not have been done without the help and support of a lot of people and I hereby would like to thank all of you.

Gunnar Hansson, first of all thank you for accepting me as your student, and introducing me to the complex world of mucin biology. Thanks for all the guidance during these years, and for always taking the time to discuss science and your general excitement for new results. I admire your dedication and unstoppable enthusiasm.

My co-supervisors **Malin Bäckström** and **Niclas Karlsson** thank you for your help and support during these years.

All my co-authors for contributing to this thesis.

A big thanks to all the past and present members of the Mucin biology group: **Ana** for never breaking the silence before Örebro on our early train trips to Stockholm. Looking forward to crawl around in the library together during the coming weeks. **Anna** your dedication is admirable. I will miss your descriptions of mucus; blobs, fluffy, gooey and sticky it's like we were discussing kanelbullar. **Catharina** for keeping the lab running, and for always being helpful. **Christian** our in office encyclopedia, one day the captain will sail out again. I will miss our fruitful discussions about science and your chess problems over a coffee cup of stone cold wine. **Elizabeth** your baking skills are memorable, good luck with your own PhD and beyond. **Evelin** I appreciate your cheerfulness (and cookies), best of luck with your proteomics studies. **Frida** it has been great to have you around for all these years, thanks for all the help with my accidental DNA work. **Karolina** thanks for all the help with the patient samples. **George** the uncrowned king of the cold-room. Thanks for cooping with my idiosyncrasy. It has been a pleasure to discuss science with you; we need more videos and fluorescence. **Karin**, I have been anxious for years to order anything, but I think I am over it. You are the best "lab mamma" one can imagine. **Hannah** I admire your energy and general excitement, it is always great to run into you in the corridor. **Hedvig** unfortunately I will miss your return to the office but I will try to clean out my desk before you return to see the mess. It has been great to share office with you, you are a great person. **Joakim** I hope that everything goes well with your own defense. It was of great support to know I was not the only one stressed in the basement. **Johanna** welcome to the group and good luck with everything. **Lauren** our "poster girl", I hope to see you in San Francisco one day. **Liisa** our trip to Japan was awesome, and you became a true friend. Sorry for pushing you on our runs! I still expect a 1:45 next year… I know the mass spec will be "safe" with you around. **Lisbeth** thanks for helping with sample collection, and for teaching me that you should sit down for a minute after a vaccination. **Malin B.** thanks for all the discussions and help with various recombinants. **Malin J.** mucin veteran, you are a wonderful person. Thanks for inviting us to your summerhouse last summer, but why did you try to boil us? **Maria S** (I hide you here) thanks for all the nice concerts, festival and other crazy nights. **Mattias** (you are one of us), I really enjoyed our fishing trips, one day I will catch a sea trout. **André** for all the great evenings playing

board games, and all those projects that looked great on paper. Good luck with your new position. **Babu** all the best in your career. **Daniel** Andra långgatan will never be the same without you. **Ida** you are cool. **Jessica** switching from glycomics to proteomics… the dark side is calling. **Karl** my one and only student, I hope you learned something from all our failures. Good luck with your own PhD. **Lisa** good luck with your new career and thanks for taking care of the Koss. **Noreen** thanks for the many samples collected. **Robert** I can still smell you bacterial cultures. **Thaher** all the best on your American adventure, hope to see you one day in khaki pants and trainers at the local baseball game. Why did you take Pearl Jam when you left? **Tina** you are such an inspiring person, reminding us all that there is more in life than science. It is amazing that you came all the way from New Zeeland to be here today.

Thomas Larsson and **Hasse Karlsson** for introducing me to mass spectrometry, it has been the basis for this thesis.

The members of **MPE**, **Dan**, **Sara**, **Niclas** and **Susann's** groups for making this a great place to work at. **Elizabeth** thanks for all the help with the recombinant proteins and nice chats. **Emma** it has been great having you around; I could use some of your relaxed attitude. **Harvey** thanks for the last minute proof reading and for introducing cider into my life. One day I will score more than two points in a game of squash. **Liaqat** congratulations with your own thesis and all the best for the future. **Richard** thanks for letting me use your apartment and for on demand birds knowledge. **Sarah** for always being happy and helpful, and all the cute little presents for Saga.

Everybody at the Proteomics core-facility and especially: **Carina**, for always taking the time to discuss everything mass spec. **Diarmuid,** hope to wake up again one day next to you in a hotel room with a pink minibar, our tours to ASMS were heroic. **Elisabeth** thanks for the support in my early MS life. **Jörgen** your dedication to OSX is admirable. **Petra** it was great fun to share an office with you back in the days.

Internet warriors **Erik** en **Peter** het is altijd een waar genoegen om een Donderdag avond te "verknallen" in jullie gezelschap.

Mijn familie: Ouders jullie zijn geweldig. Het heeft een paar jaar geduurd, maar ik denk dat het toch nog wat geworden is met jullie zoon. Bedankt voor alle steun. **Roy, Sanne, Pien** en **Kaat** het is altijd een feest jullie weer te zien, bedankt voor alles. **Olle, Ulla, Johan, Ida** and **Ellen** (welcome) thanks for everything during the years, and for welcoming me into your family.

Saga you have been a wonderful distraction during the last couple of months. You are such a cute little girl, always happy and I am enjoying every second with you.

Jenny, love of my life you're the best. I really appreciate your input and editing on all the manuscripts, thesis and applications the last couple of (boring) weeks. I'm excited to our next adventure in St. Louis. I know that it's going to be a great. "Anything to make you smile". Jag älskar dig

REFERENCES

Ahlf, D. R., Compton, P. D., Tran, J. C., Early, B. P., Thomas, P. M. & Kelleher, N. L. (2012). Evaluation of the Compact High-Field Orbitrap for Top-Down Proteomics of Human Cells. *Journal of Proteome Research* 11(8), 4308–4314.

Albert, T. K., Laubinger, W., Müller, S., Hanisch, F.-G., Kalinski, T., Meyer, F. & Hoffmann, W. (2010). Human intestinal TFF3 forms disulfide-linked heteromers with the mucus-associated FCGBP protein and is released by hydrogen sulfide. *Journal of Proteome Research* 9(6), 3108–3117.

Alpert, A. J. (1990). Hydrophilic-interaction chromatography for the separation of peptides, nucleic acids and other polar compounds. *Journal of chromatography* 499, 177–196.

Ambort, D., Johansson, M. E. V., Gustafsson, J. K., Nilsson, H. E., Ermund, A., Johansson, B. R., Koeck, P. J. B., Hebert, H. & Hansson, G. C. (2012). Calcium and pH-dependent packing and release of the gel-forming MUC2 mucin. *Proceedings of the National Academy of Sciences* 109(15), 5645–5650.

Apweiler, R., Hermjakob, H. & Sharon, N. (1999). On the frequency of protein glycosylation, as deduced from analysis of the SWISS-PROT database. *Biochimica et biophysica acta* 1473(1), 4–8.

Arumugam, M., Raes, J., Pelletier, E., Le Paslier, D., Yamada, T., Mende, D. R., Fernandes, G. R., Tap, J., Bruls, T., Batto, J.-M., Bertalan, M., Borruel, N., Casellas, F., Fernandez, L., Gautier, L., Hansen, T., Hattori, M., Hayashi, T., Kleerebezem, M., Kurokawa, K., Leclerc, M., Levenez, F., Manichanh, C., Nielsen, H. B., Nielsen, T., Pons, N., Poulain, J., Qin, J., Sicheritz-Ponten, T., Tims, S., Torrents, D., Ugarte, E., Zoetendal, E. G., Wang, J., Guarner, F., Pedersen, O., De Vos, W. M., Brunak, S., Doré, J., Antolin, M., Artiguenave, F., Blottiere, H. M., Almeida, M., Brechot, C., Cara, C., Chervaux, C., Cultrone, A., Delorme, C., Denariaz, G., Dervyn, R., Foerstner, K. U., Friss, C., Van De Guchte, M., Guedon, E., Haimet, F., Huber, W., Van Hylckama-Vlieg, J., Jamet, A., Juste, C., Kaci, G., Knol, J., Lakhdari, O., Layec, S., Le Roux, K., Maguin, E., Mérieux, A., Melo Minardi, R., M'rini, C., Muller, J., Oozeer, R., Parkhill, J., Renault, P., Rescigno, M., Sanchez, N., Sunagawa, S., Torrejon, A., Turner, K., Vandemeulebrouck, G., Varela, E., Winogradsky, Y., Zeller, G., Weissenbach, J., Ehrlich, S. D. & Bork, P. (2011). Enterotypes of the human gut microbiome. *Nature* 473(7346), 174–180.

