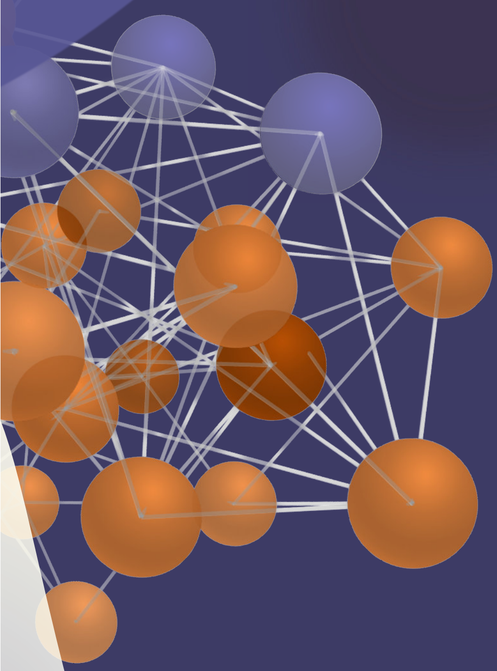
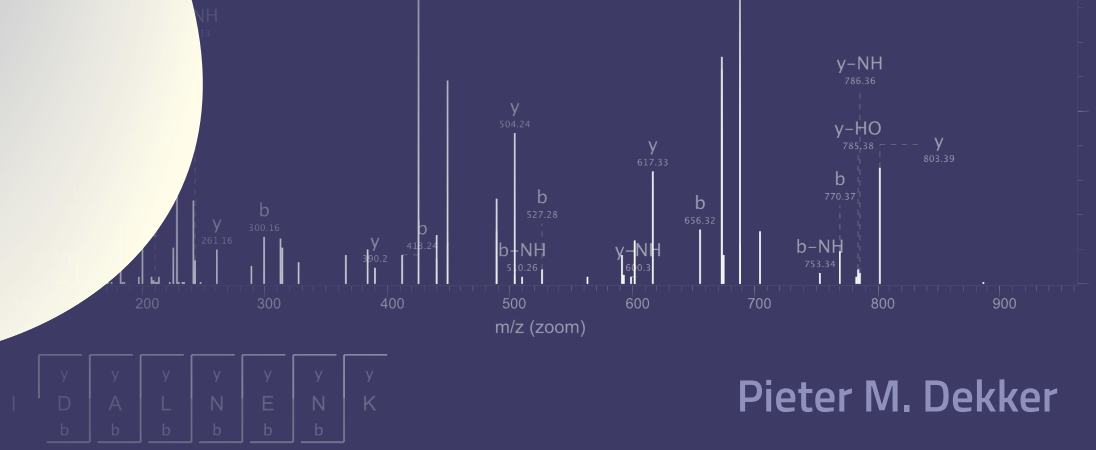


Human milk proteome and peptidome: variation, interaction, and relation with allergy



Pieter M. Dekker

Propositions

1. The human milk proteome contains more information than the abundances of the individual proteins.
   (this thesis)

2. The abundance of proteins in milk is primarily determined by their origin.
   (this thesis)

3. The literacy and concentration skills of future generations will require a drastic change in the system of scientific publishing.

4. The experience of wonder, awe, and beauty in one's field of research, helps the scientist persevere.

5. Stereotypes are incomplete stories.

6. Suppression of dissenting voices results in stronger polarization in society.

Propositions belonging to the thesis, entitled

Human milk proteome and peptidome: variation, interaction, and relation with allergy

Pieter M. Dekker
Wageningen, 13 September 2022

Human milk proteome and peptidome: variation, interaction, and relation with allergy

Pieter Matthias Dekker

Thesis committee

Promotor
Dr K.A. Hettinga
Associate professor, Food Quality and Design
Wageningen University & Research

Co-promotor
Dr E. Saccenti
Assistant professor, Laboratory of Systems and Synthetic Biology
Wageningen University & Research

Other members
Prof. Dr H.J. Wichers, Wageningen University & Research
Prof. Dr P. Zhou, Jiangnan University, Wuxi, China
Dr A. Smolinska, Maastricht University, Maastricht, The Netherlands
Dr E.A.F. van Tol, Netherlands Organization for Applied Scientific Research (TNO), Leiden, The Netherlands

This research was conducted under the auspices of the Graduate School VLAG (Advanced studies in Food Technology, Agrobiotechnology, Nutrition and Health Sciences).

Human milk proteome and peptidome: variation, interaction, and relation with allergy

Pieter M. Dekker

Thesis

submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Tuesday 13 September 2022
at 4 p.m. in the Omnia Auditorium.

Pieter M. Dekker
Human milk proteome and peptidome: variation, interaction, and relation with allergy,
206 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2022)
With references, with summary in English and Dutch

ISBN: 978-94-6447-238-7
DOI: 10.18174/570511

Contents

Chapter 1

Introduction

1.1 Human milk

Human milk is tailor-made nutrition for the vulnerable, newborn human. The mixture of macronutrients, micronutrients, and bioactive components is particularly suitable for the infant's healthy development. One of the aspects that makes human milk complex is that its composition changes drastically throughout lactation. In the first 3-5 days postpartum, the milk can be referred to as colostrum and is low in volume and high in both oligosaccharides and proteins. After colostrum, the milk undergoes a drastic change in composition (transition milk). After around 30 days postpartum, the milk is called mature milk [1]. At this stage, the composition of the milk is relatively stable. This longitudinal change in composition adapts the milk to the infant's needs in terms of nutrition and the development of the immune system, preparing the infant for the environment it is exposed to. A specific example of this is that levels of secretory immunoglobulin A (sIgA), lactoferrin, soluble CD14 receptor (sCD14), and other immune proteins are significantly higher in colostrum when compared to mature milk [2]. Besides a change throughout lactation, there is also compositional variation in human milk within-feed, within a day (diurnal), and between right and left breast [3, 4]. Furthermore, the composition differs between individuals (interindividual). Evidence shows, for example, an influence of dietary patterns of the mother and infant factors such as their sex on milk composition [5, 6].

Despite all the variation, the overall composition of mature human milk can roughly be divided into water (88%) and macro- and micronutrients that make up the remaining 12% [7]. Macronutrients are lactose (6.7-7.0%), lipids (3.5-4.8%), proteins (0.8-1.1%), and human milk oligosaccharides (HMOs) (0.5-1.5%) [7]. A visualization of the relative contribution of the various constituents is shown in Figure 1.1. Micronutrients are the components found in lower concentrations, such as, vitamins, trace elements, peptides, amino acids, metabolites, and RNA [7].

Several epidemiological studies have indicated the health effects of breastmilk on the breastfed newborn. Examples of such effects are a reduced risk of necrotizing enterocolitis, sepsis, and respiratory tract infections for exclusively breastfed infants [8–11]. Other studies have tried to link specific compositional aspects or individual components of human milk to health benefits. Some of these studies show clear relations, as is the case for Autran et al. [12], showing the effect that specific HMOs can have on the risk of developing necrotizing enterocolitis. The knowledge of these health benefits has led to the recommendation of exclusive breastfeeding in the first six months of life by most health organizations [13]. In addition, this knowledge has also led to an increase in the number of donor human milk banks, where donated milk is collected and distributed to vulnerable, e.g., preterm delivered, infants [14].

There is a growing body of literature on the composition of human milk and its effects on the infant's healthy development, but much is still unknown. This thesis focuses on elucidating the protein and peptide profile, its interindividual variation, and its possible effects on the development of the infant's immune system. This introductory chapter describes fundamental concepts for a basic understanding of human milk proteins and peptides, their analysis, and possible role in allergy development.

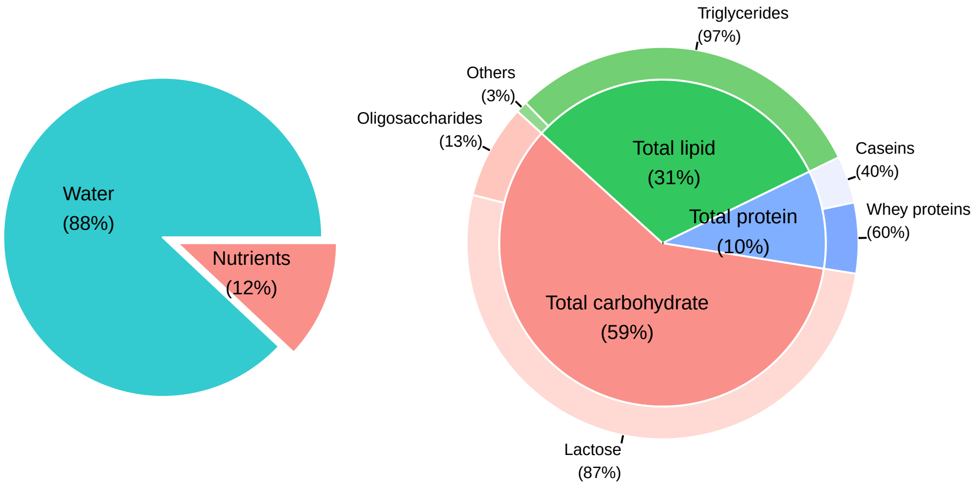


Figure 1.1: Overall composition of mature human milk (values are taken from Monaco et al. [7] and Donovan [15]).

1.2 Human milk proteins

Proteins provide the child with essential amino acids for growth and can have functions in, for example, immune response, digestion, and the transport of minerals. To date, 1500 proteins have been identified in human milk [16]. Despite the large diversity in proteins in milk, most of the total protein concentration is due to about a dozen proteins (Figure 1.2). We can distinguish 3 major classes of proteins, which are in order of relative contribution: whey proteins, caseins, and milk fat globule membrane (MFGM) proteins. This classification is based on differences in structure and characteristics of the proteins.

The whey proteins are the soluble fraction and contribute around 83% to the total

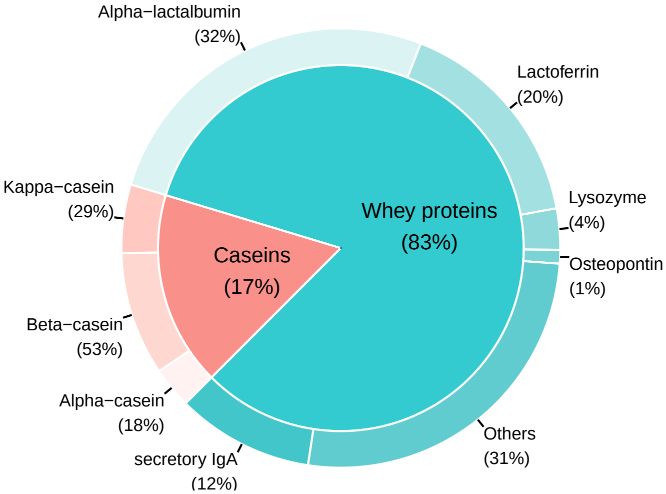


Figure 1.2: Relative contribution of major human milk proteins to the total protein content (values are taken from Donovan [15]) (IgA: immunoglobulin A).

protein content in mature milk [17]. The proteins that contribute the most to this class are α-lactalbumin, lactoferrin, and sIgA. Functions of the major whey proteins are related primarily to immune response and cell communication (see Table 1.1). These functions are also represented the most when all whey proteins are considered [18].

The caseins comprise β-casein, κ-casein, and α_{s1}-casein, and account for around 17% of the total protein concentration in mature milk [17]. However, it should be noted that the whey casein ratio shows considerable variation over lactation, ranging from 97:3 to 45:55 [17]. In addition, significant differences can be observed between studies due to the use of different analytical techniques or because the casein fraction is used instead of solely the casein subunits [17]. Caseins are organized into casein micelles, although in human milk, the structure of these micelles is less rigid than in bovine milk [19]. The casein micelles form a dispersed fraction that allows separation from the milk serum by a decrease in pH or ultracentrifugation. The functions of this group of proteins are related to metabolism and energy pathways, that is, nutrition and transport of calcium phosphate (Table 1.1).

MFGM proteins make up around 1-2% of the total protein content and comprise proteins that are part of the membrane that surrounds the lipid droplets (milk fat globule, MFG) in milk [20]. Among these proteins are well-known MFGM proteins, such as mucins, butyrophilin, butyrophilin like protein, and xanthine dehydrogenase/oxidase [21]. In addition, small amounts of various other nonmembrane proteins are also found in the MFGM protein fraction [22]. These nonmembrane proteins may associate with the membrane or be entrapped between the different membrane layers of the MFG [23]. Functions of MFGM proteins are diverse (Table 1.1), with the majority related to nutrition, immune response, and cell communication [22].

Table 1.1: Major human milk proteins with their functions (based on Donovan [15], Peterson et al. [24], and Qiu et al. [25]).

Protein class	Proteins	Main functions
Caseins	α_{s1}-casein	Nutrition
	β-casein	Nutrition, nutrient carrier
	κ-casein	Nutrition, host defense
Whey proteins	α-lactalbumin	Nutrition, nutrient carrier, host defense, prebiotic
	Lactoferrin	Nutrient carrier, intestinal development, host defense, prebiotic, cognition
	Secretory IgA	Host defense
	Osteopontin	Host defense
	Lysozyme	Host defense
Milk fat globule membrane proteins	Mucins	Antimicrobial
	Lactadherin	Host defense
	Butyrophilin	Fat secretion
	Bile salt activated lipase	Enzyme, barrier promoting

1.2.1 Protein synthesis

Proteins end up in the milk through different mechanisms, as visualized in Figure 1.3. Two major mechanisms are (1) proteins synthesized in the mammary epithelial cell (MEC), and (2) proteins originating from the blood.

The synthesis of proteins in the MECs starts with amino acids that originate from the blood. In the ribosomes of the rough endoplasmic reticulum (ER), the synthesis into proteins takes place, after which proteins are released into the ER lumen and can undergo post-translational modification (PTM) in the Golgi apparatus. From the Golgi lumina, proteins move in secretory vesicles to the apical plasma membrane, where they are released in the alveolar lumen (exocytosis, pathway A). Proteins such as caseins and α-lactalbumin are transported via this route. Other proteins, among which transmembrane proteins, associate with the MFGM and are secreted together with the MFG (pathway B) [23]. The MFGM consists of three membrane layers, of

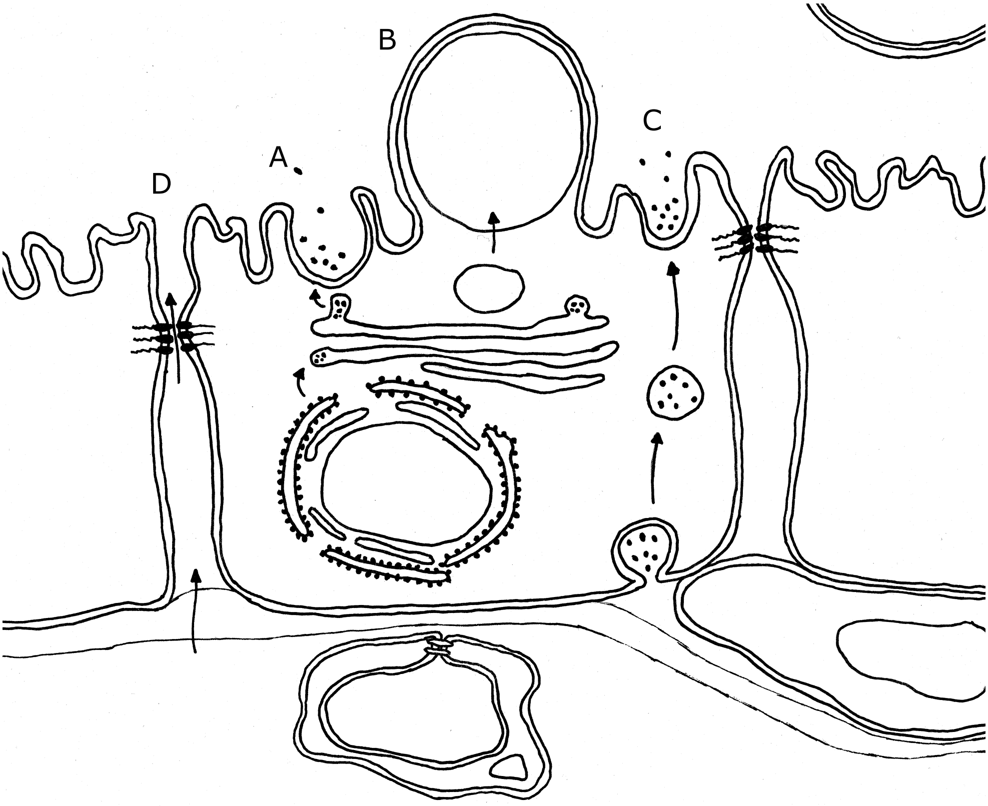


Figure 1.3: Schematic visualization of the mammary epithelial cell (MEC), secreting milk proteins via four pathways. Pathway A indicates the secretion of milk proteins synthesized in the MEC through secretory vesicles. Pathway B shows the formation and release of milk fat globules (MFGs) with three membrane layers containing milk fat globule membrane (MFGM) proteins. Pathway C indicates transcytosis of serum proteins through vesicles, and pathway D shows paracellular transport of serum proteins through disrupted tight junctions. Based on Vilotte et al. [23] and Boron et al. [26].

which the first is formed in the cytoplasm of the MEC. The other two membrane layers are added during secretion from the MECs, where MFGs are wrapped in the apical plasma membrane of the MEC [23].

Proteins that originate from blood serum are, for example, serum albumin, immunoglobulins and possibly dietary, non-human proteins (Section 1.4) [27, 28]. It is believed that these proteins enter the alveolar lumen through vesicular transport from basal to apical membrane, a pathway referred to as transcytosis (pathway C) [29]. Another mechanism for the transfer of these proteins is paracellularly (pathway D). This paracellular transfer might be enhanced when tight junctions are disrupted due to, for example, inflammation [30].

1.2.2 Factors affecting milk protein composition

As mentioned in Section 1.2.1, milk proteins result from protein synthesis in the mammary epithelium or transcytosis from the blood. Therefore, any factor that can influence cellular processes or harm the mammary epithelium, has the potential to affect the protein profile. This could be, for example, pathological conditions, energy balance, genetics, etc. Previous research has found several such factors. A significant factor is the time postpartum. As discussed before, human milk varies throughout lactation, which greatly influences the protein profile as well [31, 32]. Furthermore, gestational age has an effect, as it was found that in the first two weeks of lactation, milk for preterm delivered babies had higher concentrations of immune proteins IgA and lysozyme when compared with milk for term delivered babies [33]. The same study found an opposite effect for α-lactalbumin and β-casein, proteins important for nutrition [33].

Besides the effects of lactation stage and gestational age, a difference in protein profile was found between different ethnicities. In a study that compared milk from Chinese mothers with milk from Dutch mothers, 38 out of 166 proteins differed between these two groups, of which most differences were found in colostrum [34]. The effect of ethnicity was confirmed in a study by Zhang et al. [35], who found a difference between different ethnic groups and geographical locations within China.

The health condition of the mother is also affecting the milk proteome. Studies have shown, for example, that mothers with allergy or breast cancer have differences in levels of specific proteins [36, 37]. Zhu et al. [38] showed in a thorough study of two healthy individual mothers that protein profiles are significantly different between individuals. This emphasizes the importance of considering interindividual differences in future studies.

1.2.3 Proteolytic systems in human milk

A subgroup of the whey proteins are proteins that play a role in the breakdown of proteins. These groups of enzymes are also referred to as proteolytic systems and comprise proteases, protease activators, protease inhibitors, and protease activator inhibitors. Proteases are enzymes that can hydrolyze the peptide bond between two amino acids in a target amino acid sequence. These proteases can be activated by protease activators, but their activity can also be inhibited by protease inhibitors. In turn, the protease activators can be hindered in their action by protease activator inhibitors.

Proteases known to be present in human milk are plasmin, cathepsin D, elastase, trypsin, chymotrypsin, thrombin, kallikrein, cytosol aminopeptidase, carboxypeptidase B2, and matrix metalloproteinase [39]. A pivotal study carried out by Nielsen et al. [40] gave evidence that proteases are already active within the mammary gland. On the other end, it was hypothesized by Demers-Mathieu et al. [41] that proteases such as, for example, cathepsin D, can increase in activity upon autoactivation by the low pH in the infants' stomach.

Protease inhibitors or antiproteases can bind and inhibit proteases. The protease inhibitors can therefore affect the proteolytic activity. Some of the protease inhibitors that have been identified in human milk are: α_1-antitrypsin, α_1-antichymotrypsin, α_2-antiplasmin, α_2-macroglobulin, plasma serine protease inhibitor and antithrombin III [38].

Proteolytic systems occur in body tissues and bodily fluids and play essential roles in processes like development and host defense, but also in pathological conditions like inflammatory diseases [42]. However, their origin and function in human milk are not entirely understood, as proteolytic systems can find their origin in either blood, immune cell secretions, or synthesis within the MECs [43]. In addition, it is known that in blood, several of these proteases and protease inhibitors participate in the coagulation mechanisms [44].

All enzymes that are part of the proteolytic systems have a particular specificity, i.e., proteases can have a specificity of cleavage of target proteins at specific amino acids in the sequence, protease activators can have one specific or multiple proteases that it activates, etc. The result is a highly complex proteolytic machinery, and consequently, degradation of proteins in human milk.

1.3 Human milk peptides

The array of all protein fragments or peptides formed upon proteolytic degradation is referred to as the peptidome or degradome. Based on their amino acid sequence, individual peptides can often be linked to the protein they originate from, also referred to as the precursor protein. A broad definition of the peptidome would include any protein fragment, even if only one amino acid is removed from the C- or N-terminal of the protein. However, it is often impossible to distinguish between large protein fragments or intact proteins in the analysis of proteins and peptides. Therefore, a stricter definition of the peptidome is used by defining a range of peptide lengths.

Protein degradation in milk is mainly due to enzymatic proteolysis. To date, 4000 unique peptides have been identified in human milk [45], and this typically concerns peptides from only a selection of precursor proteins. The majority (85%) of the peptides originate from the precursor proteins β-casein, polymeric immunoglobulin receptor (PIGR), α_{s1}-casein, and osteopontin [46, 47]. This typical pattern can be linked to the high abundance of these proteins (see Figure 1.2), but also to the association of plasminogen with casein micelles, and susceptibility of these proteins or protein regions for proteolysis [47]. In addition, peptides from one precursor protein can overlap each other fully or partly, resulting in so-called “peptide ladders” (see Figure 1.4).

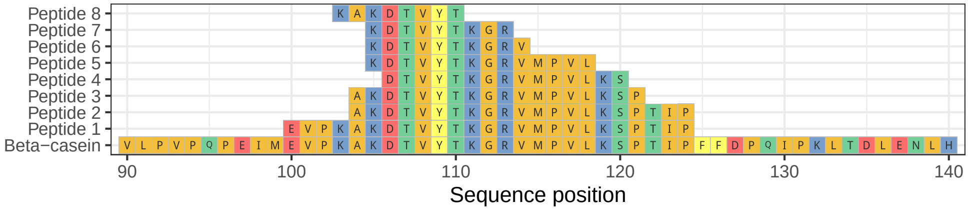


Figure 1.4: Typical peptide ladder from a selected region of the sequence of the human milk precursor protein β-casein (UniProt ID: P05184). Letters represent the different amino acids and background colors reflect their chemical characteristics: blue = positively charged, red = negatively charged, green = polar uncharged, yellow = aromatic with hydrophobic side chain, and orange = others.

It has been suggested that the proteolytic degradation of milk proteins might be a predigestion, supporting the infant’s underdeveloped digestive system. In addition, it is known that peptides can play crucial roles in biological systems. Some of the peptides have bioactive properties which might support the healthy development of the child. In a study by Nielsen et al., 306 out of 1100 identified human milk peptides were labeled as potentially bioactive [40]. Among these bioactive properties were, e.g., antimicrobial, cell-proliferation stimulating, and angiotensin-converting enzyme (ACE)-inhibitory effects [40]. Many have suggested that this can contribute positively to infant development and health. However, this often relies only on prediction of peptide bioactivity and little is known about the actual physiological significance of the peptides *in vivo*. A number of studies have begun to examine this physiological significance, and for several peptides it is now known that they can exert activity *in vivo*. One study showed, for example, the presence and activity of an antimicrobial cathelicidin peptide (LL-37) in human milk [48]. Another study showed that a group of β-casein peptides, also referred to as β-casomorphins and known for their opioid activity, promote β-cell development and could have a protective effect against the development of type 1 diabetes [49]. Immunomodulatory activity of a human milk peptide has been evidenced by Cai et al. [50], who showed that the BCCY-1 peptide from β-casein provided protection to infections in mice. These studies show the physiological significance of milk peptides for the infant's healthy development.

Whereas variation in the proteome is extensively studied, little is known about the factors influencing the peptidome. What determines the peptide profile in human milk is first of all the proteolytic activity (see Section 1.2.3). Furthermore, since the peptidome is a product of the proteome, levels of both proteins and peptides might be affected by the same factors. Several studies have found, for example, significant differences in the milk peptidome between mothers that delivered term or preterm [46, 51, 52], and differences throughout lactation [46, 53].

1.4 Non-human proteins and peptides in human milk

Evidence shows that human milk contains non-human proteins and peptides [54–56]. Among these are, for example, house dust mite allergen Der p 1 [57], wheat gliadin [58], hen's egg allergen ovomucoid [59], and bovine β-lactoglobulin (BLG) [60]. Most of these studies were carried out using assay-based immunochemical methods, in which antibodies can be prone to cross-reactivity with other proteins, resulting in possible false positive identification [61]. Nevertheless, some recent studies have confirmed the presence of these proteins and peptides with mass spectrometry (MS) techniques, which allow a direct analysis of a tryptic digest of the protein sequence [62, 63]. Zhu et al. [63] found with MS that the majority of the non-human proteins and peptides originate from the diet, especially from bovine milk products. Particular interest in these proteins and peptides was raised due to the allergenicity of several of them [64].

1.5 The burden of allergic diseases

An allergy or allergic disease can be defined as "clinical conditions associated with altered immunologic reactivity that may be either IgE mediated or non-IgE mediated" [65]. In chronic allergic diseases like atopic dermatitis (eczema), hay fever (allergic rhinitis), or allergic asthma, the immune system repeatedly responds to antigens and causes tissue damage and inflammation. Allergic diseases can bring dietary, social, and psychological burdens to the allergic individual. In the case of food allergy, for example, this can manifest itself through food anxiety or social limitations [66]. The prevalence of allergies is increasing worldwide, also bringing a heavy burden for the healthcare system. This can be seen in, e.g., the steep rise in hospital cases due to anaphylaxis, the most severe allergic reaction [67]. The overall costs of food allergy in the US have been estimated at $4184 per child, of which the most prominent part ($3457) is costs to families (e.g., allergen-free food products and lost labor productivity) [68]. To alleviate this burden and improve quality of life, there is an ongoing quest for ways to reduce the prevalence of allergic diseases, and therefore to prevent the onset of these diseases.

1.5.1 Allergy development

It is known that susceptibility to develop allergic diseases is in part determined by genetics [69]. This genetic part is believed to be multigenic, that is, specified by multiple genes. Bønnelykke et al. [70] combined evidence from allergy-related genome-wide association studies and studies of monogenic diseases, pointing to several potentially pathogenetic pathways for allergic diseases. Among these pathways are Th2 initiation/amplification, innate sensing, virus receptor, barrier integrity, signaling immune tolerance, mast cell responses, and interactions between T-cell receptors and major histocompatibility complexes [70]. Nevertheless, evidence shows that development of allergic diseases is often due to a combination of genetic susceptibility and environmental factors, as well as interactions between these two (e.g., through epigenetics) (see Figure 1.5) [69, 71]. Examples of the environmental factors are, nutrition, exposure to allergens, and factors such as pollution, which affect the function of the mucosal barrier [72].

The mucosal barrier is of crucial importance in allergy development since it is a vulnerable interface between the body and its environment. Consequently, an impaired functioning of the mucosal barrier can allow antigens to enter the systemic circulation [73]. In addition, the timing of exposure to the environmental factors has shown to be of crucial importance since, in the first months of life, the immune system and organ systems are in development, and both allergic sensitization and tolerance induction can take place [72]. So-called windows of opportunity for tolerance induction are therefore especially present in the first months of life.

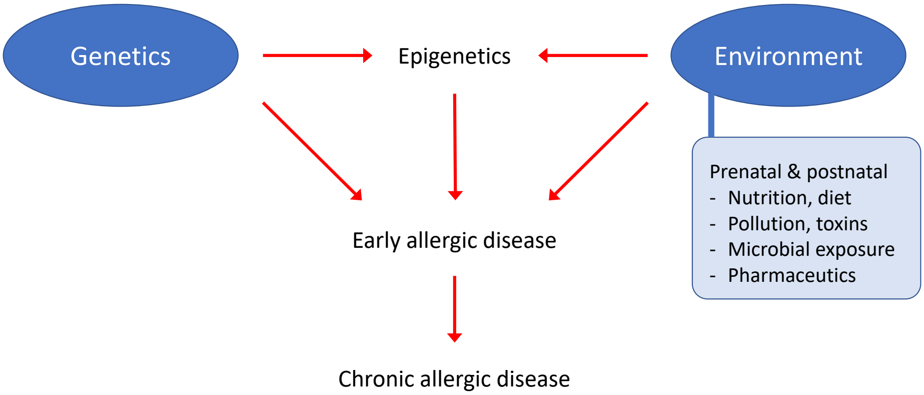


Figure 1.5: Schematic overview of factors that play a role in the development of allergic diseases (adapted from Meyers et al. [74]).

1.5.2 Human milk and allergy development

One of the major ways through which the newborn is exposed to the environment is through nutrition. Especially mucosae in the gastrointestinal tract are, through nutrition, exposed to substances that are unknown to the infant. Since human milk is considered the ideal nutrition for the newborn, research has investigated its effect on allergy development. Some authors describe a reduced risk of sensitization in exclusively breastfed infants [75–77], whereas others do not show significant effects [78], or even an increased risk in case the mother is asthmatic [79]. Considering these contradicting outcomes and the interindividual variation in human milk as discussed before, specific milk constituents could play a role in the effect on allergy development. This was also noted by Matheson et al. [80], who suggested further study of the compositional factors of human milk that could exert immunomodulatory effects. Compositional factors that have received little attention yet in relation to allergy development are proteins and peptides. There is a group of proteins, e.g., immunoglobulins and cytokines, that can play a role in the development of the immune system. In addition, there are peptides in human milk which have immune activity [50]. Human milk proteins and peptides could play different roles in the development of allergies in the breastfed infant (see Figure 1.6).

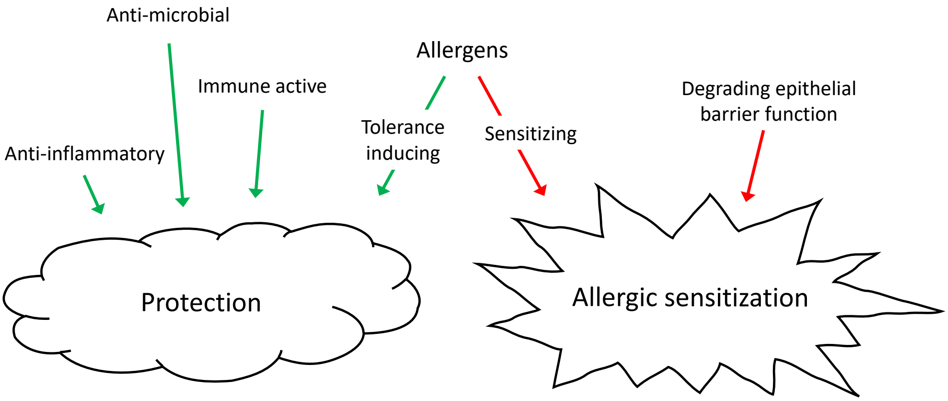


Figure 1.6: Schematic overview of potential effects of human milk proteins and peptides on the development of allergic diseases.

First, proteins such as lactoferrin and growth factors, may help in the development of the epithelial barrier in the intestines and therefore promote the barrier function [81, 82]. On the other hand, the presence of house dust mite allergen in milk might disrupt the epithelial barrier through its strong proteolytic activity [57]. Furthermore, both milk proteins and peptides can have antimicrobial activity and might provide protection through alteration of the gut microbiome [83]. Inflammation of the intestinal mucosae, might be prevented by bioactive proteins and peptides [84]. Examples of immune active proteins in human milk that may play a role are the cytokine interleukin (IL) 13 and transforming growth factor (TGF)β2, associated with risk for eczema [85], and eosinophil cationic protein and IgA, which have been associated with the occurrence of cow milk allergy [86, 87]. An immune response could also be triggered by exposure to antigens in the milk. This could either teach the immune system and prevent the development of allergy, or it could trigger an abnormal immune response and lead to sensitization. Recently, studies found that early introduction of allergens can induce immune tolerance [88, 89]. Nevertheless, this early introduction has limited effectivity and an earlier study showed that it can even lead to immune priming instead of tolerance induction [90]. Macchiaverni et al. [91] have suggested that early introduction of allergens via breastmilk might result in a more effective induction of tolerance. This is supported by a cohort study showing that infants consuming human milk with detectable egg allergen (ovalbumin) levels at 3 and 6 months postpartum, have a lower risk of developing egg allergy when compared to infants who receive milk with undetectable levels of egg allergen [92]. Furthermore, it was found in a study with mice, that protection of pups to bovine BLG sensitization was not provided only by provision of milk with BLG, but that immunization of the dams was necessary to show a strong protective effect [93]. This protective effect correlated with the levels of BLG-specific antibodies and BLG-immune complexes in the milk, suggesting that the level of these proteins and protein complexes in milk play an important role in inhibiting allergic sensitization in the infant [93].

To date, research into a possible effect of allergens in human milk on the development of tolerance or immune priming has mainly been carried out using assay-based techniques. These techniques have several limitations, among which are, (*i*) it is a targeted analysis, i.e., has a limited number of identifiable allergens, (*ii*) it is limited to the detection of a region of an allergen, and (*iii*) it is prone to cross-reactivity. From this, it results that there is a need for the confirmation of the presence of allergens in human milk using direct analysis of protein sequences with techniques such as MS. In addition, since little is known about the complete human milk proteome and the relation with the allergy status of both mother and child, research is needed to investigate this.

1.6 Proteomics and peptidomics analysis

For the analysis of proteins and peptides in complex biological samples like milk, MS is a key method. MS techniques measure the mass-to-charge (*m/z*) ratio of ionized analytes. Recent advances in this technology have led to a resolution that enables the analysis of hundreds of proteins and peptides in human milk simultaneously.

The MS technique used for proteomics analysis in this thesis is a so-called bottom-up or shotgun approach [94]. In this approach, the protein fraction of a sample is cleaned up, reduced, and alkylated. The proteins are then digested using trypsin, a protease with known cleavage specificity. The resulting tryptic peptides are then separated with liquid chromatography and analyzed with MS, which provides a data set with tryptic peptides as features. This data can be transformed into protein abundances using software that employs a database comprising the protein sequences expected to be present in the samples. It should be noted that only proteins present in this database can be identified, and it should therefore be as complete as possible [95]. Given the specificity of the protease used in the sample preparation, the database is subsequently (*in silico*) digested by the software. This results in a list with theoretical peptides, where each peptide is linked to theoretical spectra as well as fragmentation spectra. Experimental spectra can be matched with theoretical spectra, resulting in peptide identifications. The results can then be transformed into protein identifications using the uniqueness of the identified peptide sequences. A schematic overview of this proteomics approach is shown in Figure 1.7.

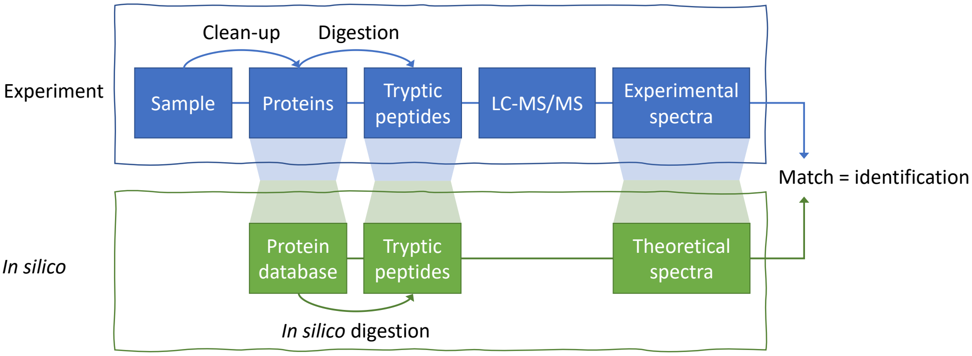


Figure 1.7: Simplified workflow of bottom-up proteomics analysis with mass spectrometry (MS), showing the steps taken in the experiment (top) and in the software-based processing of the resulting data (bottom).

A similar approach can be used for peptidomics analysis [96]. However, in the sample preparation of peptide samples, no protease is used, and thus, the *in silico* digestion of the database is set to “unspecific” cleavage. This will digest the protein sequences in the database *in silico* into all possible peptides without regarding cleavage specificity.

1.7 Analysis of human milk proteins and peptides: from univariate analysis to systems biology approaches

Along with the development of MS techniques, new tools are continuously being developed to extract and visualize relevant information from the data that these techniques generate.

A classical approach in the data analysis of proteomics and peptidomics data uses techniques such as univariate, multivariate, or pattern recognition methods. In univariate analysis, only one variable is considered at a time, aiming to find differences between groups of samples, such as in t-tests. Although univariate methods are generally favored because of the straightforward interpretation, applying them to proteomics and peptidomics would mean that significance testing is needed for hundreds of features. Consequentially, multiple testing correction needs to be applied to correct for the probability of false-positive results. Unfortunately, methods for such corrections are often too stringent, with the result that biologically relevant differences are often found to be non-significant after correction [97].

Multivariate analysis involves more than one variable and aims to find patterns and relations between multiple variables simultaneously. Examples of methods used for this are, amongst others, principal component analysis (PCA), partial least squares discriminant analysis (PLS-DA), Random Forest classification, and machine learning. Multivariate approaches can complement univariate analysis and reveal other relevant differences. This is possible because multivariate analysis can reveal (*i*) differences on how independent variables are related to each other, and (*ii*) subtle effects that are consistently present in a large set of the variables. These phenomena are explained in detail by Saccenti et al. [98].

Another way to analyze proteomics and peptidomics data is through a systems biology approach. Systems biology integrates multiple profiles of a biological system and uses these to elucidate the characteristics or pinpoint important parts of the system. Such an approach can move the focus from hypothesis-driven to discovery-driven research [99]. An example of how this can be applied to proteomics or peptidomics data is through network analysis. To approach data from a network point of view enables the consideration of individual components (nodes) together with the connections between them (edges) and their possible function in common pathways. An example of a network analysis is the comparison of networks across groups of samples through differential connectivity (Figure 1.8), which can reveal, for example, differential regulation of biological pathways between groups of samples.

In biological processes, proteins can be part of multiple pathways and can interact with other proteins within a pathway [101]. Protein synthesis itself results from a biological process, so protein abundance is also regulated by specific pathways. In response to conditions of the body such as stress, inflammation, etc., cellular changes can modify the importance of pathways and the interactions between proteins [102]. When it comes to peptides, peptide-peptide relations can be due to, for example, partly overlapping sequences (Figure 1.4), specificity of proteolytic cleavage, or biological function [103]. Furthermore, in combining proteomics and peptidomics data, relations can be present, for example, between proteases and the peptide fragments resulting from their proteolytic activity. Whereas regular chemometrics would not be able to elucidate such complex systems of interrelationships, network analysis fills this gap [104]. This gives new opportunities in the analysis of proteomics and peptidomics data, enabling the recognition of subtle changes in biological pathways.

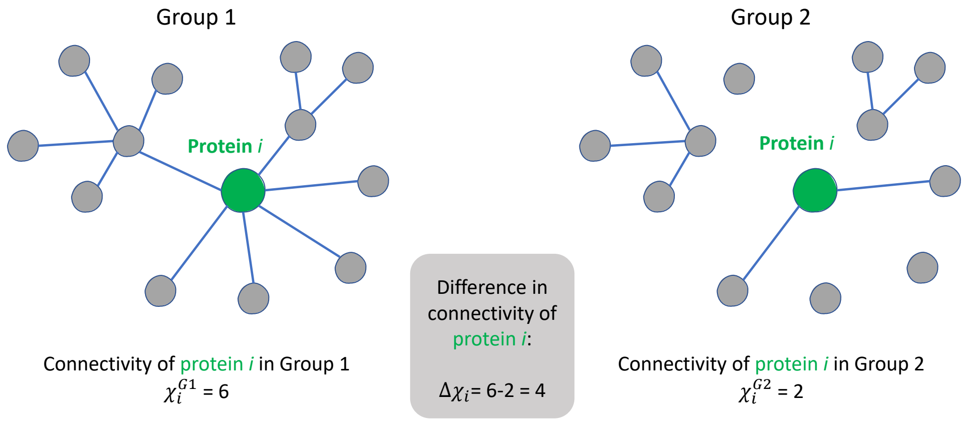

Group 1
Group 2
Protein *i*
Protein *i*
Difference in connectivity of protein *i*:
$\Delta\chi_i$= 6-2 = 4
Connectivity of protein *i* in Group 1
χ_i^{G1} = 6
Connectivity of protein *i* in Group 2
χ_i^{G2} = 2

Figure 1.8: Graphical illustration of the principle of differential connectivity in network analysis. Nodes represent features (proteins or peptides), and edges represent associations between the nodes. Two networks are shown, representing data from two different groups of samples in which protein *i* is differentially connected. Adapted from Jahagirdar et al. [100].

1.8 The CHILD Cohort Study

In 2008, a prospective longitudinal birth cohort study was launched in Canada, the CHILD Cohort Study (www.childstudy.ca) [105]. This cohort study aimed to advance knowledge about genetic and environmental determinants of atopic diseases. More than 3500 pregnant women participated, who gave birth between 2009 and 2012. Information on the children, their parents, and their environments was gathered over time through biological samples, questionnaires, home assessments, and clinical assessments. Amongst the biological samples collected in this cohort study was breast milk [106]. Together with the data gathered regarding allergy status of mother and child, this provided an excellent opportunity to investigate the human milk protein and peptide composition and their relation with allergy development.

1.9 Aim and outline of this thesis

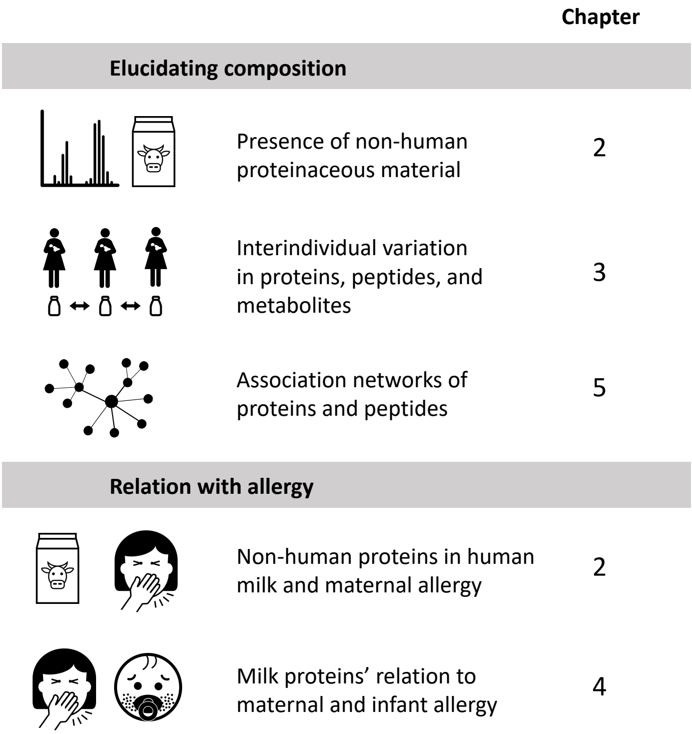
Human milk comprises a large variety of proteins and peptides that might be beneficial for the development of the infant in various ways. The milk proteome and peptidome have been extensively studied in recent years. However, several aspects of their nature are still unclear, such as the interindividual variation and the causal factors leading to the formation of the peptidome. In addition, little is known about their relation with maternal allergy and allergy development in breastfed infants. This thesis aimed to shed light on the detailed protein and peptide profile of human milk and on how they relate to allergy of both mother and child.

As illustrated in Figure 1.9, the work of this thesis is divided over several experimental chapters. In **Chapter 2**, non-human and allergenic proteins in human milk were investigated. To determine whether the levels of such proteins are dependent on maternal allergy, milk from allergic and non-allergic mothers was compared. In addition, the pathways through which these proteins can end up in the milk are discussed

The normal interindividual variation in the milk proteome and peptidome is poorly understood. **Chapter 3** provides insight into this, describing variation in 286 mature milk samples from 29 mothers (pooled per mother). The metabolome of these samples was also analyzed, adding the low molecular weight substances as potential indicators of the biological processes taking place in the mother's body.

A large study, comprising 300 milk samples from the CHILD Cohort Study, is described in **Chapter 4**. The samples used in this study were chosen in equal groups of 75, with all different combinations for both mother and child allergy status. The proteome of these samples was analyzed and compared to investigate the relation between the milk proteome and both allergy of mother and child. Furthermore, the peptidome was analyzed from the same samples, and relations between proteome and peptidome were investigated in **Chapter 5**. This was set up to elucidate the links between proteins and peptides and to obtain more insight into the factors that affect the degradation of proteins into peptides.

Lastly, the results of the **Chapters 2 - 5** are integrated and discussed in **Chapter 6**. This chapter ends with conclusions, scientific challenges, and recommendations for further research.



	Chapter
Elucidating composition	
Presence of non-human proteinaceous material	2
Interindividual variation in proteins, peptides, and metabolites	3
Association networks of proteins and peptides	5
Relation with allergy	
Non-human proteins in human milk and maternal allergy	2
Milk proteins' relation to maternal and infant allergy	4

Figure 1.9: Schematic overview of the experimental chapters in this thesis and their contribution to the aim of the study.

References

[1] Lönnerdal, B. "Nutritional and Physiologic Significance of Human Milk Proteins." In: *The American journal of clinical nutrition* 77 (2003), 1537S–1543S. DOI: 10.1093/ajcn/77.6.1537s.

[2] Trend, S. et al. "Levels of Innate Immune Factors in Preterm and Term Mothers' Breast Milk during the 1st Month Postpartum". In: *British Journal of Nutrition* 115 (2016), 1178–1193. DOI: 10.1017/S0007114516000234.

[3] Bilston-John, S. H. et al. "Daily and Within-Feed Variation of Macro- and Trace-Element Concentrations in Human Milk and Implications for Sampling". In: *Food Chemistry* 363 (2021), 130179. DOI: 10.1016/j.foodchem.2021.130179.

[4] Mitoulas, L. R. et al. "Variation in Fat, Lactose and Protein in Human Milk over 24h and throughout the First Year of Lactation". In: *British Journal of Nutrition* 88 (2002), 29–37. DOI: 10.1079/bjn2002579.

[5] Tian, H. M. et al. "Dietary Patterns Affect Maternal Macronutrient Intake Levels and the Fatty Acid Profile of Breast Milk in Lactating Chinese Mothers". In: *Nutrition* 58 (2019), 83–88. DOI: 10.1016/j.nut.2018.06.009.

[6] Baldeón, M. E. et al. "Free Amino Acid Content in Human Milk Is Associated with Infant Gender and Weight Gain during the First Four Months of Lactation". In: *Nutrients* 11 (2019), 2239. DOI: 10.3390/nu11092239.

[7] Monaco, M. H., Kim, J., and Donovan, S. M. "Human Milk: Composition and Nutritional Value". In: *Encyclopedia of Food and Health*. Elsevier, 2015, 357–362. ISBN: 978-0-12-384953-3. DOI: 10.1016/B978-0-12-384947-2.00413-X.

[8] Ip, S. et al. "Breastfeeding and Maternal and Infant Health Outcomes in Developed Countries." In: *Evidence report/technology assessment* (2007), 1–186. DOI: 10.1542/gr.18-2-15.

[9] Victora, C. G. et al. "Breastfeeding in the 21st Century: Epidemiology, Mechanisms, and Lifelong Effect". In: *The Lancet* 387 (2016), 475–490. DOI: 10.1016/S0140-6736(15)01024-7.

[10] Sullivan, S. et al. "An Exclusively Human Milk-Based Diet Is Associated with a Lower Rate of Necrotizing Enterocolitis than a Diet of Human Milk and Bovine Milk-Based Products". In: *Journal of Pediatrics* 156 (2010), 562–567.e1. DOI: 10.1016/j.jpeds.2009.10.040.

[11] Van Rossum, C. et al. *Health Effects of Breastfeeding, a Systematic Literature Review*. Tech. rep. Bilthoven: National Institute for Public Health and Environment, 2015.

[12] Autran, C. A. et al. “Human Milk Oligosaccharide Composition Predicts Risk of Necrotising Enterocolitis in Preterm Infants”. In: *Gut* 67 (2018), 1064–1070. doi: 10.1136/gutjnl-2016-312819.

[13] Meek, J. Y., Feldman-Winter, L., and Noble, L. “Optimal Duration of Breastfeeding”. In: *Pediatrics* 146 (2020), e2020021063. doi: 10.1542/peds.2020-021063.

[14] Fang, M. T. et al. “Developing Global Guidance on Human Milk Banking”. In: *Bulletin of the World Health Organization* 99 (2021), 892–900. doi: 10.2471/BLT.21.286943.

[15] Donovan, S. M. “Human Milk Proteins: Composition and Physiological Significance”. In: *Nestle Nutrition Institute Workshop Series*. Vol. 90. 2019, 93–101. doi: 10.1159/000490298.

[16] Zhu, J. and Dingess, K. A. “The Functional Power of the Human Milk Proteome”. In: *Nutrients* 11 (2019), 1834. doi: 10.3390/nu11081834.

[17] Liao, Y. et al. “Absolute Quantification of Human Milk Caseins and the Whey /Casein Ratio during the First Year of Lactation”. In: *Journal of Proteome Research* 16 (2017), 4113–4121. doi: 10.1021/acs.jproteome.7b00486.

[18] Liao, Y. et al. “Proteomic Characterization of Human Milk Whey Proteins during a Twelve-Month Lactation Period”. In: *Journal of Proteome Research* 10 (2011), 1746–1754. doi: 10.1021/pr101028k.

[19] Sood, S. M., Herbert, P. J., and Slattery, C. W. “Structural Studies on Casein Micelles of Human Milk: Dissociation of β-Casein of Different Phosphorylation Levels Induced by Cooling and Ethylenediaminetetraacetate”. In: *Journal of Dairy Science* 80 (1997), 628–633. doi: 10.3168/jds.S0022-0302(97)75980-0.

[20] Yang, M. et al. “Quantitative Proteomic Analysis of Milk Fat Globule Membrane (MFGM) Proteins in Human and Bovine Colostrum and Mature Milk Samples through iTRAQ Labeling”. In: *Food and Function* 7 (2016), 2438–2450. doi: 10.1039/c6fo00083e.

[21] Cavaletto, M., Giuffrida, M. G., and Conti, A. “The Proteomic Approach to Analysis of Human Milk Fat Globule Membrane”. In: *Clinica Chimica Acta* 347 (2004), 41–48. doi: 10.1016/j.cccn.2004.04.026.

[22] Liao, Y. et al. “Proteomic Characterization of Human Milk Fat Globule Membrane Proteins during a 12 Month Lactation Period”. In: *Journal of Proteome Research* 10 (2011), 3530–3541. doi: 10.1021/pr200149t.

[23] Vilotte, J. L. et al. "Genetics and Biosynthesis of Milk Proteins". In: *Advanced Dairy Chemistry*. Ed. by P. L. H. McSweeney and P. F. Fox. Fourth. Vol. 1A: Proteins: Basic Aspects. Boston, MA: Springer US, 2013, 431–461. ISBN: 978-1-4614-4714-6. DOI: 10.1007/978-1-4614-4714-6_14.

[24] Peterson, J. A. et al. "Structural and Functional Aspects of Three Major Glycoproteins of the Human Milk Fat Globule Membrane". In: *Advances in Experimental Medicine and Biology*. Vol. 501. 2001, 179–187. DOI: 10.1007/978-1-4615-1371-1_23.

[25] Qiu, Y. et al. "Bile Salt–Dependent Lipase Promotes the Barrier Integrity of Caco-2 Cells by Activating Wnt/β-Catenin Signaling via LRP6 Receptor". In: *Cell and Tissue Research* 383 (2021), 1077–1092. DOI: 10.1007/s00441-020-03316-4.

[26] Boron, W. F. and Boulpaep, E. L., eds. *Medical Physiology*. Third. Philadelphia: Elsevier Health Sciences, 2017. ISBN: 978-1-4557-4377-3.

[27] Burgoyne, R. D. and Duncan, J. S. "Secretion of Milk Proteins". In: *Journal of Mammary Gland Biology and Neoplasia* 3 (1998), 275–286. DOI: 10.1023/A:1018763427108.

[28] Monks, J. and Neville, M. C. "Albumin Transcytosis across the Epithelium of the Lactating Mouse Mammary Gland". In: *Journal of Physiology* 560 (2004), 267–280. DOI: 10.1113/jphysiol.2004.068403.

[29] Fox, P. F. et al. "Enzymology of Milk and Milk Products". In: *Dairy Chemistry and Biochemistry*. Cham: Springer International Publishing, 2015, 377–414. ISBN: 978-3-319-14892-2. DOI: 10.1007/978-3-319-14892-2_10.

[30] Stelwagen, K. and Singh, K. "The Role of Tight Junctions in Mammary Gland Function". In: *Journal of Mammary Gland Biology and Neoplasia* 19 (2014), 131–138. DOI: 10.1007/s10911-013-9309-1.

[31] Zhang, L. et al. "Changes over Lactation in Breast Milk Serum Proteins Involved in the Maturation of Immune and Digestive System of the Infant". In: *Journal of Proteomics* 147 (2016), 40–47. DOI: 10.1016/j.jprot.2016.02.005.

[32] Lönnerdal, B. et al. "Longitudinal Evolution of True Protein, Amino Acids and Bioactive Proteins in Breast Milk: A Developmental Perspective". In: *Journal of Nutritional Biochemistry* 41 (2017), 1–11. DOI: 10.1016/j.jnutbio.2016.06.001.

[33] Montagne, P. et al. "Immunological and Nutritional Composition of Human Milk in Relation to Prematurity and Mothers' Parity during the First 2 Weeks of Lactation". In: *Journal of Pediatric Gastroenterology and Nutrition* 29 (1999), 75–80. DOI: 10.1097/00005176-199907000-00018.

[34] Elwakiel, M. et al. "Variability of Serum Proteins in Chinese and Dutch Human Milk during Lactation". In: *Nutrients* 11 (2019), 499. DOI: 10.3390/nu11030499.

[35] Zhang, L. et al. "Geography and Ethnicity Related Variation in the Chinese Human Milk Serum Proteome". In: *Food and Function* 10 (2019), 7818–7827. DOI: 10.1039/c9fo01591d.

[36] Aslebagh, R. et al. "Proteomics Analysis of Human Breast Milk to Assess Breast Cancer Risk". In: *Electrophoresis* 39 (2018), 653–665. DOI: 10.1002/elps.201700123.

