Exploring the fine composition of Camelus milk from Kazakhstan with emphasis on protein components

Alma Ryskaliyeva

université PARIS-SACLAY

NNT 2018SACLA016

Thèse de doctorat

Exploring the fine composition of *Camelus* milk from Kazakhstan with emphasis on protein components

Thèse de doctorat de l'Université Paris-Saclay
préparée à AgroParisTech (l'Institut des sciences et industries du vivant et de l'environnement)

École doctorale n°581 **: agriculture, alimentation, biologie, environnement et santé (ABIES)**
Spécialité de doctorat: Biologie moléculaire et cellulaire

Thèse présentée et soutenue à Paris, le 12 juillet 2018, par

Alma Ryskaliyeva

Composition du Jury :

Etienne VERRIER Professeur, AgroParisTech (UMR Génétique animale & Biologie intégrative)	Président
Georg ERHARDT Professor of Animal Breeding and Genetics Justus-Liebig-University Giessen, Germany	Rapporteur
Paola SACCHI Professor of Animal Breeding and Genetics Università di Torino, Italy	Rapporteur
Yves LE LOIR Directeur de Recherches INRA, Rennes, (UMR Sciences et Technologie du Lait et de l'œuf)	Examinateur
Sophie MATTALIA Ingeenier – Head of team Institut de l'élevage (Idele), Paris, France	Examinateur
Patrice MARTIN Directeur de Recherches INRA, Jouy-en-Josas (UMR Génétique animale & Biologie intégrative)	Directeur de thèse
Gaukhar KONUSPAYEVA Associated Professor, Al-Farabi Kazakh National University, Almaty, Kazakhstan	Co-Directeur de thèse

Abstract

The present study aimed to identify, in exploring the protein fraction of camelid milks from several regions of Kazakhstan, original molecules (peptide, proteins) potentially responsible for the properties attributed to camel milk. We initially analyzed globally the composition of camel milk protein fraction (proteomic approach), focusing mainly on caseins and their molecular diversity. Regarding whey proteins, we focused our efforts on the whey acidic protein (WAP) whose protease inhibitory properties are well established and which is an originality of camelids (the only large mammal with pig expressing this protein in milk). Finally, we started to isolate extracellular vesicles from milk, which are known to carry genetic information (mRNA and microRNA) and proteins involved in the communication between cells and organisms, in order to characterize their proteome.

Nearly 180 milk samples from two camel species (*Camelus bactrianus* and *C. dromedarius,* and their hybrids) we collected at different lactation stage, age and calving number, and submitted to different proven analytical techniques and proteomic approaches (SDS-PAGE, LC-MS/MS and LC-ESI-MS). A detailed characterization of 50 protein molecules, relating to genetic variants, isoforms arising from post-translational modifications and alternative splicing events, belonging to 9 protein families (κ-, α_{s1}-, α_{s2}-, β-; and γ-CN, WAP, α-LAC, PGRP, CSA/LPO) was achieved. We reported the occurrence of two unknown isoforms (sv1 and sv2) of camel α_{s2}-CN arising from alternative splicing events. Using cDNA-sequencing, α_{s2}-CNsv1 was characterized as a splicing-in variant of an in-frame 27-nucleotide sequence, of which the presence at the genome level, flanked by canonic motifs defining an exon 13 encoding the nonapeptide ENSKKTVDM, was confirmed. Isoform α_{s2}-CNsv2, which appeared to be present at different phosphorylation levels ranging between 8P and 12P, was shown to include an additional decapeptide (VKAYQIIPNL), revealed by LC-MS/MS, encoded by a 3'-extension of exon 16. In addition, we reported, for the first time to our knowledge, the occurrence of a α_{s2}-CN phosphorylation isoform with at least one phosphorylated S/T residue that does not match with the usual canonic sequence (S/T-X-A) recognized by the mammary kinase, suggesting thereby the existence of two kinase systems involved in the phosphorylation of caseins in the mammary gland.

This study also aimed to evaluate possible differences between species (genetic variability). We demonstrated that genetic variants, which hitherto seemed to be species-

specific (*e.g*. β-CN A for Bactrian and β-CN B for dromedary), are in fact present both in *C. dromedarius* and *C. bactrianus*. Regarding camel β-CN we also determined a short isoform (946 Da lighter than the full-length β-CN) arising very likely in both genetic variants (A and B) from proteolysis by plasmin.

As far as camel WAP is concerned, we identified in *C. bactrianus* a new genetic variant (B), originating from a transition G => A, leading to a codon change (GTG/ATG) in the nucleotide sequence of cDNA, which modifies a single amino acid residue at position 12 of the mature protein (V12M). In addition, we describe the existence of a splicing variant of camel WAP, arising from an alternative usage of the canonical splice site recognized as such in the other mammalian species expressing WAP in their milk. We also report that the WAP isoform predominantly present in camelids milk, first described by Beg et al. (1986) as displaying an additional sequence of 4 amino acid residues (56VSSP59) in the peptide segment connecting the two 4-DSC domains, results from the usage of an unlikely intron cryptic splice site, extending camel exon 3 on its 5' side by 12-nucleotides. In addition, we confirm that in the camel gene encoding WAP, intron 3 is a GC-AG intron, with a GC donor site showing a compensatory effect in terms of a dramatic increase in consensus at the acceptor exon position.

Finally, using an optimized protocol, we isolated camel milk-derived EVs satisfying the typical requirements for exosomal morphology, size and protein content. We identified a thousand of different proteins representing the first comprehensive proteome of camel milk-derived EVs that appears wider than camel milk proteome, including markers associated with small extracellular vesicles, such as CD63, CD81, HSP70, HSP90, TSG101 and ADAM10. We also identified proteins present in other milk components. This is particularly the case for lactadherin/MFG-E8, Ras-related proteins or CD9 that have been reported to occur in MFG. Our results strongly suggest that milk-derived exosomes have different cellular origin.

Résumé

La présente étude visait à identifier, en explorant la fraction protéique des laits de camélidés provenant de plusieurs régions du Kazakhstan, des molécules originales (peptides, protéines) potentiellement responsables des propriétés attribuées au lait de chamelle. Nous avons d'abord analysé globalement la composition de la fraction protéique des laits (approche protéomique), en nous concentrant principalement sur les caséines et leur diversité moléculaire. S'agissant des protéines du lactosérum, nous avons concentré nos efforts sur la WAP dont les propriétés « inhibiteur de protéase » sont bien établies et qui est une originalité des camélidés (seul gros mammifère avec le porc exprimant cette protéine dans le lait). Enfin, nous avons commencé à isoler des vésicules extracellulaires du lait, qui sont connus pour porter des informations génétiques (ARNm et microARN) et des protéines impliquées dans la communication entre les cellules et les organismes, afin de caractériser leur protéome.

Près de 180 échantillons de lait de 2 espèces de camélidés (*Camelus bactrianus*, *C. dromedarius* et leurs hybrides) ont été collectés à différents stades de lactation, âge et nombre de vêlages, et soumis à différentes techniques analytiques et approches protéomiques (SDS-PAGE, LC-MS/MS et LC-ESI-MS). Cinquante molécules protéiques correspondant à des variants génétiques, des isoformes issues de modifications post-traductionnelles et d'épissages différentiels, appartenant à 9 familles de protéines (κ-, α_{s1}-, α_{s2}-, β- et γ-CN, WAP, α-LAC, PGRP, CSA / LPO) ont été caractérisées. L'existence de deux isoformes inconnues (sv1 et sv2) de la caséine α_{s2} a été observée dans les deux esèces. Ces isoformes sont des variants d'épissage consécutif pour l'un à l'intégration d'une séquence de 27 nucléotides « in frame », codant pour le nonapeptide ENSKKTVDM, dont la présence a été confirmée au niveau génomique, flanquée de motifs canoniques définissant une structure exonique. La seconde isoforme, présente à différents niveaux de phosphorylation compris entre 8P et 12P, comporte un décapeptide supplémentaire (VKAYQIIPNL), révélé par LC-MS/MS, codé par une extension 3 'de l'exon 16. En outre, nous rapportons, pour la première fois à notre connaissance, l'existence d'une isoforme de phosphorylation de la caséine αs2 présentant au moins un résidu S/T phosphorylé n'appartenant pas à la séquence canonique habituelle (S/T-X-A) reconnue par la kinase mammaire, suggérant ainsi l'existence de deux systèmes impliqués dans la phosphorylation des caséines, dans la glande mammaire.

Cette étude visait également à évaluer les différences entre espèces. Nous avons démontré que les variants génétiques, qui jusqu'ici semblaient être spécifiques d'espèce (β-CN A pour Bactrian et β-CN B pour dromadaire), sont présents chez *C. dromedarius* et *C. bactrianus*. En ce qui concerne la caséine β, nous avons également pu identifier dans les deux variants génétiques (A et B) une isoforme courte (946 Da plus légère que la caséine β) résultant très probablement d'une protéolyse par la plasmine.

S'agissant de la WAP, nous avons identifié chez *C. bactrianus* un nouveau variant génétique (B), issue d'une transition G => A conduisant à un changement de codon (GTG/ATG) dans la séquence nucléotidique de l'ARNm, qui entraine un changement d'acide aminé en position 12 de la protéine mature (V12M). Un variant résultant de l'usage du site d'épissage canonique, reconnu comme tel chez les autres mammifères exprimant la WAP dans leur lait, a été identifié. La forme majoritaire de la WAP cameline, décrite pour la première fois par Beg et al. (1986) qui présente une insertion de 4 résidus d'acides aminés (56VSSP59) dans le segment peptidique reliant les deux domaines 4-DSC, résulte de l'utilisation d'un site d'épissage cryptique intronique improbable, prolongeant l'exon 3 du gène de 12 nucléotides sur son extrémité 5 '. De plus, nous confirmons que chez les camélidés, l'intron 3 du gène spécifiant la WAP, est un intron rare de type GC-AG, avec un site donneur faible qui s'accompagne d'un effet compensatoire au site consensus de l'exon accepteur.

Finalement, en utilisant un protocole optimisé, nous avons isolé à partir des laits camelins des vésicules extracellulaires (VE) présentant toutes les caractéristiques morphologiques des exosomes (taille et contenu protéique). Nous avons identifié un millier de protéines différentes représentant un premier protéome des VE du lait de chamelle qui semble plus étendu que le protéome du lait de chamelle, y compris les marqueurs associés aux VE, tels CD63, CD81, HSP70, HSP90, TSG101 et ADAM10. Nous avons également identifié des protéines présentes dans d'autres compartiments du lait. C'est notamment le cas pour les protéines apparentées à Ras, pour la lactadhérine/MFG-E8, ou CD9 qui sont présentes dans les globules gras du lait. Nos résultats suggèrent par ailleurs fortement que les exosomes dérivés du lait ont des origines cellulaires différentes.

To my family, especially, my grandmother Zylikha, much appreciation

Contents

Abbreviations

α-LAC	α-lactalbumin
aa	Amino Acid
AL	Almaty region
B	Bactrian
bp	Base pair
BTN	Butyrophilin
CN	Casein
CSA	Camel serum albumin
D	Dromedary
Da	Dalton
EVs	Extracellular vesicles
FAS	Fatty acid synthase
GlyCAM1	Glycosylation-dependent cell adhesion molecule 1
H	Hybrid
KZ	Kyzylorda region
LDH	Lactadherin
LC-MS/MS	Liquid chromatography coupled to tandem mass spectrometry
LC-ESI-MS	Liquid chromatography-electrospray ionization-tandem mass spectrometry
LPO	Lactoperoxidase
LTF	Lactoferrin
MEC	Mammary epithelial cells
MFG	Milk fat globule
MFGM	Milk fat globule membrane
P	Phosphate group
PCR	Polymerase chain reaction
PGRP	Peptido Glycan Recognition Protein
PP3	Protease peptone component 3
PTM	Post-translational modifications
RP-HPLC	Reversed-Phase High-Performance Liquid Chromatography
SDS-PAGE	Sodium dodecyl sulfate-polyacrylamide gel electrophoresis
SH	Shymkent region
SIBLING	Small integrin-binding ligand N-linked glycoproteins
sv0, sv1, sv2 and sv3	Splicing variants 0, 1, 2 and 3
UP1 and UP2	Unknown proteins 1 and 2
WAP	Whey Acidic Protein
WP	Whey protein
XO	Xanthine oxidase
ZKO	Atyrau region
4-DSC	Four-disulfide cores

Chapter 1

General Introduction

1.1 Camelids: the other non-cattle dairy species of arid and semiarid rangelands

According to the most recent statistics, the world camel population is estimated to be about 29 millions (FAO, 2017). *Camelus dromedarius* is the most frequent and widespread domestic camel species, with 90% of the total camel population (Mohandesan et al., 2017). Camels have been domesticated in a number of arid regions, including Northern and Eastern Africa, the Arabian Peninsula and Central and South West Asia. *Camelus bactrianus* forms numerical inferiority, mostly inhabits in Mongolia, China, and Central Asia. Alternatively, there are also crossed camels (hybrids) which are found mainly in Russia, Iran, Turkmenistan, and in Kazakhstan. Kazakhstan is a specific region where both domesticated species (*Camelus dromedarius* and *Camelus bactrianus*) along with wild Bactrian camels (*Camelus ferus*) and their hybrids have been maintained in mixed herds (Nurseitova et al., 2014).

Dairy camel farming is a well-established part of a local economy in many arid and semiarid rangelands (Tulgat & Schaller, 1992). Camel dairy products provide not only food, but also give nomadic herders a rich source of income. The adoption of non-cattle species for milk provision was nevertheless significant and their importance can be associated with the adaptation of such species to specific geographical areas and also to local cultural beliefs and behaviors (Faye & Konuspayeva, 2012). Camel milk production accounts for only 0.34% (2.8 millions of tons) of world milk production (Faye & Konuspayeva, 2012), but interest for "white gold of the desert" is growing (Wernery, 2006). Situated in Central Asia, Kazakhstan dairy camel farming represents a vivid example of the vital adaptation of the regional economy. Camel milk is consumed as fresh milk and as a traditional fermented drink called *shubat*, which is very popular in Central Asia countries. However, suitability for cheese production of camel milk is being low, in spite of the availability of a specific chymosin nowadays on the market and camel cheese making is in development. To our knowledge, there is no national statistical data available on camel milk industry in Kazakhstan; on average there are 35,000 of total 180,000 camel heads reared for milk production, and about 35,000 tons (1.25% of the camel milk produced in the world) of fresh and fermented milk per year are consumable (FAO, 2017). Besides its nutritional qualities, camel milk have been reported to display potential health-promoting properties (Mati et al., 2017). Fresh and fermented camel milk are widely consumed in alternative medicine for prophylactic and curative

purposes of cancer (Korashy et al., 2012), diabetes (Agrawal et al., 2013), autoimmune diseases (Al-Ayadhi et al., 2015), and hepatitis (El-Fakharany et al., 2017).

1.2 Camel milk as a source of health promoting compounds

Milk is a complete and complex food suited to the specific offspring requirements for growth and development. Camel milk has an overall composition very similar to that of bovine milk, especially as far as macro nutrients (protein, fat and lactose) are concerned. Camel milk is characterized by a high content of vitamin C.

There is increasing substantial evidence that milk contains many health-promoting compounds, influencing physiological functions or reducing disease risk. Bioactive components in milk might come from a variety of sources, such as lipids, carbohydrates and proteins as well as for minerals or vitamins. Some are synthesized and secreted by the mammary tissue, whereas others are drawn from maternal serum and carried across the mammary epithelium by receptor-mediated transport (Walther & Sieber, 2011). Furthermore, the secretion of the milk fat globule (MFG) into the acini lumen by the mammary epithelium carries with it a collection of membrane-bound proteins and lipids that are present into the milk (Figure 1.1).

Proteins are found mostly in the aqueous phase, either in soluble (whey proteins), or colloidal (caseins) states, but also in the lipid phase, associated with the milk fat globule membrane (MFGM). Over the last century, protein research has investigated mainly the importance of essential amino acids and their relevance for nutrition and health. Some peptides, with particular amino acid sequences encrypted in camel milk proteins, which are inactive in the intact protein, may play a beneficial role in human health once they are released from milk either in vivo during normal digestion or by proteolysis during bacterial fermentation (Walther & Sieber, 2011). However, other compounds that may play a role in the health-promoting properties of camel milk have to be found. Extracellular vesicles, which are vectors of nucleotide sequences (small and long non-coding RNA) and proteins, could be also involved in these biological properties. Hence, to go further into the evaluation of the potential suitability of non-bovine milks, including camel milk, in human/infant nutrition, a detailed characterization of their protein fraction that contributes largely to the nutritional

value and technological properties of milk, as well as the successful development of camel dairy industry, is required.

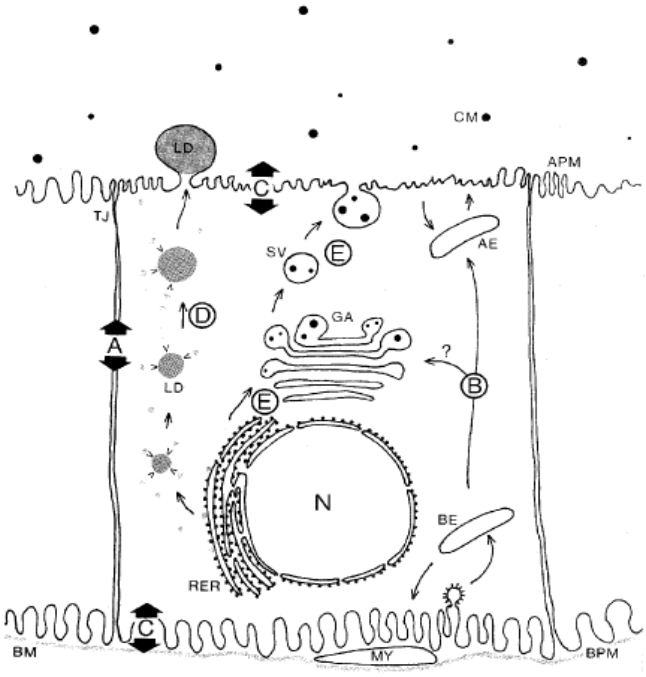


Figure 1.1. Major secretory pathways in mammary epithelial cells during lactation adapted from Mather & Keenan, 1998. (A) Paracellular route through leaky `tight' junctions; (B) Transcellular route through basal and apical endosomes and possibly through the secretory pathway; (C) Bidirectional transport of ions and small molecules via specific transporters in basal/lateral and apical plasma membranes; (D) Pathway for the assembly and secretion of milk-lipid droplets; and (E) Classical secretory pathway for the processing and secretion of milk proteins, lactose, water and ions. AE: apical endosome; APM: apical plasma membrane; BPM: basal plasma membrane; CM: casein micelle; GA: Golgi apparatus; LD: lipid droplet; N: nucleus; RER: rough endoplasmic reticulum; SV: secretory vesicle; BE: basal endosome; BM: basal membrane; TJ: `tight' junction.

1.3 The protein fractions of camel milk

Given the growing interest in camel milk, due to the health potential of its bioactive components (Al haj & Al Kanhal, 2010), the milk protein fraction of Camelids has been extensively investigated over the past 20 years and more during the last decade, with regards to casein, whey proteins and milk fat globule membrane proteins (Saadaoui et al., 2013). Whether it was on milk of one-humped Camels (*C. dromedarius)* (Alhaider et al., 2013; Elagamy et al., 1996; Ereifej et al., 2011; Erhardt et al., 2016; Felfoul et al., 2017; Hinz et al.,

2012; Kappeler et al., 1999; Merin et al., 2001; Saadaoui et al., 2013; Salmen et al., 2012; Shuiep et al., 2013; Wangoh et al., 2009) or on milk of two-humped Camels (*C. bactrianus*) (Konuspayeva et al., 2007; Ochirkhuyag et al., 1997; Yang et al., 2013), all these studies from all around the world have explored, with more or less efficient approaches, the composition of the major milk proteins.

1.3.1 Caseins

As in cow milk, *ca*. 80% of the total protein fraction of camel milk are represented by caseins (CN) that are synthesized under multi-hormonal control in the mammary gland. Associated with amorphous calcium phosphate nanoclusters they form large and stable colloidal aggregates, referred to as CN micelles. Casein micelles are present in the milk of all mammals. In bovine milk, still the most thoroughly studied milk to date, casein micelles are made of four distinct polypeptide chains: α_{s1}-, β-, α_{s2}- and κ-CN arising from the expression of four single copy autosomal genes (Figure 1.2).

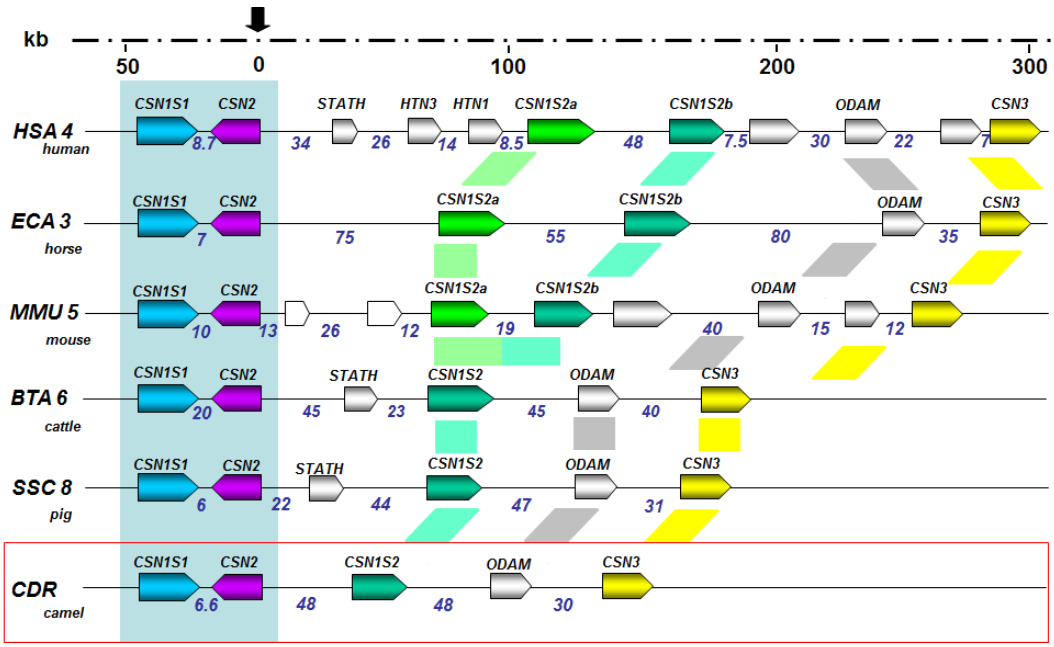


Figure 1.2. Evolution of the casein locus organization. Casein locus organization of human (*Homo sapiens*), horse (*Equus caballus*), mouse (*Mus musculus*), cattle (*Bos taurus*), pig (*Sus scrofa*) and camel (*Camelus dromedarius*) genomes (adopted from Martin et al., 2013 with additional genomic information from the NCBI) is compared. Genes are given by colored arrow boxes, showing the orientation of transcription. Putative genes based on similarity are indicated by empty boxes. Intergenic region sizes are given in kb.

Genes encoding CNs (*CSN1S1, CSN2*, *CSN1S2* and *CSN3*) are tightly linked on the same chromosome, BTA6 in cattle, CHI6 in goats (Hayes et al., 1993; Threadgill & Womack, 1990), and HSA4 in humans (Menon et al., 1992). The evolution of the CN gene cluster is postulated to have occurred by a combination of successive intra- and inter-genic exon duplications (Groenen et al., 1993; P. Martin et al., 2013; Rijnkels, 2002). In some mammals, including horse, donkey, rodents and rabbit, there are two α_{s2}-CN encoding genes differentiating in size (*CSN1S2A* and *CSN1S2B*), which may have arisen by a relatively recent gene-duplication event in rabbit (Cosenza et al., 2010; Dawson et al., 1993). However, the existence of a second α_{s2}-CN encoding gene in camel has not been reported so far.

CN micelles (Figure 1.3), which figure as calcium-transport vehicles, provide neonates with calcium at a very high concentration, which is achieved during their packaging in the secretion pathway (Semo et al., 2007). The CN micelle properties have a major influence on the technological properties of milk (Glantz et al., 2010). Micelles are characterized by a different size distribution in the milk from different mammals (Farrell et al., 2006). The average size of camel micelles is noticeably the largest: about 280 nm in diameter (Farah & Rüegg, 1989), 260 nm in goat, 190 nm in cow, and 180 nm in sheep milk (Park et al., 2007). The average diameter is inversely related to κ-CN and calcium phosphate concentrations; it has been established that large micelles are richer in calcium phosphate and smaller in κ-CN (Bornaz et al., 2009).

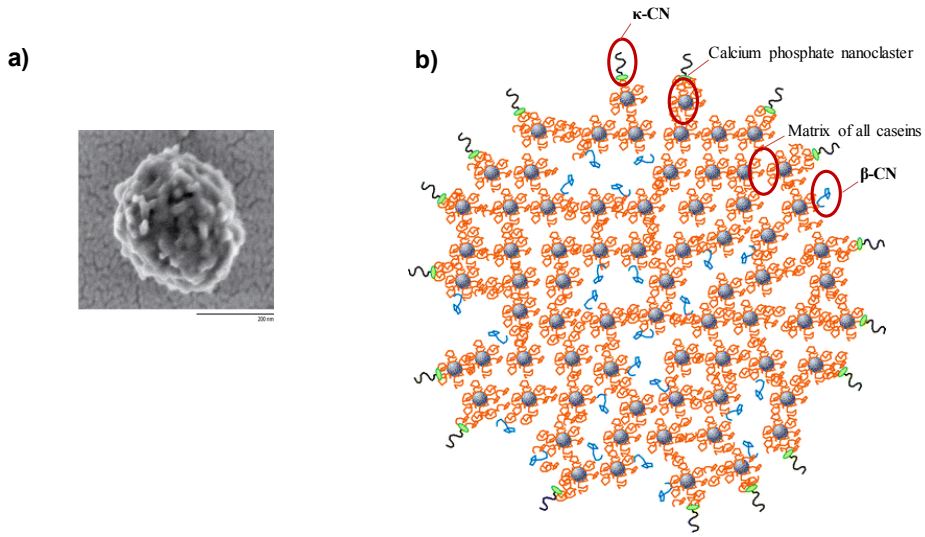


Figure 1.3. Field-emission scanning electron microscopy image of a casein micelle (a) and schematic representation of its structure (b) adopted from Dalgleish & Corredig, 2012. The α_s- and β-CNs (orange) are attached to and link the calcium phosphate nanoclusters (grey spheres). Some β-CN (blue) hydrophobically binds to other caseins and can be removed by cooling. The para κ-CN (green) and the caseinomacropeptide chains (black) are on the outermost parts of the surface.