Asara, J. M., Christofk, H. R., Freimark, L. M. & Cantley, L. C. (2008). A label-free quantification method by MS/MS TIC compared to SILAC and spectral counting in a proteomics screen. *PROTEOMICS* 8(5), 994–999.

Asker, N., Axelsson, M. A., Olofsson, S. O. & Hansson, G. C. (1998). Dimerization of the human MUC2 mucin in the endoplasmic reticulum is followed by a N-glycosylation-dependent transfer of the mono- and dimers to the Golgi apparatus. *The Journal of biological chemistry* 273(30), 18857–18863.

Ayabe, T., Satchell, D. P., Wilson, C. L., Parks, W. C., Selsted, M. E. & Ouellette, A. J. (2000). Secretion of microbicidal alpha-defensins by intestinal Paneth cells in response to bacteria. *Nature Immunology* 1(2), 113–118.

Bantscheff, M., Boesche, M., Eberhard, D., Matthieson, T., Sweetman, G. & Kuster, B. (2008). Robust and sensitive iTRAQ quantification on an LTQ Orbitrap mass spectrometer. *Molecular & Cellular Proteomics* 7(9), 1702–1713.

Bäckström, M., Ambort, D., Thomsson, E., Johansson, M. E. V. & Hansson, G. C. (2013). Increased Understanding of the Biochemistry and Biosynthesis of MUC2 and Other Gel-Forming Mucins Through the Recombinant Expression of Their Protein Domains. *Molecular Biotechnology* 54(2), 250–256.

Bell, S. L., Xu, G., Khatri, I. A., Wang, R., Rahman, S. & Forstner, J. F. (2003). N-linked

oligosaccharides play a role in disulphide-dependent dimerization of intestinal mucin Muc2. *The Biochemical journal* 373(Pt 3), 893–900.

Bennett, E. P., Mandel, U., Clausen, H., Gerken, T. A., Fritz, T. A. & Tabak, L. A. (2012). Control of mucin-type O-glycosylation: A classification of the polypeptide GalNAc-transferase gene family. *Glycobiology* 22(6), 736–756.

Bergstrom, K. S. B., Kissoon-Singh, V., Gibson, D. L., Ma, C., Montero, M., Sham, H. P., Ryz, N., Huang, T., Velcich, A., Finlay, B. B., Chadee, K. & Vallance, B. A. (2010). Muc2 protects against lethal infectious colitis by disassociating pathogenic and commensal bacteria from the colonic mucosa. *PLoS pathogens* 6(5), e1000902.

Bertaccini, D., Vaca, S., Carapito, C., Arsène-Ploetze, F., Van Dorsselaer, A. & Schaeffer-Reiss, C. (2013). An improved stable isotope N-terminal labeling approach with light/heavy TMPP to automate proteogenomics data validation: dN-TOP. *Journal of Proteome Research* 12(6), 3063–3070.

Bevins, C. L. & Salzman, N. H. (2011). Paneth cells, antimicrobial peptides and maintenance of intestinal homeostasis. *Nature Reviews Microbiology* 9(5), 356–368.

Branca, R. M. M., Orre, L. M., Johansson, H. J., Granholm, V., Huss, M., Pérez-Bercoff, Å., Forshed, J., Käll, L. & Lehtiö, J. (2013). HiRIEF LC-MS enables deep proteome coverage and unbiased proteogenomics. *Nature Methods* 11(1), 59–62.

Carlstedt, I., Herrmann, A., Karlsson, H., Sheehan, J., Fransson, L. A. & Hansson, G. C. (1993). Characterization of two different glycosylated domains from the insoluble mucin complex of rat small intestine. *Journal of Biological Chemistry* 268(25), 18771–18781.

Carlstedt, I., Lindgren, H., Sheehan, J. K., Ulmsten, U. & Wingerup, L. (1983). Isolation and characterization of human cervical-mucus glycoproteins. *The Biochemical journal* 211, 13–22.

Chelius, D. & Bondarenko, P. V. (2002). Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry. *Journal of Proteome Research* 1(4), 317–323.

Chi, A., Huttenhower, C., Geer, L. Y., Coon, J. J., Syka, J. E. P., Bai, D. L., Shabanowitz, J., Burke, D. J., Troyanskaya, O. G. & Hunt, D. F. (2007). Analysis of phosphorylation sites on proteins from Saccharomyces cerevisiae by electron transfer dissociation (ETD) mass spectrometry. *Proceedings of the National Academy of Sciences* 104(7), 2193–2198.

Choi, H. & Nesvizhskii, A. I. (2008). False Discovery Rates and Related Statistical Concepts in Mass Spectrometry-Based Proteomics. *Journal of Proteome Research* 7(1), 47–50.

Choi, S., Jeong, J., Na, S., Lee, H. S., Kim, H.-Y., Lee, K.-J. & Paek, E. (2010). New Algorithm for the Identification of Intact Disulfide Linkages Based on Fragmentation Characteristics in Tandem Mass Spectra. *Journal of Proteome Research* 9(1), 626–635.

Clowers, B. H., Dodds, E. D., Seipert, R. R. & Lebrilla, C. B. (2007). Site Determination of Protein Glycosylation Based on Digestion with Immobilized Nonspecific Proteases and Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. *Journal of Proteome Research* 6(10), 4032–4040.

Corfield, A. P., Myerscough, N., Bradfield, N., Corfield, C. D. A., Gough, M., Clamp, J. R., Durdey, P., Warren, B. F., Bartolo, D. C., King, K. R. & Williams, J. M. (1996). Colonic mucins in ulcerative colitis: evidence for loss of sulfation. *Glycoconjugate journal* 13(5), 809–822.

Costenoble, R., Picotti, P., Reiter, L., Stallmach, R., Heinemann, M., Sauer, U. & Aebersold, R. (2011). Comprehensive quantitative analysis of central carbon and amino-acid metabolism in Saccharomyces cerevisiae under multiple conditions by targeted proteomics. *Molecular Systems Biology* 7, 464.

Cox, J. & Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature*

Biotechnology 26(12), 1367–1372.
Cox, J., Neuhauser, N., Michalski, A., Scheltema, R. A., Olsen, J. V. & Mann, M. (2011). Andromeda: a peptide search engine integrated into the MaxQuant environment. *Journal of Proteome Research* 10(4), 1794–1805.
Craig, R. & Beavis, R. C. (2004). TANDEM: matching proteins with tandem mass spectra. *Bioinformatics* 20(9), 1466–1467.
Danese, S. & Fiocchi, C. (2011). Ulcerative Colitis. *New England Journal of Medicine* 365(18), 1713–1725.
Dang, L. T., Purvis, A. R., Huang, R.-H., Westfield, L. A. & Sadler, J. E. (2011). Phylogenetic and functional analysis of histidine residues essential for pH-dependent multimerization of von Willebrand factor. *Journal of Biological Chemistry* 286(29), 25763–25769.
Dayon, L., Pasquarello, C., Hoogland, C., Sanchez, J.-C. & Scherl, A. (2010). Combining low- and high-energy tandem mass spectra for optimized peptide quantification with isobaric tags. *Journal of Proteomics* 73(4), 769–777.
de Hoffmann, E. & Stroobant, V. (2013). *Mass Spectrometry*. John Wiley & Sons. ISBN 1118681940.
Derrien, M., Vaughan, E. E., Plugge, C. M. & De Vos, W. M. (2004). Akkermansia muciniphila gen. nov., sp. nov., a human intestinal mucin-degrading bacterium. *International journal of systematic and evolutionary microbiology* 54(Pt 5), 1469–1476.
Di Palma, S., Boersema, P., Heck, A. & Mohammed, S. (2011). Zwitterionic Hydrophilic Interaction Liquid Chromatography (ZIC-HILIC and ZIC-cHILIC) Provide High Resolution Separation and Increase Sensitivity in Proteome Analysis. *Analytical Chemistry* 83(9), 3440–3447.
Doucet, A. & Overall, C. M. (2008). Protease proteomics: revealing protease in vivo functions using systems biology approaches. *Molecular aspects of medicine* 29(5), 339–358.
Eckburg, P. B., Bik, E. M., Bernstein, C. N., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S. R., Nelson, K. E. & Relman, D. A. (2005). Diversity of the human intestinal microbial flora. *Science* 308(5728), 1635–1638.
Eng, J. K., McCormack, A. L. & Yates, J. R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of The American Society for Mass Spectrometry* 5(11), 976–989.
Engholm-Keller, K., Hansen, T. A., Palmisano, G. & Larsen, M. R. (2011). Multidimensional strategy for sensitive phosphoproteomics incorporating protein prefractionation combined with SIMAC, HILIC, and TiO(2) chromatography applied to proximal EGF signaling. *Journal of Proteome Research* 10(12), 5383–5397.
Ermund, A., Schütte, A., Johansson, M. E., Gustafsson, J. K. & Hansson, G. C. (2013). Studies of mucus in mouse stomach, small intestine, and colon. I. Gastrointestinal mucus layers have different properties depending on location as well as over the Peyer's patches. *American Journal of Physiology-Gastrointestinal and Liver Physiology* 305(5), G341–G347
Eyers, C. E., Lawless, C., Wedge, D. C., Lau, K. W., Gaskell, S. J. & Hubbard, S. J. (2011). CONSeQuence: prediction of reference peptides for absolute quantitative proteomics using consensus machine learning approaches. *Molecular & Cellular Proteomics* 10(11), M110.003384.
Farkas, K., Yeruva, S., Rakonczay, Z., Jr, Ludolph, L., Molnár, T., Nagy, F., Szepes, Z., Schnúr, A., Wittmann, T., Hubricht, J., Riederer, B., Venglovecz, V., Lázár, G., Király, M., Zsembery, Á., Varga, G., Seidler, U. & Hegyi, P. (2011). New therapeutic targets in ulcerative colitis: The importance of ion transporters in the human colon. *Inflammatory Bowel Diseases* 17(4), 884–898.
Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. (1989). Electrospray