[37] Hettinga, K. A. et al. "Difference in the Breast Milk Proteome between Allergic and Non-Allergic Mothers". In: *PLoS ONE* 10 (2015). Ed. by A. S. Wiley, e0122234. DOI: 10.1371/journal.pone.0122234.

[38] Zhu, J. et al. "Personalized Profiling Reveals Donor- and Lactation-Specific Trends in the Human Milk Proteome and Peptidome". In: *Journal of Nutrition* 151 (2021), 826–839. DOI: 10.1093/jn/nxaa445.

[39] Chan, L. E., Beverly, R. L., and Dallas, D. C. "The Enzymology of Human Milk". In: *Food Engineering Series*. Ed. by A. L. Kelly and L. B. Larsen. Cham: Springer International Publishing, 2021, 209–243. ISBN: 978-3-030-55482-8. DOI: 10.1007/978-3-030-55482-8_9.

[40] Nielsen, S. D., Beverly, R. L., and Dallas, D. C. "Milk Proteins Are Predigested within the Human Mammary Gland". In: *Journal of Mammary Gland Biology and Neoplasia* 22 (2017), 251–261. DOI: 10.1007/s10911-018-9388-0.

[41] Demers-Mathieu, V. et al. "Changes in Proteases, Antiproteases, and Bioactive Proteins from Mother's Breast Milk to the Premature Infant Stomach". In: *Journal of Pediatric Gastroenterology and Nutrition* 66 (2018), 318–324. DOI: 10.1097/MPG.0000000000001719.

[42] Puente, X. S. et al. "A Genomic View of the Complexity of Mammalian Proteolytic Systems". In: *Biochemical Society Transactions* 33 (2005), 331–334. DOI: 10.1042/BST0330331.

[43] Dallas, D. C., Murray, N. M., and Gan, J. "Proteolytic Systems in Milk: Perspectives on the Evolutionary Function within the Mammary Gland and the Infant". In: *Journal of Mammary Gland Biology and Neoplasia* 20 (2015), 133–147. DOI: 10.1007/s10911-015-9334-3.

[44] Kulman, J. D. and Davie, E. W. "Proteases in Blood Clotting". In: *Encyclopedia of Biological Chemistry: Second Edition*. Elsevier, 2013, 585–589. ISBN: 978-0-12-378631-9. DOI: 10.1016/B978-0-12-378630-2.00020-7.

[45] Dingess, K. A. et al. “Toward an Efficient Workflow for the Analysis of the Human Milk Peptidome”. In: *Analytical and Bioanalytical Chemistry* 411 (2019), 1351–1363. doi: 10.1007/s00216-018-01566-4.

[46] Dingess, K. A. et al. “Human Milk Peptides Differentiate between the Preterm and Term Infant and across Varying Lactational Stages”. In: *Food and Function* 8 (2017), 3769–3782. doi: 10.1039/c7fo00539c.

[47] Guerrero, A. et al. “Mechanistic Peptidomics: Factors That Dictate Specificity in the Formation of Endogenous Peptides in Human Milk”. In: *Molecular and Cellular Proteomics* 13 (2014), 3343–3351. doi: 10.1074/mcp.M113.036194.

[48] Murakami, M. et al. “Expression and Secretion of Cathelicidin Antimicrobial Peptides in Murine Mammary Glands and Human Milk”. In: *Pediatric Research* 57 (2005), 10–15. doi: 10.1203/01.PDR.0000148068.32201.50.

[49] Singh, A. et al. “The Protective Effects of Human Milk-Derived Peptides on the Pancreatic Islet Biology”. In: *Biology Open* 9 (2020). doi: 10.1242/BIO.049304.

[50] Cai, J. et al. “A Human β-Casein-Derived Peptide BCCY-1 Modulates the Innate Immune Response”. In: *Food Chemistry* 348 (2021), 129111. doi: 10.1016/j.foodchem.2021.129111.

[51] Dallas, D. C. et al. “Endogenous Human Milk Peptide Release Is Greater after Preterm Birth than Term Birth”. In: *Journal of Nutrition* 145 (2015), 425–433. doi: 10.3945/jn.114.203646.

[52] Wan, J. et al. “Peptidome Analysis of Human Skim Milk in Term and Preterm Milk”. In: *Biochemical and Biophysical Research Communications* 438 (2013), 236–241. doi: 10.1016/j.bbrc.2013.07.068.

[53] Jarmołowska, B. et al. “Changes of β-Casomorphin Content in Human Milk during Lactation”. In: *Peptides* 28 (2007), 1982–1986. doi: 10.1016/j.peptides.2007.08.002.

[54] Axelsson, I. et al. “Bovine β-Lactoglobulin in the Human Milk. A Longitudinal Study during the Whole Lactation Period”. In: *Acta Paediatrica Scandinavica* 75 (1986), 702–707. doi: 10.1111/j.1651-2227.1986.tb10277.x.

[55] Picariello, G. et al. “Excretion of Dietary Cow’s Milk Derived Peptides into Breast Milk”. In: *Frontiers in Nutrition* 6 (2019), 25. doi: 10.3389/fnut.2019.00025.

[56] Chirdo, F. G. et al. “Presence of High Levels of Non-Degraded Gliadin in Breast Milk from Healthy Mothers”. In: *Scandinavian Journal of Gastroenterology* 33 (1998), 1186–1192. doi: 10.1080/00365529850172557.

[57] Macchiaverni, P. et al. “Respiratory Allergen from House Dust Mite Is Present in Human Milk and Primes for Allergic Sensitization in a Mouse Model of Asthma”. In: *Allergy: European Journal of Allergy and Clinical Immunology* 69 (2014), 395–398. doi: 10.1111/all.12332.

[58] Troncone, R. et al. “Passage of Gliadin into Human Breast Milk”. In: *Acta Paediatrica Scandinavica* 76 (1987), 453–456. doi: 10.1111/j.1651-2227.1987.tb10498.x.

[59] Hirose, J. et al. “Occurrence of the Major Food Allergen, Ovomucoid, in Human Breast Milk as an Immune Complex”. In: *Bioscience, Biotechnology and Biochemistry* 65 (2001), 1438–1440. doi: 10.1271/bbb.65.1438.

[60] Sorva, R., Mäkinen-Kiljunen, S., and Juntunen-Backman, K. “β-Lactoglobulin Secretion in Human Milk Varies Widely after Cow’s Milk Ingestion in Mothers of Infants with Cow’s Milk Allergy”. In: *The Journal of Allergy and Clinical Immunology* 93 (1994), 787–792. doi: 10.1016/0091-6749(94)90259-3.

[61] Bertino, E. et al. “Absence in Human Milk of Bovine Beta-Lactoglobulin Ingested by the Mother. Unreliability of ELISA Measurements.” In: *Acta Biomedica de Ateneo Parmense* 68 Suppl 1 (1997), 15–9.

[62] Picariello, G. et al. “Antibody-Independent Identification of Bovine Milk-Derived Peptides in Breast-Milk”. In: *Food and Function* 7 (2016), 3402–3409. doi: 10.1039/C6FO00731G.

[63] Zhu, J. et al. “Discovery and Quantification of Nonhuman Proteins in Human Milk”. In: *Journal of Proteome Research* 18 (2019), 225–238. doi: 10.1021/acs.jproteome.8b00550.

[64] Pastor-Vargas, C. et al. “Sensitive Detection of Major Food Allergens in Breast Milk: First Gateway for Allergenic Contact during Breastfeeding”. In: *Allergy: European Journal of Allergy and Clinical Immunology* 70 (2015), 1024–1027. doi: 10.1111/all.12646.

[65] Boyce, J. A. et al. “Guidelines for the Diagnosis and Management of Food Allergy in the United States: Report of the NIAID-sponsored Expert Panel”. In: *Journal of Allergy and Clinical Immunology* 126 (2010), S1–S58. doi: 10.1016/j.jaci.2010.10.007.

[66] Dunngalvin, A. et al. “The Effects of Food Allergy on Quality of Life”. In: *Chemical Immunology and Allergy*. Vol. 101. 2015, 235–252. doi: 10.1159/000375106.

[67] Turner, P. J. et al. “Global Trends in Anaphylaxis Epidemiology and Clinical Implications”. In: *Journal of Allergy and Clinical Immunology: In Practice* 8 (2020), 1169–1176. doi: 10.1016/j.jaip.2019.11.027.

[68] Gupta, R. et al. “The Economic Impact of Childhood Food Allergy in the United States”. In: *JAMA Pediatrics* 167 (2013), 1026–1031. DOI: 10.1001/jamapediatrics.2013.2376.

[69] Holloway, J. W., Yang, I. A., and Holgate, S. T. “Genetics of Allergic Disease”. In: *Journal of Allergy and Clinical Immunology* 125 (2010), S81–S94. DOI: 10.1016/j.jaci.2009.10.071.

[70] Bønnelykke, K. et al. “Genetics of Allergy and Allergic Sensitization: Common Variants, Rare Mutations”. In: *Current Opinion in Immunology* 36 (2015), 115–126. DOI: 10.1016/j.coi.2015.08.002.

[71] Cookson, W. “The Alliance of Genes and Environment in Asthma and Allergy”. In: *Nature* 402 (1999), 5–11. DOI: 10.1038/35037002.

[72] Reynolds, L. A. and Finlay, B. B. “Early Life Factors That Affect Allergy Development”. In: *Nature Reviews Immunology* 17 (2017), 518–528. DOI: 10.1038/nri.2017.39.

[73] Gill, N., Wlodarska, M., and Finlay, B. B. “The Future of Mucosal Immunology: Studying an Integrated System-Wide Organ”. In: *Nature Immunology* 11 (2010), 558–560. DOI: 10.1038/ni0710-558.

[74] Meyers, D. A. et al. “Asthma Genetics and Personalised Medicine”. In: *The Lancet Respiratory Medicine* 2 (2014), 405–415. DOI: 10.1016/S2213-2600(14)70012-8.

[75] Kull, I. et al. “Breast-Feeding in Relation to Asthma, Lung Function, and Sensitization in Young Schoolchildren”. In: *Journal of Allergy and Clinical Immunology* 125 (2010), 1013–1019. DOI: 10.1016/j.jaci.2010.01.051.

[76] Saarinen, U. M. and Kajosaari, M. “Breastfeeding as Prophylaxis against Atopic Disease: Prospective Follow-up Study until 17 Years Old”. In: *The Lancet* 346 (1995), 1065–1069. DOI: 10.1016/S0140-6736(95)91742-X.

[77] Belderbos, M. E. et al. “Breastfeeding Modulates Neonatal Innate Immune Responses: A Prospective Birth Cohort Study”. In: *Pediatric Allergy and Immunology* 23 (2012), 65–74. DOI: 10.1111/j.1399-3038.2011.01230.x.

[78] Mihrshahi, S. et al. “The Association between Infant Feeding Practices and Subsequent Atopy among Children with a Family History of Asthma”. In: *Clinical & Experimental Allergy* 37 (2007), 671–679. DOI: 10.1111/j.1365-2222.2007.02696.x.

[79] Wright, A. L. et al. “Factors Influencing the Relation of Infant Feeding to Asthma and Recurrent Wheeze in Childhood”. In: *Thorax* 56 (2001), 192–197. DOI: 10.1136/thorax.56.3.192.

[80] Matheson, M. C., Allen, K. J., and Tang, M. L. "Understanding the Evidence for and against the Role of Breastfeeding in Allergy Prevention". In: *Clinical & Experimental Allergy* 42 (2012), 827–851. DOI: 10.1111/j.1365-2222.2011.03925.x.

[81] Hirotani, Y. et al. "Protective Effects of Lactoferrin against Intestinal Mucosal Damage Induced by Lipopolysaccharide in Human Intestinal Caco-2 Cells". In: *Journal of the Pharmaceutical Society of Japan* 128 (2008), 1363–1368. DOI: 10.1248/yakushi.128.1363.

[82] York, D. J. et al. "Human Milk Growth Factors and Their Role in Nec Prevention: A Narrative Review". In: *Nutrients* 13 (2021), 3751. DOI: 10.3390/nu13113751.

[83] Maga, E. A. et al. "Consumption of Lysozyme-Rich Milk Can Alter Microbial Fecal Populations". In: *Applied and Environmental Microbiology* 78 (2012), 6153–6160. DOI: 10.1128/AEM.00956-12.

[84] Chatterton, D. E. et al. "Anti-Inflammatory Mechanisms of Bioactive Milk Proteins in the Intestine of Newborns". In: *International Journal of Biochemistry and Cell Biology* 45 (2013), 1730–1747. DOI: 10.1016/j.biocel.2013.04.028.

[85] Munblit, D. et al. "Human Milk and Allergic Diseases: An Unsolved Puzzle". In: *Nutrients* 9 (2017), 894. DOI: 10.3390/nu9080894.

[86] Järvinen, K. M. et al. "Does Low IgA in Human Milk Predispose the Infant to Development of Cow's Milk Allergy?" In: *Pediatric Research* 48 (2000), 457–462. DOI: 10.1203/00006450-200010000-00007.

[87] Österlund, P. et al. "Eosinophil Cationic Protein in Human Milk Is Associated with Development of Cow's Milk Allergy and Atopic Eczema in Breast-Fed Infants". In: *Pediatric Research* 55 (2004), 296–301. DOI: 10.1203/01.PDR.0000106315.00474.6F.

[88] du Toit, G. et al. "Allergen Specificity of Early Peanut Consumption and Effect on Development of Allergic Disease in the Learning Early About Peanut Allergy Study Cohort". In: *Journal of Allergy and Clinical Immunology* 141 (2018), 1343–1353. DOI: 10.1016/j.jaci.2017.09.034.

[89] Perkin, M. R. et al. "Efficacy of the Enquiring About Tolerance (EAT) Study among Infants at High Risk of Developing Food Allergy". In: *Journal of Allergy and Clinical Immunology* 144 (2019), 1606–1614.e2. DOI: 10.1016/j.jaci.2019.06.045.

[90] Strobel, S. and Ferguson, A. "Immune Responses to Fed Protein Antigens in Mice. 3. Systemic Tolerance or Priming Is Related to Age at Which Antigen Is First Encountered". In: *Pediatric Research* 18 (1984), 588–594. DOI: 10.1203/00006450-198407000-00004.

[91] Macchiaverni, P. et al. "Allergen Shedding in Human Milk: Could It Be Key for Immune System Education and Allergy Prevention?" In: *Journal of Allergy and Clinical Immunology* 148 (2021), 679–688. DOI: 10.1016/j.jaci.2021.07.012.

[92] Verhasselt, V. et al. "Ovalbumin in Breastmilk Is Associated with a Decreased Risk of IgE-mediated Egg Allergy in Children". In: *Allergy: European Journal of Allergy and Clinical Immunology* 75 (2020), 1463–1466. DOI: 10.1111/all.14142.

[93] Adel-Patient, K. et al. "Prevention of Allergy to a Major Cow's Milk Allergen by Breastfeeding in Mice Depends on Maternal Immune Status and Oral Exposure during Lactation". In: *Frontiers in Immunology* 11 (2020), 1–10. DOI: 10.3389/fimmu.2020.01545.

[94] Zhang, Y. et al. "Protein Analysis by Shotgun/Bottom-up Proteomics". In: *Chemical Reviews* 113 (2013), 2343–2394. DOI: 10.1021/cr3003533.

[95] Verheggen, K. et al. "Anatomy and Evolution of Database Search Engines—a Central Component of Mass Spectrometry Based Proteomic Workflows". In: *Mass Spectrometry Reviews* 39 (2020), 292–306. DOI: 10.1002/mas.21543.

[96] Schrader, M., Schulz-Knappe, P., and Fricker, L. D. "Historical Perspective of Peptidomics". In: *EuPA Open Proteomics* 3 (2014), 171–182. DOI: 10.1016/j.euprot.2014.02.014.

[97] Pascovici, D. et al. "Multiple Testing Corrections in Quantitative Proteomics: A Useful but Blunt Tool". In: *Proteomics* 16 (2016), 2448–2453. DOI: 10.1002/pmic.201600044.

[98] Saccenti, E. et al. "Reflections on Univariate and Multivariate Analysis of Metabolomics Data". In: *Metabolomics* 10 (2014), 361–374. DOI: 10.1007/s11306-013-0598-6.

[99] Rosato, A. et al. "From Correlation to Causation: Analysis of Metabolomics Data Using Systems Biology Approaches". In: *Metabolomics* 14 (2018), 37. DOI: 10.1007/s11306-018-1335-y.

[100] Jahagirdar, S. and Saccenti, E. "On the Use of Correlation and MI as a Measure of Metabolite—Metabolite Association for Network Differential Connectivity Analysis". In: *Metabolites* 10 (2020). DOI: 10.3390/metabo10040171.

[101] Richards, A. L., Eckhardt, M., and Krogan, N. J. "Mass Spectrometry-based Protein–Protein Interaction Networks for the Study of Human Diseases". In: *Molecular Systems Biology* 17 (2021), 1–18. DOI: 10.15252/msb.20188792.

[102] Yan, P. et al. "Molecular Stressors Engender Protein Connectivity Dysfunction through Aberrant N-glycosylation of a Chaperone". In: *Cell Reports* 31 (2020), 107840. DOI: 10.1016/j.celrep.2020.107840.

[103] Lamerz, J. et al. "Correlation-Associated Peptide Networks of Human Cerebrospinal Fluid". In: *Proteomics* 5 (2005), 2789–2798. DOI: 10.1002/pmic.200401192.

[104] Kuzmanov, U. and Emili, A. "Protein-Protein Interaction Networks: Probing Disease Mechanisms Using Model Systems". In: *Genome Medicine* 5 (2013), 37. DOI: 10.1186/gm441.

[105] Subbarao, P. et al. "The Canadian Healthy Infant Longitudinal Development (CHILD) Study: Examining Developmental Origins of Allergy and Asthma". In: *Thorax* 70 (2015), 998–1000. DOI: 10.1136/thoraxjnl-2015-207246.

[106] Moraes, T. J. et al. "The Canadian Healthy Infant Longitudinal Development Birth Cohort Study: Biological Samples and Biobanking". In: *Paediatric and Perinatal Epidemiology* 29 (2015), 84–92. DOI: 10.1111/ppe.12161.

Chapter 2

Maternal allergy and the presence of non-human proteinaceous molecules in human milk

This chapter is based on:

Pieter M. Dekker, Sjef Boeren, Alet H. Wijga, Gerard H. Koppelman, Jacques J.M. Vervoort, Kasper A. Hettinga. Maternal allergy and the presence of nonhuman proteinaceous molecules in human milk.
Published in: Nutrients 12(4), 1169 (2020).
DOI: 10.3390/nu12041169

Abstract

Human milk contains proteins and/or protein fragments that originate from non-human organisms. These proteinaceous molecules, of which the secretion might be related to the mother's allergy status, could be involved in the development of the immune system of the infant. This may lead, for example, to sensitization or the induction of allergen-specific tolerance. The aim of this study was to investigate the relation between maternal allergy and the levels of non-human proteinaceous molecules in their milk. In this study, we analysed trypsin-digested human milk serum proteins of 10 allergic mothers and 10 non-allergic mothers. A search was carried out to identify peptide sequences originating from bovine or other allergenic proteins. Several methods were applied to confirm the identification of these sequences, and the differences between both groups were investigated. Out of the 78 identified non-human peptide sequences, 62 sequences matched *Bos taurus* proteins. Eight peptide sequences of bovine β-lactoglobulin had significantly higher levels in milk from allergic mothers than in milk from non-allergic mothers. Dietary bovine β-lactoglobulin may be absorbed through the intestinal barrier and secreted into human milk. This seems to be significantly higher in allergic mothers and might have consequences for the development of the immune system of their breastfed infant.

2.1 Introduction

The human milk proteome comprises more than once thought. Besides a vast number of human proteins and peptides, it also includes non-human intact proteins, large protein fragments, and peptides (later referred to as proteinaceous molecules). The presence of such molecules in human milk has been demonstrated decades ago with immunochemical analysis [1] and has recently been confirmed with mass spectrometry [2].

According to studies using mass spectrometry, the main biological source of the non-human proteinaceous molecules in human milk seems to be bovine milk. A peptidomics study demonstrated the presence of two peptides originating from bovine β-lactoglobulin (BLG) and one originating from α_{s1}-casein [3]. In a later study, this was extended with peptides from α-lactalbumin (ALA), κ-casein, β-casein, and lactoferrin [4]. Evidence for the presence of intact bovine caseins and BLG has recently been provided [2, 4, 5]. In addition to the bovine proteins and peptides, Zhu et al. [2] also identified several peptide sequences originating from other non-human species, which may include allergens. So far, only peanut allergen has been identified with high sequence coverage by liquid chromatography tandem mass spectrometry (LC-MS/MS) [6]. The presence of egg, wheat, and house dust mite (HDM) allergens in human milk, which has been demonstrated using immunochemical methods [7–9], has not been confirmed yet with LC-MS/MS analysis.

The presence of these non-human proteinaceous molecules in human milk raises the question how they end up there. It has been suggested that dietary proteins can be transferred through the intestinal barrier by both paracellular and transcellular pathways [10]. Furthermore, non-human proteins present in the mother's blood might also be transferred by these pathways through the mammary epithelia into the milk [11, 12]. Nevertheless, it remains unclear which pathways are used for the transfer of non-human proteinaceous molecules into human milk.

Several of the studies that identified non-human proteinaceous molecules in human milk report a large interindividual variation in the levels of these molecules. An explanation for this variation has not been found yet, but it does not seem to be related to dietary intake [13]. Maternal asthma and allergy could be an important factor in this variation, since it is known that e.g., atopic eczema and asthma can have an influence on intestinal barrier integrity [14–16]. This could then lead to an increased passage of dietary proteinaceous molecules through the intestinal barrier. Research to date has not yet considered the relation between maternal allergic diseases and the levels of non-human proteinaceous molecules in human milk using LC-MS/MS. The purpose of the present study was therefore to identify non-human proteinaceous molecules in human milk and to investigate if the levels of these molecules were related to maternal allergy. This could be useful for further research into the mechanisms responsible for the transfer of these molecules and the effect of these molecules on the infant's immune system.

2.2 Materials and Methods

2.2.1 Milk samples

We used data from a population-based Dutch birth cohort study: the Prevention and Incidence of Asthma and Mite Allergy (PIAMA) Study. Details of the cohort study are described elsewhere [17, 18]. In short, pregnant women were recruited from the general population during their first antenatal visit. Their children (*n* = 3963) were born in 1996/1997. Pregnant women were identified as allergic or non-allergic through a screening questionnaire. House dust samples and breastmilk samples were collected in a subgroup of the population around the child's age of three months. Breastmilk collection was done by manual pressure or by use of a breast pump. Samples were stored in small plastic cups at -80°C. Along with these samples, cat ownership and the frequency of consumption of milk and milk products by the mother was assessed using a questionnaire (Table 2.1). Maternal blood samples were collected at the child's age of one year. The study was performed in accordance with the ethical principles for medical research involving human subjects outlined in the Declaration of Helsinki. Therefore, the study protocol was approved by the Medical Ethics Committees of the participating institutes (Rotterdam MEC 132.636/1994/39 and 137.326/1994/130; Groningen MEC 94/08/92; Utrecht, MEC-TNO oordeel 95/50). All parents gave written informed consent.

The current study is based on a data-dependent LC-MS/MS proteomics data set that was obtained in a previous study [19]. It comprises mass spectrometry data of human milk serum protein samples from 10 allergic mothers and 10 non-allergic mothers from the cohort study. The number of mothers included is based on a power calculation, aiming at finding a 5-fold difference, as detailed in Hettinga et al. [19]. The selection of the allergic mothers was based on (a) self-reported (history of) asthma, current hay fever, current allergy for pets, or current allergy for house dust or house dust mite, in combination with (b) a high level of specific IgE against HDM ($\geq$3.50 kU/L) and (c) exposure to HDM allergen in mattress dust ((Der p 1 + Der f 1) >600 ng/m^2) (see Table 2.1). The selection of non-allergic mothers did not report any allergies or asthma. This group consisted of mothers with exposure to HDM allergen in mattress dust in the same range as the allergic mothers (600–2500 ng/m^2) as well as mothers with much higher exposures ($\geq$24,000 ng/m^2). The non-allergic mothers were not tested for specific IgE against house dust mite.

Table 2.1: Details on the mothers included in the sample collection, with allergy status, Der p IgE Rast-class of the allergic mothers, presence of a cat as pet and consumption of milk and dairy products.

Characteristic	Type	Non-allergic	Allergic
House dust mite allergy	Self-report	0	7
	Doctor diagnosed	0	7
House dust allergy	Self-report	0	8
	Doctor diagnosed	0	6
Allergic to pets	Self-report	0	9
	Doctor diagnosed	0	8
Asthma	Self-report	0	7
	Doctor diagnosed	0	7
House dust mite Der p IgE (Rast-class)	Class 3	NA[a]	4
	Class 4	NA[a]	5
	Class 5	NA[a]	1
Rhinitis/hay fever	Self-report	0	9
Cat as pet in the household	Presence	3	3
Consumption of milk during lactation	Not at all	2	3
	1-3x a month	0	1
	1x a week	0	0
	2-4x a week	0	3
	More than 4x a week	0	0
	1x a day	1	1
	Multiple times a day	7	2
Consumption of milk products during lactation	Not at all	2	0
	1-3x a month	0	0
	1x a week	0	0
	2-4x a week	1	0
	More than 4x a week	0	1
	1x a day	4	4
	Multiple times a day	3	5

[a] NA = Data not available.

From the milk samples, milk serum was obtained, and serum proteins were prepared for analysis with filter-aided sample preparation. In short, full scan FTMS spectra were obtained (*m/z* 380 to 1400) in positive mode on an LTQ-Orbitrap system (Thermo electron, San Jose, CA, USA). The four multiply-charged peaks with the highest intensity were recorded in the linear trap in data-dependent mode (MSMS threshold: 5000). Further details of the sample preparation and LC-MS/MS analysis have been described before [19].

The data underlying the findings presented in this paper are available on request. Requests can be submitted to the PIAMA principal investigators. Their names and e-mail addresses are listed on the PIAMA website (https://piama.iras.uu.nl/in-english/).

2.2.2 Methods

Data analysis

The raw MS/MS data was analysed using the Andromeda search engine of the MaxQuant software v1.6.1.0 [20]. Since the use of large databases in proteomic data analysis affects the sensitivity of the search [21], a complete but concise database was created for this study. This database contained human milk proteins (n = 2569), bovine milk proteins (n = 1006), and allergen proteins (n = 721). This database is provided in the Supplementary information, the fasta database. Allergens were added to the database because of their immunological relevance and bovine milk proteins because the majority of the non-human proteinaceous molecules in human milk was previously shown to originate from bovine milk [2]. The selection of human and bovine milk proteins was made based on previous data analysis of human and bovine milk protein samples (data not published) using databases with all human or bovine proteins available in UniProtKB (both downloaded from UniProt on 16-10-2018). This was complemented with data from reviews on the bovine milk and human milk proteome [22, 23]. Allergen protein sequences were obtained from UniProt on 16-10-2018 by performing a search on all proteins annotated as allergen (search term: "annotation:(type:allergen)").

The search for peptide sequences was performed three times, in which the protein database was *in silico* digested with trypsin digestion, semi-specific trypsin digestion, or unspecific digestion. The maximum number of missed cleavages was set to two in the trypsin digestion mode. In all searches, a fixed modification was set to carbamidomethylation of cysteine. Variable modifications were set to acetylation of the peptide N-term, deamidation of the side chains of asparagine and glutamine, and oxidation of methionine, with a maximum of five modifications per peptide. The identified peptides were quantified using label-free quantification (LFQ). At both peptide and protein levels, a false discovery rate of 1% was used. The peptide length was set from 6 to 35 amino acids. The precursor mass tolerance was set to 20 ppm, and fragment mass tolerance to 0.5 Da. Recalibration was carried out using a first search with a database containing common contaminants.

To remove all identifications that belong to sequences originating from human proteins, the MaxQuant output was subjected to a filtering consisting of six steps. First, all sequences originating from trypsin and keratin were removed as contaminants. Second, the reverse sequences from the decoy database were removed. Third, all sequences that had a full match with the human proteome were removed. Fourth, we removed all MS/MS scans that had a match in a separate search using only the whole human proteome database. Fifth, all sequences with an Andromeda score lower than 80 were removed to exclude low quality peptide spectrum matches (PSM). Sixth, PSMs with a second-best match to a human peptide sequence and an Andromeda score difference of <5 were removed.

Annotation

Protein entries containing an exact match with the identified and selected peptides were found using the Peptide Match service of the online Protein Information Resource [24]. This service makes use of an up-to-date UniProtKB database. Peptides were matched to this database without isoforms, where leucine and isoleucine were treated as equivalent.

Protein and organism annotation was added using a frequency of occurrence. All matching proteins and their corresponding taxonomic lineage were listed. A leading protein was selected for each peptide sequence based on the frequency of occurrence of this protein in the peptide match results. After this, a similar approach was used on the level of taxonomy, leaving the organisms with the highest number of matches to the identified peptides as leading organism or, in case of multiple organisms, the lowest common ancestor (LCA). With this approach, *Bos taurus* was preferred over e.g., *Bos mutus* as leading organism because of a higher number of identified peptides that matched the *Bos taurus* proteins.

Statistical analysis

Data analysis was carried out, and figures were made using R version 3.6.0 [25]. Missing values of LFQ intensities for the identified and selected peptide sequences were associated with levels below detection limit. Therefore, imputation was applied to log10 transformed LFQ intensities, with values from a normal distribution downshifted from the sample mean with 1.8 and with a standard deviation of 0.3.

Differences between the allergic and non-allergic group were tested using a two-sided unequal variances *t*-test and a Benjamini-Hochberg correction was applied on the resulting *p*-values. Significantly different peptides were selected with a *p*-value <0.01. An additional threshold of 0.75 was set on the difference between the means of the sample groups (log10 transformed intensity values) in order to select only significant sequences with a large between-group difference.

Confirmatory analysis

Bovine caseinate (prepared in-house), lactoferrin, BLG, ALA and bovine serum albumin (BSA) (Sigma-Aldrich, St. Louis, MO, USA) were dissolved in a 100 mM

Tris solution and digested with trypsin. For confirmation of the non-human, non-bovine peptides, 12 peptides were acquired through synthesis by Royobiotech Co., Ltd. (Shanghai, China). Protein digests and synthetic peptides were analysed one by one on the same LC-MS/MS system and with the same parameters as used for the analysis of the human milk samples [19]. A summary of the workflow and confirmation of MSMS spectra is visualized in Figure 2.1.

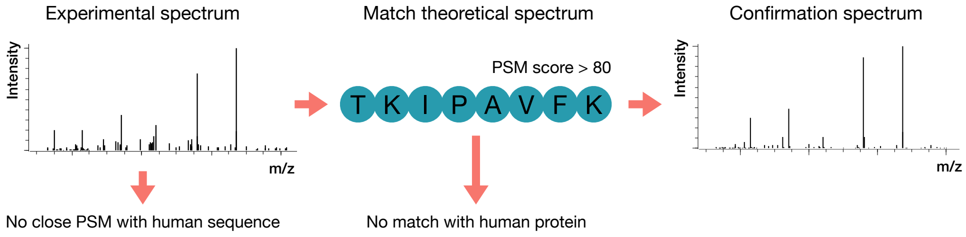


Figure 2.1: Schematic overview of the workflow used for confirmation of the identified non-human peptide sequences. After LC-MS/MS analysis, experimental MSMS spectra that matched with theoretical non-human peptide sequences were selected when there was no close peptide sequence match (PSM) with a human peptide sequence. Spectra with a PSM score <80 or a full match with the human proteome were removed. The final remaining spectra were confirmed with retention time and MSMS spectra of bovine milk, pure proteins, or synthetically acquired peptides.

2.3 Results

In this study, data-dependent shotgun proteomics data of human milk serum from 10 allergic and 10 non-allergic mothers was analysed. In a search for non-human proteins and protein fragments, the identified peptides were filtered and LFQ data was used for quantification.

2.3.1 Identification of exogenous peptides

Trypsin-digested human milk serum protein data was analysed using a database containing human milk, bovine milk, and allergen protein sequences. The identified peptide sequences were filtered to remove all human peptides and as many false positives as possible. In total, 78 non-human peptide sequences were identified in 20 samples. From these, 62 sequences had *Bos taurus* as leading organism (Table

2.2) and 16 sequences were assigned to non-bovine allergens (Table 2.3). Most of the identified peptide sequences ($n = 48$) were from trypsin-digested proteins. In addition, 10 peptides were semi-trypsin digested and 20 were not digested by trypsin.

Peptide sequences of 29 different bovine proteins were identified. From these proteins, the major bovine milk allergen, BLG, was identified with the highest sequence coverage (67%). To confirm the identification of the bovine sequences, tryptic digests of the major bovine milk proteins, BLG, BSA, α_{s1}-casein, and ALA were analysed. This led to confirmation of 20 sequences based on MS/MS spectrum and retention time. The identification of another 16 sequences was confirmed by MS/MS spectra and retention times of these sequences in a bovine milk protein data set (data set not published). The protein with the second highest sequence coverage is bovine serum albumin (BSA). The identified peptide sequences ($n = 14$) correspond to a sequence coverage of (22%). In contrast to studies from other groups that removed these peptides from their data sets because of the use of BSA as quality control in their studies [2, 4], we did not remove these peptides from our results. No evidence was found for carryover of BSA peptides in the LC system. Several BSA peptide sequences identified in human milk were not found in a trypsin-digested BSA standard solution that was used in our laboratory, indicating that the BSA-derived peptides are genuine. Considering these findings, it is likely that BSA or its proteolytic fragments are present in human milk.

From the non-bovine allergen proteins or protein fragments, proteins from *Felis catus* (domestic cat), *Equus caballus* (horse), and *Triticum aestivum* (common wheat) were identified with two or more peptide sequences. To confirm the identification of these peptide sequences, one synthesized sequence of each identified protein was acquired and analysed. From these nine peptides, two sequences were confirmed based on MS/MS spectrum and retention time. These sequences had cat and horse serum albumin as leading protein. The remaining seven sequences could not be confirmed. In several cases, the PSM of the synthesized peptide resembled the PSM of the human milk samples, but the retention time differed significantly. These PSMs are likely false positives, showing that the search for low abundant peptide sequences in shotgun proteomics is prone to finding PSMs with artefacts or co-eluted peptides. This confirms the importance of the used method in which synthesized peptides were used for confirmation.

In addition to the 16 sequences that were assigned to non-bovine allergens, 11 peptides were annotated with *Hevea brasiliensis* (Rubber tree) as leading organism. Two of these sequences were confirmed with MS/MS spectra and retention time of synthesized peptides. Nevertheless, data analysis of three other data sets of human milk shotgun proteomics did not show the presence of these peptides (data not published). Therefore, these sequences were considered as contaminant and removed from the results.

Table 2.2: All identified non-human peptide sequences that were assigned to bovine proteins, with the corresponding UniProt protein id, name of the leading protein, and *in silico* digestion mode. The number of samples in which the peptide sequence was identified is listed per group of allergic and non-allergic mothers.

Sequence	Leading proteins	Protein names	Allergic	Non-allergic	Digestion
ALPMHIR [a]	B5B0D4	β-lactoglobulin	10	2	trypsin
IDALNENK [a]	B5B0D4	β-lactoglobulin	10	5	trypsin
LIVTQTMK [a]	B5B0D4	β-lactoglobulin	10	4	trypsin
LSFNPTQLEEQCHI [b]	B5B0D4	β-lactoglobulin	10	6	trypsin
TKIPAVFK [a]	B5B0D4	β-lactoglobulin	10	0	trypsin
TPEVDDEALEK [a]	B5B0D4	β-lactoglobulin	10	2	trypsin
TPEVDDEALEKFDK [a]	B5B0D4	β-lactoglobulin	10	5	trypsin
VLVLDTDYKK [a]	B5B0D4	β-lactoglobulin	10	5	trypsin
VYVEELKPTPEGDLEILLQK [a]	B5B0D4	β-lactoglobulin	9	1	trypsin
WENDECAQK [b]	B5B0D4	β-lactoglobulin	9	1	trypsin
WENDECAQKK [b]	B5B0D4	β-lactoglobulin	4	0	trypsin
SLAMAASDISLLDAQSAPLR [b]	B5B0D4	β-lactoglobulin	6	0	semi-specific
HHIELRWK	E1BFN5	Uncharacterized protein	9	8	trypsin
QKYGVVKENVIDLTK	E1BJP1, G3MZU3	Uncharacterized proteins	0	9	semi-specific
EKESLGWQK	E1BKT9	Desmoplakin	0	2	unspecific
EHLYQENQYLEQENTQ	E1BMB1	Ninein	0	6	unspecific
QEELENRTSETNTPQGNQEY	E1BMB1	Ninein	8	3	unspecific
HEQGMDQDKN	F1MV51	APC, WNT signaling pathway regulator	10	10	unspecific
SSLSDIDQENNNNK	F1MV51	APC, WNT signaling pathway regulator	2	3	unspecific
TLQIAEIKDNSGPRSNED	F1MV51	APC, WNT signaling pathway regulator	0	2	unspecific
QNLAFVSMLNDIAAP	F1N647	Fatty acid synthase	0	1	unspecific
IQQNSSTTEKI	F2FB38	Mucin-16	6	9	unspecific
KFNITDTLMQ	F2FB38	Mucin-16	0	1	unspecific
LDQWLCEKL [b]	P00711	α-lactalbumin	4	0	trypsin
NICNISCDKFLDD	P00711	α-lactalbumin	0	1	unspecific
EKVNELSK [a]	P02662	α_{s1}-casein	7	1	trypsin

Table 2.2: *(Continued)* All identified non-human peptide sequences that were assigned to bovine proteins, with the corresponding UniProt protein id, name of the leading protein, and *in silico* digestion mode. The number of samples in which the peptide sequence was identified is listed per group of allergic and non-allergic mothers.

Sequence	Leading proteins	Protein names	Allergic	Non-allergic	Digestion
FFVAPFPEVFGK [a]	P02662	α_{s1}-casein	2	3	trypsin
HIQKEDVPSER [a]	P02662	α_{s1}-casein	10	8	trypsin
HQGLPQEVLNENLLR [a]	P02662	α_{s1}-casein	5	8	trypsin
YLGYLEQLLR [a]	P02662	α_{s1}-casein	2	5	trypsin
SCQAQPTTMAR [b]	P02668	κ-casein	9	3	trypsin
AEFVEVTK [a]	P02769	Serum albumin	7	10	trypsin
DAFLGSFLYEYSR [a]	P02769	Serum albumin	6	4	trypsin
DLGEEHFK [b]	P02769	Serum albumin	0	9	trypsin
DTHKSEIAHR [a]	P02769	Serum albumin	0	10	trypsin
DVCKNYQEAK [b]	P02769	Serum albumin	10	10	trypsin
FKDLGEEHFK [a]	P02769	Serum albumin	10	10	trypsin
HLVDEPQNLIK [a]	P02769	Serum albumin	4	9	trypsin
LVNELTEFAK [a]	P02769	Serum albumin	7	10	trypsin
QNCDQFEK [b]	P02769	Serum albumin	0	5	trypsin
RHPEYAVSVLLR [a]	P02769	Serum albumin	7	10	trypsin
SLHTLFGDELCK [b]	P02769	Serum albumin	1	8	trypsin
TCVADESHAGCEK [b]	P02769	Serum albumin	2	7	trypsin
GKYLYEIAR	P02769	Serum albumin	9	10	semi-specific
KQTALVELLK [b]	P02769	Serum albumin	2	5	unspecific
IKVMNDLSPKSNLR	P07353	Interferon gamma	2	1	semi-specific
DLKLVEQQNPK	P08037	β-1,4-galactosyltransferase 1	0	2	semi-specific
AQFVPLPVSVSVEFAVAATDCIAK [b]	P12763	α_2-HS-glycoprotein	9	0	trypsin
VNLLVDRQWQAVRNR	P15396	Ectonucleotide pyrophosphatase	10	10	trypsin
KLLNNITNDLR	P21758	Macrophage scavenger receptor	4	0	unspecific
NLLFNDNTECLAK [b]	P24627	Lactotransferrin	4	1	trypsin
NKHSNLIESQENSK	P31098, P31096	Osteopontin-K, Osteopontin	9	7	trypsin
NVTRQAYWQIHMDQ	P80209	Cathepsin D	0	3	unspecific
NGNNPNCCMNQK	P80457	Xanthine dehydrogenase/oxidase	1	0	semi-specific

Table 2.2: *(Continued)* All identified non-human peptide sequences that were assigned to bovine proteins, with the corresponding UniProt protein id, name of the leading protein, and *in silico* digestion mode. The number of samples in which the peptide sequence was identified is listed per group of allergic and non-allergic mothers.

Sequence	Leading proteins	Protein names	Allergic	Non-allergic	Digestion
EKQLPNGDWPQENISGVFNKSCA	P84466	Lanosterol synthase	5	3	unspecific
VSITCSGSSSNIGR [b]	Q1RMN8	Immunoglobulin light chain	8	5	trypsin
CASFRENVLR [b]	Q29443	Serotransferrin	10	10	trypsin
QMERALLENE	Q2HJ49	Moesin	0	3	semi-specific
NGEGQVLFETEISR	Q2TBX4	Heat shock 70 kDa protein 13	3	8	trypsin
NIIKSGSDEVQ	Q2UVX4	Complement C3	1	0	unspecific
VALNKLK	Q58D55	β-galactosidase	2	0	trypsin
VYVEQLKPTPEGDLEILLQK	Q9BDG3	β-lactoglobulin D	1	0	trypsin

[a] Confirmed by analysis of digested pure protein.

[b] Confirmed by analysis of bovine milk serum proteins.

Table 2.3: All identified non-human peptide sequences that were assigned to non-bovine allergens, with the corresponding UniProt protein id, leading organism and *in silico* digestion mode. The number of samples in which the peptide sequence was identified is listed per group of allergic and non-allergic mothers.

Sequence	Leading proteins	Leading organisms or LCA[b]	Allergic	Non-allergic	Digestion
QNWASLQPYKKL	Q08169, A0A0M9A8V0, I1VC83, A0A2A3EHG0, Q95PD7, A0A0L7RCK4, A0A310SIY9	Apidae (family) (bees)	1	2	semi-specific
RPSHQQPR	P43237, N1NEW2	Arachis (genus) (legumes)	6	4	trypsin
MQDQLDQVQK	Q8MUF6, Q9BMM8, A0A1B2YLJ8	Astigmatina (cohort) (mites)	1	5	unspecific
KELKKKVEADGEND	A0A2V1CGL9	*Cadophora* sp. DSE1049	6	4	unspecific
QIANSDEVEKI	Q24702	*Dictyocaulus viviparus*	3	6	unspecific
KCAADESAENCDK	P35747	*Equus caballus*	7	3	trypsin
LVNEVTEFAKK [a]	P35747	*Equus caballus*	10	8	trypsin
KEPERNECFLQHK [a]	P49064	*Felis catus*	5	8	trypsin
PCFSALQVDETYVPK	P49064	*Felis catus*	1	0	trypsin
YICENQDSISTK	P49064	*Felis catus*	0	5	trypsin
SALQVDETYVPK	P49064	*Felis catus*	3	4	semi-specific
KEQVARFTAGTNPK	A9QQ26	*Lycosa singoriensis*	10	10	trypsin
EQVQELR	A0A1L8GUE3, A0A3Q0GE46, A0A151P804	Tetrapoda (superclass) (4-limbed vertebrates)	2	2	trypsin
QQQTLQQILQQQ	P04723	*Triticum aestivum*	10	10	unspecific
QVLQQSSYQQLQQ	P04723	*Triticum aestivum*	0	2	unspecific
QFKPEEMTNIIK	P35083, A4KA55	*Zea mays*	8	4	semi-specific

[a] Confirmed by analysis of acquired synthesized peptide.

[b] Last common ancestor (LCA) in case of multiple leading organisms.

Another possible source of false-positive identifications could be the presence of unknown human protein variants due to point mutations. Out of the 78 identified non-human peptide sequences, 26 have one amino acid different from their human homologue. As an example, the sequence LVNELTEFAK with *Bos taurus* as leading organism has LVNEVTEFAK as homologue in *Homo sapiens*. The V → L could therefore be the result of a point mutation. Nevertheless, for all these 26 sequences, no research was found that confirmed the occurrence of these point mutations in *Homo sapiens*.

MS/MS spectra of the identified peptides and their confirmation can be found in the Supplementary information (Supplementary Figures S2.1 - S2.38).

2.3.2 Differences between allergic and non-allergic mothers

Out of the 78 non-human peptide sequences, 15 sequences were only identified in milk from allergic mothers, whereas in milk from non-allergic mothers, 11 unique sequences were identified. This difference can be largely attributed to sequences that match to bovine proteins (Table 2.2). After imputation of the LFQ data and performing a *t*-test with maternal allergy as grouping variable, 16 peptide sequences appeared to be significantly different in intensity between the two groups (Figure 2.2).

As shown in Figure 2.2, nine sequences were found to be significantly higher in intensity in milk from allergic mothers. These sequences were annotated to BLG (n = 8) and α_2-HS-glycoprotein ($n = 1$) as leading protein. The seven sequences that were significantly higher in intensity in milk from non-allergic mothers were annotated to BSA ($n = 6$) and to an uncharacterized protein ($n = 1$), with semi-specific trypsin digestion. All the significantly different sequences were annotated to proteins that originate from *Bos taurus*. As can be seen in Figure 2.3, there is a consistent difference between the two groups, indicating that the significant differences are not caused by outliers.

2.4 Discussion

The goal of this study was to identify non-human proteinaceous molecules in human milk and to investigate differences in these molecules between milk from allergic and non-allergic mothers. Out of the 78 resulting non-human peptide sequences identified in this study, 11 sequences were reported previously in human milk studies using LC-MS/MS [2, 5]. Contrary to these studies, we focused on milk serum, discarding the caseins by ultracentrifugation. This could explain the major difference with Zhu et al. [2] when it comes to the number of identified sequences matching with bovine caseins. The relatively high levels of BLG peptide sequences that we found in milk from allergic mothers explains the high sequence coverage of BLG in the current study when compared to Zhu et al. [2]. Because we removed small peptides by filter-aided sample preparation (10-20 kDa cutoff), no comparison could be made with previous peptidomics studies [3, 4]. Other qualitative differences with these previous studies can be attributed to the stricter filtering on false positives that we applied, the inclusion of serum albumins, the inclusion of semi-trypsin and non-trypsin-digested sequences, and to the inclusion of milk from allergic mothers.

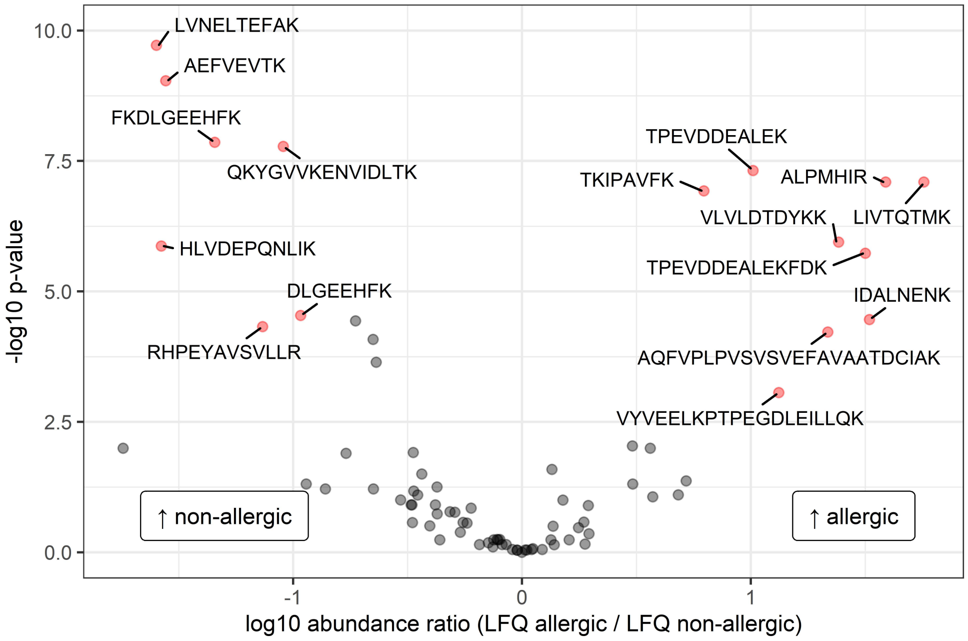


Figure 2.2: Volcano plot with the ratios of the group means of the log10 transformed LFQ intensities of the identified peptide sequences. Significantly different peptides (false discovery rate <0.01 and difference between groups > ± 0.75) are represented by filled red circles and labelled with the corresponding amino acid sequence. On the right side of the plot, the peptides with a higher level in allergic mothers are presented, and on the left side the peptides with a higher level in non-allergic mothers.

The transfer of proteinaceous molecules from the mother's intestinal tract to the mammary gland is still poorly understood, especially when it comes to intact proteins or large protein fragments. In the current study, trypsin was used to digest the

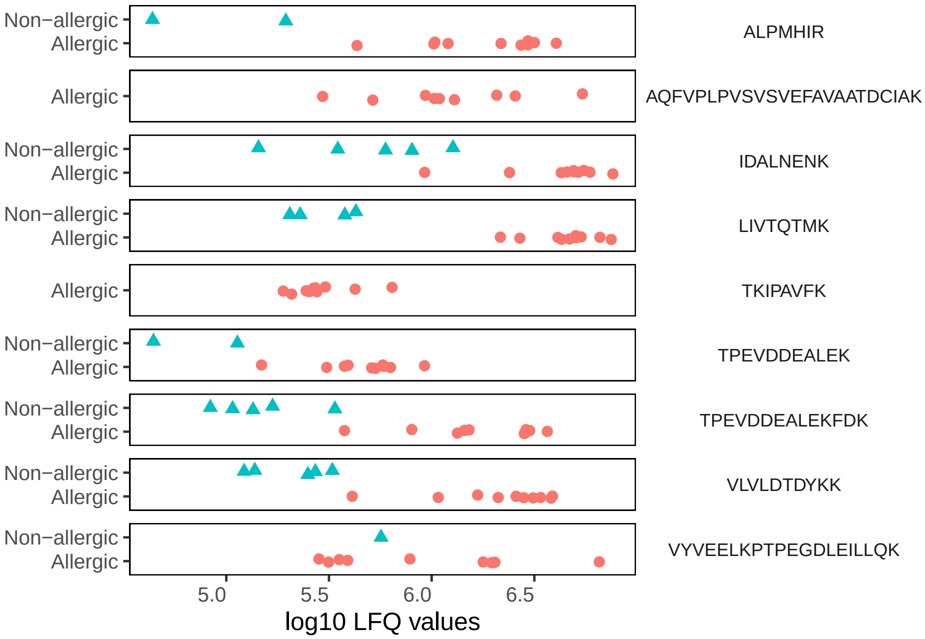


Figure 2.3: Categorical scatterplot showing non-imputed, log10 transformed LFQ intensities of peptide sequences that were found in significantly higher levels in milk from allergic mothers. Allergic mothers are represented by red circles and non-allergic mothers by blue triangles.

proteins before analysis. It can be expected that the majority of the identified peptides was digested by trypsin. Nevertheless, some peptide sequences were identified that were not, or partly, digested by trypsin. This might indicate that there are also non-human protein fragments present in human milk that were digested by other proteases than trypsin, probably before these fragments even entered the milk.

The high sequence coverage for BLG, with all but one sequence digested by trypsin, is an indication for the presence of intact BLG in human milk. This is supported by the findings of Zhu et al. [2], who, although with a lower sequence coverage, also identified BLG in human milk. In addition, several other studies reported the identification of intact BLG in human milk using immunochemical analysis [2, 26, 27].

The presence of intact BLG may be due to its relatively small size and its high resistance to pepsin digestion, as previously shown by the presence of intact BLG in jejunum samples [28]. From our results, it appears that especially proteinaceous molecules from bovine milk end up in human milk. This might relate to the high consumption rate of dairy products in the Netherlands, considering that 23% of the average dietary protein intake originates from milk and dairy products [29].

In line with this, the highly abundant bovine milk serum proteins (BLG and BSA) were identified with the highest sequence coverage. Bovine ALA, another major milk serum protein, was identified with only two peptide sequences. This low sequence coverage was expected because of its high digestibility and high homology with human ALA. This high homology reduces the number of unique peptides that can be identified from this protein. Besides the bovine peptide sequences, peptide sequences of cat and horse serum albumin were identified and confirmed by the analysis of synthesized peptides. No relation was found between the presence of cat serum albumin peptides in milk and the ownership of a cat as pet by the respective mothers. Nevertheless, it is known that mammalian serum albumins are present in animal dander and exposure to this is not limited to direct contact [30, 31]. The serum albumins of cat or horse could end up in the human digestive system by ingestion or inhalation and could subsequently be transferred to the milk. Whether these proteins are present as intact proteins or in large fragments remains unclear because of the relatively low sequence coverage that was found.

Several other studies reported the detection of other dietary allergen proteins in human milk, such as egg ovalbumin, peanut allergen, and wheat gliadin [6, 9, 13]. Peptide sequences of these three proteins were initially detected in the current study but were filtered out due to a low PSM score or to not being confirmed by analysis of synthesized peptides. This could still mean that these proteins are present in human milk but in too low concentrations for positive identification.

Several previous studies have investigated a possible difference in non-human proteins between milk from allergic (maternal history of atopic diseases) and non-allergic mothers. Høst et al. [26] and, more recently, Matangkasombut et al. [32] did not find a difference in BLG levels in milk between the two groups. Another study, also investigating BLG levels, found BLG in the milk of all allergic subjects involved and not in the milk of all non-allergic subjects [27]. Sorva et al. [33] found that BLG levels in milk of allergic and non-allergic mothers were similar after 24 h on a milk-free diet. Nevertheless, the levels of BLG tended to be higher in milk from allergic mothers one hour after consumption of 400 mL bovine milk. Surprisingly, the current study shows a significant difference concerning peptide sequences originating from BLG and BSA. A similar finding has not been reported before.

The difference between our results and the aforementioned studies can possibly be explained by the characteristics of the allergic subjects and a difference in methodology. The current study has, for example, a strict selection on both HDM-specific IgE and allergy symptoms, whereas previous studies did not elaborate on the definition they used for atopy or allergy, and in some cases, the selection of allergic subjects was based on symptoms only.

In addition, immunochemical analyses have shown to be influenced by cross-reactivity between human and bovine proteins, which make them less reliable than LC-MS/MS analysis [34, 35]. For the current study, only an indication of the frequency of consumption for milk and milk products was available (Table 2.1). Nevertheless, it seems that the allergic mothers even consumed less milk or milk products when compared to the non-allergic mothers. Therefore, our findings indicate that there is a difference between the allergic and non-allergic mothers when it comes to the transfer of bovine proteinaceous molecules from the intestinal tract to the mammary gland.

From the intestinal tract to the blood, proteins can be absorbed by both paracellular and transcellular pathways. In reviewing the literature, Reitsma et al. [10] suggested that a difference in intestinal absorption of proteins between non-sensitized and sensitized persons can take place by both pathways. Which of these two is involved in the transfer of dietary proteins into human milk is not known. One option for a transcellular pathway concerns transport of intact antigens with specific IgE via the CD23 receptor [36]. With regard to the current study, this would suggest an increased level of BLG-specific IgE in HDM allergic mothers, which seems unlikely and has not been mentioned in literature.

