

# Exploring Human Milk Dynamics: Interindividual Variation in Milk Proteome, Peptidome, and Metabolome

Pieter M. Dekker, Sjeff Boeren, Johannes B. van Goudoever, Jacques J. M. Vervoort, and Kasper A. Hettinga\*



Cite This: *J. Proteome Res.* 2022, 21, 1002–1016



Read Online

ACCESS |



Metrics & More



Article Recommendations

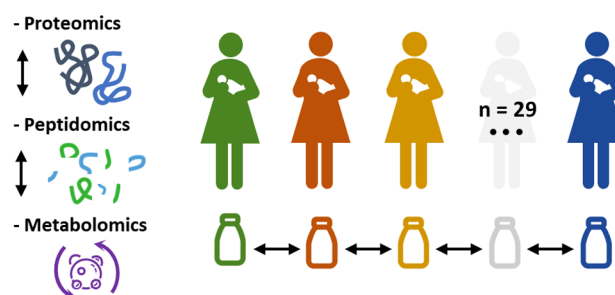


Supporting Information

**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 network 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.

**KEYWORDS:** breast milk, proteomics, peptidomics, metabolomics, proteins, peptides, metabolites, variability

## Inter-individual variation in human milk:



## 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.<sup>1</sup> Part of the proteins in human milk are synthesized in the mammary gland, for instance, caseins and  $\alpha$ -lactalbumin. Besides this, a vast number of proteins are transferred into the alveolar lumen from the systemic circulation of the mother.<sup>2</sup> Among these are for example albumin, immunoglobulin G, and even nonhuman proteins.<sup>3,4</sup>

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.<sup>5</sup> The majority of these peptides originate from the precursor protein  $\beta$ -casein. The fact that  $\beta$ -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  $\alpha$ -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.<sup>6,7</sup>

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.<sup>8,9</sup> Besides this, many recent proteomics and peptidomics studies on human milk have focused on longitudinal variation.<sup>10–15</sup> 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

**Received:** November 12, 2021

**Published:** February 1, 2022



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<sup>12</sup> and peptidome.<sup>16</sup> 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.<sup>17</sup> In recent years, weighted correlation network analysis (WGCNA) has been used to construct and analyze such coexpression networks in proteomics data.<sup>18–20</sup> 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.<sup>21</sup> 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.<sup>22,23</sup> 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.

## EXPERIMENTAL SECTION

### 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  $-18\text{ }^{\circ}\text{C}$ . Samples were picked up from homes and transported in a freezer at  $-20\text{ }^{\circ}\text{C}$  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 1. Fat content in the samples was measured by the Dutch Human Milk Bank as described by De Waard et al.<sup>24</sup>

**Table 1. Subject Demographics and Sample Characteristics**

Infant gender	female	14
	male	15
Mother BMI	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

### Proteomics

**Sample Preparation.** Human milk samples were thawed at  $4\text{ }^{\circ}\text{C}$ , and skimmed milk was obtained after centrifugation at  $1500g$  for 10 min at  $10\text{ }^{\circ}\text{C}$ . Skimmed milk was then centrifuged at  $100\text{ }000g$  for 30 min at  $30\text{ }^{\circ}\text{C}$ . Milk serum was collected, and the serum protein concentration was determined in duplicate with the Pierce bicinchoninic acid (BCA) assay (Thermo Scientific, Waltham, MA). According to these results, milk serum samples were diluted in 100 mM Tris to a concentration of  $1\text{ }\mu\text{g}/\mu\text{L}$  protein. To a  $100\text{ }\mu\text{L}$  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  $\mu$ L, containing 36  $\mu$ g 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 000g for 30 min. The filter was washed with a 50 mM ammonium bicarbonate solution. Then 100  $\mu$ L of 5 ng/ $\mu$ L 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  $\mu$ L 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.

**LC–MS/MS.** The prepared samples were analyzed with LC–MS/MS as described before.<sup>12</sup> 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 Analysis (Proteins).** The Andromeda search engine of the MaxQuant software v1.6.1.0 was used to analyze the raw LC–MS/MS data.<sup>25</sup> 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.<sup>26</sup> 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.<sup>27</sup> 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. 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.<sup>28</sup> 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<sup>29</sup> partner repository with the data set identifier PXD028280.

## Peptidomics

**Sample Preparation.** Human milk samples were thawed at 4 °C, and skimmed milk was obtained after centrifugation at 1500g 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 3000g for 10 min at 4 °C. From the resulting supernatant, 50  $\mu$ L was cleaned by solid phase extraction (SPE) on C18+ Stage tip columns (prepared in-house).<sup>30</sup> Clean-up and elution of the peptides were carried out as described before.<sup>6</sup> Lastly, peptides were reconstituted in 50  $\mu$ L of 1 mL/L formic acid in water.

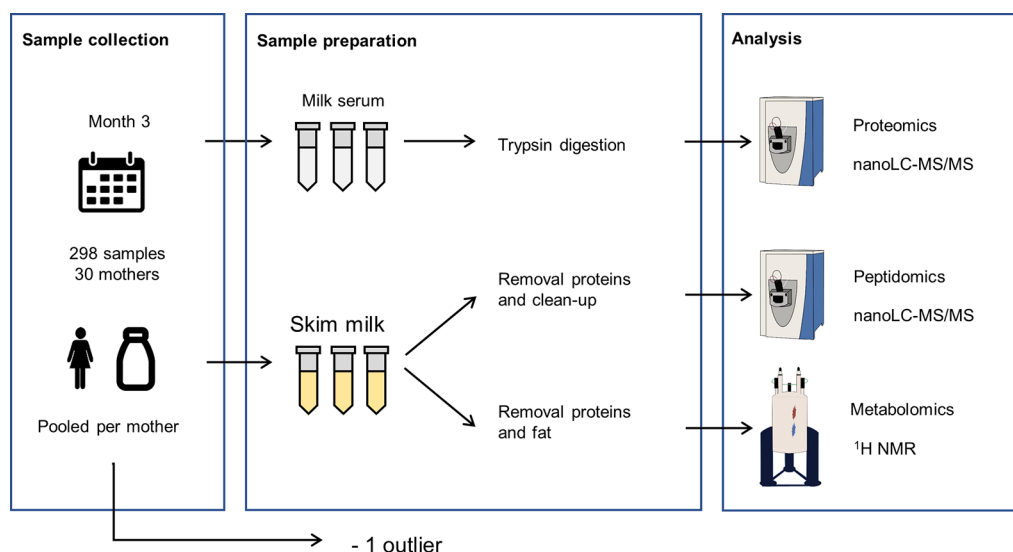
**LC–MS/MS.** Cleaned peptide samples were analyzed using a nanoLC–MS/MS mass spectrometry system (Thermo EASY nLC1000 connected to a Thermo Orbitrap XL) in which the Orbitrap was used to measure both MS and MS/MS scans. A volume of 18  $\mu$ L of sample was injected onto a 0.10  $\times$  32 mm Magic C18AQ 200A 5  $\mu$ m 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  $\mu$ L/min). Peptides were eluted from the preconcentration column onto a 0.10  $\times$  250 mm Magic C18AQ 200A 3  $\mu$ m beads analytical column (prepared in-house), and separation of the peptides took place at a flow rate of 0.5  $\mu$ L/min 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 min 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 Analysis (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<sup>29</sup> partner repository with the data set identifier PXD028294.





**Figure 1.** Schematic overview of the workflow used for the analysis of the human milk proteome, peptidome, and metabolome.

### Metabolomics

**Sample Preparation.** Human milk samples were prepared and analyzed with NMR as described in previous studies.<sup>31,32</sup> 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  $\mu$ L of supernatant was added to 500  $\mu$ L 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 <sup>1</sup>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.<sup>33</sup> Full details on parameters used for NMR analysis can be found in the Supporting Information listing S1.