ionization for mass spectrometry of large biomolecules. *Science* 246(4926), 64–71.
Fu, J., Wei, B., Wen, T., Johansson, M. E. V., Liu, X., Bradford, E., Thomsson, K. A., Mcgee, S., Mansour, L., Tong, M., Mcdaniel, J. M., Sferra, T. J., Turner, J. R., Chen, H., Hansson, G. C., Braun, J. & Xia, L. (2011). Loss of intestinal core 1–derived O-glycans causes spontaneous colitis in mice. *Journal of Clinical Investigation* 121(4), 1657–1666.
Fujiki, Y., Hubbard, A. L., Fowler, S. & Lazarow, P. B. (1982). Isolation of intracellular membranes by means of sodium carbonate treatment: application to endoplasmic reticulum. *The Journal of Cell Biology* 93(1), 97–102.
Fukuda, M. (2001). Beta-elimination for release of O-GalNAc-linked oligosaccharides from glycoproteins and glycopeptides. *Current protocols in molecular biology,* Chapter 17.
Gallien, S., Perrodou, E., Carapito, C., Deshayes, C., Reyrat, J.-M., Van Dorsselaer, A., Poch, O., Schaeffer, C. & Lecompte, O. (2009). Ortho-proteogenomics: multiple proteomes investigation through orthology and a new MS-based protocol. *Genome Research* 19(1), 128–135.
Geer, L. Y., Markey, S. P., Kowalak, J. A., Wagner, L., Xu, M., Maynard, D. M., Yang, X., Shi, W. & Bryant, S. H. (2004). Open Mass Spectrometry Search Algorithm. *Journal of Proteome Research* 3(5), 958–964.
Geiger, T., Cox, J. & Mann, M. (2010a). Proteomics on an Orbitrap benchtop mass spectrometer using all-ion fragmentation. *Molecular & Cellular Proteomics* 9(10), 2252–2261.
Geiger, T., Cox, J., Ostasiewicz, P., Wiśniewski, J. R. & Mann, M. (2010b). Super-SILAC mix for quantitative proteomics of human tumor tissue. *Nature Methods* 7(5), 383–385.
Gerber, S. A., Rush, J., Stemman, O., Kirschner, M. W. & Gygi, S. P. (2003). Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS. *Proceedings of the National Academy of Sciences* 100(12), 6940–6945.
Gillet, L. C., Navarro, P., Tate, S., Röst, H., Selevsek, N., Reiter, L., Bonner, R. & Aebersold, R. (2012). Targeted data extraction of the MS/MS spectra generated by data-independent acquisition: a new concept for consistent and accurate proteome analysis. *Molecular & Cellular Proteomics* 11(6), O111.016717.
Godl, K., Johansson, M. E. V., Lidell, M. E., Mörgelin, M., Karlsson, H., Olson, F. J., Gum, J. R., Kim, Y. S. & Hansson, G. C. (2002). The N terminus of the MUC2 mucin forms trimers that are held together within a trypsin-resistant core fragment. *The Journal of Biological Chemistry* 277(49), 47248–47256.
Gorman, J. J., Wallis, T. P. & Pitt, J. J. (2002). Protein disulfide bond determination by mass spectrometry. *Mass Spectrometry Reviews* 21(3), 183–216.
Gum, J. R., Byrd, J. C., Hicks, J. W., Toribara, N. W., Lamport, D. T. & Kim, Y. S. (1989). Molecular cloning of human intestinal mucin cDNAs. Sequence analysis and evidence for genetic polymorphism. *The Journal of biological chemistry* 264(11), 6480–6487.
Guo, Y., Nguyen, K.-A. & Potempa, J. (2010). Dichotomy of gingipains action as virulence factors: from cleaving substrates with the precision of a surgeon's knife to a meat chopper-like brutal degradation of proteins. *Periodontology 2000* 54(1), 15–44.
Gustafsson, J. K., Ermund, A., Ambort, D., Johansson, M. E. V., Nilsson, H. E., Thorell, K., Hebert, H., Sjövall, H. & Hansson, G. C. (2012a). Bicarbonate and functional CFTR channel are required for proper mucin secretion and link cystic fibrosis with its mucus phenotype. *Journal of Experimental Medicine* 209(7), 1263–1272.
Gustafsson, J. K., Ermund, A., Johansson, M. E. V., Schütte, A., Hansson, G. C. & Sjövall, H. (2012b). An ex vivo method for studying mucus formation, properties, and thickness in human colonic biopsies and mouse small and large intestinal explants. *AJP: Gastrointestinal and Liver Physiology* 302(4), G430–G438.
Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H. & Aebersold, R. (1999). Quantitative

analysis of complex protein mixtures using isotope-coded affinity tags. *Nature Biotechnology* 17(10), 994–999.

Hapfelmeier, S., Lawson, M. A. E., Slack, E., Kirundi, J. K., Stoel, M., Heikenwalder, M., Cahenzli, J., Velykoredko, Y., Balmer, M. L., Endt, K., Geuking, M. B., Curtiss, R., McCoy, K. D. & Macpherson, A. J. (2010). Reversible microbial colonization of germ-free mice reveals the dynamics of IgA immune responses. *Science (New York, NY)* 328(5986), 1705–1709.

Haslam, S. M., North, S. J. & Dell, A. (2006). Mass spectrometric analysis of N- and O-glycosylation of tissues and cells. *Current Opinion in Structural Biology* 16(5), 584–591.

Hattrup, C. L. & Gendler, S. J. (2008). Structure and function of the cell surface (tethered) mucins. *Annual Review of Physiology* 70, 431–457.

Havlis, J. & Shevchenko, A. (2004). Absolute Quantification of Proteins in Solutions and in Polyacrylamide Gels by Mass Spectrometry. *Analytical Chemistry* 76(11), 3029–3036.

Hägglund, P., Bunkenborg, J., Elortza, F., Jensen, O. N. & Roepstorff, P. (2004). A New Strategy for Identification of N-Glycosylated Proteins and Unambiguous Assignment of Their Glycosylation Sites Using HILIC Enrichment and Partial Deglycosylation. *Journal of Proteome Research* 3(3), 556–566.

Heazlewood, C. K., Cook, M. C., Eri, R., Price, G. R., Tauro, S. B., Taupin, D., Thornton, D. J., Png, C. W., Crockford, T. L., Cornall, R. J., Adams, R., Kato, M., Nelms, K. A., Hong, N. A., Florin, T. H. J., Goodnow, C. C. & McGuckin, M. A. (2008). Aberrant mucin assembly in mice causes endoplasmic reticulum stress and spontaneous inflammation resembling ulcerative colitis. *PLoS medicine* 5(3), e54.

Herbert, D. R., Yang, J.-Q., Hogan, S. P., Groschwitz, K., Khodoun, M., Munitz, A., Orekov, T., Perkins, C., Wang, Q., Brombacher, F., Urban, J. F., Rothenberg, M. E. & Finkelman, F. D. (2009). Intestinal epithelial cell secretion of RELM-beta protects against gastrointestinal worm infection. *Journal of Experimental Medicine* 206(13), 2947–2957.

Holmen Larsson, J. M., Karlsson, H., Sjovall, H. & Hansson, G. C. (2009). A complex, but uniform O-glycosylation of the human MUC2 mucin from colonic biopsies analyzed by nanoLC/MSn. *Glycobiology* 19(7), 756–766.