1.3.2 Whey proteins

Camel whey is characterized by the presence of protective proteins, which display a wide range of bioactivities (Davoodi et al., 2016), including immuno-modulating (Legrand et al., 2004), anti-carcinogenic (Habib et al., 2013), antibacterial, and antifungal activities (Kanwar et al., 2015). The WPs of camel milk mainly consist of α-lactalbumin (α-LAC), glycosylation-dependent cell adhesion molecule 1 (GlyCAM1) or lactophorin which is closely related to the bovine proteose peptone component 3 (PP3), the innate immunity Peptido Glycan Recognition Protein (PGRP) and the Whey Acidic Protein (WAP). PGRP is an intracellular component of neutrophils, which modulates anti-inflammatory reaction of the immune response (Kappeler et al., 2004). Present at a very low level in ruminant milks (Tydell et al., 2002), PGRP has been detected in mammary secretions of porcine and camel (Kappeler et al., 2004) and was shown to participate in granule-mediated killing of gram-positive and negative bacteria (Dziarski et al., 2012). Lactoferrin (LTF) interacts with lipopolysaccharides of Gram-negative bacteria whereas lysozyme C binds and hydrolyzes peptidoglycans, preferably of Gram-positive bacteria, but with a lower affinity than PGRP (Sharma et al., 2011). PP3 plays an important immunological role in the lactating camel, to prevent the occurrence of mastitis, or for its newborn by inhibiting pathogen multiplication in the respiratory and gastrointestinal tracts of the suckling young (Girardet et al., 2000). WAP plays an important role in regulating the proliferation of mammary epithelial cells by preventing elastase-type serine proteases from carrying out extracellular matrix laminin degradation. In addition, a bacteriostatic activity of rat WAP against *Staphylococcus aureus* was reported (Iwamori et al., 2010). Whereas camel α-lactalbumin (α-LAC), a small milk calcium-binding globular protein, is known to possess noticeable anticancer activity, which is determined by the ability of this protein to form complexes with oleic acid (Uversky et al., 2017). Previously it was reported, that HAMLET (human α-LAC made lethal to tumor cells), a protein lipid complex formed by α-LAC and oleic acid, induces apoptosis-like death in tumor cells (Svanborg et al., 2003).

1.3.3 Milk fat globule membrane proteins

Additionally, camel milk contains proteins from milk fat globules which represent 1–4% of total protein fraction (Cavaletto et al., 2008). Due to the functional and nutritional properties, increasing attention is being paid to the components of MFGM, especially to their protein

components (Yang et al. 2015). Thus, MFGM proteins are known to be involved in many biological functions, such as inhibition of pathogen adhesion and participation in antimicrobial defense (Smolenski et al., 2007). Large-scale studies have been published for goat (Cebo et al., 2010), ovine (Pisanu et al., 2011), and bovine MFGM proteins (Reinhardt et al., 2012). More recently proteomic profiling of the MFGM from camel (*C. dromedarius*) milk has been performed by Saadaoui et al. (2013). In result, 322 proteins associated with the dromedary MFGM, such as major MFGM proteins including fatty acid synthase (FAS), xanthine oxidase (XO), butyrophilin (BTN), and lactadherin (LDH/MFG-E8), were identified. Due to the secretion process, the protein composition of MFGM reflected those of the ER and apical plasma membrane. The MFGM proteomic dataset also contained a large number of cytoplasmic proteins as found in other studies. Thus, the MFGM can reflect dynamic changes within the mammary epithelial cells (MEC) and may provide a "snapshot" of mammary gland biology under particular conditions.

1.4 Factors responsible for the molecular complexity of milk proteins

Milk protein polymorphism is a unique biological paradigm, which helps to understand protein transport, micelle formation and organization, biodiversity and evolution, the release of bioactive peptides with implications in human health. Therefore, there is a need to obtain insight into the primary structure and the way of modifications of proteins, characterize the polymorphism at the protein and mRNA levels for better understanding the genetic basis of milk quality in dairy animals (Claverol et al., 2003). Genetic variants and post-translational modifications (PTM) of some camel milk proteins were reported by several previous studies (Pauciullo et al., 2013; Shuiep et al., 2013).

1.4.1 Genetic variants

Genetic differences can be due to point mutations such as single nucleotide polymorphisms or due to DNA rearrangements such as structure variations, insertions or deletions. When changes occur in regulatory region, alteration may occur at the transcription level, and may result in different levels of expression and different amounts of proteins. When non-synonymous changes occur in coding regions, this will result in aa substitutions. One or more differences in aa sequence result in different protein variants, which may possess

different physical and chemical qualities. Recently, the impact of milk protein variants on milk composition, production and on technological properties has been reported (Lodes et al., 1996). Effects of milk proteins on human nutrition and health have further increased the interest in milk protein variants and genetic determination. In camel milk, genetic variants of the two caseins (α_{s1}- and β-CNs) arising from single-point mutations have been detected so far. Shuiep et al. (2013) demonstrated the genetic variation of camel α_{s1}-CN at the protein level leading (E30D) to the expression of two distinct variants called A and C. Pauciullo et al. (2014) described two genetic variants A and B of camel β-CN resulting from aa exchange in position 186 (M186I).

1.4.2 Alternative splicing

In mammals, alternative splicing is a major mechanism for the enhancement of transcriptome and proteome diversity that greatly expands the repertoire of protein function (Keren et al., 2010). Splicing of precursor mRNA (pre-mRNA) is a crucial regulatory stage in the pathway of gene expression: introns are removed and exons are ligated to form mRNA. The inclusion of different exons in mRNA - alternative splicing - results in the generation of different isoforms from a single gene (Keren et al., 2010). The sequences required for splicing in higher eukaryotes consist of conserved elements at the 5' and 3' splice sites and a weakly conserved element, the branch point sequence, at the site of lariat formation (Martin & Leroux, 1992). Each of these elements seems to play multiple roles in the splicing reaction, which takes place in a large ribonuclear protein complex called the spliceosome, and progress through a two-step pathway. First, cleavage occurs at the 5' splice site and the intron 5' end is joined to a 2' OH of an adenosyl phosphate residue (A*) at the branch point sequence: YTRA*Y, generating the lariat intermediate (Harris & Senapathy, 1990; Zhuang & Weiner, 1989). Second, cleavage occurs at the 3' splice junction with concomitant ligation between the two contiguous exons. In addition, exon sequences play a role in splice site selection (Reed & Maniatis, 1986). However, the weak conservation of abovementioned elements in higher eukaryotes does not usually allow the prediction of some sequences, which are recognized and used.

Such an alternative splicing event has been first demonstrated in goat α_{s1}-CN, the alternatively processed transcript of which was lacking of exons 9, 10 and 11 together encoding 37 aa residues (Leroux et al., 1992). Multiple forms of α_{s1}-CN have been also reported in sheep (Chianese et al., 1996; Ferranti et al., 1999), in camel (Kappeler et al., 1998)

and later in lama (Pauciullo & Erhardt, 2015). In camel α_{s1}-CN, two cDNAs (short and long) encoding two protein isoforms of 207 and 215 aa were described (Kappeler et al., 1998). The nucleotide sequence of the most frequent variant transcript was shown to be deleted of an octapeptide (EQAYFHLE) encoded by exon 16.

1.4.3 Post-translational modifications - Phosphorylation

The term PTM (Post-Translational Modifications) denotes changes in the polypeptide chain due to either the addition or removal of distinct chemical moieties to amino acid residues, proteolytic processing of the protein termini, or the introduction of covalent cross-links between domains of the protein. PTMs are involved in most cellular processes including the maintenance of protein structure and integrity, regulation of metabolism and defense processes, and in cellular recognition events and morphology changes (Larsen et al., 2006). Phosphorylation of proteins is one of the most frequent PTM in eukaryotic cells. It has become a common knowledge that phosphorylation of CN occurs at S or T aa residues in tripeptide sequences S/T-X-A where X represents any aa residue and A is an acidic aa residue (Mercier, 1981). This consensus sequence is recognized by FAM20C, a Golgi CN-kinase, which phosphorylates secreted phosphoproteins, including both CN and members of the small integrin-binding ligand N-linked glycoproteins (SIBLING) protein family, which modulate biomineralization (Ishikawa et al., 2012). PTM, occurring in the endoplasmic reticulum and/or Golgi complex after synthesis of the polypeptide chain, play a critical role in micelle formation and stability (Holland, 2008).

1.5 Extracellular vesicles

Milk is usually consider as a complex biological liquid in which supramolecular structures (casein micelles and milk fat globules) are found beside minerals, vitamins and soluble proteins (whey proteins) as well as cells. It was recently shown that milk contains also extracellular vesicles that are released by cells as mediators of intercellular communication. Indeed, cell communicate with neighboring cells or with distant cells through the secretion of extracellular vesicles (Tkach & Théry, 2016). Phospholipid bilayer-enclosed extracellular vesicles (EVs) are naturally generated and released from several cell domains of life (*Bacteria, Archaea, Eukarya*) into the extracellular space under physiological and pathological conditions (Delcayre et al., 2005; G. Raposo, 1996). EVs are commonly classified according to their sub-cellular origin into three major subtypes, such as

microvesicles, exosomes, and apoptotic bodies. Contents of vesicles vary with respect to mode of biogenesis, cell type, and physiologic conditions (Abels & Breakefield, 2016). Exosomes represent the smallest population among EVs, ranging in size from 30 to 150 nm in diameter (Hromada et al., 2017). They are generated inside multivesicular bodies in the endosomal compartment during the maturation of early late endosomes and are secreted when these compartments fuse with the plasma membrane (Figure 1.4) (van der Pol et al., 2012). Found in all biofluids exosomes harbor different cargos as a function of cell type and physiologic state (Abels & Breakefield, 2016).

Milk is the sole source of nutrients for the newborn and very young offspring, as well as being an important means to transfer immune components from the mother to the newborn of which the immune system is immature (Abels & Breakefield, 2016; Hromada et al., 2017). Milk is therefore thought to play an important role in the development of the immune system of the offspring. Milk is also a source of delivers molecules, via exosomes and/or microvesicles, acting on immune modulation of neonates due to their specific proteins, mRNA, long non-coding RNA and miRNA contents. Exosomes and wider EVs have come in the limelight as biological entities containing unique proteins, lipids, and genetic material. It was shown that the RNA contained in these vesicles could be transferred from one cell to another, through an emerging mode of cell-to-cell (Colombo et al., 2014; Simons & Raposo, 2009). RNAs conveyed by EVs are translated into proteins within transformed cells (mRNA), and/or are involved in regulatory functions (miRNA). For this reason, EVs are recognized as potent vehicles for intercellular communication, capable for transferring messages of signaling molecules, nucleic acids, and pathogenic factors (Kabani & Melki, 2016).

Over the last decade, EVs were widely explored as biological nanovesicles for the development of new diagnostic and therapeutic applications as a promising source for new biomarkers in various diseases (Kanada et al., 2015). For example, exosomes secreted by dendritic cells have been shown to carry MHC-peptide complexes allowing efficient activation of T lymphocytes, thus displaying immunotherapeutic potential as promoters of adaptive immune responses (Keller et al., 2006). Recently, cell culture studies showed that bovine milk-derived EVs act as a carrier for chemotherapeutic/chemopreventive agents against lung tumor xenografts *in vivo* (Munagala et al., 2016). Nevertheless, their physiological relevance has been difficult to evaluate because their origin, biogenesis and secretion mechanisms remained enigmatic.

Despite a significant number of publications describing the molecular characteristics and investigating the potential biological functions of milk-derived exosomes (Reinhardt et al., 2012; van Herwijnen et al., 2016), there are only one dealing with exosomes derived from camel milk (Yassin et al., 2016). These authors report for the first time isolation and characterization using proteomic (SDS-PAGE and western blot analysis) and transcriptomic analyses exosomes from dromedary milk at different lactation stages. However, there is no comprehensive investigation on exosomal protein variations and variability in composition between individual camels. Milk-derived EVs from Bactrian and hybrid milks have never been explored before.

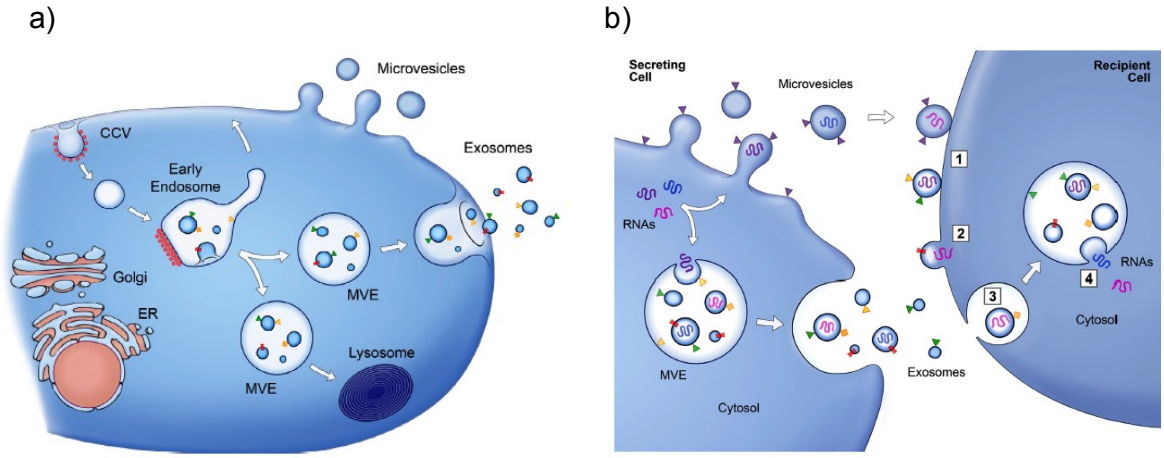


Figure 1.4. a) Release of microvesicles (MVs) and exosomes adapted from Raposo & Stoorvogel, 2013. MVs bud directly from the plasma membrane, whereas exosomes are represented by small vesicles of different sizes that are formed as the intraluminal vesicle by budding into early endosomes and MVEs and are released by fusion of multivesicular endosome (MVEs) with the plasma membrane. **b)** Schematic of protein and RNA transfer by EVs adapted from Graça Raposo and Stoorvogel (2013). Membrane-associated (triangles) and transmembrane proteins (rectangles) and RNAs (curved symbols) are selectively incorporated into the intraluminal vesicle of MVEs or into MVs budding from the plasma membrane. MVEs fuse with the plasma membrane to release exosomes into the extracellular milieu. MVs and exosomes may dock at the plasma membrane of a target cell (1). Bound vesicles may either fuse directly with the plasma membrane (2) or be endocytosed (3). Endocytosed vesicles may then fuse with the delimiting membrane of an endocytic compartment (4). Both pathways result in the delivery of proteins and RNA into the membrane or cytosol of the target cell.

1.6 Aim and outline of this study

The main objective of this thesis was to investigate the fine protein composition of *Camelus* (Bactrian, dromedary and hybrids) milks coming from different regions of Kazakhstan and of milk-derived extracellular vesicles with special emphasis on their protein contents, combining the most innovative proteomic and molecular biology approaches. We expected: i) identifying known and eventually unknown proteins from camel milk exhibiting potentially bio-activities properties; ii) providing a solid foundation for health allegations and to get a better understanding of the mechanisms involved in the true or expected effect of camel milk on human health; iii) asserting or not the originality of camel milk amongst the other dairy species. Thus, to gain an insight into the molecular diversity of camel milk proteins, we report in **Chapter 2** a complete profiling of the milk protein fraction, including in-depth characterization of the camel caseins and whey proteins comprising variants related to genetic polymorphisms, splicing defects, phosphorylation levels. In **Chapter 3**, we report the characterization of two unknown camel α_{s2}-CN splicing isoforms, resulting from translation of mRNAs yielded during the processing of primary transcripts encoding α_{s2}-CN. In **Chapter 4**, we describe a new genetic variant of camel WAP (variant B) and a splicing variant, arising from the usage of an unlikely intron cryptic splice site. In addition, we report the occurrence of a GC-AG intron (intron 3) in the camel gene encoding WAP. In **Chapter 5,** we provide results on the isolation and in-depth morphological and proteome characterization of camel milk-derived extracellular vesicles. **Chapter 6** was devoted to a general discussion of the results obtained during the course of this thesis focusing onto the consequences in terms of bioactive properties of camel milks.

References

Abels, E. R., & Breakefield, O. (2016). Introduction to Extracellular Vesicles: Biogenesis, RNA Cargo Selection, Content, Release, and Uptake. *Cellular and Molecular Neurobiology*, *36*(3), 301–312. https://doi.org/10.1007/s10571-016-0366-z

Agrawal, R. P., Tantia, P., Jain, S., Agrawal, R., & Agrawal, V. (2013). Camel milk: a possible boon for type 1 diabetic patients. *Cellular and Molecular Biology (Noisy-Le-Grand, France)*, *59*(1), 99–107.

Al-Ayadhi, L. Y., Halepoto, D. M., Al-Dress, A. M., Mitwali, Y., & Zainah, R. (2015). Behavioral Benefits of Camel Milk in Subjects with Autism Spectrum Disorder. *Journal of the College of Physicians and Surgeons--Pakistan : JCPSP*, *25*(11), 819–823. https://doi.org/11.2015/JCPSP.819823

Al haj, O. A., & Al Kanhal, H. A. (2010). Compositional, technological and nutritional aspects of dromedary camel milk. *International Dairy Journal*. https://doi.org/10.1016/j.idairyj.2010.04.003

Alhaider, A., Abdelgader, A. G., Turjoman, A. A., Newell, K., Hunsucker, S. W., Shan, B., … Duncan, M. W. (2013). Through the eye of an electrospray needle: Mass spectrometric identification of the major peptides and proteins in the milk of the one-humped camel (Camelus dromedarius). *Journal of Mass Spectrometry*, *48*(7), 779–794. https://doi.org/10.1002/jms.3213

Bornaz, S., Sahli, A., Attalah, A., & Attia, H. (2009). Physicochemical characteristics and renneting properties of camels' milk: A comparison with goats', ewes' and cows' milks. *International Journal of Dairy Technology*, *62*(4), 505–513. https://doi.org/10.1111/j.1471-0307.2009.00535.x

Cavaletto, M., Giuffrida, M. G., & Conti, A. (2008). Milk Fat Globule Membrane Components - A Proteomic Approach. In 600 (Ed.), *Bioactive Components of Milk* (pp. 129–142). Springer.

Cebo, C., Caillat, H., Bouvier, F., & Martin, P. (2010). Major proteins of the goat milk fat globule membrane. *Journal of Dairy Science*. https://doi.org/10.3168/jds.2009-2638

Chianese, L., Garro, G., Mauriello, R., Laezza, P., Ferranti, P., & Addeo, F. (1996). Occurrence of five as1-casein variants in ovine milk. *Journal of Dairy Research*, *63*, 49–59.

Claverol, S., Burlet-Schiltz, O., Gairin, J. E., & Monsarrat, B. (2003). Characterization of protein variants and post-translational modifications: ESI-MSn analyses of intact proteins eluted from polyacrylamide gels. *Molecular & Cellular Proteomics : MCP*, *2*(8), 483–493. https://doi.org/10.1074/mcp.T300003-MCP200

Colombo, M., Raposo, G., & Théry, C. (2014). Biogenesis, Secretion, and Intercellular Interactions of Exosomes and Other Extracellular Vesicles. *Annual Review of Cell and Developmental Biology*, *30*(1), 255–289. https://doi.org/10.1146/annurev-cellbio-101512-122326

Cosenza, G., Pauciullo, A., Annunziata, A. L., Rando, A., Chianese, L., Marletta, D., … Ramunno, L. (2010). Identification and characterization of the donkey CSN1S2 I and II CDNAS. *Italian Journal of Animal Science*, *9*(2). https://doi.org/10.4081/ijas.2010.e40

Dalgleish, D. G., & Corredig, M. (2012). The Structure of the Casein Micelle of Milk and Its Changes During Processing. *Annual Review of Food Science and Technology*. https://doi.org/10.1146/annurev-food-022811-101214

Davoodi, S. H., Shahbazi, R., Esmaeili, S., Sohrabvandi, S., Mortazavian, A. M., Jazayeri, S., & Taslimi, A. (2016). Health-related aspects of milk proteins. *Iranian Journal of Pharmaceutical Research*.

Dawson, S. P., Wilde, C. J., Tighe, P. J., & Mayer, R. J. (1993). Characterization of two novel casein transcripts in rabbit mammary gland. *The Biochemical Journal*, *296 (Pt 3*, 777–84. https://doi.org/10.1042/bj2960777

Delcayre, A., Shu, H., & Le Pecq, J. B. (2005). Dendritic cell-derived exosomes in cancer immunotherapy: Exploiting nature's antigen delivery pathway. *Expert Review of Anticancer Therapy*. https://doi.org/10.1586/14737140.5.3.537

Dziarski, R., Kashyap, D. R., & Gupta, D. (2012). Mammalian Peptidoglycan Recognition Proteins Kill Bacteria by Activating Two-Component Systems and Modulate Microbiome and Inflammation. *Microbial Drug Resistance*, *18*(3), 280–285.

https://doi.org/10.1089/mdr.2012.0002

El-Fakharany, E. M., El-Baky, N. A., Linjawi, M. H., Aljaddawi, A. A., Saleem, T. H., Nassar, A. Y., … Redwan, E. M. (2017). Influence of camel milk on the hepatitis C virus burden of infected patients. *Experimental and Therapeutic Medicine*, *13*(4), 1313–1320. https://doi.org/10.3892/etm.2017.4159

Elagamy, E. I., Ruppanner, R., Ismail, A., Champagne, C. P., & Assaf, R. (1996). Purification and characterization of lactoferrin, lactoperoxidase, lysozyme and immunoglobulins from camel's milk. *International Dairy Journal*, *6*(2), 129–145. https://doi.org/10.1016/0958-6946(94)00055-7

Ereifej, K. I., Alu'datt, M. H., Alkhalidy, H. A., Alli, I., & Rababah, T. (2011). Comparison and characterisation of fat and protein composition for camel milk from eight Jordanian locations. *Food Chemistry*, *127*(1), 282–289. https://doi.org/10.1016/j.foodchem.2010.12.112

Erhardt, G., Shuiep, E. T. S., Lisson, M., Weimann, C., Wang, Z., El Zubeir, I. E. Y. M., & Pauciullo, A. (2016). Alpha S1-casein polymorphisms in camel (Camelus dromedarius) and descriptions of biological active peptides and allergenic epitopes. *Tropical Animal Health and Production*, *48*(5), 879–887. https://doi.org/10.1007/s11250-016-0997-6

FAO. (2017). FAOSTAT. Retrieved from http://www.fao.org/faostat/en/#home

Farah, Z., & Rüegg, M. W. (1989). The size distribution of casein micelles in camel milk. *Food Microstructure*, *8*, 211–216.