**Data Analysis (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.

### Statistical Analysis

All statistical analyses were performed using R version 4.0.1.<sup>27</sup> 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).<sup>34</sup>

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.<sup>20</sup> 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 Supplemental Figure S2) and a power of 8 for the peptidomics data (Supplemental Figure S4). 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  $\rho$ ). 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.<sup>35</sup> 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.<sup>36</sup> 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.<sup>37</sup>

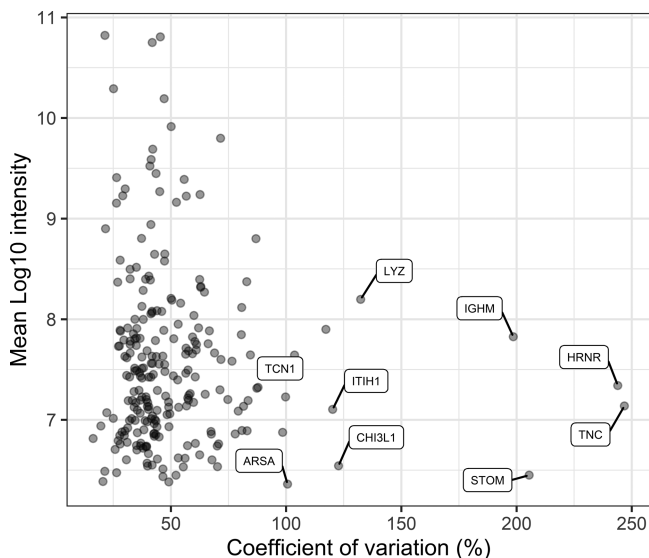
## RESULTS AND DISCUSSION

In this study, the human milk peptidome, proteome, and metabolome of 29 mothers were analyzed (see Figure 1). With the resulting data, the interindividual variation in mature human milk was investigated.

### 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 Supplemental Table S1.

As shown in Figure 2, only 4.2% ( $n = 10$ ) of the identified proteins show an extensive overall CV of >100%. From this



**Figure 2.** Distribution of the variation of proteins. The overall CV (%) on the x-axis versus the mean log<sub>10</sub> of the LFQ intensities for each identified protein on the y-axis. Proteins with the largest variation are labeled with their respective gene code.

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.<sup>14,38</sup> 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 Supplemental Figure S1A). Furthermore, it was found by Zhang et al. that the

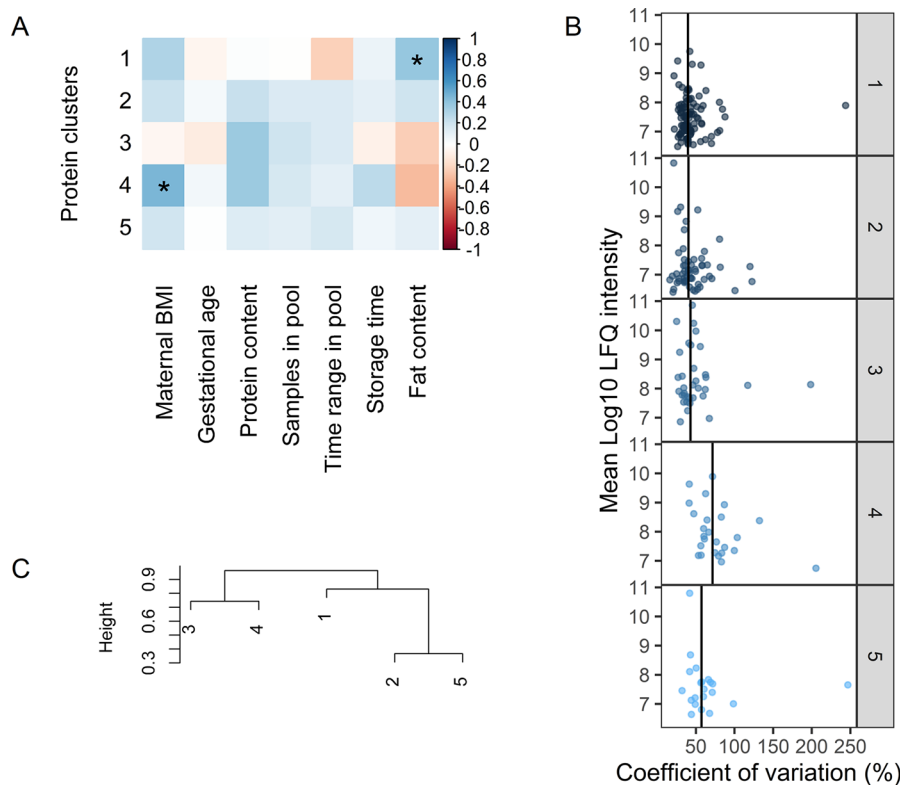
interindividual variation also surpasses the intraindividual variation in the proteome of human milk.<sup>12</sup> This indicates that the major contributor to the overall variation is interindividual variation. This is a pattern also found for the proteome of other body fluids.<sup>39–41</sup>

A few proteins show a remarkably large interindividual variation (Table 2). From these, the first four will be discussed

**Table 2. Top 10 Proteins with the Largest Interindividual Variation (CV)**

protein ID	protein name	gene	mean log <sub>10</sub> 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	Interalpha-trypsin inhibitor heavy chain H1	ITIH1	7.1	120.2
Q6ZW64	cDNA FLJ41552 fis	NA	7.9	117.2
P20061	Transcobalamin-1	TCN1	7.6	103.7
P15289	Arylsulfatase A	ARSA	6.4	100.6

more in detail. Tenascin (TNC) is well-known for its neutralizing effect on HIV.<sup>42</sup> Whereas the decrease of TNC over time was shown to become stable after 30 days postpartum,<sup>43</sup> 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  $\mu\text{g}/\text{mL}$ .<sup>44</sup> 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.<sup>45</sup> Mills et al. showed, in line with this, that airway epithelial cells generate TNC in response to viral infection.<sup>46</sup> 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.<sup>47</sup> 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 (e.g., 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.<sup>48</sup> 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.<sup>49</sup> A significant difference was observed in HRNR staining of murine mammary tissue during lactation and at the onset of involution.<sup>49</sup> 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.<sup>31</sup> Nevertheless, little is known about the function or expression of this protein in human milk,



**Figure 3.** (A) Association of eigenproteins with subject and sample characteristics using spearman correlation. Significant correlations are annotated with \* ( $p < 0.05$ ). (B) Interindividual 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.

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 2). Using ELISA, two studies showed a large interindividual variation of IgM in the first 2 weeks of lactation.<sup>50,51</sup> Although there is a gradual decrease of this protein over lactation,<sup>12,52</sup> it was found that its interindividual variation in mature milk is larger than the other immunoglobulins.<sup>53</sup>

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.<sup>22,23</sup> 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.<sup>54</sup>