Another transcellular pathway is through enterocytes and involves degradation of the protein in lysosomes [37]. A recent study using Caco-2 cell monolayers showed that casein fragments survive transfer by this pathway but that BLG seems to be completely degraded [38]. Therefore, this pathway seems unlikely. A third transcellular pathway is via M cells, and it has been suggested that BLG can be transferred through this pathway without degradation [39]. Nevertheless, there is no evidence that transport through these pathways is increased in allergic mothers. A prerequisite for the uptake of proteinaceous molecules through the paracellular pathway is an impaired intestinal barrier. Reitsma et al. [10] pointed out that sensitized persons have an increased level of mast cells that release IgE-induced tryptase.

The tryptase affects the tight junctions in the intestinal barrier, leading to an increased permeability, which could allow the passage of proteinaceous molecules. Nevertheless, this pathway is linked to food hypersensitivity reactions where the location of sensitization is in the intestinal tract itself. In patients with HDM allergy, sensitization occurs primarily in the respiratory tract. However, Calderón et al. [40] suggested that HDM sensitization can be systemic and could cause reactions in other parts of the body. Such a lung-gut crosstalk is plausible, considering the evidence showing that the mucosal immune system can be considered as a system-wide organ [41]. In reviewing the literature, Zhu et al. [42] suggested, in line with this hypothesis, the role of a thymic stromal lymphopoietin-mediated pathway that is induced by

HDM allergen sensitization, which might promote the breakdown of the epithelial barrier in the intestinal tract.

Several other factors could be involved in the disruption of the intestinal barrier. Firstly, it has been shown that, independent of atopy, asthma can be associated with an increased intestinal barrier permeability [14, 43]. In the current study, seven out of 10 of the allergic mothers reported asthma (Table 2.1). Secondly, Tulic et al. [44] showed that HDM is often present in the gut and that its cysteine protease Der p 1 causes disruption of the epithelial barrier. This disruption appeared to be similar for HDM-sensitized and HDM-non-sensitized subjects. Nevertheless, due to an inflammatory response, it might be possible that recovery of the intestinal barrier dysfunction is delayed or incomplete in allergic subjects. This might then explain the permeability of the intestinal barrier in the allergic subjects in the current study, the majority of whom have HDM allergy. It should be noted that the majority of the *in vivo* studies on intestinal permeability make use of small inert molecules and their passage through the intestine. It has been shown that an increased transfer of these molecules through the intestinal barrier does not necessarily correlate with the transfer of antigens [45]. In addition, a previous study showed, using ELISA, that levels of BLG in human milk were not related to intestinal barrier permeability [33]. Therefore, more research is needed to elucidate whether the increased barrier permeability caused by these factors indeed leads to an increased passage of proteinaceous molecules.

After passage through the intestinal barrier, it is expected that the non-human proteinaceous molecules enter the blood stream and are subsequently transferred through the mammary epithelium into the alveolar lumen. This transfer seems to take place through a one-way transcytotic pathway. Monks et al. [12] showed the role of this pathway in the transfer of extracellular serum albumin in mice and suggested that this is the same pathway that is involved in the transfer of IgA. After transfer to the milk, the non-human proteinaceous molecules end up in the digestive system of the infant. Worth noting is that Hettinga et al. [19], in analyzing the same data but focusing on human proteins, found significantly higher levels of several protease inhibitors in human milk from allergic mothers. These protease inhibitors (cystatin C, inter-α-trypsin inhibitors, and serine-protease inhibitors) are potentially active against enzymes that hydrolyze BLG. Consequently, the human milk composition of allergic mothers might reduce the hydrolysis of BLG in the infant's intestinal tract.

Since the current study does not include absolute quantification, the exact level of these molecules in human milk remains unclear. Regardless of their level, it is known that bovine milk proteins in human milk can have an effect on the infant. Several reported cases described non-IgE-mediated food protein-induced enterocolitis syndrome caused by bovine milk proteins in exclusively breastfed infants [46, 47]. In all these cases, the infant had a positive family history for atopy and clinical manifestations were resolved after the mother strictly eliminated cow's milk from her diet. It remains unclear whether non-human proteinaceous molecules in human milk can have an effect on the development of the immune system of the breastfed infant beyond causing allergic symptoms. Verhasselt et al. [48] showed, using a mouse model, that antigen transfer through breastmilk induced tolerance and protection from allergic asthma. Translating this to BLG, it is known that BLG-derived peptides can be HLA-DRB1-restricting, a characteristic that might support oral tolerance development [49]. In line with this, Peters et al. [50] showed recently that early introduction of cow's milk was associated with a reduced risk of cow's milk allergy. The presence of higher levels of BLG or its derived peptides in human milk of allergic mothers might therefore have a protective effect on further allergy development. Nevertheless, evidence remains speculative, and a direct relation needs to be investigated.

Interestingly, BSA peptide sequences were found in significantly lower levels in milk from allergic mothers. Previous research with rats showed that intact BSA can pass the intestinal epithelium [51]. Nevertheless, the difference found between the two groups is difficult to interpret. The most likely but speculative explanation is a specific pathway that is activated in healthy mothers but that is negatively regulated in allergic mothers.

2.5 Conclusions

In conclusion, in the present study, a significant difference in levels of non-human proteinaceous molecules in human milk of allergic and non-allergic mothers has been observed. Sequences from BLG appeared in higher levels and sequences from BSA in lower levels in milk from allergic mothers when compared to milk from non-allergic mothers. These findings suggest that there is a difference in transfer of proteinaceous molecules through the intestinal barrier of allergic mothers, allowing dietary proteins to enter the bloodstream and ultimately the milk. This study has raised important questions about the role that these proteinaceous molecules might play in the development of the immune system of infants.

Supplementary information

The following supplementary information is available and can be accessed through the QR code in Figure 2.4: Supplementary Figures S2.1-S2.38: MS/MS spectra from all non-human peptide sequences identified and confirmed in this study. Supplementary information, fasta database: fasta database containing all protein sequences and identifiers used for the data analysis in this study.

Figure 2.4: Scan this QR code to access the supplementary information, or visit https://figshare.com/s/a9f79bed463de3e8157c.

References

[1] Kilshaw, P. J. and Cant, A. J. "The Passage of Maternal Dietary Proteins into Human Breast Milk". In: *International Archives of Allergy and Applied Immunology* 75 (1984), 8–15. DOI: 10.1159/000233582.

[2] Zhu, J. et al. "Discovery and Quantification of Nonhuman Proteins in Human Milk". In: *Journal of Proteome Research* 18 (2019), 225–238. DOI: 10.1021/acs.jproteome.8b00550.

[3] Picariello, G. et al. "Antibody-Independent Identification of Bovine Milk-Derived Peptides in Breast-Milk". In: *Food and Function* 7 (2016), 3402–3409. DOI: 10.1039/C6FO00731G.

[4] Picariello, G. et al. "Excretion of Dietary Cow's Milk Derived Peptides into Breast Milk". In: *Frontiers in Nutrition* 6 (2019), 25. DOI: 10.3389/fnut.2019.00025.

[5] Coscia, A. et al. "Detection of Cow's Milk Proteins and Minor Components in Human Milk Using Proteomics Techniques". In: *The Journal of Maternal-Fetal & Neonatal Medicine* 25 (2012), 49–51. DOI: 10.3109/14767058.2012.715015.

[6] Schocker, F. et al. "Prospective Investigation on the Transfer of Ara h 2, the Most Potent Peanut Allergen, in Human Breast Milk". In: *Pediatric Allergy and Immunology* 27 (2016), 348–355. DOI: 10.1111/pai.12533.

[7] Pastor-Vargas, C. et al. "Sensitive Detection of Major Food Allergens in Breast Milk: First Gateway for Allergenic Contact during Breastfeeding". In: *Allergy: European Journal of Allergy and Clinical Immunology* 70 (2015), 1024–1027. DOI: 10.1111/all.12646.

[8] Macchiaverni, P. et al. "Respiratory Allergen from House Dust Mite Is Present in Human Milk and Primes for Allergic Sensitization in a Mouse Model of Asthma". In: *Allergy: European Journal of Allergy and Clinical Immunology* 69 (2014), 395–398. DOI: 10.1111/all.12332.

[9] Vance, G. H. et al. "Exposure of the Fetus and Infant to Hens' Egg Ovalbumin via the Placenta and Breast Milk in Relation to Maternal Intake of Dietary Egg". In: *Clinical & Experimental Allergy* 35 (2005), 1318–1326. DOI: 10.1111/j.1365-2222.2005.02346.x.

[10] Reitsma, M. et al. "Protein Transport across the Small Intestine in Food Allergy". In: *Molecular Nutrition and Food Research* 58 (2014), 194–205. DOI: 10.1002/mnfr.201300204.

[11] Benn, C. S. et al. "Mammary Epithelial Paracellular Permeability in Atopic and Non-Atopic Mothers versus Childhood Atopy". In: *Pediatric Allergy and Immunology* 15 (2004), 123–126. DOI: 10.1046/j.1399-3038.2003.00138.x.

[12] Monks, J. and Neville, M. C. "Albumin Transcytosis across the Epithelium of the Lactating Mouse Mammary Gland". In: *Journal of Physiology* 560 (2004), 267–280. DOI: 10.1113/jphysiol.2004.068403.

[13] Chirdo, F. G. et al. "Presence of High Levels of Non-Degraded Gliadin in Breast Milk from Healthy Mothers". In: *Scandinavian Journal of Gastroenterology* 33 (1998), 1186–1192. DOI: 10.1080/00365529850172557.

[14] Benard, A. et al. "Increased Intestinal Permeability in Bronchial Asthma". In: *Journal of Allergy and Clinical Immunology* 97 (1996), 1173–1178. DOI: 10.1016/S0091-6749(96)70181-1.

[15] Caffarelli, C. et al. "Elimination Diet and Intestinal Permeability in Atopic Eczema: A Preliminary Study". In: *Clinical & Experimental Allergy* 23 (1993), 28–31. DOI: 10.1111/j.1365-2222.1993.tb02480.x.

[16] Majamaa, H. and Isolauri, E. "Evaluation of the Gut Mucosal Barrier: Evidence for Increased Antigen Transfer in Children with Atopic Eczema". In: *Journal of Allergy and Clinical Immunology* 97 (1996), 985–990. DOI: 10.1016/S0091-6749(96)80074-1.

[17] Brunekreef, B. et al. "The Prevention and Incidence of Asthma and Mite Allergy (PIAMA) Birth Cohort Study: Design and First Results". In: *Pediatric Allergy and Immunology* 13 (2003), 55–60. DOI: 10.1034/j.1399-3038.13.s.15.1.x.

[18] Wijga, A. H. et al. "Cohort Profile: The Prevention and Incidence of Asthma and Mite Allergy (PIAMA) Birth Cohort". In: *International Journal of Epidemiology* 43 (2014), 527–535. DOI: 10.1093/ije/dys231.

[19] Hettinga, K. A. et al. "Difference in the Breast Milk Proteome between Allergic and Non-Allergic Mothers". In: *PLoS ONE* 10 (2015). Ed. by A. S. Wiley, e0122234. DOI: 10.1371/journal.pone.0122234.

[20] Cox, J. and Mann, M. "MaxQuant Enables High Peptide Identification Rates, Individualized p.p.b.-Range Mass Accuracies and Proteome-Wide Protein Quantification". In: *Nature Biotechnology* 26 (2008), 1367–1372. DOI: 10.1038/nbt.1511.

[21] Verheggen, K. et al. "Anatomy and Evolution of Database Search Engines—a Central Component of Mass Spectrometry Based Proteomic Workflows". In: *Mass Spectrometry Reviews* 39 (2020), 292–306. DOI: 10.1002/mas.21543.

[22] D'Alessandro, A., Zolla, L., and Scaloni, A. "The Bovine Milk Proteome: Cherishing, Nourishing and Fostering Molecular Complexity. An Interactomics and Functional Overview". In: *Molecular BioSystems* 7 (2011), 579–597. doi: 10.1039/c0mb00027b.

[23] D'Alessandro, A., Scaloni, A., and Zolla, L. "Human Milk Proteins: An Interactomics and Updated Functional Overview". In: *Journal of Proteome Research* 9 (2010), 3339–3373. doi: 10.1021/pr100123f.

[24] Chen, C. et al. "A Fast Peptide Match Service for UniProt Knowledgebase". In: *Bioinformatics* 29 (2013), 2808–2809. doi: 10.1093/bioinformatics/btt484.

[25] Development Team Core. *R. A Language and Environment for Statistical Computing*. 2020.

[26] Høst, A. et al. "Bovine β-1actoglobulin in Human Milk from Atopic and Non-Atopic Mothers. Relationship to Maternal Intake of Homogenized and Unhomogenized Milk". In: *Clinical & Experimental Allergy* 20 (1990), 383–387. doi: 10.1111/j.1365-2222.1990.tb02798.x.

[27] Axelsson, I. et al. "Bovine β-Lactoglobulin in the Human Milk. A Longitudinal Study during the Whole Lactation Period". In: *Acta Paediatrica Scandinavica* 75 (1986), 702–707. doi: 10.1111/j.1651-2227.1986.tb10277.x.

[28] Sanchón, J. et al. "Protein Degradation and Peptide Release from Milk Proteins in Human Jejunum. Comparison with in Vitro Gastrointestinal Simulation". In: *Food Chemistry* 239 (2018), 486–494. doi: 10.1016/j.foodchem.2017.06.134.

[29] European Food Safety Authority. "Scientific Opinion on Dietary Reference Values for Protein". In: *EFSA Panel on Dietetic Products, Nutrition and Allergies* (2012). doi: 10.2903/j.efsa.2012.2557.

[30] Liccardi, G. et al. "Role of Sensitization to Mammalian Serum Albumin in Allergic Disease". In: *Current Allergy and Asthma Reports* 11 (2011), 421–426. doi: 10.1007/s11882-011-0214-7.

[31] Zahradnik, E. and Raulf, M. "Animal Allergens and Their Presence in the Environment". In: *Frontiers in Immunology* 5 (2014), 76. doi: 10.3389/fimmu.2014.00076.

[32] Matangkasombut, P. et al. "Detection of β-Lactoglobulin in Human Breast-Milk 7 Days after Cow Milk Ingestion". In: *Paediatrics and International Child Health* 37 (2017), 199–203. doi: 10.1080/20469047.2017.1289310.

[33] Sorva, R., Mäkinen-Kiljunen, S., and Juntunen-Backman, K. "β-Lactoglobulin Secretion in Human Milk Varies Widely after Cow's Milk Ingestion in Mothers of Infants with Cow's Milk Allergy". In: *The Journal of Allergy and Clinical Immunology* 93 (1994), 787–792. DOI: 10.1016/0091-6749(94)90259-3.

[34] Bertino, E. et al. "Absence in Human Milk of Bovine Beta-Lactoglobulin Ingested by the Mother. Unreliability of ELISA Measurements." In: *Acta Biomedica de Ateneo Parmense* 68 Suppl 1 (1997), 15–9.

[35] Restani, P. et al. "Evaluation of the Presence of Bovine Proteins in Human Milk as a Possible Cause of Allergic Symptoms in Breast-Fed Children". In: *Annals of Allergy, Asthma and Immunology* 84 (2000), 353–360. DOI: 10.1016/S1081-1206(10)62786-X.

[36] Bevilacqua, C. et al. "Food Allergens Are Protected from Degradation during CD23-mediated Transepithelial Transport". In: *International Archives of Allergy and Immunology* 135 (2004), 108–116. DOI: 10.1159/000080653.

[37] Caillard, I. and Tomé, D. "Transport of β-Lactoglobulin and α-Lactalbumin in Enterocyte-like Caco-2 Cells". In: *Reproduction, Nutrition, Development* 35 (1995), 179–188. DOI: 10.1016/0926-5287(96)80190-0.

[38] Sakurai, N. et al. "Apical-to-Basolateral Transepithelial Transport of Cow's Milk Caseins by Intestinal Caco-2 Cell Monolayers: MS-based Quantitation of Cellularly Degraded a- and b-Casein Fragments". In: *Journal of Biochemistry* 164 (2018), 113–125. DOI: 10.1093/jb/mvy034.

[39] Rytkönen, J. et al. "Enterocyte and M-cell Transport of Native and Heat-Denatured Bovine β-Lactoglobulin: Significance of Heat Denaturation". In: *Journal of Agricultural and Food Chemistry* 54 (2006), 1500–1507. DOI: 10.1021/jf052309d.

[40] Calderón, M. A. et al. "Respiratory Allergy Caused by House Dust Mites: What Do We Really Know?" In: *Journal of Allergy and Clinical Immunology* 136 (2015), 38–48. DOI: 10.1016/j.jaci.2014.10.012.

[41] Gill, N., Wlodarska, M., and Finlay, B. B. "The Future of Mucosal Immunology: Studying an Integrated System-Wide Organ". In: *Nature Immunology* 11 (2010), 558–560. DOI: 10.1038/ni0710-558.

[42] Zhu, T. H. et al. "Epithelial Barrier Dysfunctions in Atopic Dermatitis: A Skin–Gut–Lung Model Linking Microbiome Alteration and Immune Dysregulation". In: *British Journal of Dermatology* 179 (2018), 570–581. DOI: 10.1111/bjd.16734.

[43] Hijazi, Z. et al. “Intestinal Permeability Is Increased in Bronchial Asthma”. In: *Archives of Disease in Childhood* 89 (2004), 227–229. doi: 10.1136/adc.2003.027680.

[44] Tulic, M. K. et al. “Presence of Commensal House Dust Mite Allergen in Human Gastrointestinal Tract: A Potential Contributor to Intestinal Barrier Dysfunction”. In: *Gut* 65 (2016), 757–766. doi: 10.1136/gutjnl-2015-310523.

[45] Ménard, S., Cerf-Bensussan, N., and Heyman, M. “Multiple Facets of Intestinal Permeability and Epithelial Handling of Dietary Antigens”. In: *Mucosal Immunology* 3 (2010), 247–259. doi: 10.1038/mi.2010.5.

[46] Sopo, S. M. et al. “Chronic Food Protein-Induced Enterocolitis Syndrome Caused by Cow’s Milk Proteins Passed through Breast Milk”. In: *International Archives of Allergy and Immunology* 164 (2014), 207–209. doi: 10.1159/000365104.

[47] Vergara Perez, I. and Vila Sexto, L. “Suspected Severe Acute Food Protein–Induced Enterocolitis Syndrome Caused by Cow’s Milk through Breast Milk”. In: *Annals of Allergy, Asthma and Immunology* 121 (2018), 245–246. doi: 10.1016/j.anai.2018.04.023.

[48] Verhasselt, V. et al. “Breast Milk-Mediated Transfer of an Antigen Induces Tolerance and Protection from Allergic Asthma”. In: *Nature Medicine* 14 (2008), 170–175. doi: 10.1038/nm1718.

[49] Gouw, J. W. et al. “Identification of Peptides with Tolerogenic Potential in a Hydrolysed Whey-Based Infant Formula”. In: *Clinical & Experimental Allergy* 48 (2018), 1345–1353. doi: 10.1111/cea.13223.

[50] Peters, R. L. et al. “Early Exposure to Cow’s Milk Protein Is Associated with a Reduced Risk of Cow’s Milk Allergic Outcomes”. In: *Journal of Allergy and Clinical Immunology: In Practice* 7 (2019), 462–470.e1. doi: 10.1016/j.jaip.2018.08.038.

[51] Warshaw, A. L., Walker, W. A., and Isselbacher, K. J. “Protein Uptake by the Intestine: Evidence for Absorption of Intact Macromolecules”. In: *Gastroenterology* 66 (1974), 987–992. doi: 10.1016/S0016-5085(74)80174-5.

Chapter 3

Exploring human milk dynamics: inter-individual variation in milk proteome, peptidome and metabolome

This chapter is based on:

Pieter M. Dekker, Sjef Boeren, Johannes B. van Goudoever, Jacques J.M. Vervoort, Kasper A. Hettinga. Exploring human milk dynamics: inter-individual variation in milk proteome, peptidome and metabolome.
Published in: Journal of Proteome Research 21 (4), 1002-1016 (2022).
DOI: 10.1021/acs.jproteome.1c00879

Abstract

Human milk is a dynamic biofluid, and its detailed composition receives increasing attention. While most studies focus on changes over time or differences between maternal characteristics, interindividual variation receives little attention. Nevertheless, a comprehensive insight into this can help interpret human milk studies and help human milk banks provide targeted milk for recipients. This study aimed to map interindividual variation in the human milk proteome, peptidome, and metabolome and to investigate possible explanations for this variation. A set of 286 milk samples was collected from 29 mothers in the third month postpartum. Samples were pooled per mother, and proteins, peptides, and metabolites were analyzed. A substantial coefficient of variation (>100%) was observed for 4.6% and 36.2% of the proteins and peptides, respectively. In addition, using weighted correlation analysis (WGCNA), 5 protein and 11 peptide clusters were obtained, showing distinct characteristics. With this, several associations were found between the different data sets and with specific sample characteristics. This study provides insight into the dynamics of human milk protein, peptide, and metabolite composition. In addition, it will support future studies that evaluate the effect size of a parameter of interest by enabling a comparison with natural variability.

3.1 Introduction

Human milk is a dynamic biofluid. Its composition depends on, for example, lactation stage and health status of the mother. Proteins are one of the main constituents of human milk and have been shown to be involved in the growth and the healthy development of the infant. To date, the composition of the human milk proteome is well established. The most recent studies on this have reported up to 1500 proteins in human milk [1]. Part of the proteins in human milk are synthesized in the mammary gland, for instance, caseins and α-lactalbumin. Besides this, a vast number of proteins are transferred into the alveolar lumen from the systemic circulation of the mother [2]. Among these are for example albumin, immunoglobulin G, and even non-human proteins [3, 4].

Already before excretion of the milk, proteolysis of proteins takes place, resulting in the human milk peptidome. This peptidome has been shown to comprise more than 4000 unique peptides [5]. The majority of these peptides originate from the precursor protein β-casein. The fact that β-casein is overrepresented in the human milk peptidome is first of all due to its abundance. Besides this, its open and flexible structure makes it prone to proteolytic digestion. Other proteins that are abundant in milk, such as α-lactalbumin, have a closed and globular structure, resulting in a lower contribution of these precursor proteins to the peptidome. Within the human milk peptidome, a substantial number of peptides was found to be a bioactive peptide itself or to be a precursor for a bioactive peptide [6, 7].

Researchers have pursued evaluation of the presence of biomarkers in the peptidome or proteome, investigating the relation with factors such as breast cancer risk or maternal allergy [8, 9]. Besides this, many recent proteomics and peptidomics studies on human milk have focused on longitudinal variation [10–15]. These studies provide the evidence that the human milk proteome changes over lactation according to functionality, that is, from a direct defense mechanism toward the reinforcement for an independent immune system. So far, however, there has been little discussion about interindividual variation in the human milk proteome and peptidome. In the few studies on this done so far, it was established that variation between individual mothers is greater than longitudinal variation. This was observed to be valid for both the human milk proteome [12] and peptidome [16]. From this, the questions arise: what is the extent of this interindividual variation and what is its origin? In addition, it emphasizes the challenge in investigating relations between composition and other parameters such as maternal characteristics. If the interindividual variation is not considered in those investigations, the relevance of differences found between groups of samples will be hard to interpret and can easily be overestimated.

Besides the importance of mapping the interindividual variation in the human milk proteome and peptidome, it remains a challenge to understand the mechanisms underlying this interindividual variation. Part of this variation might be explained by biological processes in the human body, of which indicators might be found in low molecular weight substances, that is, metabolites. An example of this could be the relation between, for example, free amino acids and protein synthesis. Nevertheless, both proteomics and peptidomics analyses result in hundreds of features, giving rise to a challenge in turning data into biologically relevant information. Synthesis and secretion of milk proteins are regulated by biological pathways, and proteins can function interactively in different biological pathways. Therefore, protein coexpression networks can provide useful information on protein relationships and involvement in biological pathways [17].

In recent years, weighted correlation network analysis (WGCNA) has been used to construct and analyze such coexpression networks in proteomics data [18–20]. Peptides, on the other hand, are intermediates in the proteolytic degradation of proteins. In complex samples, such as human milk, peptides originate from dozens of precursor proteins. Peptide levels can be interdependent due to, for example, partly overlapping sequences (peptide-ladders) or specificity of proteolytic cleavage. Grouping peptides based on correlation in intensities can unveil patterns of proteolytic degradation [21]. In approaching these complex data, WGCNA can be used to identify clusters of associated proteins and peptides. In short, the goals with this WGCNA approach were (1) to elucidate whether interindividual variation was specific for certain biological functions or pathways, (2) to shed light on protein-protein and peptide-peptide associations, (3) to investigate associations of proteins and peptides with sample characteristics, and (4) to investigate whether protein and peptide intensities were associated with metabolite levels.

In the current study, we investigated the variation in human milk proteome, peptidome, and metabolome in pooled human milk samples from 29 healthy mothers taken in the third month of lactation. Longitudinal variation in the human milk proteome is the largest in the first month, where a transition takes place from colostrum to mature milk. In the third month of lactation, it is known that longitudinal variation due to the maturation of the milk has leveled out [22, 23]. This time point was therefore chosen as a representation of mature human milk. Samples were analyzed, and interindividual variation for all three omics analyses were reported. Furthermore, relations between the three omics data sets were studied using WGCNA to find underlying reasons for the interindividual variation.

3.2 Experimental section

3.2.1 Sample material

Human milk samples were obtained from healthy mothers donating breastmilk to the Dutch Human Milk Bank (Amsterdam, The Netherlands). Donating mothers were subjected to a preliminary screening by the milk bank, and the milk was collected according to standardized procedures (http://www.moedermelkbank.nl). Informed consent was provided by all mothers to use remnants of the donated milk for scientific research.

A selection of 298 samples was made, donated by 30 different mothers, in the third month postpartum. Latter criterion was chosen to avoid influence of the large longitudinal variation present in milk in the first weeks postpartum. Subsequently, samples were pooled per mother. One of these pooled samples was removed from our selection due to its distinct peptide profile in combination with a low fat and carbohydrate content. These observations indicate the occurrence of mastitis; consequently, the sample was considered an outlier. After this removal, the sample set comprised 29 pooled samples from a total of 286 milk samples. The number of samples included in the pooled samples ranged from 5 to 16, with a time range from 2 to 28 days. Milk was obtained by manual or pump expression at home and collected in a polypropylene bottle. After collection, samples were stored immediately at -18°C. Samples were picked up from homes and transported in a freezer at -20°C to the milk bank where they were stored at the same temperature. Detailed information on the samples included in this study can be found in Table 3.1. Fat content in the samples was measured by the Dutch Human Milk Bank as described by De Waard et al [24].

3.2.2 Proteomics

Sample preparation

Human milk samples were thawed at 4°C, and skimmed milk was obtained after centrifugation at 1500*g* for 10 min at 10°C. Skimmed milk was then centrifuged at 100,000*g* for 30 min at 30°C. Milk serum was collected, and the serum protein concentration was determined in duplicate with the Pierce bicinchonic acid (BCA) assay (Thermo Scientific, Waltham, MA). According to these results, milk serum samples were diluted in 100 mM Tris to a concentration of 1 μg/μL protein. To a 100 μL diluted milk sample, a final concentration of 15 mM dithiothreitol was added and subsequently incubated at 45°C for 30 min. After disulfide bonds were reduced, the sample was transferred into 6 M urea, and alkylation of the reduced cysteine residues was obtained by addition of 20 mM acrylamide and 10 min incubation at room temperature. From this alkylated protein sample, 180 μL, containing 36 μg of protein, was transferred to a Pall 3K omega filter (10–20 kDa cutoff, OD003C34; Pall, Washington, NY, USA) and the samples were centrifuged at 12,000*g* for 30 min. The filter was washed with a 50 mM ammonium bicarbonate solution. Then 100 μL of 5 ng/μL sequencing grade trypsin was added, and digestion took place overnight under mild shaking at room temperature. The filter with the digested proteins was centrifuged and washed with 100 μL of 1 mL/L formic acid solution. The pH of the final peptide solution was set to around 3 using a 10% trifluoroacetic acid solution.

Table 3.1: Subject demographics and sample characteristics. Classes of body mass index (BMI) are normal (BMI <25), overweight (BMI >25 and <30), and obese (BMI >30).

Infant		
Sex (*n*)	Female	14
	Male	15
Mother		
BMI classification (*n*)	Normal	18
	Overweight	7
	Obese	4
Age (years)	Median	32.4
	Range	26-42.8
Milk samples		
Samples included in pool	Median	9
	Range	5-16
Time (days) between first and last sample in pool	Median	10
	Range	2-28
Total protein concentration (mg/mL)	Mean	9.4
	Range	8.6-10.2

LC-MS/MS

The prepared samples were analyzed with LC-MS/MS as described before [12]. In short, an LTQ-Orbitrap XL system (Thermo electron, San Jose, CA, USA) was used to obtain full scan FTMS spectra in positive mode (*m/z* 380 to 1400). MS/MS scans of the four multiply-charged peaks with the highest intensity were recorded in the linear trap in data-dependent mode and with an MS/MS threshold of 5000.

Data processing (proteins)

The Andromeda search engine of the MaxQuant software v1.6.1.0 was used to analyze the raw LC-MS/MS data [25]. A database (n = 4296) was used comprising the major human and bovine milk proteins as well as allergen proteins. Detailed information on the creation of this database as well as the database itself can be found in a previous study [26]. *In silico* digestion was carried out with trypsin digestion with a maximum of 2 missed cleavages per peptide sequence. Peptide length was set to a minimum of 6 and a maximum of 35 amino acids, and a fixed modification was set to acrylamide on cysteines to account for the alkylation. A false discovery rate (FDR) of 1% was used at the peptide and protein level. Furthermore, a precursor mass tolerance was set to 20 ppm and fragment mass tolerance to 0.5 Da. Recalibration was carried out with a first search using a database with common contaminants.

Further data analysis was carried out, and figures were made using R version 4.0.1 [27]. First, identifications were filtered to exclude matches with the decoy database, potential contaminants, proteins only identified with modified peptides, proteins only identified with one peptide and proteins identified in less than 10 out of 29 samples. For each identified protein group, a leading protein was selected based on peptide count, Swiss-Prot review status, and availability of gene ontology annotation. Label-free quantification (LFQ) intensities were used to analyze the data further and were imputed (described below), transformed with logarithm base 10, and corrected for the dilution factor.

Imputation of missing values was carried out with the Gibbs sampler-based GSimp algorithm, designed for the imputation of left censored missing values [28]. For annotation purposes, a leading protein was selected for protein groups with more than one protein. If a protein group included one or more reviewed proteins, the first reviewed protein was selected as leading protein. If no reviewed protein was included, the protein with the most extensive GO annotation was selected as leading protein.

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE [29] partner repository with the data set identifier PXD028280.

3.2.3 Peptidomics

Sample preparation

Human milk samples were thawed at 4°C, and skimmed milk was obtained after centrifugation at 1500*g* for 10 min at 10°C. Proteins were precipitated by addition of an equal volume of 200 g/L trichloroacetic acid in milli-Q water and subsequent centrifugation at 3000*g* for 10 min at 4°C. From the resulting supernatant, 50 μL was cleaned by solid phase extraction (SPE) on C18+ Stage tip columns (prepared in-house) [30]. Clean-up and elution of the peptides were carried out as described before [6]. Lastly, peptides were reconstituted in 50 μL of 1 mL/L formic acid in water.

LC-MS/MS

Cleaned peptide samples were analyzed using a nanoLC-MS/MS mass spectrometry system (Thermo EASY nLC1000 connected to a Thermo Orbitrap XL) in which the Orbitrap was used to measure both MS and MS/MS scans. A volume of 18 μL of sample was injected onto a 0.10×32 mm Magic C18AQ 200A 5 μm beads (Bruker Nederland B.V.) preconcentration column (prepared in-house) at a constant pressure of 800 bar (normally resulting in a flow of ca. 11 μL/min). Peptides were eluted from the preconcentration column onto a 0.10×250 mm Magic C18AQ 200A 3 μm beads analytical column (prepared in-house), and separation of the peptides took place at a flow rate of 0.5 μL/minute with a gradient of acetonitrile. In 50 min, the gradient increased from 5% to 30% acetonitrile in water with 1 mL/L formic acid, followed by a 3-minute cleaning of the column by a fast increase to 50% acetonitrile. Between preconcentration and analytical column, a P777 Upchurch microcross was positioned, with a stainless-steel needle fitted into the waste line. Using this needle, a 3.5 kV electrospray potential was applied to the eluent. Full scan positive mode FTMS spectra were obtained with the Orbitrap at a resolution of 15,000 and within the range of *m/z* 280 to 1400. For the most abundant doubly and triply charged peaks in the FTMS scans, CID (isolation width 2 *m/z*, 28% normalized collision energy, activation Q 0.25, and activation time 15 ms) MS/MS scans were recorded in data-dependent mode at a resolution of 7500 in the Orbitrap as well (MS/MS threshold 10,000, 45 s exclusion duration).

Data processing (peptides)

Raw LC-MS/MS data files from peptidomics analysis were processed similar to the proteomics data, with some differences. *In silico* digestion was carried out with unspecific digestion settings and a peptide length set to a minimum of 8 and a maximum of 25 amino acids. Variable modifications were set to acetylation of the protein N-term, oxidation of methionine, deamidation of asparagine and glutamine, and phosphorylation of serine and threonine. A maximum of 5 variable modifications were allowed per peptide sequence. LFQ intensities were used to further analyze the data and imputed with the same algorithm as the proteomics data.

Filtering was applied on the MaxQuant output to reduce the number of false positives. Identifications that were removed matched with the decoy database, matched with contaminants, were only identified with a modification, or were identified in less than 10 out of 29 samples. Imputation and selection of a leading protein was carried out using the same approach as for the proteomics data.

The mass spectrometry peptidomics data have been deposited to the ProteomeXchange Consortium via the PRIDE [29] partner repository with the data set identifier PXD028294.

3.2.4 Metabolomics

Sample preparation

Human milk samples were prepared and analyzed with NMR as described in previous studies [31, 32]. In brief, samples were thawed at 4°C and centrifuged for 30 min at a speed of 12,000 rpm (Eppendorf centrifuge 5424, Eppendorf AG, Hamburg, Germany). Next, 500 μL of supernatant was added to 500 μL of deuterated chloroform, and this was thoroughly mixed for 30 min. This mixture was again centrifuged for 15 min at 10,000 rpm. The aqueous top layer was obtained and with an equal volume of phosphate buffer (pH = 7) transferred to a Pall 3K omega filter (10–20 kDa cutoff, OD003C34; Pall, Washington, NY, USA). The filtrate obtained by centrifugation at 10,000 rpm for 30 min was transferred to a 3 mm NMR tube.

NMR analysis

NMR measurements were carried out using a Bruker Avance III NMR spectrometer with a 600 MHz/54 mm UltraShielded Plus magnet. The spectrometer was equipped with a CryoPlatform cryogenic system for cooling, a BCU-05 cooling unit, and with an ATM automatic tuning and matching unit (Bruker Biospin, Rheinstetten, Germany). Samples were measured in ^{1}H NMR tubes of 3 mm (Bruker matching system). One-dimensional nuclear Overhauser effect spectroscopy (NOESY) spectra were obtained at a temperature of 300 K. All obtained spectra were corrected with automatic baseline correction and aligned to the resonance of alanine (1.484 ppm). The Human Metabolome Database version 4 (http://hmdb.ca) and published literature were used for the assignment of metabolites to the spectra [33]. Full details on parameters used for NMR analysis can be found in the Supporting information listing S1.

Data processing (metabolomics)

NMR data were aligned, and the water region was removed. To minimize overlap in the spectra, NMR resonances were specifically integrated by careful selection of peaks. A selection of one NMR resonance was made in case a metabolite was represented by multiple resonances in the NMR spectra. Nonoverlapping peaks were chosen for further data analysis. In case baseline correction resulted in negative intensities, a value of 0.0001 was imposed to replace these. All NMR resonances were scaled to unit variance before correlations were investigated.

3.2.5 Statistical analysis

All statistical analyses were performed using R version 4.0.1 [27]. Interindividual variation was calculated as coefficient of variation (CV), which is also known as relative standard deviation and expressed as percentage.

Weighted Correlation Network Analysis (WGCNA)

To reduce dimensionality and to elucidate patterns of cross-correlation present in the proteomics and peptidomics data, a weighted correlation network analysis was carried out using the WGCNA package for R (version 1.70.3) [34].

With this analysis, a set of clusters was obtained for each data set, where each cluster consists of highly correlating proteins or peptides. Details of WGCNA applied to proteomics data were described by Pei et al [20]. In brief, a correlation matrix was obtained using the biweight midcorrelation measure. From this, a signed and weighted network was created, to which a soft-thresholding power was applied. This soft-thresholding power was chosen based on the approximation of scale-free topology. As shown in the supporting information, a power of 5 was chosen for the proteomics data (see Supplementary Figure S3.1) and a power of 8 for the peptidomics data (Supplementary Figure S3.2). By applying this power, noise was removed, and the strength of correlations was enhanced. After this, topology overlap metrics were calculated from the network and were subjected to hierarchical clustering. Clusters were obtained from the dendrogram using the cutreeDynamic function with a minimum cluster size of 15. The eigenvalues of the clusters (later referred to as eigenproteins and eigenpeptides) were used to investigate relations between the data sets.

Relationships between characteristics of the samples or mothers, eigenproteins, eigenpeptides, and metabolites were assessed using Spearman's rank correlation (denoted with ρ). To calculate statistical significance, the "corPvaluestudent" function from the WGCNA package was used, which provides Student asymptotic p-values.

Gene overrepresentation

To investigate whether protein clusters resulting from the WGCNA were characterized by specific gene ontology (GO) annotations, a GO overrepresentation analysis was carried out using the R package ClusterProfiler, version 3.16.1 [35]. The “enrichGO” function was used in combination with the “compareCluster” function with as background all identified proteins. GO terms were obtained from the org.Hs.eg.db package [36]. On the output of the overrepresentation analysis, the “simplify” function was applied to remove redundant GO annotations. For this, the “Wang” measure, and a similarity cutoff of 0.7, was used. Overrepresentation was visualized using dot plots in which the GO annotations with the top 3 most significant GO terms were shown.

Sequence logos

Sequence logos were created for the P1 and P1’ positions of the peptides’ N- and C-terminal ends. This was done based on both frequency and intensity of the amino acid in the P1 and P1’ position using the R package ggseqlogo, version 0.1 [37].

3.3 Results and Discussion

In this study, the human milk peptidome, proteome, and metabolome of 29 mothers were analyzed (see Figure 3.1). With the resulting data, the interindividual variation in mature human milk was investigated.

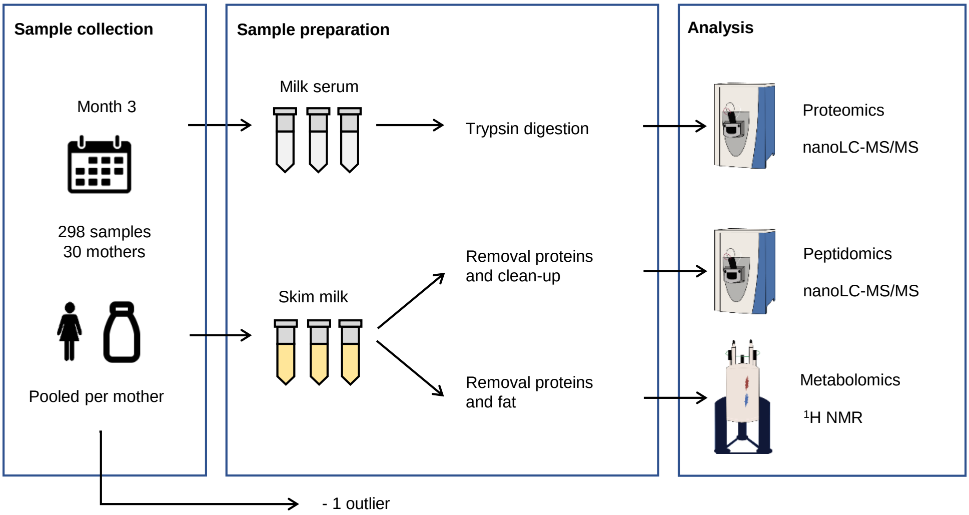


Figure 3.1: Schematic overview of the workflow used for the analysis of the human milk proteome, peptidome, and metabolome.

3.3.1 Proteomics

After analysis and subsequent filtering, 237 proteins were identified and quantified with label-free quantification (LFQ). The number of identified proteins per sample ranged from 110 to 228. From these 237 proteins, 84% were identified in more than half of the 29 samples. An overview of all identified proteins can be found in Supplementary Table S3.1.

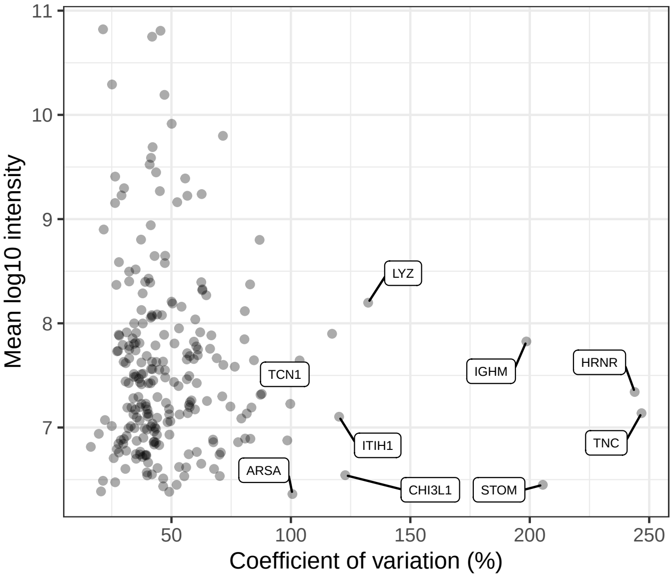


Figure 3.2: Distribution of the variation of proteins. The overall coefficient of variation (CV) on the x-axis versus the mean log10 of the LFQ intensities for each identified protein on the y-axis. Proteins with the largest variation are labeled with their respective gene code.

As shown in Figure 3.2, only 4.2% (n = 10) of the identified proteins show an extensive overall CV >100%. From this figure, it can also be noted that high abundant proteins show a relatively low variation between samples when compared with low abundant proteins. This corresponds with previous studies in which a relatively low variation was found for the most abundant human milk proteins [14, 38]. In addition, it was observed that the overall variation in proteins (median CV = 42.8%) surpasses to a great extent the technical variation (median CV = 20.4%) (see Supplementary Figure S3.3A). Furthermore, it was found by Zhang et al. that the interindividual variation also surpasses the intraindividual variation in the proteome of human milk [12]. This indicates that the major contributor to the overall variation is interindividual variation. This is a pattern also found for the proteome of other body fluids [39–41].

Table 3.2: Top 10 proteins with the largest interindividual coefficient of variation (CV).

Protein ID	Protein name	Gene	Mean log10 intensity	CV (%)
P24821	Tenascin	TNC	7.1	246.8
Q86YZ3	Hornerin	HRNR	7.3	243.9
P27105	Stomatin	STOM	6.5	205.5
P01871	Immunoglobulin heavy constant mu	IGHM	7.8	198.6
P61626	Lysozyme C	LYZ	8.2	132.3
P36222	Chitinase-3-like protein 1	CHI3L1	6.5	122.8
P19827	Inter-α-trypsin inhibitor heavy chain H1	ITIH1	7.1	120.2
Q6ZW64	cDNA FLJ41552 fis		7.9	117.2
P20061	Transcobalamin-1	TCN1	7.6	103.7
P15289	Arylsulfatase A	ARSA	6.4	100.6

Few proteins show a remarkably large interindividual variation (Table 3.2). From these, the first four will be discussed more in detail. Tenascin (TNC) is well-known for its neutralizing effect on HIV [42]. Whereas the decrease of TNC over time was shown to become stable after 30 days postpartum [43], it was found that the concentration of TNC in milk from HIV negative mothers (21 to 46 days postpartum) can range from around 0.1 to more than 100 μg/mL [44]. This is in line with the large variation observed in the current study, where all donating mothers tested HIV negative. Although little is known about the expression of TNC in milk, it is known that TNC synthesis is rapidly induced in many tissues in response to pathological stress and inflammation [45]. Mills et al. showed, in line with this, that airway epithelial cells generate TNC in response to viral infection [46]. Furthermore, Sur et al. showed recently that exosomes in plasma from COVID-19 patients contain significantly increased levels of TNC, triggering pro-inflammatory cytokine signaling [47]. Hence, it can be hypothesized that a higher level of TNC in milk might be an attempt to protect the offspring against the transmission of viral infections from the mother. Alternatively, high TNC levels could indicate an inflammatory response in the mother (for example, mammary gland), although the donors were reported healthy at the time of the donations. For the second protein, hornerin (HRNR), it is known that it is expressed in regenerating and psoriatic skin [48]. Nevertheless, none of the donating mothers mentioned psoriasis as an underlying disease in the current study. In addition, HRNR has also been found to be differently expressed in breast epithelial cells that are in different stages of mammary development [49].

A significant difference was observed in HRNR staining of murine mammary tissue during lactation and at the onset of involution [49]. This might be due to epithelial cell turnover or apoptosis and could explain the large interindividual variation for this protein. The third protein, stomatin (STOM), is a protein found in the plasma membrane associated with lipid rafts. The presence of STOM in milk is dependent on energy balance in the lactation of cows [31]. Nevertheless, little is known about the function or expression of this protein in human milk, and a cause for its large interindividual variation remains speculative. The fourth protein, immunoglobulin M (IgM), is secreted into milk as sIgM and is secreted in the same way as secretory immunoglobulin A (sIgA). Whereas most other identified immunoglobulins show a CV <100%, cDNA FLJ41552 fis (UniProt ID: Q6ZW64) is highly similar to the constant region of IgA and has a CV >100% as well (Table 3.2). Using ELISA, two studies showed a large interindividual variation of IgM in the first 2 weeks of lactation [50, 51]. Although there is a gradual decrease of this protein over lactation [12, 52], it was found that its interindividual variation in mature milk is larger than the other immunoglobulins [53].

It should be noted that in the current study, pooled samples were used from the third month postpartum. It is known that, in this month, longitudinal variation due to the maturation of the milk has leveled out [22, 23]. The influence of intraindividual variation is therefore expected to be minor. Nevertheless, in case of large intraindividual variation due to single outliers before pooling, the effect on interindividual variation is reduced due to the pooling of the samples [54].

To examine whether proteins with high interindividual variation relate to specific biological processes or sample characteristics, a coexpression network was constructed using weighted correlation network analysis (WGCNA). With this, a set of 5 protein clusters was identified (see Supplementary Figures S3.1 and S3.4).

As can be seen in Figure 3.3B, the largest interindividual variation is present in cluster 4 with a median CV = 71.6% and containing medium abundant proteins. This cluster includes nonmicellar caseins and milk fat globule membrane (MFGM) related proteins such as butyrophilin, lactadherin, lipoprotein lipase, lysozyme C, platelet glycoprotein 4, stomatin, and mucins. This suggests the coabundance of proteins involved in the pathway of MFGM secretion by the mammary epithelial cell. When comparing the gene annotations of the clusters (Figure 3.4), cluster 4 contains specifically proteins annotated with lipid storage and phagocytosis, biological processes typical for MFGM proteins [55]. As can be seen in Figure 3.3A, a positive relation was found between protein cluster 4 and maternal BMI ($\rho = 0.45$, $p = 0.01$). The strongest correlation between individual proteins in this cluster and BMI was found with the antiadhesive protein podocalyxin (PODXL). A study by Crujeiras et al. showed that PODXL is negatively associated with methylation levels in subcutaneous adipose tissue and suggested that there may be an epigenetic

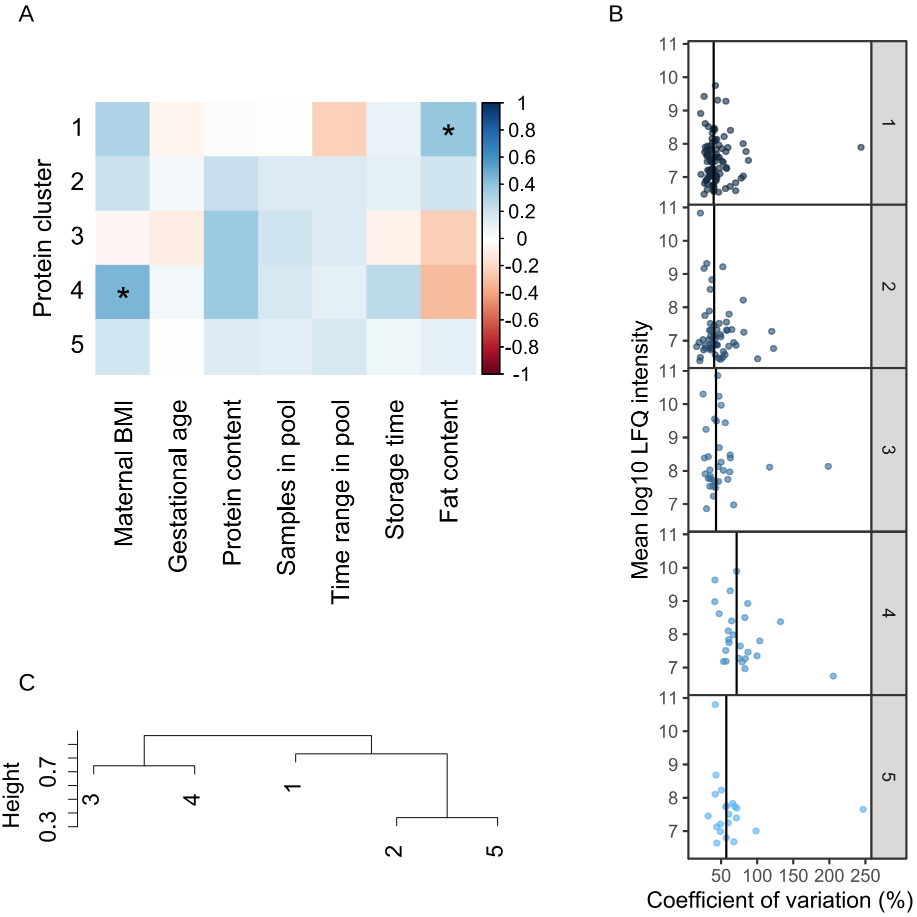


Figure 3.3: (A) Association of eigenproteins with subject and sample characteristics using Spearman correlation. Significant correlations are annotated with * (p <0.05). (B) Interindividual coefficient of variation (CV) in proteins per WGCNA cluster. Vertical lines indicate the median CV of the cluster. (C) Hierarchical clustering of the eigenproteins of each cluster.

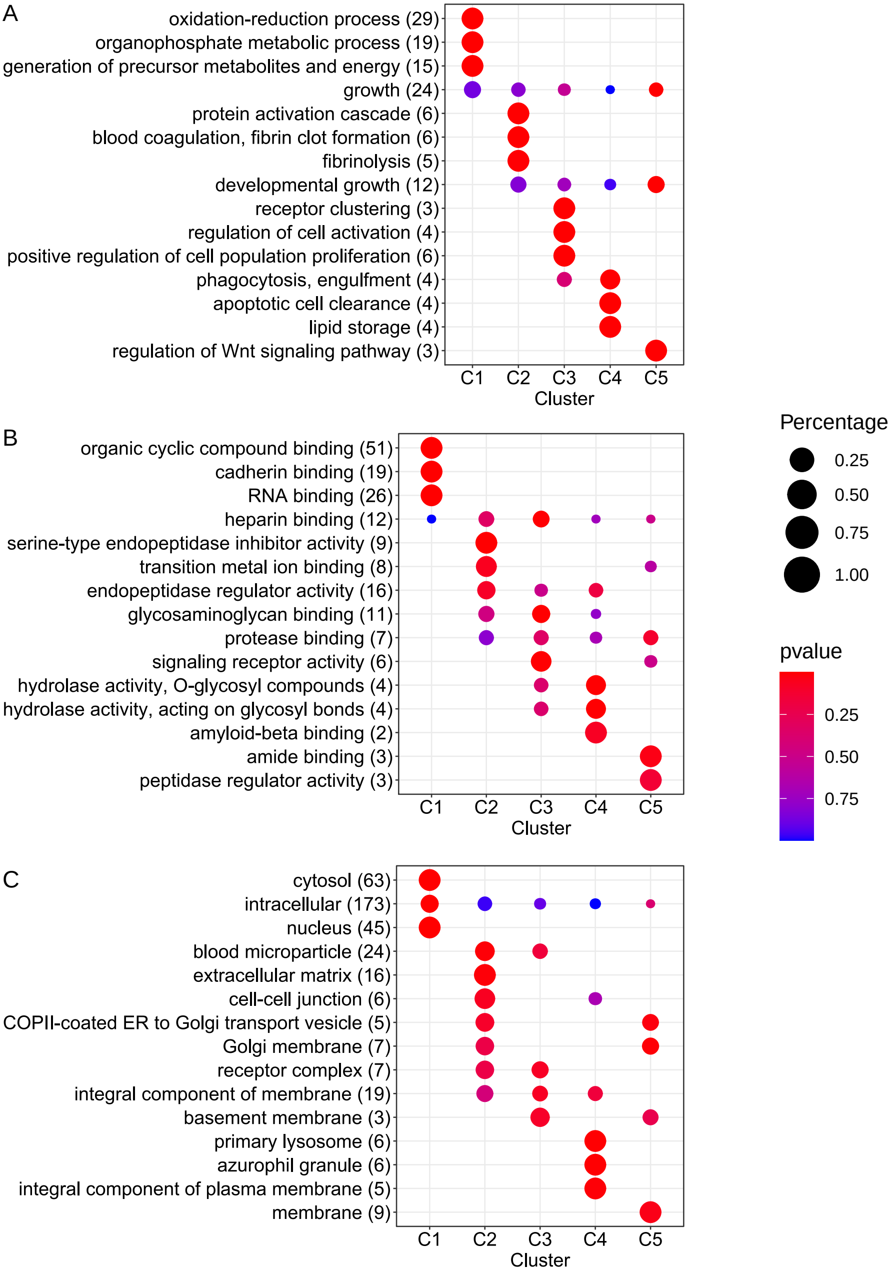


Figure 3.4: *(Caption on next page.)*

Figure 3.4: Overrepresented GO annotations in each WGCNA cluster of the proteomics data, with (A) biological processes, (B) molecular functions, and (C) cellular components.

regulation associated with obesity [56]. Nevertheless, further research is needed to investigate the relation between PODXL in human milk and maternal BMI.

From the other clusters, clusters 1 and 2 show a similar pattern with low abundant proteins and relatively low variation (median CVs of 39.7 and 40.4%, respectively). Cluster 1 comprises the majority of the proteins, with annotations showing involvement in energy pathways and metabolism (Figure 3.4A), and has a positive association with total fat content ($\rho = 0.39$, $p = 0.04$) (see Figure 3.3A). It has been suggested that there might be a common regulation for lipid and protein synthesis in milk [57]. Although the measurements of total protein and fat content in this study do not show a correlation, the observed association might be due to a selective pathway, concerning a selection of the proteins present in human milk.