Farrell, H. M., Malin, E. L., Brown, E. M., & Qi, P. X. (2006). Casein micelle structure: What can be learned from milk synthesis and structural biology? *Current Opinion in Colloid and Interface Science*. https://doi.org/10.1016/j.cocis.2005.11.005

Faye, B., & Konuspayeva, G. (2012). The sustainability challenge to the dairy sector - The growing importance of non-cattle milk production worldwide. *International Dairy Journal*. https://doi.org/10.1016/j.idairyj.2011.12.011

Felfoul, I., Jardin, J., Gaucheron, F., Attia, H., & Ayadi, M. A. (2017). Proteomic profiling of camel and cow milk proteins under heat treatment. *Food Chemistry*, *216*, 161–169.

https://doi.org/10.1016/j.foodchem.2016.08.007

Ferranti, P., Lilla, S., Chianese, L., & Addeo, F. (1999). Alternative nonallelic deletion is constitutive of ruminant α(s1)- casein. *Journal of Protein Chemistry*, *18*(5), 595–602. https://doi.org/10.1023/A:1020659518748

Girardet, J. M., Saulnier, F., Gaillard, J. L., Ramet, J. P., & Humbert, G. (2000). Camel (camelus dromedarius) milk PP3: evidence for an insertion in the amino-terminal sequence of the camel milk whey protein. *Biochemistry and Cell Biology = Biochimie et Biologie Cellulaire*, *78*(1), 19–26. https://doi.org/10.1139/o99-067

Glantz, M., Devold, T. G., Vegarud, G. E., Lindmark Månsson, H., Stålhammar, H., & Paulsson, M. (2010). Importance of casein micelle size and milk composition for milk gelation. *Journal of Dairy Science*, *93*(4), 1444–1451. https://doi.org/10.3168/jds.2009-2856

Groenen, M. A. M., Dijkhof, R. J. M., Verstege, A. J. M., & van der Poel, J. J. (1993). The complete sequence of the gene encoding bovine α2-casein. *Gene*, *123*(2), 187–193. https://doi.org/10.1016/0378-1119(93)90123-K

Habib, H. M., Ibrahim, W. H., Schneider-Stock, R., & Hassan, H. M. (2013). Camel milk lactoferrin reduces the proliferation of colorectal cancer cells and exerts antioxidant and DNA damage inhibitory activities. *Food Chemistry*, *141*(1), 148–152. https://doi.org/10.1016/j.foodchem.2013.03.039

Harris, N. L., & Senapathy, P. (1990). Distribution and consenus of branch point signals in eukaryotic genes: A computerized statistical analysis. *Nucleic Acids Research*, *18*(10), 3015. https://doi.org/10.1093/nar/18.10.3015

Hayes, H., Petit, E., Bouniol, C., & Popescu, P. (1993). Localization of the α-S2-casein gene (CASAS2) to the homoeologous cattle, sheep, and goat chromosomes 4 by in situ hybridization. *Cytogenetic and Genome Research*, *64*(3–4), 281–285. https://doi.org/10.1159/000133593

Hinz, K., O'Connor, P. M., Huppertz, T., Ross, R. P., & Kelly, A. L. (2012). Comparison of the principal proteins in bovine, caprine, buffalo, equine and camel milk. *Journal of Dairy Research*, *79*(02), 185–191. https://doi.org/10.1017/S0022029912000015

Holland, J. W. (2008). Post-Translational Modifications of Caseins. In *Milk Proteins* (pp. 107–132). https://doi.org/10.1016/B978-0-12-374039-7.00004-0

Hromada, C., Mühleder, S., Grillari, J., Redl, H., & Holnthoner, W. (2017). Endothelial extracellular vesicles-promises and challenges. *Frontiers in Physiology*. https://doi.org/10.3389/fphys.2017.00275

Ishikawa, H. O., Xu, A., Ogura, E., Manning, G., & Irvine, K. D. (2012). The raine syndrome protein FAM20C is a golgi kinase that phosphorylates bio-mineralization proteins. *PLoS ONE*, *7*(8). https://doi.org/10.1371/journal.pone.0042988

Iwamori, T., Nukumi, N., Itoh, K., Kano, K., Naito, K., Kurohmaru, M., … Tojo, H. (2010). Bacteriostatic activity of Whey Acidic Protein (WAP). *The Journal of Veterinary Medical Science / the Japanese Society of Veterinary Science*, *72*(5), 621–5. https://doi.org/10.1292/jvms.08-0331

Kabani, M., & Melki, R. (2016). More than just trash bins? Potential roles for extracellular vesicles in the vertical and horizontal transmission of yeast prions. *Current Genetics*. https://doi.org/10.1007/s00294-015-0534-6

Kanada, M., Bachmann, M. H., Hardy, J. W., Frimannson, D. O., Bronsart, L., Wang, A., … Contag, C. H. (2015). Differential fates of biomolecules delivered to target cells via extracellular vesicles. *Proceedings of the National Academy of Sciences*, 201418401. https://doi.org/10.1073/pnas.1418401112

Kanwar, J. R., Roy, K., Patel, Y., Zhou, S. F., Singh, M. R., Singh, D., … Kanwar, R. K. (2015). Multifunctional iron bound lactoferrin and nanomedicinal approaches to enhance its bioactive functions. *Molecules*. https://doi.org/10.3390/molecules20069703

Kappeler, S., Ackermann, M., Farah, Z., & Puhan, Z. (1999). Sequence analysis of camel (Camelus dromedarius) lactoferrin. *International Dairy Journal*, *9*(7), 481–486. https://doi.org/10.1016/S0958-6946(99)00117-X

Kappeler, S., Farah, Z., & Puhan, Z. (1998). Sequence analysis of Camelus dromedarius milk caseins. *The Journal of Dairy Research*, *65*(2), 209–222. https://doi.org/10.1017/S0022029997002847

Kappeler, S., Heuberger, C., Farah, Z., & Puhan, Z. (2004). Expression of the peptidoglycan recognition protein, PGRP, in the lactating mammary gland. *Journal of Dairy Science*, *87*(8), 2660–8. https://doi.org/10.3168/jds.S0022-0302(04)73392-5

Keller, S., Sanderson, M. P., Stoeck, A., & Altevogt, P. (2006). Exosomes: From biogenesis and secretion to biological function. *Immunology Letters*. https://doi.org/10.1016/j.imlet.2006.09.005

Keren, H., Lev-Maor, G., & Ast, G. (2010). Alternative splicing and evolution: Diversification, exon definition and function. *Nature Reviews Genetics*. https://doi.org/10.1038/nrg2776

Konuspayeva, G., Faye, B., Loiseau, G., & Levieux, D. (2007). Lactoferrin and immunoglobulin contents in camel's milk (Camelus bactrianus, Camelus dromedarius, and Hybrids) from Kazakhstan. *Journal of Dairy Science*, *90*(1), 38–46. https://doi.org/10.3168/jds.S0022-0302(07)72606-1

Korashy, H. M., Maayah, Z. H., Abd-Allah, A. R., El-Kadi, A. O. S., & Alhaider, A. a. (2012). Camel milk triggers apoptotic signaling pathways in human hepatoma HepG2 and breast cancer MCF7 cell lines through transcriptional mechanism. *Journal of Biomedicine & Biotechnology*, *2012*, 1–9. https://doi.org/10.1155/2012/593195

Larsen, M. R., Trelle, M. B., Thingholm, T. E., & Jensen, O. N. (2006). Analysis of posttranslational modifications of proteins by tandem mass spectrometry. *BioTechniques*. https://doi.org/10.2144/000112201

Legrand, D., Elass, E., Pierce, A., & Mazurier, J. (2004). Lactoferrin and host defence: An overview of its immuno-modulating and anti-inflammatory properties. *BioMetals*. https://doi.org/10.1023/B:BIOM.0000027696.48707.42

Leroux, C., Mazure, N., & Martin, P. (1992). Mutations away from splice site recognition sequences might cis-modulate alternative splicing of goat α(s1)-casein transcripts. Structural organization of the relevant gene. *Journal of Biological Chemistry*, *267*(9), 6147–6157.

Lodes, A. Krause, I. Buchberger, J. Aumann, J. et al. (1996). The influence of genetic variants of milk proteins on the compositional and technological properties of milk. 1. Casein

micelle size and the content of non-glycosylated kappa-casein. *Milchwissenschaft*, *51*(7), 368–373.

Martin, P., Cebo, C., & Miranda, G. (2013). Interspecies comparison of milk proteins: Quantitative variability and molecular diversity. In *Advanced Dairy Chemistry: Volume 1A: Proteins: Basic Aspects, 4th Edition* (pp. 387–429). https://doi.org/10.1007/978-1-4614-4714-6_13

Martin, P., & Leroux, C. (1992). Exon-skipping is responsible for the 9 amino acid residue deletion occurring near the N-terminal of human β-casein. *Biochemical and Biophysical Research Communications*, *183*(2), 750–757. https://doi.org/10.1016/0006-291X(92)90547-X

Mather, I. H., & Keenan, T. W. (1998). The cell biology of milk secretion: Historical notes. *Journal of Mammary Gland Biology and Neoplasia*. https://doi.org/10.1023/A:1018755225291

Mati, A., Senoussi-Ghezali, C., Si Ahmed Zennia, S., Almi-Sebbane, D., El-Hatmi, H., & Girardet, J. M. (2017). Dromedary camel milk proteins, a source of peptides having biological activities – A review. *International Dairy Journal*. https://doi.org/10.1016/j.idairyj.2016.12.001

Menon, R. S., Chang, Y. F., Jeffers, K. F., Jones, C., & Ham, R. G. (1992). Regional localization of human β-casein gene (CSN2) to 4pter-q21. *Genomics*, *13*(1), 225–226. https://doi.org/10.1016/0888-7543(92)90227-J

Mercier, J. C. (1981). Phosphorylation of caseins, present evidence for an amino acid triplet code posttranslationally recognized by specific kinases. *Biochimie*. https://doi.org/10.1016/S0300-9084(81)80141-1

Merin, U., Bernstein, S., Bloch-Damti, A., Yagil, R., Van Creveld, C., Lindner, P., & Gollop, N. (2001). A comparative study of milk serum proteins in camel (Camelus dromedarius) and bovine colostrum. *Livestock Production Science*, *67*(3), 297–301. https://doi.org/10.1016/S0301-6226(00)00198-6

Mohandesan, E., Speller, C. F., Peters, J., Uerpmann, H. P., Uerpmann, M., De Cupere, B., … Burger, P. A. (2017). Combined hybridization capture and shotgun sequencing for

ancient DNA analysis of extinct wild and domestic dromedary camel. *Molecular Ecology Resources*, *17*(2), 300–313. https://doi.org/10.1111/1755-0998.12551

Munagala, R., Aqil, F., Jeyabalan, J., & Gupta, R. C. (2016). Bovine milk-derived exosomes for drug delivery. *Cancer Letters*, *371*(1), 48–61. https://doi.org/10.1016/j.canlet.2015.10.020

Nurseitova, M., Konuspayeva, G., & Jurjanz, S. (2014). Comparison of dairy performances between dromedaries, bactrian and crossbred camels in the conditions of South Kazakhstan. *Emirates Journal of Food and Agriculture*, *26*(4), 366–370. https://doi.org/10.9755/ejfa.v26i4.17271

Ochirkhuyag, B., Chobert, J. M., Dalgalarrondo, M., Choiset, Y., & Haertlé, T. (1997). Characterization of caseins from Mongolian yak, khainak, and bactrian camel\n. *Le Lait*, *77*(5), 601–613. https://doi.org/10.1051/lait:1997543

Park, Y. W., Juarez, M., Ramos, M., & Haenlein, G. F. W. (2007). Physico-chemical characteristics of goat and sheep milk. *Goat and Sheep Milk*, *68*(1–2), 88–113. https://doi.org/DOI: 10.1016/j.smallrumres.2006.09.013

Pauciullo, A., & Erhardt, G. (2015). Molecular characterization of the llamas (Lama glama) casein cluster genes transcripts (CSN1S1, CSN2, CSN1S2, CSN3) and regulatory regions. *PLoS ONE*, *10*(4). https://doi.org/10.1371/journal.pone.0124963

Pauciullo, A., Giambra, I. J., Iannuzzi, L., & Erhardt, G. (2014). The β-casein in camels: Molecular characterization of the CSN2 gene, promoter analysis and genetic variability. *Gene*, *547*(1), 159–168. https://doi.org/10.1016/j.gene.2014.06.055

Pauciullo, A., Shuiep, E. S., Cosenza, G., Ramunno, L., & Erhardt, G. (2013). Molecular characterization and genetic variability at κ-casein gene (CSN3) in camels. *Gene*, *513*(1), 22–30. https://doi.org/10.1016/j.gene.2012.10.083

Pisanu, S., Ghisaura, S., Pagnozzi, D., Biosa, G., Tanca, A., Roggio, T., … Addis, M. F. (2011). The sheep milk fat globule membrane proteome. *Journal of Proteomics*. https://doi.org/10.1016/j.jprot.2010.11.011

Raposo, G. (1996). B lymphocytes secrete antigen-presenting vesicles. *Journal of*

Experimental Medicine, *183*(3), 1161–1172. https://doi.org/10.1084/jem.183.3.1161

Raposo, G., & Stoorvogel, W. (2013). Extracellular vesicles: Exosomes, microvesicles, and friends. *Journal of Cell Biology*. https://doi.org/10.1083/jcb.201211138

Reed, R., & Maniatis, T. (1986). A role for exon sequences and splice-site proximity in splice-site selection. *Cell*, *46*(5), 681–690. https://doi.org/10.1016/0092-8674(86)90343-0

Reinhardt, T. A., Lippolis, J. D., Nonnecke, B. J., & Sacco, R. E. (2012). Bovine milk exosome proteome. *Journal of Proteomics*, *75*(5), 1486–1492. https://doi.org/10.1016/j.jprot.2011.11.017

Rijnkels, M. (2002). Multispecies comparison of the casein gene loci and evolution of casein gene family. *Journal of Mammary Gland Biology and Neoplasia*. https://doi.org/10.1023/A:1022808918013

Saadaoui, B., Henry, C., Khorchani, T., Mars, M., Martin, P., & Cebo, C. (2013). Proteomics of the milk fat globule membrane from Camelus dromedarius. *Proteomics*, *13*(7), 1180–1184. https://doi.org/10.1002/pmic.201200113

Salmen, S. H., Abu-Tarboush, H. M., Al-Saleh, A. A., & Metwalli, A. A. (2012). Amino acids content and electrophoretic profile of camel milk casein from different camel breeds in Saudi Arabia. *Saudi Journal of Biological Sciences*, *19*(2), 177–183. https://doi.org/10.1016/j.sjbs.2011.12.002

Semo, E., Kesselman, E., Danino, D., & Livney, Y. D. (2007). Casein micelle as a natural nano-capsular vehicle for nutraceuticals. *Food Hydrocolloids*, *21*(5–6), 936–942. https://doi.org/10.1016/j.foodhyd.2006.09.006

Sharma, P., Dube, D., Singh, A., Mishra, B., Singh, N., Sinha, M., … Singh, T. P. (2011). Structural basis of recognition of pathogen-associated molecular patterns and inhibition of proinflammatory cytokines by camel peptidoglycan recognition protein. *Journal of Biological Chemistry*, *286*(18), 16208–16217. https://doi.org/10.1074/jbc.M111.228163

Shuiep, E. T. S., Giambra, I. J., El Zubeir, I. E. Y. M., & Erhardt, G. (2013). Biochemical and molecular characterization of polymorphisms of αs1-casein in Sudanese camel (Camelus

dromedarius) milk. *International Dairy Journal*, *28*(2), 88–93. https://doi.org/10.1016/j.idairyj.2012.09.002

Simons, M., & Raposo, G. (2009). Exosomes--vesicular carriers for intercellular communication. *Current Opinion in Cell Biology*, *21*, 575–581. https://doi.org/10.1016/j.ceb.2009.03.007

Smolenski, G., Haines, S., Kwan, F. Y. S., Bond, J., Farr, V., Davis, S. R., … Wheeler, T. T. (2007). Characterisation of host defence proteins in milk using a proteomic approach. *Journal of Proteome Research*. https://doi.org/10.1021/pr0603405

Svanborg, C., Ågerstam, H., Aronson, A., Bjerkvig, R., Düringer, C., Fischer, W., … Svensson, M. (2003). HAMLET kills tumor cells by an apoptosis-like mechanism - Cellular, molecular, and therapeutic aspects. *Advances in Cancer Research*. https://doi.org/10.1016/S0065-230X(03)88302-1

Threadgill, D. W., & Womack, J. E. (1990). Genomic analysis of the major bovine milk protein genes. *Nucleic Acids Research*, *18*(23), 6935–42. https://doi.org/10.1093/nar/18.23.6935

Tkach, M., & Théry, C. (2016). Communication by Extracellular Vesicles: Where We Are and Where We Need to Go. *Cell*. https://doi.org/10.1016/j.cell.2016.01.043

Tulgat, R., & Schaller, G. B. (1992). Status and distribution of wild Bactrian camels Camelus bactrianus ferus. *Biological Conservation*, *62*(1), 11–19. https://doi.org/10.1016/0006-3207(92)91147-K

Tydell, C., Yount, N., Tran, D., Yuan, J., & Selsted, M. E. (2002). Isolation, characterization, and antimicrobial properties of bovine oligosaccharide-binding protein. A microbicidal granule protein of eosinophils and neutrophils. *Journal of Biological Chemistry*, *277*(22), 19658–19664. https://doi.org/10.1074/jbc.M200659200

Uversky, V. N., El-Fakharany, E. M., Abu-Serie, M. M., Almehdar, H. A., & Redwan, E. M. (2017). Divergent Anticancer Activity of Free and Formulated Camel Milk α-Lactalbumin. *Cancer Investigation*, *35*(9), 610–623. https://doi.org/10.1080/07357907.2017.1373783

van der Pol, E., Boing, A. N., Harrison, P., Sturk, A., & Nieuwland, R. (2012). Classification, Functions, and Clinical Relevance of Extracellular Vesicles. *Pharmacological Reviews*, *64*(3), 676–705. https://doi.org/10.1124/pr.112.005983

van Herwijnen, M. J. C., Zonneveld, M. I., Goerdayal, S., Nolte – 't Hoen, E. N. M., Garssen, J., Stahl, B., … Wauben, M. H. M. (2016). Comprehensive Proteomic Analysis of Human Milk-derived Extracellular Vesicles Unveils a Novel Functional Proteome Distinct from Other Milk Components. *Molecular & Cellular Proteomics*. https://doi.org/10.1074/mcp.M116.060426

Walther, B., & Sieber, R. (2011). Bioactive proteins and peptides in foods. *International Journal for Vitamin and Nutrition Research*, *81*(2–3), 181–192. https://doi.org/10.1024/0300-9831/a000054

Wangoh, J., Farah, Z., & Puhan, Z. (1998). Iso-electric focusing of camel milk proteins. *International Dairy Journal*, *8*(7), 617–621. https://doi.org/10.1016/S0958-6946(98)00092-2

Wernery, U. (2006). Camel milk, the white gold of the desert. *Journal of Camel Practice and Research*.

Yang, Y., Bu, D., Zhao, X., Sun, P., Wang, J., & Zhou, L. (2013). Proteomic analysis of cow, yak, buffalo, goat and camel milk whey proteins: Quantitative differential expression patterns. *Journal of Proteome Research*, *12*(4), 1660–1667. https://doi.org/10.1021/pr301001m

Yang, Y., Zheng, N., Zhao, X., Zhang, Y., Han, R., Ma, L., … Wang, J. (2015). Data from proteomic characterization and comparison of mammalian milk fat globule proteomes by iTRAQ analysis. *Data in Brief*. https://doi.org/10.1016/j.dib.2014.12.004

Yassin, A. M., Abdel Hamid, M. I., Farid, O. A., Amer, H., & Warda, M. (2016). Dromedary milk exosomes as mammary transcriptome nano-vehicle: Their isolation, vesicular and phospholipidomic characterizations. *Journal of Advanced Research*, *7*(5), 749–756. https://doi.org/10.1016/j.jare.2015.10.003

Zhuang, Y., & Weiner, A. M. (1989). A compensatory base change in human U2 snRNA can suppress a branch site mutation. *Genes & Development*, *3*(10), 1545–1552. https://doi.org/10.1101/gad.3.10.1545

Chapter 2

Combining different proteomic approaches to resolve complexity of the milk protein fraction of dromedary, Bactrian camels and hybrids, from different regions of Kazakhstan

Alma Ryskaliyeva[1], Céline Henry[2], Guy Miranda[1], Bernard Faye[3], Gaukhar Konuspayeva[4] and Patrice Martin[1]

[1]INRA, UMR GABI, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France
[2]INRA, MICALIS Institute, Plateforme d'Analyse Protéomique Paris Sud-Ouest (PAPPSO), Université Paris-Saclay, 78350 Jouy-en-Josas, France
[3]CIRAD, UMR SELMET, 34398 Montpellier, France
[4]Al-Farabi Kazakh State University, Biotechnology department, 050040 Almaty, Kazakhstan

PLOS ONE 13(5):e0197026. (2018) https://doi.org/10.1371/journal.pone.0197026

Abstract

Nutritional suitability of milk is not only related to gross composition, but is also strongly affected by the microheterogeniety of the protein fraction. Hence, to go further into the evaluation of the potential suitability of non-bovine milks in human/infant nutrition it is necessary to have a detailed characterization of their protein components. Combining proven proteomic approaches (SDS-PAGE, LC-MS/MS and LC-ESI-MS) and cDNA sequencing, we provide here in depth characterization of the milk protein fraction of dromedary and Bactrian camels, and their hybrids, from different regions of Kazakhstan. A total 391 functional groups of proteins were identified from 8 camel milk samples. A detailed characterization of 50 protein molecules, relating to genetic variants and isoforms arising from post-translational modifications and alternative splicing events, belonging to nine protein families (κ-, α_{s1}-, α_{s2}-, β-; and γ-CN, WAP, α-LAC, PGRP, CSA/LPO) was achieved by LC-ESI-MS. The presence of two unknown proteins UP1 (22,939 Da) and UP2 (23,046 Da) was also reported as well as the existence of a β-CN short isoform (946 Da lighter than the full-length β-CN), arising very likely in both genetic variants (A and B) from proteolysis by plasmin. In addition, we report, for the first time to our knowledge, the occurrence of a α_{s2}-CN phosphorylation isoform with 12P groups within two recognition motifs, suggesting thereby the existence of two kinase systems involved in the phosphorylation of caseins in the mammary gland. Finally, we demonstrate that genetic variants, which hitherto seemed to be species- specific (*e.g.* β-CN A for Bactrian and β-CN B for dromedary), are in fact present both in *Camelus dromedarius* and *C. bactrianus*.

Key words: *Camelus dromedarius, Camelus bactrianus,* hybrids, milk, casein, whey proteins, post-translational modifications, splicing, genetic polymorphism, phosphorylation, proteomics

2.1 Introduction

According to the most recent statistics, the world camel population is estimated to be about 29 million (FAO, 2017). *Camelus dromedarius* is the most frequent and widespread domestic camel species composing 90% of the total camel population (Mohandesan et al., 2017). Camels have been domesticated in a number of arid regions, including Northern and Eastern Africa, the Arabian Peninsula and Central and South West Asia. *Camelus bactrianus* forms numerical inferiority, mostly inhabits in Mongolia, China, and Central Asia. Alternatively, there are also crossed camels (hybrids) which are found mainly in Russia, Iran, Turkmenistan, and in Kazakhstan.

Kazakhstan is a specific region where both domesticated species (*C. dromedarius* and *C. bactrianus*) are maintained in mixed herds (Nurseitova et al., 2014). There are about 35,000 camel heads reared in this country for milk production (FAO, 2017). Camel milk is consumed as fresh milk and as a traditional fermented drink called *shubat*, which is very popular in Central Asia countries. Besides nutritional qualities, camel fresh and fermented milk have been reported to display potential health-promoting properties (Agrawal et al., 2003; Al-Ayadhi & Elamin, 2013; El-Fakharany et al., 2017; Korashy et al., 2012; Manaer et al., 2015; Sboui et al., 2010) which depend very heavily on its unique protein content.

Advanced improvement in proteomic techniques allow nowadays obtaining a precise image of the protein fraction of milk. Recently, proteomic approaches, based on mass spectrometry (Alhaider et al., 2013) and isobaric tag for relative and absolute quantification (Yang et al., 2013), have been used to analyze the proteome of dromedary camel milk and Bactrian camel milk whey, respectively. These techniques were useful to gain knowledge on the detection, quantification and characterization of camel milk proteins. These studies confirm that camel milk is a rich source of biologically active proteins and peptides (Hsieh et al., 2015; Mati et al., 2017).

Whey proteins which were reported to display a wide range of bioactivities (Davoodi et al., 2016), including immuno-modulating (Legrand et al., 2004), anti-carcinogenic (Habib et al., 2013), antibacterial, and antifungal activities (Kanwar et al., 2015), account for 20% of total camel milk proteins. Pattern-recognition proteins, such as the peptidoglycan recognition protein (PGRP), an intracellular component of neutrophils, modulate anti-inflammatory reaction of the immune response (Kappeler et al., 2004). LTF interacts with

lipopolysaccharides of Gram-negative bacteria whereas lysozyme C binds and hydrolyzes peptidoglycans, preferably of Gram-positive bacteria, but with a lower affinity than PGRP (Sharma et al., 2011). Present at a very low level in ruminant milks (Tydell et al., 2002), PGRP has been detected in mammary secretions of porcine and camel (Kappeler et al., 2004) and was shown to participate in granule-mediated killing of gram-positive and negative bacteria (Dziarski et al., 2012). Proteose peptone component 3 (PP3 or Lactophorin or GlyCAM1) plays an important immunological role in the lactating camel, to prevent the occurrence of mastitis, or for its newborn by inhibiting pathogen multiplication in the respiratory and gastrointestinal tracts of the suckling young (Girardet et al., 2000). Likewise, camel milk contains the whey acidic protein (WAP), also found in rodents and lagomorphs (Hennighausen & Sippel, 1982). The biological function of this protein is unknown. However, proteins such as elafin and antileukoproteinase 1, containing WAP domains, are known to function as protease inhibitor involved in the immune defense of multiple epithelia and has been identified as candidate molecular markers for several cancers (Bouchard et al., 2006).

As in cow milk, *ca*. 80% of the total protein fraction of camel milk are represented by caseins (CN) that are synthesized under multi-hormonal control in the mammary gland of mammals. Associated with amorphous calcium phosphate nanoclusters they form large and stable colloidal aggregates, the so-called CN micelles, which figure as calcium-transport vehicles. These CN micelles provide neonates with calcium at a very high concentration, which is achieved during their packaging in the secretion pathway (McMahon & Oommen, 2013). Recently it was reported that α_{s1}- and α_{s2}-CN display molecular chaperone-like activity inhibiting CN aggregation and triggering micelle structure (Sakono et al., 2011).

However, there is no comprehensive investigation on milk protein variations and variability in composition between individual camels. In addition, proteomic studies did not consider the molecular diversity of each type of protein, arising from genetic polymorphisms (mutations), defects in the processing of primary transcripts and post-translational modifications (PTM) such as phosphorylation, factors that significantly have a pronounced impact on protein structure, and finally on milk properties. Milk protein polymorphism is a unique biological paradigm that could help to understand CN intracellular transport, micelle formation and organization, biodiversity and evolution (Martin et al., 2013), the release of bioactive peptides with implications in human health (Balteanu et al., 2013).

Therefore, to gain an insight into the molecular diversity of camel milk proteins, we design a comprehensive strategy combining classical (SDS-PAGE) and advanced proteomic approaches (LC-MS/MS, LC-ESI-MS), as well as cDNA sequencing. Here we report a complete profiling of the milk protein fraction of Bactrian and dromedary camels from Kazakhstan, including a detailed characterization of camel CN and WPs including variants related to genetic polymorphisms, splicing defects, phosphorylation levels. In addition, we introduce a reference point for further investigation in camel milk protein polymorphism.