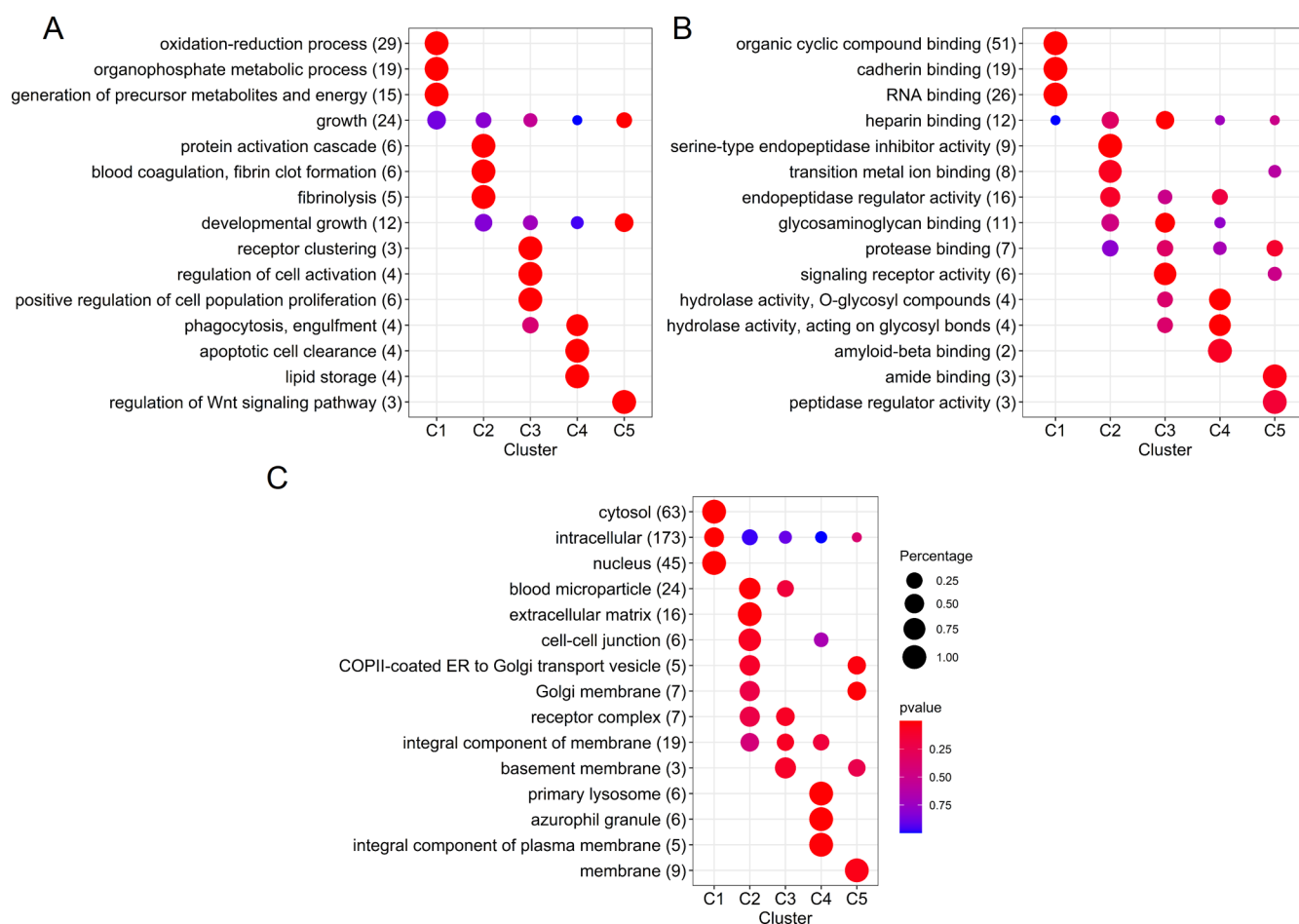
To examine whether proteins with high interindividual variation relate to specific biological processes or sample characteristics, a coexpression network was constructed using weighted correlation network analysis (WGCNA). With this, a set of 5 protein clusters was identified (see Supplemental Figures S2 and S3).

As can be seen in Figure 3B, the largest interindividual variation is present in cluster 4 with a median CV of 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 4), cluster 4 contains specifically proteins annotated with lipid storage and phagocytosis, biological processes typical for MFGM proteins.<sup>55</sup> As can be seen in Figure 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 regulation associated with obesity.<sup>56</sup> 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 4A), and has a positive association with total fat content ( $\rho = 0.39$ ,  $p = 0.04$ ) (see Figure 3A). It has been suggested that there might be a common regulation for lipid and protein synthesis in milk.<sup>57</sup> 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 4B), among which are  $\alpha$ -1-antitrypsin (SERPINA1), plasma protease C1 inhibitor



**Figure 4.** Overrepresented GO annotations in each WGCNA cluster of the proteomics data, with (A) biological processes, (B) molecular functions, and (C) cellular components.

(SERPING1), and  $\alpha$ -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 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.<sup>14,58</sup> However, such correlation was not observed in the current study, possibly because the focus was on interindividual variation, and Elwakiel et al.<sup>14</sup> 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 3C) and does not seem to be characterized by large protein groups with unique biological processes or molecular functions (Figure 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.

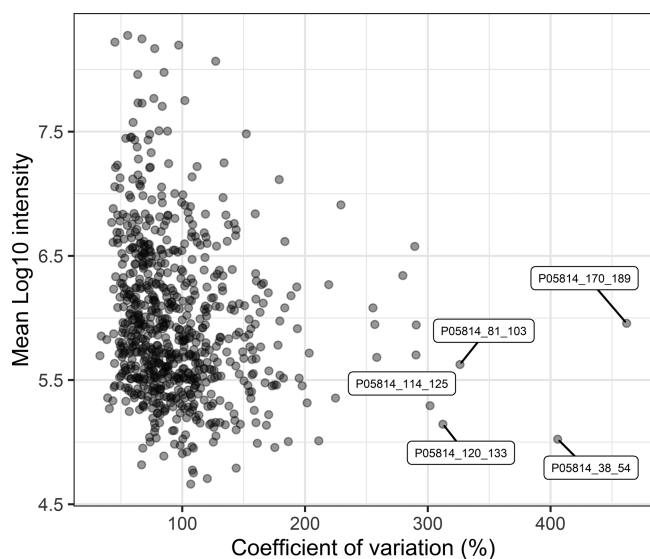
## Peptidomics

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  $\beta$ -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.<sup>59</sup> This pattern is typical for human milk peptidomics and corresponds to previous findings.<sup>6,60</sup> An overview of all identified peptides can be found in Supplemental Table S2.

As shown in Figure 5, 36.2% ( $n = 268$ ) of the identified peptides show an overall CV of >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 Supplemental Figure S1B). 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





**Figure 5.** Distribution of the variation of peptides. The overall CV (%) on the *x*-axis versus the mean log<sub>10</sub> 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.

in the technical replicates, where the median and maximum CV for proteins are 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, the 10 peptides with the highest interindividual variation are shown. All but one of these peptides are from the precursor protein  $\beta$ -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.

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.<sup>61</sup> Nevertheless, future research is needed to confirm this. In addition, the effect of single outliers 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 Supplemental Figures S4 and S5). 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 4, several of the clusters are distinctly dominated

**Table 4. Peptide Clusters with Their Size and Dominating Precursor Proteins**

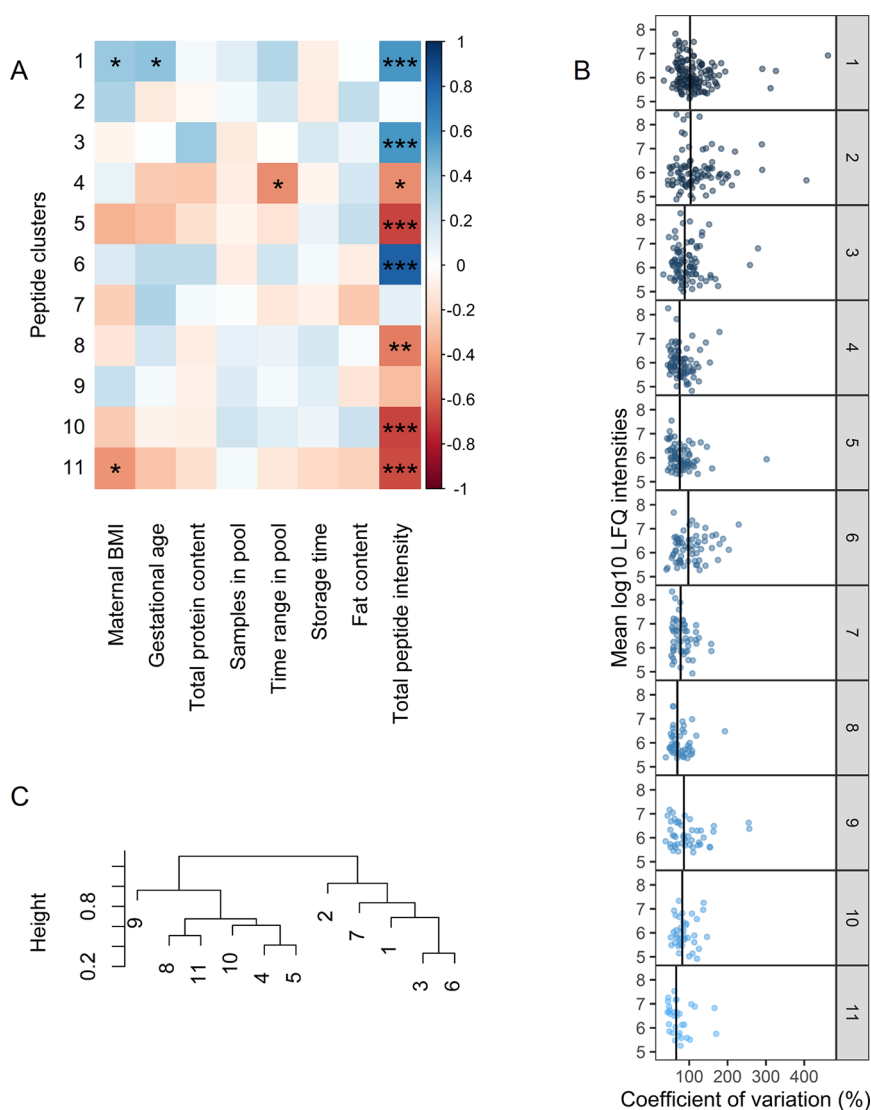
cluster label	cluster size	top precursor protein ID	top precursor protein name	number peptides per precursor protein
1	132	P05814	Beta-casein	98
		P47710	Alpha-S1-casein	7
2	91	P05814	Beta-casein	39
		P10451	Osteopontin	15
3	83	P05814	Beta-casein	39
		P47710	Alpha-S1-casein	13
4	77	P01833	Polymeric immunoglobulin receptor	34
		P05814	Beta-casein	23
5	74	Q99541	Perilipin-2	22
		P15941	Mucin-1	19
6	62	P05814	Beta-casein	40
		P10451	Osteopontin	6
		P47710	Alpha-S1-casein	6
7	53	P05814	Beta-casein	33
		P47710	Alpha-S1-casein	10
8	49	P01833	Polymeric immunoglobulin receptor	20
		P10451	Osteopontin	14
9	47	P01833	Polymeric immunoglobulin receptor	26
		P07498	Kappa-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