Hooper, L. V. & Gordon, J. I. (2001). Glycans as legislators of host-microbial interactions: spanning the spectrum from symbiosis to pathogenicity. *Glycobiology* 11(2), 1R–10R.

Hsu, J.-L., Huang, S.-Y., Shiea, J.-T., Huang, W.-Y. & Chen, S.-H. (2005). Beyond quantitative proteomics: signal enhancement of the a1 ion as a mass tag for peptide sequencing using dimethyl labeling. *Journal of Proteome Research* 4(1), 101–108.

Hu, Q., Noll, R. J., Li, H., Makarov, A., Hardman, M. & Graham Cooks, R. (2005). The Orbitrap: a new mass spectrometer. *Journal of Mass Spectrometry* 40(4), 430–443.

Huang, R.-H., Wang, Y., Roth, R., Yu, X., Purvis, A. R., Heuser, J. E., Egelman, E. H. & Sadler, J. E. (2008). Assembly of Weibel-Palade body-like tubules from N-terminal domains of von Willebrand factor. *Proceedings of the National Academy of Sciences* 105(2), 482–487.

Huang, Z. H., Shen, T., Wu, J., Gage, D. A. & Watson, J. T. (1999). Protein sequencing by matrix-assisted laser desorption ionization-postsource decay-mass spectrometry analysis of the N-Tris(2,4,6-trimethoxyphenyl)phosphine-acetylated tryptic digests. *Analytical Biochemistry* 268(2), 305–317.

Huang, Z. H., Wu, J., Roth, K. D., Yang, Y., Gage, D. A. & Watson, J. T. (1997). A picomole-scale method for charge derivatization of peptides for sequence analysis by mass spectrometry. *Analytical Chemistry* 69(2), 137–144.

Huddleston, M. J., Bean, M. F. & Carr, S. A. (1993). Collisional fragmentation of glycopeptides by electrospray ionization LC/MS and LC/MS/MS: methods for selective detection of glycopeptides in protein digests. *Analytical Chemistry* 65(7), 877–884.

Jensen, P. H., Karlsson, N. G., Kolarich, D. & Packer, N. H. (2012). Structural analysis of N- and

O-glycans releasedfrom glycoproteins. *Nature Protocols* 7(7), 1299–1310 Nature Publishing Group.

Jensen, P. H., Kolarich, D. & Packer, N. H. (2010). Mucin-type O-glycosylation - putting the pieces together. *FEBS Journal* 277(1), 81–94.

Johansen, F.-E. & Kaetzel, C. S. (2011). Regulation of the polymeric immunoglobulin receptor and IgA transport: new advances in environmental factors that stimulate pIgR expression and its role in mucosal immunity. *Mucosal Immunology* 4(6), 598–602.

Johansson, M. E. V., Ambort, D., Pelaseyed, T., Schütte, A., Gustafsson, J. K., Ermund, A., Subramani, D. B., Holmén-Larsson, J. M., Thomsson, K. A., Bergström, J. H., van der Post, S., Rodriguez-Piñeiro, A. M., Sjövall, H., Bäckström, M. & Hansson, G. C. (2011). Composition and functional role of the mucus layers in the intestine. *Cellular and molecular life sciences* 68(22), 3635–3641.

Johansson, M. E. V., Gustafsson, J. K., Holmén-Larsson, J., Jabbar, K. S., Xia, L., Xu, H., Ghishan, F. K., Carvalho, F. A., Gewirtz, A. T., Sjövall, H. & Hansson, G. C. (2014). Bacteria penetrate the normally impenetrable inner colon mucus layer in both murine colitis models and patients with ulcerative colitis. *Gut* 63(2), 281–291.

Johansson, M. E. V., Phillipson, M., Petersson, J., Velcich, A., Holm, L. & Hansson, G. C. (2008). The inner of the two Muc2 mucin-dependent mucus layers in colon is devoid of bacteria. *Proceedings of the National Academy of Sciences* 105(39), 15064–15069.

Johansson, M. E. V., Thomsson, K. A. & Hansson, G. C. (2009). Proteomic Analyses of the Two Mucus Layers of the Colon Barrier Reveal That Their Main Component, the Muc2 Mucin, Is Strongly Bound to the Fcgbp Protein. *Journal of Proteome Research* 8(7), 3549–3557.

Johnson, R. S., Martin, S. A. & Biemann, K. (1988). Collision-induced fragmentation of (M + H)+ ions of peptides. Side chain specific sequence ions. *International Journal of Mass Spectrometry and Ion Processes* 86, 137–154.

Kaczmarczyk, A., Thuveson, M. & Fries, E. (2002). Intracellular coupling of the heavy chain of pre-alpha-inhibitor to chondroitin sulfate. *The Journal of Biological Chemistry* 277(16), 13578–13582.

Karam, S. M. (1999). Lineage commitment and maturation of epithelial cells in the gut. *Frontiers in bioscience* 4, D286–98.

Karas, M. & Hillenkamp, F. (1988). Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical Chemistry* 60(20), 2299–2301.

Karlsson, N. G., Herrmann, A., Karlsson, H., Johansson, M. E., Carlstedt, I. & Hansson, G. C. (1997). The glycosylation of rat intestinal Muc2 mucin varies between rat strains and the small and large intestine. A study of O-linked oligosaccharides by a mass spectrometric approach. *The Journal of Biological Chemistry* 272(43), 27025–27034.

Käll, L., Storey, J. D., MacCoss, M. J. & Noble, W. S. (2008). Assigning significance to peptides identified by tandem mass spectrometry using decoy databases. *Journal of Proteome Research* 7(1), 29–34.

Kebarle, P. (2000). A brief overview of the present status of the mechanisms involved in electrospray mass spectrometry. *Journal of mass spectrometry : JMS* 35(7), 804–817.

Keller, A., Nesvizhskii, A. I., Kolker, E. & Aebersold, R. (2002). Empirical Statistical Model To Estimate the Accuracy of Peptide Identifications Made by MS/MS and Database Search. *Analytical Chemistry* 74(20), 5383–5392.

Khan, Z., Bloom, J. S., Garcia, B. A., Singh, M. & Kruglyak, L. (2009). Protein quantification across hundreds of experimental conditions. *Proceedings of the National Academy of Sciences* 106(37), 15544–15548.

Khor, B., Gardet, A. & Xavier, R. J. (2011). Genetics and pathogenesis of inflammatory bowel disease. *Nature* 474(7351), 307–317.

Kim, M.-S., Pinto, S. M., Getnet, D., Nirujogi, R. S., Manda, S. S., Chaerkady, R., Madugundu, A. K., Kelkar, D. S., Isserlin, R., Jain, S., Thomas, J. K., Muthusamy, B., Leal-Rojas, P., Kumar, P., Sahasrabuddhe, N. A., Balakrishnan, L., Advani, J., George, B., Renuse, S., Selvan, L. D. N., Patil, A. H., Nanjappa, V., Radhakrishnan, A., Prasad, S., Subbannayya, T., Raju, R., Kumar, M., Sreenivasamurthy, S. K., Marimuthu, A., Sathe, G. J., Chavan, S., Datta, K. K., Subbannayya, Y., Sahu, A., Yelamanchi, S. D., Jayaram, S., Rajagopalan, P., Sharma, J., Murthy, K. R., Syed, N., Goel, R., Khan, A. A., Ahmad, S., Dey, G., Mudgal, K., Chatterjee, A., Huang, T.-C., Zhong, J., Wu, X., Shaw, P. G., Freed, D., Zahari, M. S., Mukherjee, K. K., Shankar, S., Mahadevan, A., Lam, H., Mitchell, C. J., Shankar, S. K., Satishchandra, P., Schroeder, J. T., Sirdeshmukh, R., Maitra, A., Leach, S. D., Drake, C. G., Halushka, M. K., Prasad, T. S. K., Hruban, R. H., Kerr, C. L., Bader, G. D., Iacobuzio-Donahue, C. A., Gowda, H. & Pandey, A. (2014). A draft map of the human proteome. *Nature* 509(7502), 575–581.

Kleifeld, O., Doucet, A., auf dem Keller, U., Prudova, A., Schilling, O., Kainthan, R. K., Starr, A. E., Foster, L. J., Kizhakkedathu, J. N. & Overall, C. M. (2010). Isotopic labeling of terminal amines in complex samples identifies protein N-termini and protease cleavage products. *Nature Biotechnology* 28(3), 281–288.