2.2 Materials and Methods

2.2.1 Ethics statements

All animal studies were carried out in compliance with European Community regulations on animal experimentation (European Communities Council Directive 86/609/EEC) and with the authorization of the Kazakh Ministry of Agriculture. Milk sampling was performed in appropriate conditions supervised by a veterinary accredited by the French Ethics National Committee for Experimentation on Living Animals. No endangered or protected animal species were involved in this study. No specific permissions or approvals were required for this study with the exception of the rules of afore-mentioned European Community regulations on animal experimentation, which were strictly followed.

2.2.2 Milk samples collection and preparation

In total 181 raw milk samples (Table 2.1.) were collected during morning milking on healthy dairy camels belonging to two camel species: *C. bactrianus* (n=72) and *C. dromedarius* (n=65), and their hybrids (n=42), at different lactation stages, ranging between 30 and 90 days postpartum. Bactrian camels were originating from Kazakh type whereas dromedary camels were from Turkmen Arvana breed. Unfortunately, the information about the nature and the level of hybridization of hybrids was not available. All species are well adapted to the local environment of Kazakhstan.

Camels grazed on four various natural pastures with the distance more than 3,500 kms between the regions at extreme points of Kazakhstan: Almaty (AL) at the foot of Tien Shan Mountain, Shymkent (SH) along deserts Kyzylkum and Betpak-Dala, Kyzylorda (KZ) on the edge of the steppe, and Atyrau (ZKO) at the mouth of the Caspian Sea (Figure 2.1). Whole-

milk samples were centrifuged at 2,500 *g* for 20 min at 4°C (Allegra X-15R, Beckman Coulter, France) to separating fat from skimmed milk. Samples were quickly frozen and stored at -80°C (fat) and -20°C (skimmed milk) until analysis.

Table 2.1. Camel milk samples collected (n = 181) in the 3 species of the 4 regions of Kazakhstan

ID	Region	Coding	Bactrian (B)	Dromedary (D)	Hybrid (H)	Total number of camels for each region
1	Almaty	AL B/D/H	13	20	1	34
2	Shymkent	SH B/D/H	20	21	20	61
3	Kyzylorda	KZ B/D/H	18	16	20	54
4	Atyrau	ZKO B/D/H	21	8	3	32

Figure 2.1. Geographical location of camel milk sampling

2.2.3 Selection of milk samples for analysis

Of the 181 milk samples collected, 63, including *C. bactrianus* (n=19), *C. dromedarius* (n=20), and hybrids (n=24) from four different regions of Kazakhstan were selected for SDS-PAGE analysis (Figure 2.2). Each Bactrian and dromedary camel group formed by 5 animals, except Bactrians of Atyrau regions (n=4). For hybrids, there were 4 groups comprising 10 animals (Kyzylorda and Shymkent regions), whereas there were only 1 and 3 animals for Almaty and Atyrau regions, respectively. This selection was based on lactation stages and number of parities (from 2 to 14) of each camel group composed by the species and grazing regions. It should be emphasized that data available on animals: breed, age, lactation stage and calving number, were estimated by a local veterinarian, since no registration of camels in farms is maintained. Due to the lack of sufficient information, dromedary milk samples (n=5) from Almaty region were excluded from subsequent analyses. Then, 8 of the 58 remaining milk samples from three different regions (*C. bactrianus*, n=3, *C. dromedarius*, n=3, and hybrids, n=2) exhibiting the most representative SDS-PAGE patterns were analyzed by LC-MS/MS after a tryptic digestion of excised gel bands. Additionally, 30 milk samples (*C. bactrianus*, n=10; *C. dromedarius*, n=10; hybrids, n=10), taken from the 63 milks analyzed by SDS-PAGE, were analyzed by LC-ESI-MS (Bruker Daltonics).

2.2.4 Coomassie blue (Bradford) protein assay

To estimate the concentration of total protein in a milk sample the Coomassie Blue Protein Assay was used (Bradford, 1976). Absorbance at 590 nm was measured using the UV-Vis spectrophotometer (UVmini-1240, Shimadzu). The reference standard curve was done with commercial bovine serum albumin (BSA) powder dissolved in MilliQ water and diluted to a concentration of 1 mg/mL. Series of dilutions (0.1, 0.2, 0.4, 0.6, and 0.8 µg/µL) were prepared from the stock solution, in duplicate to ensure the protein concentration is within the range of the assay.

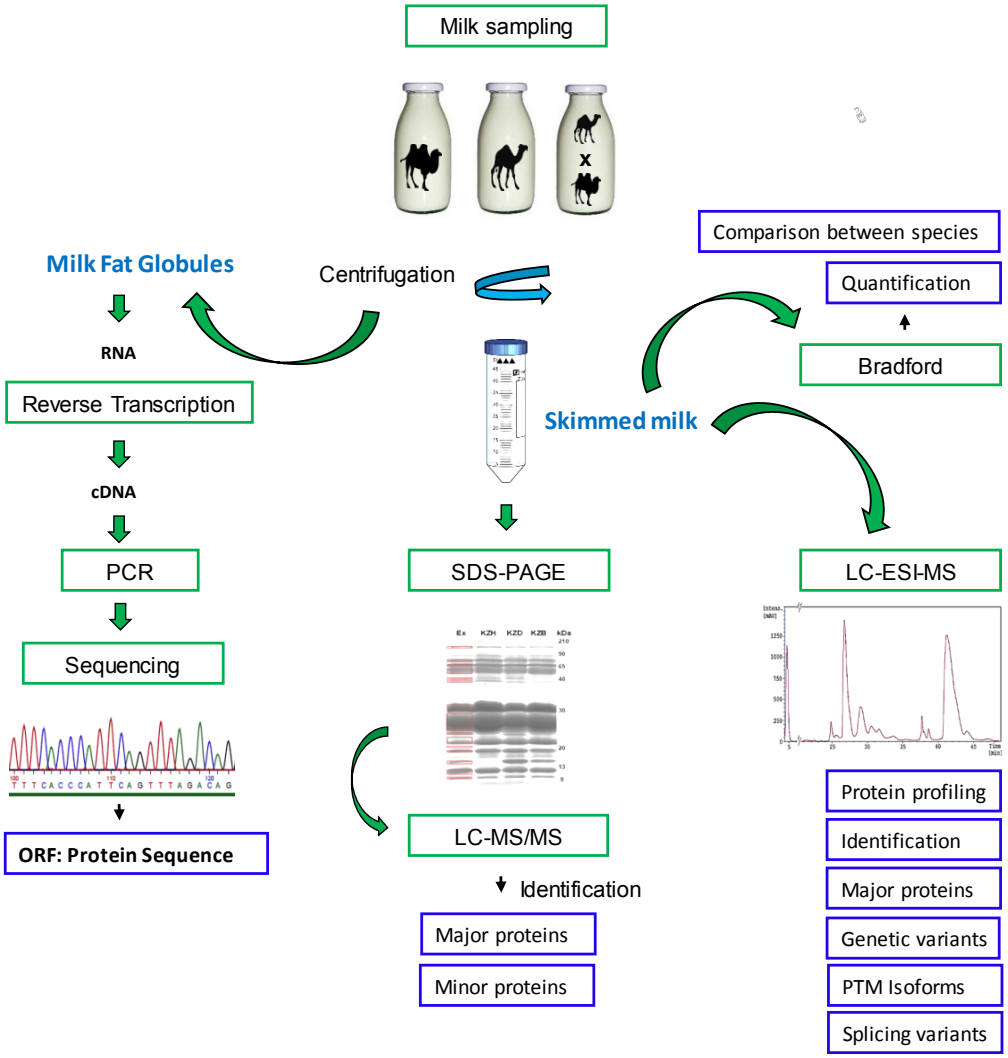


Figure 2.2. Diagram of the experimental scheme designed for quantification and identification of camel milk proteins

2.2.5 1D sodium dodecyl sulfate polyacrylamide gel electrophoresis (SDS-PAGE)

Both major and low-abundant proteins resolved by SDS-PAGE were identified after excision by mass analysis of the tryptic hydrolysate. The method used in the study was based on that from Laemmli (Laemmli, 1970). Twenty-five micrograms of each individual skimmed milk sample were loaded into 12.5% acrylamide resolving gel and subjected to electrophoresis. Samples were prepared with Laemmli Lysis-Buffer (Sigma-Aldrich). Separations were performed in a vertical electrophoresis apparatus (Bio-Rad, Marnes-la-

Coquette, France). After GelCode Blue Safe Protein staining and gel scanning using Image Scanner iii (Epson Expression™ 10,000 XL, Sweden), resolved bands were excised from the gel and submitted to digestion by trypsin. Thereafter, tryptic peptides were analyzed by LC-MS/MS.

2.2.6 Identification of proteins by LC-MS/MS analysis

In order to identify the main protein contained in each electrophoretic band, mono dimensional electrophoresis (1D SDS-PAGE) followed by trypsin digestion and by LC-MS/MS analysis, was used essentially as described (Saadaoui et al., 2014). Briefly, after a 10 cm migration of samples in such an 1D SDS-PAGE, the 16 main electrophoretic bands (1.5 mm^3) were cut on each gel lane, transferred into 96-well microtiter plates (FrameStar, 4titude, 0750/Las). Reduction of disulfide bridges of proteins was carried out by incubating at 37°C for one hour with dithiothreitol (DTT, 10 mM, Sigma), meanwhile the alkylation of free cysteinyl residues with iodoacetamide (IAM, 50 mM, Sigma) at RT for 45 min in total obscurity. After gel pieces were washed twice, first, with 100 µL 50% ACN/50 mM NH_4HCO_3 and then with 50 µL ACN, they were finally dried. The hydration was performed at 37°C overnight using digestion buffer 400 ng lys-C protease + trypsin. Hereby, peptides were extracted with 50% ACN/0.5% TFA and then with 100% ACN. Peptide solutions were dried in a concentrator and finally dissolved into 70 µL 2% ACN in 0.08% TFA. The identification of peptides was obtained using UltiMate™ 3000 RSLCnano System (Thermo Fisher Scientific) coupled either to LTQ Orbitrap XL™ Discovery mass spectrometer or QExactive (Thermo Fischer Scientific). Four µL of each sample was injected with flow of 20 µL/min on a precolumn cartridge (stationary phase: C18 PepMap 100, 5 µm; column: 300 µm x 5 mm) and desalted with a loading buffer 2% ACN and 0.08% TFA. After 4 min, the precolumn cartridge was connected to the separating RSLC PepMap C18 column (stationary phase: RSLC PepMap 100, 2 µm; column: 75 µm x 150 mm). Elution buffers were A: 2% ACN in 0.1% formic acid (HCOOH) and B: 80% ACN in 0.1% HCOOH. The peptide separation was achieved with a linear gradient from 0 to 35% B for 34 min at 300 nL/min. One run took 42 min, including the regeneration and the equilibration steps at 98% B.

Peptide ions were analyzed using Xcalibur 2.1 with the following machine set up in CID mode: 1) full MS scan in Orbitrap with a resolution of 15 000 (scan range [m/z] = 300-1600) and 2) top 8 in MS/MS using CID (35% collision energy) in Ion Trap. Analyzed

charge states were set to 2-3, the dynamic exclusion to 30 s and the intensity threshold was fixed at 5.0 x 10^2.

Raw data were converted to mzXML by MS convert (ProteoWizard version 3.0.4601). UniProtKB Cetartiodactyla database was used (157,113 protein entries, version 2015), in conjunction with contaminant databases were searched by algorithm X!TandemPiledriver (version 2015.04.01.1) with the software X!TandemPipeline (version 3.4) developed by the PAPPSO platform (http://pappso.inra.fr/bioinfo/). The protein identification was run with a precursor mass tolerance of 10 ppm and a fragment mass tolerance of 0.5 Da. Enzymatic cleavage rules were set to trypsin digestion ("after R and K, unless P follows directly after") and no semi-enzymatic cleavage rules were allowed. The fix modification was set to cysteine carbamido methylation and methionine oxidation was considered as a potential modification. Results were filtered using inbuilt X!TandemParser with peptide *E*-value of 0.05, a protein *E*-value of -2.6, and a minimum of two peptides.

2.2.7 LC-ESI-MS

Fractionation of camel milk proteins and determination of their molecular masses, performed by coupling RP-HPLC to ESI-MS (micrOTOF^{TM} II focus ESI-TOF mass spectrometer; Bruker Daltonics), were essentially as described (Saadaoui et al., 2014). In total 20 µL of skimmed milk samples were first clarified by the addition of 230 µL of clarification solution 0.1 M bis-Tris buffer pH 8.0, containing 8 M urea, 1.3% trisodium citrate, and 0.3% DTT. Clarified milk samples (25 µL) were directly injected onto a Biodiscovery C5 reverse phase column (300 Å pore size, 3 µm, 150 x 2.1 mm; Supelco, France). The mobile phase of the column corresponded to a gradient mixture of Solvent A (H_2O/TFA 100:0.25, v/v) and Solvent B (ACN/TFA 100:0.20, v/v). Elution was achieved using a linear gradient from 5% to 27% B in 20 min, from 27% to 33% B in 0.1 min, from 33% to 34% B in 11.1 min, from 34% to 40% B in 0.1 min, from 40% to 41% B in 14.9 min, and from 41% to 90% B in 0.1 min. This gradient elution was followed by an isocratic elution at 90% B for 4.9 min, and a linear return to 5% B in 0.1 min. The temperature of the column was adjusted to 52^{o}C and the flow rate to 0.2 mL/min. Eluted peaks were detected by UV-absorbance at 214 nm. The liquid effluent was introduced to the mass spectrometer. Positive ion mode was used, and mass scans were acquired over a mass-to-charge ratio (m/z) ranging between 600 and 3000 Da.

The LC/MS system was controlled by the HyStar software (Bruker Daltonics). Peak profiles from UV 214 nm and Extracted Ion Chromatograms (EIC), multicharged ion spectra, deconvoluted spectra and determination of masses were obtained with DataAnalysis Version 4.0 SP1 software (Bruker Daltonics).

2.2.8 Milk fat globule collection and RNA extraction

Total RNA was extracted from MFG fraction stored at -80°C using Trizol (Invitrogen) following the protocol from the manufacturer as described by Brenaut et al., (2012).

2.2.9 First-strand cDNA synthesis and PCR amplification

First-strand cDNA was synthesized from 5 to 10 ng of total RNA primed with oligo(dT)20 and random primers (3:1, vol/vol) using Superscript III reverse transcriptase (Invitrogen Life Technologies Inc., Carlsbad, CA) according to the manufacturer's instructions. One microliter of 2 U/µL RNase H (Invitrogen Life Technologies) was then added and the reaction mix was incubated for 20 min at 37°C to remove RNA from heteroduplexes. Single-strand cDNA thus obtained was stored at -20°C. cDNA samples covering the entire coding regions of caseins were amplified. PCR was performed in an automated thermocycler GeneAmp® PCR System 2,400 (Perkin-Elmer, Norwalk, USA) with GoTaq® G2 Flexi DNA Polymerase Kit (Promega Corporation, USA). Reactions were carried out with 0.2 mL thin-walled PCR tubes with flat cap strips (Thermo Scientific, UK), in 50 µL volumes containing 5X Green or Colorless GoTag® Flexi Buffer, $MgCl_2$ Solution 25 mM, PCR Nucleotide Mix 10 mM each, GoTag® G2 Flexi DNA Polymerase (5 U/µL), 10 mM each oligonucleotide primer, template DNA and nuclease-free water, up to the final volume. Primer pairs, purchased from Eurofins (Eurofins genomics, Germany), were designed using published *Camelus* nucleic acid sequence (NCBI, NM_001303566.1). Sequencing of PCR fragments was performed with primer pairs used for PCR and sequenced from both strands, according to the Sanger method by Eurofins.

2.3 Results

2.3.1 Total protein content

Using the Bradford assay for estimating the protein concentration in milk samples, we observed that the highest protein concentration occurred with Bactrian camel milk samples, but the difference was slight comparing with crossed camel species. The total protein value in raw camel milk from Shymkent region was estimated to be *ca.* 33 g/L (33.15 ± 6.64 g/L) for *C. bactrianus* (n=5), and 31 g/L (30.83 ± 5.82 g/L) for *C. dromedarius* (n=7), whereas hybrids (n=9) displayed an intermediate value 31.5 g/L (31.43 ± 4.56 g/L). On average, Bactrian milk was considered to have a higher total protein content than that of Dromedary (Zhao et al., 2015) and hybrid milks. Our results are in agreement with data reported previously by Konuspayeva et al., (2009). No significant differences were found across species from different geographical locations.

2.3.2 Identification of main milk proteins from 1D SDS-PAGE by LC-MS/MS

After first adjusting protein concentrations at the same value, 63 individual camel milk samples were separated onto SDS-PAGE. The comparative analysis of whole milk samples by SDS-PAGE displayed rather similar electrophoretic profiles with related migration characteristics and the same apparent molecular weights between individual milk samples of different species and regions. A typical gel pattern from which proteins were identified in individual *C. bactrianus*, *C. dromedarius* and hybrid milk samples of Kyzylorda region is shown in Figure 2.3.

Sixteen main bands relatively well-resolved were excised from the electrophoretic pattern. The most intense band observed around 26 kDa was identified as β-CN. Quantitative analyses on camel milk proteins carried out before have demonstrated significantly higher amounts of β-CN compared to the homologous bovine CN (Kappeler et al., 2003). The most representative other bands were characterized as being: WAP (12.5 kDa), α-LAC (14.3 kDa), GlyCAM 1 (15.4 kDa and 17.2 kDa), κ-CN (20.3 kDa), PGRP (21.3 kDa), α_{s2}-CN (22.9 kDa), α_{s1}-CN (25.7 kDa), neutrophil gelatinase (28.3 kDa), lipoprotein lipase (46.5 kDa), perilipin-2 (47.2 kDa), butyrophilin (51.0 kDa), amine oxidase (55.3 kDa), lactadherin (56.2 kDa), heat

shock protein (70.0 kDa), LTF (77.1 kDa), lactoperoxidase (87.7 kDa), and xanthine oxidase (150 kDa). Masses mentioned above correspond to theoretical masses of proteins identified on the basis of tryptic profiles after LC-MS/MS analysis. Globally, the electrophoretic patterns of Kazakh camel milk samples agree with those reported recently for Israelian and Tunisian camel milk samples (Felfoul et al., 2017; Merin et al., 2001). However, surprisingly the prominent fact was the apparent absence in Kazakh milk samples of camel serum albumin (CSA), the major WP with a molecular mass equal to 66.0 kDa in camel colostrum (Merin et al., 2001). By contrast, this protein has been successfully identified, with the best *E*-value, in Tunisian fresh milk samples (Felfoul et al., 2017).

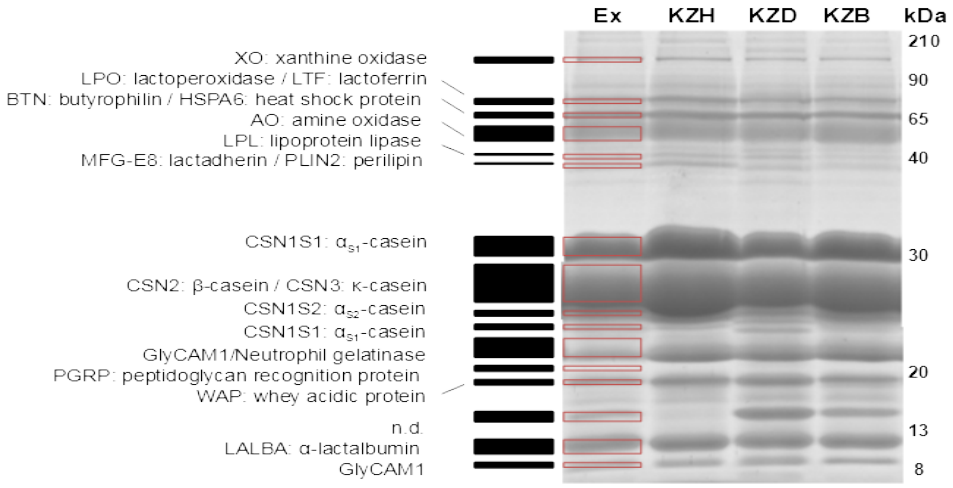


Figure 2.3. 1D SDS-PAGE pattern of *C. bactrianus* (KZB), *C. dromedarius* (KZD) and hybrid (KZH) skimmed milk samples of Kyzylorda (KZ) region. In the given example pattern (Ex), red frames and black boxes aligned correspond to electrophoretic bands that were excised from the gel and subsequently analyzed for protein identification, after tryptic digestion, by LC-MS/MS. Molecular weight markers from 210 to 8 kDa are indicated at the right of the gel.

2.3.3 Qualitative proteome of camel skimmed milk by LC-MS/MS

We took advantage of LC-MS/MS analysis to identify proteins in electrophoretic bands to go further into the description of the protein fraction of camel milk. Indeed, for each band analyzed by LC-MS/MS, between 10 and 70 different proteins were identified. In such a way, using UniprotKB taxonomy cetartiodactyla (SwissProt + Trembl) database, a total of 391 functional groups of proteins (proteins belonging to a same group share common

peptides) were identified after LC-MS/MS analysis of 8 camel milk samples (S1 Table). A set of 235 proteins was observed as common to the 8 milk samples. As example, a list of the first 70 common proteins found in milk samples of the three species from Shymkent region is given in Table 2.2.

Eight proteins were identified as authentically matching with proteins in *C. dromedarius* protein database, two with *C. bactrianus* protein database, 46 with *C. ferus* protein database, and the remaining (n=14) with the other mammalian species such as*, Lama guanicoe, Bos taurus, Sus scrofa* and *Ovis aries* protein databases. Immune-related proteins such as GlyCAM1, lactadherin (MFG-E8), and LTF, as well as milk fat globule membrane (MFGM)-enriched proteins such as xanthine oxidase (XO), butyrophilin (BTN), actin, ras-related protein Rab-18, ADP-ribosylation factor 1, tyrosine-protein kinase and GTP-binding protein SAR1b, were detected. Likewise, proteins originating from blood such as serpin A3-1, apolipoprotein A-1, α-1-antitrypsin like protein, α-1-acid glycoprotein, β-2-microglobulin, complement C3-like protein were found in all milk samples analyzed.