by peptides from certain precursor proteins. Furthermore, there are several significant correlations between eigenpeptides and sample characteristics (Figure 6).

**Table 3. Top 10 Peptides with the Largest Interindividual Variation (CV)**

sequence	protein ID	peptide range	mean log <sub>10</sub> intensity	CV (%)
SVPQPKVLPQPQVVPYPQR	P05814	170–189	6.0	461.9
KVKHEDQQQGEDEHQDK	P05814	38–54	5.0	405.8
ILPLAQPAVVLPVQPPEIMEVPK	P05814	81–103	5.6	326.2
SPTIPFFDPQIPKL	P05814	120–133	5.1	312.3
VMPVLKSPTIPF	P05814	114–125	5.3	301.8
SVPQPKVLPQPQVVPYPQ	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



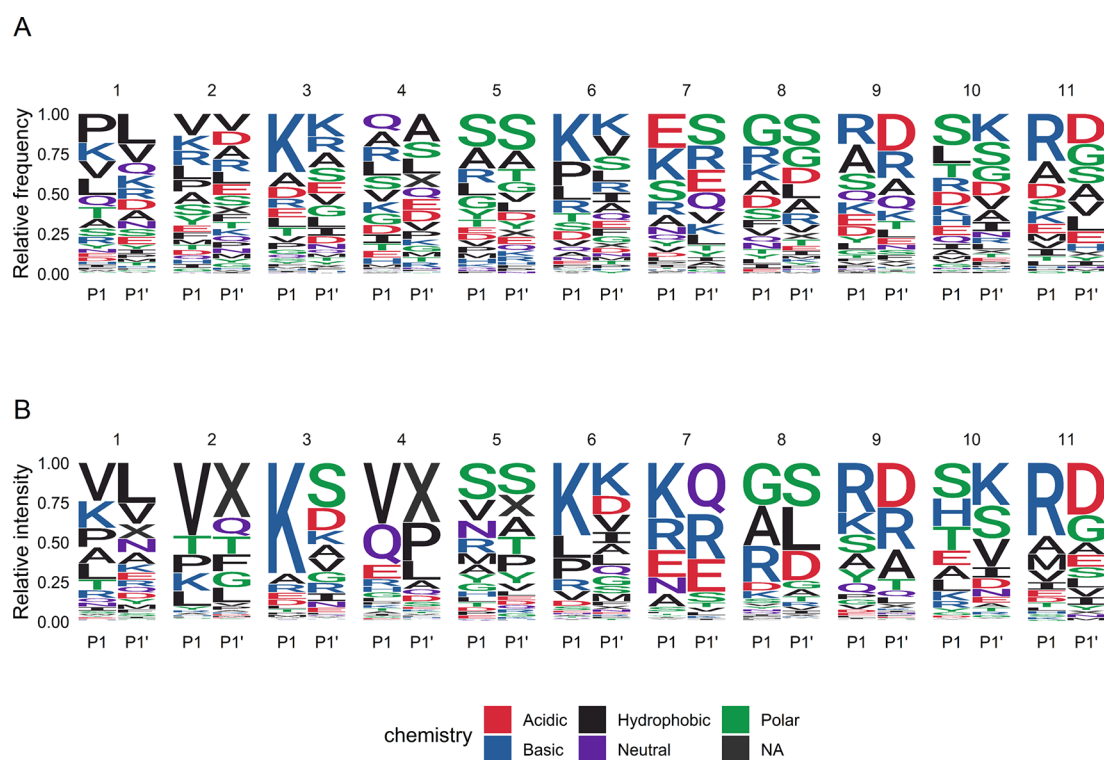


**Figure 6.** (A) Association of peptide clusters with subject and sample characteristics using Spearman correlation. Significant correlations are annotated with \* ( $p < 0.05$ ), \*\* ( $p < 0.01$ ), or \*\*\* ( $p < 0.001$ ). (B) Interindividual variation (%) in peptides per WGCNA cluster. Vertical lines indicate the median CV of the cluster. (C) Hierarchical clustering of the eigenpeptides of each cluster.

The largest interindividual variation can be observed in clusters 1, 2, 3, and 6 (median CV of 102.4, 103.7, 88.4 and 97.6%, respectively), which are dominated by peptides from  $\beta$ -casein,  $\alpha_{S1}$ -casein, and osteopontin (Table 4).

Several peptides with large variation (Table 3) show coabundance in the first three clusters. For example, two of these peptides (SVPQPKVLPPIPQQVVPYPQR and SVPQPKVLPPIPQQVVPYPQ) 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 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  $\beta$ -casein and  $\alpha_{S1}$ -

casein. Nevertheless, this cluster has a much lower median variation (77.7%), and the  $\beta$ -casein peptides in this cluster are exclusively from the N-terminal end of the protein (sequence position 16 to 54, see Supplemental Table S2). This suggests that the N-terminal of  $\beta$ -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 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 6A). This was also observed for cluster 2, even though both clusters comprise several highly abundant peptides (Figure 6B). This suggests that higher proteolytic activity seems specific for certain proteins and even protein regions. Peptide clusters dominated by  $\beta$ -casein and  $\alpha_{S1}$ -casein are associated more with each other than with the clusters dominated by other precursor proteins (Figure 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.



**Figure 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".

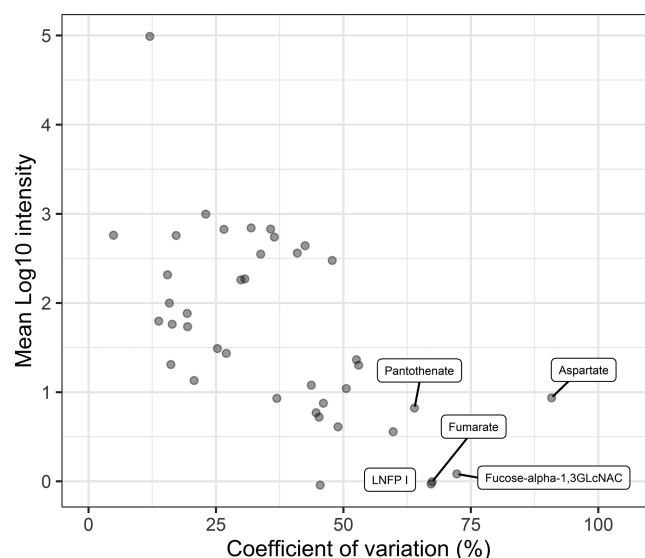
Clusters 4, 8, 9, and 11 are dominated by peptides from PIGR, with a median CV of 74.9, 69.2, 86.1, and 66.1%, respectively. Clusters 5 and 10 are dominated by peptides from MFGM proteins (median CV of 75.5 and 81.8%, respectively). The clustering of peptides of MFGM proteins is in line with Giuffrida et al., who proposed a specific mechanism for proteolysis of MFGM proteins by proteolytic enzymes in the alveolar cell membranes.<sup>62</sup> This is further supported by the fact that most peptides in these clusters do not match the specificity of plasmin (Figure 7). As shown in Figure 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  $\beta$ -casein and  $\alpha_{S1}$ -casein, possibly driven by plasmin activity and leading to the largest interindividual variation.