Kobayashi, K., Ogata, H., Morikawa, M., Iijima, S., Harada, N., Yoshida, T., Brown, W. R., Inoue, N., Hamada, Y., Ishii, H., Watanabe, M. & Hibi, T. (2002). Distribution and partial characterisation of IgG Fc binding protein in various mucin producing cells and body fluids. *Gut* 51(2), 169–176.

Kolarich, D., Jensen, P. H., Altmann, F. & Packer, N. H. (2012). Determination of site-specific glycan heterogeneity on glycoproteins. *Nature Protocols* 7(7), 1285–1298.

Komiya, T., Tanigawa, Y. & Hirohashi, S. (1999). Cloning and Identification of the Gene Gob-5, Which Is Expressed in Intestinal Goblet Cells in Mice. *Biochemical and biophysical research communications* 255(2), 347–351.

Köcher, T., Pichler, P., Schutzbier, M., Stingl, C., Kaul, A., Teucher, N., Hasenfuss, G., Penninger, J. M. & Mechtler, K. (2009). High Precision Quantitative Proteomics Using iTRAQ on an LTQ Orbitrap: A New Mass Spectrometric Method Combining the Benefits of All. *Journal of Proteome Research* 8(10), 4743–4752.

Krüger, M., Moser, M., Ussar, S., Thievessen, I., Luber, C. A., Forner, F., Schmidt, S., Zanivan, S., Fässler, R. & Mann, M. (2008). SILAC mouse for quantitative proteomics uncovers kindlin-3 as an essential factor for red blood cell function. *Cell* 134(2), 353–364.

Kryuchkov, F., Verano-Braga, T., Hansen, T. A., Sprenger, R. R. & Kjeldsen, F. (2013). Deconvolution of mixture spectra and increased throughput of peptide identification by utilization of intensified complementary ions formed in tandem mass spectrometry. *Journal of Proteome Research* 12(7), 3362–3371.

Kumar, C. & Mann, M. (2009). Bioinformatics analysis of mass spectrometry-based proteomics data sets. *FEBS Letters* 583(11), 1703–1712.

Kumar, P., Luo, Q., Vickers, T. J., Sheikh, A., Lewis, W. G., Fleckenstein, J. M. & Payne, S. M. (2014). EatA, an Immunogenic Protective Antigen of Enterotoxigenic Escherichia coli, Degrades Intestinal Mucin. *Infection and Immunity* 82(2), 500–508.

LaPointe, L. C., Dunne, R., Brown, G. S., Worthley, D. L., Molloy, P. L., Wattchow, D. & Young, G. P. (2008). Map of differential transcript expression in the normal human large intestine. *Physiological Genomics* 33(1), 50–64.

Larsson, J. M. H., Karlsson, H., Crespo, J. G., Johansson, M. E. V., Eklund, L., Sjövall, H. & Hansson, G. C. (2011). Altered O-glycosylation profile of MUC2 mucin occurs in active ulcerative colitis and is associated with increased inflammation. *Inflammatory Bowel Diseases* 17(11), 2299–2307.

Lau, E., Lam, M. P. Y., Siu, S. O., Kong, R. P. W., Chan, W. L., Zhou, Z., Huang, J., Lo, C. & Chu,

I. K. (2011). Combinatorial use of offline SCX and online RP-RP liquid chromatography for iTRAQ-based quantitative proteomics applications. *Molecular BioSystems* 7(5), 1399–1408.

Lewis, J. D., Chuai, S., Nessel, L., Lichtenstein, G. R., Aberra, F. N. & Ellenberg, J. H. (2008). Use of the noninvasive components of the mayo score to assess clinical response in Ulcerative Colitis. *Inflammatory Bowel Diseases* 14(12), 1660–1666.

Lidell, M. E., Johansson, M. E. V. & Hansson, G. C. (2003a). An autocatalytic cleavage in the C terminus of the human MUC2 mucin occurs at the low pH of the late secretory pathway. *The Journal of Biological Chemistry* 278(16), 13944–13951.

Lidell, M. E., Johansson, M. E. V., Mörgelin, M., Asker, N., Gum, J. R., Kim, Y. S. & Hansson, G. C. (2003b). The recombinant C-terminus of the human MUC2 mucin forms dimers in Chinese-hamster ovary cells and heterodimers with full-length MUC2 in LS 174T cells. *The Biochemical journal* 372(Pt 2), 335–345.

Lidell, M. E., Moncada, D. M., Chadee, K. & Hansson, G. C. (2006). Entamoeba histolytica cysteine proteases cleave the MUC2 mucin in its C-terminal domain and dissolve the protective colonic mucus gel. *Proceedings of the National Academy of Sciences* 103(24), 9298–9303.

Link, A. J., Eng, J., Schieltz, D. M., Carmack, E., Mize, G. J., Morris, D. R., Garvik, B. M. & Yates, J. R. (1999). Direct analysis of protein complexes using mass spectrometry. *Nature Biotechnology* 17(7), 676–682.

Liu, F., van Breukelen, B. & Heck, A. J. R. (2014). Facilitating protein disulfide mapping by a combination of pepsin digestion, electron transfer higher energy dissociation (EThcD), and a dedicated search algorithm SlinkS. *Molecular & Cellular Proteomics* 13(10), 2776–2786.

Liu, H., Finch, J. W., Lavallee, M. J., Collamati, R. A., Benevides, C. C. & Gebler, J. C. (2007). Effects of column length, particle size, gradient length and flow rate on peak capacity of nano-scale liquid chromatography for peptide separations. *Journal of Chromatography A* 1147(1), 30–36.

Liu, H., Sadygov, R. G. & Yates, J. R. (2004). A Model for Random Sampling and Estimation of Relative Protein Abundance in Shotgun Proteomics. *Analytical Chemistry* 76(14), 4193–4201.

Loomes, K. M., Senior, H. E., West, P. M. & Roberton, A. M. (1999). Functional protective role for mucin glycosylated repetitive domains. *European journal of biochemistry* 266(1), 105–111.

Löffler, A., Doucey, M. A., Jansson, A. M., Müller, D. R., de Beer, T., Hess, D., Meldal, M., Richter, W. J., Vliegenthart, J. F. & Hofsteenge, J. (1996). Spectroscopic and protein chemical analyses demonstrate the presence of C-mannosylated tryptophan in intact human RNase 2 and its isoforms. *Biochemistry* 35(37), 12005–12014.

Luckey, T. D. (1972). Introduction to intestinal microecology. *The American journal of clinical nutrition* 25(12), 1292–1294.

Lundgren, D. H., Hwang, S.-I., Wu, L. & Han, D. K. (2010). Role of spectral counting in quantitative proteomics. *Expert review of proteomics* 7(1), 39–53.

Luo, Q., Shen, Y., Hixson, K. K., Zhao, R., Yang, F., Moore, R. J., Mottaz, H. M. & Smith, R. D. (2005). Preparation of 20-μm-i.d. Silica-Based Monolithic Columns and Their Performance for Proteomics Analyses. *Analytical Chemistry* 77(15), 5028–5035.

MacCoss, M. J., Wu, C. C. & Yates, J. R. (2002). Probability-based validation of protein identifications using a modified SEQUEST algorithm. *Analytical Chemistry* 74(21), 5593–5599.

Makarov, A. (2000). Electrostatic axially harmonic orbital trapping: a high-performance technique of mass analysis. *Analytical Chemistry* 72(6), 1156–1162.

Mank, M., Stahl, B. & Boehm, G. (2004). 2,5-Dihydroxybenzoic acid butylamine and other ionic liquid matrixes for enhanced MALDI-MS analysis of biomolecules. *Analytical Chemistry* 76(10), 2938–2950.

Mann, M. & Jensen, O. N. (2003). Proteomic analysis of post-translational modifications. *Nature Biotechnology* 21(3), 255–261.

Mann, M. & Wilm, M. (1994). Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Analytical Chemistry* 66(24), 4390–4399.

Manza, L. L., Stamer, S. L., Ham, A.-J. L., Codreanu, S. G. & Liebler, D. C. (2005). Sample preparation and digestion for proteomic analyses using spin filters. *PROTEOMICS* 5(7), 1742–1745.

Marcus, K., Schäfer, H., Klaus, S., Bunse, C., Swart, R. & Meyer, H. E. (2007). A new fast method for nanoLC-MALDI-TOF/TOF-MS analysis using monolithic columns for peptide preconcentration and separation in proteomic studies. *Journal of Proteome Research* 6(2), 636–643.

Martin, S. E., Shabanowitz, J., Hunt, D. F. & Marto, J. A. (2000). Subfemtomole MS and MS/MS peptide sequence analysis using nano-HPLC micro-ESI fourier transform ion cyclotron resonance mass spectrometry. *Analytical Chemistry* 72(18), 4266–4274.