Cluster 2 is characterized by proteins with serine-type endopeptidase inhibitor activity (Figure 3.4B), among which are α_1-antitrypsin (SERPINA1), plasma protease C1 inhibitor (SERPING1), and α_2-macroglobulin (A2M), proteins also involved in blood coagulation. In addition, this cluster contains a majority of other blood originating proteins such as albumin (ALB) and haptoglobin (HP) (Figure 3.4C). This indicates a coabundance, and possibly related/shared pathways, of these proteins and the protease systems present in milk.

Cluster 3 comprises many of the immune proteins of milk such as polymeric immunoglobulin receptor (PIGR), immunoglobulin A (IgA), immunoglobulin M (IgM), J chain, and lactoferrin (LF). This cluster comprises, in general, medium and high abundant proteins with low variation (median CV = 43%). Several lines of research have suggested a relation between immune proteins in milk and their degradation by proteases [14, 58]. However, such correlation was not observed in the current study, possibly because the focus was on interindividual variation, and Elwakiel et al. [14] studied longitudinal (intraindividual) variation, observing large variation for these proteins over lactation.

Cluster 5 has a median CV of 43% and, in general, low abundant proteins. This cluster is closely related to cluster 2 (see Figure 3.3C) and does not seem to be characterized by large protein groups with unique biological processes or molecular functions (Figure 3.4).

Overall, these results indicate that there is a high interindividual variation in several specific proteins as well as in a cluster of coabundant proteins containing MFGM related proteins and nonmicellar caseins.

3.3.2 Peptidomics

For the analysis of the peptides, proteins were removed from the milk by precipitation. LC-MS/MS analysis of the supernatant and subsequent filtering of the data resulted in the identification of 740 peptides originating from 23 different precursor proteins. The number of peptides identified per sample ranged from 440 to 637. A major part of the identified peptides (38.4%) originated from β-casein, followed by polymeric immunoglobulin receptor (PIGR) (16.5%) and osteopontin (8.5%). This overrepresentation of peptides from a few proteins is probably due to the high abundance of these proteins in combination with direct or indirect association with plasminogen and sensitivity for proteolysis [59]. This pattern is typical for human milk peptidomics and corresponds to previous findings [6, 60]. An overview of all identified peptides can be found in Supplementary Table S3.2.

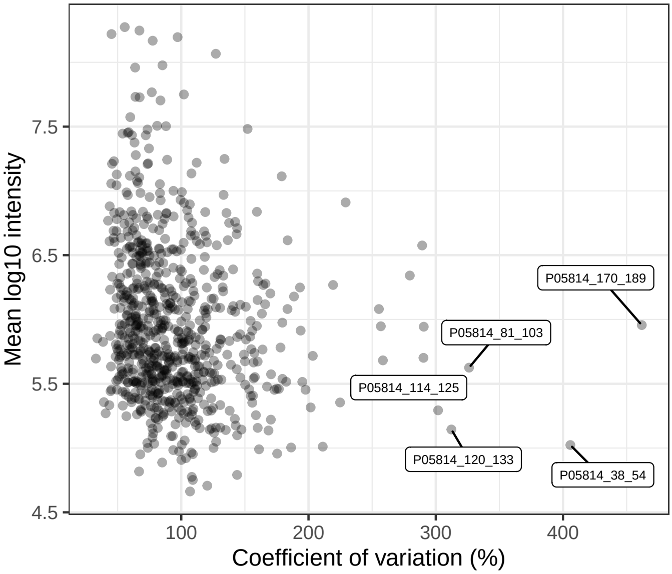


Figure 3.5: Distribution of the variation of peptides. The overall coefficient of variation (CV) on the x-axis versus the mean log10 of the LFQ intensities for each identified peptide on the y-axis. Peptides with the largest variation are labeled with the UniProt ID of their respective precursor protein and their range in the protein sequence.

As shown in Figure 3.5, 36.2% ($n = 268$) of the identified peptides show an overall CV >100%. This figure shows that, like the proteomics data, high abundant peptides show a relatively low variation compared with the lower abundant peptides. In addition, the overall variation (median CV = 85.2%) is substantially larger than the technical variation (median CV = 21.8%) (see Supplementary Figure S3.3B). This indicates that, also for the peptidome, the major contributor to the overall variation is interindividual variation.

It can be noted that the overall variation in peptides is larger than the variation in proteins. This difference is not observed in the technical replicates, where the median and maximum CV for proteins is 20.4 and 107% and for peptides 21.8 and 101%. Therefore, it can be concluded that interindividual variation in the human milk peptidome is substantially larger than in the proteome.

In Table 3.3, the 10 peptides with the highest interindividual variation are shown. All but one of these peptides are from the precursor protein β-casein. Proteolysis of human milk proteins depends on a complex system of proteases, protease inhibitors, and other factors. Therefore, it can be hypothesized that highly variable peptides are, for example, to a variable extent, further degraded depending on the balances in the proteolytic systems. Most of these peptides are relatively long and originate from a region in the protein sequence that is heavily hydrolyzed. Many possible precursor and product peptides of these peptides were also identified, indicating that further proteolysis of these peptides is highly variable and results in their large interindividual variation.

Table 3.3: Top 10 peptides with the largest interindividual coefficient of variation (CV).

Sequence	Protein ID	Peptide range	Mean log10 intensity	CV (%)
SVPQPKVLPIPQQVVPYPQR	P05814	170-189	6.0	461.9
KVKHEDQQQGEDEHQDK	P05814	38-54	5.0	405.8
ILPLAQPAVVLPVPQPEIMEVPK	P05814	81-103	5.6	326.2
SPTIPFFDPQIPKL	P05814	120-133	5.1	312.3
VMPVLKSPTIPF	P05814	114-125	5.3	301.8
SVPQPKVLPIPQQVVPYPQ	P05814	170-188	5.9	290.6
SDISNPTAHENYEKNNVMLQW	P47710	165-185	5.7	290.3
GRVMPVLKSPTIPF	P05814	112-125	6.6	289.3
LAPVHNPISV	P05814	217-226	6.3	279.6
DTVYTKGRVMPVLKSPTIPF	P05814	106-125	5.7	258.4

Little is known about longitudinal variation in the human milk peptide profile. However, it is expected that there is less variation in the third month postpartum due to the maturation of the milk. This is in line with the observation that the activity of plasmin, the main human milk protease, decreases over time [61]. Nevertheless, future research is needed to confirm this. In addition, the effect of single outliers

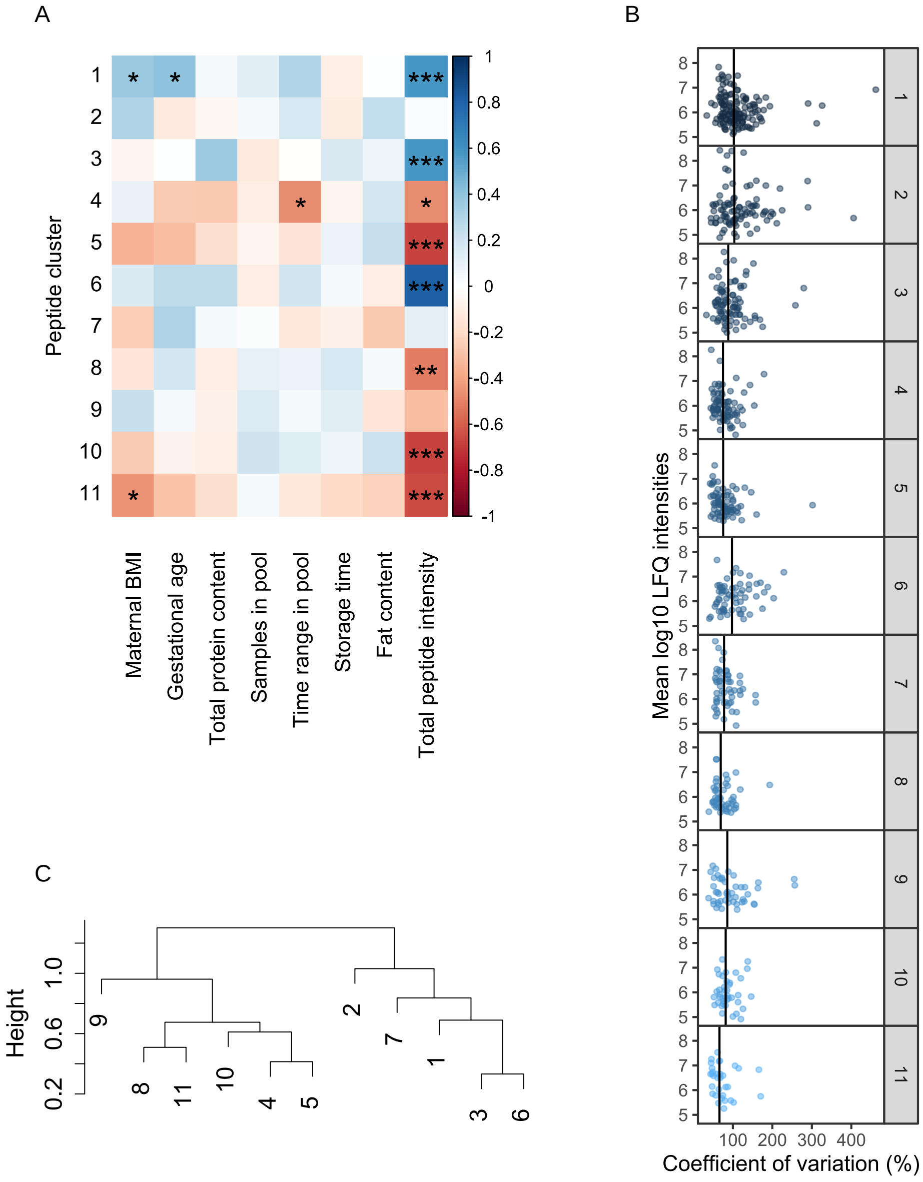


Figure 3.6: *(Caption on next page.)*

Figure 3.6: (A) Association of peptide clusters with subject and sample characteristics using Spearman correlation. Significant correlations are annotated with * (p <0.05), ** (p <0.01), or *** (p <0.001). (B) interindividual variation (%) in peptides per WGCNA cluster. Vertical lines indicate the median coefficient of variation (CV) of the cluster. (C) Hierarchical clustering of the eigenpeptides of each cluster.

before pooling, that is, large intraindividual variation, will be less reflected in the interindividual variation due to the pooling of the samples.

Similar to the proteomics data, WGCNA was applied to the peptidomics data, resulting in 11 clusters of coabundant peptides (see Supplementary Figures S3.2 and S3.5). Although to our knowledge, this is the first time WGCNA has been applied to peptidomics data, clustering of coabundant peptides might provide insights into the different factors that affect the proteolytic degradation of proteins in milk. As can be noted from Table 3.4, several of the clusters are distinctly dominated by peptides from certain precursor proteins. Furthermore, there are several significant correlations between eigenpeptides and sample characteristics (Figure 3.6).

The largest interindividual variation can be observed in clusters 1, 2, 3, and 6 (median CV = 102.4, 103.7, 88.4 and 97.6%, respectively), which are dominated by peptides from β-casein, α_{s1}-casein, and osteopontin (Table 3.4).

Several peptides with large variation (Table 3.3) show coabundance in the first three clusters. For example, two of these peptides (SVPQPKVLPIPQQVVPYPQR and SVPQPKVLPIPQQVVPYPQ) occur in cluster 1 and are only different in one amino acid. This coabundance of peptides shows that the level of a certain peptide can depend on the level of a larger, precursor peptide. When it comes to the responsible proteases, this could mean that further digestion of precursor peptides by, for example, nonspecific carboxypeptidases is dependent on cleavage of the proteins by a more specific protease such as plasmin. It is known that plasmin cleaves preferentially with lysine (K) or arginine (R) in the P1 position. From Figure 3.7A, it can be seen that from the clusters dominated by casein peptides, especially clusters 3 and 6 are characterized by many peptides with K or R in the P1 position, matching plasmin specificity. Besides clusters 1, 2, 3, and 6, cluster 7 is also dominated by peptides from β-casein and α_{s1}-casein. Nevertheless, this cluster has a much lower median variation (77.7%), and the β-casein peptides in this cluster are exclusively from the N-terminal end of the protein (sequence position 16 to 54, see Supplementary Table S3.2). This suggests that the N-terminal of β-casein is proteolyzed with different driving factors and lower interindividual variation than the rest of the sequence. Since cleavage specificity of this cluster is not unique (Figure 3.7), factors such as structure, peptidase activity, or protease inhibition might cause the difference with the other clusters. Nevertheless, proteolysis resulting in the peptides in cluster 7 does not seem to be influenced by total proteolytic activity (Figure 3.6A). This was also observed for cluster 2, even though both clusters comprise several highly abundant peptides (Figure 3.6B). This suggests that higher proteolytic activity seems specific for certain proteins and even protein regions. Peptide clusters dominated by β-casein and α_{s1}-casein are associated more with each other than with the clusters dominated by other precursor proteins (Figure 3.6C). This indicates that the degradation of caseins is distinct from the degradation of other proteins, which might be due to their association in micelles.

Table 3.4: Peptide clusters with their size and dominating precursor proteins.

Cluster label	Cluster size	Top precursor protein ID	Top precursor protein name	Number of peptides per precursor protein
1	132	P05814	β-casein	98
		P47710	α_{s1}-casein	7
2	91	P05814	β-casein	39
		P10451	Osteopontin	15
3	83	P05814	β-casein	39
		P47710	α_{s1}-casein	13
4	77	P01833	Polymeric immunoglobulin receptor	34
		P05814	β-casein	23
5	74	Q99541	Perilipin-2	22
		P15941	Mucin-1	19
6	62	P05814	β-casein	40
		P10451	Osteopontin	6
		P47710	α_{s1}-casein	6
7	53	P05814	β-casein	33
		P47710	α_{s1}-casein	10
8	49	P01833	Polymeric immunoglobulin receptor	20
		P10451	Osteopontin	14
9	47	P01833	Polymeric immunoglobulin receptor	26
		P07498	κ-casein	7
10	40	Q13410	Butyrophilin subfamily 1 member A1	18
		Q99541	Perilipin-2	8
11	32	P01833	Polymeric immunoglobulin receptor	13
		Q13410	Butyrophilin subfamily 1 member A1	7

Clusters 4, 8, 9, and 11 are dominated by peptides from PIGR, with a median CV = 74.9, 69.2, 86.1, and 66.1%, respectively. Clusters 5 and 10 are dominated by peptides from MFGM proteins (median CV = 75.5 and 81.8%, respectively). The clustering of peptides of MFGM proteins is in line with Giuffrida et al., who proposed a specific mechanism for proteolysis of MFGM proteins by proteolytic enzymes in the alveolar cell membranes [62]. This is further supported by the fact that most peptides in these clusters do not match the specificity of plasmin (Figure 3.7). As shown in Figure 3.6, several clusters dominated by peptides from PIGR or MFGM related proteins (5, 8, 10, and 11) show a strong negative correlation with total peptide intensity. As was observed before, higher proteolytic activity seems to attribute mainly to an increase in the intensity of peptides originating from β-casein and α_{s1}-casein, possibly driven by plasmin activity and leading to the largest interindividual variation.

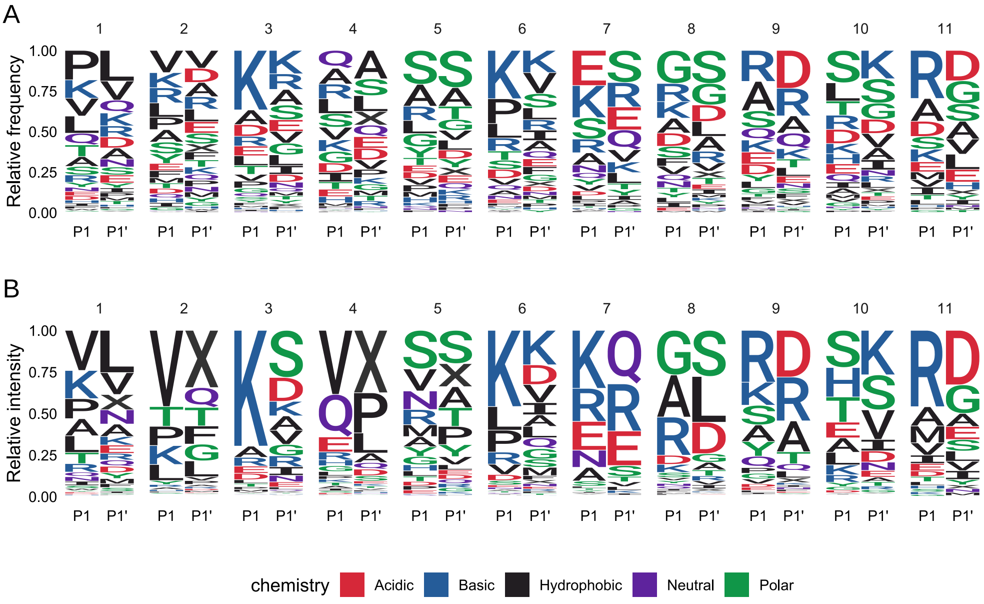


Figure 3.7: (A) Relative frequencies, and (B) relative intensities of amino acids in P1 and P1' position for each peptide cluster. When P1 or P1' position is the C-terminal or N-terminal end of the protein sequence, the empty position is annotated with “X”.

Taken together, these results show that the largest interindividual variation is present in peptides of β-casein, α_{s1}-casein, and osteopontin. With WGCNA, 11 distinct clusters of peptides were obtained, showing differences in characteristics related to proteolytic degradation, such as precursor proteins, cleavage patterns, and association with total peptide intensity.

3.3.3 Metabolomics

Metabolomics analysis with NMR resulted in the identification of 40 metabolites, among which were fatty acids, free amino acids, oligosaccharides, and other small molecules. A detailed list of all identified metabolites can be found in Supplementary Table S3.3.

As shown in Figure 3.8, similar to the proteome and peptidome, metabolites with high intensity also show low interindividual variation. Overall variation between the samples is for the majority of the metabolites larger than the technical variation (see Supplementary Figure S3.3C).

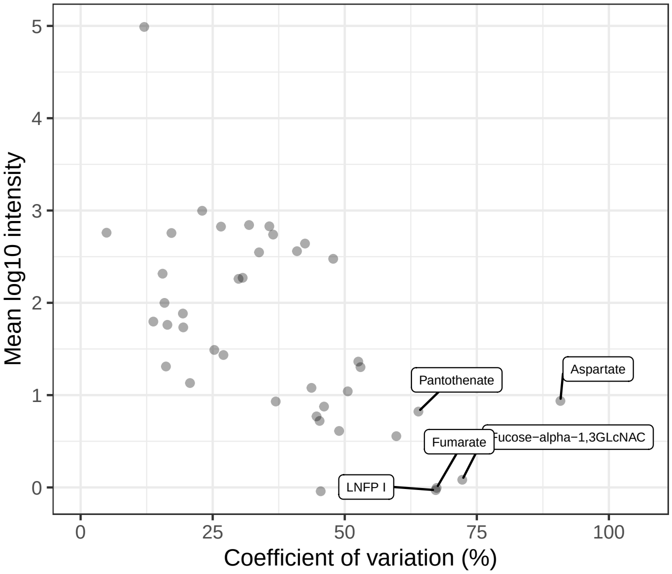


Figure 3.8: Distribution of the variation of metabolites. The overall coefficient of variation (CV) on the x-axis versus the mean log10 of the intensities for each identified metabolite on the y-axis. Metabolites with the largest variation are labeled.

All metabolites identified show a CV <100%, a low variation compared to the proteomics and peptidomics data. The variation in metabolites is similar to the results reported by Smilowitz et al., with the notable exception of butyrate and formate [63]. Smilowitz et al. found a CV of 77.3 and 121% for these metabolites, whereas in the current study, these metabolites had a CV of 4.9 and 36.9%, respectively. One explanation for this difference might be the sample collection. Whereas in the current study, samples were pooled, Smilowitz et al. used nonpooled samples. Larger variation in single samples versus smaller variation in pooled samples might be due to high intraindividual variation, that is, variation between different feedings.

It is proposed from several studies that a large part of the interindividual variation in the human milk oligosaccharide (HMO) metabolome is due to secretor status and Lewis blood type [63–66]. On the basis of intensities of 2’-fucosyllactose (2’FL) and lacto-n-fucopentaose I (LNFP I), 3 out of the 29 donating mothers in the current study are proposed as nonsecretors (Se-) (see Supplementary Figure S3.6). On the basis of intensities of 3’FL, LNFP III, and lactodifucotetraose (LDFT), 2 out of the 29 mothers are proposed to be Lewis negative (Le-), of which one was also Se- (see Supplementary Figure S3.6). Removal of the Se- and Le- samples (n = 4) from the calculations shows decreases in the interindividual variation of the HMO metabolites (see Table 3.5 and Supplementary Table S3.3). Nevertheless, it should be noted that there remains a substantial interindividual variation in for example, LNFP I, 3’-FL, and fucose-α-1,3-GLcNAC. This is an important finding considering the important role of HMOs in the healthy development of the infant.

Table 3.5: Metabolites with high (right) and low (left) interindividual coefficient of variation (CV) together with the interindividual variation in samples from mothers that are proposed to be secretors (Se(+)) as well as Lewis positive (Le(+)).

Metabolite	CV (%)	CV (%) Se(+)Le(+)	Metabolite	CV (%)	CV (%) Se(+)Le(+)
Butyrate	4.9	4.8	Lysine	214.1	209.8
Lactose	12.0	9.5	Aspartate	90.8	89.3
cis-Aconitate	13.8	11.2	Fucose-α-1,3GLcNAC	72.2	58.1
Alanine	15.5	16.1	Fumarate	67.4	64.7
Acetate	15.9	17.0	LNFP I	67.2	63.0
Valine	16.2	17.1	Pantothenate	63.9	64.1
Lactate	16.4	16.0	CMP	59.8	52.2
Urea	17.2	17.0	3’-FL	53.0	41.1
Lacto-N-difucohexaose II	19.4	13.8	LDFT	52.6	44.6
Methionine	19.4	18.7	2’-FL	50.6	37.2

Several metabolites show associations with subject and sample characteristics (Figure 3.9). First, a strong negative relation between glycerophosphocholine (GPC) and BMI (ρ = -0.52, p = 0.003) is present. It was found that in serum of patients with metabolic abnormal obesity, GPC is significantly decreased [67]. Future research is necessary to show whether this holds for human milk as well.

Second, cytidine monophosphate (CMP), cytidine triphosphate and diphosphate (CTP/CDP), GPC, and phosphocholine (PC) all show a significant negative correla-

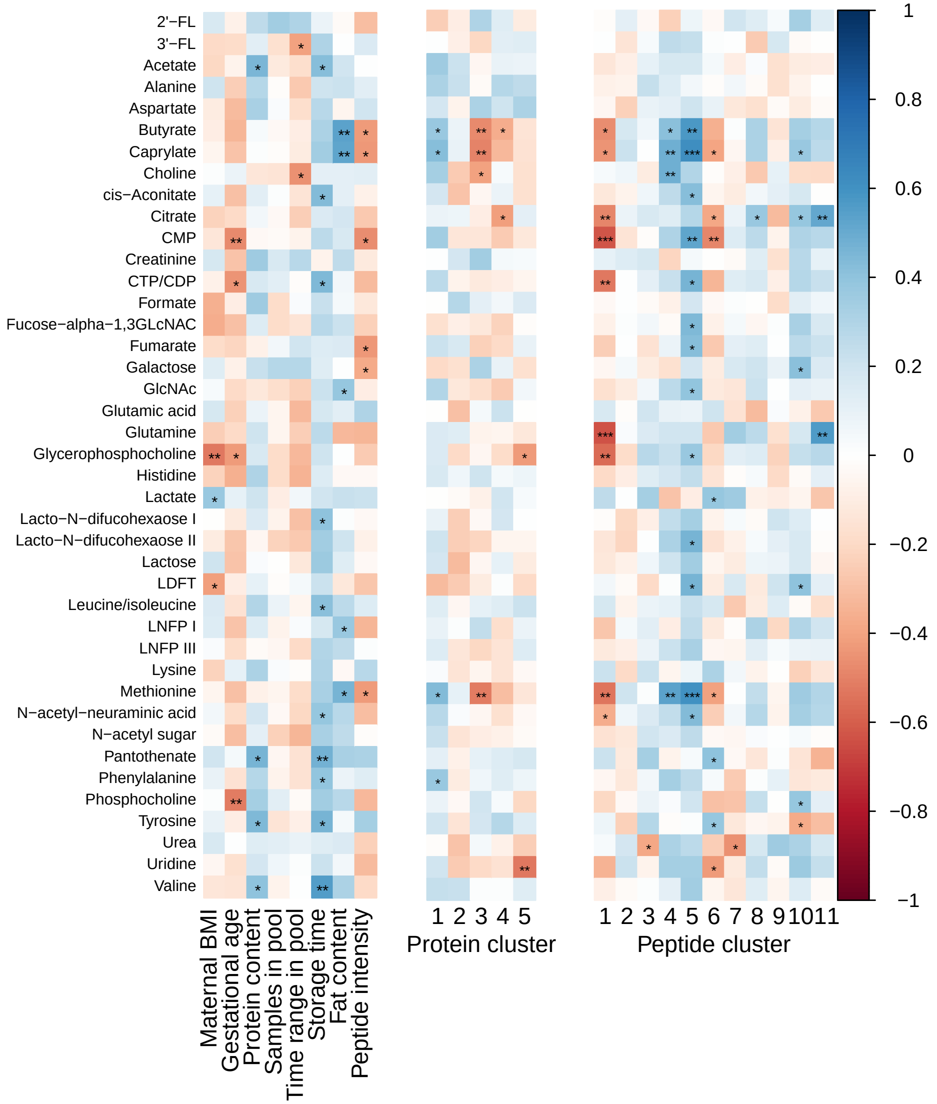


Figure 3.9: Associations of metabolites with subject and sample characteristics (left) eigenproteins (middle) and eigenpeptides (right). Significant correlations are annotated with * (p <0.05), ** (p <0.01), or *** (p <0.001).

tion with maternal age (ρ <-0.42, p <0.022). These metabolites play an important role in the synthesis of the cellular membranes. Therefore, their negative correlation with maternal age might be a marker of the known decrease in mammary epithelial cell proliferation at a higher age [68, 69]. This also accords with Wei et al., who found that PC in bovine milk is negatively correlated with energy balance and proposed a relation with cell proliferation [70].

Third, the fatty acids butyrate (C4:0) and caprylate (C8:0), and the amino acid methionine show a positive relation with fat content (ρ >0.53, p <0.003) and a negative relation with total peptide intensity (ρ <-0.44, p <0.018). This relation with methionine could point to the involvement of this amino acid in fat synthesis. Qi et al. found that methionine promotes fat synthesis through the SNAT2-PI3K signaling pathway in bovine mammary epithelial cells [71]. If this pathway is also present in humans, it would explain the correlations observed in the current study.

Fourth, a negative relation was found between total peptide intensity and butyrate, caprylate, CMP, fumarate, galactose, LNFP I, and methionine (ρ >-0.43, p <0.02). A higher intensity of these metabolites might indicate a lower proteolytic activity in the milk. Although little is known about milk metabolites and their relation with proteolytic activity, it is known that butyrate can stimulate the secretion of plasminogen activator inhibitor 1 in colonic epithelium [72]. Knowing that plasminogen activation needs to precede plasmin activity, this might also hold for the mammary epithelium, causing a decrease in proteolytic activity in the secreted milk.

As can be noted from Figure 3.9, several metabolites show a positive association with the storage time of the samples. From previous research, it is known that butyrate and acetate levels can be affected by storage time [73]. However, the strongest associations were found with pantothenate (vitamin B5) (ρ = 0.48, p = 0.009) and valine (ρ = 0.55, p = 0.002). An increase of pantothenate during frozen storage contradicts the findings of Goldsmith et al., who reported a decrease [74]. On the other hand, the association with valine could point to continued proteolysis during storage, resulting in more FAA. Nevertheless, no strong positive associations were found with peptide or protein clusters (Figures 3.6A and 3.3A, respectively). Therefore, further research on the influence of storage time on the metabolome of human milk is needed for more insight into this.

3.3.4 Relation among omics data sets

To identify associations between the proteomics and peptidomics data, eigenproteins were compared with eigenpeptides (Figure 3.10). From this, several associations were found, of which the most notable will be discussed.

First, it can be observed that protein cluster 4, comprising caseins and MFGM proteins, relates positively with the β-casein-dominated peptide clusters 1 and 6 (ρ

>0.38, p <0.04). From this, it can be hypothesized that the interindividual variation of part of the β-casein peptides is related to the amount of nonmicellar β-casein present in the milk. Second, protein cluster 2, which contains most serine protease inhibitors (SERPINs), relates positively with peptide cluster 2 (ρ = 0.47, p = 0.01). This peptide cluster is dominated by β-casein peptides but does not relate to total peptide intensity (Table 3.4 and Figure 3.6A). This points to SERPIN inhibition of serine proteases, such as thrombin and plasmin, responsible for further degradation of these peptides.

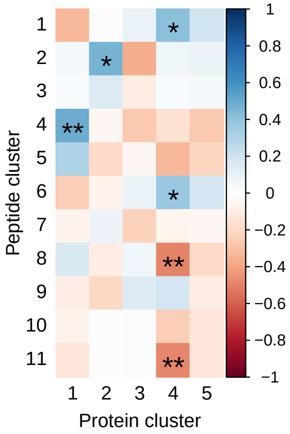


Figure 3.10: Association of eigenproteins with eigenpeptides. Significant correlations are annotated with * (p <0.05) or ** (p <0.01).

In the association of eigenproteins and eigenpeptides with metabolites (Figure 3.9), it was found that butyrate, caprylate, and methionine are negatively associated with protein cluster 3, which contains the majority of the immune-related proteins. In addition, these metabolites associate positively with peptide cluster 5, which is dominated by peptides from MFGM related proteins (perilipin-2 and mucin-1) and negatively with peptide cluster 1. Qi et al. reported that methionine was not only found to promote fat synthesis but also to promote protein synthesis and cell proliferation through the same SNAT2-PI3K signaling pathway [71]. The proteolysis of MFGM related proteins by specific enzymes in the alveolar cell membranes, as discussed before, might therefore be related to cell proliferation. In addition, several other metabolites relate positively with peptide cluster 5 including CMP, GPC, fucose-GlcNac, and LNFP I. Most of these metabolites relate negatively with total peptide intensity, suggesting that higher proteolytic activity correlates with metabolic changes and a decrease in peptides from MFGM proteins.

Therefore, it seems that changes in the metabolome can explain part of the interindividual variation in the human milk proteome and peptidome. Nevertheless, these findings raise intriguing questions regarding the nature of especially the human milk peptidome and deserve further investigation.

3.4 Conclusion

In this study, pooled human milk samples were used to investigate the interindividual variation in proteome, peptidome, and metabolome. The largest interindividual variation was observed in the peptidome (median CV 85.2%), after which follows the proteome (median CV 42.8%) and the metabolome (median CV 36.1%). Nevertheless, the majority of proteins, peptides, and metabolites show interindividual variation with a CV <100%. With the WGCNA algorithm, 5 protein clusters and 11 peptide clusters were obtained, each with distinct characteristics. Using these WGCNA clusters, several associations were found between the data sets and with sample characteristics, giving insight into the causes of interindividual variation. Since the donating mothers in this study are generally healthy, the interindividual variation observed in this study can be considered a normal variation. The findings reported in this study can help in the interpretation of effect sizes in future omics studies since these can now be compared to the natural variability.

Supplementary information

The following supplementary information is available and can be accessed through the QR code in Figure 3.11: Additional results from technical replicates; additional figures on WGCNA data analysis of both proteomics and peptidomics data; boxplots showing proposed nonsecretor and Lewis negative mothers included in the study; tables with all identified proteins, peptides, and metabolites; parameters of NMR analysis.

Figure 3.11: Scan this QR code to access the supplementary information, or visit https://figshare.com/s/977aa820aa4d17e72faa.

References

[1] Zhu, J. et al. “Discovery and Quantification of Nonhuman Proteins in Human Milk”. In: *Journal of Proteome Research* 18 (2019), 225–238. doi: 10.1021/acs.jproteome.8b00550.

[2] Tuma, P. L. and Hubbard, A. L. “Transcytosis: Crossing Cellular Barriers”. In: *Physiological Reviews* 83 (2003), 871–932. doi: 10.1152/physrev.00001.2003.

[3] Bernard, H. et al. “Peanut Allergens Are Rapidly Transferred in Human Breast Milk and Can Prevent Sensitization in Mice”. In: *Allergy: European Journal of Allergy and Clinical Immunology* 69 (2014), 888–897. doi: 10.1111/all.12411.

[4] Wall, S. K. et al. “Blood-Derived Proteins in Milk at Start of Lactation: Indicators of Active or Passive Transfer”. In: *Journal of Dairy Science* 98 (2015), 7748–7756. doi: 10.3168/jds.2015-9440.

[5] Dingess, K. A. et al. “Toward an Efficient Workflow for the Analysis of the Human Milk Peptidome”. In: *Analytical and Bioanalytical Chemistry* 411 (2019), 1351–1363. doi: 10.1007/s00216-018-01566-4.

[6] Dingess, K. A. et al. “Human Milk Peptides Differentiate between the Preterm and Term Infant and across Varying Lactational Stages”. In: *Food and Function* 8 (2017), 3769–3782. doi: 10.1039/c7fo00539c.

[7] Ferranti, P. et al. “Casein Proteolysis in Human Milk: Tracing the Pattern of Casein Breakdown and the Formation of Potential Bioactive Peptides”. In: *Journal of Dairy Research* 71 (2004), 74–87. doi: 10.1017/S0022029903006599.

[8] Aslebagh, R. et al. “Proteomics Analysis of Human Breast Milk to Assess Breast Cancer Risk”. In: *Electrophoresis* 39 (2018), 653–665. doi: 10.1002/elps.201700123.

[9] Hettinga, K. A. et al. “Difference in the Breast Milk Proteome between Allergic and Non-Allergic Mothers”. In: *PLoS ONE* 10 (2015). Ed. by A. S. Wiley, e0122234. doi: 10.1371/journal.pone.0122234.

[10] Liao, Y. et al. “Proteomic Characterization of Human Milk Fat Globule Membrane Proteins during a 12 Month Lactation Period”. In: *Journal of Proteome Research* 10 (2011), 3530–3541. doi: 10.1021/pr200149t.

[11] Zhang, Q. et al. “Quantitative Analysis of the Human Milk Whey Proteome Reveals Developing Milk and Mammary-Gland Functions across the First Year of Lactation”. In: *Proteomes* 1 (2013), 128–158. doi: 10.3390/proteomes1020128.

[12] Zhang, L. et al. "Changes over Lactation in Breast Milk Serum Proteins Involved in the Maturation of Immune and Digestive System of the Infant". In: *Journal of Proteomics* 147 (2016), 40–47. DOI: 10.1016/j.jprot.2016.02.005.

[13] Gao, X. et al. "Temporal Changes in Milk Proteomes Reveal Developing Milk Functions". In: *Journal of Proteome Research* 11 (2012), 3897–3907. DOI: 10.1021/pr3004002.

[14] Elwakiel, M. et al. "Variability of Serum Proteins in Chinese and Dutch Human Milk during Lactation". In: *Nutrients* 11 (2019), 499. DOI: 10.3390/nu11030499.

[15] Zhu, J. et al. "Personalized Profiling Reveals Donor- and Lactation-Specific Trends in the Human Milk Proteome and Peptidome". In: *Journal of Nutrition* 151 (2021), 826–839. DOI: 10.1093/jn/nxaa445.

[16] Campanhon, I. B. et al. "Protective Factors in Mature Human Milk: A Look into the Proteome and Peptidome of Adolescent Mothers' Breast Milk". In: *British Journal of Nutrition* 122 (2019), 1377–1385. DOI: https://doi.org/10.1017/s0007114519002447.

[17] Vella, D. et al. "From Protein-Protein Interactions to Protein Co-Expression Networks: A New Perspective to Evaluate Large-Scale Proteomic Data". In: *Eurasip Journal on Bioinformatics and Systems Biology* 2017 (2017), 6. DOI: 10.1186/s13637-017-0059-z.

[18] Mantini, G. et al. "Co-Expression Analysis of Pancreatic Cancer Proteome Reveals Biology and Prognostic Biomarkers". In: *Cellular Oncology* 43 (2020), 1147–1159. DOI: 10.1007/s13402-020-00548-y.

[19] Zou, X. et al. "Quantitative Proteomics and Weighted Correlation Network Analysis of Tear Samples in Adults and Children with Diabetes and Dry Eye". In: *Translational Vision Science and Technology* 9 (2020), 1–15. DOI: 10.1167/tvst.9.13.8.

[20] Pei, G., Chen, L., and Zhang, W. "WGCNA Application to Proteomic and Metabolomic Data Analysis". In: *Methods in Enzymology* 585 (2017), 135–158. DOI: 10.1016/bs.mie.2016.09.016.

[21] Lamerz, J. et al. "Correlation-Associated Peptide Networks of Human Cerebrospinal Fluid". In: *Proteomics* 5 (2005), 2789–2798. DOI: 10.1002/pmic.200401192.

[22] Lönnerdal, B. et al. "Longitudinal Evolution of True Protein, Amino Acids and Bioactive Proteins in Breast Milk: A Developmental Perspective". In: *Journal of Nutritional Biochemistry* 41 (2017), 1–11. DOI: 10.1016/j.jnutbio.2016.06.001.

[23] Zhang, J. et al. "Longitudinal Changes in the Concentration of Major Human Milk Proteins in the First Six Months of Lactation and Their Effects on Infant Growth". In: *Nutrients* 13 (2021), 1476. doi: 10.3390/nu13051476.

[24] De Waard, M. et al. "Holder-Pasteurized Human Donor Milk: How Long Can It Be Preserved?" In: *Journal of Pediatric Gastroenterology and Nutrition* 66 (2018), 479–483. doi: 10.1097/MPG.0000000000001782.

[25] Cox, J. and Mann, M. "MaxQuant Enables High Peptide Identification Rates, Individualized p.p.b.-Range Mass Accuracies and Proteome-Wide Protein Quantification". In: *Nature Biotechnology* 26 (2008), 1367–1372. doi: 10.1038/nbt.1511.

[26] Dekker, P. M. et al. "Maternal Allergy and the Presence of Nonhuman Proteinaceous Molecules in Human Milk". In: *Nutrients* 12 (2020), 1169. doi: 10.3390/nu12041169.

[27] Development Team Core. *R. A Language and Environment for Statistical Computing*. 2020.

[28] Wei, R. et al. "GSimp: A Gibbs Sampler Based Left-Censored Missing Value Imputation Approach for Metabolomics Studies". In: *PLoS Computational Biology* 14 (2018). Ed. by J. Nielsen, e1005973. doi: 10.1371/journal.pcbi.1005973.

[29] Vizcaíno, J. A. et al. "2016 Update of the PRIDE Database and Its Related Tools". In: *Nucleic Acids Research* 44 (2016), D447–D456. doi: 10.1093/nar/gkv1145.

[30] Lu, J. et al. "Filter-Aided Sample Preparation with Dimethyl Labeling to Identify and Quantify Milk Fat Globule Membrane Proteins". In: *Journal of Proteomics* 75 (2011), 34–43. doi: 10.1016/j.jprot.2011.07.031.

[31] Lu, J. et al. "Changes in Milk Proteome and Metabolome Associated with Dry Period Length, Energy Balance, and Lactation Stage in Postparturient Dairy Cows". In: *Journal of Proteome Research* 12 (2013), 3288–3296. doi: 10.1021/pr4001306.

[32] Xu, W. et al. "Relationship between Energy Balance and Metabolic Profiles in Plasma and Milk of Dairy Cows in Early Lactation". In: *Journal of Dairy Science* 103 (2020), 4795–4805. doi: 10.3168/jds.2019-17777.

[33] Wishart, D. S. et al. "HMDB 4.0: The Human Metabolome Database for 2018". In: *Nucleic Acids Research* 46 (2018), D608–D617. doi: 10.1093/nar/gkx1089.

[34] Langfelder, P. and Horvath, S. "WGCNA: An R Package for Weighted Correlation Network Analysis". In: *BMC Bioinformatics* 9 (2008), 559. doi: 10.1186/1471-2105-9-559.

[35] Yu, G. et al. “ClusterProfiler: An R Package for Comparing Biological Themes among Gene Clusters”. In: *OMICS A Journal of Integrative Biology* 16 (2012), 284–287. DOI: 10.1089/omi.2011.0118.

[36] Carlson, M. et al. “Genome Wide Annotation for Human”. In: *R package version 3.8.2* 3 (2019). DOI: 10.18129/B9.bioc.org.Hs.eg.db.

[37] Wagih, O. “Ggseqlogo: A Versatile R Package for Drawing Sequence Logos”. In: *Bioinformatics* 33 (2017). Ed. by J. Hancock, 3645–3647. DOI: 10.1093/bioinformatics/btx469.

[38] Zhang, L. et al. “Geography and Ethnicity Related Variation in the Chinese Human Milk Serum Proteome”. In: *Food and Function* 10 (2019), 7818–7827. DOI: 10.1039/c9fo01591d.

[39] Hu, Y. et al. “Comparative Proteomic Analysis of Intra- and Interindividual Variation in Human Cerebrospinal Fluid”. In: *Molecular and Cellular Proteomics* 4 (2005), 2000–2005. DOI: 10.1074/mcp.M500207-MCP200.

[40] Nagaraj, N. and Mann, M. “Quantitative Analysis of the Intra-and Inter-Individual Variability of the Normal Urinary Proteome”. In: *Journal of Proteome Research* 10 (2011), 637–645. DOI: 10.1021/pr100835s.

[41] Zhong, W. et al. “Whole-Genome Sequence Association Analysis of Blood Proteins in a Longitudinal Wellness Cohort”. In: *Genome Medicine* 12 (2020), 53. DOI: 10.1186/s13073-020-00755-0.

[42] Fouda, G. G. et al. “Tenascin-C is an Innate Broad-Spectrum, HIV-1-neutralizing Protein in Breast Milk”. In: *Proceedings of the National Academy of Sciences of the United States of America* 110 (2013), 18220–18225. DOI: 10.1073/pnas.1307336110.

[43] Molinari, C. E. et al. “Longitudinal Analysis of Protein Glycosylation and β-Casein Phosphorylation in Term and Preterm Human Milk during the First 2 Months of Lactation”. In: *British Journal of Nutrition* 110 (2013), 105–115. DOI: 10.1017/S0007114512004588.

[44] Mansour, R. G. et al. “The Presence and Anti-HIV-1 Function of Tenascin C in Breast Milk and Genital Fluids”. In: *PLoS ONE* 11 (2016). Ed. by S. Gantt, e0155261. DOI: 10.1371/journal.pone.0155261.

[45] Midwood, K. S. et al. “Tenascin-C at a Glance”. In: *Journal of Cell Science* 129 (2016), 4321–4327. DOI: 10.1242/jcs.190546.

[46] Mills, J. T. et al. “Airway Epithelial Cells Generate Pro-Inflammatory Tenascin-C and Small Extracellular Vesicles in Response to TLR3 Stimuli and Rhinovirus Infection”. In: *Frontiers in Immunology* 10 (2019), 1–12. DOI: 10.3389/fimmu.2019.01987.

[47] Sur, S. et al. “Exosomes from COVID-19 Patients Carry Tenascin-C and Fibrinogen-β in Triggering Inflammatory Signals in Cells of Distant Organ”. In: *International Journal of Molecular Sciences* 22 (2021), 1–11. doi: 10.3390/ijms22063184.

[48] Takaishi, M. et al. “Identification of Human Hornerin and Its Expression in Regenerating and Psoriatic Skin”. In: *Journal of Biological Chemistry* 280 (2005), 4696–4703. doi: 10.1074/jbc.M409026200.

[49] Fleming, J. M. et al. “Hornerin, an S100 Family Protein, Is Functional in Breast Cells and Aberrantly Expressed in Breast Cancer”. In: *BMC Cancer* 12 (2012), 266. doi: 10.1186/1471-2407-12-266.

[50] Koenig, Á. et al. “Immunologic Factors in Human Milk: The Effects of Gestational Age and Pasteurization”. In: *Journal of Human Lactation* 21 (2005), 439–443. doi: 10.1177/0890334405280652.

[51] Lis-Kuberka, J., Berghausen-Mazur, M., and Orczyk-Pawiłowicz, M. “Lactoferrin and Immunoglobulin Concentrations in Milk of Gestational Diabetic Mothers”. In: *Nutrients* 13 (2021), 1–18. doi: 10.3390/nu13030818.

[52] Goonatilleke, E. et al. “Human Milk Proteins and Their Glycosylation Exhibit Quantitative Dynamic Variations during Lactation”. In: *Journal of Nutrition* 149 (2019), 1317–1325. doi: 10.1093/jn/nxz086.

[53] Affolter, M. et al. “Temporal Changes of Protein Composition in Breast Milk of Chinese Urban Mothers and Impact of Caesarean Section Delivery”. In: *Nutrients* 8 (2016), 504. doi: 10.3390/nu8080504.

[54] Diz, A. P., Truebano, M., and Skibinski, D. O. “The Consequences of Sample Pooling in Proteomics: An Empirical Study”. In: *Electrophoresis* 30 (2009), 2967–2975. doi: 10.1002/elps.200900210.

[55] Cao, X. et al. “Quantitative N-glycoproteomics of Milk Fat Globule Membrane in Human Colostrum and Mature Milk Reveals Changes in Protein Glycosylation during Lactation”. In: *Food and Function* 9 (2018), 1163–1172. doi: 10.1039/c7fo01796k.

[56] Crujeiras, A. B. et al. “DNA Methylation Map in Circulating Leukocytes Mirrors Subcutaneous Adipose Tissue Methylation Pattern: A Genome-Wide Analysis from Non-Obese and Obese Patients”. In: *Scientific Reports* 7 (2017), 41903. doi: 10.1038/srep41903.

[57] Qi, L. et al. “Effects of Saturated Long-Chain Fatty Acid on mRNA Expression of Genes Associated with Milk Fat and Protein Biosynthesis in Bovine Mammary Epithelial Cells”. In: *Asian-Australasian Journal of Animal Sciences* 27 (2014), 414–421. doi: 10.5713/ajas.2013.13499.

[58] Chowanadisai, W. and Lönnerdal, B. “*A*1-Antitrypsin and Antichymotrypsin in Human Milk: Origin, Concentrations, and Stability”. In: *American Journal of Clinical Nutrition* 76 (2002), 828–833. doi: 10.1093/ajcn/76.4.828.

[59] Dallas, D. C. et al. “Extensive in Vivo Human Milk Peptidomics Reveals Specific Proteolysis Yielding Protective Antimicrobial Peptides”. In: *Journal of Proteome Research* 12 (2013), 2295–2304. doi: 10.1021/pr400212z.

[60] Guerrero, A. et al. “Mechanistic Peptidomics: Factors That Dictate Specificity in the Formation of Endogenous Peptides in Human Milk”. In: *Molecular and Cellular Proteomics* 13 (2014), 3343–3351. doi: 10.1074/mcp.M113.036194.

[61] Dallas, D. C. et al. “Endogenous Human Milk Peptide Release Is Greater after Preterm Birth than Term Birth”. In: *Journal of Nutrition* 145 (2015), 425–433. doi: 10.3945/jn.114.203646.

[62] Giuffrida, M. G. et al. “Proteolysis of Milk Fat Globule Membrane Proteins in Preterm Milk: A Transient Phenomenon with a Possible Biological Role?” In: *International Journal of Immunopathology and Pharmacology* 21 (2008), 959–967. doi: 10.1177/039463200802100420.

[63] Smilowitz, J. T. et al. “The Human Milk Metabolome Reveals Diverse Oligosaccharide Profiles”. In: *Journal of Nutrition* 143 (2013), 1709–1718. doi: 10.3945/jn.113.178772.

[64] Dessì, A. et al. “Metabolomics of Breast Milk: The Importance of Phenotypes”. In: *Metabolites* 8 (2018), 79. doi: 10.3390/metabo8040079.

[65] Praticò, G. et al. “Exploring Human Breast Milk Composition by NMR-based Metabolomics”. In: *Natural Product Research* 28 (2014), 95–101. doi: 10.1080/14786419.2013.843180.

[66] Wang, A. et al. “The Milk Metabolome of Non-Secretor and Lewis Negative Mothers”. In: *Frontiers in Nutrition* 7 (2021), 1–9. doi: 10.3389/fnut.2020.576966.

[67] Chen, H. H. et al. “The Metabolome Profiling and Pathway Analysis in Metabolic Healthy and Abnormal Obesity”. In: *International Journal of Obesity* 39 (2015), 1241–1248. doi: 10.1038/ijo.2015.65.

[68] Li, C. M. C. et al. “Aging-Associated Alterations in Mammary Epithelia and Stroma Revealed by Single-Cell RNA Sequencing”. In: *Cell Reports* 33 (2020), 108566. doi: 10.1016/j.celrep.2020.108566.

[69] Raafat, A. et al. “Effects of Age and Parity on Mammary Gland Lesions and Progenitor Cells in the FVB/N-RC Mice”. In: *PLoS ONE* 7 (2012). Ed. by E. Katz, e43624. doi: 10.1371/journal.pone.0043624.

[70] Xu, W. et al. “Metabolomics of Milk Reflects a Negative Energy Balance in Cows”. In: *Journal of Proteome Research* 19 (2020), 2942–2949. DOI: 10.1021/acs.jproteome.9b00706.

[71] Qi, H. et al. “Methionine Promotes Milk Protein and Fat Synthesis and Cell Proliferation via the SNAT2-PI3K Signaling Pathway in Bovine Mammary Epithelial Cells”. In: *Journal of Agricultural and Food Chemistry* 66 (2018), 11027–11033. DOI: 10.1021/acs.jafc.8b04241.

[72] Gibson, P. R. et al. “Butyrate Is a Potent Inhibitor of Urokinase Secretion by Normal Colonic Epithelium in Vitro”. In: *Gastroenterology* 107 (1994), 410–419. DOI: 10.1016/0016-5085(94)90166-X.

[73] Wu, J. et al. “NMR-based Metabolite Profiling of Human Milk: A Pilot Study of Methods for Investigating Compositional Changes during Lactation”. In: *Biochemical and Biophysical Research Communications* 469 (2016), 626–632. DOI: 10.1016/j.bbrc.2015.11.114.

[74] Goldsmith, S. J. et al. “Effects of Processing and Storage on the Water-Soluble Vitamin Content of Human Milk”. In: *Journal of Food Science* 48 (1983), 994–995. DOI: 10.1111/j.1365-2621.1983.tb14951.x.

Chapter 4

The human milk proteome and allergy of mother and child: exploring associations with protein levels and protein network connectivity

Pieter M. Dekker, Meghan B. Azad, Sjef Boeren, Piushkumar J. Mandhane, Theo J. Moraes, Elinor Simons, Padmaja Subbarao, Stuart E. Turvey, Edoardo Saccenti, Kasper A. Hettinga

Manuscript submitted

Abstract

Background: The human milk proteome comprises a vast amount of proteins with immunomodulatory functions, but it is not clear how this relates to an allergy of the mother or allergy development in the breastfed infant. This study aimed to explore the relation between the human milk proteome and allergy of both mother and child.

Methods: Proteins were analyzed in milk samples from a subset of 300 mother-child dyads from the Canadian CHILD Cohort Study, selected based on maternal and child allergy phenotypes. For this selection, the definition of "allergy" included food allergy, eczema, allergic rhinitis, and asthma. Proteins were analyzed with non-targeted shotgun proteomics using filter-aided sample preparation (FASP) and nanoLC-Orbitrap-MS/MS. Protein abundances, obtained with label-free relative quantification, were compared using multiple statistical approaches, including univariate, multivariate, and network analyses.

Results: Using univariate analysis, we observed a trend that milk for infants who develop an allergy by 3 years of age contains higher levels of immunoglobulin chains, irrespective of the allergy status of the mother. This suggests a difference in the milk's immunological potential, which might be involved in the development of the infant's immune system.

Furthermore, network analysis showed overall stronger connectivity of proteins in the milk of allergic mothers and milk for infants who would ultimately develop an allergy. This difference in connectivity was especially noted for proteins involved in the translation machinery and may be due to the physiological status of the mother, which is reflected in the interconnectedness of proteins in the milk. In addition, it was shown that network analysis complements the other methods for data analysis by revealing subtler associations between the milk proteome and mother-child allergy status.

Conclusion: Together, these findings give new insight into the human milk proteome as it relates to the allergy status of mother and child, and inspire new research directions into the complex interplay of the mother-milk-infant triad.

4.1 Introduction

Having an allergy can strongly impact someone's quality of life in terms of dietary, social, and psychological factors. In addition, the burden of allergic diseases for healthcare is increasing in western countries [1]. In the attempt to decrease these socioeconomic burdens, a primary concern is to determine where the development of allergic diseases is triggered (window of opportunity) and how this can be prevented. This centers around the early years, as allergic diseases often begin to manifest in the first years of life, and healthy development of the infant's immune system is crucial for later immune health [2].

The role of human milk in the development of allergies has received considerable attention in recent years [3–5]. Breastfed babies receive a spectrum of healthy nutrients through human milk, in a stage of life that is crucial for the development of the immune system. Several components in human milk have functional properties that could play a role in immune development, such as antioxidant, antibacterial, and immunomodulating properties [6]. The effect of breastfeeding on the development of allergic diseases is complex and has been the subject of several epidemiological studies in the last decades [7–9], although meta-analyses do not show conclusive evidence for an allergy preventing effect of breastfeeding [10, 11]. For example, Kull et al. [8] showed that, exclusively breastfed (4 months or more) children in the general population, had a reduced risk of sensitization and asthma compared to children breastfed for less than 4 months, while Mihrshahi et al. [9] reported no significant association between onset of atopy and duration of exclusive breastfeeding. One explanation for these contradicting findings could be differences in the definition of the outcomes. However, it could also be due to the individual-specific composition of human milk which relates to, amongst other factors, maternal genetics, maternal diet, maternal nutrition stores, time of gestation, and time of lactation [12, 13].

It is possible that specific components in human milk with levels determined by individual-specific factors could influence the development of the immune system of the breastfed child. Proteins are a particularly important group of such human milk components with immunomodulatory potential, including immunoglobulins (Igs), cytokines, and dietary antigens.

Thus far, several studies have demonstrated the importance of human milk proteins for the development of the infant's immune system [14–17]. In a cohort study including 398 children, Munblit et al. [14] found that higher levels of transforming growth factor (TGF)$\beta 2$ in human milk were related to a higher occurrence of eczema in the infant. Österlund et al. [15] showed that eosinophil cationic protein (ECP), a marker of eosinophil degranulation, had higher levels in human milk consumed by children that develop cow milk allergy or atopic dermatitis. In another study, Järvinen et al. [16] reported that infants who received human milk with low levels of total immunoglobulin A (IgA) were more likely to develop cow milk allergy. A more recent study conducted by Michel et al. [17] showed interdependencies between maternal allergy status, risk of allergy development in the infant, and IgA, TGFβ1, and TGFβ2 levels in human milk.