Table 2.2. Top 70 proteins identified by LC-MS/MS from individual *C. bactrianus* (B), *C. dromedarius* (D) and hybrid (H) milk samples of Shymkent region

ID	Accession	Description	Mr	(-) log *E*-value			Coverage, %			Number of Spectra*		
			kDa	B	D	H	B	D	H	B	D	H
1	O97943-2	Short isoform of Alpha-S1-casein *(C. dromedarius)*	25,7	453,63	535,3	585,79	88	92	92	574	679	842
2	A0A077SL35	Beta-casein *(C. bactrianus)*	26,1	299,05	285,89	395,72	77	79	79	542	507	745
3	O97944	Alpha-S2-casein *(C. dromedarius)*	22,9	285,16	320,65	276,39	70	70	66	357	451	376
4	W6GH05	Lactoferrin *(C. dromedarius)*	77,1	723,25	557,5	1174,16	85	80	88	253	157	829
5	P15522-2	Isoform B of Glycosylation-dependent cell adhesion molecule 1 *(C. dromedarius)*	15,4	112,39	189,45	192,31	64	70	69	220	359	358
6	L0P304	Kappa-casein *(C. bactrianus)*	20,3	127,64	149,66	194,03	51	53	51	217	222	329
7	S9WF76	Lactadherin-like protein *(C. ferus)*	45,6	263,03	350,07	501,73	51	53	57	162	236	436
8	P00710	Alpha-lactalbumin *(C. dromedarius)*	14,3	288,2	277,9	329,27	83	80	83	161	159	175
9	Q9GK12	Peptidoglycan recognition protein 1 *(C. dromedarius)*	21,3	308,94	359,49	336,8	73	73	79	112	156	140
10	S9Z0L8	Amine oxidase [flavin-containing] *(C. ferus)*	55,3	233,97	231,94	277,32	74	70	75	110	138	190
11	S9Y4T1	Xanthine dehydrogenase/oxidase *(C. ferus)*	150	270,07	435,86	330,26	40	45	41	79	138	121
12	S9X3X3	Butyrophilin subfamily 1 member A1 *(C. ferus)*	51,09	95,37	140,19	162,3	47	50	51	61	107	111
13	S9X4G0	Neutrophil gelatinase-associated lipocalin-like protein *(C. ferus)*	28,3	104,88	98,15	165,14	48	45	62	44	33	82

14	S9X1L5	Lipoprotein lipase isoform 3 (Fragment) *(C. ferus)*	46,5	107,66	133,72	54,73	43	55	27	43	56	29
15	S9YK74	Perilipin *(C. ferus)*	47,2	116,47	169,78	167,59	56	60	52	39	67	46
16	P09837	Whey acidic protein *(C. dromedarius)*	12,5	42,97	59,92	82,11	55	75	84	36	43	73
17	S9X4X6	Uncharacterized protein *(C. ferus)*	43,1	47,87	14	17,5	30	13	16	36	9	16
18	S9X7Q1	Lactoperoxidase isoform 1 preproprotein *(C. ferus)*	87,7	83,1	62,59	65,16	26	22	20	25	19	17
19	S9XDK9	Complement C3-like protein *(C. ferus)*	262,7	87,48	44,5	292,06	12	4	24	25	8	69
20	S9XR87	Beta-2-microglobulin *(C. ferus)*	14,8	36,87	36,15	40,76	43	43	43	23	28	44
21	O18831	Growth/differentiation factor 8 *(S. scrofa)*	42,7	44,62	49,29	67,44	28	24	32	23	22	17
22	P68103	Elongation factor 1-alpha 1 *(B. taurus)*	50	58,23	42,59	78,3	31	27	31	20	12	24
23	S9YCI6	Peptidyl-prolyl cis-trans isomerase *(C. ferus)*	23,8	44,54	43,1	71,24	54	54	57	20	21	22
24	S9XP75	Monocyte differentiation antigen CD14 *(C. ferus)*	29,7	49,75	44,58	73,91	27	27	27	17	16	22
25	S9WCV2	Sulfhydryl oxidase *(C. ferus)*	72,2	31,64	76,49	14,43	18	27	8	16	26	6
26	S9YC53	Alpha-1-antitrypsin-like protein *(C. ferus)*	51,9	64,76	64,84	51,84	28	31	32	14	21	15
27	S9YS49	Putative E3 ubiquitin-protein ligase Roquin *(C. ferus)*	158,5	56,65	90,28	8,84	10	14	2	14	29	3
28	T0NN97	Uncharacterized protein *(C. ferus)*	151,4	38,86	53,95	36,72	12	13	11	13	16	15
29	A0A0F6YEF6	Anti-HCV NS3/4A serine protease immoglobulin heavy chain (Fragment) *(C. dromedarius)*	13,4	38,24	35	52,95	24	24	24	12	10	26
30	S9X9X0	Vitelline membrane outer layer protein 1-like protein *(C. ferus)*	24,2	44,94	52,11	75,95	51	50	53	12	14	31
31	S9X358	Tissue alpha-L-fucosidase *(C. ferus)*	42,7	25,83	23,27	24,46	23	18	18	12	12	7
32	S9WS72	Sodium-dependent phosphate transport protein 2B-like protein *(C. ferus)*	75,4	24,99	15,41	24,7	18	7	9	11	8	8
33	S9X2V0	UDP-Gal:betaGlcNAc beta 1,4-galactosyltransferase 1, membrane-bound form-like protein *(C. ferus)*	37,3	30,18	13,09	38,99	23	11	33	11	7	15
34	S9XSQ6	Vitamin D-binding protein-like protein *(C. ferus)*	49,3	41,66	109,25	18,01	24	45	8	10	30	3
35	S9WGZ9	Rab GDP dissociation inhibitor beta isoform 1 *(C. ferus)*	35,7	27,75	22,32	55,78	32	25	43	10	8	15
36	P19120	Heat shock cognate 71 kDa protein *(B. taurus)*	71,1	37,14	60,92	111,21	17	24	35	9	16	30
37	S9XA25	Ezrin isoform 5-like protein (Fragment) *(C. ferus)*	71	24,32	28,23	47,46	13	10	20	9	7	13
38	S9XI30	Uncharacterized protein *(C. ferus)*	22	25,16	66,23	20,66	38	54	39	9	31	8
39	S9Y2X0	Platelet glycoprotein 4 *(C. ferus)*	38,3	28,14	40,27	55,15	15	17	29	9	8	17
40	G9F6X8	Protein disulfide-isomerase *(S. scrofa)*	56,3	20,09	20,62	21,76	17	15	9	9	10	3
41	S9WZP7	Serpin A3-8 *(C. ferus)*	74,9	15,26	28,06	72,32	11	12	20	8	12	28
42	A7MBJ4	Receptor-type tyrosine-protein phosphatase F *(B. taurus)*	211,1	23,69	21,19	20,54	6	6	4	8	8	8
43	S9YFG2	Complement factor D *(C. ferus)*	38,1	37,99	33,7	44,44	14	14	14	8	9	13

44	S9WPL9	Uncharacterized protein *(C. ferus)*	84,6	31,09	39,23	57,64	12	10	17	7	6	18
45	S9XE02	Beta-2-glycoprotein 1-like protein *(C. ferus)*	30,9	46,22	64,63	48,09	42	43	27	7	12	5
46	B8XH67	Na(+)/H(+) exchange regulatory cofactor NHE-RF *(S. scrofa)*	39,2	20,94	37,33	32,94	22	22	30	7	9	11
47	Q28452	Quinone oxidoreductase *(L. guanicoe)*	35,1	23,01	24,97	51,04	28	25	43	7	5	11
48	S9YU13	Glutathione S-transferase-like protein *(C. ferus)*	27,7	34,68	8,59	22	21	15	19	7	3	4
49	P00727-2	Isoform 2 of Cytosol aminopeptidase *(B. taurus)*	53,9	14,58	13,07	35,48	15	9	18	6	4	7
50	S9WUC8	Ig kappa chain V-II region RPMI 6410-like protein *(C. ferus)*	26,7	22,72	12,39	32,25	19	10	27	6	4	15
51	S9WRY0	L-lactate dehydrogenase B chain isoform 1-like protein *(C. ferus)*	30,2	25,51	28,95	37,66	22	22	22	6	7	6
52	S9X5V9	Fc of IgG binding protein (Fragment) *(C. ferus)*	254,9	20,45	65,48	10,4	3	9	2	6	16	4
53	B5B0D4	Major allergen beta-lactoglobulin *(B. taurus)*	19,8	9,94	54,67	12,3	21	59	25	6	27	7
54	S9XE13	Uncharacterized protein *(C. ferus)*	82,6	12,16	32,28	71,26	5	10	15	5	17	22
55	S9Y8C6	Phosphoglucomutase 1 isoform 3-like protein *(C. ferus)*	68,3	11,37	18,27	42,46	9	11	22	5	6	11
56	S9Y3S5	Olfactory receptor *(C. ferus)*	108,3	19,83	37,9	43,59	4	6	7	5	9	11
57	S9XT33	Lipopolysaccharide-binding protein *(C. ferus)*	47,5	11,05	9,43	16,89	12	10	17	5	4	7
58	S9YL21	Apolipoprotein A-I *(C. ferus)*	22,5	19,05	68,11	82,98	27	49	57	5	12	14
59	S9Y5X2	Cell death activator CIDE-A-like protein *(C. ferus)*	41,2	16,09	22,41	15,11	11	14	10	4	11	4
60	S9XC74	Osteopontin isoform OPN-c *(C. ferus)*	34,6	8,45	12,77	34,67	8	11	23	4	10	24
61	T0NLV9	Epoxide hydrolase 1 *(C. ferus)*	54,3	7,1	11,53	12,54	9	14	10	4	6	4
62	S9WKD1	Ribonuclease 4 *(C. ferus)*	26,8	11,56	25,5	29,63	12	22	23	3	7	9
63	S9WDV3	Fibrinogen gamma chain isoform gamma-B *(C. ferus)*	50,5	7,22	33,1	76,88	6	49	37	3	12	33
64	S9XLJ3	Brain-specific serine protease 4-like protein *(C. ferus)*	44,6	7,52	8,26	9,3	10	10	12	3	3	4
65	S9WY98	Sodium/glucose cotransporter 1 *(C. ferus)*	78,3	9,01	5,72	14,27	4	3	5	3	2	5
66	S9WX48	Alpha-1-acid glycoprotein *(C. ferus)*	22,9	6,65	7,25	29,92	14	8	44	3	2	17
67	W5P9V5	Uncharacterized protein *(O. aries)*	85,3	7,14	7,74	8,54	3	3	3	3	3	3
68	T0NLF0	Vitronectin *(C. ferus)*	56,2	11,18	7,8	25,19	7	4	10	3	2	4
69	Q0IIG8	Ras-related protein Rab-18 *(B. taurus)*	22,9	3,49	15,33	21,19	10	28	24	2	5	4
70	S9YNY9	Nucleobindin-1 *(C. ferus)*	53	4,54	48,27	20,28	5	41	21	2	16	9

Molecular masses (M_r) of proteins are expressed in kDa, E-value in log, coverage in %. Spectra indicates the number of spectra permitting the identification of proteins. Major proteins identified in excised gel bands after SDS-PAGE are given in bold type.

*abundance of each protein was estimated from spectral count. The number of spectra of *C. bactrianus* (B) classified the table.

2.3.4 Camel milk protein profiling by LC-ESI-MS

Thirty individual milk samples, including *C. bactrianus* (n=10), *C. dromedarius* (n=10), and hybrids (n=10) taken from the 58 milk samples analyzed in SDS-PAGE were submitted to LC-ESI-MS analysis. Milk proteins separated by RP-HPLC were identified based on their molecular mass, arising from ESI-MS. Putative genetic variants and post-translational (glycosylation and phosphorylation) isoforms were determined by deconvoluting multiple charged ion spectra in a real mass scale. Knowing their primary structures, it is possible to determine molecular masses of non post-translationally modified proteins, and then we can precisely know the mass of phosphorylation isoforms resulting from the addition of phosphate groups (±79.98 Da). Likewise, masses of isoforms arising from cryptic splice site usage, usually leading to the loss of the first codon (CAG) of an exon specifying a glutaminyl residue (-128 Da), are easily deduced. A camel mass reference database was thus created for the main milk proteins by combining the data available from *C. dromedarius*, *C. bactrianus, C. ferus, and Lama glama* milk protein sequences published in UniProtKB (ExPASy SIB Bioinformatics Resource Portal) and the National Centre for Biotechnology Information (NCBI).

To illustrate the efficiency of such an approach, a typical protein profile obtained with a milk from a hybrid camel sampled in Kyzylorda region is given in Figure 2.4. The analysis of molecular isoforms, identified from mass data, are reported in Table 2.3, in which experimental and theoretical molecular masses of camel milk proteins are given and confronted. The mass accuracy has allowed distinguishing about 50 protein molecules corresponding to isoforms belonging to nine protein families, eluted from the reverse-phase column as 15 peaks.

In peak I, the two molecular masses (21,157 Da and 21,184 Da) found were associated with glycoforms of κ-CN. The molecular mass of 21,157 Da corresponds to mono-phosphorylated variant A of κ-CN with tri-saccharides ((GaN-Ga-SA2) x 3 or (GaN-Ga) + (GaN-Ga-SA3) x 2, or (GaN-Ga-SA) + (GaN-Ga-SA2) + (GaN-Ga-SA3)). The molecular mass of 21,184 Da was expected to be non-phosphorylated variant B of κ-CN with penta-saccharides ((GaN-Ga) x 3 + (GaN-Ga-SA2) x 2, or (GaN-Ga) + (GaN-Ga-SA) x 4, or (GaN-Ga) x 2 + (GaN-Ga-SA) x 2 + (GaN-Ga-SA2), or (GaN-Ga) x 3 + (GaN-Ga-SA) + (GaN-Ga-SA3)). Peak II contained molecules of which the molecular masses (18,210 Da and 18,236 Da) were identified as non-phosphorylated variant B of κ-CN along with the A variant

modified at its N-terminal residue to form a pyro-glutamic acid (pyro-E), which is formed spontaneously by cyclization of the N-terminal E residue. The two molecular masses: 12,564 Da and 12,644 Da, detected in peak III, were assigned to the WAP peptide chain without or with one P group, respectively. Peaks IV, V, and VI were shown to contain α_{s1}-CN. The molecular mass of 23,878 Da observed in peak IV was interpreted as being a short isoform (201-residues) of α_{s1}-CN variant A with 4P groups, arising from exons 13' and 16 skipping events in the mature mRNA during the course of primary transcripts splicing, resulting in deleted sequences (residues E112-Q117 and E155-E162). Despite identification of only one splicing isoform with 4P groups (23,878 Da) in this milk sample, isoforms with 3P and 5P, along with cryptic splice site usage were identified in several other milk samples. Peak V consisted of three relative groups of three masses with sequential increments (s.i.) of 80 Da: 24,547 Da - 24,707 Da, 24,675 Da - 24,835 Da, and 24,689 Da - 24,849 Da. The mass difference (128 Da) between the first and the second group (Table 2.3.) corresponds to the loss of glutaminyl residue 83 (ΔQ83), encoded by the first codon (CAG) of exon 11. As reported previously (Kappeler et al., 1998), 24,755 Da was identified as the short isoform (207-residues) of the α_{s1}-CN variant A originating from exon 16 skipping during the course of the primary transcript processing. The mass difference (14 Da) between the second (24,675 Da) and the third (24,689 Da) group is due to the aa substitution E30D reported by Shuiep et al., (2013) characterizing the C variant. Thus, it is concluded that the third mass group gathers α_{s1}-CN short isoforms (207-residues) of variant C, with 5P, 6P and 7P, respectively, described in *C. dromedarius*. While cryptic splice site isoforms (ΔQ83) of variant C, with different phosphorylation levels, were not found in the milk sample shown at Figure 2.4, they were successfully found in several milk samples. Whereas, α_{s1}-CN short isoform was systematically present in all camel milk samples with 5, 6 and 7P (Table 2.3.), by contrast, α_{s1}-CN short isoforms of variant C occurred in some milk samples with 4P (24,611 Da) and up to 9P (24,929 Da). Herein, α_{s1}-CN short isoforms of variants A and C carrying 6P groups are isoforms with the highest mass signal intensity values 50,634 *vs.* 47,392, respectively.

Peak VI was more complex to interpret. Masses found in this peak belonged to four different molecular mass groups: 14,430 Da (ascribed to α-LAC), 22,939-23,099 Da (s.i. of 80 Da), 25,646 Da and 25,693-25,773 Da (s.i. of 80 Da), and 25,787 Da. Masses around 23 kDa (22,939-23,099 Da), with a mass increment of two P groups (160 Da), were not referenced to any protein in our database. These findings strongly suggest the existence of an additional uncharacterized phosphorylated protein, namely UP1, which remains to be

identified. The third mass group, 25,646 Da and 25,693-25,773 Da, corresponds to a mixture of two long isoforms (214 and 215 aa residues, respectively) of α_{s1}-CN variant C with 5P and 6P (25,693-25,773 Da) which differs from variant A by an aa substitution (E30D) in the mature protein (Erhardt et al., 2016). The mass of 25,646 Da corresponds to a 214 aa residues isoform of α_{s1}-CN variant C ($_{\Delta}$Q83), with 6P. The last molecular mass (25,787 Da) found in this peak was related to the mature variant A of α_{s1}-CN bearing 6P groups, which is by far much less abundant than the short α_{s1}-CN A-6P isoform (intensity of the mass signals: 3,472 *vs.* 50,634).

The four subsequent peaks (VII, VIII, IX, and X) all contained α_{s2}-CN molecules, with phosphorylation levels ranging between 7P (21,825 Da, peak VII) and 12P (22,226 Da, peak X). Observed molecular masses of 21,825-21,984 Da were in perfect concordance with those predicted for α_{s2}-CN displaying 7P and 9P, whereas α_{s2}-CN with 8P (21,906 Da) was the most frequent isoform. In addition, the mass of 23,179 Da in peak VII probably corresponds to the UP1 found in fraction VI with one more P group. Masses ranging between 21,986 and 22,226 Da (s.i. of 80 Da) found in peaks VIII, IX, and X were related to α_{s2}-CN variant A with 9P to 12P. These results suggest three more potential phosphorylation sites than reported by Kappeler et al., (1998) who mentioned a maximum of 9 S residues phosphorylated in camel α_{s2}-CN. More recently, Felfoul et al., (2017) detected two α_{s2}-CN isoforms with 10 and 11P groups in camel milk. Interestingly, peak X contains a second uncharacterized protein (UP2) with a molecular mass of 23,046 Da, not referring to any mass in our database for camel milk proteins. Such a mass was found in all camel milk samples analyzed so far (n=30). This suggests the possible existence of a further phosphoprotein in camel milk, very likely a CN, since two putative related isoforms with two (23,206 Da) and three (23,286 Da) additional P groups were detected in peak XI, in which the most abundant mass found (19,143 Da) was attributed to PGRP.

In the hybrid from Kyzylorda region (Table 2.3.), masses found in peak XII ranged between 66,481 and 67,342 Da. The most abundant masses 66,481 Da and 66,512 Da might be related to CSA of which the theoretical mass (peptide sequence predicted from the *C. dromedarius* genome, NCBI Accession number XP_010981066.1) is 66,477 Da. The mass differences of 4 Da and 35 Da could be attributed to putative genetic polymorphisms. The molecular weight reported by Felfoul et al., (2017) from fresh camel milk was estimated as 66,600 Da. However, one cannot exclude that such masses could correspond to LPO

depending on cleavage sites of the propeptide, when comparing with bovine LPO and human myeloperoxidase (Dull et al., 1990).

Molecular masses of 24,793 - 24,953 Da (s.i. of 80 Da) found in peak XIII, were ascribed to β-CN variant A with 2P, 3P and 4P, first described in the *C. bactrianus*. Molecular masses of 24,891-24,970 Da, which differ from β-CN A-3P and 4P by a 18 Da, correspond to β-CN variant B, first described in *C. dromedarius*. The mass difference of 18 Da between variants A and B is due to the M186I substitution. Isoforms of β-CN with 4P predominate whatever the milk sample and the genetic variant were, with equivalent intensity values of the mass signal for variants A and B, exemplified by a heterozygous hybrid camel: 84,494 *vs*. 87,973, respectively. In addition, the molecular mass of 24,842 Da, observed in peak XIII, corresponds to a splicing variant of β-CN B-4P. Such an isoform, which was so far considered as typical to the dromedary camel, was also found in hybrids and Bactrian camels. It is due to a cryptic splice site usage leading to the loss of the first codon (CAG) of exon 6, encoding residue Q29 in the protein.

Surprisingly, in the next peak (XIV), molecular masses around 24,000 Da (23,878 Da to 24,024 Da) were observed. Given the elution time and the mass range, these masses were very likely relative to the β-CN fraction, especially since a 18 Da mass differential existing between the pair of molecular masses (24,006 Da and 24,024 Da), is consistent with the occurrence of β-CN variants A and B, in both species. The important mass reduction, - 946 Da, relatively to the full-length β-CN, is hypothesized to be due to the cleavage by plasmin of the first seven N-terminal residues (1REKEEFK7) of the mature protein, given that this heptapeptide accounts for 947 Da. Furthermore, molecular masses equivalent to 23,878 Da and 23,895 Da are supposed to originate in the cryptic splice site usage (ΔQ29), previously mentioned. Finally, in the last peak (XV) mass values 12,357 Da and 12,376 Da again with the mass difference in 18 Da were observed. These masses correspond very likely to camel γ_2-CN A and B (12,357 Da *vs.* 12,376 Da, respectively), which are degradation products of β-CN (Beg et al., 1986).

This extensive analysis shows that mass accuracy provided by LC-ESI-MS was effective to allow protein identification of most of the protein isoforms by comparison of masses observed experimentally to theoretical molecular masses, and sufficiently powerful to recognize post-translational modifications (PTM) such as phosphorylation of CN, as well as genetic variants and long and short isoforms due to splicing inaccuracies.

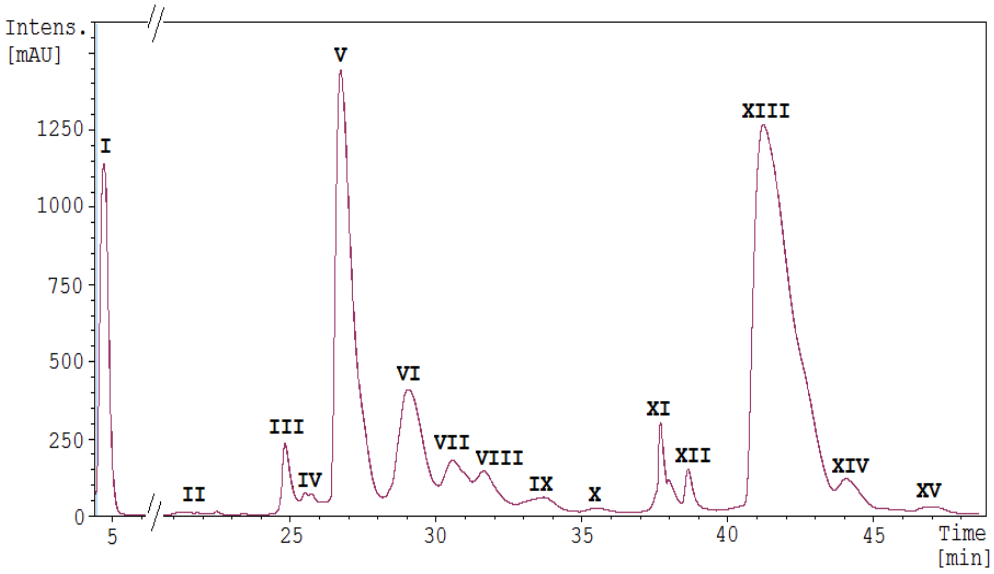


Figure 2.4. LC-ESI-MS profile of clarified crossed camel milk of Kyzylorda region. Nine major milk protein fractions were identified in the following order: peak I and II contained glycosylated ant natural isoforms of κ-CN; peak III: WAP; peaks IV, V: α_{s1}-CN; peak VI: α-LAC, α_{s1}-CN and UP1; peak VII: α_{s2}-CN and UP1; peaks VIII, IX, and X α_{s2}-CN along with UP2 in peak X; peak XI: PGRP and UP2; peak XII: CSA/LPO; peaks XIII and XIV: β-CN, and peak XV: γ_2-CN

Table 2.3. Identification of Camel milk protein **(**hybrid from Kyzylorda region) from observed molecular masses using LC-ESI-MS

Peak	Ret.Time, min	Observed M_r, Da	Theoretical M_r, Da	Protein description	UniProt/ NCBI GenBank Accession number	Intensity
I	4.50	21,157	21,158	κ-CN A, 1P, (GaN-Ga-SA2)x3*, pyro-E		1,361
		21,184	**21,182**	**κ-CN B, 0P, (GaN-Ga)x3 + (GaN-Ga-SA2)x2**, pyro-E**		**5,810**
II	18.61	18,210	18,210	κ-CN B, 0P ?	L0P304	161
		18,236	18,235	κ-CN A, 0P, pyro-E	P79139	72
III	24.32	12,564	12,564	WAP, 0P	P09837	1,756
		12,644	12,644	WAP, 1P		1,575
IV	24.97	23,878	23,878	α_{s1}-CN A - short isoform (Δex 16 and Δex 13'), 4P		242
V	26.23	24,547	24,547	α_{s1}-CN C -short isoform (Δex 16), 5P, splice variant (ΔQ83)		4,885
		24,627	**24,627**	**α_{s1}-CN C - short isoform (Δex 16), 6P, splice variant (ΔQ83)**		**21,606**
		24,707	24,707	α_{s1}-CN C - short isoform (Δex 16), 7P, splice variant (ΔQ83)		6,990
		24,675	24,675	α_{s1}-CN C - short isoform (Δex 16), 5P		9,441
		24,755	**24,755**	**α_{s1}-CN C - short isoform** (Δex 16)**, 6P**	**K7DXB9**	**47,392**
		24,835	24,835	α_{s1}-CN C - short isoform (Δex 16), 7P		7,046
		24,689	24,689	α_{s1}-CN A - short isoform (Δex 16), 5P		9,748
		24,768	**24,769**	**α_{s1}-CN A - short isoform** (Δex 16)**, 6P**	**O97943-2**	**50,634**
		24,849	24,849	α_{s1}-CN A - short isoform (Δex 16), 7P		6,909
VI	28.53	**14,430**	**14,430**	**α-LAC**	**P00710**	**17,797**
		22,939	n/a	Uncharacterized protein 1 (UP1)	n/a***	2,701
		23,020	n/a	UP1+80Da	n/a	2,489
		23,099	n/a	UP1+160Da	n/a	1,079
		25,646	25,645	α_{s1}-CN C, 6P, splice variant (ΔQ83)		3,501
		25,693	25,693	α_{s1}-CN C, 5P		564
		25,773	**25,773**	**α_{s1}-CN C, 6P**		**7,880**
		25,787	**25,787**	**α_{s1}-CN A, 6P**	**O97943-1**	**3,472**
VII	30.05	21,825	21,826	α_{s2}-CN, 7P		552
		21,906	**21,906**	**α_{s2}-CN, 8P**	**O9794**	**5,242**
		21,984	21,986	α_{s2}-CN, 9P		403
		23,178	n/a	UP1+240Da	n/a	1,256
VIII	31.11	21,986	21,986	α_{s2}-CN, 9P	O97944	356
		22,066	**22,066**	**α_{s2}-CN, 10P**		**4,790**
IX	33.18	22,066	22,066	α_{s2}-CN, 10P		148
		22,145	**22,146**	**α_{s2}-CN, 11P**		**1,964**
X	35.05	22,226	22,226	α_{s2}-CN, 12P		894
		23,046	n/a	Uncharacterized protein 2 (UP2)	n/a	231
XI	37.16	**19,143**	**19,143**	**PGRP**	**Q9GK12**	**7,207**
		23,206	n/a	UP2+160Da	n/a	1,592
		23,286	n/a	UP2+240Da	n/a	735
XII	38.09	66,481	66,477	CSA ?	XP_010981066.1	1,096
			66,491	LPO ?	Q9GJW6	
		66,512	n/a	CSA ? LPO?		2,663

		67,342	n/a	CSA ? LPO?		1,010
XIII	40.67	24,746	24,745	β-CN A, 3P, splice variant (ΔQ29)		2,073
		24,793	24,792	β-CN A, 2P		5,469
		24,825	24,825	β-CN A, 4P, splice variant (ΔQ29)		9,586
		24,873	24,872	β-CN A, 3P		10,177
		24,953	**24,953**	**β-CN A, 4P**	**A0A077SL35**	**84,494**
		24,842	24,842	β-CN B, 4P, splice variant (ΔQ29)		10,029
		24,891	24,890	β-CN B, 3P		10,365
		24,970	**24,971**	**β-CN B, 4P**	**Q9TVD0**	**87,973**
XIV	43.71	23,878	23,878	β-CN A-short isoform (Δ946 Da), 4P, splice variant (ΔQ29)		707
		23,963	23,958	β-CN A-short isoform (Δ946 Da), 5P, splice variant (ΔQ29)		244
		23,929	23,926	β-CN A-short isoform (Δ946 Da), 3P		438
		24,006	**24,006**	**β-CN A-short isoform (Δ946 Da), 4P**		**9,026**
		23,895	23,896	β-CN B-short isoform (Δ946 Da), 4P, splice variant (ΔQ29)		625
		24,024	**24,024**	**β-CN B-short isoform (Δ946 Da), 4P**		**5,545**
XV	47.02	12, 357	12,358	γ2-CN A, 0P		1,473
		12,376	12,376	γ2-CN B, 0P		1,065

Major proteins within each peak are in bold. Proteins and isoforms previously described are on grey background

*(GaN-Ga-SA2) x 3, or (GaN-Ga) + (GaN-Ga-SA3) x 2, or (GaN-Ga-SA) + (GaN-Ga-SA2) + (GaN-Ga-SA3)

**(GaN-Ga) x 3 + (GaN-Ga-SA2) x 2, or (GaN-Ga) + (GaN-Ga-SA) x 4, or (GaN-Ga) x 2 + (GaN-Ga-SA) x 2 + (GaN-Ga-SA2), or (GaN-Ga) x 3 + (GaN-Ga-SA) + (GaN-Ga-SA3)

***n/a - not applicable

2.3.5 Multiple spliced variants of CSN1S1

To confirm the occurrence of *CSN1S1* multiple splice variants, we took advantage of the possibility to extract RNA from milk fat globules to sequence PCR fragments of cDNA encoding α_{s1}-CN. Three different *CSN1S1* transcripts were found in each species and both genetic variants A and C. The nucleotide sequence of the most frequent variant transcript was shown to be deleted of exon 16, encoding the octapeptide EQAYFHLE. Besides, we also observed an isoform displaying the same sequence in which the first codon of exon 11 was lacking. Finally, a full-length transcript including exon 16 and the first codon of exon 11 was also detected, at a lower concentration.