McIlwain, S., Tamura, K., Kertesz-Farkas, A., Grant, C. E., Diament, B., Frewen, B., Howbert, J. J., Hoopmann, M. R., Käll, L., Eng, J. K., MacCoss, M. J. & Noble, W. S. (2014). Crux: rapid open source protein tandem mass spectrometry analysis. *Journal of Proteome Research* 13(10), 4488–4491.

Meiring, H., Van der Heeft, E., Hove, Ten, G. & De Jong, A. (2002). Nanoscale LC-MS (n): technical design and applications to peptide and protein analysis. *Journal of Separation Science* 25(9), 557–568.

Mikesh, L. M., Ueberheide, B., Chi, A., Coon, J. J., Syka, J. E. P., Shabanowitz, J. & Hunt, D. F. (2006). The utility of ETD mass spectrometry in proteomic analysis. *Biochimica et biophysica acta* 1764(12), 1811–1822.

Nakshabendi, I. M., McKee, R., Downie, S., Russell, R. I. & Rennie, M. J. (1999). Rates of small intestinal mucosal protein synthesis in human jejunum and ileum. *The American journal of physiology* 277(6 Pt 1), E1028–31.

Neilson, K. A., Ali, N. A., Muralidharan, S., Mirzaei, M., Mariani, M., Assadourian, G., Lee, A., van Sluyter, S. C. & Haynes, P. A. (2011). Less label, more free: approaches in label-free quantitative mass spectrometry. *PROTEOMICS* 11(4), 535–553.

Nesvizhskii, A. I. & Aebersold, R. (2004). Analysis, statistical validation and dissemination of large-scale proteomics datasets generated by tandem MS. *Drug discovery today* 9(4), 173–181.

Neverova, I. & Van Eyk, J. E. (2005). Role of chromatographic techniques in proteomic analysis. *Journal of Chromatography B* 815(1-2), 51–63.

Nilsson, J., Rüetschi, U., Halim, A., Hesse, C., Carlsohn, E., Brinkmalm, G. & Larson, G. (2009). Enrichment of glycopeptides for glycan structure and attachment site identification. *Nature Methods* 6(11), 809–811.

Olsen, J. V., Macek, B., Lange, O., Makarov, A., Horning, S. & Mann, M. (2007). Higher-energy C-trap dissociation for peptide modification analysis. *Nature Methods* 4(9), 709–712.

Olsen, J. V., Ong, S.-E. & Mann, M. (2004). Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Molecular & Cellular Proteomics* 3(6), 608–614.

Ong, S.-E. & Mann, M. (2005). Mass spectrometry–based proteomics turns quantitative. *Nature Chemical Biology* 1(5), 252–262.

Ong, S.-E., Blagoev, B., Kratchmarova, I., Kristensen, D. B., Steen, H., Pandey, A. & Mann, M. (2002). Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics. *Molecular & Cellular Proteomics* 1(5), 376–386.

Park, C. Y., Klammer, A. A., Käll, L., MacCoss, M. J. & Noble, W. S. (2008). Rapid and accurate peptide identification from tandem mass spectra. *Journal of Proteome Research* 7(7), 3022–3027.

Patel, S. D., Rajala, M. W., Rossetti, L., Scherer, P. E. & Shapiro, L. (2004). Disulfide-dependent multimeric assembly of resistin family hormones. *Science* 304(5674), 1154–1158.

Pelaseyed, T., Zäch, M., Petersson, A. C., Svensson, F., Johansson, D. G. A. & Hansson, G. C. (2013). Unfolding dynamics of the mucin SEA domain probed by force spectroscopy suggest that it acts as a cell-protective device. *The FEBS journal* 280(6), 1491–1501.

Peng, J., Elias, J. E., Thoreen, C. C., Licklider, L. J. & Gygi, S. P. (2003). Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome. *Journal of Proteome Research* 2(1), 43–50.

Perez-Vilar, J. & Hill, R. L. (1999). The structure and assembly of secreted mucins. *The Journal of Biological Chemistry* 274(45), 31751–31754.

Perez-Vilar, J., Randell, S. H. & Boucher, R. C. (2004). C-Mannosylation of MUC5AC and MUC5B Cys subdomains. *Glycobiology* 14(4), 325–337.

Perkins, D. N., Pappin, D. J., Creasy, D. M. & Cottrell, J. S. (1999). Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20(18), 3551–3567.

Picotti, P., Bodenmiller, B., Mueller, L. N., Domon, B. & Aebersold, R. (2009). Full dynamic range proteome analysis of S. cerevisiae by targeted proteomics. *Cell* 138(4), 795–806.

Poulsen, J. W., Madsen, C. T., Young, C., Poulsen, F. M. & Nielsen, M. L. (2013). Using guanidine-hydrochloride for fast and efficient protein digestion and single-step affinity-purification mass spectrometry. *Journal of Proteome Research* 12(2), 1020–1030.

Proc, J. L., Kuzyk, M. A., Hardie, D. B., Yang, J., Smith, D. S., Jackson, A. M., Parker, C. E. & Borchers, C. H. (2010). A Quantitative Study of the Effects of Chaotropic Agents, Surfactants, and Solvents on the Digestion Efficiency of Human Plasma Proteins by Trypsin. *Journal of Proteome Research* 9(10), 5422–5437.

Puente, X. S., Sánchez, L. M., Overall, C. M. & López-Otín, C. (2003). Human and mouse proteases: a comparative genomic approach. *Nature Reviews Genetics* 4(7), 544–558

Pullan, R. D., Thomas, G. A., Rhodes, M., Newcombe, R. G., Williams, G. T., Allen, A. & Rhodes, J. (1994). Thickness of adherent mucus gel on colonic mucosa in humans and its relevance to colitis. *Gut* 35(3), 353–359.

Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F., Yamada, T., Mende, D. R., Li, J., Xu, J., Li, S., Li, D., Cao, J., Wang, B., Liang, H., Zheng, H., Xie, Y., Tap, J., Lepage, P., Bertalan, M., Batto, J.-M., Hansen, T., Le Paslier, D., Linneberg, A., Nielsen, H. B., Pelletier, E., Renault, P., Sicheritz-Ponten, T., Turner, K., Zhu, H., Yu, C., Li, S., Jian, M., Zhou, Y., Li, Y., Zhang, X., Li, S., Qin, N., Yang, H., Wang, J., Brunak, S., Doré, J., Guarner, F., Kristiansen, K., Pedersen, O., Parkhill, J., Weissenbach, J., Antolin, M., Artiguenave, F., Blottiere, H., Borruel, N., Bruls, T., Casellas, F., Chervaux, C., Cultrone, A., Delorme, C., Denariaz, G., Dervyn, R., Forte, M., Friss, C., Van De Guchte, M., Guedon, E., Haimet, F., Jamet, A., Juste, C., Kaci, G., Kleerebezem, M., Knol, J., Kristensen, M., Layec, S., Le Roux, K., Leclerc, M., Maguin, E., Melo Minardi, R., Oozeer, R., Rescigno, M., Sanchez, N., Tims, S., Torrejon, T., Varela, E., De Vos, W., Winogradsky, Y., Zoetendal, E., Bork, P., Ehrlich, S. D. & Wang, J. (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464(7285), 59–65.

Quenzer, T. L., Emmett, M. R., Hendrickson, C. L., Kelly, P. H. & Marshall, A. G. (2001). High sensitivity Fourier transform ion cyclotron resonance mass spectrometry for biological analysis with nano-LC and microelectrospray ionization. *Analytical Chemistry* 73(8), 1721–1725.

Ridley, C., Kouvatsos, N., Raynal, B. D., Howard, M., Collins, R. F., Desseyn, J.-L., Jowitt, T. A., Baldock, C., Davis, C. W., Hardingham, T. E. & Thornton, D. J. (2014). Assembly of the

respiratory mucin MUC5B: a new model for a gel-forming mucin. *Journal of Biological Chemistry* 289(23), 16409–16420.

Robbe, C., Capon, C., Coddeville, B. & Michalski, J. (2004). Structural diversity and specific distribution of O-glycans in normal human mucins along the intestinal tract. *The Biochemical journal* 384(Pt 2), 307–316.

Robbe, C., Capon, C., Maes, E., Rousset, M., Zweibaum, A., Zanetta, J. & Michalski, J. (2003). Evidence of regio-specific glycosylation in human intestinal mucins. *Journal of Biological Chemistry* 278(47), 46337.

Roepstorff, P. & Fohlman, J. (1984). Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomedical mass spectrometry* 11(11), 601.