In addition to these studies showing the importance of human milk proteins, some studies show the presence of dietary allergens in human milk and their possible relation with maternal and infant allergy status. It was shown, for example, that a bovine milk allergen is present in higher levels in milk from allergic mothers [18] and that the presence of such allergens in the milk might result in tolerance induction [19]. In addition, it was shown by Adel-Patient et al. [20] that sensitized mice who were exposed to bovine β-lactoglobulin (BLG) during lactation transferred protection for this allergen to their offspring at a level that correlated with BLG-specific antibodies in the milk.

The research to date has been mostly limited to targeted, assay-based protein analysis, with a small number of identified proteins. As a result, little is currently known about the relation between the complete human milk proteome and the allergy status of mother and child. We set out to investigate this, using human milk samples from a subset of the Canadian CHILD Cohort Study, a general population birth cohort [21]. This subset included 300 mother-child dyads, equally distributed across four groups representing all possible combinations of allergy of both mother and child. The human milk proteome of these samples was analyzed with a shotgun/bottom-up proteomics workflow, meaning that proteins are analyzed through the identification of peptides that are released from the protein through trypsin digestion. The resulting data was investigated using univariate analysis, exploratory multivariate analysis, classification models, and network analysis (see Figure 4.1).

Whereas in univariate analysis, the focus is on the abundance of the individual proteins, a systems biology approach with network analysis enables considering interconnections between proteins. A protein network is a graphic representation of proteins (nodes) and their associations (edges) expressed by a similarity measure such as correlation coefficients. Edges between nodes can therefore provide information on the interdependence of proteins in pathways and expression [22, 23]. Analysis of protein networks is essential in a thorough investigation of a possible relation between the milk proteome and a pathological condition such as allergy because proteins are pivotal components in sometimes interconnected biological pathways and often play a role through interaction with other proteins [24]. Comparison of associations between proteins across conditions such as allergy status can be carried out using differential network analysis. Such analysis of differences in protein interactions can elucidate and provide a better understanding of molecular mechanisms than univariate analyses focusing on the abundance of individual proteins [25].

This study aimed to explore the relation between the complete human milk proteome on the one hand and both maternal allergy and child allergy development on the other, and is the first to undertake an untargeted analysis of the human milk proteome, examining its relation with the allergy status of both mother and child.

4.2 Materials and methods

4.2.1 Study design CHILD cohort

This study included a subset of $n = 300$ mother-child dyads originating from the CHILD cohort (https://www.childstudy.ca) [21]. In the CHILD Cohort Study, pregnant mothers were recruited from the general population from four locations in Canada (Vancouver, Edmonton, Manitoba, and Toronto). The study was carried out following the Declaration of Helsinki, and local Human Research Ethics Boards approved the study protocols. All parents involved in the study provided written informed consent at enrollment.

The selection of the 300 mother-child dyads for the current study was made based on the allergy status of the mother and the child (Figure 4.1). Based on a 2×2 factorial design including allergy of both mother and child, four equal-sized groups ($n = 75$) were created (allergic mother and child, allergic mother and non-allergic child, non-allergic mother and allergic child, non-allergic mother and child). These 4 groups are later referred to as “mother-child allergy groups.” The groups were matched for lactation stage, maternal age, maternal BMI, secretor status, ethnicity, and infant sex.

4.2.2 Definition of allergy

The definition of maternal allergy included at least one self-reported diagnosis of allergic disease, including asthma, food allergy, hay fever, or skin allergy, at the time of enrollment during pregnancy.

The definition of child allergy included atopic sensitization (1 or 3 years of age) with one or more of the following: atopic dermatitis (1 or 3 years of age), recurrent wheezing (1 year of age), asthma (3 years of age), rhinitis (3 years of age), or food allergy (3 years of age). Atopic sensitization was determined using standardized skin prick tests, including six inhalant allergens (*Alternaria alternata*, cat hair, dog epithelium, house dust mites [*Dermatophagoides pteronyssinus* and *Dermatophagoides farinae*], and German cockroach) and four food allergens (bovine milk, egg, peanut, and soybean). According to the criteria described by Williams et al. [26], atopic dermatitis was assessed by pediatricians of the CHILD study. At three years of age, the CHILD study physician made a careful assessment of the child's clinical history. Diagnoses recorded as "yes" and "possible" were considered positive for the purpose of defining whether the child had any of asthma, allergic rhinitis, food allergy, or atopic dermatitis. A detailed description of the assessments of allergic sensitization and diseases has been given before [27].

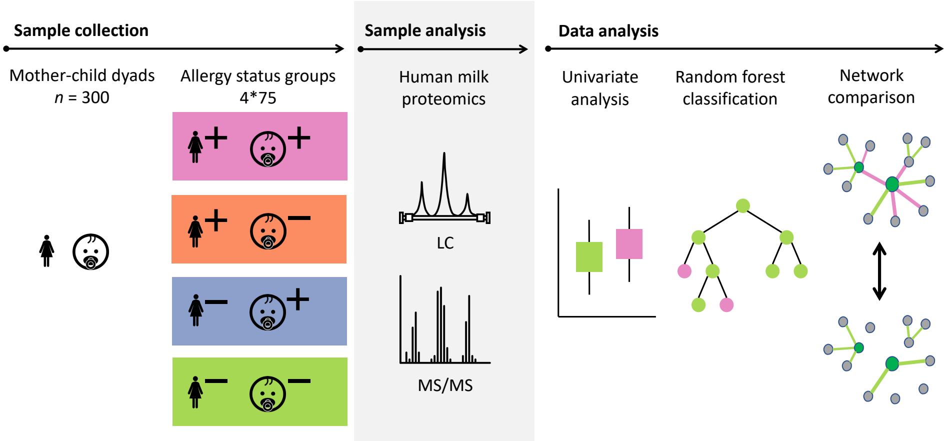


Figure 4.1: Schematic overview showing the sample set from the CHILD Cohort Study and the approach that was used for the analysis of the data. Proteins in a selection of 300 human milk samples from mother-child dyads with different allergy status (+ indicates allergy, - indicates no allergy) were analyzed using mass spectrometry. The data analysis was carried out using univariate analysis, classification models, and network comparison. Probabilistic Context Likelihood of Relatedness on Correlation (PCLRC), differential connectivity, and Covariance Simultaneous Component Analysis (COVSCA) were used for the network comparisons.

4.2.3 Sample collection

Milk samples were collected according to the CHILD protocol [28]. In short, foremilk and hindmilk samples were collected from several feedings during a day and were pooled to minimize within feed variation and diurnal variation. Samples were collected between 6 and 35 weeks post-partum (median = 15.6 weeks, interquartile range (IQR) = 4.6). Samples were stored at 4°C in the home refrigerator and within 24 hours, picked up and transported on ice to the CHILD laboratory. There, samples were aliquoted and stored until further analysis at -80°C. Further transport of the samples was done on dry ice.

4.2.4 Sample preparation

Skimmed milk was obtained by centrifugation at 10,000*g* and 4°C for 30 minutes. Then, skimmed milk was again centrifuged at 1,000*g* and 4°C for 10 minutes to remove any remaining lipids. Skimmed milk samples were prepared with filter-aided sample preparation (FASP) for protein analysis in randomized order as previously described [29].

In addition to the samples from the CHILD Cohort Study, aliquots of a pooled human milk sample were added as a control for technical variation. This sample comprised multiple aliquots of pooled human milk samples from the Dutch Human Milk Bank (Amsterdam, The Netherlands).

4.2.5 LC-MS/MS analysis

Trypsin digested proteins were analyzed with LC-MS/MS as described before, with minor adjustments [30]. In short, 1.5 - 4 μL of tryptic peptide solution was loaded onto a 0.10 × 250 mm ReproSil-Pur 120 C18-AQ 1.9 μm beads analytical column (prepared in-house) at 825 bar. A gradient from 9 to 34% acetonitrile in water with 0.1% formic acid in 50 min (Thermo nLC1000) was used. Full scan FTMS spectra were obtained using a Q-Exactive HFX (Thermo electron, San Jose, CA, USA) in positive mode between 380 and 1400 *m/z*.

The 25 most abundant positively charged peaks (2-5) in the MS scan were fragmented (HCD) with an isolation width of 1.2 *m/z* and 24% normalized collision energy. MSMS scans were recorded in data-dependent mode with a threshold of 1.2×10^5 and 15 s exclusion for the selected *m/z* ± 10 ppm. Samples were analyzed with a technical replicate added randomly to each 7 injections.

4.2.6 Data processing

The Andromeda search engine of the MaxQuant software v1.6.17.0 [31] was used to analyze the raw LC-MS/MS data. For this, a database was created by an initial MaxQuant run using the full human proteome (downloaded from UniProtKB on 20-01-2021, n = 194,237) [32]. Protein identifiers obtained as identification from this initial run were used to create a human milk database for a second run (n = 24,175), in which also a cow milk protein (n = 1,006) and an allergen protein database (n = 721) were added [18].

In MaxQuant, digestion specificity was set to Trypsin/P, with maximally 2 missed cleavages. A fixed propionamide modification was set for cysteines and variable modifications for acetylation of the peptide N-term, deamidation of the side chains of asparagine and glutamine, and oxidation of methionine, with a maximum of five modifications per peptide were set.

Label-free quantification (LFQ) was used to obtain protein abundances. Per identified protein group, a leading protein was selected as described before [18] and proteins were manually annotated with keywords using the UniProt KB database [32] (accession date: 21-02-2022).

4.2.7 Statistical methods

Missing data

In dealing with the missing values in the proteomics data, identifications were first filtered with the requirement that proteins have a minimum of 25 valid values in at least one of the four sample groups. In practice this resulted in a minimum of 66 and a median of 215 valid values. This way of filtering the data prevented the removal of proteins that had only valid values in one of the four sample groups. Following this, the remaining missing values were imputed using the GSimp package [33]. This Gibbs sampler-based algorithm imputes missing values with the assumption that missing values are not at random (MNAR) and left-censored.

Univariate analysis

The Kruskal-Wallis test was applied to deduce differences in protein abundance between the milk from mothers in the different mother-child allergy groups [34]. Resulting p-values were corrected for multiple testing using Benjamini-Hochberg correction [35]. After correction, an adjusted p-value <0.05 was considered significant. Dunn's multiple comparison test [36] was applied to determine differences between specific groups and also these p-values were corrected for multiple hypothesis testing using Benjamini-Hochberg correction.

Principal Component Analysis

For unsupervised data exploration, Principal Component Analysis (PCA) [37] was applied on the 300 × 647 data matrix (samples × proteins), using the FactoMineR package for R [38]. This enabled investigation of the data structure and the possible presence of patterns in protein abundance that cause differentiation between samples from groups with different allergy status. Data was scaled to unit variance before analysis.

Random Forest modeling

Random Forest [39] classification models were built using the R package "randomForest" [40] as described before [41]. Six different models were built to discriminate between the different mother-child allergy groups, covering all pairwise comparisons of allergic/non-allergic mothers with allergic/non-allergic children. The significance of the reported results was assessed with a permutation test using 1000 permutations.

Network inference and analysis

Probabilistic Context Likelihood of Relatedness on Correlation (PCLRC). Protein-protein association networks were built using the Probabilistic Context Likelihood of Relatedness on Correlation (PCLRC) algorithm [42]. This algorithm provides a robust estimation of correlation, using resampling and a modified version of the Context Likelihood of Relatedness (CLR) algorithm [43] to remove nonsignificant background correlations.

With resampling (*n* iterations = 1000), 75% of each dataset was randomly selected and subjected to the CLR algorithm. Resulting from this was a matrix with a probabilistic measure p_{ij} for each correlation between proteins r_{ij}, where *i* and *j* indicate the *i*-th and *j*-th element in the Spearman correlation matrix. Correlations were retained if p_{ij} >0.99 and all other correlations were replaced with 0.

$$r_{ij} = \begin{cases} r_{ij} & \text{if } p_{ij} \geq 0.99 \\ 0 & \text{if } p_{ij} < 0.99 \end{cases}$$

Networks were built for each different mother-child allergy group, resulting in a total of 4 protein networks. The connectivity of a protein *i* in network *a* with mother-child allergy status *S* is defined according to:

$$\chi_i^{a\in S} = \left(\sum_{j=1}^{J} |r_{ij}| \right) - 1$$

Whereas the differential connectivity between two networks *a* and *b*, with different mother-child allergy statuses *S1* and *S2*, is calculated by:

$$\Delta_i^{a\in S1, b\in S2} = \chi_i^{a\in S1} - \chi_i^{b\in S2}$$

All *p*-values for differential connectivity were adjusted for multiple testing with Benjamini-Hochberg correction [35]. Significant differential connectivities (*p* <0.05) were considered for further analysis and interpretation.

Covariance Simultaneous Component Analysis (COVSCA). To explore comprehensively the (dis)similarity among the protein association networks, Covariance Simultaneous Component Analysis (COVSCA) was used [44]. With this approach, differences and commonalities between the different networks can be modeled.

In comparing networks with COVSCA, each network becomes a point in the component space. Thus, the method enables a representation and visualization of multiple networks in a way that is similar to PCA. Points (protein association networks) that are close to each other in the R-dimensional space share similar characteristics, i.e., similar correlation patterns. Furthermore, the loadings of the components give the relative contribution of each protein in shaping the observed network differences.

COVSCA, initially developed for modeling multiple covariance matrices at the same time, can also be used for the adjacency matrices resulting from the PCLRC. The K matrices are modeled as a combination of low dimensional prototypes ($L \ll K$):

$$\mathbf{S}_k = \sum_{l=1}^{L} c_{kl}\mathbf{Z}_l\mathbf{Z}_l^T$$

In this, $c_{kl} \geq 0 (l = 1, 2, ..., L)$ are weight coefficients, and $\mathbf{Z}_l\mathbf{Z}_l^T$ are prototypical symmetric matrices consisting of loading $\mathbf{Z}$ of size $J \times R_l$ that hold simultaneously for all S_k.

Two rank-1 prototype matrices were used to fit the model, resulting in one set of loadings per component. This was chosen as the best compromise between goodness of fit (68%) and the complexity of the COVSCA model (rank and number of the prototypical matrices). COVSCA loadings were transformed to z-scores and loadings with $z > |2|$ were further investigated.

Overrepresentation analysis

The GORILLA (Gene Ontology enRIchment anaLysis and visuaLizAtion tool) (http://cbl-gorilla.cs.technion.ac.il/) tool [45] was used for overrepresentation analysis of gene ontology (GO) annotations in selections of proteins that were differentially connected. The tool was used in two list mode where all proteins identified in the current study were used as background set. All p-values reported were corrected with Benjamini-Hochberg correction [35], and considered significant with p <0.05.

4.3 Results

Proteomic analysis of all samples led to a total of 1629 identified proteins before filtering on missing values. After filtering these proteins on the requirement of being identified $\geq$ 25 times in at least one of the four mother-child allergy groups, 647 proteins remained for further data analysis. In this filtered dataset, the number of identified proteins per sample ranged between 256 and 586 (median = 458). The major milk proteins α-lactalbumin, albumin, lactoferrin, β-casein, and α_{s1}-casein, were in all analyzed samples among the top 15 most abundant proteins. A complete overview of the 647 identified proteins can be found in Supplementary File S4.1.

4.3.1 Univariate analysis

Differences in protein intensities between the different mother-child allergy groups were assessed with Kruskal-Wallis tests. After correction for multiple hypothesis testing, no significant differences were found among the four groups (Table 4.1). Kruskal-Wallis outcomes with uncorrected p <0.05 were further assessed with post-hoc tests (Dunn's) and subsequent correction for multiple testing, which resulted in 23 proteins that showed a difference between the groups with adjusted p <0.05 (Table 4.1).

Most of these differences (n = 15) were found between the non-allergic group (M-C-) and the group where only the child ultimately developed an allergy (M-C+). Proteins that differed between these groups were primarily Ig chains (11 out of 15) and were mostly higher in abundance in the group where the mother was non-allergic and the child developed an allergy (Figure 4.2). Additionally, 4 of these Igs show also higher levels in milk from allergic mothers with children who did not develop an allergy.

Further investigation of all identified Ig proteins showed that the mean abundance of these proteins is in general lower in the groups where mother, child or both are allergic, when compared to the non-allergic group (Figure 4.3). This effect is the clearest in the comparison of the group where only the child developed an allergy with the group where both mother and child are non-allergic. Out of 81 Ig proteins, 75 proteins have a mean abundance that is higher in the group where the child developed an allergy.

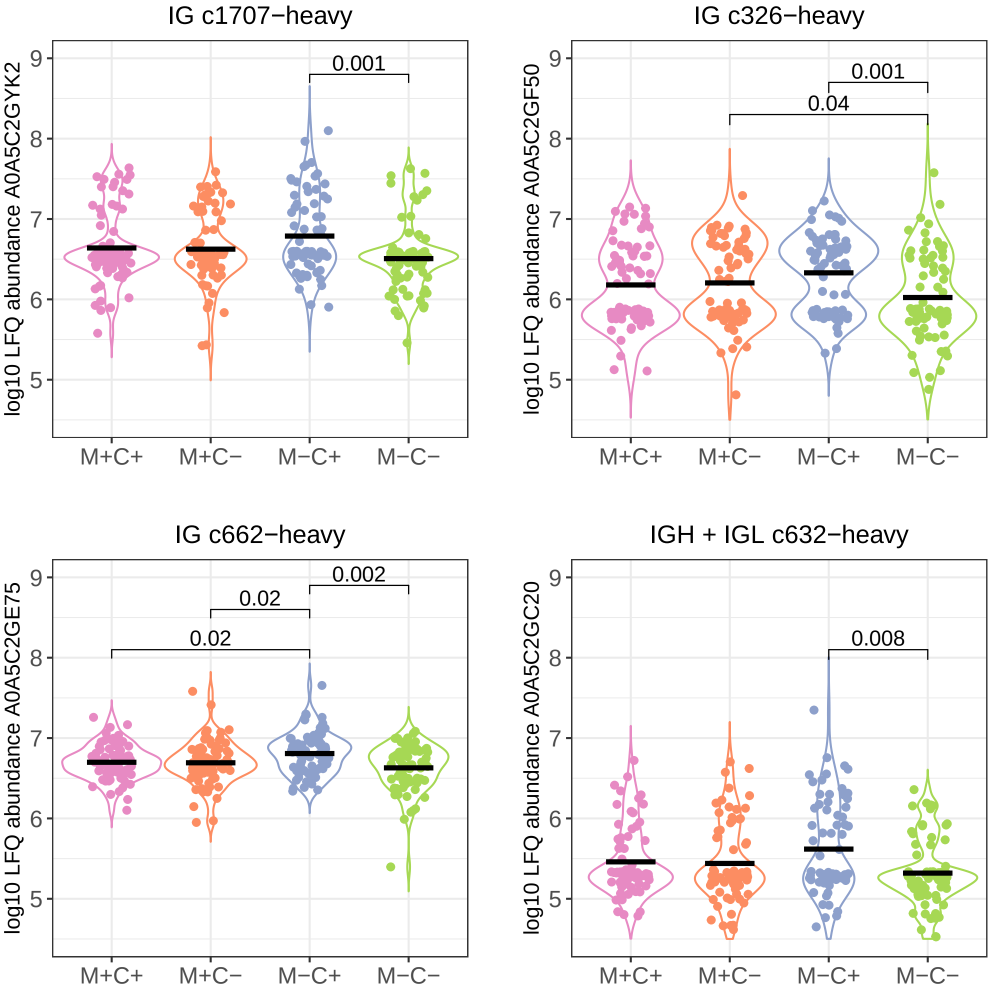


Figure 4.2: Violin plots visualizing the differences in abundance of the 4 most significantly different immunoglobulin (Ig) chains between the different allergy status groups from the CHILD Cohort Study. Differences between groups are indicated with *p*-values from Dunn's post-hoc tests, and means of each group are shown with black, horizontal lines. In the labeling of the groups, M indicates mother, C indicates child, + indicates allergy, and - indicates no allergy.

Table 4.1: Results of univariate analysis (Kruskal-Wallis) with subsequent post-hoc test (Dunn's) for the comparison of protein abundance in milk from allergic (M+) and non-allergic (M-) mothers, with children who developed an allergy (C+) and did not develop an allergy (C-) in the CHILD Cohort Study. The trend indicates higher (⇑) or lower (⇓) abundance in the first group in the comparison. Listed are all proteins with uncorrected *p*-value <0.05 (Kruskal-Wallis) and corrected *p*-value <0.05 (Dunn's), sorted by the mother-child allergy groups in the comparison.

					Comparison			
Leading protein	UniProt ID	Keyword	*p*-value[a]	Adjusted *p*-value[a]	Group 1	Group 2	Adjusted *p*-value[b]	Trend
Sodium-dependent phosphate transport protein 2B	O95436	Transport	0.008	0.867	M-/C-	M-/C+	0.014	⇑
V2-7 protein	A2MYD4	Immunoglobulin	0.029	0.957	M-/C-	M-/C+	0.041	⇓
IGL c1836-light	A0A5C2G0A5	Immunoglobulin	0.014	0.957	M-/C-	M-/C+	0.019	⇓
IGL c2315-light	A0A5C2G2Y4	Immunoglobulin	0.031	0.957	M-/C-	M-/C+	0.037	⇓
IGH + IGL c632-heavy	A0A5C2GC20	Immunoglobulin	0.012	0.944	M-/C-	M-/C+	0.008	⇓
IG c662-heavy	A0A5C2GE75	Immunoglobulin	0.002	0.453	M-/C-	M-/C+	0.002	⇓
IG c326-heavy	A0A5C2GF50	Immunoglobulin	0.003	0.453	M-/C-	M-/C+	0.001	⇓
IG c849-heavy	A0A5C2GF92	Immunoglobulin	0.020	0.957	M-/C-	M-/C+	0.025	⇓
IG c279-heavy	A0A5C2GLS6	Immunoglobulin	0.011	0.944	M-/C-	M-/C+	0.008	⇓
IG c1707-heavy	A0A5C2GYK2	Immunoglobulin	0.002	0.453	M-/C-	M-/C+	0.001	⇓
Methyltransferase-like protein 9	H3BM54	Methyltransferase	0.001	0.453	M-/C-	M-/C+	0.028	⇓
Delta-1-pyrroline-5-carboxylate synthase	P54886	Proline biosynthesis	0.026	0.957	M-/C-	M-/C+	0.039	⇓
Prosaposin variant	Q53FJ5	Lipid metabolism	0.028	0.957	M-/C-	M-/C+	0.018	⇓
Immunoglobulin heavy	Q9NPP6	Immunoglobulin	0.032	0.957	M-/C-	M-/C+	0.036	⇓
IgG L chain	S6BAR0	Immunoglobulin	0.023	0.957	M-/C-	M-/C+	0.026	⇓
N90-VRC38.07 heavy	A0A1W6IYI6	Immunoglobulin	0.033	0.957	M-/C-	M+/C-	0.019	⇓
V2-7 protein	A2MYD4	Immunoglobulin	0.029	0.957	M-/C-	M+/C-	0.041	⇓
IGL c1836-light	A0A5C2G0A5	Immunoglobulin	0.014	0.957	M-/C-	M+/C-	0.019	⇓
IG c326-heavy	A0A5C2GF50	Immunoglobulin	0.003	0.453	M-/C-	M+/C-	0.043	⇓
IG c849-heavy	A0A5C2GF92	Immunoglobulin	0.020	0.957	M-/C-	M+/C-	0.025	⇓
Nephronectin	Q6UXI9	Calcium binding	0.032	0.957	M-/C-	M+/C-	0.021	⇑

Table 4.1: *(Continued)* Results of univariate analysis (Kruskal-Wallis) with subsequent post-hoc test (Dunn's) for the comparison of protein abundance in milk from allergic (M+) and non-allergic (M-) mothers, with children who developed an allergy (C+) and did not develop an allergy (C-) in the CHILD Cohort Study. The trend indicates higher (⇑) or lower (⇓) abundance in the first group in the comparison. Listed are all proteins with uncorrected p-value <0.05 (Kruskal-Wallis) and corrected p-value <0.05 (Dunn's), sorted by the mother-child allergy groups in the comparison.

					Comparison			
Leading protein	UniProt ID	Keyword	p-value[a]	Adjusted p-value[a]	Group 1	Group 2	Adjusted p-value[b]	Trend
Phospholipid hydroperoxide glutathione peroxidase	P36969	Peroxidase	0.028	0.957	M-/C-	M+/C+	0.021	⇓
Sodium-dependent phosphate transport protein 2B	O95436	Transport	0.008	0.867	M-/C-	M+/C+	0.014	⇑
Hornerin	Q86YZ3	Keratinization	0.024	0.957	M-/C-	M+/C+	0.020	⇑
Galectin-3-binding protein	Q08380	Cell adhesion	0.037	0.999	M-/C+	M+/C-	0.023	⇑
IGL c2315-light	A0A5C2G2Y4	Immunoglobulin	0.031	0.957	M-/C+	M+/C-	0.037	⇑
IG c662-heavy	A0A5C2GE75	Immunoglobulin	0.002	0.453	M-/C+	M+/C-	0.017	⇑
Methyltransferase-like protein 9	H3BM54	Methyltransferase	0.001	0.453	M-/C+	M+/C-	0.034	⇑
Alpha-S1-casein	A0A0J9YVR3	Milk protein	0.035	0.993	M-/C+	M+/C+	0.043	⇓
IG c662-heavy	A0A5C2GE75	Immunoglobulin	0.002	0.453	M-/C+	M+/C+	0.017	⇑
Methyltransferase-like protein 9	H3BM54	Methyltransferase	0.001	0.453	M-/C+	M+/C+	0.000	⇑
Ectonucleoside triphosphate diphosphohydrolase 3	O75355	Hydrolase	0.022	0.957	M-/C+	M+/C+	0.040	⇓
Protein FAM3C	Q92520	Cytokine	0.008	0.867	M-/C+	M+/C+	0.009	⇑
IgG L chain	S6BAR0	Immunoglobulin	0.023	0.957	M-/C+	M+/C+	0.026	⇑

[a] from Kruskal-Wallis tests

[b] from Dunn's post-hoc tests

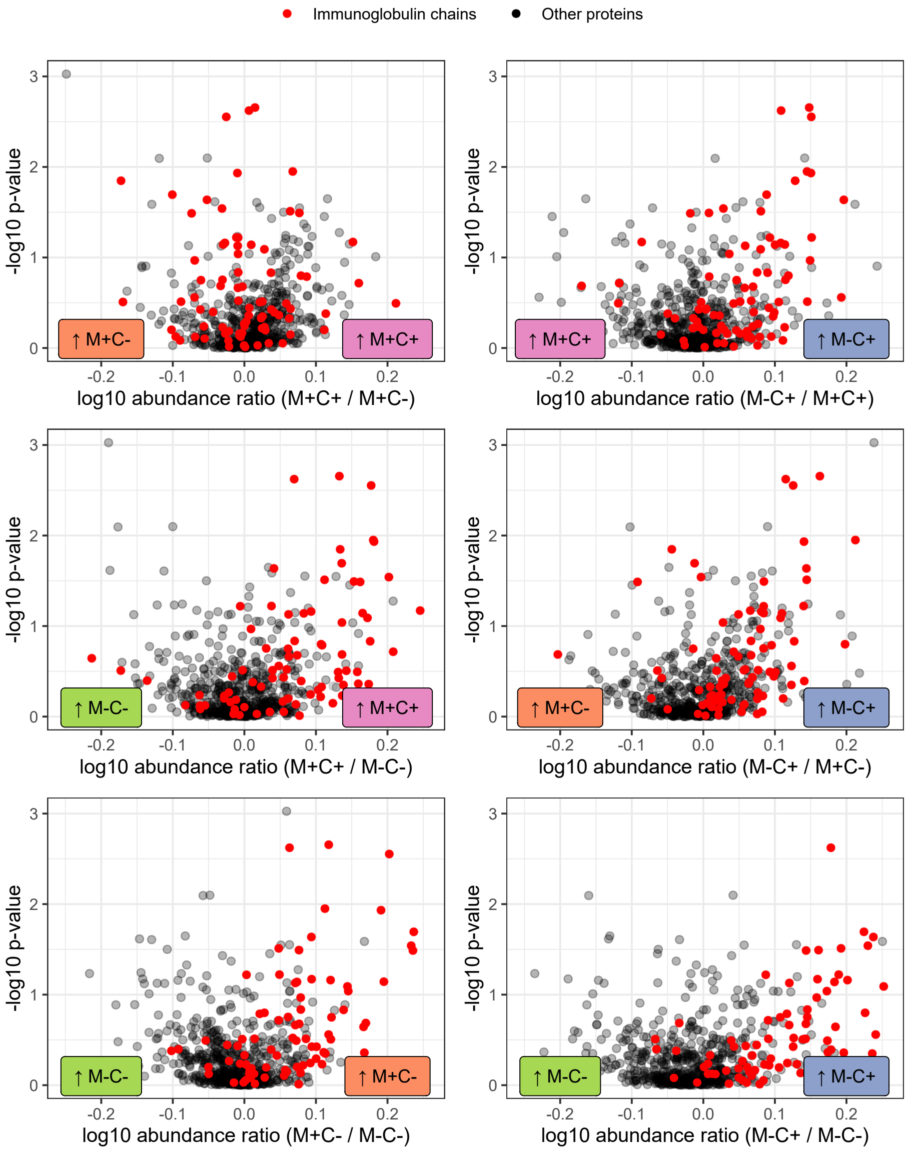


Figure 4.3: *(Caption on next page.)*

Figure 4.3: Volcano plots visualizing the trend in immunoglobulin levels in milk from different mother-child allergy status groups from the CHILD Cohort Study. Each data point represents one protein, with on the x-axes the ratio of the means of the log10 transformed label-free quantification (LFQ). Immunoglobulin-related proteins are represented by red and other proteins with grey dots. Colored labels on left and right side of x = 0 indicate in which mother-child allergy status group the mean abundance of the respective proteins is higher. In the labeling of the groups, M indicates mother, C indicates child, + indicates allergy, and - indicates no allergy.

4.3.2 Non-human proteins

In the current study, several non-human proteins were identified (n = 9), among which were albumin from dog, horse, and cat, as well as bovine α_{s1}-casein and BLG (Table 4.2). However, the majority of these proteins were only found with few tryptic peptides in a low number of samples and filtered out before further data analysis. Additional non-human proteins of potential interest that were included in the database, but not identified in any samples, include allergens from, for example, peanut, egg, and dust mite.

4.3.3 Multivariate exploratory analysis

To explore whether patterns in the abundance of proteins allow a differentiation of the different groups of allergy status, PCA was performed. The visualization of all samples using the first two components of the PCA shows that there is no separation between the groups of different allergy status of mother and child (Figure 4.4), suggesting no major global differences in the protein profiles of these four groups.

Table 4.2: Identified non-human tryptic peptides in human milk samples from the CHILD Cohort Study (n = 150 allergic mothers and 150 non-allergic mothers).

Sequence	UniProt ID	Leading protein	Organism	Identified in n (%) samples from allergic mothers	Identified in n (%) samples from non-allergic mothers	Identification score[a]
KQTALVELLK	P49822	Albumin	Bos taurus (Bovine)	11 (7)	18 (12)	87.3
LVNELTEFAK	P02769	Albumin	Bos taurus (Bovine)	100 (67)	103 (69)	125.0
EKVNELSK	P02662	α_{s1}-casein	Bos taurus (Bovine)	2 (1)	3 (2)	149.7
HIQKEDVPSER	P02662	α_{s1}-casein	Bos taurus (Bovine)	4 (3)	1 (1)	93.6
IDALNENK	P02754	β-lactoglobulin	Bos taurus (Bovine)	12 (8)	11 (7)	89.8
LISVDTEHSNIYLQNGPNR	F1N076	Ceruloplasmin	Bos taurus (Bovine)	28 (19)	32 (21)	203.6
MFTTAPDQVDKENEDFQESNK	F1N076	Ceruloplasmin	Bos taurus (Bovine)	2 (1)	3 (2)	88.0
SSQDLQPR	Q0P5H7	Probable arginine–tRNA ligase	Bos taurus (Bovine)	32 (21)	32 (21)	81.4
FPKADFAEISK	P49822	Albumin	Canis lupus familiaris (Dog)	4 (3)	4 (3)	90.2
LVNEVTEFAKK	Q5XLE4	Albumin	Equus caballus (Horse)	147 (98)	139 (93)	124.1
AEFAEISK	P49064	Albumin	Felis catus (Cat)	73 (49)	82 (55)	84.9
AFKAWSVAR	P49064	Albumin	Felis catus (Cat)	100 (67)	109 (73)	98.3
EVCKNYQEAK	P49064	Albumin	Felis catus (Cat)	94 (63)	95 (63)	96.5
YICENQDSISTK	P49064	Albumin	Felis catus (Cat)	17 (11)	11 (7)	85.4

[a] Score from the MaxQuant output indicating the quality of the identification of the peptide. A higher score represents a better identification.

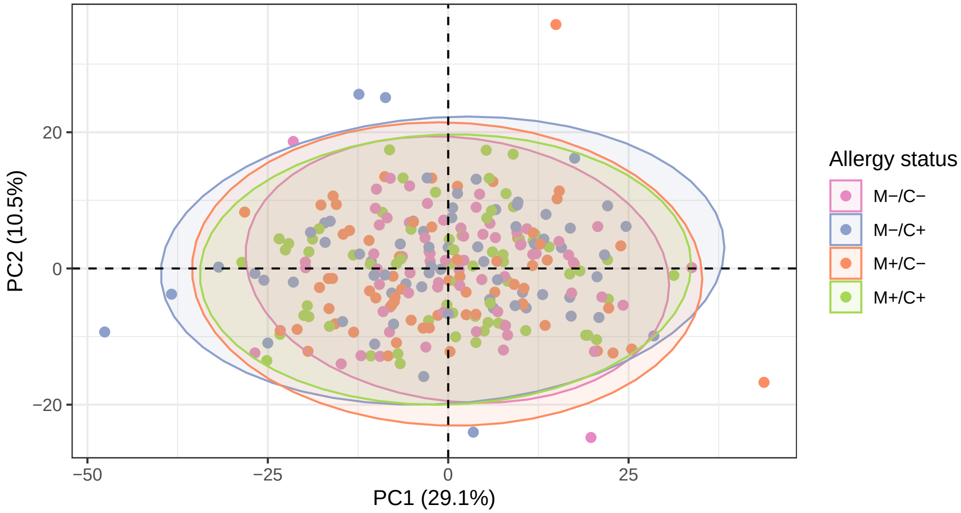


Figure 4.4: Scatter plot of principal component analysis (PCA) representing the human milk protein profile of mother-child dyads from the CHILD Cohort Study. Each point represents one dyad.

4.3.4 Prediction of allergy status using Random Forest models

Random Forest classification was used to discriminate the samples of the different mother-child allergy groups based on the milk protein profile. Two-group models were built for all combinations of maternal allergy status and child allergy status. From the results shown in Table 4.3, it can be noted that all classification models have low discriminating power and that it was therefore not possible to discriminate between the groups. The best accuracy (60%, considered "poor") was obtained for the model that discriminates between the group where only child developed an allergy and the group where both mother and child were non-allergic. Together, this indicates that differences in the human milk proteome between the four groups are weakly reflected in protein abundances, as was also shown by the univariate analysis.

Table 4.3: Outcome of Random Forest models on human milk proteins for the discrimination of groups with different allergy statuses from the CHILD Cohort Study. Comparisons of the groups are labelled according to allergy status, with allergic (M+) and non-allergic (M-) mothers, and allergic (C+) and non-allergic (C-) children.

Comparison					
Group 1	Group 2	Accuracy (%), (*p*-value)	Specificity (%), (*p*-value)	Sensitivity (%), (*p*-value)	AUROC, (*p*-value)
M+/C+	M+/C-	46.7 (0.62)	49.3 (0.39)	44.0 (0.76)	51.7 (0.81)
M+/C+	M-/C+	50.7 (0.34)	45.3 (0.62)	56.0 (0.10)	53.2 (0.63)
M+/C+	M-/C-	50.0 (0.38)	53.3 (0.18)	46.7 (0.60)	51.1 (0.86)
M+/C-	M-/C+	49.3 (0.41)	50.7 (0.29)	48.0 (0.52)	51.4 (0.84)
M+/C-	M-/C-	54.7 (0.14)	56.0 (0.11)	53.3 (0.23)	50.6 (0.94)
M-/C+	M-/C-	60.0 (0.01)	58.7 (0.05)	61.3 (0.01)	58.6 (0.22)

4.3.5 Network analysis

Next, differential network analysis was applied to investigate whether maternal allergy status or the development of allergy in the child is reflected in the milk protein profile in more subtle ways.

Network inference

The protein-protein association networks (Supplementary Figure S4.1) of the different mother-child allergy groups were used to calculate the connectivity of each protein in each mother-child allergy group. The PCLRC algorithm retained mostly positive associations and connectivity represents the number of associations per protein. A comparison of the protein connectivity is visualized in Figure 4.5. What can be observed from this is a pattern that for the groups where at least one of mother or child is allergic, there is stronger interconnectivity between milk proteins when compared to the group where both mother and child are non-allergic.

To investigate this pattern further, proteins with differential connectivity >50 were selected as differentially connected proteins and further investigated (Supplementary File S4.1). This selection was made in a trade-off between the complexity and interpretability of the proteins with high differential connectivity. The selection resulted in 173, 171, and 153 proteins for comparison of non-allergic mother and child with respectively (*i*) allergic mother and non-allergic child, (*ii*) non-allergic mother and allergic child, and (*iii*) allergic mother and child. From these proteins, 95 proteins occurred in all three selections, showing a similarity in differential connectivity. Interestingly, GO overrepresentation analysis of these proteins showed

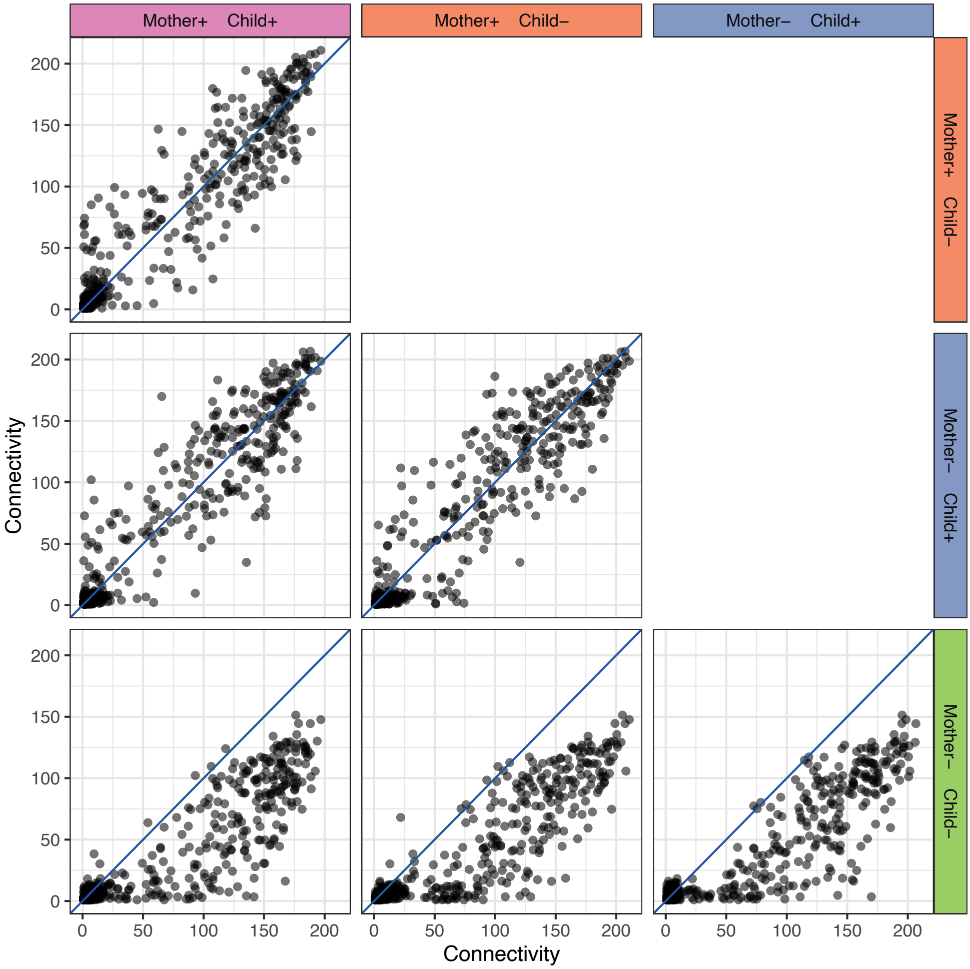


Figure 4.5: Human milk protein connectivity in the different mother-child allergy groups from the CHILD Cohort Study. Each subplot represents a pairwise comparison of protein connectivity in two mother-child allergy groups and each dot represents a single protein. Protein connectivity is obtained from the adjacency matrices build with the PCLRC algorithm and all groups are compared with one another in each subplot. In the labeling of the groups, + indicates allergy and - indicates no allergy.

a significant overrepresentation of proteins involved in translation initiation (p = 1.08×10^{-15}). This overrepresentation is due to ribosomal proteins (n = 24) and translation initiation factors (EIF3A, EIF4A1, EIF5A).

None of the differentially connected proteins showed a difference in level between the different mother-child allergy groups with univariate analysis, indicating the complementarity of these two approaches.

Network modeling

In addition to pairwise comparison of networks, a simultaneous comparison was carried out using COVSCA. With COVSCA, similarities and differences in correlation patterns can be analyzed for a set of networks. In the visualization of the results of COVSCA, each network is a data point in the component space (see Figure 4.6).

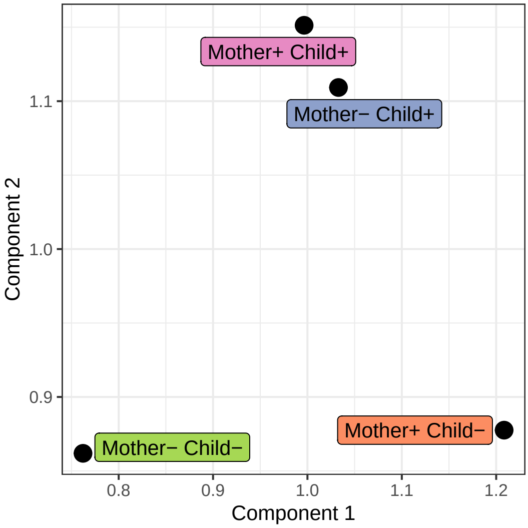


Figure 4.6: Score plot of the COVSCA model for the protein correlation network obtained using PCLRC of different groups based on maternal and child allergy status in the CHILD Cohort Study. Each point represents a protein-protein association network of one mother-child allergy group (+ indicates allergy, and - indicates no allergy). Protein importance for each component is shown in Figure 4.7.

In this comparison, the networks of the four different mother-child allergy groups were compared. From this, it can first be observed that the group with both non-allergic mothers and non-allergic children shows differences in correlation patterns

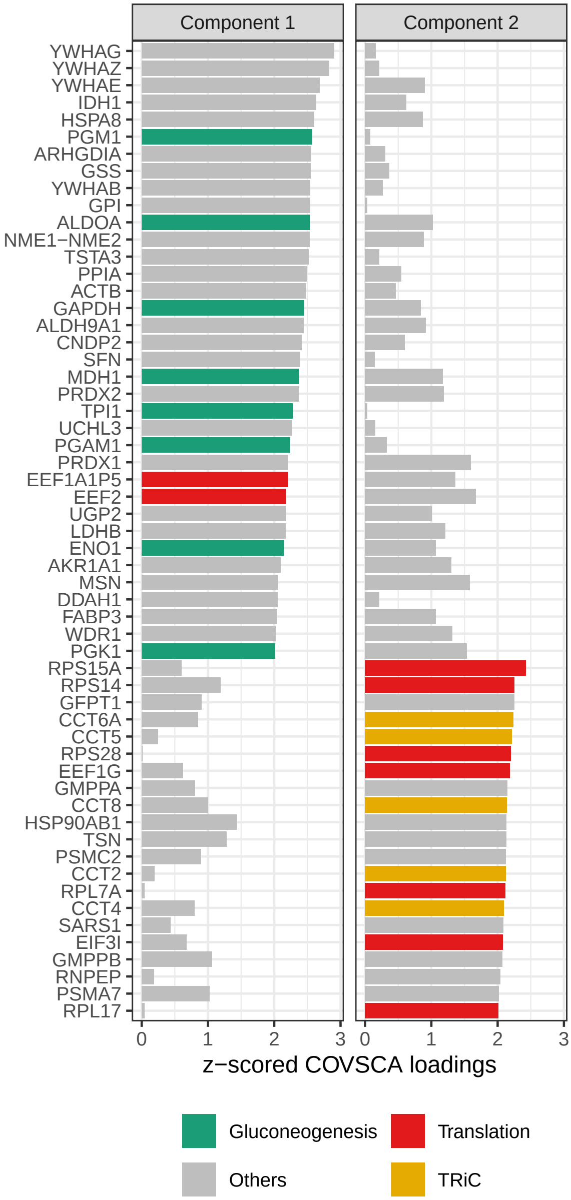


Figure 4.7: *(Caption on next page.)*

Figure 4.7: COVSCA loadings of the COVSCA model of different groups based on maternal and child allergy status in the CHILD Cohort Study. Loadings indicate the importance of each protein for the differences or similarities in correlation patterns observed in the COVSCA score plot (Figure 4.6). Proteins are labeled with gene IDs along the y-axis, and colors indicate shared gene ontology annotations (TRiC: tailless complex polypeptide 1 ring complex).

with all other groups. These differences are present in both COVSCA components. Second, the group where only mothers are allergic shows a difference in network correlation patterns with the non-allergic group on Component 1. Thirdly, groups comprising children who developed an allergy show similarities in correlation patterns on both components.

To investigate these observations further, the loadings of the COVSCA model with $z > |2|$ were examined. These loadings represent the proteins that contributed the most to the difference in correlation patterns between the different networks.

The loadings for Component 1 (see Figure 4.7) are overrepresented by proteins involved in gluconeogenesis ($p = 0.0003$), the synthesis of glucose. This component accounts for separation between the non-allergic group and the groups where either mother, child, or both are allergic. The second component, which drives the separation of the groups on allergy status of the child, shows a significant overrepresentation of proteins involved in the positive regulation of DNA biosynthetic processes ($p = 0.0013$). This is mainly due to 5 members of the tailless complex polypeptide 1 ring complex (TRiC or CCT). In addition, several proteins involved in translation processes show differences in correlation patterns on this component.

4.4 Discussion

We investigated the associations of human milk proteins with maternal and child allergy. Using univariate analysis, predictive modeling and network analysis, several relevant differences and distinctive patterns were found between groups of different allergy status.

4.4.1 Differences in immunoglobulin levels between groups with different allergy statuses

Several proteins showed differences in level when the different mother-child allergy groups were compared. Although none were statistically significant after traditional correction for multiple testing, this does not necessarily imply they are biologically insignificant. It is widely acknowledged that correction methods for multiple hypothesis testing can be too stringent for bottom-up proteomics data [46, 47] because each protein is represented by multiple tryptic peptides. Therefore, we reported both corrected and uncorrected p-values, and discuss the findings.

Most of the differences in protein abundance were found between non-allergic mothers with non-allergic children and the group where only the child developed an allergy. This was also reflected in the accuracy of the Random Forest classification model for these two groups, which was the highest (60%) among all models. The differentially abundant proteins were mainly Ig variable domains. These results point to differences in specific Igs in milk consumed by children who ultimately develop an allergy, and these differences did not seem to be directly linked to the mother's allergy status. This raises two important questions for future research: (*i*) why do these mothers secrete these specific Igs in higher levels in their milk, and (*ii*) could the development of allergy in the child be related to these Igs?

Regarding the first question, the findings in this study show that, regardless of maternal allergy status, milk for children who ultimately develop an allergy contained higher levels of specific Igs. Possibly, other factors that lead to allergy development in the child, such as, health conditions, genetics, dietary patterns, or environmental exposures, also lead to higher levels of Igs in the milk. Another possibility is that infants who would develop an allergy somehow cause higher levels of specific Igs in the milk of the mother. Further research is required to explore these possibilities.

When it comes to the second question, there is contradicting evidence. It has been shown that higher levels of specific Igs in human milk could help in the healthy development of the child's immune system. For example, a study conducted by Ohsaki et al. [48] showed that ovalbumin-specific IgG immune complexes in human milk fed to mice induced tolerance. A study by Lupinek et al. [49] complements this by showing that allergen-specific IgG originating from cord blood or breast milk seemed to protect against allergic sensitization. Nevertheless, Järvinen et al. [16] showed that cow's milk specific IgA levels in human milk did not correlate with the development of cow's milk allergy in the child.

Unfortunately, more details on the function or specificity of the identified Ig variable domains are not available. A complete analysis of the sequence diversity of the antibody repertoire could be done with targeted approaches [50–52], but was outside the scope of the current study.

Notably, soluble CD14, a protein in human milk that may be protective against the development of food allergies [53, 54], was not different between the mother-child allergy groups in our study (uncorrected $p = 0.53$). This and other contradictions with prior studies could be related to our clinical definition of allergy. For example, in a previous study, significant differences were observed in comparing milk from mothers with house dust mite allergy and non-allergic mothers [55]. These differences concerned especially protease inhibitors and apolipoproteins. We did not find these proteins to be different in abundance, which is possibly due to differences in the definition of allergy. Hettinga et al. [55] used a rather strict definition of house dust mite allergy, combined with high immunoglobulin E (IgE) levels in the blood and high environmental exposure to house dust mite, whereas we applied a more heterogeneous definition encompassing diagnosis of multiple allergic conditions.

Relatively few non-human proteins were identified in a low number of samples, and no apparent differences were observed between the different mother-child allergy groups (Table 4.2). Nevertheless, some studies have argued that non-human proteins, especially allergens, play an important role in allergy development [19, 56, 57]. Data from several sources show that most of these proteins originate from the diet and especially from cow's milk or cow's milk products [18, 58]. The difference between prior studies and the current study might be due to differences in, e.g., maternal consumption of dairy products.

4.4.2 Distinctive patterns of connectivity for groups with different allergy statuses

A particularly novel aspect of our study was the network analyses, which demonstrated distinctive association patterns between proteins when groups with different allergy statuses are compared. This points to differences in pathway regulations that are specific for each group. Our most striking finding is the overall lower connectivity observed in the group where both mother and child are non-allergic. This overall difference in connectivity might reflect maternal lifestyle, environmental exposures, or health. For example, a recent study by Yan et al. [59] showed that disease-associated stress brought about the remodeling of protein pathways, leading to a proteome-wide increase in interaction strength and change in connections. Although such an increase in connectivity has not been described before regarding the human milk proteome and allergy, there is evidence showing that allergies are linked to systemic inflammation [60, 61]. Such a state of systemic inflammation might, in turn, result in a change in protein connectivity in the human milk proteome.

Interestingly, in the COVSCA model, we observed that correlation patterns of proteins involved in gluconeogenesis were important for separation between the non-allergic group and the groups where either mother, child, or both are allergic. This points to differences in the regulation of glucose synthesis in the mammary epithelial cells, which could reflect a competition between immune and epithelial cells for glucose, as it is known that during an immune response, immune cells need more glucose [62].

We also detected an overrepresentation of proteins involved in the translation machinery among the differentially connected proteins, which suggests a difference or dysregulation in translation machinery in allergic and non-allergic mothers with children who will develop an allergy. In addition, COVSCA loadings show different correlation patterns between these groups for proteins from the TRiC/CCT complex, which plays an essential role in protein folding and proteostasis [63]. How these TRiC/CCT proteins and proteins from the translation machinery end up in the milk is not known, but they might originate from cells present in the milk [64]. Their difference in connectivity among the different groups might then be due to, for example, different types of cells or a different regulation in these cells. The latter would be in line with Calvano et al. [65], who found that in blood leukocytes from patients with systemic inflammation, there are dysregulations in, amongst others, elongation initiation factors and ribosomal proteins. This could explain the stronger connectivity of the protein synthesis machinery in milk from allergic mothers, who possibly have a higher level of systemic inflammation. Nevertheless, stronger connectivity was also observed in milk from non-allergic mothers with children who would develop an allergy. No studies were found that could explain this observation, and further research should be undertaken to investigate and clarify this.

4.4.3 Limitations and strengths

Although bottom-up proteomics has many advantages, it also has limitations, including the dependence on a database (that is, protein sequences not in the database cannot be identified). This poses a challenge, for example, in the identification of the variable regions of the Igs, of which many sequences are not available in databases. Another limitation is the large number of identifications resulting from these techniques, which requires stringent multiple hypothesis testing in classical univariate data analysis. Finally, although a relatively large sample size was used in the current study, it included considerable clinical heterogeneity in the definition of 'allergy.' The distinct profile of the non-allergic group in both univariate and network analysis suggests that this group is the most homogeneous compared to the allergy groups. It is therefore possible that the clinical heterogeneity of allergy has obscured the effect resulting from specific allergy phenotypes (e.g., food allergies or asthma) if these would have distinct associations with milk proteins.

In summary, this study set out to investigate the human milk proteome and relations with both maternal allergy status and child allergy development. The results show trends in differential abundances of immune-related proteins between the mother-child allergy groups, suggesting a possible difference in the immunological potential of the milk. However, an attempt to exploit these differences to build Random Forest classification models resulted in low predictive power. This was confirmed with multivariate exploratory analysis that did not show differences in the data structure for the different mother-child allergy groups. Interestingly, using a network approach, that enables investigation of protein-protein associations, significant differences were found among the different mother-child allergy groups. The major finding was an overall stronger connectivity of proteins in the milk of allergic mothers and milk for infants who would ultimately develop an allergy, showing that the allergy status of either mother, child, or both is reflected in the interconnectedness of the milk proteins. Collectively, these results show that network analysis complements univariate analysis and classification models to reveal subtler relationships between maternal-child allergy phenotypes and the human milk proteome. Specifically, the network analysis points to a difference in the regulation of pathways for translation and protein folding in these groups, possibly reflecting the physiological state of the mother. Further research is warranted to investigate these associations and the implicated biological pathways to understand their possible functional role in allergy development and prevention.

Supplementary information

The following supplementary information is available and can be accessed through the QR code in Figure 4.8: Network representations of the proteins in milk from the different mother-child allergy status groups from the CHILD Cohort Study (Suplementary Figure S4.1), and a complete overview of the identified proteins (Supplementary File S4.1).

Figure 4.8: Scan this QR code to access the supplementary information, or visit https://figshare.com/s/10a912f058a29df91f39.

References

[1] Turner, P. J. et al. “Global Trends in Anaphylaxis Epidemiology and Clinical Implications”. In: *Journal of Allergy and Clinical Immunology: In Practice* 8 (2020), 1169–1176. DOI: 10.1016/j.jaip.2019.11.027.

[2] Lloyd, C. M. and Saglani, S. “Development of Allergic Immunity in Early Life”. In: *Immunological Reviews* 278 (2017), 101–115. DOI: 10.1111/imr.12562.

[3] Matheson, M. C., Allen, K. J., and Tang, M. L. “Understanding the Evidence for and against the Role of Breastfeeding in Allergy Prevention”. In: *Clinical & Experimental Allergy* 42 (2012), 827–851. DOI: 10.1111/j.1365-2222.2011.03925.x.

[4] Munblit, D. et al. “Human Milk and Allergic Diseases: An Unsolved Puzzle”. In: *Nutrients* 9 (2017), 894. DOI: 10.3390/nu9080894.