Taken together, these results show that the largest interindividual variation is present in peptides of  $\beta$ -casein,  $\alpha_{S1}$ -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.

### 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 Supplemental Table S3.

As shown in Figure 8, similar to the proteome and peptidome, metabolites with high intensity also show low interindividual variation. Overall variation between the samples



**Figure 8.** Distribution of the variation of metabolites. The overall 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.

is for the majority of the metabolites larger than the technical variation (see Supplemental Figure S1C).

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.<sup>63</sup> 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.<sup>63–66</sup> 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 Supplemental Figure S6). 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 Supplemental Figure S6). Removal of the Se− and Le− samples ( $n = 4$ ) from the calculations shows decreases in the interindividual variation of the HMO metabolites (see Table 5 and Supplemental Table

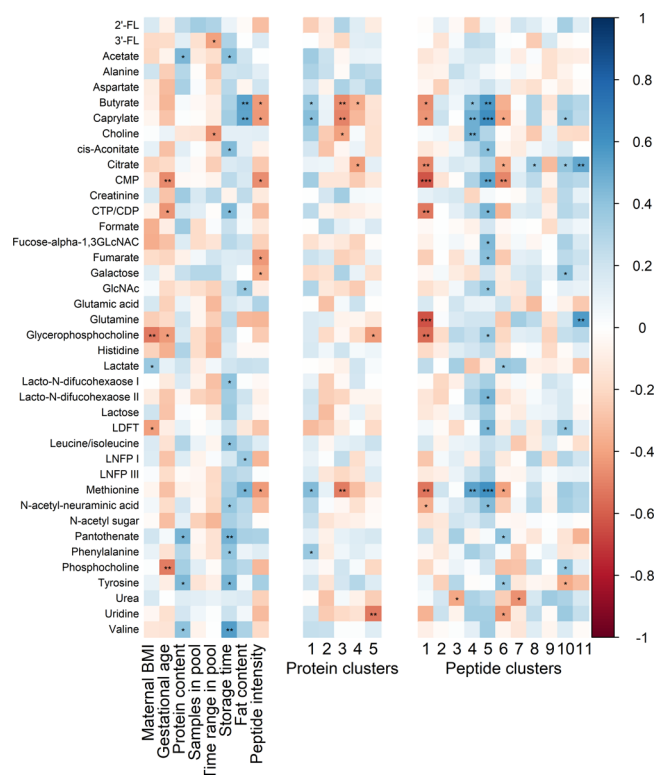
**Table 5. Metabolites with Highest (top 10 rows) and Lowest (bottom 10 rows) Interindividual Variation (CV) Together with Interindividual Variation in Samples from Mothers Proposed to Be Secretors (Se(+)) as Well as Lewis Positive (Le(+))**

metabolite	CV (%)	CV (%) Se(+)/Le(+)
Aspartate	90.8	89.3
Fucose-alpha-1,3GLcNAC	72.2	58.1
Fumarate	67.4	64.7
LNFP I	67.2	63.0
Pantothenate	63.9	64.1
CMP	59.8	52.2
3'-FL	53.0	41.1
LDFT	52.6	44.6
2'-FL	50.6	37.2
Histidine	48.9	50.6
Methionine	19.4	18.7
Lacto-N-difucohexaose II	19.4	13.8
Urea	17.2	17.0
Lactate	16.4	16.0
Valine	16.2	17.1
Acetate	15.9	17.0
Alanine	15.5	16.1
cis-Aconitate	13.8	11.2
Lactose	12.0	9.5
Butyrate	4.9	4.8

S3). Nevertheless, it should be noted that there remains a substantial interindividual variation in, for example, LNFP I, 3'-FL, and fucose- $\alpha$ -1,3-GLcNAC. This is an important finding considering the important role of HMOs in the healthy development of the infant.

Several metabolites show associations with subject and sample characteristics (Figure 9). First, a strong negative relation between glycerophosphocholine (GPC) and BMI ( $\rho = -0.52$ ,  $p = 0.003$ ) is present. It was found that in serum of patients with metabolic abnormal obesity, GPC is significantly decreased.<sup>67</sup> 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 correlation with



**Figure 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 \*\*\* ( $p < 0.001$ ).

maternal age ( $\rho < -0.42$ ,  $p < 0.022$ ). 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.<sup>68,69</sup> 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.<sup>70</sup>

Third, the fatty acids butyrate (C4:0) and caprylate (C8:0), and the amino acid methionine show a positive relation with fat content ( $\rho > 0.53$ ,  $p < 0.003$ ) and a negative relation with total peptide intensity ( $\rho < -0.44$ ,  $p < 0.018$ ). 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.<sup>71</sup> If this pathway is also present in humans, it would explain the correlations observed in the current study.

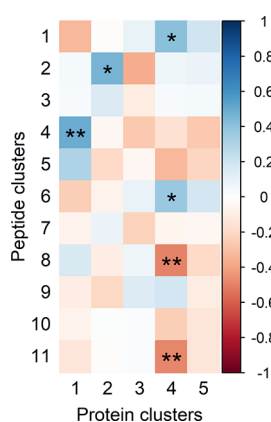
Fourthly, a negative relation was found between total peptide intensity and butyrate, caprylate, CMP, fumarate, galactose, LNFP I, and methionine ( $\rho > -0.43$ ,  $p < 0.02$ ). A higher intensity of these metabolites might indicate a lower proteolytic activity in the milk. Although little is known 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.<sup>72</sup> 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 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.<sup>73</sup> However, the strongest associations were found with pantothenate (vitamin B5) ( $\rho = 0.48$ ,  $p = 0.009$ ) and valine ( $\rho = 0.55$ ,  $p = 0.002$ ). An increase of pantothenate during frozen storage contradicts the findings of Goldsmith et al., who reported a decrease.<sup>74</sup> 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 6A and 3A, respectively). Therefore, further research on the influence of storage time on the metabolome of human milk is needed for more insight into this.

### Relation among Omics Data Sets

To identify associations between the proteomics and peptidomics data, eigenproteins were compared with eigenpeptides (Figure 10). From this, several associations were found, of which the most notable will be discussed.



**Figure 10.** Association of eigenproteins with eigenpeptides. Significant correlations are annotated with \* ( $p < 0.05$ ) or \*\* ( $p < 0.01$ ).

First, it can be observed that protein cluster 4, comprising caseins and MFGM proteins, relates positively with the  $\beta$ -casein-dominated peptide clusters 1 and 6 ( $\rho > 0.38$ ,  $p < 0.04$ ). From this, it can be hypothesized that the interindividual variation of part of the  $\beta$ -casein peptides is related to the amount of nonmicellar  $\beta$ -casein present in the milk. Second, protein cluster 2, which contains most serine protease inhibitors (SERPINs), relates positively with peptide cluster 2 ( $\rho = 0.47$ ,  $p = 0.01$ ). This peptide cluster is dominated by  $\beta$ -casein peptides but does not relate to total peptide intensity (Table 4 and Figure 6A). This points to SERPIN inhibition of serine proteases, such as thrombin and plasmin, responsible for further degradation of these peptides.

In the association of eigenproteins and eigenpeptides with metabolites (Figure 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.<sup>71</sup> 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.

## 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.

## ASSOCIATED CONTENT

### Supporting Information

The Supporting Information is available free of charge at <https://pubs.acs.org/doi/10.1021/acs.jproteome.1c00879>.