Rose, R. J., Damoc, E., Denisov, E., Makarov, A. & Heck, A. J. R. (2012). High-sensitivity Orbitrap mass analysis of intact macromolecular assemblies. *Nature Methods* 9(11), 1084–1086.

Ross, P. L., Huang, Y. N., Marchese, J. N., Williamson, B., Parker, K., Hattan, S., Khainovski, N., Pillai, S., Dey, S., Daniels, S., Purkayastha, S., Juhasz, P., Martin, S., Bartlet-Jones, M., He, F., Jacobson, A. & Pappin, D. J. (2004). Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents. *Molecular & Cellular Proteomics*, 3(12), 1154–1169.

Rousseau, K., Kirkham, S., Johnson, L., Fitzpatrick, B., Howard, M., Adams, E. J., Rogers, D. F., Knight, D., Clegg, P. & Thornton, D. J. (2008). Proteomic analysis of polymeric salivary mucins: no evidence for MUC19 in human saliva. *The Biochemical journal* 413(3), 545.

Ruiz-Perez, F. & Nataro, J. P. (2014). Bacterial serine proteases secreted by the autotransporter pathway: classification, specificity, and role in virulence. *Cellular and molecular life sciences* 71(5), 745–770.

Ryle, J. P. & Sanger, F. (1955). Disulphide interchange reactions. *The Biochemical journal* 60(4), 535–540.

Sadler, J. E. (1998). Biochemistry and genetics of von Willebrand factor. *Annual Review of Biochemistry* 67, 395–424.

Sadygov, R. G., Cociorva, D. & Yates, J. R. (2004). Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. *Nature Methods* 1(3), 195–202.

Saitoh, O., Kojima, K., Sugi, K., Matsuse, R., Uchida, K., Tabata, K., Nakagawa, K., Kayazawa, M., Hirata, I. & Katsu, K. (1999). Fecal eosinophil granule-derived proteins reflect disease activity in inflammatory bowel disease. *The American journal of Gastroenterology* 94(12), 3513–3520.

Sandle, G. I. (1998). Salt and water absorption in the human colon: a modern appraisal. *Gut* 43(2), 294–299.

Sartor, R. B. (2008). Microbial influences in inflammatory bowel diseases. *Gastroenterology* 134(2), 577–594.

Schmidt, A., Kellermann, J. & Lottspeich, F. (2005). A novel strategy for quantitative proteomics using isotope-coded protein labels. *PROTEOMICS* 5(1), 4–15.

Schwanhäusser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W. & Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature* 473(7347), 337–342.

Shen, Y., Zhao, R., Belov, M. E., Conrads, T. P., Anderson, G. A., Tang, K., Pasa-Tolić, L., Veenstra, T. D., Lipton, M. S., Udseth, H. R. & Smith, R. D. (2001). Packed capillary reversed-phase liquid chromatography with high-performance electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry for proteomics. *Analytical Chemistry* 73(8), 1766–1775.

Shevchenko, A., Tomas, H., Havlis, J., Olsen, J. V. & Mann, M. (2006). In-gel digestion for mass spectrometric characterization of proteins and proteomes. *Nature Protocols* 1(6), 2856–2860.

Shevchenko, A., Wilm, M., Vorm, O. & Mann, M. (1996). Mass spectrometric sequencing of

proteins silver-stained polyacrylamide gels. *Analytical Chemistry* 68(5), 850–858.
Singh, A. K., Xia, W., Riederer, B., Juric, M., Li, J., Zheng, W., Cinar, A., Xiao, F., Bachmann, O., Song, P., Praetorius, J., Aalkjaer, C. & Seidler, U. (2013). Essential role of the electroneutral Na +HCO 3−cotransporter NBCn1 in murine duodenal acid/base balance and colonic mucus layer build-up in vivo. *The Journal of Physiology* 591(Pt 8), 2189–2204.
Singh, P., Panchaud, A. & Goodlett, D. R. (2010). Chemical Cross-Linking and Mass Spectrometry As a Low-Resolution Protein Structure Determination Technique. *Analytical Chemistry* 82(7), 2636–2642.
Smits, A. H., Jansen, P. W. T. C., Poser, I., Hyman, A. A. & Vermeulen, M. (2012). Stoichiometry of chromatin-associated protein complexes revealed by label-free quantitative mass spectrometry-based proteomics. *Nucleic Acids Research* 41(1), e28–e28.
Staubach, F., Künzel, S., Baines, A. C., Yee, A., McGee, B. M., Bäckhed, F., Baines, J. F. & Johnsen, J. M. (2012). Expression of the blood-group-related glycosyltransferase B4galnt2 influences the intestinal microbiota in mice. *The ISME journal* 6(7), 1345–1355.
Steen, H. & Mann, M. (2004). The ABC"s (and XYZ"s) of peptide sequencing. *Nature reviews Molecular cell biology* 5(9), 699–711.
Steentoft, C., Vakhrushev, S. Y., Joshi, H. J., Kong, Y., Vester-Christensen, M. B., Schjoldager, K. T.-B. G., Lavrsen, K., Dabelsteen, S., Pedersen, N. B., Marcos-Silva, L., Gupta, R., Bennett, E. P., Mandel, U., Brunak, S., Wandall, H. H., Levery, S. B. & Clausen, H. (2013). Precision mapping of the human O-GalNAc glycoproteome through SimpleCell technology. *The EMBO Journal* 32(10), 1478–1488.
Steentoft, C., Vakhrushev, S. Y., Vester-Christensen, M. B., Schjoldager, K. T.-B. G., Kong, Y., Bennett, E. P., Mandel, U., Wandall, H., Levery, S. B. & Clausen, H. (2011). Mining the O-glycoproteome using zinc-finger nuclease-glycoengineered SimpleCell lines. *Nature Methods* 8(11), 977–982.
Sury, M. D., Chen, J.-X. & Selbach, M. (2010). The SILAC fly allows for accurate protein quantification in vivo. *Molecular & Cellular Proteomics* 9(10), 2173–2183.
Swaney, D. L., Wenger, C. D. & Coon, J. J. (2010). Value of using multiple proteases for large-scale mass spectrometry-based proteomics. *Journal of Proteome Research* 9(3), 1323–1329.
Syka, J. E. P., Coon, J. J., Schroeder, M. J., Shabanowitz, J. & Hunt, D. F. (2004). Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry. *Proceedings of the National Academy of Sciences* 101(26), 9528–9533.
Tateno, H., Yabe, R., Sato, T., Shibazaki, A., Shikanai, T., Gonoi, T., Narimatsu, H. & Hirabayashi, J. (2012). Human ZG16p recognizes pathogenic fungi through non-self polyvalent mannose in the digestive system. *Glycobiology* 22(2), 210–220.
Taylor, J. A. & Johnson, R. S. (1997). Sequence database searches via de novo peptide sequencing by tandem mass spectrometry. *Rapid communications in mass spectrometry* 11(9), 1067–1075.
Thakur, S. S., Geiger, T., Chatterjee, B., Bandilla, P., Fröhlich, F., Cox, J. & Mann, M. (2011). Deep and highly sensitive proteome coverage by LC-MS/MS without prefractionation. *Molecular & Cellular Proteomics* 10(8), M110.003699.
Thompson, A., Schäfer, J., Kuhn, K., Kienle, S., Schwarz, J., Schmidt, G., Neumann, T., Johnstone, R., Mohammed, A. K. A. & Hamon, C. (2003). Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS. *Analytical Chemistry* 75(8), 1895–1904.
Toribara, N. W., Gum, J. R., Culhane, P. J., Lagace, R. E., Hicks, J. W., Petersen, G. M. & Kim, Y. S. (1991). MUC-2 human small intestinal mucin gene structure. Repeated arrays and polymorphism. *Journal of Clinical Investigation* 88(3), 1005–1013.
Tran, J. C., Zamdborg, L., Ahlf, D. R., Lee, J. E., Catherman, A. D., Durbin, K. R., Tipton, J. D., Vellaichamy, A., Kellie, J. F., Li, M., Wu, C., Sweet, S. M. M., Early, B. P., Siuti, N., LeDuc, R.

D., Compton, P. D., Thomas, P. M. & Kelleher, N. L. (2011). Mapping intact protein isoforms in discovery mode using top-down proteomics. *Nature* 480(7376), 254–258 Nature Publishing Group.

Turano, C., Coppari, S., Altieri, F. & Ferraro, A. (2002). Proteins of the PDI family: Unpredicted non-ER locations and functions. *Journal of Cellular Physiology* 193(2), 154–163.

Tysk, C., Lindberg, E., Järnerot, G. & Flodérus-Myrhed, B. (1988). Ulcerative colitis and Crohn's disease in an unselected population of monozygotic and dizygotic twins. A study of heritability and the influence of smoking. *Gut* 29(7), 990–996.