[5] Burris, A. D., Pizzarello, C., and Järvinen, K. M. “Immunologic Components in Human Milk and Allergic Diseases with Focus on Food Allergy”. In: *Seminars in Perinatology* 45 (2021), 151386. DOI: 10.1016/j.semperi.2020.151386.

[6] Field, C. J. “The Immunological Components of Human Milk and Their Effect on Immune Development in Infants”. In: *Journal of Nutrition* 135 (2005), 1–4. DOI: 10.1093/jn/135.1.1.

[7] Saarinen, U. M. and Kajosaari, M. “Breastfeeding as Prophylaxis against Atopic Disease: Prospective Follow-up Study until 17 Years Old”. In: *The Lancet* 346 (1995), 1065–1069. DOI: 10.1016/S0140-6736(95)91742-X.

[8] Kull, I. et al. “Breast-Feeding in Relation to Asthma, Lung Function, and Sensitization in Young Schoolchildren”. In: *Journal of Allergy and Clinical Immunology* 125 (2010), 1013–1019. DOI: 10.1016/j.jaci.2010.01.051.

[9] Mihrshahi, S. et al. “The Association between Infant Feeding Practices and Subsequent Atopy among Children with a Family History of Asthma”. In: *Clinical & Experimental Allergy* 37 (2007), 671–679. DOI: 10.1111/j.1365-2222.2007.02696.x.

[10] Annesi-Maesano, I. et al. “Allergic Diseases in Infancy: I - Epidemiology and Current Interpretation”. In: *World Allergy Organization Journal* 14 (2021), 100591. DOI: 10.1016/j.waojou.2021.100591.

[11] Victora, C. G. et al. “Breastfeeding in the 21st Century: Epidemiology, Mechanisms, and Lifelong Effect”. In: *The Lancet* 387 (2016), 475–490. DOI: 10.1016/S0140-6736(15)01024-7.

[12] Andreas, N. J., Kampmann, B., and Mehring Le-Doare, K. "Human Breast Milk: A Review on Its Composition and Bioactivity". In: *Early Human Development* 91 (2015), 629–635. doi: 10.1016/j.earlhumdev.2015.08.013.

[13] Bravi, F. et al. "Impact of Maternal Nutrition on Breast-Milk Composition: A Systematic Review". In: *American Journal of Clinical Nutrition* 104 (2016), 646–662. doi: 10.3945/ajcn.115.120881.

[14] Munblit, D. et al. "Immune Components in Human Milk Are Associated with Early Infant Immunological Health Outcomes: A Prospective Three-Country Analysis". In: *Nutrients* 9 (2017), 532. doi: 10.3390/nu9060532.

[15] Österlund, P. et al. "Eosinophil Cationic Protein in Human Milk Is Associated with Development of Cow's Milk Allergy and Atopic Eczema in Breast-Fed Infants". In: *Pediatric Research* 55 (2004), 296–301. doi: 10.1203/01.PDR.0000106315.00474.6F.

[16] Järvinen, K. M. et al. "Does Low IgA in Human Milk Predispose the Infant to Development of Cow's Milk Allergy?" In: *Pediatric Research* 48 (2000), 457–462. doi: 10.1203/00006450-200010000-00007.

[17] Michel, L. et al. "Novel Approach to Visualize the Inter-Dependencies between Maternal Sensitization, Breast Milk Immune Components and Human Milk Oligosaccharides in the LIFE Child Cohort". In: *PLoS ONE* 15 (2020). Ed. by L. Yeruva, e0230472. doi: 10.1371/journal.pone.0230472.

[18] Dekker, P. M. et al. "Maternal Allergy and the Presence of Nonhuman Proteinaceous Molecules in Human Milk". In: *Nutrients* 12 (2020), 1169. doi: 10.3390/nu12041169.

[19] Macchiaverni, P. et al. "Allergen Shedding in Human Milk: Could It Be Key for Immune System Education and Allergy Prevention?" In: *Journal of Allergy and Clinical Immunology* 148 (2021), 679–688. doi: 10.1016/j.jaci.2021.07.012.

[20] Adel-Patient, K. et al. "Prevention of Allergy to a Major Cow's Milk Allergen by Breastfeeding in Mice Depends on Maternal Immune Status and Oral Exposure during Lactation". In: *Frontiers in Immunology* 11 (2020), 1–10. doi: 10.3389/fimmu.2020.01545.

[21] Subbarao, P. et al. "The Canadian Healthy Infant Longitudinal Development (CHILD) Study: Examining Developmental Origins of Allergy and Asthma". In: *Thorax* 70 (2015), 998–1000. doi: 10.1136/thoraxjnl-2015-207246.

[22] Rosato, A. et al. "From Correlation to Causation: Analysis of Metabolomics Data Using Systems Biology Approaches". In: *Metabolomics* 14 (2018), 37. doi: 10.1007/s11306-018-1335-y.

[23] Saccenti, E. and Svensson, M. “Systems Biology and Biomarkers in Necrotizing Soft Tissue Infections”. In: *Advances in Experimental Medicine and Biology*. Ed. by A. Norrby-Teglund, M. Svensson, and S. Skrede. Vol. 1294. Cham: Springer International Publishing, 2020, 167–186. ISBN: 978-3-030-57616-5. DOI: 10.1007/978-3-030-57616-5_11.

[24] Richards, A. L., Eckhardt, M., and Krogan, N. J. “Mass Spectrometry-based Protein–Protein Interaction Networks for the Study of Human Diseases”. In: *Molecular Systems Biology* 17 (2021), 1–18. DOI: 10.15252/msb.20188792.

[25] Kuzmanov, U. and Emili, A. “Protein-Protein Interaction Networks: Probing Disease Mechanisms Using Model Systems”. In: *Genome Medicine* 5 (2013), 37. DOI: 10.1186/gm441.

[26] Williams, H. C. et al. “The U.K. Working Party’s Diagnostic Criteria for Atopic Dermatitis. III. Independent Hospital Validation”. In: *British Journal of Dermatology* 131 (1994), 406–416. DOI: 10.1111/j.1365-2133.1994.tb08532.x.

[27] Tran, M. M. et al. “Predicting the Atopic March: Results from the Canadian Healthy Infant Longitudinal Development Study”. In: *Journal of Allergy and Clinical Immunology* 141 (2018), 601–607.e8. DOI: 10.1016/j.jaci.2017.08.024.

[28] Moraes, T. J. et al. “The Canadian Healthy Infant Longitudinal Development Birth Cohort Study: Biological Samples and Biobanking”. In: *Paediatric and Perinatal Epidemiology* 29 (2015), 84–92. DOI: 10.1111/ppe.12161.

[29] Dekker, P. M. et al. “Exploring Human Milk Dynamics: Interindividual Variation in Milk Proteome, Peptidome, and Metabolome”. In: *Journal of Proteome Research* 21 (2021), 1002–1016. DOI: 10.1021/acs.jproteome.1c00879.

[30] Liu, Y. et al. “Lactococcus Lactis Mutants Obtained from Laboratory Evolution Showed Elevated Vitamin K2 Content and Enhanced Resistance to Oxidative Stress”. In: *Frontiers in Microbiology* 12 (2021). DOI: 10.3389/fmicb.2021.746770.

[31] Cox, J. and Mann, M. “MaxQuant Enables High Peptide Identification Rates, Individualized p.p.b.-Range Mass Accuracies and Proteome-Wide Protein Quantification”. In: *Nature Biotechnology* 26 (2008), 1367–1372. DOI: 10.1038/nbt.1511.

[32] Bateman, A. et al. “UniProt: The Universal Protein Knowledgebase in 2021”. In: *Nucleic Acids Research* 49 (2021), D480–D489. DOI: 10.1093/nar/gkaa1100.

[33] Wei, R. et al. “GSimp: A Gibbs Sampler Based Left-Censored Missing Value Imputation Approach for Metabolomics Studies”. In: *PLoS Computational Biology* 14 (2018). Ed. by J. Nielsen, e1005973. DOI: 10.1371/journal.pcbi.1005973.

[34] Wilcoxon, F. “Individual Comparisons by Ranking Methods”. In: *Biometrics Bulletin* 1 (1945), 80. DOI: 10.2307/3001968.

[35] Benjamini, Y. and Hochberg, Y. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1995), 289–300. DOI: 10.1111/j.2517-6161.1995.tb02031.x.

[36] Dunn, O. J. “Multiple Comparisons Using Rank Sums”. In: *Technometrics* 6 (1964), 241–252. DOI: 10.1080/00401706.1964.10490181.

[37] Hotelling, H. “Analysis of a Complex of Statistical Variables into Principal Components”. In: *Journal of Educational Psychology* 24 (1933), 417–441. DOI: 10.1037/h0071325.

[38] Lê, S., Josse, J., and Husson, F. “FactoMineR: An R Package for Multivariate Analysis”. In: *Journal of Statistical Software* 25 (2008), 1–18. DOI: 10.18637/jss.v025.i01.

[39] Breiman, L. “Random Forests”. In: *Machine Learning* 45 (2001), 5–32. DOI: 10.1023/A:1010933404324.

[40] Liaw, A. and Wiener, M. “Classification and Regression by randomForest”. In: *R News* 2 (2002), 18–22.

[41] Vignoli, A. et al. “Differential Network Analysis Reveals Molecular Determinants Associated with Blood Pressure and Heart Rate in Healthy Subjects”. In: *Journal of Proteome Research* 20 (2021), 1040–1051. DOI: 10.1021/acs.jproteome.0c00882.

[42] Saccenti, E. et al. “Probabilistic Networks of Blood Metabolites in Healthy Subjects as Indicators of Latent Cardiovascular Risk”. In: *Journal of Proteome Research* 14 (2015), 1101–1111. DOI: 10.1021/pr501075r.

[43] Akhand, M. A. H. et al. “Context Likelihood of Relatedness with Maximal Information Coefficient for Gene Regulatory Network Inference”. In: *2015 18th International Conference on Computer and Information Technology (ICCIT)*. IEEE, 2015, 312–316. ISBN: 978-1-4673-9930-2. DOI: 10.1109/ICCITechn.2015.7488088.

[44] Smilde, A. K. et al. “Covariances Simultaneous Component Analysis: A New Method within a Framework for Modeling Covariances”. In: *Journal of Chemometrics* 29 (2015), 277–288. DOI: 10.1002/cem.2707.

[45] Eden, E. et al. “GOrilla: A Tool for Discovery and Visualization of Enriched GO Terms in Ranked Gene Lists”. In: *BMC Bioinformatics* 10 (2009), 48. doi: 10.1186/1471-2105-10-48.

[46] Handler, D. C. and Haynes, P. A. “Statistics in Proteomics: A Meta-Analysis of 100 Proteomics Papers Published in 2019”. In: *Journal of the American Society for Mass Spectrometry* 31 (2020), 1337–1343. doi: 10.1021/jasms.9b00142.

[47] Pascovici, D. et al. “Multiple Testing Corrections in Quantitative Proteomics: A Useful but Blunt Tool”. In: *Proteomics* 16 (2016), 2448–2453. doi: 10.1002/pmic.201600044.

[48] Ohsaki, A. et al. “Maternal IgG Immune Complexes Induce Food Allergen-Specific Tolerance in Offspring”. In: *Journal of Experimental Medicine* 215 (2018), 91–113. doi: 10.1084/jem.20171163.

[49] Lupinek, C. et al. “Maternal Allergen-Specific IgG Might Protect the Child against Allergic Sensitization”. In: *Journal of Allergy and Clinical Immunology* 144 (2019), 536–548. doi: 10.1016/j.jaci.2018.11.051.

[50] Snapkov, I. et al. “Progress and Challenges in Mass Spectrometry-Based Analysis of Antibody Repertoires”. In: *Trends in Biotechnology* xx (2021), 1–19. doi: 10.1016/j.tibtech.2021.08.006.

[51] Iversen, R. et al. “Strong Clonal Relatedness between Serum and Gut IgA despite Different Plasma Cell Origins”. In: *Cell Reports* 20 (2017), 2357–2367. doi: 10.1016/j.celrep.2017.08.036.

[52] van Keulen, B. J. et al. “Human Milk from Previously Covid-19-Infected Mothers: The Effect of Pasteurization on Specific Antibodies and Neutralization Capacity”. In: *Nutrients* 13 (2021), 1645. doi: 10.3390/nu13051645.

[53] Labéta, M. O. et al. “Innate Recognition of Bacteria in Human Milk Is Mediated by a Milk- Derived Highly Expressed Pattern Recognition Receptor, Soluble CD14”. In: *Journal of Experimental Medicine* 191 (2000), 1807–1812. doi: 10.1084/jem.191.10.1807.

[54] Friedman, N. J. and Zeiger, R. S. “The Role of Breast-Feeding in the Development of Allergies and Asthma”. In: *Journal of Allergy and Clinical Immunology* 115 (2005), 1238–1248. doi: 10.1016/j.jaci.2005.01.069.

[55] Hettinga, K. A. et al. “Difference in the Breast Milk Proteome between Allergic and Non-Allergic Mothers”. In: *PLoS ONE* 10 (2015). Ed. by A. S. Wiley, e0122234. doi: 10.1371/journal.pone.0122234.

[56] Schweitzer, M. et al. “Early Oral Exposure to House Dust Mite Allergen through Breast Milk: A Potential Risk Factor for Allergic Sensitization and Respiratory Allergies in Children”. In: *Journal of Allergy and Clinical Immunology* 139 (2017), 369–372.e10. doi: 10.1016/j.jaci.2016.07.021.

[57] Pastor-Vargas, C. et al. “Sensitive Detection of Major Food Allergens in Breast Milk: First Gateway for Allergenic Contact during Breastfeeding”. In: *Allergy: European Journal of Allergy and Clinical Immunology* 70 (2015), 1024–1027. doi: 10.1111/all.12646.

[58] Zhu, J. et al. “Discovery and Quantification of Nonhuman Proteins in Human Milk”. In: *Journal of Proteome Research* 18 (2019), 225–238. doi: 10.1021/acs.jproteome.8b00550.

[59] Yan, P. et al. “Molecular Stressors Engender Protein Connectivity Dysfunction through Aberrant N-glycosylation of a Chaperone”. In: *Cell Reports* 31 (2020), 107840. doi: 10.1016/j.celrep.2020.107840.

[60] Qi, S. et al. “Effect of Nasal Allergen Challenge in Allergic Rhinitis on Mitochondrial Function of Peripheral Blood Mononuclear Cells”. In: *Annals of Allergy, Asthma and Immunology* 118 (2017), 367–369. doi: 10.1016/j.anai.2016.11.026.

[61] Czarnowicki, T. et al. “Diverse Activation and Differentiation of Multiple B-cell Subsets in Patients with Atopic Dermatitis but Not in Patients with Psoriasis”. In: *Journal of Allergy and Clinical Immunology* 137 (2016), 118–129.e5. doi: 10.1016/j.jaci.2015.08.027.

[62] Habel, J. and Sundrum, A. “Mismatch of Glucose Allocation between Different Life Functions in the Transition Period of Dairy Cows”. In: *Animals* 10 (2020), 1–21. doi: 10.3390/ani10061028.

[63] Grantham, J. “The Molecular Chaperone CCT/TRiC: An Essential Component of Proteostasis and a Potential Modulator of Protein Aggregation”. In: *Frontiers in Genetics* 11 (2020), 1–7. doi: 10.3389/fgene.2020.00172.

[64] Trend, S. et al. “Leukocyte Populations in Human Preterm and Term Breast Milk Iied by Multicolour Flow Cytometry”. In: *PLoS ONE* 10 (2015). Ed. by M. Sperandio, e0135580. doi: 10.1371/journal.pone.0135580.

[65] Calvano, S. E. et al. “A Network-Based Analysis of Systemic Inflammation in Humans”. In: *Nature* 437 (2005), 1032–1037. doi: 10.1038/nature03985.

Chapter 5

Network insights in proteomics and peptidomics of human milk

Pieter M. Dekker, Sjef Boeren, Edoardo Saccenti, Kasper A. Hettinga

Manuscript in preparation to be submitted

Abstract

Proteins and peptides in human milk have bioactive potential to benefit the newborn and support healthy development. However, in ongoing investigations of the health benefits of proteins and peptides, many questions remain unanswered about the nature of these components, how they are formed, and how they end up in the milk. This study aimed to explore and elucidate the complexity of the human milk proteome and peptidome and to investigate associations between these. Proteins and peptides were analyzed with non-targeted nanoLC-Orbitrap-MS/MS in a selection of 300 milk samples from the CHILD Cohort Study. Protein and peptide abundances were integrated, and a network was inferred using Gaussian graphical modeling (GGM), allowing an investigation of direct associations. We showed that signatures of (*i*) specific mechanisms of transport of different groups of proteins, (*ii*) proteolytic degradation by proteases and aminopeptidases, and (*iii*) coagulation and complement activation are present in human milk. These results show the value of an integrated approach in evaluating large-scale omics data sets and provide valuable information for studies that aim to associate protein or peptide profiles from biofluids such as milk with physiological characteristics.

5.1 Introduction

Proteins in human milk have a wide variety of biological functions, ranging from nutrition to immune modulation [1]. Their synthesis can occur in the mammary epithelial cells (MECs), which is the case for the major milk proteins, such as the caseins and α-lactalbumin (ALA) [2]. Other proteins are believed to be synthesized in other parts of the body and are subsequently transferred towards and through the MECs [3]. Shared mechanisms of transfer, location of synthesis, or functioning in the same biological pathways can result in interdependencies between proteins [4].

Parts of the proteins' amino acid sequence can, once detached from the original sequence, exert a completely different biological and biochemical activity. This detachment can occur during proteolytic degradation, resulting in peptides and free amino acids. In human milk, proteolytic degradation starts already when milk is secreted into the alveolar lumen [5] and is due to proteolytic systems comprising proteases, protease activators, and protease inhibitors [6]. Active proteases, such as plasmin (PLG) and kallikrein, hydrolyze peptide bonds between amino acids in the protein sequence, disrupting the protein's primary structure [7].

It is known that peptides play a significant role in many cellular processes in the body, for example, acting as hormones, cytokines, or growth factors [8, 9]. Nevertheless, the role of peptides in human milk is not entirely understood yet. Studies have shown that some of the peptides can exert specific bioactivities, such as immunomodulatory, antimicrobial, antioxidative, or angiotensin-converting enzyme (ACE) inhibitory effects [10, 11]. This could be beneficial for the protection of the mammary gland against infection but also have health benefits for the breastfed infant [12]. Although proteolytic degradation in the digestive system could degrade bioactive peptides, specific peptide sequences might be protected against, or resistant to, further proteolytic degradation. In addition, new bioactive peptides may be formed upon enzymatic digestion in the infants' gastrointestinal tract from either intact proteins or larger peptides [11].

To date, several studies have investigated the human milk peptidome from a mechanistic perspective [13, 14], focusing on cleavage patterns and protease specificity. Although this has provided valuable insights into the human milk peptidome, much is still unknown. Since peptides are a product of larger peptides or proteins, and since the proteolytic systems themselves are part of the proteome, it is expected that relationships exist between the proteome and peptidome. Analysis of these relationships in an integrated approach is an important step in increasing knowledge about the proteolytic activity in human milk.

This study aimed to investigate associations between proteins, between peptides, and across proteins and peptides in human milk. For this, proteomics, and peptidomics data from 300 human milk samples were subjected to network analysis using Gaussian graphic modeling (GGM), and observed associations were discussed. The resulting pairwise partial correlations enable a distinction between indirect and direct associations by adjusting for the contribution of all remaining variables [15]. The rationale behind this approach is that the associations observed in the GGM network can provide information about the biological function of the proteins and peptides and how they are formed or end up in the milk.

5.2 Materials and methods

5.2.1 Sample collection

The CHILD Cohort Study is a Canadian national population-based cohort (https://www.childstudy.ca) in which information was collected over time from parents, infants, and their environment [16]. Pregnant mothers were recruited from the general population from Vancouver, Edmonton, Manitoba, and Toronto. Local Human Research Ethics Boards approved the study protocols, and the study was carried out following the Declaration of Helsinki. All parents provided written informed consent at the time of enrollment in the study.

Human milk samples from a selection of 300 mother-child dyads from the CHILD Cohort Study were used. The selection of these samples was made based on the allergy status of the mother and the infant, including equal numbers of different combinations of mother-child allergy statuses.

Milk samples were collected according to the CHILD protocol [17]. In short, foremilk and hindmilk samples were collected from several feedings during a day and were pooled to minimize within feed variation and diurnal variation. Samples were collected between 6 and 35 weeks post-partum (median = 15.6 weeks, interquartile range (IQR) = 4.6). Samples were stored at 4°C in the home refrigerator and within 24 hours, picked up and transported on ice to the CHILD laboratory. There, samples were aliquoted and stored until further analysis at -80°C. Further transport of the samples was done on dry ice.

5.2.2 Proteomics

Sample preparation

Skimmed milk was obtained by centrifugation at 10,000*g* and 4°C for 30 minutes. Then, skimmed milk was again centrifuged at 1000*g* and 4°C for 10 minutes to remove any remaining lipids. Skimmed milk samples were prepared with filter-aided sample preparation for protein analysis as described before [18].

In addition to the samples from the CHILD Cohort Study, aliquots of a pooled human milk sample were added as a control for technical variation. This sample comprised multiple aliquots of pooled human milk samples from the Dutch Human Milk Bank (Amsterdam, The Netherlands).

LC-MS/MS analysis

Trypsin digested proteins were analyzed with LC-MS/MS as described before, with minor adjustments [19]. In short, 1.5 - 4 μL of tryptic peptide solution was loaded onto a 0.10×250 mm ReproSil-Pur 120 C18-AQ 1.9 μm beads analytical column (prepared in-house) at 825 bar. A gradient from 9 to 34% acetonitrile in water with 0.1% formic acid in 50 min (Thermo nLC1000) was used. Full scan FTMS spectra were obtained using a Q-Exactive HFX (Thermo electron, San Jose, CA, USA) in positive mode between 380 and 1400 *m/z* at resolution 60,000.

The 25 most abundant positively charged peaks (2-5) in the MS scan were isolated and fragmented (HCD) with an isolation width of 1.2 *m/z* and 24% normalized collision energy. MSMS scans were recorded at resolution 15,000 in data-dependent mode with a threshold of 1.2×10^5 and 15 s exclusion for the selected *m/z* ± 10 ppm. Samples were analyzed with a technical replicate added randomly to each 7 injections.

Data processing

The Andromeda search engine of the MaxQuant software v1.6.17.0 was used to analyze the raw LC-MS/MS data [20]. For this, a database was created by an initial MaxQuant run using the full human proteome (downloaded from UniProtKB on 20-01-2021, n = 194,237) [21]. Protein identifiers obtained as identification from this initial run were used to create a human milk database for a second run (n = 24,175), in which also a cow milk protein (n = 1006) and an allergen protein database (n = 721) were added [22].

In MaxQuant, digestion specificity was set to Trypsin/P, with maximally 2 missed cleavages. A fixed propionamide modification was set for cysteines and variable modifications for acetylation of the peptide N-term, deamidation of the side chains of asparagine and glutamine, and oxidation of methionine, with a maximum of five modifications per peptide were set. For each identified protein group, a leading protein was selected as described elsewhere [22].

5.2.3 Peptidomics

Sample preparation

Skimmed milk samples were prepared for peptide analysis as previously described [18]. In short, proteins were removed using precipitation. For this, an equal volume of 200 g/L trichloroacetic acid in milli-Q water was added, followed by centrifugation at 3000*g* for 10 minutes at 4°C. From the supernatant that was obtained, 50 μL was cleaned up using solid phase extraction (SPE) on C18+ Stage tip columns (prepared in-house), as previously described [23, 24]. Eluted peptides were reconstituted in 50 μL of 1 mL/L formic acid in water.

LC-MS/MS analysis

Peptides were analyzed with LC-MS/MS, using the same method as for the protein analysis described above. For the peptidomics analysis, 4 μL of peptide solution was loaded onto the column. Samples were analyzed with a technical replicate added randomly to each 7 injections.

Data processing

The raw LC-MS/MS data files from the peptide analysis were processed similarly as the proteomics data. Differences were the digestion specificity which was set to unspecific with variable modifications for acetylation of the protein N-term, deamidation of the side chains of asparagine and glutamine, and oxidation of methionine, with a maximum of five modifications per peptide. The sequence database which was created for the processing of the proteomics data containing human milk, cow milk and allergen proteins, was used (as described above). Peptide length was set to a minimum of 8 and a maximum of 25 amino acids.

5.2.4 Statistical methods

Statistical analysis and visualizations were, unless specified differently, carried out using R version 4.0.1 [25].

Missing data

MaxQuant proteinGroups (proteomics) or peptides (peptidomics) result files were filtered so that only proteins and peptides that were identified in more than half (>150) of the samples were retained. In this way, a selection of the most prevalent and abundant proteins and peptides was used for further data analysis. Following this, 3 samples were omitted as outliers due to their distinct peptide profile. These samples showed a total peptide abundance several magnitudes higher than the average, possibly due to the occurrence of mastitis. In the remaining data, missing values were imputed using the GSimp package with default parameters [26]. This package uses a Gibbs sampler-based algorithm to impute missing values with the assumption that missing values are not at random (MNAR) and left censored.

Graphical Gaussian modelling (GGM) network analysis

To investigate associations within and between the datasets, network analysis was applied on a combined data matrix, comprising proteins (n = 456) and peptides (n = 1455) in 297 samples.

To build the network, partial correlations were estimated using Gaussian graphical modeling (GGM). The GGMs were built with a shrinkage-based regularization approach for which the *ggm.estimator.pcor* function from the GeneNet package for R was used [27].

This function estimates the partial correlation coefficients in a pairwise manner. Partial correlation coefficients ρ_{ij} describe the pairwise correlation between protein or peptide X_{i} and X_{j} after accounting for their correlation with all other proteins and peptides. This approach accounts for confounders and covariates, indirect associations that are often present in omics data sets. Therefore, this approach enabled the study of direct associations between proteins, peptides, as well as across proteins and peptides.

For the inference of the network, only significant edges were used. To determine the significance of the edges, the built-in empirical Bayes local false discovery rate (fdr) statistic was used [28]. Edges were considered significant if the probability of their “presence” was larger than 0.9 (which is equal to a local fdr <0.1).

Network visualization and clustering

Adjacency matrices with partial correlations from the GGMs were visualized in networks using Cytoscape v3.9.1 [29]. In the GGM network, proteins and peptides are presented as nodes, and GGM-estimated, significant partial correlations are the edges between the nodes. Subsequent clustering of networks was performed using the Leiden algorithm [30], through the clusterMaker2 plugin for Cytoscape [31]. For this clustering, Constant Potts Model was used as quality function in combination with a resolution parameter of 10^{-3}, β value 0.01, and 1000 iterations. Clusters comprising more than 3 nodes were retained for further investigation.

Overrepresentation analysis

To determine whether protein clusters were overrepresented with specific gene ontology (GO) annotations, the GORILLA tool (Gene Ontology enRIchment anaLysis and visuaLizAtion tool) (http://cbl-gorilla.cs.technion.ac.il/) [32] was used. For this, the two-list mode was used, with all identified proteins as background set. *P*-values were corrected with Benjamini-Hochberg correction [33]. An adjusted *p*-value <0.05 was considered significant.

5.3 Results

The LC-MS/MS analysis resulted in the identification of 1629 proteins and 9192 peptides originating from 48 precursor proteins.

After filtering the data on the requirement of identification in more than half of the samples, 456 proteins and 1455 peptides remained. The peptides still originated from 48 precursor proteins. The relative contribution of the precursor proteins to the peptidome showed that the majority of the peptides originated from β-casein (38.5%), polymeric immunoglobulin receptor (PIGR) (10.5%), and butyrophilin subfamily 1 member A1 (BTN1A1) (8.5%), a similar pattern found in previous studies [18, 24].

5.3.1 Network analysis

An association network was inferred by the generation of GGMs. Edges were drawn in the network if partial correlations were significant (local fdr <0.1). This resulted in an initial network comprising 1861 nodes and 16609 edges (corresponding to 0.91% of all possible edges).

With the Leiden algorithm for community detection, 117 clusters were found (Figure 5.1). From all clusters, 42 included both proteins and peptides, whereas 7 clusters only comprised proteins, and 68 clusters only comprised peptides. Most connections (94.9%) were observed between features from the same data set, that is, cross-associations between proteins and between peptides.

Associations between proteins

GO annotation of proteins was used to investigate the overrepresentation of annotations in clusters of associated proteins. An overview of the clusters for which the proteins showed significant overrepresentation of annotations can be found in Table 5.1.

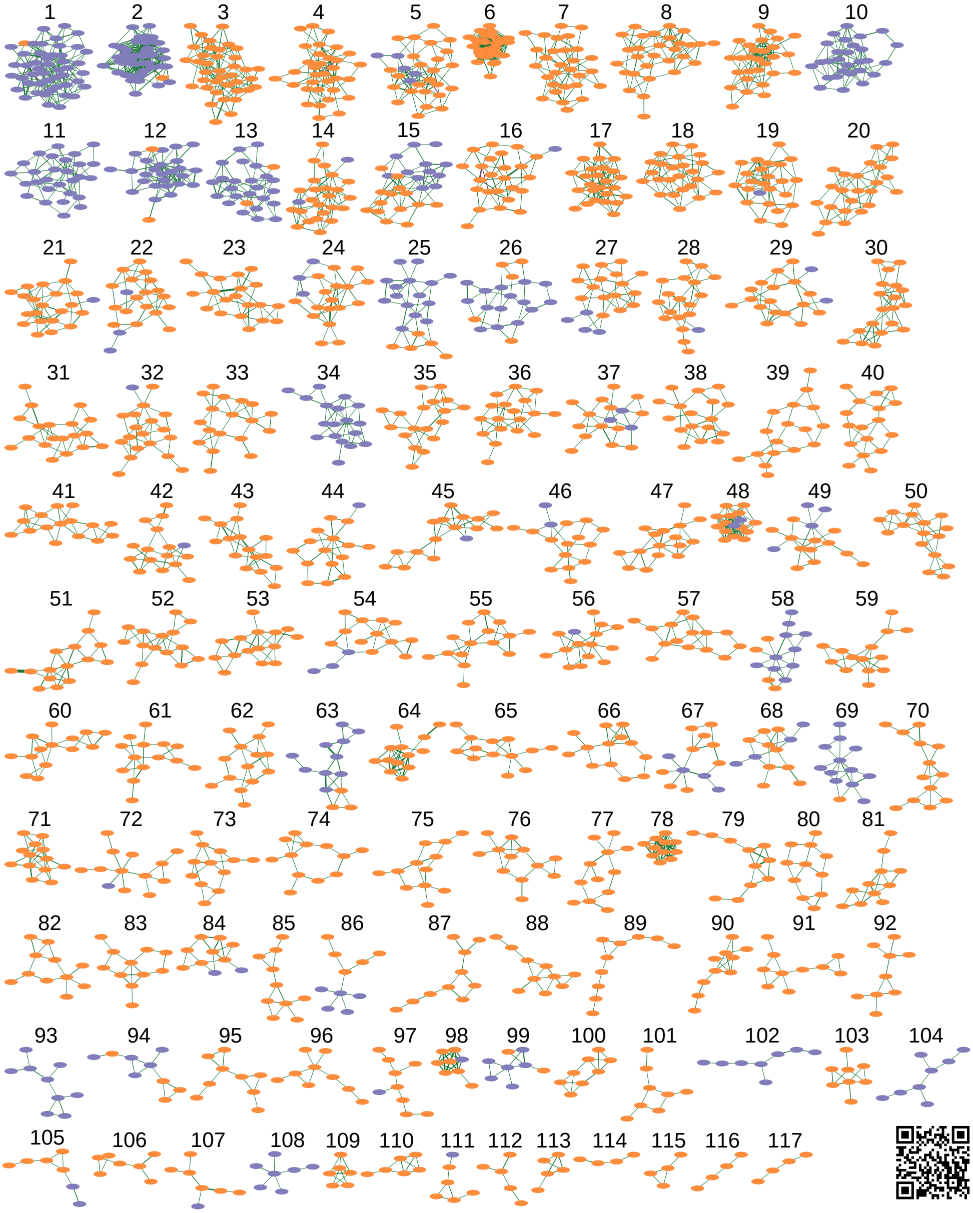


Figure 5.1: *(Caption on next page.)*

Figure 5.1: Network representation of associations between proteins and peptides, calculated with Gaussian graphical models (GGMs). Purple nodes represent proteins, and orange nodes represent peptides. The thickness of the edges is proportional to the partial correlation coefficients from the GGMs. A digital version of this image can be accessed by scanning the QR code.

A significant overrepresentation was found for Cluster 2 of proteins annotated with pentameric immunoglobulin M (IgM) complex (adjusted $p = 0.045$). Nevertheless, it is important to note that annotation was only available for 6 out of the 46 proteins in this cluster. This is because most of these proteins are variable regions of immunoglobulins (Igs), for which GO annotation is not available. Besides the variable regions, this cluster also comprises the heavy constant regions of IgM and IgA, as well as the Ig J chain, which links multimeric IgA and IgM. Therefore, the associations between heavy chains, light chains, J chain, and variable regions indicate their relation as substructures of antibodies. The associations between IgA and IgM also indicate their common origin since IgA and IgM in milk are produced mainly in the plasma cells in the mammary tissue [34]. Subsequent transepithelial transport of these proteins through PIGR results in their secretion in milk.

Cluster 10 shows a significant overrepresentation ($p = 4.84 \times 10^{-9}$) of proteins that are commonly located in blood microparticles, which are microvesicles found in blood. The cluster comprises 28 proteins, of which 17 are annotated as a component of blood microparticles. Among these are, for example, serum albumin (ALB), the major milk protease PLG, and protease inhibitors (5 serine protease inhibitors (SERPINs) and 2 inter-α-trypsin inhibitors (ITIs)). It is generally assumed that PLG, which has an important role in blood coagulation, is blood-derived and transported into the milk from the systemic circulation [35]. In addition, a recent study has shown that one of the SERPINs in this cluster, SERPINA1 (also referred to as α_1-antitrypsin (ATA1)), is synthesized in the liver and enters the milk via direct transmission from the systemic circulation [3]. Considering the overrepresentation of proteins typically found in blood with functional characteristics in blood coagulation, it can be hypothesized that this cluster represents proteins originating from the systemic circulation, all being transported via a transcellular or paracellular pathway through the mammary epithelium.

Proteins that are known to be part of the milk fat globule membrane (MFGM) were jointly present in Cluster 11 [36]. Within the epithelial cell, milk fat globules (MFGs) are surrounded by a single layer membrane which comprises proteins, such as, lactadherin (MFGE8), xanthine dehydrogenase/oxidase (XDH), fatty acid-binding protein (FABP3), and BTN1A1. These proteins are believed to support the

Table 5.1: Overview of significant overrepresentations of Gene Ontology (GO) annotations in the protein clusters shown in Figure 5.1.

Cluster	Category[a]	Annotation	Adjusted p-value	Responsible proteins
2	CC	Pentameric IgM immunoglobulin complex	4.50×10^{-2}	IGHM, IGJ
10	CC	Blood microparticle	4.84×10^{-9}	A1BG, C9, SERPING1, HP, ORM1, ORM2, CP, ITIH1, CFH, ITIH4, PLG, SERPINC1, ALB, VTN, SERPINF2, AFM, HPX
	MF	Serine-type endopeptidase inhibitor activity	2.64×10^{-4}	ITIH1, ANXA2, ITIH4, SERPINC1, SERPING1, SERPINF1, SERPINF2, SERPINA1
	BP	Acute inflammatory response	3.56×10^{-4}	ORM2, ITIH4, SERPINC1, SERPINF2, HP, ORM1, SERPINA1
11	CC	Bounding membrane of organelle	1.25×10^{-2}	STOM, STX3, RAB2A, SNAP23, YKT6, PLIN3, SAR1A, RAB11B, CNP, RAB18, EHD4, GLIPR2, ITPR2
12	CC	Lysosome	6.83×10^{-6}	HEXB, LGMN, CTSD, CTSB, GRN, CTSZ, FUCA1, CTSS, ARSA
13	BP	Protein folding in endoplasmic reticulum	1.93×10^{-3}	CALR, HSP90B1, P4HB, HSPA5, PDIA3
15	CC	Apical plasma membrane	9.30×10^{-5}	CIB1, SLC9A3R1, RDX, EZR, MSN, PODXL
	BP	Regulation of cytoplasmic transport	1.29×10^{-2}	RDX, EZR, MSN
26	MF	Adenyl nucleotide binding	2.30×10^{-2}	DYNC1H1, EPRS, CCT3, FBP16, WARS, CIT, UBA1
63	CC	Chylomicron	8.88×10^{-6}	APOE, APOA2, APOA1, APOB, APOA4
	MF	Intermembrane lipid transfer activity	2.31×10^{-6}	APOE, APOA2, APOA1, APOB, APOA4
	BP	Intermembrane lipid transfer	6.33×10^{-6}	APOE, APOA2, APOA1, APOB, APOA4
93	CC	Ribonucleoprotein complex	2.29×10^{-7}	RPSA, RPL34, RPL6, RPL5, PSMA1, RPS16, RPL3, RPL15, RPS18
	MF	Structural constituent of ribosome	2.71×10^{-7}	RPSA, RPL34, RPL6, RPL5, RPS16, RPL3, RPL15, RPS18
	BP	Protein localization to endoplasmic reticulum	2.97×10^{-7}	RPSA, RPL6, RPL34, RPL5, RPS16, RPL3, RPL15, RPS18
108	CC	Ribosomal subunit	1.76×10^{-3}	RPS4X, RPS9, RPLP2, RPL13, RPL28
	MF	Structural constituent of ribosome	2.44×10^{-3}	RPS4X, RPS9, RPLP2, RPL13, RPL28
	BP	Protein targeting to membrane	1.78×10^{-3}	RPS4X, RPS9, RPLP2, RPL13, RPL28

[a] BP = Biological process, MF = Molecular function, and CC = Cellular component.

MFG in moving towards and binding to the apical plasma membrane [37], which forms the outer bilayer of the MFGM after secretion. The clustering of the typical MFGM proteins (Cluster 11) and the GO overrepresentation of the bounding membrane of organelle as a cellular component ($p = 0.0125$) confirms that these proteins are related to the membrane and thus have a common origin.

Cluster 12 comprises 5 cathepsins (B, C, D, Z, and S), as well as, amongst others, progranulin (GRN), N-acetylglucosamine-6-sulfatase (GNS), and legumain (LGMN). These are proteins typically found in the lysosomal lumen [38], which is also revealed by the GO overrepresentation of the lysosome as a cellular component for this cluster ($p = 6.83 \times 10^{-6}$). In addition, the strong associations observed between these proteins suggest that these proteins are released into the alveolar lumen through a common mechanism such as lysosomal exocytosis [39].

The proteins ezrin (EZR), radixin (RDX), and moesin (MSN) form together the ERM protein family. These ERM proteins can bind with the Na(+)/H(+) exchange regulatory cofactor (NHERF1), and it is known that both ERM and NHERF1 can act as a crosslinker between the actin cytoskeleton and cell membranes by interaction with the intracellular domain of the apical membrane protein podocalyxin (PODXL) [40, 41]. Together, these proteins play an essential role in tissue integrity [42]. The association of these proteins in Cluster 15 suggests a loss of apical membrane from the MECs. Surprisingly, these proteins do not cluster with the typical MFGM proteins (Cluster 11), even though the outer bilayer of the MFGM is formed from the apical membrane. This suggests that the apical membrane found in human milk does not originate only from the MFGM. One explanation for this is the frozen storage of the samples, which results in damaging of cells present in the milk and consequently a release of parts of the apical membrane. A study carried out by Qu et al. [43] confirms this by showing that frozen storage results in increased levels of, amongst others, EZR, MSN, and NHERF1 in milk.

Cluster 63 shows an overrepresentation of chylomicron as cellular location of the proteins ($p = 8.88 \times 10^{-6}$). The cluster comprises 5 different apolipoproteins, including apolipoproteins A1, A2, A4, B100, and E. It was shown in a study with mice, that lipoprotein-particles can be transferred from serum, deliver cholesterol in the MEC, and be secreted into the milk [44]. Considering the overrepresentation of apolipoproteins, Cluster 63 might be an indicator of this mechanism.

Interestingly, this cluster also includes the major α_{s1}-casein variant, while other caseins and major milk proteins are found in Cluster 1. Unlike the other casein subunits, α_{s1}-casein does not decrease over lactation like β-casein and κ-casein [45]. In addition, it was found that this protein is not uniquely expressed in the mammary gland but also in monocytes [46]. The level of this protein in milk might therefore be dependent on other factors and possibly related to the transfer of lipoprotein-particles from the systemic circulation.

Clusters 93 and 108 both show an overrepresentation of ribosomal constituents. Very little is known about why ribosomal proteins are present in milk. They might originate from exosomes, apoptosis of MECs, or intact or damaged cells present in the milk. Nevertheless, their association shows that their levels in milk depend on similar driving factors and possibly share the same secretion mechanism or origin.

The largest protein cluster, Cluster 1, did not show a significant overrepresentation. This cluster comprises the major milk proteins, among which are, for example, ALA, β-casein, κ-casein, lactoferrin (LF), and PIGR. It is known that these proteins are synthesized in the mammary gland [2], a process which is regulated by lactogenic hormones [47]. Considering the strong associations between these proteins, it can be assumed that their expression is related to hormonal regulation of protein synthesis and can be distinguished from the other proteins.

Overall, results indicate that the abundance of the majority of the proteins in human milk depends primarily on the pathway of entering the milk.

Associations between peptides

It is apparent that clusters with peptides often comprise peptide ladders, differing only a few amino acids from the neighboring peptides (Table 5.2 and 5.3). These peptides are presumably formed by aminopeptidases, which cleave a single amino acid off a peptide sequence (exoproteolysis). Several proteins with aminopeptidase activity were identified in the proteomics data of this study, which are in order of average abundance: cytosol aminopeptidase (LAP3), dipeptidyl peptidase 2 (DPP7), aminopeptidase B (RNPEP), leukotriene A-4 hydrolase (LTA4H), and aminopeptidase N (ANPEP). Although not all these aminopeptidases have been identified before in human milk, the activity of aminopeptidases in human milk has been evidenced [48]. The strong association observed between peptides of a peptide ladder suggests that this type of proteolytic degradation occurs in an abundance-dependent manner where the formation of a peptide depends on the abundance of its precursor.

Before cleavage of larger peptides by aminopeptidases is possible, initial proteolytic degradation of proteins needs to occur (endoproteolysis). It has been suggested that endogenous proteases, especially PLG, carry out such proteolysis [13]. PLG is a highly specific protease that hydrolyzes the peptide bond between lysine (K) or arginine (R) in the P1 position and any other amino acid in the P1' position. Interestingly, this matches with several outer C-terminal or N-terminal positions of the peptide ladders observed in the clusters (Table 5.3). Further allocation of proteases to endoproteolytic cleavage sites remains speculative due to the overlapping specificity of proteases as well as the presence of less specific proteases. Nevertheless, the observed associations reveal signatures of endoproteolytic and exoproteolytic degradation of proteins and direct future studies in further investigation of

Table 5.2: Overview of clusters that comprise peptide ladders with more than 5 overlapping peptides. Rows are sorted on precursor protein and sequence position.

Cluster	UniProt ID	Protein name	Sequence range covered	Number of peptides	Average peptide length
3	Q13410	Butyrophilin subfamily 1 member A1	504 - 526	20	14.8
6	Q13410	Butyrophilin subfamily 1 member A1	489 - 526	29	17.3
7	P01833	Polymeric immunoglobulin receptor	605 - 647	7	17.7
8	P05814	β-casein	112* - 161	20	19.8
9	P05814	β-casein	34* - 54*	17	15.8
16	P01833	Polymeric immunoglobulin receptor	601 - 639	6	16.7
17	P07498	κ-casein	63 - 109*	22	17.3
20	P05814	β-casein	16 - 40*	23	16.0
21	P05814	β-casein	97 - 119*	8	15.6
27	P15941	Mucin-1	1207 - 1242	14	15.7
29	P01833	Polymeric immunoglobulin receptor	593* - 622*	8	17.2
30	P10451	Osteopontin	155 - 174	6	16.0
40	P01833	Polymeric immunoglobulin receptor	598* - 639	17	20.3
43	P01833	Polymeric immunoglobulin receptor	622 - 647	9	21.0
45	P05814	β-casein	199 - 226	12	16.9
48	P0C0L5	Complement C4-B	1429* - 1449	9	17.0
50	P05814	β-casein	88 - 113*	16	20.7
52	P05814	β-casein	145 - 175*	14	13.6
53	P05814	β-casein	127 - 148	6	13.7
55	P01833	Polymeric immunoglobulin receptor	604* - 643	8	15.5
57	P47710	α_{s1}-casein	26 - 51*	15	20.2
62	P05814	β-casein	151 - 197	13	17.6
64	P01833	Polymeric immunoglobulin receptor	572* - 603*	9	20.6
66	Q13410	Butyrophilin subfamily 1 member A1	71 - 94*	10	15.1
70	P05814	β-casein	33 - 59	9	21.4
71	P10451	Osteopontin	216 - 246	6	22.0
76	P05814	β-casein	199 - 226	6	18.2
78	P05814	β-casein	34* - 57	11	17.7
83	P19835	Bile salt-activated lipase	21 - 39*	6	16.0
85	P19835	Bile salt-activated lipase	24 - 41	6	13.3
88	P05814	β-casein	16 - 40*	8	19.5
90	P10451	Osteopontin	169* - 203*	8	16.9
91	P10451	Osteopontin	173* - 192	10	14.7
95	P05814	β-casein	104* - 121	9	14.0
98	Q14512	Fibroblast growth factor-binding protein 1	27* - 51*	6	21.7
100	P05814	β-casein	19 - 38*	8	14.5
103	P05814	β-casein	100 - 120	8	15.9

* Cleavage position with the specificity matching the protease plasmin.

the role of these two mechanisms in shaping the human milk peptidome.

Table 5.3: The 3 largest peptide clusters with peptides from a single precursor protein including their sequence positions.

Cluster	UniProt ID	Protein name	Sequence range	Peptide sequence
6	Q13410	Butyrophilin subfamily 1 member A1	489 - 513	QDLSKEIPLSPMGEDSAPRDADTLH.............
			491 - 513	..LSKEIPLSPMGEDSAPRDADTLH.............
			492 - 513	...SKEIPLSPMGEDSAPRDADTLH.............
			493 - 504	....KEIPLSPMGEDS......................
			493 - 507	....KEIPLSPMGEDSAPR...................
			493 - 508	....KEIPLSPMGEDSAPRD..................
			493 - 510	....KEIPLSPMGEDSAPRDAD................
			493 - 511	....KEIPLSPMGEDSAPRDADT...............
			493 - 513	....KEIPLSPMGEDSAPRDADTLH.............
			493 - 514	....KEIPLSPMGEDSAPRDADTLHS............
			493 - 515	....KEIPLSPMGEDSAPRDADTLHSK...........
			494 - 504	.....EIPLSPMGEDS......................
			494 - 507	.....EIPLSPMGEDSAPR...................
			494 - 513	.....EIPLSPMGEDSAPRDADTLH.............
			494 - 514	.....EIPLSPMGEDSAPRDADTLHS............
			494 - 515	.....EIPLSPMGEDSAPRDADTLHSK...........
			495 - 504	......IPLSPMGEDS......................
			495 - 507	......IPLSPMGEDSAPR...................
			495 - 508	......IPLSPMGEDSAPRD..................
			495 - 513	......IPLSPMGEDSAPRDADTLH.............
			495 - 515	......IPLSPMGEDSAPRDADTLHSK...........
			496 - 508	.......PLSPMGEDSAPRD..................
			496 - 513	.......PLSPMGEDSAPRDADTLH.............
			498 - 507	.........SPMGEDSAPR...................
			498 - 508	.........SPMGEDSAPRD..................
			498 - 513	.........SPMGEDSAPRDADTLH.............
			499 - 513	..........PMGEDSAPRDADTLH.............
			501 - 513	............GEDSAPRDADTLH.............
			502 - 526	.............EDSAPRDADTLHSKLIPTQPSQGAP
17	P07498	κ-casein	63 - 78	YQRRPAIAINNPYVPR...............................
			66 - 77	...RPAIAINNPYVP................................
			66 - 78	...RPAIAINNPYVPR...............................
			79 - 100	................TYYANPAVVRPHAQIPQRQYLP.........
			79 - 89	................TYYANPAVVRP....................
			79 - 94	................TYYANPAVVRPHAQIP...............
			79 - 96	................TYYANPAVVRPHAQIPQR.............
			81 - 96	..................YANPAVVRPHAQIPQR.............
			82 - 106	...................ANPAVVRPHAQIPQRQYLPNSHPPT...
			82 - 97	...................ANPAVVRPHAQIPQRQ............

Table 5.3: *(Continued)* The 3 largest peptide clusters with peptides from a single precursor protein including their sequence positions.

Cluster	UniProt ID	Protein name	Sequence range	Peptide sequence
			83 - 100	....................NPAVVRPHAQIPQRQYLP.........
			83 - 96	....................NPAVVRPHAQIPQR.............
			83 - 97	....................NPAVVRPHAQIPQRQ............
			86 - 106	.......................VVRPHAQIPQRQYLPNSHPPT...
			86 - 107	.......................VVRPHAQIPQRQYLPNSHPPTV..
			86 - 108	.......................VVRPHAQIPQRQYLPNSHPPTVV.
			86 - 109	.......................VVRPHAQIPQRQYLPNSHPPTVVR
			91 - 109	............................AQIPQRQYLPNSHPPTVVR
			93 - 107	..............................IPQRQYLPNSHPPTV..
			93 - 108	..............................IPQRQYLPNSHPPTVV.
			93 - 109	..............................IPQRQYLPNSHPPTVVR
			99 - 109	....................................LPNSHPPTVVR
20	P05814	*β*-casein	16 - 32	RETIESLSSSEESITEY........
			16 - 33	RETIESLSSSEESITEYK.......
			16 - 34	RETIESLSSSEESITEYKQ......
			16 - 37	RETIESLSSSEESITEYKQKVE...
			17 - 32	.ETIESLSSSEESITEY........
			17 - 33	.ETIESLSSSEESITEYK.......
			17 - 34	.ETIESLSSSEESITEYKQ......
			17 - 36	.ETIESLSSSEESITEYKQKV....
			17 - 37	.ETIESLSSSEESITEYKQKVE...
			17 - 38	.ETIESLSSSEESITEYKQKVEK..
			17 - 40	.ETIESLSSSEESITEYKQKVEKVK
			18 - 38	..TIESLSSSEESITEYKQKVEK..
			20 - 32	....ESLSSSEESITEY........
			21 - 32	.....SLSSSEESITEY........
			21 - 33	.....SLSSSEESITEYK.......
			21 - 37	.....SLSSSEESITEYKQKVE...
			23 - 32	.......SSSEESITEY........
			23 - 33	.......SSSEESITEYK.......
			23 - 37	.......SSSEESITEYKQKVE...
			24 - 37	........SSEESITEYKQKVE...
			25 - 32	.........SEESITEY........
			25 - 33	.........SEESITEYK.......
			26 - 37	..........EESITEYKQKVE...

Associations between proteins and peptides

Contrary to expectations, it can be noted that from all identified proteins with potential protease activity ($n = 21$), only 5 appear in a cluster with peptides. The most probable explanation for the lack of strong associations between proteases and peptide clusters is the fact that the abundance of a protease is not necessarily equal to or related to its proteolytic activity in the natural milk environment. This can be due to, for example, the protease being present in the zymogen or inactive state, the pH of the milk, or the inhibition of proteases through protease inhibitors.

Although most of the observed associations are associations between molecular features of the same type, that is, between proteins and between peptides, several interesting associations were found between proteins and peptides and will be discussed.

It was found that fibronectin (FN1) and the fibrinogen chains that make up the fibrinogen complex (α (FGA), β (FGB), and γ (FGG)), associated strongly with fibrinogen peptides (Cluster 5 in Figure 5.2). Fibrinogen is a protein complex synthesized in the liver, which plays, together with FN1, a central role in blood coagulation. The coagulation is activated when fibrinopeptides are cleaved off enzymatically from both FGA and FGB by thrombin, resulting in the formation of fibrin and fibrin clots [49]. Surprisingly, the fibrinopeptide of FGA was identified in the peptide data, suggesting that fibrinogen chains present in human milk can occur in the activated form, that is, as fibrin. Degradation of fibrin takes place through proteolysis by PLG [50].

Several of the degradation products can also be observed in Cluster 5, which is an indicator of fibrinolysis to prevent clot formation [51]. This proteolysis is carried out by activated PLG. From Cluster 5 in Figure 5.2, it can be noted that FGA is more degraded, with 28 identified peptides, whereas FGB and FGG have 1 and 3 identified peptides, respectively. This matches with the fact that the FGA chain is cleaved first in the degradation of fibrin [52]. Additionally, α_2-macroglobulin (A2M) appears in Cluster 5. A2M is a protease inhibitor which is known to regulate the degradation of fibrin, by inhibition of PLG.

Together this shows the presence and association of several components and degradation products of blood coagulation in human milk. The origin of these proteins and peptides remains a question. One explanation might be that they are blood-derived, and indirectly end up in the milk through, for example, damage of skin tissue. Nevertheless, it is more probable that they are part of the standard human milk composition, since FGA was identified in 299 of the 300 samples. This also agrees with a study by Green et al. [53] which investigated PLG-deficient mice and suggested that an accumulation of fibrin in the mammary gland could block mammary ducts and ultimately induce involution. Our observation of associations between fibrinogen chains and their degradation products, suggests that, if more fibrinogen is present in the milk, more degradation takes place. From this, it can be hypothesized that the fibrinolysis pathway in milk is present to prevent blocked ducts, and therefore, to maintain lactation.