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 (PDF)

## AUTHOR INFORMATION

### Corresponding Author

Kasper A. Hettinga – Food Quality and Design Group, Wageningen University & Research, 6708 WG Wageningen, The Netherlands; [orcid.org/0000-0002-9017-4447](https://orcid.org/0000-0002-9017-4447); Phone: +31-317-48-2401; Email: [kasper.hettinga@wur.nl](mailto:kasper.hettinga@wur.nl)

### Authors

Pieter M. Dekker – Food Quality and Design Group, Wageningen University & Research, 6708 WG Wageningen, The Netherlands; Laboratory of Biochemistry, Wageningen University & Research, 6708 WE Wageningen, The Netherlands; [orcid.org/0000-0002-0189-3213](https://orcid.org/0000-0002-0189-3213)

Sjef Boeren – Laboratory of Biochemistry, Wageningen University & Research, 6708 WE Wageningen, The Netherlands



Johannes B. van Goudoever – Department of Pediatrics, Amsterdam UMC Vrije Universiteit Emma Children's Hospital, 1081 Amsterdam, The Netherlands

Jacques J. M. Vervoort – Laboratory of Biochemistry, Wageningen University & Research, 6708 WE Wageningen, The Netherlands; [orcid.org/0000-0002-0091-5687](https://orcid.org/0000-0002-0091-5687)

Complete contact information is available at:

<https://pubs.acs.org/10.1021/acs.jproteome.1c00879>

## Notes

The authors declare no competing financial interest.

The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE<sup>29</sup> partner repository with the data set identifier PXD028280. The mass spectrometry peptidomics data have been deposited to the ProteomeXchange Consortium via the PRIDE<sup>29</sup> partner repository with the data set identifier PXD028294.

## ACKNOWLEDGMENTS

The authors thank Dagmar Dokter and Yi Guo for their help in the analysis of the peptidome and proteome. The authors acknowledge the funding provided by The Netherlands Organization for Scientific Research (NWO-TTW) (project number 15299).

## DEDICATION

The authors dedicate the paper to the bright memory of Jacques J. M. Vervoort, a friend and recognized scientist, who passed away in July 2021.

## REFERENCES

- (1) Zhu, J.; Garrigues, L.; Van den Toorn, H.; Stahl, B.; Heck, A. J. Discovery and quantification of nonhuman proteins in human milk. *J. Proteome Res.* **2018**, *18*, 225–238.
- (2) Tuma, P. L.; Hubbard, A. L. Transcytosis: Crossing cellular barriers. *Physiol. Rev.* **2003**, *83*, 871–932.
- (3) Bernard, H.; Ah-Leung, S.; Drumare, M. F.; Feraudet-Tarisse, C.; Verhasselt, V.; Wal, J. M.; Créminon, C.; Adel-Patient, K. Peanut allergens are rapidly transferred in human breast milk and can prevent sensitization in mice. *Allergy Eur. J. Allergy Clin. Immunol.* **2014**, *69*, 888–897.
- (4) Wall, S. K.; Gross, J. J.; Kessler, E. C.; Villez, K.; Bruckmaier, R. M. Blood-derived proteins in milk at start of lactation: Indicators of active or passive transfer. *J. Dairy Sci.* **2015**, *98*, 7748–7756.
- (5) Dingess, K. A.; van den Toorn, H. W.; Mank, M.; Stahl, B.; Heck, A. J. Toward an efficient workflow for the analysis of the human milk peptidome. *Anal. Bioanal. Chem.* **2019**, *411*, 1351–1363.
- (6) Dingess, K. A.; de Waard, M.; Boeren, S.; Vervoort, J.; Lambers, T. T.; Van Goudoever, J. B.; Hettinga, K. Human milk peptides differentiate between the preterm and term infant and across varying lactational stages. *Food Funct.* **2017**, *8*, 3769–3782.
- (7) Ferranti, P.; Traisci, M. V.; Picariello, G.; Nasi, A.; Boschi, V.; Siervo, M.; Falconi, C.; Chianese, L.; Addeo, F. Casein proteolysis in human milk: Tracing the pattern of casein breakdown and the formation of potential bioactive peptides. *J. Dairy Res.* **2004**, *71*, 74–87.
- (8) Aslebagh, R.; Channaveerappa, D.; Arcaro, K. F.; Darie, C. C. Proteomics analysis of human breast milk to assess breast cancer risk. *Electrophoresis* **2018**, *39*, 653–665.
- (9) Hettinga, K. A.; Reina, F. M.; Boeren, S.; Zhang, L.; Koppelman, G. H.; Postma, D. S.; Vervoort, J. J.; Wijga, A. H. Difference in the breast milk proteome between allergic and non-allergic mothers. *PLoS One* **2015**, *10*, e0122234.
- (10) Liao, Y.; Alvarado, R.; Phinney, B.; Lönnerdal, B. Proteomic characterization of human milk fat globule membrane proteins during a 12 month lactation period. *J. Proteome Res.* **2011**, *10*, 3530–3541.
- (11) Zhang, Q.; Cundiff, J. K.; Maria, S. D.; McMahon, R. J.; Woo, J. G.; Davidson, B. S.; Morrow, A. L. Quantitative analysis of the human milk whey proteome reveals developing milk and Mammary-Gland functions across the first year of lactation. *Proteomes* **2013**, *1*, 128–158.
- (12) Zhang, L.; de Waard, M.; Verheijen, H.; Boeren, S.; Hageman, J. A.; van Hooijdonk, T.; Vervoort, J.; van Goudoever, J. B.; Hettinga, K. Changes over lactation in breast milk serum proteins involved in the maturation of immune and digestive system of the infant. *J. Proteomics* **2016**, *147*, 40–47.
- (13) Gao, X.; McMahon, R. J.; Woo, J. G.; Davidson, B. S.; Morrow, A. L.; Zhang, Q. Temporal changes in milk proteomes reveal developing milk functions. *J. Proteome Res.* **2012**, *11*, 3897–3907.
- (14) Elwakiel, M.; Boeren, S.; Hageman, J. A.; Szeto, I. M.; Schols, H. A.; Hettinga, K. A. Variability of serum proteins in Chinese and Dutch human milk during lactation. *Nutrients* **2019**, *11*, 499.
- (15) Zhu, J.; Dingess, K. A.; Mank, M.; Stahl, B.; Heck, A. J. Personalized profiling reveals donor- and lactation-specific trends in the human milk proteome and peptidome. *J. Nutr.* **2021**, *151*, 826–839.
- (16) Campanhon, I. B.; da Silva, M. R. S.; de Magalhães, M. T. Q.; Zingali, R. B.; Bezerra, F. F.; Torres, A. G. Protective factors in mature human milk: a look into the proteome and peptidome of adolescent mothers' breast milk. *Br. J. Nutr.* **2019**, *122*, 1377–1385.
- (17) Vella, D.; Zoppis, I.; Mauri, G.; Mauri, P.; Di Silvestre, D. From protein-protein interactions to protein co-expression networks: a new perspective to evaluate large-scale proteomic data. *Eurasip J. Bioinforma. Syst. Biol.* **2017**, *2017*, 6.
- (18) Mantini, G.; Vallés, A. M.; le Large, T. Y.; Capula, M.; Funel, N.; Pham, T. V.; Piersma, S. R.; Kazemier, G.; Bijlsma, M. F.; Giovannetti, E.; Jimenez, C. R. Co-expression analysis of pancreatic cancer proteome reveals biology and prognostic biomarkers. *Cell. Oncol.* **2020**, *43*, 1147–1159.
- (19) Zou, X.; Wang, S.; Zhang, P.; Lu, L.; Zou, H. Quantitative proteomics and weighted correlation network analysis of tear samples in adults and children with diabetes and dry eye. *Transl. Vis. Sci. Technol.* **2020**, *9*, 1–15.
- (20) Pei, G.; Chen, L.; Zhang, W. WGCNA application to proteomic and metabolomic data analysis. *Methods Enzymol.* **2017**, *585*, 135–158.
- (21) Lamerz, J.; Selle, H.; Scapozza, L.; Cramer, R.; Schulz-Knappe, P.; Mohring, T.; Kellmann, M.; Khamenia, V.; Zucht, H. D. Correlation-associated peptide networks of human cerebrospinal fluid. *Proteomics* **2005**, *5*, 2789–2798.
- (22) Lönnerdal, B.; Erdmann, P.; Thakkar, S. K.; Sauser, J.; Destailats, F. Longitudinal evolution of true protein, amino acids and bioactive proteins in breast milk: a developmental perspective. *J. Nutr. Biochem.* **2017**, *41*, 1–11.
- (23) Zhang, J.; Zhao, A.; Lai, S.; Yuan, Q.; Jia, X.; Wang, P.; Zhang, Y. Longitudinal changes in the concentration of major human milk proteins in the first six months of lactation and their effects on infant growth. *Nutrients* **2021**, *13*, 1476.
- (24) De Waard, M.; Mank, E.; Van Dijk, K.; Schoonderwoerd, A.; Van Goudoever, J. B. Holder-pasteurized human donor milk: How long can it be preserved? *J. Pediatr. Gastroenterol. Nutr.* **2018**, *66*, 479–483.
- (25) Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26*, 1367–1372.
- (26) Dekker, P. M.; Boeren, S.; Wijga, A. H.; Koppelman, G. H.; Vervoort, J. J.; Hettinga, K. A. Maternal allergy and the presence of nonhuman proteinaceous molecules in human milk. *Nutrients* **2020**, *12*, 1169.