Uhlén, M., Björling, E., Agaton, C., Szigyarto, C. A.-K., Amini, B., Andersen, E., Andersson, A.-C., Angelidou, P., Asplund, A., Asplund, C., Berglund, L., Bergström, K., Brumer, H., Cerjan, D., Ekström, M., Elobeid, A., Eriksson, C., Fagerberg, L., Falk, R., Fall, J., Forsberg, M., Björklund, M. G., Gumbel, K., Halimi, A., Hallin, I., Hamsten, C., Hansson, M., Hedhammar, M., Hercules, G., Kampf, C., Larsson, K., Lindskog, M., Lodewyckx, W., Lund, J., Lundeberg, J., Magnusson, K., Malm, E., Nilsson, P., Odling, J., Oksvold, P., Olsson, I., Oster, E., Ottosson, J., Paavilainen, L., Persson, A., Rimini, R., Rockberg, J., Runeson, M., Sivertsson, A., Sköllermo, A., Steen, J., Stenvall, M., Sterky, F., Strömberg, S., Sundberg, M., Tegel, H., Tourle, S., Wahlund, E., Waldén, A., Wan, J., Wernérus, H., Westberg, J., Wester, K., Wrethagen, U., Xu, L. L., Hober, S. & Pontén, F. (2005). A human protein atlas for normal and cancer tissues based on antibody proteomics. *Molecular & Cellular Proteomics* 4(12), 1920–1932.

van der Flier, L. G. & Clevers, H. (2009). Stem cells, self-renewal, and differentiation in the intestinal epithelium. *Annual Review of Physiology* 71, 241–260.

van der Post, S., Subramani, D. B., Bäckström, M., Johansson, M. E. V., Vester-Christensen, M. B., Mandel, U., Bennett, E. P., Clausen, H., Dahlén, G., Sroka, A., Potempa, J. & Hansson, G. C. (2013). Site-specific O-glycosylation on the MUC2 mucin protein inhibits cleavage by the Porphyromonas gingivalis secreted cysteine protease (RgpB). *The Journal of Biological Chemistry* 288(20), 14636–14646.

Van der Sluis, M., De Koning, B. A. E., De Bruijn, A. C. J. M., Velcich, A., Meijerink, J. P. P., Van Goudoever, J. B., Büller, H. A., Dekker, J., Van Seuningen, I., Renes, I. B. & Einerhand, A. W. C. (2006). Muc2-deficient mice spontaneously develop colitis, indicating that MUC2 is critical for colonic protection. *Gastroenterology* 131(1), 117–129.

Venable, J. D., Dong, M.-Q., Wohlschlegel, J., Dillin, A. & Yates, J. R. (2004). Automated approach for quantitative analysis of complex peptide mixtures from tandem mass spectra. *Nature Methods* 1(1), 39–45.

Verdugo, P., Aitken, M., Langley, L. & Villalon, M. J. (1991). Molecular mechanism of product storage and release in mucin secretion. II. The role of extracellular Ca++. *The American review of respiratory disease* 144(3 Pt 2), 625–633.

Vischer, U. M. & Wagner, D. D. (1994). von Willebrand factor proteolytic processing and multimerization precede the formation of Weibel-Palade bodies. *Blood* 83(12), 3536–3544.

Vuckovic, D., Dagley, L. F., Purcell, A. W. & Emili, A. (2013). Membrane proteomics by high performance liquid chromatography-tandem mass spectrometry: Analytical approaches and challenges. *PROTEOMICS* 13(3-4), 404–423.

Waanders, L. F., Hanke, S. & Mann, M. (2007). Top-down quantitation and characterization of SILAC-labeled proteins. *Journal of The American Society for Mass Spectrometry* 18(11), 2058–2064.

Wang, H. & Hanash, S. M. (2003). Multi-dimensional liquid phase based separations in proteomics. *Journal of chromatography B, Analytical technologies in the biomedical and life sciences* 787(1), 11–18.

Washburn, M. P., Wolters, D. & Yates, J. R. (2001). Large-scale analysis of the yeast proteome by

multidimensional protein identification technology. *Nature Biotechnology* 19(3), 242–247.
Wiesner, J., Premsler, T. & Sickmann, A. (2008). Application of electron transfer dissociation (ETD) for the analysis of posttranslational modifications. *PROTEOMICS* 8(21), 4466–4483.
Wilhelm, M., Schlegl, J., Hahne, H., Moghaddas Gholami, A., Lieberenz, M., Savitski, M. M., Ziegler, E., Butzmann, L., Gessulat, S., Marx, H., Mathieson, T., Lemeer, S., Schnatbaum, K., Reimer, U., Wenschuh, H., Mollenhauer, M., Slotta-Huspenina, J., Boese, J.-H., Bantscheff, M., Gerstmair, A., Faerber, F. & Kuster, B. (2014). Mass-spectrometry-based draft of the human proteome. *Nature* 509(7502), 582–587.
Wilm, M. (2011). Principles of electrospray ionization. *Molecular & Cellular Proteomics* 10(7), M111.009407.
Wiśniewski, J. R. & Mann, M. (2012). Consecutive proteolytic digestion in an enzyme reactor increases depth of proteomic and phosphoproteomic analysis. *Analytical Chemistry* 84(6), 2631–2637.
Wiśniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. (2009). Universal sample preparation method for proteome analysis. *Nature Methods* 6(5), 359–362.
Wu, C. C., MacCoss, M. J., Howell, K. E. & Yates, J. R. (2003). A method for the comprehensive proteomic analysis of membrane proteins. *Nature Biotechnology* 21(5), 532–538.
Wu, Y., Wang, F., Liu, Z., Qin, H., Song, C., Huang, J., Bian, Y., Wei, X., Dong, J. & Zou, H. (2014). Five-plex isotope dimethyl labeling for quantitative proteomics. *Chemical Communications* 50(14), 1708.
Wuhrer, M., Catalina, M. I., Deelder, A. M. & Hokke, C. H. (2007). Glycoproteomics based on tandem mass spectrometry of glycopeptides. *Journal of chromatography B, Analytical technologies in the biomedical and life sciences* 849(1), 115–128.
Wuhrer, M., de Boer, A. R. & Deelder, A. M. (2009). Structural glycomics using hydrophilic interaction chromatography (HILIC) with mass spectrometry. *Mass Spectrometry Reviews* 28(2), 192–206.
Xu, P., Duong, D. M. & Peng, J. (2009). Systematical Optimization of Reverse-Phase Chromatography for Shotgun Proteomics. *Journal of Proteome Research* 8(8), 3944–3950.
Yang, B., Wu, Y.-J., Zhu, M., Fan, S.-B., Lin, J., Zhang, K., Li, S., Chi, H., Li, Y.-X., Chen, H.-F., Luo, S.-K., Ding, Y.-H., Wang, L.-H., Hao, Z., Xiu, L.-Y., Chen, S., Ye, K., He, S.-M. & Dong, M.-Q. (2012). Identification of cross-linked peptides from complex samples. *Nature Methods* 9(9), 904–906.
Yurtsever, Z., Sala-Rabanal, M., Randolph, D. T., Scheaffer, S. M., Roswit, W. T., Alevy, Y. G., Patel, A. C., Heier, R. F., Romero, A. G., Nichols, C. G., Holtzman, M. J. & Brett, T. J. (2012). Self-cleavage of human CLCA1 protein by a novel internal metalloprotease domain controls calcium-activated chloride channel activation. *Journal of Biological Chemistry* 287(50), 42138–42149.
Zhang, Y., Wen, Z., Washburn, M. P. & Florens, L. (2009). Effect of Dynamic Exclusion Duration on Spectral Count Based Quantitative Proteomics. *Analytical Chemistry* 81(15), 6317–6326.
Zielinska, D. F., Gnad, F., Wisniewski, J. R. & Mann, M. (2010). Precision mapping of an in vivo N-glycoproteome reveals rigid topological and sequence constraints. *Cell*. 141(5), 897–907.
Zoetendal, E. G., Wright, von, A., Vilpponen-Salmela, T., Ben-Amor, K., Akkermans, A. D. L. & de Vos, W. M. (2002). Mucosa-Associated Bacteria in the Human Gastrointestinal Tract Are Uniformly Distributed along the Colon and Differ from the Community Recovered from Feces. *Applied and environmental microbiology* 68(7), 3401–3407.
Zubarev, R. A., Kelleher, N. L. & McLafferty, F. W. (1998). Electron Capture Dissociation of Multiply Charged Protein Cations. A Nonergodic Process. *Journal of the American Chemical Society* 120, 3265–3266.