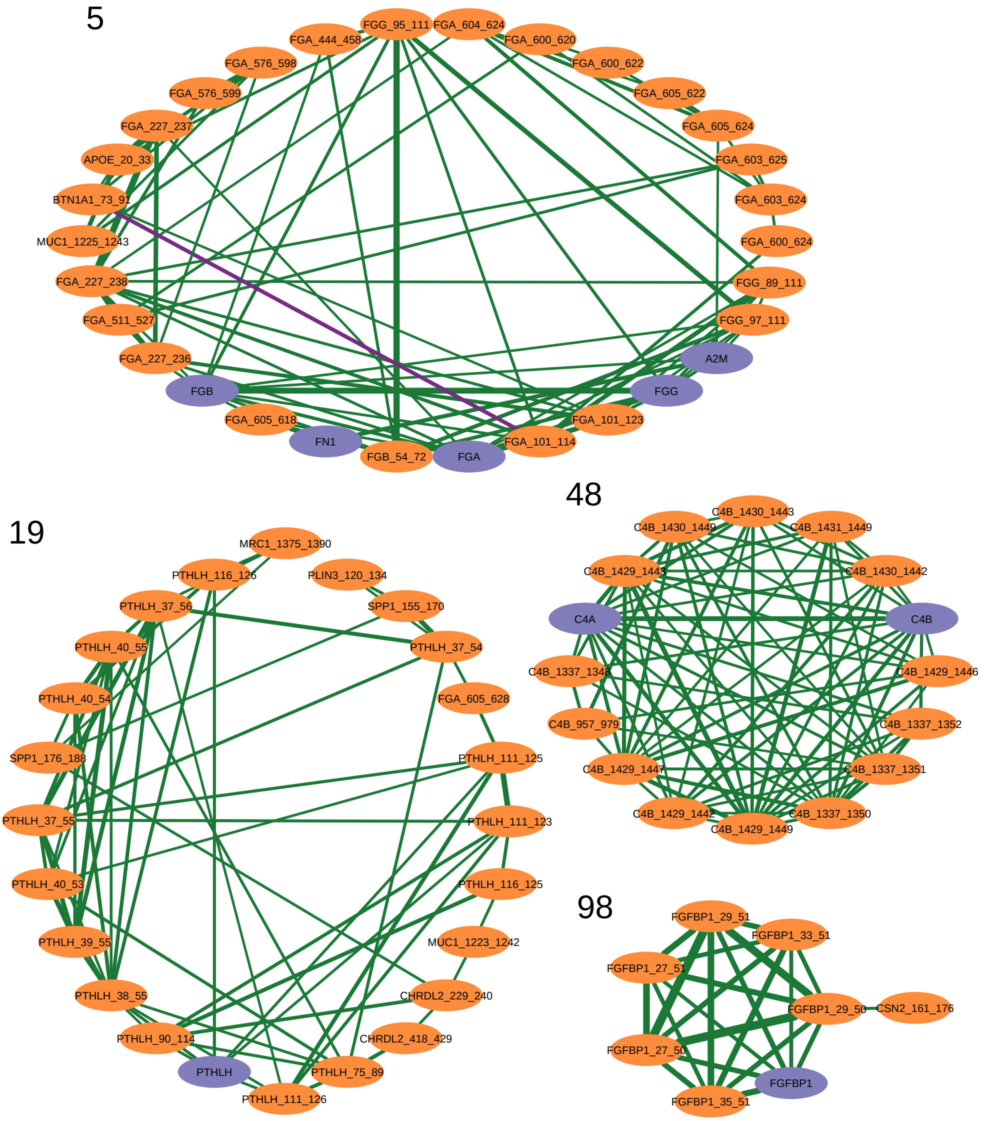


Figure 5.2: Network representation (circular layout) of a selection of associations between proteins and peptides, calculated with Gaussian graphical models (GGMs) and clustered with the Leiden clustering algorithm. Purple nodes represent proteins, and orange nodes represent peptides. The thickness of the edges is proportional to the partial correlation coefficients from the GGMs. The selection of clusters is made from Figure 5.1 with corresponding cluster labels.

Cluster 19 comprises, among others, parathyroid hormone-related protein (PTHLH or PTHrP) and 15 of its peptides. It has been suggested that PTHLH is involved in the regulation of calcium transport through the mammary gland [54]. After synthesis, PTHLH is degraded into three secretory forms, ranging from sequence position 37-72, 74-130, 143-175, respectively (signal peptide is included in the numbering of the sequence positions) [55]. It can be noted from Cluster 19 in Figure 5.2, that peptides derived from all three secretory forms of PTHLH were identified and associated with the precursor protein. Although the functions of the different secretory forms of PTHLH in human milk are not known yet, our results suggest that they are all present in secretory form in the milk and that their abundance depends on the abundance of intact PTHLH.

Cluster 48 (Figure 5.2) shows the association between peptides from complement C4 and the intact C4 isotypes C4A and C4B. These proteins are part of the complement system, a set of proteins, enzymes, and receptors found in blood that plays a key role in the innate immune system's defense against pathogens. Several other proteins from this complex were identified in the proteomics data, among which are C3, C7, C9, plasma protease C1 inhibitor (SERPING1), and complement factors I (CFI) and H (CFH). The presence of complement proteins in human milk has been evidenced before [56] and can boost protective mechanisms of the infants' mucosae [57]. The identification of C4 in the current study covers regions between sequence positions 23 and 1716 (sequence coverage = 77%), whereas the total length of C4 is 1744 amino acids. This provides evidence for the presence of intact C4 in human milk. C4 can participate in the classical and lectin complement pathways and is cleaved into fragments upon activation [58]. In the peptide fraction, all but one of the C4 peptides that were identified originate from a specific region (between positions 1337 and 1449), which is the C-terminal part of the C4b fragment (position 757-1446). This C-terminal region of C4b is cleaved off in the formation of the C4d fragment (position 957-1336). Together, this shows that the identified C4 peptides are byproducts of the activation cascade of C4 [51]. The association of these peptides with intact C4 suggests that C4 activation in human milk is dependent on the abundance of intact C4.

Fibroblast growth factor-binding protein 1 (FGFBP1) is a protein that can bind fibroblast growth factors (FGFs), a family of cell signaling proteins, and release them from the extracellular matrix. All identified FGFBP1 peptides ($n = 6$) originate from the N-terminal region of the protein (between positions 24 and 51), which is also covered in the identification of the protein. Of the cleavage sites of the peptides, 15 out of 24 have lysine in position P1, suggesting that PLG is responsible for most of the cleavages. The strong association between the peptides and their intact protein (Cluster 98 in Figure 5.2) suggests a specific proteolytic degradation which is not related to degradation of other proteins in milk. Such degradation might be related to the role of FGFBP1 in protecting FGF against degradation [59], but this remains speculative, since no previous studies were found on proteolytic degradation of FGFBP1.

Overall, the protein-peptide associations revealed several mechanisms of specific proteolytic degradation that takes place in human milk. Specifically, degradation of fibrin(ogen), PTHLH, complement C4, and FGFBP1, showed to be associated with the abundance of their precursor protein and different from proteolytic degradation from the major precursor proteins in milk.

5.4 Conclusions

This study used a network approach to assess associations between the human milk peptide and protein profile. Strong associations were found especially between proteins and between peptides, and some across proteins and peptides. Furthermore, the used network approach revealed clusters of proteins in human milk that could be linked to their transport mechanisms through the mammary epithelium. In addition, associations between peptides elucidated the proteolytic degradation through aminopeptidases, which showed to be dependent on the abundance of the precursor peptide. Lastly, associations observed between coagulation and complement activation proteins, their proteolytic enzymes, and their peptides suggests that these two pathways are activated in milk.

References

[1] Donovan, S. M. “Human Milk Proteins: Composition and Physiological Significance”. In: *Nestle Nutrition Institute Workshop Series*. Vol. 90. 2019, 93–101. doi: 10.1159/000490298.

[2] Vilotte, J. L. et al. “Genetics and Biosynthesis of Milk Proteins”. In: *Advanced Dairy Chemistry*. Ed. by P. L. H. McSweeney and P. F. Fox. Fourth. Vol. 1A: Proteins: Basic Aspects. Boston, MA: Springer US, 2013, 431–461. isbn: 978-1-4614-4714-6. doi: 10.1007/978-1-4614-4714-6_14.

[3] Jager, S. et al. “Proteoform Profiles Reveal That Alpha-1-Antitrypsin in Human Serum and Milk Is Derived from a Common Source”. In: *Frontiers in Molecular Biosciences* 9 (2022), 1–10. doi: 10.3389/fmolb.2022.858856.

[4] Vella, D. et al. “From Protein-Protein Interactions to Protein Co-Expression Networks: A New Perspective to Evaluate Large-Scale Proteomic Data”. In: *Eurasip Journal on Bioinformatics and Systems Biology* 2017 (2017), 6. doi: 10.1186/s13637-017-0059-z.

[5] Nielsen, S. D., Beverly, R. L., and Dallas, D. C. “Milk Proteins Are Predigested within the Human Mammary Gland”. In: *Journal of Mammary Gland Biology and Neoplasia* 22 (2017), 251–261. doi: 10.1007/s10911-018-9388-0.

[6] Dallas, D. C., Murray, N. M., and Gan, J. “Proteolytic Systems in Milk: Perspectives on the Evolutionary Function within the Mammary Gland and the Infant”. In: *Journal of Mammary Gland Biology and Neoplasia* 20 (2015), 133–147. doi: 10.1007/s10911-015-9334-3.

[7] Kelly, A. L. and Larsen, Lotte Bach, eds. *Agents of Change Enzymes in Milk and Dairy Products*. First. Food Engineering Series. Cham: Springer, 2021. isbn: 978-3-030-55481-1.

[8] Schulte, I. et al. “Peptides in Body Fluids and Tissues as Markers of Disease”. In: *Expert Review of Molecular Diagnostics* 5 (2005), 145–157. doi: 10.1586/14737159.5.2.145.

[9] Foreman, R. E. et al. “Peptidomics: A Review of Clinical Applications and Methodologies”. In: *Journal of Proteome Research* 20 (2021), 3782–3797. doi: 10.1021/acs.jproteome.1c00295.

[10] Nielsen, S. D. et al. “Release of Functional Peptides from Mother’s Milk and Fortifier Proteins in the Premature Infant Stomach”. In: *PLoS ONE* 13 (2018). doi: 10.1371/journal.pone.0208204.

[11] Dallas, D. C. et al. “A Peptidomic Analysis of Human Milk Digestion in the Infant Stomach Reveals Protein-Specific Degradation Patterns”. In: *Journal of Nutrition* 144 (2014), 815–820. doi: 10.3945/jn.113.185793.

[12] Wada, Y. and Lönnerdal, B. “Bioactive Peptides Derived from Human Milk Proteins: An Update”. In: *Current Opinion in Clinical Nutrition and Metabolic Care* 23 (2020), 217–222. doi: 10.1097/MCO.0000000000000642.

[13] Guerrero, A. et al. “Mechanistic Peptidomics: Factors That Dictate Specificity in the Formation of Endogenous Peptides in Human Milk”. In: *Molecular and Cellular Proteomics* 13 (2014), 3343–3351. doi: 10.1074/mcp.M113.036194.

[14] Khaldi, N. et al. “Predicting the Important Enzymes in Human Breast Milk Digestion”. In: *Journal of Agricultural and Food Chemistry* 62 (2014), 7225–7232. doi: 10.1021/jf405601e.

[15] Krumsiek, J. et al. “Gaussian Graphical Modeling Reconstructs Pathway Reactions from High-Throughput Metabolomics Data”. In: *BMC Systems Biology* 5 (2011). doi: 10.1186/1752-0509-5-21.

[16] Subbarao, P. et al. “The Canadian Healthy Infant Longitudinal Development (CHILD) Study: Examining Developmental Origins of Allergy and Asthma”. In: *Thorax* 70 (2015), 998–1000. doi: 10.1136/thoraxjnl-2015-207246.

[17] Moraes, T. J. et al. “The Canadian Healthy Infant Longitudinal Development Birth Cohort Study: Biological Samples and Biobanking”. In: *Paediatric and Perinatal Epidemiology* 29 (2015), 84–92. doi: 10.1111/ppe.12161.

[18] Dekker, P. M. et al. “Exploring Human Milk Dynamics: Interindividual Variation in Milk Proteome, Peptidome, and Metabolome”. In: *Journal of Proteome Research* 21 (2021), 1002–1016. doi: 10.1021/acs.jproteome.1c00879.

[19] Liu, Y. et al. “Lactococcus Lactis Mutants Obtained from Laboratory Evolution Showed Elevated Vitamin K2 Content and Enhanced Resistance to Oxidative Stress”. In: *Frontiers in Microbiology* 12 (2021). doi: 10.3389/fmicb.2021.746770.

[20] Cox, J. and Mann, M. “MaxQuant Enables High Peptide Identification Rates, Individualized p.p.b.-Range Mass Accuracies and Proteome-Wide Protein Quantification”. In: *Nature Biotechnology* 26 (2008), 1367–1372. doi: 10.1038/nbt.1511.

[21] Bateman, A. et al. “UniProt: The Universal Protein Knowledgebase in 2021”. In: *Nucleic Acids Research* 49 (2021), D480–D489. doi: 10.1093/nar/gkaa1100.

[22] Dekker, P. M. et al. “Maternal Allergy and the Presence of Nonhuman Proteinaceous Molecules in Human Milk”. In: *Nutrients* 12 (2020), 1169. doi: 10.3390/nu12041169.

[23] Lu, J. et al. “Filter-Aided Sample Preparation with Dimethyl Labeling to Identify and Quantify Milk Fat Globule Membrane Proteins”. In: *Journal of Proteomics* 75 (2011), 34–43. doi: 10.1016/j.jprot.2011.07.031.

[24] Dingess, K. A. et al. “Human Milk Peptides Differentiate between the Preterm and Term Infant and across Varying Lactational Stages”. In: *Food and Function* 8 (2017), 3769–3782. doi: 10.1039/c7fo00539c.

[25] Development Team Core. *R. A Language and Environment for Statistical Computing*. 2020.

[26] Wei, R. et al. “GSimp: A Gibbs Sampler Based Left-Censored Missing Value Imputation Approach for Metabolomics Studies”. In: *PLoS Computational Biology* 14 (2018). Ed. by J. Nielsen, e1005973. doi: 10.1371/journal.pcbi.1005973.

[27] Schäfer, J., Opgen-Rhein, R., and Strimmer, K. “Reverse Engineering Genetic Networks Using the {GeneNet} Package”. In: *R News* 6 (2006), 50–53.

[28] Efron, B. “Large-Scale Simultaneous Hypothesis Testing: The Choice of a Null Hypothesis”. In: *Journal of the American Statistical Association* 99 (2004), 96–104. doi: 10.1198/016214504000000089.

[29] Shannon, P. et al. “Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks”. In: *Genome Research* 13 (2003), 2498–2504. doi: 10.1101/gr.1239303.

[30] Traag, V. A., Waltman, L., and van Eck, N. J. “From Louvain to Leiden: Guaranteeing Well-Connected Communities”. In: *Scientific Reports* 9 (2019), 1–12. doi: 10.1038/s41598-019-41695-z.

[31] Morris, J. H. et al. “ClusterMaker: A Multi-Algorithm Clustering Plugin for Cytoscape”. In: *BMC Bioinformatics* 12 (2011), 1–14. doi: 10.1186/1471-2105-12-436.

[32] Eden, E. et al. “GOrilla: A Tool for Discovery and Visualization of Enriched GO Terms in Ranked Gene Lists”. In: *BMC Bioinformatics* 10 (2009), 48. doi: 10.1186/1471-2105-10-48.

[33] Benjamini, Y. and Hochberg, Y. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1995), 289–300. doi: 10.1111/j.2517-6161.1995.tb02031.x.

[34] Hurley, W. L. and Theil, P. K. “Perspectives on Immunoglobulins in Colostrum and Milk”. In: *Nutrients* 3 (2011), 442–474. doi: 10.3390/nu3040442.

[35] Kelly, A. L., O’Flaherty, F., and Fox, P. F. “Indigenous Proteolytic Enzymes in Milk: A Brief Overview of the Present State of Knowledge”. In: *International Dairy Journal* 16 (2006), 563–572. doi: 10.1016/j.idairyj.2005.10.019.

[36] Vanderghem, C. et al. “Study on the Susceptibility of the Bovine Milk Fat Globule Membrane Proteins to Enzymatic Hydrolysis and Organization of Some of the Proteins”. In: *International Dairy Journal* 21 (2011), 312–318. doi: 10.1016/j.idairyj.2010.12.006.

[37] Monks, J. et al. “Xanthine Oxidoreductase Mediates Membrane Docking of Milk-Fat Droplets but Is Not Essential for Apocrine Lipid Secretion”. In: *Journal of Physiology* 594 (2016), 5899–5921. doi: 10.1113/JP272390.

[38] Lübke, T., Lobel, P., and Sleat, D. E. “Proteomics of the Lysosome”. In: *Biochimica et Biophysica Acta - Molecular Cell Research* 1793 (2009), 625–635. doi: 10.1016/j.bbamcr.2008.09.018.

[39] Tancini, B. et al. “Lysosomal Exocytosis: The Extracellular Role of an Intracellular Organelle”. In: *Membranes* 10 (2020), 1–21. doi: 10.3390/membranes10120406.

[40] Le Tran, N., Wang, Y., and Nie, G. “Podocalyxin in Normal Tissue and Epithelial Cancer”. In: *Cancers* 13 (2021). doi: 10.3390/cancers13122863.

[41] Jiang, L. et al. “CLIC Proteins, Ezrin, Radixin, Moesin and the Coupling of Membranes to the Actin Cytoskeleton: A Smoking Gun?” In: *Biochimica et Biophysica Acta - Biomembranes* 1838 (2014), 643–657. doi: 10.1016/j.bbamem.2013.05.025.

[42] Horrillo, A. et al. “Loss of Endothelial Barrier Integrity in Mice with Conditional Ablation of Podocalyxin (Podxl) in Endothelial Cells”. In: *European Journal of Cell Biology* 95 (2016), 265–276. doi: 10.1016/j.ejcb.2016.04.006.

[43] Qu, J. et al. “Changes in Bioactive Proteins and Serum Proteome of Human Milk under Different Frozen Storage”. In: *Food Chemistry* 352 (2021), 129436. doi: 10.1016/j.foodchem.2021.129436.

[44] Monks, J. et al. “A Lipoprotein-Containing Particle Is Transferred from the Serum across the Mammary Epithelium into the Milk of Lactating Mice”. In: *Journal of Lipid Research* 42 (2001), 686–696. doi: 10.1016/s0022-2275(20)31630-8.

[45] Liao, Y. et al. “Absolute Quantification of Human Milk Caseins and the Whey /Casein Ratio during the First Year of Lactation”. In: *Journal of Proteome Research* 16 (2017), 4113–4121. doi: 10.1021/acs.jproteome.7b00486.

[46] Vordenbäumen, S. et al. "Casein α S1 Is Expressed by Human Monocytes and Upregulates the Production of GM-CSF via P38 MAPK". In: *The Journal of Immunology* 186 (2011), 592–601. doi: 10.4049/jimmunol.1001461.

[47] Rhoads, R. E. and Grudzien-Nogalska, E. "Translational Regulation of Milk Protein Synthesis at Secretory Activation". In: *Journal of Mammary Gland Biology and Neoplasia* 12 (2007), 283–292. doi: 10.1007/s10911-007-9058-0.

[48] Demers-Mathieu, V. et al. "Analysis of Milk from Mothers Who Delivered Prematurely Reveals Few Changes in Proteases and Protease Inhibitors across Gestational Age at Birth and Infant Postnatal Age". In: *Journal of Nutrition* 147 (2017), 1152–1159. doi: 10.3945/jn.116.244798.

[49] Scheraga, H. A. "The Thrombin-Fibrinogen Interaction". In: *Biophysical Chemistry* 112 (2004), 117–130. doi: 10.1016/j.bpc.2004.07.011.

[50] Horan, J. T. and Francis, C. W. "Fibrin Degradation Products, Fibrin Monomer and Soluble Fibrin in Disseminated Intravascular Coagulation". In: *Seminars in Thrombosis and Hemostasis* 27 (2001), 657–666. doi: 10.1055/s-2001-18870.

[51] Koomen, J. M. et al. "Direct Tandem Mass Spectrometry Reveals Limitations in Protein Profiling Experiments for Plasma Biomarker Discovery". In: *Journal of Proteome Research* 4 (2005), 972–981. doi: 10.1021/pr050046x.

[52] Kirschbaum, N. E. and Budzynski, A. Z. "A Unique Proteolytic Fragment of Human Fibrinogen Containing the Aα COOH-terminal Domain of the Native Molecule". In: *Journal of Biological Chemistry* 265 (1990), 13669–13676. doi: 10.1016/s0021-9258(18)77401-2.

[53] Green, K. A. et al. "Lack of Plasminogen Leads to Milk Stasis and Premature Mammary Gland Involution during Lactation". In: *Developmental Biology* 299 (2006), 164–175. doi: 10.1016/j.ydbio.2006.07.021.

[54] Seki, K. et al. "Parathyroid-Hormone-Related Protein in Human Milk and Its Relation to Milk Calcium". In: *Gynecologic and Obstetric Investigation* 44 (1997), 102–106. doi: 10.1159/000291496.

[55] Plawner, L. L. et al. "Cell Type-Specific Secretion of Parathyroid Hormone-Related Protein via the Regulated versus the Constitutive Secretory Pathway". In: *Journal of Biological Chemistry* 270 (1995), 14078–14084. doi: 10.1074/jbc.270.23.14078.

[56] Ogundele, M. O. "Role and Significance of the Complement System in Mucosal Immunity: Particular Reference to the Human Breast Milk Complement". In: *Immunology and Cell Biology* 79 (2001), 1–10. doi: 10.1046/j.1440-1711.2001.00976.x.

[57] Noel, G. et al. “Human Breast Milk Enhances Intestinal Mucosal Barrier Function and Innate Immunity in a Healthy Pediatric Human Enteroid Model”. In: *Frontiers in Cell and Developmental Biology* 9 (2021), 1–15. doi: 10.3389/fcell.2021.685171.

[58] Wang, H. and Liu, M. “Complement C4, Infections, and Autoimmune Diseases”. In: *Frontiers in Immunology* 12 (2021), 1–15. doi: 10.3389/fimmu.2021.694928.

[59] Huang, W. et al. “Sox12, a Direct Target of FoxQ1, Promotes Hepatocellular Carcinoma Metastasis through up-Regulating Twist1 and FGFBP1”. In: *Hepatology* 61 (2015), 1920–1933. doi: 10.1002/hep.27756.

Chapter 6

Discussion

6.1 Introduction

Human milk comprises a variety of proteins that can potentially contribute to the development of the infant's immune system. In addition, peptides, the degradation products of the proteins, can exert immunomodulatory effects and could therefore also play a role in this. The levels of proteins and peptides in milk are determined by different factors, resulting in complex compositional profiles with substantial interindividual differences.

The account given in this thesis focuses on obtaining a better understanding of the complexity and variation in the protein and peptide profile in human milk and how this relates to the allergy status of both mother and child. These topics were split up accordingly into two overarching objectives (**Chapter 1** and Figure 1.9).

The first objective was to investigate the characteristics and interindividual variation of the proteins and peptides in human milk. This was achieved by:

- Investigating the presence of non-human proteins (**Chapter** 2).
- Exploring the interindividual variation in proteins and peptides (**Chapter 3**).
- Studying the associations between and across proteins and peptides (**Chapter 5**).

The second objective was to relate the protein and peptide profile with the allergy status of both mother and child. This was achieved by:

- Investigating the relation between maternal allergy and non-human proteins in the milk (**Chapter** 2).
- Examining the relation between maternal and child allergy and the complete protein profile (**Chapter** 4).

In this chapter, the results of the previous chapters are combined, and further data analysis is added in order to discuss the findings in relation to the objectives. Finally, the chapter ends with conclusions, implications, and recommendations for future studies.

6.2 Methodological considerations

6.2.1 Added value of integrative system biology approach

In this thesis, several different methods were used to analyze milk omics data. In **Chapter** 2 differences were observed in, among others, the abundance of bovine β-lactoglobulin (BLG) when comparing milk from non-allergic and allergic mothers using classical data analysis. This finding generated several new hypotheses that would need validation, for example, regarding the unknown mechanisms of transport of these proteins through the mother's body. With the approach used in this chapter, relevant differences were revealed that resulted in the generation of several new hypotheses.

A drawback of classical methods, such as univariate analysis, is the problem of multiple testing corrections. The number of variables in typical omics data ranges from several hundred to thousands. Consequently, applying univariate analysis to such data will have a higher probability of false-positive results, for which correction is needed. However, although correction methods decrease the probability of false positives, they can also increase the probability of false negatives. In addition, correction for multiple testing in proteomics data can be less effective due to several specific factors, as highlighted by Pascovici et al. [1]. Together, the use of corrected *p*-values as "cookie-cutter" evidence can restrict the effectiveness of univariate analysis in revealing relevant differences between groups of samples.

In an attempt to go beyond the approach of classical data analysis, the field of systems biology offers a deductive or top-down approach that allows contextualization of results and can therefore guide the biological interpretation and generation of hypotheses [2]. In this approach, patterns extracted from omics profiles are used to generate hypotheses on the regulatory mechanisms that play a role in cellular systems [3]. Ultimately, this approach aims to understand the behavior of a system by defining interactions among its different components [3]. Examples of typical systems biology tools used in this thesis are network inference and network analysis. Using these tools, it was shown that the associations of proteins and peptides in networks help decipher the complexity of the human milk proteome and peptidome and provide additional information on top of univariate and multivariate analysis.

In **Chapter 3**, weighted correlation network analysis (WGCNA) [4] was used to investigate whether specific patterns of interindividual variation occur in subgroups of correlating proteins and peptides. The network analysis allowed clustering of associated proteins and peptides, showing that, for example, high interindividual variation is especially present in clusters of associated peptides originating from the mid and C-terminal regions of β-casein. Peptides from the N-terminal region showed a distinct pattern of cross-correlations from the rest of the protein sequence and showed an overall lower interindividual variation. Therefore, the application of network analysis gave context to the observed interindividual variation of peptides by revealing differences in proteolytic degradation between different proteins and between different sequence regions within proteins.

The results presented in **Chapter 4** show how network analysis complements univariate analysis, multivariate analysis, and classification models. Protein connectivity revealed that proteins functioning in the translation machinery are differently connected in the different groups of mother-child allergy status. This observation could not be made based on the outcome of the other data analyses.

Further evidence of the added value of an integrative approach in the data analysis was given in **Chapter 5**. Omics data were integrated using associations obtained from Gaussian graphical models (GGMs) [5], which were used to estimate partial correlations. Partial correlation is a measure for the degree of association corrected for the other variables present in the data. Subsequent network inference allowed, therefore, for the investigation of direct associations between proteins and peptides. This method revealed, amongst others, that the abundance of proteins in the milk is more dependent on their mechanism of transport and origin than on their function, thereby providing relevant insights into what determines the human milk protein and peptide composition.

Together, systems biology-based analysis of omics data provides insights into the biological mechanisms of the secretion of proteins into milk and the proteolytic degradation of these proteins into peptides. This shows the potential of this approach to enhance the interpretation of omics data and to generate hypotheses based on mechanistic insights. Therefore, it is recommended to apply this in future studies so that results can be put in context and more insight is obtained into biological mechanisms.

6.2.2 Limitations of shotgun proteomics and data processing

The use of shotgun or bottom-up proteomics in this study, in which proteins are digested by an enzyme with strict cleavage specificity before analysis, has pros and cons.

First of all, the untargeted shotgun proteomics and peptidomics analysis used in this study provides great opportunities for further data investigation. Because of the database-dependent data processing (see Section 1.6 and Figure 1.7), identification of proteins and peptides is restricted to the sequences provided in the database. As illustrated by Knudsen et al. [6] in response to Bromenshenk et al. [7], it is of crucial importance that this database is as complete as possible. Nevertheless, a compromise needs to be made between completeness and conciseness since a database that is too large results in a very strict false discovery rate filtering and consequently in false negative identifications. In the processing of the raw data for this thesis, databases were used comprising human milk proteins, bovine milk proteins and allergens, covering as much as possible the sequences that were expected to be present in the samples. However, it should be noted that in the data of **Chapters 4** and **5**, for example, the average percentage of identified MS/MS spectra was only 7% and 2% in protein and peptide data, respectively. Although low identification rates are rather common for milk proteomics and peptidomics, in average

MS experiments identification rates are around 25% [8]. Further development of data processing tools, new knowledge about the protein sequences present in human milk, and knowledge about their post-translational modifications, could help to improve the identification rate. This would only require a re-processing of the raw data which is an advantage of the untargeted analysis. However, it is important to note that such re-processing of the data brings challenges as well. Especially for large-scale peptidomics data (**Chapter 5**), extensive data processing is needed, for which a High Performance Computer (HPC) infrastructure is essential. Nevertheless, even with the use of an HPC, possibilities are limited. Especially, options such as widening the range of the peptide lengths and adding variable amino acid modifications increase the processing time exponentially.

Second, a limitation for this thesis was the reduction, alkylation, and trypsin digestion of proteins before analysis. Although intrinsic to bottom-up proteomics, this did not allow for the analysis of intact immunoglobulin (Ig) isotypes. Reduction of proteins caused a breakdown of the Ig structure by disruption of the disulfide bridges that connect the different heavy and light chains. After tryptic digestion and LC-MS/MS analysis, peptides originating from the constant, heavy chains are isotype-specific. However, peptides originating from the variable regions are not isotype-specific and cannot be matched with their respective isotype. In addition, the non-covalent antigen-antibody interactions are also disrupted during the sample preparation [9]. The importance of these antigen-antibody complexes for immune system development has been established by recent studies, showing that antigen-IgG complexes in the milk can be transferred across the gut epithelial barrier by the neonatal Fc receptor (FcRn) and might protect the child against sensitization [10–12]. Although both antigens and antibody-chains were identified in this thesis, the LC-MS/MS results did not allow for a distinction between free antigen and antigen in complex. Future investigations will be needed to provide this level of information.

6.3 Characteristics of the human milk protein and peptide profile

6.3.1 Presence of non-human proteins and peptides

One of the compositional characteristics of human milk, which has received little attention, is the presence of non-human proteins and peptides. Our findings confirm the presence of non-human proteins in human milk and indicate that these are present intact or as large fragments (**Chapter** 2). Furthermore, most of these proteins are bovine proteins, with BLG being identified with the highest sequence coverage (67%).

In the analysis carried out for **Chapter 3** and **Chapter 4**, the resolution of the used LC-MS/MS systems increased, leading to more protein identifications. Surprisingly, however, this did not lead to the identification of more non-human proteins. Although non-human proteins were still identified, they were found in fewer samples and with lower sequence coverage. In **Chapter 3**, 11 non-human proteins were identified (Table 6.1), of which 6 proteins were also identified in **Chapter 2**.

Table 6.1: Identified non-human proteins in the data from **Chapter 3**.

UniProt ID	Leading protein	Organism	Identified with *n* peptides	Identified in *n* samples	Identification score[a]
A0A140T897	Albumin	Bos taurus (Bovine)	31	11	183.9
E1BF59	Plectin	Bos taurus (Bovine)	9	2	8.4
F1MWI1	Clusterin	Bos taurus (Bovine)	2	1	2.7
P02754	β-lactoglobulin	Bos taurus (Bovine)	5	4	9.7
P24627	Lactotransferrin	Bos taurus (Bovine)	3	1	6.2
P49822	Albumin	Canis lupus familiaris (Dog)	6	7	8.3
P49064	Albumin	Felis catus (Cat)	4	7	31.8

[a] Score from the MaxQuant output indicating the quality of the identification of the protein. A higher score represents a better identification.

The number of samples in which these proteins were identified, as well as the sequence coverage of the identified proteins, was limited. This is due to their low abundance and might be due to the health status of the donating mothers since it was shown in **Chapter 2** that allergic mothers shed more non-human proteins in their milk than non-allergic mothers. Because all donating mothers participating in **Chapter 3** underwent a preliminary health screening, it was expected that the majority of them were non-allergic. This is also in line with the fact that cow, dog, and cat albumin were the most frequently identified non-human proteins in **Chapter 3** since these proteins were also frequently identified in the samples of the non-allergic mothers in **Chapter 2** (Table 2.3).

However, if allergy status determined the presence of the non-human proteins in milk, a clear difference would have been expected between the sample groups of different maternal allergy statuses from the CHILD Cohort Study analyzed for **Chapter 4**. Nevertheless, only 14 non-human tryptic peptides from 8 unique non-human proteins were identified in that chapter, with identifications almost equally divided over the samples of allergic and non-allergic mothers. A commonality between the different chapters is the detection of bovine and cat albumin and bovine BLG. The lower number of identified bovine milk proteins in the milk samples from allergic mothers in the CHILD Cohort Study might be due to lower consumption of milk and dairy products in Canada (188 kg milk per capita in 2013) when compared to the Netherlands (341 kg milk per capita in 2013) [13]. Another explanation might be the fact that in **Chapter 2**, a strict definition of house dust mite (HDM) allergy was used, namely (a) self-reported (history of) asthma, current hay fever, current allergy for pets, or current allergy for house dust or HDM in combination with (b) a high level of specific IgE against HDM (≥3.50 kU/L) and (c) high exposure to HDM allergen in mattress dust ((Der p 1 + Der f 1) >600 ng/m^2). Supposing that this specific allergy and allergy definition plays a role in the observed difference in non-human proteins in the milk, the hypothesis posed in **Chapter 2** might be true, in where it was argued that there could be a higher permeability in the epithelial barrier of the intestinal tract of HDM-sensitized mothers. This would then allow an increased passage of non-human proteins into the systemic circulation, possibly mediated by a disruption of the intestinal barrier by HDM allergen Der p 1 (see Figure 6.1).

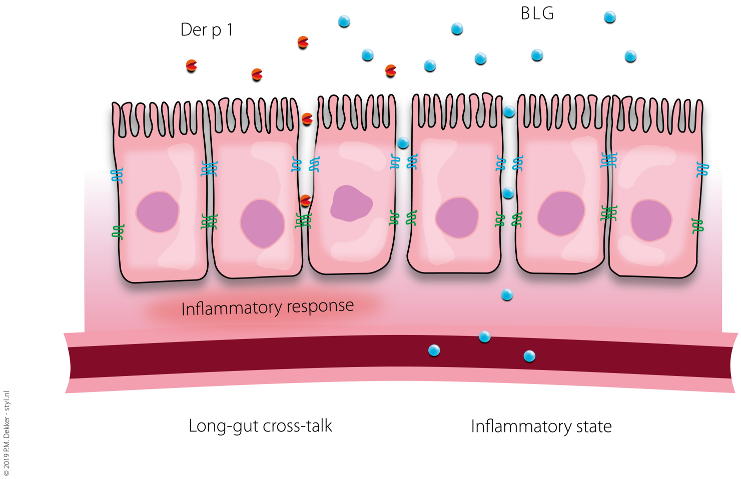


Figure 6.1: Schematic overview of the hypothesized mechanism through which non-human proteins, such as β-lactoglobulin (BLG), could pass the intestinal epithelial barrier and enter the systemic circulation of the mother. Epithelial junctions could be disrupted through both proteolytic activity of Der p 1 and inflammatory response, which subsequently allows paracellular passage of dietary proteins.

Thus far, it has been argued that non-human proteins or large protein fragments are present in human milk and originate especially from the diet. It can be hypothesized that the transfer of this proteinaceous material across the intestinal and mammary epithelium is increased upon specific pathophysiological status of the mother, as in the case of HDM allergy.

In **Chapter 2**, only the protein fraction of the milk was analyzed, discarding the peptides during the sample preparation. However, this peptide fraction was analyzed separately from the proteins in **Chapters 3** and **5**. Additional analysis of the data from these chapters revealed several non-human peptides (Tables 6.2 and 6.3, respectively). Interestingly, all non-human peptides matched with bovine protein sequences, and the majority had β-casein, α_{s1}-casein, or BLG as a precursor. Peptides from these proteins have been identified before in the peptide fraction of human milk [14, 15]. However, only 4 out of the 48 non-human peptides that were identified matched those found in these previous studies. In addition, a large difference can be observed between the peptides from **Chapters 3** and **5**, with only 2 peptides identified in both chapters. One explanation for these differences among studies might be interindividual differences. It was shown in a recent study from Caira et al. [16] that profiles of bovine milk-derived peptides in blood plasma showed extensive qualitative and quantitative variability between individuals, even though prior washout and consumption of milk were the same for all subjects. The presence of these peptides in human milk depends on many factors that can be individual-dependent, such as intestinal digestion, transfer through the intestinal epithelium, possible further digestion by blood proteases, transfer to the mammary gland, transport through the mammary epithelium.

Another factor contributing to the differences between studies is the number of samples combined with the ”match-between-runs” (MBR) algorithm used in the data processing [17]. Briefly, MBR addresses the missing value problem of bottom-up proteomics by inferring identifications of one sample to the other samples using *m/z*, charge state, and retention time. Once a peptide is identified in one sample, the application of MBR increases the likelihood of identification of this peptide in the other samples. Therefore, an increase in sample size with interindividual differences causes an increase in the total number of identifications. This could explain the differences between **Chapters 3** and **5** (Tables 6.2 and 6.3), in which 29 and 297 samples were analyzed, respectively.

Table 6.2: Identified non-human peptides in the peptidomics data from **Chapter 3**. All peptides match with bovine (Bos taurus) proteins and have an identification score >80.

Sequence	UniProt ID	Leading protein	Identification score[a]	Identified in *n* samples
HIQKEDVPSER[b]	P02662	α_{s1}-casein	122.7	19
LRLKKYKVPQL[c]	P02662	α_{s1}-casein	121.8	1
KVPQLEIVPN[c]	P02662	α_{s1}-casein	96.3	21
TDAPSFSDIPNPI[c]	P02662	α_{s1}-casein	155.2	22
QPVNITVQESSSSGPSSMTA	Q28110	Low affinity immunoglobulin gamma Fc region receptor II	85.6	26
FQSEEQQQTEDELQDK[c]	T1T0C1	β-casein	111.4	12
VYPFPGPIPN[b]	T1T0C1	β-casein	108.2	22
YQEPVLGPVRGP[b]	T1T0C1	β-casein	141.9	1
YQEPVLGPVRGPF[b]	T1T0C1	β-casein	117.7	1
YQEPVLGPVRGPFPII[b]	T1T0C1	β-casein	105.2	1
YQEPVLGPVRGPFPIIV[b]	T1T0C1	β-casein	127.4	22

[a] Score from the MaxQuant output indicating the quality of the identification of the peptide. A higher score represents a better identification.

[b] Peptides that have been identified in blood serum by Caira et al. [16].

[c] Peptides of which a precursor peptide has been identified in blood serum by Caira et al. [16].

Table 6.3: Identified non-human peptides in peptide data from **Chapter 5**. All peptides match with bovine (Bos taurus) proteins and have an identification score >80.

Sequence	UniProt ID	Leading protein	Identification score[a]	Identified in *n* samples
FLDDDLTDDIMCVK	P00711	α-lactalbumin	151.2	26
DDDLTDDIMCVK	P00711	α-lactalbumin	117.7	1
FQSEEQQQTEDELQDK[c]	P02666	β-casein	190.4	1
VVPPFLQPEV[c]	P02666	β-casein	88.9	226
TLTDVENLHLPLPLLQ[b]	P02666	β-casein	310.3	1
TDVENLHL[c]	P02666	β-casein	86.1	8
TDVENLHLPLPLLQ[b]	P02666	β-casein	168.3	2
DVENLHLPLPLLQ[b]	P02666	β-casein	143.2	1
YQEPVLGPVRGPFP[b]	P02666	β-casein	104.8	2
YQEPVLGPVRGPFPIIV[b]	P02666	β-casein	124.1	31
EPVLGPVRGPFPIIV[b]	P02666	β-casein	99.8	10
SLAMAASDISLL	P02754	β-lactoglobulin	189.1	2
VYVEELKPTPEGDLE[c]	P02754	β-lactoglobulin	92.0	1
VYVEELKPTPEGDLEI[c]	P02754	β-lactoglobulin	160.2	2
VYVEELKPTPEGDLEIL[c]	P02754	β-lactoglobulin	170.9	2

Table 6.3: *(Continued)* Identified non-human peptides in peptide data from **Chapter 5**. All peptides match with bovine (Bos taurus) proteins and have an identification score >80.

Sequence	UniProt ID	Leading protein	Identification score[a]	Identified in *n* samples
VYVEELKPTPEGDLEILLQK	P02754	β-lactoglobulin	103.1	2
YVEELKPTPEGDLEIL[c]	P02754	β-lactoglobulin	117.8	6
VEELKPTPEGDLE[b]	P02754	β-lactoglobulin	147.4	9
VEELKPTPEGDLEI[b]	P02754	β-lactoglobulin	194.2	1
VEELKPTPEGDLEIL[b]	P02754	β-lactoglobulin	163.8	21
VEELKPTPEGDLEILLQK	P02754	β-lactoglobulin	174.8	5
EELKPTPE[c]	P02754	β-lactoglobulin	112.4	2
EELKPTPEGDLE[c]	P02754	β-lactoglobulin	168.8	5
EELKPTPEGDLEI[c]	P02754	β-lactoglobulin	194.9	2
EELKPTPEGDLEIL[c]	P02754	β-lactoglobulin	133.8	10
ELKPTPEGDLEIL[b]	P02754	β-lactoglobulin	132.0	4
IDALNENK[c]	P02754	β-lactoglobulin	110.8	6
DALNENKVLVL	P02754	β-lactoglobulin	90.9	2
TPEVDDEALEK[b]	P02754	β-lactoglobulin	232.8	16
TPEVDDEALEKF[b]	P02754	β-lactoglobulin	213.3	2
TPEVDDEALEKFDK[b]	P02754	β-lactoglobulin	213.7	2
EVDDEALEKFDK[c]	P02754	β-lactoglobulin	103.9	2
EQLLDNFHLMAESSEDLP	P24591	Insulin-like growth factor-binding protein 1	81.5	34
QLLDNFHLMAESSEDLP	P24591	Insulin-like growth factor-binding protein 1	175.2	141
ISSSSSAEERREIH	Q28085	Complement factor H	99.8	1
QTSLSPDLSQESLSPDL	Q28107	Coagulation factor V	121.0	18
QTALSPDLSQESLSPDLGQT	Q28107	Coagulation factor V	87.2	44
ETLVGYSMVGCQRAMLAN	Q71SP7	Fatty acid synthase	84.7	9
ETLEYVEAHGTGTKVGDPQELNG	Q71SP7	Fatty acid synthase	82.6	71

[a] Score from the MaxQuant output indicating the quality of the identification of the peptide. A higher score represents a better identification.

[b] Peptides that have been identified in blood serum by Caira et al. [16].

[c] Peptides of which a precursor peptide has been identified in blood serum by Caira et al. [16].

A recent study by Caira et al. [16] showed that a wide variety of bovine milk-derived peptides could be detected in human blood after consumption of bovine milk. From the non-human peptides identified in **Chapter 3** (Table 6.2), 6 peptides were identified in blood, and another 4 peptides matched with a precursor peptide identified in blood [16]. Of the non-human peptides identified in **Chapter 5** (Table 6.3), 13 peptides were identified in blood, and another 14 peptides matched with a precursor peptide identified in blood [16].

When peptide intensities from non-human peptides in **Chapter 3** were summed into groups of precursor proteins, a strong correlation (Spearman $\rho = 0.85$) was observed between bovine α_{s1}-casein and β-casein peptides, suggesting that peptides from different bovine milk proteins are transported by the same mechanism.

Together, this confirms that non-human peptides are present in the systemic circulation and can cross both intestinal and mammary epithelial barriers to finally end up in the alveolar lumen of the mammary gland. Nevertheless, the function of these peptides in the milk remains elusive.

6.3.2 Interindividual variation in protein abundance

Protein profiles of milk are different per individual. It was shown in **Chapter 3** that the extent of these interindividual differences has a pattern in which few (<5%), low abundant proteins show substantial interindividual variation (coefficient of variation, CV >100%). The proteome had a median interindividual CV = 42.8%. In the attempt to relate variation in the proteome to maternal characteristics, an association was found between one protein cluster and body mass index (BMI). Nevertheless, the determining factors of most of the variation remained an unanswered question.

Considering also the interindividual variation in proteins identified in **Chapter 4**, it can be noted that this is larger (median CV = 94%) and for some proteins extreme (maximum CV = 674%) (see Figure 6.2). Several factors can explain these differences. First, there is more diversity amongst the participating mothers in the CHILD Cohort Study in terms of, for example, ethnicity, lifestyle, and health status. Second, there is a difference in the pooling of the samples between the chapters. In **Chapter 3**, the samples were pooled over a period ranging between 2 and 28 days within the third month post-partum. However, the samples used for **Chapter 4** were all pooled samples from fore and hindmilk of several feedings of only a single day. Therefore, the latter samples might represent more temporal changes than the samples from **Chapter 3**. Third, more proteins were identified in **Chapter 4** (647 versus 286 in **Chapter 3**). This difference concerns especially low-abundant proteins which show more interindividual variation in general (Figure 3.2).

To investigate whether interindividual variation was different between groups of different maternal and child allergy statuses, the CV was also calculated for each group separately. Comparing the CVs of the various groups revealed interesting differences (Figure 6.3). First, it can be noted that proteins with low interindividual variation (CV <50%) are consistent among the groups, showing that low interindividual variation is consistent, irrespective of mother-child allergy status. Second, several histones, histone H3.2 (H3C15), histone H4 (H4C1), and histone H2B (H2BC15), show large differences in interindividual variation between the groups of different mother-child allergy statuses. Extracellular histones are damage-associated molecular pattern molecules (DAMPs) in cellular processes and can be pathophysiological indicators for several diseases [18]. Although little is known about the presence and function of histones in human milk, it was recently shown that they could be hydrolyzed by proteolytic activity of secretory immunoglobulin A (sIgA) in the milk [19]. Interestingly, the heavy chain of IgA shows a high variation (CV = 361%) in the group where both mother and infant are allergic and less in the other groups (CV <183%). A similar pattern can be observed for tenascin-C (TNC), an antiviral protein. It was hypothesized in **Chapter 3** that higher expression of TNC in milk could protect the offspring against viral infections, or alternatively, indicate an inflammatory response of the mother. The difference in variation of the aforementioned proteins within the groups of mother-child allergy status could be related to the heterogeneous pathophysiology of allergies within a group.

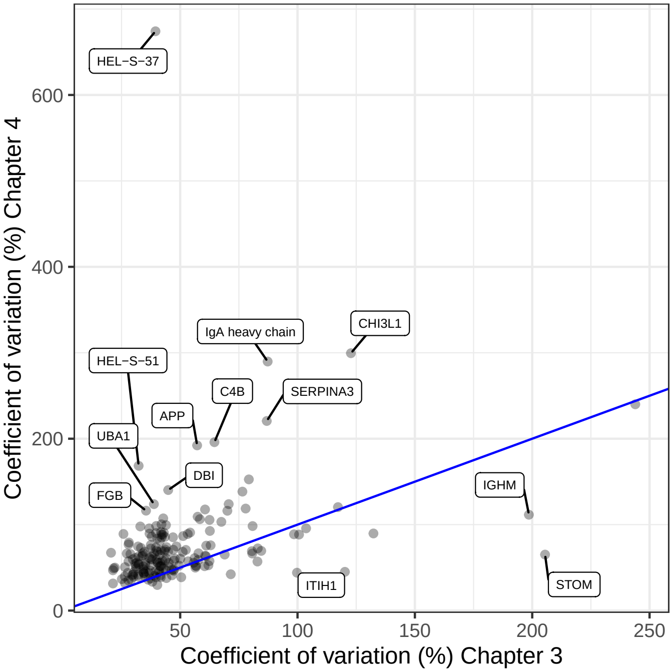


Figure 6.2: Comparison of the interindividual variation observed for proteins identified in **Chapter 3** and **Chapter 4**. Each point represents a protein, with on the axes the coefficients of variation (CV) from both chapters. Proteins located near the blue diagonal line have a similar interindividual variation in both chapters. Labeled proteins have a difference in coefficient of variation (CV) >75%, and are labeled with gene IDs or, if not available, protein names.

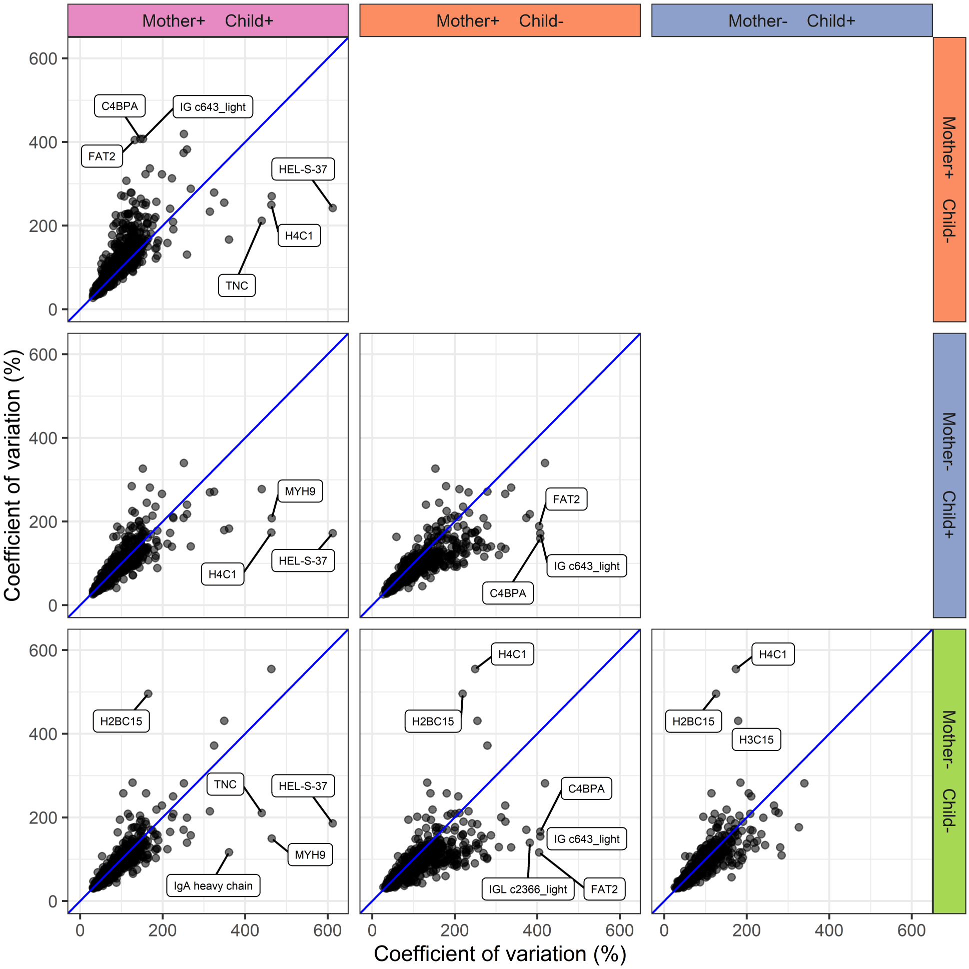


Figure 6.3: Comparison of the interindividual variation in proteins observed in the groups of different mother-child allergy status from **Chapter 4**. Each dot represents a single protein. In the labeling of the groups, + indicates allergy and - indicates no allergy. Labeled proteins have a difference in coefficient of variation (CV) >200%, and are labeled with gene IDs or, if not available, protein names.

In a further investigation of the interindividual differences observed in the data of **Chapter 4**, metadata and protein intensities were used to explore which maternal or infant characteristics contribute the most to the interindividual variation. For this analysis, the R package "variancePartition" [20] was used, which allows the quantification of multiple sources of variation in gene expression data using linear mixed models. In this analysis, the categorical variables ethnicity (Caucasian, Asian, First Nation, Other), study site (Vancouver, Edmonton, Manitoba, Toronto), infant sex, and allergy status of mother and child were considered random effects. In addition, the continuous variables maternal age, maternal BMI, lactation stage, and total protein concentration were considered fixed effects. This analysis showed that, after correcting for the other variables, the total protein concentration explains a median of 10.3% of the variation in the proteins, followed by ethnicity with a median of 1.4% of the variation (Figure 6.4).

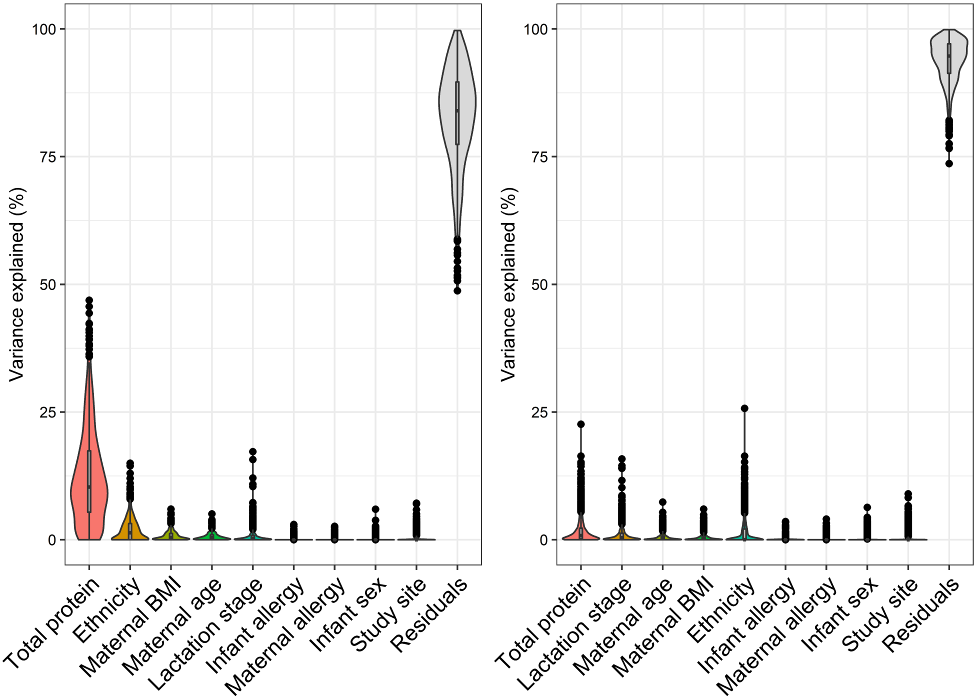


Figure 6.4: Contribution of each variable to the total variance in (left) the proteomics data from **Chapter 4**, and (right) the peptidomics data from **Chapter 5**.

The contribution of the total protein concentration to the variation in the whole proteome is an interesting result. It is known that a few major milk proteins are the largest contributors to the total protein concentration of human milk (Section 1.2 and Figure 1.2). Among the top 10 proteins for which the variation is best explained by the total protein concentration were the major milk proteins lactotransferrin (LF) and polymeric immunoglobulin receptor (PIGR). Regarding LF, this corresponds with Czosnykowska-Łukacka et al. [21], who reported a positive correlation between the concentration of LF and the total protein concentration. The fact that total protein concentration contributes substantially to the variation in the whole proteome suggests that the same factors involved in the regulation of LF synthesis could also be involved in regulating other proteins.

Although the 1.4% contribution of ethnicity seems low, this is a significant contribution. To deduce differences in protein intensity between the milk from Caucasian (n = 177) and the milk from Asian mothers (n = 83), Mann-Whitney U tests [22] were applied. The resulting p-values were corrected for multiple hypothesis testing using Benjamini-Hochberg correction [23], where after correction, a p-value <0.05 was considered significant. With this analysis, 302 significantly different proteins were found. Of these, 53 were higher in abundance in milk from Asian mothers, and 249 were more abundant in milk from Caucasian mothers (Figure 6.5).

What is surprising is that among the proteins that were higher abundant in milk from Asian mothers are all identified Ig heavy chains (IgA, IgM, and IgG4), as well as 16 other Ig chains. Elwakiel et al. [24] who investigated differences in milk protein composition from Dutch and Chinese mothers, also found differences in Ig-chains, although their levels were not consistently higher in milk from Chinese mothers. In addition to the Ig chains, gene ontology (GO) overrepresentation analysis with Benjamini-Hochberg correction reveals an overrepresentation of proteins from the lysosomal lumen ($p = 0.0002$) and the extracellular region ($p = 0.0002$). This overrepresentation is, amongst others, due to plasma serine protease inhibitor (SERPINA5), alpha-1-antichymotrypsin (SERPINA3), cathepsins C (CTSC), D (CTSD), and S (CTSS), which were all more abundant in milk from Asian mothers.

Regarding the proteins that are more abundant in the milk of Caucasian mothers, overrepresentation analysis points to the cytosol ($p = 1.3 \times 10^{-25}$) and the ribosomal subunit ($p = 8.6 \times 10^{-10}$). The most probable explanation for this is that more cells are present in milk from Caucasian mothers. Cell membranes of these cells are disrupted during frozen storage of the samples, resulting in a release of cell content from the cells upon thawing. This explanation corresponds with the findings of Qu et al. [25], who also showed a drastic increase of ribosomal proteins in milk upon frozen storage.