2.4 Discussion

Given the growing interest in camel milk, due to the health potential of its bioactive components (Al haj & Al Kanhal, 2010) and frequently reported high anti-microbial activity (El-Agamy, 2009), over the past 20 years and even more during the last decade, the milk

protein fraction of Camelids, from all around the world has been extensively investigated (Alhaider et al., 2013; El-Agamy et al., 2009; Ereifej et al., 2011; Erhardt et al., 2016; Felfoul et al., 2017; Hinz et al., 2012; Kappeler et al., 1999; Konuspayeva et al., 2007; Merin et al., 2001; Ochirkhuyag et al., 1997; Salmen et al., 2012; Shuiep et al., 2013; Wangoh et al., 1998; Yang et al., 2013; Youcef et al., 2009). All these studies have explored, with more or less efficient approaches, the composition of the major milk proteins. However, the molecular diversity of these major proteins had not yet been studied. Then, our main objective was i) to provide, if not a comprehensive, at least an in-depth description of the protein fraction of camel milk; ii) to go further into an extensive analysis of the molecular diversity of major milk proteins from Camelids (*C. dromedarius*, *C. bactrianus,* and hybrids) sampled from different regions in Kazakhstan. For these purposes, different proteomic tools and methodological approaches were applied. For short, up to 391 protein species were identified in cumulating LC-MS/MS analyses of 8 individual *Camelus* milk, and the extensive characterization of CN and WP polymorphisms, using LC-ESI-MS, revealed a minimum of 50 molecular species.

2.4.1 Interspecies in-depth proteomic analysis of camel milk proteins

To our knowledge, the number of proteins identified in this study was relatively higher compared to the numbers reported in previous studies on the camel proteome (Alhaider et al., 2013; Yang et al., 2013). The largest camel milk proteome determined so far comprised about 238 proteins including some known camel proteins and heavy-chain immunoglobulins (Alhaider et al., 2013). In the abovementioned study carried out on *C. dromedarius*, proteins were identified from 2D SDS-PAGE with subsequent matrix-assisted laser desorption/ionization (MALDI) time-of-flight mass spectrometry analysis. However, it should be mentioned that several of the 238 proteins identified matched with the same protein in different species. Hence, at most *ca.* 140 proteins may be considered as unique. By comparison, in the present study a total of 391 unique protein species were determined from LC-MS/MS analyses of *C. bactrianus* (n=3), *C. dromedarius* (n=3), and hybrids (n=2), sampled from three different regions (Atyrau, Shymkent and Kyzylorda). Proteins such as flavin monoamine oxidase, perilipin 2, neutrophil gelatinase-associated lipocalin-like protein, brain-specific serine protease 4-like protein and others, which were not determined

previously, were successfully detected. Conversely, about 30 proteins identified by Alhaider and co-workers (2013) were not found in our study.

However, as for other mammals, CN represent the major protein fraction of camel milk (80%), among which β-CN is the most abundant (Pauciullo et al., 2014). Quantitative analyses performed by Kappeler et al., (2003) on camel milk CN have demonstrated significant higher amounts of β-CN (15 g/L *vs.* 10 g/L) compared to the homologous bovine β-CN and significant lower amounts of κ-CN (0.8 g/L *vs.* 3.5 g/L). Regarding relative proportions, as previously reported (Kappeler et al., 1998), α_{s1}-, α_{s2}-, β- and κ-CN contribute to about 22%, 9.5%, 65%, and 3.5% of total CN, respectively. Taking into account the 30 milk samples analyzed in LC-ESI-MS, relative proportions of individual CN, estimated from the mass signal intensity of each CN family (summing the mass signal of its phosphorylation and splicing isoforms) relatively to the sum of mass signal intensities of all CN families (considering that ionizing properties of caseins and their isoforms are comparable), were 37% α_{s1}-CN, 6.1% α_{s2}-CN, 53.1% β-CN, and 3.8% κ-CN. These values varied considerably compared to those reported previously by Kappeler et al., (1998) essentially as far as α_{s1}-CN and β-CN are concerned. Whereas α_{s1}-CN accounts for 36.1% for *C. bactrianus,* it reaches 37.4% and 37.6% in *C. dromedarius* and hybrids, respectively (Table 2.4.). Percentage of α_{s1}-CN calculated in our study was 15% higher than the value reported by Kappeler et al., (1987). Such an increase is compensated in part by a decrease of 12% of β-CN. The small amount of κ-CN observed is probably underestimated, since most of the highly glycosylated isoforms were not detected. However, this is in agreement with the fact that the size distribution of CN micelles is inversely related to κ-CN content (Bijl et al., 2014; Ostersen et al., 1997), since camel CN micelles are the largest, ranging in size between 280-550 nm (Bornaz et al., 2009).

Even though, there are 2 potential phosphorylation sites in κ-CN (S141 and S159) conserved and phosphorylated in sheep and goats (Martin et al., 2013) only isoforms with a single or no P group in the first chromatographic peak comprising glycosylated isoforms with 3 or 5 carbohydrate motifs were detected. Five glycosylated isoforms of camel κ-CN ranging in size between *ca.* 24 and 25.9 kDa were found in camel milk using 2D SDS-PAGE (Hinz et al., 2012).

In addition, γ_2-CN, a C-terminal product resulting from a highly specific proteolysis of β-CN by the natural milk protease (plasmin) was successfully found in the milk samples analyzed. Previously published data suggested that the proportion of γ-CN in total CN

fraction is highest at the beginning and the end of lactation, and in very low yielding animals (Ostersen et al., 1997). The molecular masses observed in this study (12,357 Da and 12,376 Da) were lower from those previously observed by Kappeler et al., (1998): 13.9 kDa, 15.7 kDa and 15.75 kDa.

Immune-related proteins such as GlyCAM1, MFGE8 and LTF were detected in camel milk. GlyCAM1, also named lactophorin or PP3 is a cysteine free protein, which belongs to the family of GlyCAM-type molecules (Beg et al., 1987). Two splicing variants A and B were distinguished in camel milk (Kappeler et al., 1999). Variant A encoding 137 aa residues has a M_r of 15.7 kDa, while variant B encoding 122 aa residues has a M_r of 13.8 kDa. The primary structure of Variant A reveals 54% identity with a protein isolated from bovine milk (Sørensen & Petersen, 1993). Until late, it has been claimed that camel GlyCAM1 is neither glycosylated nor phosphorylated as bovine GlyCAM1. However, Girardet et al., (2000) suggested the probable existence of one O-glycosylation site (16TDT18) in variant A of which the apparent M_r was estimated as 22.5 kDa from SDS-PAGE. Using the same approach, two bands were found, in which we identified GlyCAM1 from LC-MS/MS analysis 22 kDa and 10 kDa, corresponding probably to the glycosylated and putatively phosphorylated isoform of GlyCAM1 observed by Girardet et al., (2000), and to a product of proteolysis, respectively. Surprisingly, no molecular masses corresponding to camel GlyCAM1 A and B were identified by LC-ESI-MS analysis. Likewise, LC-ESI-MS did not permit to detect LTF, even though, SDS-PAGE and LC-MS/MS data confirm its presence in analyzed camel milk samples. On the other hand, molecular masses ranging between 74,338 Da-79,621 Da could be attributed to camel LTF of which the theoretical mass reported by Kappeler et al., (1999) for the mature protein (689 aa residues long) without PTM is 75,250 Da. Therefore, the mass difference observed is very likely attributable to PTM. In addition, Konuspayeva et al., (2005) reported that the level of LTF is affected by seasonal variations.

Elsewhere, MFGM-enriched proteins such as XO, BTN, fatty acid synthase, actin, ras-related protein Rab-18, ADP-ribosylation factor 1, tyrosine-protein kinase, GTP-binding protein SAR1b were identified in Kazakh camel milk samples in accordance with previous results obtained with *C. dromedarius* (Saadaoui et al., 2013) and *C. bactrianus* (Yang et al., 2013) milk samples. Surprisingly, whereas BTN was present in all milk samples, it seems to be absent in *C. bactrianus* from Atyrau region. This could be due to the way the band in the electrophoresis gel was cut, since BTN was found in the other seven samples analyzed. Regarding proteins originating from blood, such as serpin A3-1, apolipoprotein A-1, α-1-

antitrypsin like protein, α-1-acid glycoprotein, β-2-microglobulin, complement C3-like protein, they were also found in Kazakh camel milks, in agreement with findings of Yang et al., (2013) reported for Bactrian camels from China. By contrast, as mentioned in the Results section, no trace of CSA was found in Kazakh milk samples from LC-MS/MS analyses, whereas its presence is suspected from LC-ESI-MS.

A heat shock protein (HSPA6 also called HSP70B') occurred at rank 23 amongst the first third of the most represented proteins in Kazakh camel milks (Table 2.2.). Expression of heat shock proteins, including HSP70 is increased during heat stress and involved in defense against dehydration or thermal stress in arid environments (Rhoads et al., 2013; Sharma et al., 2013). The entire sequence of this protein has been deduced from the nucleotide sequence of a full-length cDNA in *C. dromedarius* (Elrobh et al., 2011). Comprising 643 aa residues, the camel protein, of which the M_r is 70,543 Da in agreement with the molecular mass estimated from SDS-PAGE, shares a high similarity (94% identity) with cow and pig HSP70.

Against all expectations, peptides with sequence similarity with bovine *β*-lactoglobulin, the major allergen in bovine milk, were identified in the 8 camel milk samples (Bactrian, dromedary and hybrids) from Kazakhstan, analyzed by LC-MS/MS. The coverage percentage ranged between 30 and 60% in individual milk samples, and reached 71% cumulating all the peptides found. Five peptides related to bovine *β*-lactoglobulin were also detected by Alhaider et al., (2013) in camel milk from Saudi Arabia and the United States. Youcef et al., (2009) revealed a weak cross reaction between dromedary WPs and IgG anti bovine *β*-lactoglobulin. Such findings disagree with the usually admitted notion that *β*-lactoglobulin is absent in camel milk (Hinz et al., 2012; Restani et al., 1999). Even though we cannot exclude a possible contamination by bovine milk (unlikely with the 8 camel milk samples analyzed by LC-MS/MS) or the presence in camel milk of a Progesterone Associated Endometrial Protein (PAEP) displaying strong similarities with *β*-lactoglobulin. However, significant similarities between human PAEP and the peptides having allowed the identification of *β*-lactoglobulin in *C. bactrianus* milk, were not found.

Table 2.4. Relative proportion of each CN expressed in %, estimated from the mass signal intensity of each CN family relatively to the sum of mass signal intensities of all CN families in the three camel species

	κ-CN		**α_{s1}-CN**		**α_{s2}-CN**		**β-CN**	
	m	σ	m	σ	m	σ	m	σ
Bactrian	3,09	1,89	36,09	2,33	7,13	1,49	53,68	2,08
Dromedary	3,63	2,13	37,39	3,89	5,79	0,98	53,19	3,46
Hybrid	4,77	3,01	37,57	3,03	5,25	1,56	52,41	4,18

m = mean; σ = standard deviation

2.4.2 Molecular diversity of camel caseins: genetic polymorphism and alternative splicing

Regarding camel α_{s1}-CN, the situation is particularly confusing. Kappeler et al., (1998) first described two cDNA (short and long), encoding two protein isoforms of 207 and 215 aa, named A and B variants. The A variant corresponds to the short isoform (207 aa), in which the octapeptide 155EQAYFHLE162 encoded by exon 16 was missing, whereas this octapeptide is present in the 215 aa-long isoforms. In our study, two isoforms long and short showing a 1,018 Da mass difference were found, in which the short isoform was the major component (*ca.* 90%) of total camel α_{s1}-CN. Such an alternative splicing event has been first reported in goats (Leroux et al., 1992), sheep (Chianese et al., 1996; P. Ferranti et al., 1995) and later in lama (Pauciullo & Erhardt, 2015). In addition, we observed the existence of two distinct genetic variants called A and C, arising from the E30D aa substitution, as previously reported by Shuiep et al., (2013). Since, variants A and B described by Kappeler et al., (1998) displayed a E aa residue in position 30 of the mature peptide chain, it becomes obvious that Kappeler's A and B variants derived in fact from a single allele, of which the primary transcript is subject to exon 16 skipping during the splicing process. In other words, the B variant is nothing other than a splicing variant of a single allele that we propose to call *CSN1S1*A*.

Recently, Erhardt et al., (2016) reported in *C. dromedarius* from different regions of Sudan, the existence of a further variant, called D, clearly displaying a different IEF behavior. Excluding this D variant, which was not precisely characterized, there are α_{s1}-CN long and short non-allelic isoforms arising from alternative splicing of a single primary transcript and only two perfectly characterized genetic variants A and C resulting from a single G>T

nucleotide substitution in exon 4 and leading to E30D aa substitution. This molecular diversity is becoming more complex due to different phosphorylation levels ranging between 5-8P groups (see thereafter) and due to isoforms arising from cryptic splice site usage (Boumahrou et al., 2011; Ferranti et al., 1999; Leroux et al., 1992), leading to the loss of a Q residue corresponding to the first codon of exon 11. Results from cDNA sequencing substantiate this.

Electrophoretic and LC-MS analyses as well as cDNA sequencing confirmed that β-CN occurs as two genetic variants A and B, with the aa substitution M186I (yielding a -18 Da mass difference). The most frequent form of β-CN had 4P groups, one P group more than reported for Somali, Turkana and Pakistani camels by Kappeler et al., (1998). Surprisingly, in Kazakh populations, a second series of β-CN components with lower molecular masses (mass difference: -946 Da), relatively to the full-length β-CN were found. This phenomenon, observed with both genetic variants, might be due to the cleavage by plasmin of the first seven N-terminal residues (REKEEFK) of the mature protein. A mass difference of 947 Da was observed between the native full-length protein with 4P (24,953 Da and 24,971 Da for A and B variants, respectively) and the plasmin cleavage product at the same phosphorylation level (24,006 Da and 24,024 Da for A and B variants, respectively). The occurrence of a K residue in position 7 of the mature β-CN does not occur in any other species, of which the N-terminal sequence is known (Martin et al., 2013). However, our results strongly suggest that the peptide bond 7K-T8 is sensible to plasmin that is, like trypsin, a serine protease. Indeed, REKEEFK was present amongst tryptic peptides identified in LC-MS/MS analysis.

There is another even less probable possibility, involving the deletion of exon 5 that encodes 8 aa residues (ESITHINK for a mass of 923 Da), since a similar event was previously characterized from mare (Miranda et al., 2004) and donkey (Cunsolo et al., 2017) milks. However, sequencing of camel β-CN cDNA has not revealed any deletion in the mRNA encoding this protein (results not shown), consistently with Kappeler et al., (1998) who only reported a full-length sequence for β-CN, conversely to α_{s1}-CN. Since in our study we were not able to provide any further confirmation of the presence of shorter mRNA of camel β-CN in which exon 5 is spliced out, we give preference to the cleavage by plasmin of the first seven N-terminal residues of β-CN rather than an alternative splicing process.

Surprisingly, two so far uncharacterized proteins (UP1 and UP2) with molecular masses around 23,000 Da and different phosphorylation levels were observed, suggesting they

are possibly proteins related to CN. However, to prove this hypothesis further research for in depth characterization of these proteins is necessary.

2.4.3 Post-translational modifications of milk proteins: phosphorylation of caseins

Among the various approaches developed in proteomics, electrospray ionization (ESI) mass spectrometry (MS) is eminently suitable for studying PTM, including phosphorylation and glycosylation, since the technique provides molecular mass determination of native proteins. Phosphorylation of proteins is one of the most frequent PTM in eukaryotic cells. It has become a common knowledge that phosphorylation of CN occurs at S or T aa residues in tripeptide sequences S/T-X-A where X represents any aa residue and A is an acidic residue (Mercier, 1981). This consensus sequence is recognized by FAM20C, a Golgi CN-kinase, which phosphorylates secreted phosphoproteins, including both CN and members of the small integrin-binding ligand N-linked glycoproteins (SIBLING) protein family, which modulate biomineralization (Ishikawa et al., 2012). Each phosphorylation event adds 79.98 Da to the molecular mass of the peptide chain (Larsen et al., 2006). (Larsen et al., 2006)It was predicted with high confidence 8 probably phosphorylated S residues in α_{s1}-CN (S18, S68, S70, S71, S72, S73, S193, and S202), 9 potential phosphorylated S residues in α_{s2}-CN (S8, S9, S10, S32, S53, S108, S110, S113, and S121), 4 S residues in β-CN (S15, S17, S18, and S19), and 2 S residues in κ-CN (S141 and S159). However, up to 9P residues per α_{s1}-CN molecule were observed whatever the genetic variant is. Theoretically, given the S/T-X-A consensus rule, there are 4 T residues that could be phosphorylated (T55, T80, T153, and T196), leading to a maximum of 12 P groups per molecule. Therefore, we can put forward that at least one of the four T residues is phosphorylated in the α_{s1}-CN-9P.

With 11 potentially phosphorylated aa residues matching the S/T-X-A motif (Figure 2.5), camel α_{s2}-CN displays the highest phosphorylation level, in agreement with Felfoul et al., (2017), who reported recently 11P groups. To reach such a phosphorylation level, besides the nine SerP, two putative ThrP (T118 and T132) have to be phosphorylated. In all the Kazakh milk samples analyzed in LC-ESI-MS we found α_{s2}-CN with 12 P groups, as the molecular mass of 22,226 Da observed corresponds to the mass of the peptide backbone (21,266 Da) increased by 960 Da, a mass increment which coincides with 12 P groups. That means that at least another S/T residue that does not match with the canonic sequence

recognized by the mammary kinase(s), is potentially phosphorylated. According to Allende et al., (1995) the sequence S/T-X-X-A follow-through with the minimum requirements for phosphorylation by the CN-kinase II (CK2). It is critical to highlight in this regard that E or D in this site can be replaced by SerP or ThrP. Two T residues, namely T39 and T129 in the camel α_{s2}-CN fully meet the requirements of the above-mentioned motif (Figure 2.5) and could be phosphorylated. Such an event is the only hypothesis to reach 12P for camel α_{s2}-CN. Since these two kinases are very likely secreted, the idea that phosphorylation at T39/T129 may occur in the extracellular environment cannot be excluded. This warrants further investigation. Fam20C, which is very likely the major secretory pathway protein kinase (Tagliabracci et al., 2015), might be responsible for the phosphorylation of S and T residues within S/T-X-A motif, whereas a CK2-type kinase might be responsible for phosphorylation of T residue within an S/T-X-X-A motif. This is in agreement with the hypothesis put forward by Bijl et al., (2014) and Fang et al., (2016), who suggest from phenotypic correlations and hierarchical clustering the existence of at least 2 regulatory systems for phosphorylation of α_s-CN. Elsewhere, bovine milk osteopontin which is a multiphosphorylated glycoprotein also found in bone, was shown to contain 27 SerP and one ThrP (Sørensen et al., 1995). Twenty five SerP and one ThrP were located in S/T-X-E/S(P)/D motifs, whereas two SerP were found in the sequence S-X-X-E/S(P).

```
         10          20          30          40          50
KHEMDQGSSS EESINVSQQK FKQVKKVAIH PSKEDICSTF CEEAVRNIKE
         60          70          80          90         100
VESAEVPTEN KISQFYQKWK FLQYLQALHQ GQIVMNPWDQ GKTRAYPFIP
        110         120         130         140         150
TVNTEQLSIS EESTEVPTEE STEVFTKKTE LTEEEKDHQK FLNKIYQYYQ
        160         170
TFLWPEYLKT VYQYQKTMTP WNHIKRYF
```

Figure 2.5. Amino acid sequence of mature camel α_{s2}-CN with potential phosphorylation sites. Seryl and Threonyl residues matching the S/T-X-A motif are in red and blue, respectively, and underlined. Threonyl residues matching the S/T-X-X-A motif are in green and underlined

2.5 Conclusions

In this study, six main findings combining proven proteomic and molecular biology approaches are provided. The first one is an enhancing of our knowledge of camel milk protein composition. The second one is deciphering the extreme complexity of camel CN fraction due to PTM (phosphorylation) and splicing events (exon skipping and cryptic splice

site usage). The third finding is the detection of two unknown proteins, UP1 and UP2 that remain to be characterized. In addition, we provide results substantiating the possible existence of a camel β-lactoglobulin. However, this result requires further investigation, currently in progress in the laboratory. Afterwards, we report for the first time the presence of α_{s2}-CN-12P, and short isoforms of β-CN probably arising from proteolysis by plasmin, the natural protease of milk. The ultimate finding is the demonstration that genetic variants, which hitherto seemed specific to a species (β-CN A for Bactrian and β-CN B for dromedary), are in fact present in both *dromedarius* and *bactrianus*.

Acknowledgements

The study was carried out within the Bolashak International Scholarship of the first author, funded by the JSC «Center for International Programs» (Kazakhstan). The research was partly supported by a grant from the Ministry of Education and Science of the Republic of Kazakhstan under name "Proteomic investigation of camel milk" #1729/GF4, which is duly appreciated. The authors thank all Kazakhstani camel milk farms for rendering help in sample collection, as well as PAPPSO and @BRIDGe teams at INRA (Jouy-en-Josas, France) for providing necessary facilities and technical support.

References

Agrawal, R. P., Swami, S. C., Beniwal, R., Kochar, D. K., Sahani, M. S., Tuteja, F. C., & Ghouri, S. K. (2003). Effect of camel milk on glycemic control, risk factors and diabetes quality of life in type-1 diabetes: A randomised prospective controlled study. *Journal of Camel Practice and Research*, *10*(1), 45–50.

Al-Ayadhi, L. Y., & Elamin, N. E. (2013). Camel milk as a potential therapy as an antioxidant in Autism Spectrum Disorder (ASD). *Evidence-Based Complementary and Alternative Medicine : ECAM*, *2013*, 602834. https://doi.org/10.1155/2013/602834

Al haj, O. A., & Al Kanhal, H. A. (2010). Compositional, technological and nutritional aspects of dromedary camel milk. *International Dairy Journal*. https://doi.org/10.1016/j.idairyj.2010.04.003

Alhaider, A., Abdelgader, A. G., Turjoman, A. A., Newell, K., Hunsucker, S. W., Shan, B.,

… Duncan, M. W. (2013). Through the eye of an electrospray needle: Mass spectrometric identification of the major peptides and proteins in the milk of the one-humped camel (Camelus dromedarius). *Journal of Mass Spectrometry*, *48*(7), 779–794. https://doi.org/10.1002/jms.3213

Allende, J. E., & Allende, C. C. (1995). Protein kinases. 4. Protein kinase CK2: an enzyme with multiple substrates and a puzzling regulation. *The FASEB Journal : Official Publication of the Federation of American Societies for Experimental Biology*, *9*(5), 313–323.