- (27) Core Development Team. *R: A language and environment for statistical computing*; R Foundation for Statistical Computing: Vienna, Austria, 2020. <http://www.r-project.org>.
- (28) Wei, R.; Wang, J.; Jia, E.; Chen, T.; Ni, Y.; Jia, W. GSimp: AGibbs sampler based left-censored missing value imputation approach for metabolomics studies. *PLoS Comput. Biol.* **2018**, *14*, e1005973.
- (29) Vizcaino, J. A.; Csordas, A.; Del-Toro, N.; Dianas, J. A.; Griss, J.; Lavidas, I.; Mayer, G.; Perez-Riverol, Y.; Reisinger, F.; Tertent, T.; Xu, Q. W.; Wang, R.; Hermjakob, H. 2016 update of the PRIDE database and its related tools. *Nucleic Acids Res.* **2016**, *44*, D447–D456.
- (30) Lu, J.; Boeren, S.; de Vries, S. C.; van Valenberg, H. J.; Vervoort, J.; Hettinga, K. Filter-aided sample preparation with dimethyl labeling to identify and quantify milk fat globule membrane proteins. *J. Proteomics* **2011**, *75*, 34–43.
- (31) Lu, J.; Antunes Fernandes, E.; Páez Cano, A. E.; Vinitwatanakun, J.; Boeren, S.; van Hooijdonk, T.; van Kneegsel, A.; Vervoort, J.; Hettinga, K. A. Changes in milk proteome and metabolome associated with dry period length, energy balance, and lactation stage in postparturient dairy cows. *J. Proteome Res.* **2013**, *12*, 3288–3296.
- (32) Xu, W.; Vervoort, J.; Saccenti, E.; Kemp, B.; van Hooij, R. J.; van Kneegsel, A. T. Relationship between energy balance and metabolic profiles in plasma and milk of dairy cows in early lactation. *J. Dairy Sci.* **2020**, *103*, 4795–4805.
- (33) Wishart, D. S.; et al. HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Res.* **2018**, *46*, D608–D617.
- (34) Langfelder, P.; Horvath, S. WGCNA: An R package for weighted correlation network analysis. *BMC. Bioinformatics* **2008**, *9*, 559.
- (35) Yu, G.; Wang, L. G.; Han, Y.; He, Q. Y. ClusterProfiler: An R package for comparing biological themes among gene clusters. *Omi. A J. Integr. Biol.* **2012**, *16*, 284–287.
- (36) Carlson, M.; Falcon, S.; Pages, H.; Li, N. Genome wide annotation for human. *R Package version 3.8.2* **2019**, 1.
- (37) Wagih, O. Ggseqlogo: A versatile R package for drawing sequence logos. *Bioinformatics* **2017**, *33*, 3645–3647.
- (38) Zhang, L.; Ma, Y.; Yang, Z.; Jiang, S.; Liu, J.; Hettinga, K. A.; Lai, J.; Zhou, P. Geography and ethnicity related variation in the Chinese human milk serum proteome. *Food Funct.* **2019**, *10*, 7818–7827.
- (39) Hu, Y.; Malone, J. P.; Fagan, A. M.; Townsend, R. R.; Holtzman, D. M. Comparative proteomic analysis of intra- and interindividual variation in human cerebrospinal fluid. *Mol. Cell. Proteomics* **2005**, *4*, 2000–2005.
- (40) Nagaraj, N.; Mann, M. Quantitative analysis of the intra- and inter-individual variability of the normal urinary proteome. *J. Proteome Res.* **2011**, *10*, 637–645.
- (41) Zhong, W.; Gummeson, A.; Tebani, A.; Karlsson, M. J.; Hong, M. G.; Schwenk, J. M.; Edfors, F.; Bergström, G.; Fagerberg, L.; Uhlén, M. Whole-genome sequence association analysis of blood proteins in a longitudinal wellness cohort. *Genome Med.* **2020**, *12*, 53.
- (42) Fouda, G. G.; Jaeger, F. H.; Amos, J. D.; Ho, C.; Kunz, E. L.; Anasti, K.; Stamper, L. W.; Liebl, B. E.; Barbas, K. H.; Ohashi, T.; Moseley, M. A.; Liao, H. X.; Erickson, H. P.; Alam, S. M.; Permar, S. R. Tenascin-C is an innate broad-spectrum, HIV-1-neutralizing protein in breast milk. *Proc. Natl. Acad. Sci. U. S. A* **2013**, *110*, 18220–18225.
- (43) Molinari, C. E.; Casadio, Y. S.; Hartmann, B. T.; Arthur, P. G.; Hartmann, P. E. Longitudinal analysis of protein glycosylation and  $\beta$ -casein phosphorylation in term and preterm human milk during the first 2 months of lactation. *Br. J. Nutr.* **2013**, *110*, 105–115.
- (44) Mansour, R. G.; Stamper, L.; Jaeger, F.; McGuire, E.; Fouda, G.; Amos, J.; Barbas, K.; Ohashi, T.; Alam, S. M.; Erickson, H.; Permar, S. R. The presence and anti-HIV-1 function of tenascin C in breast milk and genital fluids. *PLoS One* **2016**, *11*, e0155261.
- (45) Midwood, K. S.; Chiquet, M.; Tucker, R. P.; Orend, G. Tenascin-C at a glance. *J. Cell Sci.* **2016**, *129*, 4321–4327.
- (46) Mills, J. T.; Schwenzer, A.; Marsh, E. K.; Edwards, M. R.; Sabroe, I.; Midwood, K. S.; Parker, L. C. Airway epithelial cells generate pro-inflammatory tenascin-C and small extracellular vesicles in response to TLR3 stimuli and rhinovirus infection. *Front. Immunol.* **2019**, *10*, 1–12.
- (47) Sur, S.; Khatun, M.; Steele, R.; Isbell, T. S.; Ray, R.; Ray, R. B. Exosomes from COVID-19 patients carry tenascin-C and fibrinogen- $\beta$  in triggering inflammatory signals in cells of distant organ. *Int. J. Mol. Sci.* **2021**, *22*, 1–11.
- (48) Takaishi, M.; Makino, T.; Morohashi, M.; Huh, N. H. Identification of human hornerin and its expression in regenerating and psoriatic skin. *J. Biol. Chem.* **2005**, *280*, 4696–4703.
- (49) Fleming, J. M.; Ginsburg, E.; Oliver, S. D.; Goldsmith, P.; Vonderhaar, B. K. Hornerin, an S100 family protein, is functional in breast cells and aberrantly expressed in breast cancer. *BMC Cancer* **2012**, *12*, 266.
- (50) Koenig, Á.; de Albuquerque Diniz, E. M.; Correia Barbosa, S. F.; Costa Vaz, F. A. Immunologic factors in human milk: The effects of gestational age and pasteurization. *J. Hum. Lact.* **2005**, *21*, 439–443.
- (51) Lis-Kuberka, J.; Berghausen-Mazur, M.; Orczyk-Pawilowicz, M. Lactoferrin and immunoglobulin concentrations in milk of gestational diabetic mothers. *Nutrients* **2021**, *13*, 1–18.
- (52) Goonatilake, E.; Huang, J.; Xu, G.; Wu, L.; Smilowitz, J. T.; German, J. B.; Lebrilla, C. B. Human milk proteins and their glycosylation exhibit quantitative dynamic variations during lactation. *J. Nutr.* **2019**, *149*, 1317–1325.
- (53) Affolter, M.; Garcia-Rodenas, C. L.; Vinyes-Pares, G.; Jenni, R.; Roggero, I.; Avanti-Nigro, O.; de Castro, C. A.; Zhao, A.; Zhang, Y.; Wang, P.; Thakkar, S. K.; Favre, L. Temporal changes of protein composition in breast milk of Chinese urban mothers and impact of caesarean section delivery. *Nutrients* **2016**, *8*, 504.
- (54) Diz, A. P.; Truebano, M.; Skibinski, D. O. The consequences of sample pooling in proteomics: An empirical study. *Electrophoresis* **2009**, *30*, 2967–2975.
- (55) Cao, X.; Kang, S.; Yang, M.; Li, W.; Wu, S.; Han, H.; Meng, L.; Wu, R.; Yue, X. Quantitative N-glycoproteomics of milk fat globule membrane in human colostrum and mature milk reveals changes in protein glycosylation during lactation. *Food Funct.* **2018**, *9*, 1163–1172.
- (56) Crujeiras, A. B.; Diaz-Lagares, A.; Sandoval, J.; Milagro, F. I.; Navas-Carretero, S.; Carreira, M. C.; Gomez, A.; Hervas, D.; Monteiro, M. P.; Casanueva, F. F.; Esteller, M.; Martinez, J. A. DNA methylation map in circulating leukocytes mirrors subcutaneous adipose tissue methylation pattern: A genome-wide analysis from non-obese and obese patients. *Sci. Rep.* **2017**, *7*, 41903.
- (57) Qi, L.; Yan, S.; Sheng, R.; Zhao, Y.; Guo, X. Effects of saturated long-chain fatty acid on mRNA expression of genes associated with milk fat and protein biosynthesis in bovine mammary epithelial cells. *Asian-Australasian J. Anim. Sci.* **2014**, *27*, 414–421.
- (58) Chowanadisai, W.; Lönnnerdal, B.  $\alpha$ 1-antitrypsin and antichymotrypsin in human milk: Origin, concentrations, and stability. *Am. J. Clin. Nutr.* **2002**, *76*, 828–833.
- (59) Dallas, D. C.; Guerrero, A.; Khaldi, N.; Castillo, P. A.; Martin, W. F.; Smilowitz, J. T.; Bevins, C. L.; Barile, D.; German, J. B.; Lebrilla, C. B. Extensive in vivo human milk peptidomics reveals specific proteolysis yielding protective antimicrobial peptides. *J. Proteome Res.* **2013**, *12*, 2295–2304.
- (60) Guerrero, A.; Dallas, D. C.; Contreras, S.; Chee, S.; Parker, E. A.; Sun, X.; Dimapasoc, L.; Barile, D.; German, J. B.; Lebrilla, C. B. Mechanistic peptidomics: Factors that dictate specificity in the formation of endogenous peptides in human milk. *Mol. Cell. Proteomics* **2014**, *13*, 3343–3351.
- (61) Dallas, D. C.; Smink, C. J.; Robinson, R. C.; Tian, T.; Guerrero, A.; Parker, E. A.; Smilowitz, J. T.; Hettinga, K. A.; Underwood, M. A.; Lebrilla, C. B.; Bruce German, J.; Barile, D. Endogenous human milk peptide release is greater after preterm birth than term birth. *J. Nutr.* **2015**, *145*, 425–433.