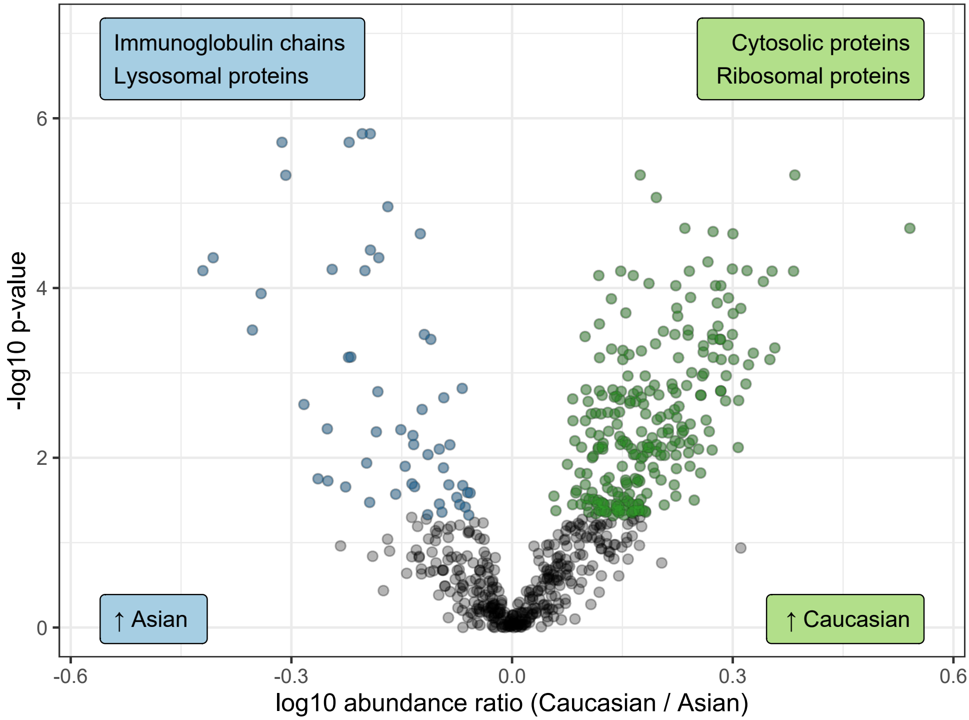


Figure 6.5: Volcano plot visualizing the differences in protein abundances in milk from Caucasian and Asian mothers from the CHILD Cohort Study (data from **Chapter 4**). Each data point represents one protein, with on the x-axes the ratio of the means of the log10 transformed label-free quantification (LFQ). Proteins with p <0.05 are represented by colored dots and the other proteins with grey dots. Colored labels on left and right side of x = 0 indicate in which ethnicity the mean abundance of the respective proteins is higher.

6.3.3 Interindividual variation in peptides

The interindividual variation in the peptide profile was more extensive when compared to the protein profile. In **Chapter 3**, a CV >100% was observed for 36% of all identified peptides and a median CV = 85%. Similar to the proteomics data, variation in peptides showed to be associated with maternal characteristics since two peptide clusters were associated with BMI and one with gestational age. In the same way as the protein data, a comparison was made between the interindividual variation of the peptides from **Chapter 3** and **Chapter 5** (Figure 6.6).

It can be observed that also for the peptides, there is a larger interindividual variation in **Chapter 5**, with a median CV = 162%. The high interindividual variation in the peptidome compared to the proteome (also observed in **Chapter 3**) suggests that the factors determining the milk peptidome are more individual-dependent. Among these factors are, for example, the presence, activation, and inhibition of proteases and aminopeptidases.

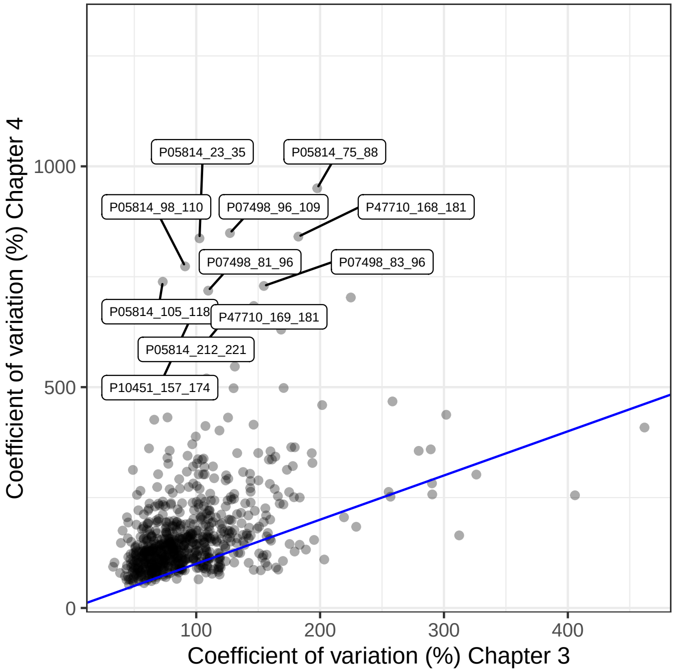


Figure 6.6: Comparison of the interindividual variation observed for the peptides identified in **Chapter 3** and **Chapter 5**. Each point represents a peptide, with on the axes the coefficients of variation (CV) from both chapters. Peptides located close to the blue diagonal line have a similar interindividual variation in both chapters. Labeled peptides have a difference in coefficient of variation (CV) >500% and are labeled with the UniProt ID of their respective precursor protein and their range in the protein sequence.

The sources of variation in the peptidomics data were quantified similarly to the proteomics data. This shows that after correcting for the other variables, the total protein concentration explains a median of 0.8% of the variation in the peptides, followed by the lactation stage with a median of 0.4% of the variation (Figure 6.4).

Interestingly, although a small part of the variation in the peptidome can be explained by ethnicity (median = 0.02%), there are several peptides whose variation seems related to ethnicity. This is especially interesting considering the fact that among the proteins that showed to be significantly different between milk from Caucasian and Asian mothers were several proteases and protease inhibitors. Mann-Whitney U tests on the peptide data with subsequent Benjamini-Hochberg correction for multiple hypothesis testing revealed 626 significantly different peptides in the comparison of milk from Caucasian and Asian mothers (Figure 6.7).

Of these peptides, 280 were more abundant in milk from Asian mothers, mainly derived from β-casein (n = 196) and osteopontin (n = 32). The peptides which were more abundant in milk from Caucasian mothers (n = 346) were derived to a lesser extent from β-casein (n = 60) and also from butyrophilin subfamily 1 member A1 (BTN1A1) (n = 78), perilipin-2 (PLIN2) (n = 54), PIGR (n = 36), and perilipin-3 (PLIN3) (n = 13). BTN1A1, PLIN2, and PLIN3 are typical milk fat globule membrane (MFGM) proteins, which showed to have a different proteolytic degradation mechanism in **Chapter 3**. This could point to more milk fat globules (MFGs) in milk from Caucasian mothers, leading to more MFGM proteins, and more degradation of these proteins into peptides. However, few MFGM proteins were more abundant in milk from Caucasian mothers. It can thus be suggested that there is a difference in proteolytic degradation in milk from mothers of different ethnicities, leading to significant differences in the milk peptidome.

In line with our work, the next step would be to simultaneously map both inter-individual and intra-individual variation in the human milk protein profile. This has been done in a small-scale study by Zhu et al. [26], including samples from only two mothers. However, a more extensive study population, including different ethnicities, is needed to create a more comprehensive map of the extent of the different sources of variation.

6.3.4 Interactions between proteins and peptides

By integrating proteomics and peptidomics data, several strong associations were found across proteins and peptides in **Chapter 5**. Associations were found especially across proteins and peptides that are part of specific pathways: coagulation and complement activation. In addition, associations between proteins and between peptides were investigated, revealing clusters of proteins sharing the same mechanism of transport and clusters of peptides with commonalities in degradation mechanism.

Several of the associations observed in **Chapter 5** shed light on results reported in the other experimental chapters. First, clustering of Ig constant and variable chains revealed their common source and probable relatedness in Ig isotype. This supports

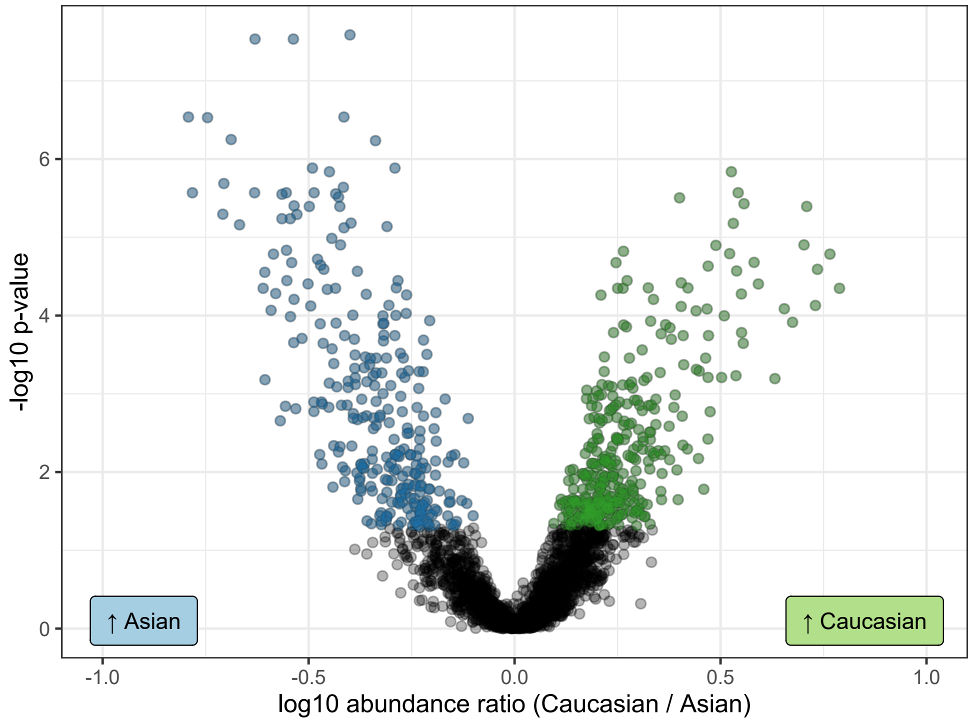


Figure 6.7: Volcano plot visualizing the differences in peptide abundances in milk from Caucasian and Asian mothers from the CHILD Cohort Study (data from **Chapter 5**). Each data point represents one peptide, with on the x-axes the ratio of the means of the log10 transformed label-free quantification (LFQ). Peptides with $p < 0.05$ are represented by colored dots and the other proteins with grey dots. Colored labels on left and right side of x = 0 indicate in which ethnicity the mean abundance of the respective peptides is higher.

the further discussion of the Ig variable chains in Section 6.4.2 about how these proteins might be related to allergy development in the child. Second, proteins from which their association in a cluster revealed their common origin from lysosomal exocytosis (for example, the different cathepsins) were also more abundant in milk from Asian mothers. This suggests a different regulation of lysosomal exocytosis in the mammary gland from Asian mothers, which possibly leads to a difference in the peptide profile (Section 6.3.3). Third, the associations observed between complement proteins and their peptides as well as the associations between coagulation proteins and their degradation products, suggests that these coordinated cascades, which are common in blood, are also activated in milk.

Together, this shows that a systems biology approach in which associations are considered, can reveal more of the complexity underlying the synthesis and mechanisms of transport of proteins, and their degradation into peptides.

6.4 Human milk proteins & peptides and allergy

6.4.1 Proteins and maternal allergy

No significant relation was found between maternal allergy status and the abundance of proteins in the milk (**Chapter 4**). Regarding protein connectivity and correlation patterns, overall stronger connectivity was observed for the groups where the mother was allergic. However, this pattern could not be distinguished from the group where the mother was non-allergic, and the child ultimately developed an allergy. In **Chapter 2**, however, it was found that maternal allergy status influenced milk protein composition and showed higher levels of bovine proteins, as discussed before. Furthermore, another study showed that cystatin C, endogenous protease inhibitors, and apolipoproteins were more abundant in milk from allergic mothers [27]. This was not confirmed by the findings from **Chapter 4**. A probable explanation for this might be the difference in allergy definition (as discussed in more detail in Section 6.3.1). In light of our findings from **Chapter 5**, where protease inhibitors appear associated with other blood-derived proteins (Figure 5.1, cluster 10), and apolipoproteins with cystatin C (Figure 5.1, cluster 63), it can be hypothesized that the observation of Hettinga et al. [27] was due to an increased transfer of blood-derived proteins in HDM allergic mothers.

Taken together, these findings and the findings discussed in Section 6.3.1 support the recommendation for future studies to investigate specific definitions of maternal allergy in relation to both endogenous and non-human proteins in the milk.

6.4.2 Proteins and child allergy

Moving on now to consider allergy development in the infant, a trend was observed that Ig chains are more abundant in milk for infants who would ultimately develop an allergy (**Chapter 4**).

In mature milk (3 months postpartum), secretory IgA is the major Ig isotype, contributing around 81% to the total Ig concentration, followed by IgG (11%) and secretory IgM (8%) [28]. In the prevention of sensitization of the infant, sIgA and IgG are the most important isotypes [29]. Igs of one isotype have differences in the sequence of the antigen-binding variable domain. This domain can bind with the respective antigens and result in the formation of allergen-Ig complexes. The following part of this section will discuss the relevance of total isotype levels, allergen-specific Ig, and allergen-Ig complexes, respectively.

Ig isotypes

Our findings could point to a higher abundance of total Ig or Ig isotypes in the milk for children who would develop an allergy. However, this would have resulted in significant differences in especially the abundance of Ig isotype-specific heavy chains, which was not observed in our data (Table 4.1). In literature, it has been suggested that low IgA levels in milk might lead to allergy development in the offspring [30], but a more recent study showed, in line with our observations, that total levels of IgA in milk are not associated with allergy outcomes at 2 or 5 years of age [31].

Allergen-specific Ig

It was found that the Ig-chains that were more abundant in milk for infants who ultimately developed an allergy were especially the variable regions of Igs (Table 4.1). Unfortunately, due to the nature of the MS analysis used in our experiments, these variable regions could not be linked to their respective antibody (Section 6.2.2). Nevertheless, considering the contribution of IgA to the total Ig pool, as well as the strong associations between most of the identified variable regions and the heavy constant regions of IgA, IgM, and J-chain (**Chapter 5**), it is more likely that the identified variable regions are derived from IgA than from the other isotypes. Nevertheless, confirmation of this is required in future studies.

The variable regions of the Igs could represent relevant differences in allergen-specific Ig in the milk. Regarding IgG, this would contradict previous studies, which show that higher levels of allergen-specific IgG are associated with a lower risk of allergy [32, 33]. For allergen-specific IgA, there is conflicting evidence in the literature, showing associations with both a lower risk of allergy development [34] and a higher risk of allergy development [31, 35].

Allergen-Ig complexes

One explanation for the contradicting results of the aforementioned studies might be whether these allergen-specific Igs are present in a complex with an allergen or not. Several lines of evidence showed the importance of allergen-Ig complexes in milk for the education of the infant's immune system [10, 12, 32]. It has been suggested that IgA might decrease allergic sensitization through binding with free allergens, keeping the allergens away from the intestinal epithelium [36]. Allergen-IgG complexes, on the other hand, can be transferred through the infants' intestinal barrier by the FcRn, which allows interaction with T cells and could ultimately lead to tolerance induction [37]. In mice studies, it was shown that allergen-exposed, sensitized dams, have more ovalbumin-IgG and BLG-IgG complexes in their milk [11, 12, 32]. Evidence was provided that, through these Ig complexes, infants might be protected against allergic sensitization at 5 years of age [10].

Given that allergen-Ig complexes in milk are important for tolerance induction in the breastfed infant, it is recommended for further research to comprehensively map the diversity of the Ig pool in human milk, including Ig isotypes, allergen-specific Igs, allergen-Ig complexes, and free allergen.

Both IgA and IgM showed substantial interindividual variation in **Chapter 3** (their heavy constant regions have a CV of 117% and 199%, respectively). As discussed in Section 6.3.2, the data from **Chapter 4** shows a distinctive pattern in the interindividual variation of IgA between the groups of different allergy statuses. The variation in IgA was more prominent in the group where both mother and child were allergic, which might be due to more clinical heterogeneity in this group.

In addition, the heavy constant regions of all human milk Ig isotypes and several variable regions were found to be significantly higher in abundance in milk from Asian mothers when compared with Caucasian mothers. Nevertheless, the different mother-child allergy status groups of the CHILD Cohort Study were matched for ethnicity. Furthermore, additional comparisons with ANOVA and adjustment for ethnicity as a covariate did not reveal significantly different proteins between the groups.

Protein connectivity

Lastly, higher protein connectivity was found in milk from the mother-child allergy status groups where at least one of mother or child was allergic. This differential connectivity concerned especially ribosomal proteins and other proteins involved in translation processes (**Chapter 4**). In Section 6.3.2, it was observed that ribosomal proteins and proteins from the cytosol are higher abundant in milk from Caucasian mothers when compared to Asian mothers, most likely due to more cells being present in their milk. The difference in connectivity of these proteins between the groups of different mother-child allergy statuses could, therefore, possibly indicate a difference in the cells present in the milk. However, further research will be needed to elucidate this.

6.4.3 Peptides and allergy

Univariate analysis was applied to the peptidomics data similarly to the proteomics data (as described in **Chapter 4**) to investigate whether the milk peptide profile was related to mother and child allergy status. Kruskal-Wallis tests with subsequent correction for multiple hypothesis testing showed no significantly different peptides between the groups of different mother-child allergy statuses. To explore whether the non-significant differences found were relevant, post-hoc tests were performed on peptides with an uncorrected p <0.05. Outcomes from the post-hoc tests were adjusted for multiple testing, and peptides with an adjusted p-value <0.05 ($n = 66$) were searched for in the Milk Bioactive Peptide Database (MBPDB) [38]. Nevertheless, none of the 66 peptides matched with bioactive peptides in the database.

To investigate a possible relation between overall proteolytic activity and mother-child allergy status, the intensity of all peptides was summed per sample, and groups were compared. An ANOVA did not reveal significant differences between the groups ($p = 0.96$).

In summary, this shows that, contrary to expectations, maternal allergy does not significantly affect the milk peptide profile. Neither does the milk peptide profile significantly affect allergy development in the child. Further investigation of the peptidomics data with, for example, network analysis would require a better understanding of the meaning of peptide-peptide associations in the milk peptidome.

6.5 Conclusions

This thesis has provided a deeper insight into the human milk proteome and peptidome. It can be concluded that:

- Non-human proteins and peptides are present in human milk and originate mainly from bovine milk.

- There is substantial interindividual variation in the human milk proteome and even more in the peptidome.

- The abundance of proteins in milk is primarily determined by the mechanism through which the proteins are transported through the mammary epithelium.

- The abundance of peptides in the milk is, in the case of the so-called "peptide ladders", to a great extent determined by the level of the precursor peptides.

- There are significant differences in the human milk proteome and peptidome between mothers from different ethnic backgrounds.

Furthermore, insights have been obtained into the relation between human milk proteome and peptidome and the allergy status of both mother and child. It can be concluded that:

- House dust mite allergy, in combination with high dairy consumption, leads to higher levels of dairy proteins in human milk when compared with non-allergic mothers.
- There is a trend that milk for infants who ultimately developed an allergic disease contained higher levels of variable regions of immunoglobulins.
- Milk from healthy mothers with infants who did not develop an allergy showed lower protein connectivity than milk from mothers who were allergic or from whom the infant developed an allergy.
- No relation was found between maternal allergy status or child allergy development and the human milk peptidome.

6.6 Implications and future perspectives

The findings presented in this thesis have a number of practical implications. It was shown in this thesis that human milk can contain non-human, allergenic proteins and potentially allergenic peptides. This knowledge is of high importance for human milk banks considering the fact that the milk they collect is meant for the extra vulnerable, the preterm infants. Although the risk of this for the newborn is unknown, it is recommended to prevent allergic sensitization and, therefore, to pool samples of multiple donors to dilute the allergen if present. In addition, this knowledge should be transferred to medical doctors in nursing care so that well thought out measures can be taken when exclusively breastfed infants start to show signs of allergic sensitization or food protein-induced enterocolitis [39].

Second, considering the differences that were found in milk from mothers of different ethnicities, I recommend further investigation of the relevance of these differences for the infant's healthy development. Outcomes of this will be relevant for (*i*) human milk banks who can consider matching donor ethnicity with recipient ethnicity, and (*ii*) producers of infant formula, who can adapt their formulation to the ethnic background of the consumers.

As research into mother-child allergy status and the relation with proteins and peptides will continue, a narrow definition of allergy is recommended for future studies. In addition, a comprehensive analysis of the Ig isotypes, allergen-specific Igs, allergen-Ig complexes, and free allergen is recommended. From our observations and recent literature findings [10, 11, 40], these components are the most promising to reveal relevant associations with child allergy development. Lastly, it is recommended to investigate the possible role of the intact or damaged cells present in the milk. The possible relations between the differential connectivity of especially ribosomal proteins among the different allergy groups may be related to differences in cells present in the milk.

References

[1] Pascovici, D. et al. “Multiple Testing Corrections in Quantitative Proteomics: A Useful but Blunt Tool”. In: *Proteomics* 16 (2016), 2448–2453. DOI: 10.1002/pmic.201600044.

[2] Rosato, A. et al. “From Correlation to Causation: Analysis of Metabolomics Data Using Systems Biology Approaches”. In: *Metabolomics* 14 (2018), 37. DOI: 10.1007/s11306-018-1335-y.

[3] Saccenti, E. and Svensson, M. “Systems Biology and Biomarkers in Necrotizing Soft Tissue Infections”. In: *Advances in Experimental Medicine and Biology*. Ed. by A. Norrby-Teglund, M. Svensson, and S. Skrede. Vol. 1294. Cham: Springer International Publishing, 2020, 167–186. ISBN: 978-3-030-57616-5. DOI: 10.1007/978-3-030-57616-5_11.

[4] Langfelder, P. and Horvath, S. “WGCNA: An R Package for Weighted Correlation Network Analysis”. In: *BMC Bioinformatics* 9 (2008), 559. DOI: 10.1186/1471-2105-9-559.

[5] Schäfer, J., Opgen-Rhein, R., and Strimmer, K. “Reverse Engineering Genetic Networks Using the {GeneNet} Package”. In: *R News* 6 (2006), 50–53.

[6] Knudsen, G. M. and Chalkley, R. J. “The Effect of Using an Inappropriate Protein Database for Proteomic Data Analysis”. In: *PLoS ONE* 6 (2011). DOI: ARTNe2087310.1371/journal.pone.0020873.

[7] Bromenshenk, J. J. et al. “Iridovirus and Microsporidian Linked to Honey Bee Colony Decline”. In: *PLoS ONE* 5 (2010), e13181. DOI: 10.1371/journal.pone.0013181.

[8] Griss, J. et al. “Recognizing Millions of Consistently Unidentified Spectra across Hundreds of Shotgun Proteomics Datasets”. In: *Nature Methods* 13 (2016), 651–656. DOI: 10.1038/nmeth.3902.

[9] van Oss, C. J., Good, R. J., and Chaudhury, M. K. “Nature of the Antigen-Antibody Interaction. Primary and Secondary Bonds: Optimal Conditions for Association and Dissociation”. In: *Journal of Chromatography B: Biomedical Sciences and Applications* 376 (1986), 111–119. DOI: 10.1016/S0378-4347(00)80828-2.

[10] Lupinek, C. et al. “Maternal Allergen-Specific IgG Might Protect the Child against Allergic Sensitization”. In: *Journal of Allergy and Clinical Immunology* 144 (2019), 536–548. DOI: 10.1016/j.jaci.2018.11.051.

[11] Adel-Patient, K. et al. "Prevention of Allergy to a Major Cow's Milk Allergen by Breastfeeding in Mice Depends on Maternal Immune Status and Oral Exposure during Lactation". In: *Frontiers in Immunology* 11 (2020), 1–10. DOI: 10.3389/fimmu.2020.01545.

[12] Ohsaki, A. et al. "Maternal IgG Immune Complexes Induce Food Allergen-Specific Tolerance in Offspring". In: *Journal of Experimental Medicine* 215 (2018), 91–113. DOI: 10.1084/jem.20171163.

[13] FAOSTAT. *Food Supply-Livestock and Fish Primary Equivalent*. 2013. URL: http://www.fao.org/faostat/en/#data/CL.

[14] Picariello, G. et al. "Antibody-Independent Identification of Bovine Milk-Derived Peptides in Breast-Milk". In: *Food and Function* 7 (2016), 3402–3409. DOI: 10.1039/C6FO00731G.

[15] Picariello, G. et al. "Excretion of Dietary Cow's Milk Derived Peptides into Breast Milk". In: *Frontiers in Nutrition* 6 (2019), 25. DOI: 10.3389/fnut.2019.00025.

[16] Caira, S. et al. "In Vivo Absorptomics: Identification of Bovine Milk-Derived Peptides in Human Plasma after Milk Intake". In: *Food Chemistry* 385 (2022), 132663. DOI: 10.1016/j.foodchem.2022.132663.

[17] Tyanova, S., Temu, T., and Cox, J. "The MaxQuant Computational Platform for Mass Spectrometry-Based Shotgun Proteomics". In: *Nature Protocols* 11 (2016), 2301–2319. DOI: 10.1038/nprot.2016.136.

[18] Chen, R. et al. "Release and Activity of Histone in Diseases". In: *Cell Death and Disease* 5 (2014), 1–9. DOI: 10.1038/cddis.2014.337.

[19] Kompaneets, I. Y. et al. "Secretory Immunoglobulin A from Human Milk Hydrolyzes 5 Histones and Myelin Basic Protein". In: *Journal of Dairy Science* 105 (2022), 950–964. DOI: 10.3168/jds.2021-20917.

[20] Hoffman, G. E. and Schadt, E. E. "variancePartition: Interpreting Drivers of Variation in Complex Gene Expression Studies". In: *BMC Bioinformatics* 17 (2016), 17–22. DOI: 10.1186/s12859-016-1323-z.

[21] Czosnykowska-Łukacka, M. et al. "Lactoferrin in Human Milk of Prolonged Lactation". In: *Nutrients* 11 (2019), 2350. DOI: 10.3390/nu11102350.

[22] Wilcoxon, F. "Individual Comparisons by Ranking Methods". In: *Biometrics Bulletin* 1 (1945), 80. DOI: 10.2307/3001968.

[23] Benjamini, Y. and Hochberg, Y. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 57 (1995), 289–300. DOI: 10.1111/j.2517-6161.1995.tb02031.x.

[24] Elwakiel, M. et al. “Variability of Serum Proteins in Chinese and Dutch Human Milk during Lactation”. In: *Nutrients* 11 (2019), 499. doi: 10.3390/nu11030499.

[25] Qu, J. et al. “Changes in Bioactive Proteins and Serum Proteome of Human Milk under Different Frozen Storage”. In: *Food Chemistry* 352 (2021), 129436. doi: 10.1016/j.foodchem.2021.129436.

[26] Zhu, J. et al. “Personalized Profiling Reveals Donor- and Lactation-Specific Trends in the Human Milk Proteome and Peptidome”. In: *Journal of Nutrition* 151 (2021), 826–839. doi: 10.1093/jn/nxaa445.

[27] Hettinga, K. A. et al. “Difference in the Breast Milk Proteome between Allergic and Non-Allergic Mothers”. In: *PLoS ONE* 10 (2015). Ed. by A. S. Wiley, e0122234. doi: 10.1371/journal.pone.0122234.

[28] Goonatilleke, E. et al. “Human Milk Proteins and Their Glycosylation Exhibit Quantitative Dynamic Variations during Lactation”. In: *Journal of Nutrition* 149 (2019), 1317–1325. doi: 10.1093/jn/nxz086.

[29] Shamji, M. H. et al. “The Role of Allergen-Specific IgE, IgG and IgA in Allergic Disease”. In: *Allergy: European Journal of Allergy and Clinical Immunology* 76 (2021), 3627–3641. doi: 10.1111/all.14908.

[30] Järvinen, K. M. et al. “Does Low IgA in Human Milk Predispose the Infant to Development of Cow’s Milk Allergy?” In: *Pediatric Research* 48 (2000), 457–462. doi: 10.1203/00006450-200010000-00007.

[31] Kuitunen, M., Kukkonen, A. K., and Savilahti, E. “Impact of Maternal Allergy and Use of Probiotics during Pregnancy on Breast Milk Cytokines and Food Antibodies and Development of Allergy in Children until 5 Years”. In: *International Archives of Allergy and Immunology* 159 (2012), 162–170. doi: 10.1159/000336157.

[32] Mosconi, E. et al. “Breast Milk Immune Complexes Are Potent Inducers of Oral Tolerance in Neonates and Prevent Asthma Development”. In: *Mucosal Immunology* 3 (2010), 461–474. doi: 10.1038/mi.2010.23.

[33] Nakata, K. et al. “The Transfer of Maternal Antigen-Specific IgG Regulates the Development of Allergic Airway Inflammation Early in Life in an FcRn-dependent Manner”. In: *Biochemical and Biophysical Research Communications* 395 (2010), 238–243. doi: 10.1016/j.bbrc.2010.03.170.

[34] Savilahti, E. et al. “IgA Antibodies, TGF-β1 and -B2, and Soluble CD14 in the Colostrum and Development of Atopy by Age 4”. In: *Pediatric Research* 58 (2005), 1300–1305. doi: 10.1203/01.pdr.0000183784.87452.c6.

[35] Böttcher, M. F. et al. "Total and Allergen-Specific Immunoglobulin a Levels in Saliva in Relation to the Development of Allergy in Infants up to 2 Years of Age". In: *Clinical & Experimental Allergy* 32 (2002), 1293–1298. doi: 10.1046/j.1365-2222.2002.01470.x.

[36] Atyeo, C. and Alter, G. "The Multifaceted Roles of Breast Milk Antibodies". In: *Cell* 184 (2021), 1486–1499. doi: 10.1016/j.cell.2021.02.031.

[37] Yoshida, M. et al. "Human Neonatal Fc Receptor Mediates Transport of IgG into Luminal Secretions for Delivery of Antigens to Mucosal Dendritic Cells". In: *Immunity* 20 (2004), 769–783. doi: 10.1016/j.immuni.2004.05.007.

[38] Nielsen, S. D. et al. "Milk Bioactive Peptide Database: A Comprehensive Database of Milk Protein-Derived Bioactive Peptides and Novel Visualization". In: *Food Chemistry* 232 (2017), 673–682. doi: 10.1016/j.foodchem.2017.04.056.

[39] Vergara Perez, I. and Vila Sexto, L. "Suspected Severe Acute Food Protein–Induced Enterocolitis Syndrome Caused by Cow's Milk through Breast Milk". In: *Annals of Allergy, Asthma and Immunology* 121 (2018), 245–246. doi: 10.1016/j.anai.2018.04.023.

[40] Demers-Mathieu, V. and Lavangnananda, S. "Restricting Cow's Milk in the Maternal Diet Reduces the Titers of β-Lactoglobulin-Specific IgG Antibodies in Human Milk". In: *Breastfeeding Medicine* XX (2022), 1–5. doi: 10.1089/bfm.2021.0242.

Summary

To provide newborns with nutrition and protection, the mother's body produces a complex biofluid, human milk. Among the many constituents of this milk are proteins, a diverse group of molecules with a variety of functions. Such functions can be the provision of nutrients for growth, digestion, transport of minerals, or immune response. In addition, some proteins function as proteases and cleave other proteins into smaller fragments, the peptides. Peptides can have bioactive functions different from the functions of the protein they originate from, such as antimicrobial activity or cell proliferation stimulation. Proteins and peptides in human milk have been a topic of ongoing investigation. Still, many questions regarding their origin, biological functions, and effect on the health of the breastfed infant remained unanswered. In this thesis, we investigated the protein and peptide composition and their relation with maternal and child allergy.

One aspect of the protein composition of human milk is the presence of non-human proteins. This was the main topic of **Chapter** 2, where we analyzed milk proteomics data of 10 allergic and 10 non-allergic mothers. The most striking result that emerged from this was that mothers with house dust mite allergy shed more β-lactoglobulin in their milk when compared to non-allergic mothers. In addition, we identified non-human sequences from various organisms, but the primary source was shown to be bovine milk. The identification of these sequences was confirmed with several different methods. Together, these findings suggested a difference in the transfer of dietary proteins or protein fragments through the intestinal barrier of allergic mothers, allowing dietary proteins to enter the bloodstream and, ultimately, the milk.

The normal variation in the proteome and peptidome among different individuals is important information that can support the interpretation of differences between groups of samples. We mapped this interindividual variation in **Chapter 3** using a set of 286 samples from 29 mothers, pooled per mother. Substantial interindividual variation was observed for a selection of proteins and peptides, whereas the peptides showed an overall larger variation. In addition to the proteome and the peptidome, we also analyzed the metabolome in these samples. The metabolites, which can be seen as indicators of biological processes in the human body, were used to investigate whether patterns of variation could be explained by biological processes. For this investigation, we applied weighted correlation network analysis and examined the association between clusters of proteins, clusters of peptides, and

metabolites. The findings of this study provide insight into the dynamics of human milk protein, peptide, and metabolite composition.

In **Chapter 4**, we continued the investigation of the proteome and the relation with allergy of the mother and allergy development in the child. In a collaboration with the Canadian CHILD Cohort Study, we analyzed 300 milk samples from mother-child dyads with different allergy statuses. For the data analysis, different methods were employed, including univariate analysis, multivariate exploratory analysis, Random Forest classification models, and network analysis. We found a difference between the samples of the different groups of mother-child allergy status, with a trend that immunoglobulin chains are more abundant in milk from allergic mothers and milk meant for infants who would ultimately develop an allergy. Furthermore, we also found that the protein connectivity, that is, the associations proteins have with other proteins, is less in milk from non-allergic mothers with infants who would not develop an allergy. We suggested that this might be due to a dysregulation of cellular processes or a difference in cells present in the milk. Finally, it was noteworthy that the network analysis revealed subtle but relevant information that was not brought to light using the classical methods for data analysis.

After studying the proteome of the samples of the CHILD Cohort Study, we integrated this data with the peptidome and explored protein-protein, peptide-peptide, and protein-peptide associations with a network approach in **Chapter 5**. This study revealed that such associations give important information about the biological processes and mechanisms involved in the synthesis and secretion of proteins into milk. Also, we were able to differentiate between patterns of endoproteolytic and exoproteolytic activity using peptide-peptide associations. Regarding protein-peptide associations, we demonstrated specific digestion of several proteins, whereas the abundance of the peptides depends on the abundance of these proteins. This was observed for, among others, complement proteins and coagulation proteins. Together, these results confirm the value of an integrated approach in the evaluation of large-scale omics data sets and provide valuable information with regard to the biological factors that determine the protein and peptide profile of human milk.

In **Chapter 6** we showed the results of additional data analysis and we integrated and discussed these with the results of the preceding chapters. We derived from this integration that the presence of non-human proteins is dependent on a specific pathological state of the mother (house dust mite allergy), but that non-human peptides are commonly found in the milk and match with dietary peptides found in blood. Furthermore, we showed that the interindividual variation in both protein and peptide profile is, for a significant part, due to ethnicity. We also examined the relation between proteins in the milk and the development of allergy in the breastfed infant and highlighted that our findings point to differences in immunoglobulins and cells present in the milk. Lastly, we reported a negative result, showing that no relation was found between the peptide profile of the milk and allergy status of mother or allergy development in the child.

Concluding, we have demonstrated in this thesis that the human milk protein and peptide profiles include exogenous proteinaceous components, have substantial interindividual variation, and relate to ethnicity of the mother. Additionally, we showed that there are differences in the milk proteome of mother-child dyads with an allergy when compared with mother-child dyads without allergy. The analysis of the human milk protein and peptide profile undertaken in this thesis, has extended the current knowledge and shows the need of a systems approach to discover their complexity.

Samenvatting

Om de pasgeborene voeding en bescherming te bieden, produceert het lichaam van de moeder een complexe vloeistof, moedermelk. Eén van de vele bestanddelen van moedermelk zijn de eiwitten, een diverse groep moleculen met uiteenlopende functies. Deze functies zijn bijvoorbeeld het leveren van voedingsstoffen voor de groei, hulp bij vertering, het transport van mineralen, of een rol in immuunrespons. Daarnaast fungeren sommige eiwitten als proteasen en splitsen ze andere eiwitten op in kleinere fragmenten, de peptiden. Peptiden kunnen bioactieve eigenschappen hebben die verschillen van de functies van het eiwit waarvan zij afkomstig zijn. Dit kan bijvoorbeeld antimicrobiële activiteit zijn of stimulering van de celgroei. Eiwitten en peptiden in moedermelk zijn door de jaren heen al grondig bestudeerd. Toch blijven er nog veel vragen onbeantwoord over hun oorsprong, biologische functies en effect op de gezondheid van de zuigeling. In dit proefschrift onderzochten we de peptide- en eiwitsamenstelling van moedermelk en hoe deze zich verhoudt tot allergie van de moeder en de ontwikkeling van allergie bij het kind.

Eén aspect van de eiwitsamenstelling van moedermelk is de aanwezigheid van niet-humane eiwitten. Dit was het belangrijkste onderwerp van **Hoofdstuk 2**, waar we de eiwitsamenstelling van melk van 10 allergische en 10 niet-allergische moeders analyseerden. Het meest opvallende resultaat dat hieruit naar voren kwam was dat moeders met huisstofmijtallergie meer β-lactoglobuline in hun melk hadden in vergelijking met niet-allergische moeders. Daarnaast vonden we nog andere niet-humane eiwitten en eiwitfragmenten in de melk, afkomstig van verschillende organismen. De belangrijkste bron hiervan bleek koemelk te zijn. De identificatie van deze eiwitten en eiwitfragmenten werd op meerdere manieren bevestigd. Samen doen deze bevindingen vermoeden dat er tussen allergische en niet-allergische moeders een verschil is in het transport van eiwitten of eiwitfragmenten door de darmbarrière, waardoor voedingseiwitten in de bloedbaan terecht kunnen komen en uiteindelijk in de melk.

De normale variatie in de peptide- en eiwitsamenstelling tussen verschillende individuen is belangrijke informatie die de interpretatie van verschillen tussen groepen van melkmonsters kan ondersteunen. We hebben deze interindividuele variatie in kaart gebracht in **Hoofdstuk 3** aan de hand van een reeks van 286 monsters afkomstig van 29 moeders, samengevoegd per moeder. Hierin werd aanzienlijke interindividuele variatie waargenomen voor een selectie van eiwitten en peptiden, waarbij de peptiden over het algemeen een grotere variatie vertoonden. Naast de eiwitten en peptiden hebben we ook de metabolieten in deze monsters geanalyseerd. Metabolieten kunnen worden gezien als indicatoren van biologische processen in het menselijk lichaam en werden in deze studie gebruikt om te onderzoeken of patronen in variatie konden worden verklaard door specifieke biologische processen. Hiervoor hebben we in de data-analyse gebruik gemaakt van gewogen correlatie netwerkanalyse, waarmee we de associatie tussen clusters van eiwitten, clusters van peptiden, en metabolieten konden onderzoeken. De uitkomsten van deze studie geven inzicht in de dynamische samenstelling van moedermelkeiwitten, -peptiden en -metabolieten.

In **Hoofdstuk 4** zetten we het onderzoek voort naar de eiwitsamenstelling en het mogelijke verband met allergie van de moeder en allergie-ontwikkeling bij het kind. In samenwerking met de Canadese CHILD Cohort Studie, analyseerden we 300 melkmonsters van moeder-kind paren met verschillende allergiestatussen. Voor de data-analyse werden verschillende methoden gebruikt, waaronder univariate analyse, multivariate verkennende analyse, Random Forest classificatiemodellen, en netwerkanalyse. We vonden onder andere een trend dat immunoglobulinemoleculen meer aanwezig zijn in melk van allergische moeders gegeven aan zuigelingen die uiteindelijk een allergie ontwikkelen. Bovendien vonden we ook dat de eiwitconnectiviteit, dat wil zeggen, de mate waarin de intensiteit van eiwitten geassocieerd kan worden met die van andere eiwitten, minder is in melk van niet-allergische moeders met zuigelingen die geen allergie ontwikkelen op latere leeftijd. Wij suggereerden dat dit zou kunnen komen door een ontregeling van cellulaire processen of een verschil in cellen die in de melk aanwezig zijn. Ten slotte was het opmerkelijk dat met behulp van de netwerkanalyse subtiele maar relevante informatie aan het licht kwam die niet met de klassieke methoden voor data-analyse aan het licht was gekomen.

Nadat we de eiwitsamenstelling van de monsters van de CHILD Cohort Studie hadden bestudeerd, voegden we deze gegevens samen met resultaten van peptide-analyse en onderzochten we met een netwerkbenadering de eiwit-eiwit, peptide-peptide, en eiwit-peptide associaties in **Hoofdstuk 5**. Deze studie toonde aan dat dergelijke associaties belangrijke informatie kunnen geven over de biologische processen en mechanismen die betrokken zijn bij de synthese en secretie van eiwitten in melk. Bovendien konden we aan de hand van peptide-peptide-associaties een onderscheid maken tussen patronen van endoproteolytische en exoproteolytische activiteit. Wat betreft de eiwit-peptide associaties toonden we specifieke afbraak van enkele eiwitten aan, waarbij de hoeveelheid van de peptiden afhangt van de hoeveelheid van deze eiwitten. Dit werd onder meer waargenomen bij eiwitten van het complementsysteeem en stollingsfactoren. Samen bevestigen deze resultaten de waarde van een geïntegreerde aanpak bij de evaluatie van grootschalige omics-datasets. Daarnaast leveren ze waardevolle informatie op met betrekking tot de biologische factoren die het eiwit- en peptideprofiel van moedermelk bepalen.

In **Hoofdstuk 6** hebben we de resultaten van aanvullende data-analyse laten zien, en deze samengevoegd en besproken met de resultaten van de voorgaande hoofdstukken. Uit deze integratie hebben we afgeleid dat de aanwezigheid van niet-humane eiwitten in moedermelk afhankelijk is van een specifieke pathologische toestand van de moeder (huisstofmijtallergie). In het geval van peptiden van niet-humane eiwitten geldt dat deze algemeen worden aangetroffen in de moedermelk en overeenkomen met in het bloed aangetroffen peptiden afkomstig uit voeding. Daarnaast lieten we zien dat de interindividuele variatie in zowel het eiwitals in het peptideprofiel voor een belangrijk deel te wijten is aan etniciteit. We onderzochten de relatie tussen eiwitten in de melk en de ontwikkeling van allergie bij zuigelingen nog verder, wat benadrukte dat onze bevindingen wijzen op verschillen in immunoglobulinen en cellen die in de melk aanwezig zijn. Tenslotte hebben wij laten zien dat er geen verband is gevonden tussen het peptideprofiel van de melk en de allergiestatus van de moeder of de ontwikkeling van allergie bij het kind.

Concluderend hebben we in dit proefschrift aangetoond dat de eiwit- en peptideprofielen van moedermelk niet-humane componenten bevatten, aanzienlijke interindividuele variatie hebben, en verband houden met de etniciteit van de moeder. Bovendien zagen we dat er verschillen waren in de melkeiwitsamenstelling van moeder-kind paren met een allergie in vergelijking met moeder-kind paren zonder allergie. De analyse van het eiwit- en peptideprofiel van moedermelk in dit proefschrift heeft de bestaande kennis uitgebreid en toont de noodzaak aan van een systeembenadering om de complexiteit ervan te ontdekken.

Acknowledgements

Reading the previous chapters might give rise to the illusion that everything during the PhD trajectory was about proteins, peptides, and data analysis. However, behind and besides the results of the PhD were many persons that were very important and without whom this book would not be there in its current form. I am pleased to use some pages to express my gratitude to these persons.

First and foremost, **Teresa**. We were "just" friends when I started my PhD, and now we are married. The time spent with you was like an adventure parallel to the PhD. We jumped from a plane in the Netherlands, went parasailing in Mexico, traveled several times to your country of "eternal spring", got married during a pandemic, and managed to go through the IND migration procedure. Besides these "crazy" things, you brought many less crazy (and even nicer) things into my life. Your optimism, support, love, and faith motivated me greatly, and without you, I would not have finished this thesis. Thanks for everything. Love will never fail!

Het wordt wel eens gezegd dat je een leven lang leert. In mijn geval leek dat voor een lange tijd wel erg letterlijk te gebeuren. Eerst MBO, toen HBO, een pre-master, een master, en toen nog een PhD. Zonder de steun van mijn lieve ouders **Chris** en **Riek** was dit nooit mogelijk geweest. Ik verbaas me er wel eens over dat jullie daar altijd achter stonden en het nooit hebben afgeraden. Bedankt daarvoor, maar ook voor alle andere steun die jullie gaven. Jullie hebben vaak bezoekjes gebracht aan Wageningen wat niet naast de deur is en ook hebben jullie me tijdens de studie wel 5 of 6 keer helpen verhuizen (ik ben de tel kwijt). Het was fijn om te merken dat jullie altijd interesse hadden in wat ik tijdens de PhD aan het doen was, ook al ging het soms misschien jullie pet te boven. Dit geldt ook voor mijn broers en zussen, **Ankie** en **Arjan**, **Inge** en **Joost**, en **Hans** en **Helen**, bedankt voor jullie interesse, dat werkte motiverend! Het was altijd een welkome afleiding om jullie te bezoeken, en zonder zo nu en dan een voorleesverhaaltje, watergevecht, of stoeipartij met mijn neven en nichten **Roos**, **Sem**, **Thom**, **Ezra**, **Cil**, **Hanne**, **Luuk**, **Judah**, **Juul**, **Jop**, **Max**, en **Noah**, zouden de achterliggende 4 jaar PhD een stuk saaier zijn geweest. Dankjewel!

Joost, jou wil ik nog specifiek bedanken voor het ontwerpen van de cover, uitnodiging, en de hulp bij het drukken. Super dat je dit wilde doen!

A mi familia Guatemalteca, **Giovanni** y **Lilian**, **Erick** y **Suzette**, y **Andres**, muchas gracias por su apoyo y cariño. Nuestras visitas a Guatemala están llenas de buenos recuerdos. Gracias por su comprensión las veces que tuve que trabajar ahí. Un fuerte abrazo y saludos a toda la familia Castillo y familia Mazariegos.

One of the most important factors in a PhD project is the supervisor. Before applying for the PhD position, I asked a bit around to find out what people thought of **Kasper Hettinga** as a PhD supervisor. All responses I got were very positive, and looking back, they turned out to be correct. Kasper, thank you for your supervision. You were always available, giving helpful comments, asking good questions, getting legal agreements approved and signed by all parties, and also just for a chat about running, science, bread, etc. Thanks for giving me the opportunity to do this PhD and for all the support.

I am saddened that I cannot continue here to personally express my gratitude to **Jacques Vervoort**, my second supervisor. On 19 July 2021, we received with great sadness the unexpected news that Jacques had passed away. Jacques contributed significantly to all the work reported in this thesis, and I missed him and his input in the last stretch. Besides the valuable scientific input, Jacques was able to give "no-nonsense" feedback and always had a down-to-earth view on basically everything.

Many thanks go to **Edoardo Saccenti** for suddenly taking up the role of the second supervisor. It was very kind that you were willing to help out. Your expertise and input were very valuable, you were always available for questions, and you had so much patience in explaining the network analysis. Thanks a lot!

A large part of the PhD project relied on a collaboration with the CHILD Cohort Study. First, I would like to thank CHILD Study Investigator **Meghan Azad** for all the help and guidance. It was great to learn from you in the discussions about the sample selections, the talks about the results, and your input for the manuscript. Of course I would also like to thank **all members and staff of the CHILD Cohort Study**. These include participant families, research staff, administrative staff, volunteers, lab technicians, statisticians, and clinical staff at McMaster University, University of Manitoba, University of Alberta, University of Toronto, and the University of British Columbia.

The collaboration with the CHILD Cohort Study contributed significantly to the around 950 samples that were analyzed with LC-MS/MS during this PhD project. Sometimes, this kept the LC-MS/MS busy for a month. This would not have been possible without the help of **Sjef Boeren**, who carried out the LC-MS/MS analyses. Thank you for all your support, Sjef!

During the PhD project, we had regular meetings with the NWO user committee, where I was given the opportunity to show preliminary results. All members provided very helpful input and therefore I would like to thank: **Wynand Alkema**, **Jolanda Boer**, **Lauren Brink**, **Janna van Diepen**, **Hans van Goudoever**, **Gabriela Gross**, **Anita Hartog**, **Tim Lambers**, **Sarmauli Manurung**, **Colleen McMilin**, and **Alet Wijga**. In addition, I would like to thank the NWO program officers that were involved in the project, **Lisette Krul**, **Margot Beukers**, and **Vincent van Hensbergen**.

One of the things I enjoyed the most during the PhD was (co-)supervising thesis projects. Thanks **Dagmar**, **Huizi**, **Tong**, **Yi**, **Tamar**, and **Zuomin**, for giving me this opportunity and for your help through your thesis projects!

It was very nice to have the opportunity to be working in both Food Quality and Design (FQD) and the Laboratory of Biochemistry (BIC) and I thank **Vincenzo Fogliano** and **Dolf Wijers** for these great working environments. I also thank the secretaries, **Corine**, **Kimberley**, **Lisa**, and **Laura**, who were always available to help out with administrative issues of any kind.

Since there were many colleagues, I continue with mentioning a selection of colleagues with whom I worked together in one way or the other, shared the office with, or shared food with. In case your name is not mentioned and you are reading this, thank you for being a colleague!

First, a special thanks to **Zhijun** and **Lijiao** for being friends and colleagues at the same time! It was great to share dinners and play games, hope we can continue doing this!

I also want to thank all former and current colleagues of the dairy group, especially **Sara**, **Eva**, **Ling**, **Hannah**, **Qing Ren**, **Sine**, **Naomi**, **Etske**, **Hein**, **Swantje**, **Shiksha**, **Jiaying**, **Julie**, **Zekun**, **Huifang**, and **Peiheng**. It was nice to have the monthly meetings where we could share ideas and discuss results!

Thanks also to **Tiny**, **Hao**, **Mostafa**, **Mohammad** (thanks for the heritage food!), **Hongwei**, **Qing Han**, **Jilu**, **Fabiola**, **Annelies**, **Ruth**, **Oluranti**, **Ying Lyu**, **Jianing**, **Diana**, and **Sumanth**. It was great to be colleagues!

Thanks to my friends **Chris**, **Maarten**, and all other friends in Wageningen who supported in different but invaluable ways.

Last in order, but not in importance, I would like to thank my paranymphs **Peiheng** and **Shiksha**. Thanks so much for your help!

About the author

Pieter Dekker was born on 22 December 1989 in Vlissingen. He studied Analytical Chemistry in secondary vocational education (Vlissingen) and higher professional education (Breda), after which he continued with a Master's programme in Food Technology in Wageningen. During the MSc, he did an internship at the School of Food and Nutritional Sciences at University College Cork in Ireland and a research thesis at the Laboratory of Food Chemistry at Wageningen University. After completing the MSc programme he worked for a year as a junior researcher in the department of Food Authenticity at the Wageningen Food Safety Research institute. In 2018, he started as a PhD candidate at the Food Quality and Design group and the Laboratory of Biochemistry. During the PhD project, he investigated the human milk protein and peptide profile and the relation of this to maternal allergy and allergy development in infants. The results of his research are presented in this thesis.

List of publications

This thesis:

Pieter M. Dekker, Sjef Boeren, Alet H. Wijga, Gerard H. Koppelman, Jacques J. Vervoort, Kasper A. Hettinga (2020). “Maternal allergy and the presence of non-human proteinaceous molecules in human milk.” *Nutrients, 12(4), 1169.*

Pieter M. Dekker, Sjef Boeren, Johannes B. van Goudoever, Jacques J. Vervoort, Kasper A. Hettinga (2022). “Exploring human milk dynamics: inter-individual variation in milk proteome, peptidome and metabolome.” *Journal of Proteome Research, 21(4), 1002-1016.*

Others:

Silvia Andini, **Pieter M. Dekker**, Harry Gruppen, Carla Araya-Cloutier, Jean-Paul Vincken (2019). “Modulation of glucosinolate composition in Brassicaceae seeds by germination and fungal elicitation.” *Journal of agricultural and food chemistry, 67(46), 12770-12779.*

Saskia M. van Ruth, **Pieter M. Dekker**, Erwin Brouwer, Maikel Rozijn, Sara W. Erasmus, Dara Fitzpatrick (2019). “The sound of salts by broadband acoustic resonance dissolution spectroscopy.” *Food Research International, 116, 1047-1058.*

Saskia M. van Ruth, Frans Hettinga, **Pieter M. Dekker**, Dara Fitzpatrick (2019). "The sound of the sand from the Dutch shores." *Applied Acoustics, 154, 1-10.*

Zhijun Wang, Sara W. Erasmus, **Pieter M. Dekker**, Boli Guo, Jetse J. Stoorvogel, Saskia M. van Ruth (2020). “Linking growing conditions to stable isotope ratios and elemental compositions of Costa Rican bananas (Musa spp.).” *Food Research International, 129, 108882.*

Saskia M. van Ruth, Joris van der Veeken, **Pieter M. Dekker**, Pieternel A. Luning, Wim Huisman (2020). “Feeding fiction: Fraud vulnerability in the food service industry.” *Food Research International, 133, 109158.*

Overview of completed training activities

Discipline-specific activities

Masterclass Dairy Protein Biochemistry	VLAG	2018
Summerschool MaxQuant / Perseus	Max Planck Society	2018
Masterclass Advanced Proteomics	VLAG	2019
Big Data Analysis in the Life Sciences	VLAG	2019
IMGC Milk genomics and health (student award)	IMGC	2019
IMGC Milk genomics and health	IMGC	2020
19th Human Proteome Organization World Congress	HUPO	2020
IMGC Milk genomics and health	IMGC	2021

General courses

Essentials of scientific writing & presenting	WGS	2018
Research data management	WGS	2018
Reviewing a scientific paper	WGS	2018
PhD workshop carousel	WGS	2018
VLAG PhD week	VLAG	2018
Bridging across cultural differences	WGS	2018
Philosophy and Ethics of Food Science and Technology	VLAG	2019
Scientific publishing	WGS	2019
Applied statistics	VLAG	2019
Supervising BSc & MSc thesis students	ESC	2019
Workshop Introduction to LaTeX	PE&RC	2019
Start to teach	ESC	2020
Teaching in university science laboratories	University of Amsterdam/Coursera	2021
Lecturing	ESC	2021

Other activities

Preparation of research proposal	FQD	2018
Food, Biotechnology, Development (CPT93803)	WUR, CPT-CID	2018
Scientific meetings, seminars, colloquia	FQD	2018-2022

The research described in this thesis was financially supported by the Dutch Research Council (NWO) with project No. 15299.

Financial support from Wageningen University for printing this thesis is gratefully acknowledged.

Cover design by Joost Heikoop || www.styl.nl
Printed by Print.com B.V. on FSC-certified paper.