Balteanu, V. A., Carsai, T. C., & Vlaic, A. (2013). Identification of an intronic regulatory mutation at the buffalo αs1-casein gene that triggers the skipping of exon 6. *Molecular Biology Reports*, *40*(7), 4311–4316. https://doi.org/10.1007/s11033-013-2518-2

Beg, O. U., Bahr-Lindström, H. von, Zaidi, Z. H., & Jörnvall, H. (1987). Characterization of a heterogeneous camel milk whey non-casein protein. *FEBS Letters*, *216*(2), 270–274. https://doi.org/10.1016/0014-5793(87)80704-4

Beg, O. U., von Bahr-Lindström, H., Zaidi, Z. H., & Jörnvall, H. (1986). Characterization of a camel milk protein rich in proline identifies a new β-casein fragment. *Regulatory Peptides*, *15*(1), 55–61. https://doi.org/10.1016/0167-0115(86)90075-3

Bijl, E., de Vries, R., van Valenberg, H., Huppertz, T., & van Hooijdonk, T. (2014). Factors influencing casein micelle size in milk of individual cows: Genetic variants and glycosylation of κ-casein. *International Dairy Journal*, *34*(1), 135–141. https://doi.org/10.1016/j.idairyj.2013.08.001

Bijl, E., van Valenberg, H. J. F., Huppertz, T., van Hooijdonk, A. C. M., & Bovenhuis, H. (2014). Phosphorylation of αS1-casein is regulated by different genes. *Journal of Dairy Science*, *97*(11), 7240–7246. https://doi.org/10.3168/jds.2014-8061

Bornaz, S., Sahli, A., Attalah, A., & Attia, H. (2009). Physicochemical characteristics and renneting properties of camels' milk: A comparison with goats', ewes' and cows' milks. *International Journal of Dairy Technology*, *62*(4), 505–513. https://doi.org/10.1111/j.1471-0307.2009.00535.x

Bouchard, D., Morisset, D., Bourbonnais, Y., & Tremblay, G. M. (2006). Proteins with whey-

acidic-protein motifs and cancer. *Lancet Oncology*. https://doi.org/10.1016/S1470-2045(06)70579-4

Boumahrou, N., Bevilacqua, C., Beauvallet, C., Miranda, G., Andrei, S., Rebours, E., … Martin, P. (2011). Evolution of major milk proteins in Mus musculus and Mus spretus mouse species: a genoproteomic analysis. *BMC Genomics*, *12*(1), 80. https://doi.org/10.1186/1471-2164-12-80

Bradford, M. M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein using the principle of protein dye binding. *Analytical Biochemistry*, *72*, 248–254. https://doi.org/10.1016/0003-2697(76)90527-3

Brenaut, P., Bangera, R., Bevilacqua, C., Rebours, E., Cebo, C., & Martin, P. (2012). Validation of RNA isolated from milk fat globules to profile mammary epithelial cell expression during lactation and transcriptional response to a bacterial infection. *Journal of Dairy Science*, *95*(10), 6130–6144. https://doi.org/10.1016/j.jmb.2011.04.044

Chianese, L., Garro, G., Mauriello, R., Laezza, P., Ferranti, P., & Addeo, F. (1996). Occurrence of five αs1-casein variants in ovine milk. *Journal of Dairy Research*, *63*, 49–59.

Cunsolo, V., Saletti, R., Muccilli, V., Gallina, S., Di Francesco, A., & Foti, S. (2017). Proteins and bioactive peptides from donkey milk: The molecular basis for its reduced allergenic properties. *Food Research International*. https://doi.org/10.1016/j.foodres.2017.07.002

Davoodi, S. H., Shahbazi, R., Esmaeili, S., Sohrabvandi, S., Mortazavian, A. M., Jazayeri, S., & Taslimi, A. (2016). Health-related aspects of milk proteins. *Iranian Journal of Pharmaceutical Research*.

Dull, T. J., Uyeda, C., Strosberg, A. D., Nedwin, G., & Seilhamer, J. J. (1990). Molecular cloning of cDNAs encoding bovine and human lactoperoxidase. *DNA and Cell Biology*, *9*(7), 499–509. https://doi.org/10.1089/dna.1990.9.499

Dziarski, R., Kashyap, D. R., & Gupta, D. (2012). Mammalian Peptidoglycan Recognition Proteins Kill Bacteria by Activating Two-Component Systems and Modulate Microbiome and Inflammation. *Microbial Drug Resistance*, *18*(3), 280–285.

https://doi.org/10.1089/mdr.2012.0002

El-Agamy, E. I. (2009). Bioactive Components in Camel Milk. In *Bioactive Components in Milk and Dairy Products* (pp. 159–194). https://doi.org/10.1002/9780813821504.ch6

El-Agamy, E. I., Nawar, M., Shamsia, S. M., Awad, S., & Haenlein, G. F. W. (2009). Are camel milk proteins convenient to the nutrition of cow milk allergic children? *Small Ruminant Research*, *82*(1), 1–6. https://doi.org/10.1016/j.smallrumres.2008.12.016

El-Fakharany, E. M., El-Baky, N. A., Linjawi, M. H., Aljaddawi, A. A., Saleem, T. H., Nassar, A. Y., … Redwan, E. M. (2017). Influence of camel milk on the hepatitis C virus burden of infected patients. *Experimental and Therapeutic Medicine*, *13*(4), 1313–1320. https://doi.org/10.3892/etm.2017.4159

Elrobh, M. S., Alanazi, M. S., Khan, W., Abduljaleel, Z., Al-Amri, A., & Bazzi, M. D. (2011). Molecular cloning and characterization of cDNA encoding a putative stress-induced heat-shock protein from Camelus dromedarius. *International Journal of Molecular Sciences*, *12*(7), 4214–36. https://doi.org/10.3390/ijms12074214

Ereifej, K. I., Alu'datt, M. H., Alkhalidy, H. A., Alli, I., & Rababah, T. (2011). Comparison and characterisation of fat and protein composition for camel milk from eight Jordanian locations. *Food Chemistry*, *127*(1), 282–289. https://doi.org/10.1016/j.foodchem.2010.12.112

Erhardt, G., Shuiep, E. T. S., Lisson, M., Weimann, C., Wang, Z., El Zubeir, I. E. Y. M., & Pauciullo, A. (2016). Alpha S1-casein polymorphisms in camel (Camelus dromedarius) and descriptions of biological active peptides and allergenic epitopes. *Tropical Animal Health and Production*, *48*(5), 879–887. https://doi.org/10.1007/s11250-016-0997-6

Fang, Z. H., Visker, M. H. P. W., Miranda, G., Delacroix-Buchet, A., Bovenhuis, H., & Martin, P. (2016). The relationships among bovine αS-casein phosphorylation isoforms suggest different phosphorylation pathways. *Journal of Dairy Science*, *99*(10), 8168–8177. https://doi.org/10.3168/jds.2016-11250

FAO. (2017). FAOSTAT. Retrieved from http://www.fao.org/faostat/en/#home

Felfoul, I., Jardin, J., Gaucheron, F., Attia, H., & Ayadi, M. A. (2017). Proteomic profiling of

camel and cow milk proteins under heat treatment. *Food Chemistry*, *216*, 161–169. https://doi.org/10.1016/j.foodchem.2016.08.007

Ferranti, P., Lilla, S., Chianese, L., & Addeo, F. (1999). Alternative nonallelic deletion is constitutive of ruminant α(s1)- casein. *Journal of Protein Chemistry*, *18*(5), 595–602. https://doi.org/10.1023/A:1020659518748

Ferranti, P., Malorni, A., Nitti, G., Laezza, P., Pizzano, R., Chianese, L., & Addeo, F. (1995). Primary structure of ovine alpha s1-caseins: localization of phosphorylation sites and characterization of genetic variants A, C and D. *Journal of Dairy Research*, *62*(2), 281–296. https://doi.org/10.1017/S0022029900030983

Girardet, J. M., Saulnier, F., Gaillard, J. L., Ramet, J. P., & Humbert, G. (2000). Camel (camelus dromedarius) milk PP3: evidence for an insertion in the amino-terminal sequence of the camel milk whey protein. *Biochemistry and Cell Biology = Biochimie et Biologie Cellulaire*, *78*(1), 19–26. https://doi.org/10.1139/o99-067

Habib, H. M., Ibrahim, W. H., Schneider-Stock, R., & Hassan, H. M. (2013). Camel milk lactoferrin reduces the proliferation of colorectal cancer cells and exerts antioxidant and DNA damage inhibitory activities. *Food Chemistry*, *141*(1), 148–152. https://doi.org/10.1016/j.foodchem.2013.03.039

Hennighausen, L. G., & Sippel, A. E. (1982). Mouse whey acidic protein is a novel member of the family of “four-disulfide core” proteins. *Nucleic Acids Research*, *10*(8), 2677–2684. https://doi.org/10.1093/nar/10.8.2677

Hinz, K., O’Connor, P. M., Huppertz, T., Ross, R. P., & Kelly, A. L. (2012). Comparison of the principal proteins in bovine, caprine, buffalo, equine and camel milk. *Journal of Dairy Research*, *79*(02), 185–191. https://doi.org/10.1017/S0022029912000015

Hsieh, C. C., Hernández-Ledesma, B., Fernández-Tomé, S., Weinborn, V., Barile, D., & De Moura Bell, J. M. L. N. (2015). Milk proteins, peptides, and oligosaccharides: Effects against the 21st century disorders. *BioMed Research International*. https://doi.org/10.1155/2015/146840

Ishikawa, H. O., Xu, A., Ogura, E., Manning, G., & Irvine, K. D. (2012). The raine syndrome protein FAM20C is a golgi kinase that phosphorylates bio-mineralization proteins. *PLoS*

ONE, *7*(8). https://doi.org/10.1371/journal.pone.0042988

Kanwar, J. R., Roy, K., Patel, Y., Zhou, S. F., Singh, M. R., Singh, D., … Kanwar, R. K. (2015). Multifunctional iron bound lactoferrin and nanomedicinal approaches to enhance its bioactive functions. *Molecules*. https://doi.org/10.3390/molecules20069703

Kappeler, S., Ackermann, M., Farah, Z., & Puhan, Z. (1999). Sequence analysis of camel (Camelus dromedarius) lactoferrin. *International Dairy Journal*, *9*(7), 481–486. https://doi.org/10.1016/S0958-6946(99)00117-X

Kappeler, S., Farah, Z., & Puhan, Z. (1998). Sequence analysis of Camelus dromedarius milk caseins. *The Journal of Dairy Research*, *65*(2), 209–222. https://doi.org/10.1017/S0022029997002847

Kappeler, S., Farah, Z., & Puhan, Z. (1999). Alternative Splicing of Lactophorin mRNA from Lactating Mammary Gland of the Camel (Camelus dromedarius). *J Dairy Sci*, *82*(November 1999), 2084–2093. https://doi.org/10.3168/jds.S0022-0302(99)75450-0

Kappeler, S., Farah, Z., & Puhan, Z. (2003). 5′-Flanking Regions of Camel Milk Genes Are Highly Similar to Homologue Regions of Other Species and Can be Divided into Two Distinct Groups. *Journal of Dairy Science*, *86*(2), 498–508. https://doi.org/http://dx.doi.org/10.3168/jds.S0022-0302(03)73628-5

Kappeler, S., Heuberger, C., Farah, Z., & Puhan, Z. (2004). Expression of the peptidoglycan recognition protein, PGRP, in the lactating mammary gland. *Journal of Dairy Science*, *87*(8), 2660–8. https://doi.org/10.3168/jds.S0022-0302(04)73392-5

Konuspayeva, G., Faye, B., & Loiseau, G. (2009). The composition of camel milk: A meta-analysis of the literature data. *Journal of Food Composition and Analysis*. https://doi.org/10.1016/j.jfca.2008.09.008

Konuspayeva, G., Faye, B., Loiseau, G., & Levieux, D. (2007). Lactoferrin and immunoglobulin contents in camel's milk (Camelus bactrianus, Camelus dromedarius, and Hybrids) from Kazakhstan. *Journal of Dairy Science*, *90*(1), 38–46. https://doi.org/10.3168/jds.S0022-0302(07)72606-1

Konuspayeva, G., Serikbayeva, A., Loiseau, G., Narmuratova, M., & Faye, B. (2005).

Lactoferrin of camel milk of Kazakhstan. In *Desertification Combat and Food Safety: the Added Value of Camel Producers* (Vol. 362, pp. 158–167).

Korashy, H. M., Maayah, Z. H., Abd-Allah, A. R., El-Kadi, A. O. S., & Alhaider, A. a. (2012). Camel milk triggers apoptotic signaling pathways in human hepatoma HepG2 and breast cancer MCF7 cell lines through transcriptional mechanism. *Journal of Biomedicine & Biotechnology*, *2012*, 1–9. https://doi.org/10.1155/2012/593195

Laemmli, U. K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature*. https://doi.org/10.1038/227680a0

Larsen, M. R., Trelle, M. B., Thingholm, T. E., & Jensen, O. N. (2006). Analysis of posttranslational modifications of proteins by tandem mass spectrometry. *BioTechniques*. https://doi.org/10.2144/000112201

Legrand, D., Elass, E., Pierce, A., & Mazurier, J. (2004). Lactoferrin and host defence: An overview of its immuno-modulating and anti-inflammatory properties. *BioMetals*. https://doi.org/10.1023/B:BIOM.0000027696.48707.42

Leroux, C., Mazure, N., & Martin, P. (1992). Mutations away from splice site recognition sequences might cis-modulate alternative splicing of goat α(s1)-casein transcripts. Structural organization of the relevant gene. *Journal of Biological Chemistry*, *267*(9), 6147–6157.

Manaer, T., Yu, L., Zhang, Y., Xiao, X. J., & Nabi, X. H. (2015). Anti-diabetic effects of shubat in type 2 diabetic rats induced by combination of high-glucose-fat diet and low-dose streptozotocin. *Journal of Ethnopharmacology*, *169*, 269–274. https://doi.org/10.1016/j.jep.2015.04.032

Martin, P., Cebo, C., & Miranda, G. (2013). Interspecies comparison of milk proteins: Quantitative variability and molecular diversity. In *Advanced Dairy Chemistry: Volume 1A: Proteins: Basic Aspects, 4th Edition* (pp. 387–429). https://doi.org/10.1007/978-1-4614-4714-6_13

Mati, A., Senoussi-Ghezali, C., Si Ahmed Zennia, S., Almi-Sebbane, D., El-Hatmi, H., & Girardet, J. M. (2017). Dromedary camel milk proteins, a source of peptides having biological activities – A review. *International Dairy Journal*.

https://doi.org/10.1016/j.idairyj.2016.12.001

McMahon, D. J., & Oommen, B. S. (2013). Casein micelle structure, functions, and interactions. In *Advanced Dairy Chemistry: Volume 1A: Proteins: Basic Aspects, 4th Edition* (pp. 185–209). https://doi.org/10.1007/978-1-4614-4714-6_6

Mercier, J. C. (1981). Phosphorylation of caseins, present evidence for an amino acid triplet code posttranslationally recognized by specific kinases. *Biochimie*. https://doi.org/10.1016/S0300-9084(81)80141-1

Merin, U., Bernstein, S., Bloch-Damti, A., Yagil, R., Van Creveld, C., Lindner, P., & Gollop, N. (2001). A comparative study of milk serum proteins in camel (Camelus dromedarius) and bovine colostrum. *Livestock Production Science*, *67*(3), 297–301. https://doi.org/10.1016/S0301-6226(00)00198-6

Miranda, G., Mahé, M. F., Leroux, C., & Martin, P. (2004). Proteomic tools characterize the protein fraction of Equidae milk. In *Proteomics* (Vol. 4, pp. 2496–2509). https://doi.org/10.1002/pmic.200300765

Mohandesan, E., Speller, C. F., Peters, J., Uerpmann, H. P., Uerpmann, M., De Cupere, B., … Burger, P. A. (2017). Combined hybridization capture and shotgun sequencing for ancient DNA analysis of extinct wild and domestic dromedary camel. *Molecular Ecology Resources*, *17*(2), 300–313. https://doi.org/10.1111/1755-0998.12551

Nurseitova, M., Konuspayeva, G., & Jurjanz, S. (2014). Comparison of dairy performances between dromedaries, bactrian and crossbred camels in the conditions of South Kazakhstan. *Emirates Journal of Food and Agriculture*, *26*(4), 366–370. https://doi.org/10.9755/ejfa.v26i4.17271

Ochirkhuyag, B., Chobert, J. M., Dalgalarrondo, M., Choiset, Y., & Haertlé, T. (1997). Characterization of caseins from Mongolian yak, khainak, and bactrian camel. *Le Lait*, *77*(5), 601–613. https://doi.org/10.1051/lait:1997543

Ostersen, S., Foldager, J., & Hermansen, J. E. (1997). Effects of stage of lactation, milk protein genotype and body condition at calving on protein composition and renneting properties of bovine milk. *The Journal of Dairy Research*, *64*(2), 207–219. https://doi.org/10.1017/S0022029996002099

Pauciullo, A., & Erhardt, G. (2015). Molecular characterization of the llamas (Lama glama) casein cluster genes transcripts (CSN1S1, CSN2, CSN1S2, CSN3) and regulatory regions. *PLoS ONE*, *10*(4). https://doi.org/10.1371/journal.pone.0124963

Pauciullo, A., Giambra, I. J., Iannuzzi, L., & Erhardt, G. (2014). The β-casein in camels: Molecular characterization of the CSN2 gene, promoter analysis and genetic variability. *Gene*, *547*(1), 159–168. https://doi.org/10.1016/j.gene.2014.06.055

Restani, P., Gaiaschi, a, Plebani, a, Beretta, B., Cavagni, G., Fiocchi, a, … Galli, C. L. (1999). Cross-reactivity between milk proteins from different animal species. *Clinical and Experimental Allergy : Journal of the British Society for Allergy and Clinical Immunology*, *29*(7), 997–1004. https://doi.org/cea563 [pii]

Rhoads, R. P., Baumgard, L. H., Suagee, J. K., & Sanders, S. R. (2013). Nutritional Interventions to Alleviate the Negative Consequences of Heat Stress. *Advances in Nutrition: An International Review Journal*, *4*(3), 267–276. https://doi.org/10.3945/an.112.003376

Saadaoui, B., Bianchi, L., Henry, C., Miranda, G., Martin, P., & Cebo, C. (2014). Combining proteomic tools to characterize the protein fraction of llama (Lama glama) milk. *Electrophoresis*, *35*(10), 1406–1418. https://doi.org/10.1002/elps.201300383

Saadaoui, B., Henry, C., Khorchani, T., Mars, M., Martin, P., & Cebo, C. (2013). Proteomics of the milk fat globule membrane from Camelus dromedarius. *Proteomics*, *13*(7), 1180–1184. https://doi.org/10.1002/pmic.201200113

Sakono, M., Motomura, K., Maruyama, T., Kamiya, N., & Goto, M. (2011). Alpha casein micelles show not only molecular chaperone-like aggregation inhibition properties but also protein refolding activity from the denatured state. *Biochemical and Biophysical Research Communications*, *404*(1), 494–497. https://doi.org/10.1016/j.bbrc.2010.12.009

Salmen, S. H., Abu-Tarboush, H. M., Al-Saleh, A. A., & Metwalli, A. A. (2012). Amino acids content and electrophoretic profile of camel milk casein from different camel breeds in Saudi Arabia. *Saudi Journal of Biological Sciences*, *19*(2), 177–183. https://doi.org/10.1016/j.sjbs.2011.12.002

Sboui, A., Khorchani, T., Djegham, M., Agrebi, A., Elhatmi, H., & Belhadj, O. (2010). Anti-

diabetic effect of camel milk in alloxan-induced diabetic dogs: A dose-response experiment. *Journal of Animal Physiology and Animal Nutrition*, *94*(4), 540–546. https://doi.org/10.1111/j.1439-0396.2009.00941.x

Sharma, P., Dube, D., Singh, A., Mishra, B., Singh, N., Sinha, M., … Singh, T. P. (2011). Structural basis of recognition of pathogen-associated molecular patterns and inhibition of proinflammatory cytokines by camel peptidoglycan recognition protein. *Journal of Biological Chemistry*, *286*(18), 16208–16217. https://doi.org/10.1074/jbc.M111.228163

Sharma, S., Ramesh, K., Hyder, I., Uniyal, S., Yadav, V. P., Panda, R. P., … Sarkar, M. (2013). Effect of melatonin administration on thyroid hormones, cortisol and expression profile of heat shock proteins in goats (Capra hircus) exposed to heat stress. *Small Ruminant Research*, *112*(1–3), 216–223. https://doi.org/10.1016/j.smallrumres.2012.12.008

Shuiep, E. T. S., Giambra, I. J., El Zubeir, I. E. Y. M., & Erhardt, G. (2013). Biochemical and molecular characterization of polymorphisms of αs1-casein in Sudanese camel (Camelus dromedarius) milk. *International Dairy Journal*, *28*(2), 88–93. https://doi.org/10.1016/j.idairyj.2012.09.002

Sørensen, E. S., & Petersen, T. E. (1993). Purification and characterization of three proteins isolated from the proteose peptone fraction of bovine milk. *Journal of Dairy Research*, *60*, 189–197. https://doi.org/10.1017/S0022029900027503

Sørensen, E. S., Petersen, T. E., & Højrup, P. (1995). Posttranslational modifications of bovine osteopontin: Identification of twenty-eight phosphorylation and three O-glycosylation sites. *Protein Science*, *4*(10), 2040–2049. https://doi.org/10.1002/pro.5560041009

Tagliabracci, V. S., Wiley, S. E., Guo, X., Kinch, L. N., Durrant, E., Wen, J., … Dixon, J. E. (2015). A Single Kinase Generates the Majority of the Secreted Phosphoproteome. *Cell*, *161*(7), 1619–1632. https://doi.org/10.1016/j.cell.2015.05.028

Tydell, C., Yount, N., Tran, D., Yuan, J., & Selsted, M. E. (2002). Isolation, characterization, and antimicrobial properties of bovine oligosaccharide-binding protein. A microbicidal granule protein of eosinophils and neutrophils. *Journal of Biological Chemistry*, *277*(22), 19658–19664. https://doi.org/10.1074/jbc.M200659200

Wangoh, J., Farah, Z., & Puhan, Z. (1998). Iso-electric focusing of camel milk proteins. *International Dairy Journal*, *8*(7), 617–621. https://doi.org/10.1016/S0958-6946(98)00092-2

Yang, Y., Bu, D., Zhao, X., Sun, P., Wang, J., & Zhou, L. (2013). Proteomic analysis of cow, yak, buffalo, goat and camel milk whey proteins: Quantitative differential expression patterns. *Journal of Proteome Research*, *12*(4), 1660–1667. https://doi.org/10.1021/pr301001m

Youcef, N., Saidi, D., Mezemaze, F., El-Mecherfi, K. E., Kaddouri, H., Negaoui, H., … Kheroua, O. (2009). Cross reactivity between dromedary whey proteins and IgG anti bovine α-lactalbumin and anti bovine β-lactoglobulin. *American Journal of Applied Sciences*, *6*(8), 1448–1452. https://doi.org/10.3844/ajassp.2009.1448.1452

Zhao, D. B., Bai, Y. H., & Niu, Y. W. (2015). Composition and characteristics of Chinese Bactrian camel milk. *Small Ruminant Research*. https://doi.org/10.1016/j.smallrumres.2015.04.008

Chapter 3

Alternative splicing events expand molecular diversity of camel CSN1S2 increasing its ability to generate potentially bioactive peptides

Alma Ryskaliyeva[1], Céline Henry[2], Guy Miranda[1], Bernard Faye[3], Gaukhar Konuspayeva[4] and Patrice Martin[1]

[1]INRA, UMR GABI, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France
[2]INRA, MICALIS Institute, Plateforme d'Analyse Protéomique Paris Sud-Ouest (PAPPSO), Université Paris-Saclay, 78350 Jouy-en-Josas, France
[3]CIRAD, UMR SELMET, 34398 Montpellier, France
[4]Al-Farabi Kazakh State University, Biotechnology department, 050040 Almaty, Kazakhstan

Scientific Reports (2019). 9:5243 10.1038/s41598-019-41649-5

Abstract

In a previous study on camel milk from Kazakhstan, we reported the occurrence of two unknown proteins (UP1 and UP2) with different levels of phosphorylation. Here we show that UP1 and UP2 are isoforms of camel α_{s2}-CN (α_{s2}-CNsv1 and α_{s2}-CNsv2, respectively) arising from alternative splicing events. First described as a 178 amino-acids long protein carrying eight phosphate groups, the major camel α_{s2}-CN isoform (called here α_{s2}-CN) has a molecular mass of 21,906 Da. α_{s2}-CNsv1, a rather frequent (35%) isoform displaying a higher molecular mass (+1,033 Da), is present at four phosphorylation levels (8P to 11P). Using cDNA-sequencing, α_{s2}-CNsv1 was shown to be a variant arising from the splicing-in of an in-frame 27-nucleotide sequence encoding the nonapeptide ENSKKTVDM, for which the presence at the genome level was confirmed. α_{s2}-CNsv2, which appeared to be present at 8P to 12P, was shown to include an additional decapeptide (VKAYQIIPNL) revealed by LC-MS/MS, encoded by a 3'-extension of exon 16. Since milk proteins represent a reservoir of biologically active peptides, the molecular diversity generated by differential splicing might increase its content. To evaluate this possibility, we searched for bioactive peptides encrypted in the different camel α_{s2}-CN isoforms, using an *in silico* approach. Several peptides, putatively released from the C-terminal part of camel α_{s2}-CN isoforms after *in silico* digestion by proteases from the digestive tract, were predicted to display anti-bacterial and antihypertensive activities.

Key words: camel milk, protein, casein, polymorphism, post-translational modifications, bioactive peptides

3.1 Introduction

Recently, combining different proteomic approaches, the complexity of camel milk proteins was resolved to provide a detailed characterization of fifty protein molecules belonging to the 9 main milk protein families, including caseins: κ-, α_{s2}-, α_{s1}- and β-CN and two unknown proteins (UP1 and UP2), exhibiting molecular masses around 23,000 Da (Ryskaliyeva et al., 2018). Since UP1 and UP2 co-eluted in RP-HPLC with α_s-CN and displayed different phosphorylation levels, it was tempting to consider that these proteins could originate in CN. However, based on their molecular weight, UP1 and UP2 could be larger isoforms of α_{s2}-CN or smaller isoforms of α_{s1}-CN.

However, the hypothesis of an additional casein in camel milk encoded by a supplementary gene could not be ruled out. Indeed, genes encoding CN are tightly linked on the same chromosome, BTA6 in cattle, CHI6 in goats (Hayes, Petit, Bouniol, & Popescu, 1993; Threadgill & Womack, 1990) and HSA4 in humans (Menon, Chang, Jeffers, Jones, & Ham, 1992). The evolution of the CN gene cluster (Figure 3.1) is postulated to have occurred by a combination of successive intra- and inter-genic exon duplications (Groenen, Dijkhof, Verstege, & van der Poel, 1993; P. Martin, Cebo, & Miranda, 2013; Rijnkels, Elnitski, Miller, & Rosen, 2003). In some mammals, including horses, donkeys, rodents and rabbits, there are two α_{s2}-CN encoding genes differentiating in size (*CSN1S2*-like or *CSN1S2A* and *CSN1S2B*), which may have arisen by a gene-duplication event that has occurred prior to the split of Eutherian mammalian species (Groenen et al., 1993). The second *CSN1S2*-like gene was lost in the Artiodactyla, including the camel, while further divergence occurred in both copies in the other species. In humans, there are also two *CSN1S2* genes albeit no evidence of protein expression exists (Rijnkels et al., 2003).