(62) Giuffrida, M. G.; Cavaletto, M.; Lamberti, C.; Dellavalle, G.; Fabris, C.; Conti, A.; Sabatino, G.; Testa, T.; Coscia, A.; Giuliani, F.; Bertino, E. Proteolysis of milk fat globule membrane proteins in preterm milk: A transient phenomenon with a possible biological role? *Int. J. Immunopathol. Pharmacol.* **2008**, *21*, 959–967.

(63) Smilowitz, J. T.; O'Sullivan, A.; Barile, D.; German, J. B.; Lönnerdal, B.; Slupsky, C. M. The human milk metabolome reveals diverse oligosaccharide profiles. *J. Nutr.* **2013**, *143*, 1709–1718.

(64) Dessì, A.; Briana, D.; Corbu, S.; Gavrili, S.; Marincola, F. C.; Georgantzi, S.; Pintus, R.; Fanos, V.; Malamitsi-Puchner, A. Metabolomics of breast milk: The importance of phenotypes. *Metabolites* **2018**, *8*, 79.

(65) Praticò, G.; Capuani, G.; Tomassini, A.; Baldassarre, M. E.; Delfini, M.; Miccheli, A. Exploring human breast milk composition by NMR-based metabolomics. *Nat. Prod. Res.* **2014**, *28*, 95–101.

(66) Wang, A.; Koleva, P.; du Toit, E.; Geddes, D. T.; Munblit, D.; Prescott, S. L.; Eggesbø, M.; Johnson, C. C.; Wegienka, G.; Shimojo, N.; Campbell, D.; Kozyrskyj, A. L.; Slupsky, C. M. The milk metabolome of non-secretor and Lewis negative mothers. *Front. Nutr.* **2021**, *7*, 1–9.

(67) Chen, H. H.; Tseng, Y. J.; Wang, S. Y.; Tsai, Y. S.; Chang, C. S.; Kuo, T. C.; Yao, W. J.; Shieh, C. C.; Wu, C. H.; Kuo, P. H. The metabolome profiling and pathway analysis in metabolic healthy and abnormal obesity. *Int. J. Obes.* **2015**, *39*, 1241–1248.

(68) Li, C. M. C.; Shapiro, H.; Tsiobikas, C.; Selfors, L. M.; Chen, H.; Rosenbluth, J.; Moore, K.; Gupta, K. P.; Gray, G. K.; Oren, Y.; Steinbaugh, M. J.; Guerriero, J. L.; Pinello, L.; Regev, A.; Brugge, J. S. Aging-associated alterations in mammary epithelia and stroma revealed by single-cell RNA sequencing. *Cell Rep.* **2020**, *33*, 108566.

(69) Raafat, A.; Strizzi, L.; Lashin, K.; Ginsburg, E.; McCurdy, D.; Salomon, D.; Smith, G. H.; Medina, D.; Callahan, R. Effects of age and parity on mammary gland lesions and progenitor cells in the FVB/N-RC mice. *PLoS One* **2012**, *7*, e43624.

(70) Xu, W.; Van Kneysel, A.; Saccenti, E.; Van Hoeij, R.; Kemp, B.; Vervoort, J. Metabolomics of milk reflects a negative energy balance in cows. *J. Proteome Res.* **2020**, *19*, 2942–2949.

(71) Qi, H.; Meng, C.; Jin, X.; Li, X.; Li, P.; Gao, X. Methionine promotes milk protein and fat synthesis and cell proliferation via the SNAT2-PI3K signaling pathway in bovine mammary epithelial cells. *J. Agric. Food Chem.* **2018**, *66*, 11027–11033.

(72) Gibson, P. R.; Rosella, O.; Rosella, G.; Young, G. P. Butyrate is a potent inhibitor of urokinase secretion by normal colonic epithelium in vitro. *Gastroenterology* **1994**, *107*, 410–419.

(73) Wu, J.; Domellöf, M.; Zivkovic, A. M.; Larsson, G.; Öhman, A.; Nording, M. L. NMR-based metabolite profiling of human milk: A pilot study of methods for investigating compositional changes during lactation. *Biochem. Biophys. Res. Commun.* **2016**, *469*, 626–632.

(74) Goldsmith, S. J.; Eitenmiller, R. R.; Toledo, R. T.; Barnhart, H. M. Effects of processing and storage on the water-soluble vitamin content of human milk. *J. Food Sci.* **1983**, *48*, 994–995.