Alternative splicing is a process by which multiple mRNA isoforms are generated. It is a powerful means to extend protein diversity. Such a process which is another possibility to increase the number of molecular species has been frequently reported to occur, as far as caseins are concerned, especially α_s-CN (Leroux, Mazure, & Martin, 1992; L. Ramunno et al., 2001; Luigi Ramunno et al., 2005), without really knowing whether it is a fortuitous or a scheduled event to expand molecular diversity and functionality of milk proteins. To substantiate the hypothesis according to which UP1 and UP2 might originate in CN and more precisely in α_s-CN, we undertook characterizing more precisely these proteins.

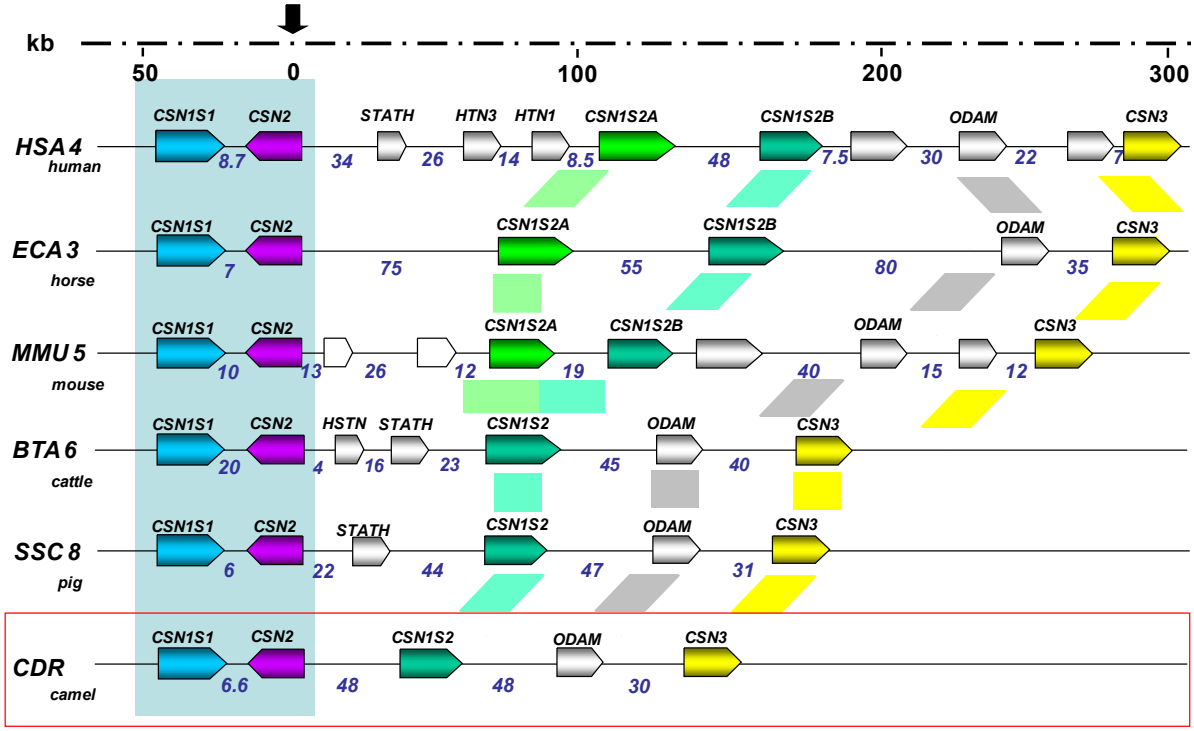


Figure 3.1. Evolution of the casein locus organization. Casein locus organization of human (*Homo sapiens*), horse (*Equus caballus*), mouse (*Mus musculus*), cattle (*Bos taurus*), pig (*Sus scrofa*) and camel (*Camelus dromedarius*) genomes (adapted from Martin, Cebo and Miranda (2013) and Lefèvre *et al.*(2009) with additional genomic information from the NCBI) was compared. Genes are given as colored arrow boxes, showing the orientation of transcription. Putative genes based on similarity are indicated as empty boxes. Intergenic region sizes are given in kb.

In addition to their nutritional value, an increasing number of therapeutic effects and a variety of potential activities (Marcone, Belton, & Fitzgerald, 2017; Mohanty, Mohapatra, Misra, & Sahu, 2016) are attributed to milk proteins as well as to milk-derived bioactive peptides encrypted in milk protein sequences Meisel, 2004). Caseins, and especially α_s-CN, have been shown to be a reservoir of bioactive peptides (Clare & Swaisgood, 2000; Farrell, Malin, Brown, & Mora-Gutierrez, 2009; Meisel, 2004), it is therefore legitimate to wonder whether these so far unknown and putatively derived α_s-CN sequences could be responsible for the occurrence of novel bioactive peptides accounting for the original properties of camel milk. Recent studies have indeed shown that healing properties assigned to camel milk, which is consumed fresh or fermented and traditionally used for the treatment of tuberculosis, gastroenteritis, and allergies, in many countries, are proved (Mati et al., 2017). Whereas there is a substantial literature on bioactive peptides derived from bovine milk proteins (Meisel, 2004) and more or less comprehensive databases of milk bioactive peptides exist (Kumar et

al., 2015; Minkiewicz, Dziuba, Iwaniak, Dziuba, & Darewicz, 2008; Nielsen, Beverly, Qu, & Dallas, 2017; Théolier, Fliss, Jean, & Hammami, 2014), studies aiming at identifying peptides derived from camel milk proteins having potential health-promoting activities are scarce. Investigations mainly focused on caseins (α_{s1}-, β- and κ-CN), and data available to date mostly concern *in vitro* and *in silico* antioxidant, antihypertensive and antimicrobial activities (Erhardt et al., 2016; Mati et al., 2017). Therefore, using an *in silico* approach, we searched for potential biological activities of sequences generated from alternative splicing of primary transcript encoding α_s-CN.

3.2 Methods

3.2.1 Ethics Statements

All animal studies were carried out in compliance with European Community regulations on animal experimentation (European Communities Council Directive 86/609/EEC) and with the authorization of the Kazakh Ministry of Agriculture. Milk sampling was supervised by a veterinarian accredited by the French Ethics National Committee for Experimentation on Living Animals.

3.2.2 Milk Sample Collection and Preparation

Raw milk samples were collected during morning milking on healthy dairy camels belonging to two species: *C. bactrianus* (n=72) and *C. dromedarius* (n=65), and hybrids (n=42) at different lactation stages, ranging between 30 and 90 days postpartum. Camels grazed on four various natural pastures from different regions of Kazakhstan, namely Almaty (AL), Shymkent (SH), Kyzylorda (KZ), and Atyrau (ZKO). Whole-milk samples were centrifuged at 3,000 *g* for 30 min at 4^{o}C (Allegra X-15R, Beckman Coulter, France) to separate fat from skimmed milk. Samples were quickly frozen and stored at -80°C (fat) and -20°C (skimmed milk) until analysis.

3.2.3 Selection of Milk Samples

Thirty milk samples: *C. bactrianus* (n=10), *C. dromedarius* (n=10), and hybrids (n=10)) were selected for LC-ESI-MS analysis from the 179 camel milks collected in a previous study (Ryskaliyeva et al., 2018), based on lactation stages and number of parities

(from 2 to 14). The most representative eight milk samples (*C. bactrianus*, n=3, *C. dromedarius*, n=3, and hybrids, n=2) were analyzed by LC-MS/MS (LTQ-Orbitrap Discovery, Thermo Fisher Scientific) after a tryptic digestion of bands, excised from each track, between 20 and 30 kDa of SDS-PAGE.

3.2.4 RNA Extraction from Milk Fat Globules

Total RNA was extracted from MFG using TRIzol® and TRIzol® LS solutions (Invitrogen, Life Technologies), respectively, according to the original manufacturer's protocol modified as described by Brenaut *et al*. (2012).

3.2.5 First-Strand cDNA Synthesis and PCR Amplification

First-strand cDNA was synthesized from 5 to 10 ng of total RNA primed with oligo(dT)20 and random primers (3:1, vol/vol) using Superscript III reverse transcriptase (Invitrogen Life Technologies Inc., Carlsbad, CA) as described previously (Ryskaliyeva et al., 2018). Primer pairs, purchased from Eurofins (Eurofins genomics, Germany), were designed using published *Camelus* nucleic acid sequences (NCBI, NM_001303566.1 for α_{s1}-CN and NM_001303561.1 for α_{s2}-CN). The forward primers for α_{s1}-CN and α_{s2}-CN amplification were 5'-CTTACCTGCCTTGTGGCTGT-3' (starting from nucleotide 61, located in exon 2 of α_{s1}-CN mRNA) and 5'-TCATTTTTACCTGCCTTTTGGCTGT-3' (starting from nucleotide 71, located in exon 2 of α_{s2}-CN mRNA), respectively. The reverse primers were 5'-GTGGAGGAGAAATTTAGAGCAT-3' (terminating at nucleotide 751 of α_{s1}-CN mRNA located in the last exon) and 5'-CGATTTTCCAGTTGAGCCATA-3' (terminating at nucleotide 692 of α_{s2}-CN mRNA located in the last exon), respectively. Thus, the amplified fragments cover regions of 691 nucleotides for α_{s1}-CN and 622 nucleotides for α_{s2}-CN, including the sequence coding the mature proteins, with genomic reference to the published sequences (NCBI, NM_001303566.1 for α_{s1}-CN and NM_001303561.1 for α_{s2}-CN). Five (two *C. bactrianus,* one *C. dromedarius*, and two hybrids) samples representative of the 30 camel milks analyzed in LC-MS, were selected for amplification of α_{s1}-CN and α_{s2}-CN cDNA by RT-PCR and sequencing. Amplicons were sequenced from both strands with primers used for PCR according to the Sanger method by Eurofins (Eurofins genomics, Germany).

3.2.6 Identification of proteins and validation of peptides by LC-MS/MS Analysis

In order to identify the different α_{s1}-CN and α_{s2}-CN isoforms, mono dimensional electrophoresis (1D SDS-PAGE), followed by trypsin digestion and LC-MS/MS analysis, was used. After a long migration (10 cm) in 1D SDS-PAGE, bands (1.5 mm^3) migrating in the range of 20-30 kDa, were cut on each of the eight gel lanes, and analyzed as described by Henry *et al.* (2015) and Saadaoui *et al.* (2014).

3.2.7 LC-ESI-MS

Fractionation of camel milk proteins and determination of their molecular masses were performed by coupling RP-HPLC to ESI-MS (micrOTOF^{TM} II focus ESI-TOF mass spectrometer; Bruker Daltonics). Twenty µL of skimmed milk samples were clarified by addition of 230 µL of clarification solution 0.1 M bis-Tris buffer pH 8.0, containing 8 M urea, 1.3% trisodium citrate, and 0.3% DTT. Clarified milk samples (25 µL) were directly injected onto a Biodiscovery C5 reverse phase column (300 Å pore size, 3µm, 150x2.1mm; Supelco, France) and analyzed as described by Miranda *et al.* (2004).

3.2.8 *In silico* release of Peptides using PeptideCutter and BIOPEP analyses

Protein sequences of α_{s2}-CN from *Bos taurus* (entry P02663), *Lama glama* (entry A0A0D6DR01) and *Camelus dromedarius* (entry O97944 and new sequences identified in the present study) were selected from the Protein Knowledge Base (UniProtKB, ExPASy Bioinformatics Resource Portal) available at www.uniprot.org. Each sequence was then subjected to *in silico* release of peptides by pepsin (pH 1.3), pepsin+trypsin and pepsin+trypsin+chymotrypsin using "PeptideCutter", a resource available at www.expasy.org. Thereafter, each α_{s2}-CN sequence was entered in the "PeptideCutter". After cutting the sequences, a list of probable peptides with cleavage sites, length and amino acid sequence of peptides was established. BIOPEP analyses were then performed at https://omictools.com/biopep-tool by selecting the available option "Peptide Prediction Software Tools". Peptide Structure Prediction/AHTpin (Kumar et al., 2015) and Antimicrobial Peptide Prediction/ Antimicrobial Peptide Scanner (Veltri, Kamath, & Shehu,

2018) (AMP Scanner Vr.2) sections were used one by one for prediction of the peptides with the sought properties.

3.3 Results and Discussion

3.3.1 What gene(s) do UP1 and UP2 arise from?

The mass accuracy has allowed distinguishing about fifty protein molecules corresponding to isoforms of 9 protein families (κ-CN, WAP, α_{s1}-CN, α-LAC, α_{s2}-CN, PGRP, LPO/CSA, β-CN and γ2-CN) from LC-MS analysis as shown in Figure 3.2. The presence of two unknown proteins UP1 and UP2 with different phosphorylation levels was reported in our previous study[8]. Regarding UP1, molecular masses ranged between 22,939 and 23,179 Da, whereas UP2 masses ranged between 23,046 Da and 23,366 Da (Table 3.1), with successive increments of 80 Da (mass of one phosphate group). The eluting range of these two proteins was between 28.53-37.16 min, within the elution times of α_{s1}- and α_{s2}-CN, which confirms our first hypothesis about their α_s-CN origin. However, UP1 and UP2 masses exceeded the observed mass of the major isoform of α_{s2}-CN with 8P (21,906 Da) by 1,033 Da and 1,300 Da, respectively, and were lighter than the C variant of α_{s1}-CN-6P (25,773 Da) by 2,834 Da and 2,567 Da, respectively (Ryskaliyeva et al., 2018). Even though it was not possible to exclude a splicing event leading to the inclusion of an additional exon sequence in the α_{s2}-CN mRNA, the most probable hypothesis was the occurrence of exon-skipping event(s) affecting α_{s1}-CN mRNA and, leading to the loss of a peptide sequence accounting for a reduction of at least 2,567 Da. A possible scenario was the skipping of exon 3 on the short isoform of α_{s1}-CN C already impacted by a cryptic splice site usage (ΔCAG encoding Q83). The molecular mass of the protein proceeding from such a messenger (23,205 Da) corresponded to the mass of UP2 +160 Da (23,206 Da). However, sequencing cDNA encoding α_{s1}-CN isoforms failed to reveal the existence of a messenger in which exon 3 was lacking. Therefore, the alternative possibility, in other words the α_{s2}-CN avenue, had to be explored.

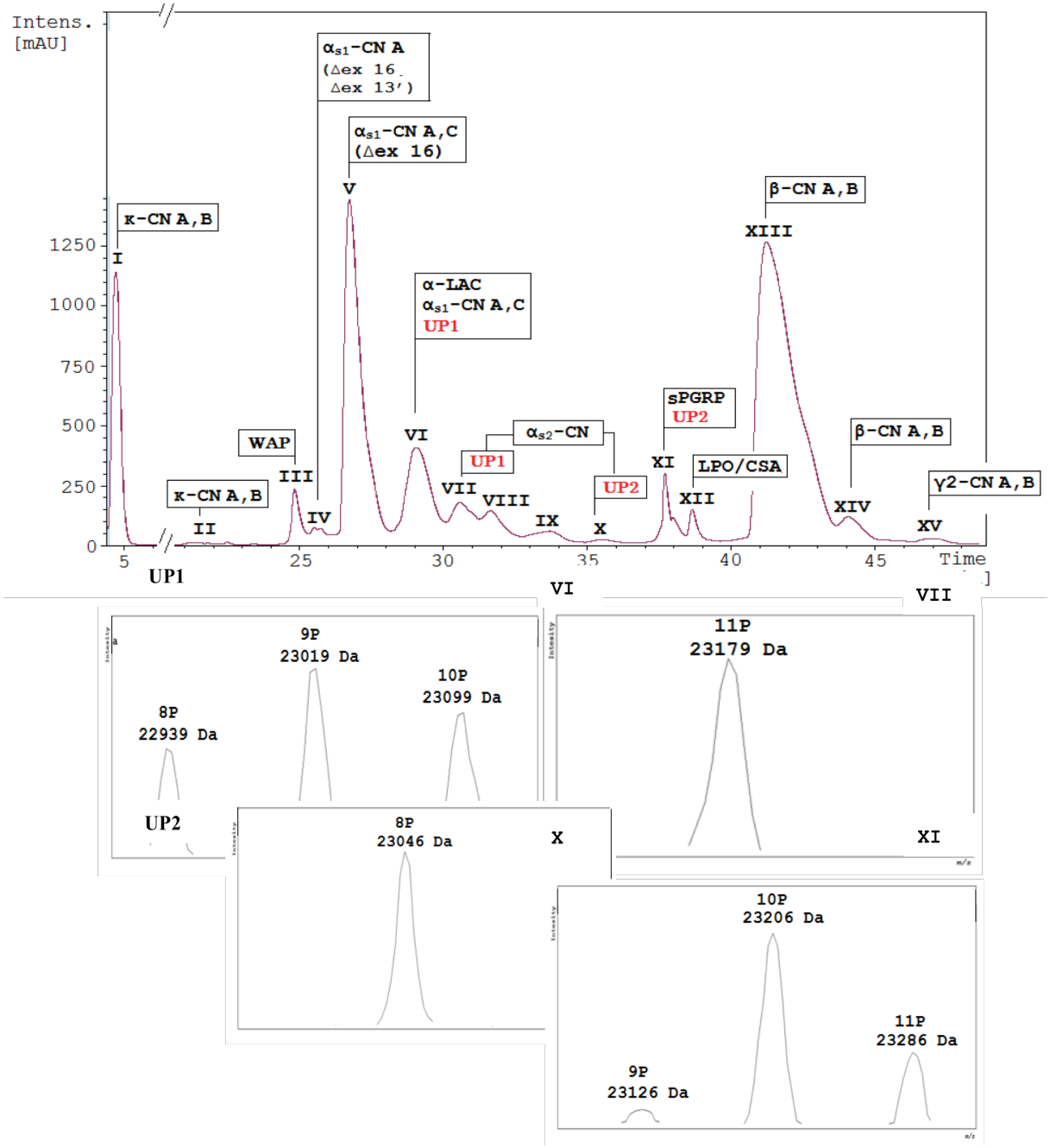


Figure 3.2. LC-ESI-MS profile of dromedary milk proteins. The chromatogram displays the presence of 15 major milk protein fractions labeled from I to XV, with retention times from 4.50 to 48.71 min, respectively. Deconvolution of multicharged ions spectra with emphasis on phosphorylation degrees (P) of two unknown proteins (UP1 and UP2) which are related to chromatographic peaks VI and VII, X and XI respectively.

Table 3.1. Analysis of molecular masses contained in peaks V-XI of dromedary milk sample from the Shymkent region

Peak	Ret.Time, min	Observed M_r, Da	Theoretical M_r, Da	Protein description	UniProt accession	Intensity
V	26,31	24,547	24,547	α_{s1}-CN C -short isoform (Δex 16), 5P, splice variant (ΔQ83)		3,954
		24,561	24,561	α_{s1}-CN A - short isoform (Δex 16), 5P, splice variant (ΔQ83)		4,385
		24,627	24,627	α_{s1}-CN C - short isoform (Δex 16), 6P, splice variant (ΔQ83)		16,348
		24,640	24,641	α_{s1}-CN A - short isoform (Δex 16), 6P, splice variant (ΔQ83)		17,422
		24,675	24,675	α_{s1}-CN C - short isoform (Δex 16), 5P		7,758
		24,689	24,689	α_{s1}-CN A - short isoform (Δex 16), 5P		8,004
		24,722	24,721	α_{s1}-CN A - short isoform (Δex 16), 7P, splice variant (ΔQ83)		4,453
		24,755	**24,755**	**α_{s1}-CN C - short isoform (Δex 16), 6P**	**K7DXB9**	**34,653**
		24,768	**24,769**	**α_{s1}-CN A - short isoform (Δex 16), 6P**	**O97943-2**	**37,452**
		24,835	24,835	α_{s1}-CN C - short isoform (Δex 16), 7P		5,026
		24,849	24,849	α_{s1}-CN A - short isoform (Δex 16), 7P		4,851
VI	28.80	**14,430**	**14,430**	**α-LAC**	**P00710**	**12,948**
		22,939	**n/a***	**UP1**	**n/a**	**2,676**
		23,019	n/a	UP1, +80 Da		2,408
		23,099	n/a	UP1, +160 Da		958
		25,645	25,645	α_{s1}-CN C, 6P, splice variant (ΔQ83)		1,736
		25,659	25,659	α_{s1}-CN A, 6P, splice variant (ΔQ83)		1,057
		25,693	25,693	α_{s1}-CN C, 5P		916
		25,772	**25,773**	**α_{s1}-CN C, 6P**		**5,014**
		25,787	25,787	α_{s1}-CN A, 6P	O97943-1	1,509
VII	30.07	21,826	21,825	α_{s2}-CN, 7P		709
		21,906	**21,905**	**α_{s2}-CN, 8P**	**O97944**	**4,222**
		21,985	21,986	α_{s2}-CN, 9P		289
		23,179	n/a	UP1, +240 Da		1,430
VIII	31.26	21,986	21,985	α_{s2}-CN, 9P	O97944	866
		22,066	22,065	α_{s2}-CN, 10P		3,682
IX	33.04	22,066	22,065	α_{s2}-CN, 10P		120
		22,146	22,145	α_{s2}-CN, 11P		1,408
X	34.85	22,226	22,225	α_{s2}-CN, 12P		806
		23,046	n/a	UP2	n/a	295
XI	37.15	**19,143**	**19,143**	**PGRP**	**Q9GK12**	**3,659**
		23,126	n/a	UP2, +80 Da		150
		23,206	**n/a**	**UP2, +160 Da**		**1,162**
		23,286	n/a	UP2, +240 Da		940

n/a - not applicable

3.3.2 UP1 and UP2: new camel α_{s2}-CN splicing variants

Amplification of camel α_{s2}-CN cDNA revealed the presence of a major PCR fragment (*ca.* 620 bp) and several minor PCR products differing in size between *ca.* 670 bp and 710 bp (supplementary data S1). Sequencing of PCR fragments generated two different nucleotide sequences: first identical from the forward primer to nucleotide 359, and then overlapping and shifted by 27 nucleotides (Figure 3.3). The main sequence corresponded to the 193-aa α_{s2}-CN (including the signal peptide) reported by Kappeler *et al.* (1998). The second sequence, with weaker signals, showed the insertion of the following sequence: GAA AAT TCA AAA AAG ACT GTT GAT ATG, between exons 12' and 14. Thus, this insertion introduced an additional peptide sequence (ENSKKTVDM), identical to the aa sequence encoded by exon 13 in the bovine *CSN1S2* gene (Figure 3.4). The level of exon 13 conservation in both species appeared to be extremely high. This exon is also present in the predicted sequence of the *CSN1S2* gene from the *Camelus ferus* genome (NCBI Reference Sequence: XP_014418048.1) and the lama gene transcript (GenBank: LK999989.1) with two point mutations. The first mutation concerning the fourth codon (AAA=>AAT) is silent and the second one, that is a missense mutation, regards the last codon (ACG => ATG), leading to T => M substitution (Pauciullo & Erhardt, 2015). Exon 13 is present in one of the two copies of the *CSN1S2* gene of most mammalian species. In mice, rats and rabbits the aa sequence encoded by this exon is present in CSN1S2-like or CSN1S2A) protein but not in CSN1S2B (Rijnkels, 2002). The insertion of this sequence leads to the increasing of the molecular mass of α_{s2}-CN by 1,033 Da, exactly the mass difference observed between α_{s2}-CN-8P and UP1.

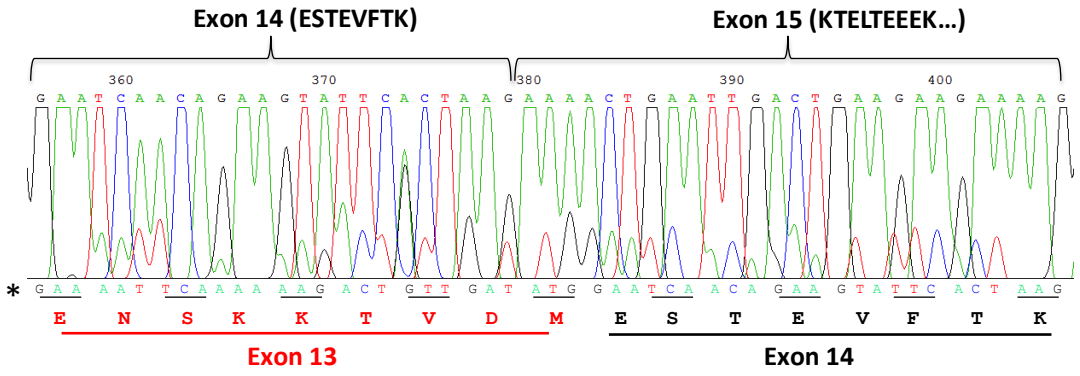


Figure 3.3 Sequence of *C. dromedarius* α_{s2}-CN cDNA spanning exons 14 and 15 (main sequence). A secondary sequence (*) identified by manual reading of overlapping weak signals is given below the main sequence, showing the existence of transcripts, in which exon 13 is included. The corresponding aa sequence is given below. cDNA sequences encoding CSN1S2sv1 were submitted to NCBI Genbank with the following submission IDs: BankIt2160486 Seq1 MK077758 (*C. bactrianus*) and BankIt2160533 Seq1 MK077759 (*C. dromedarius*).

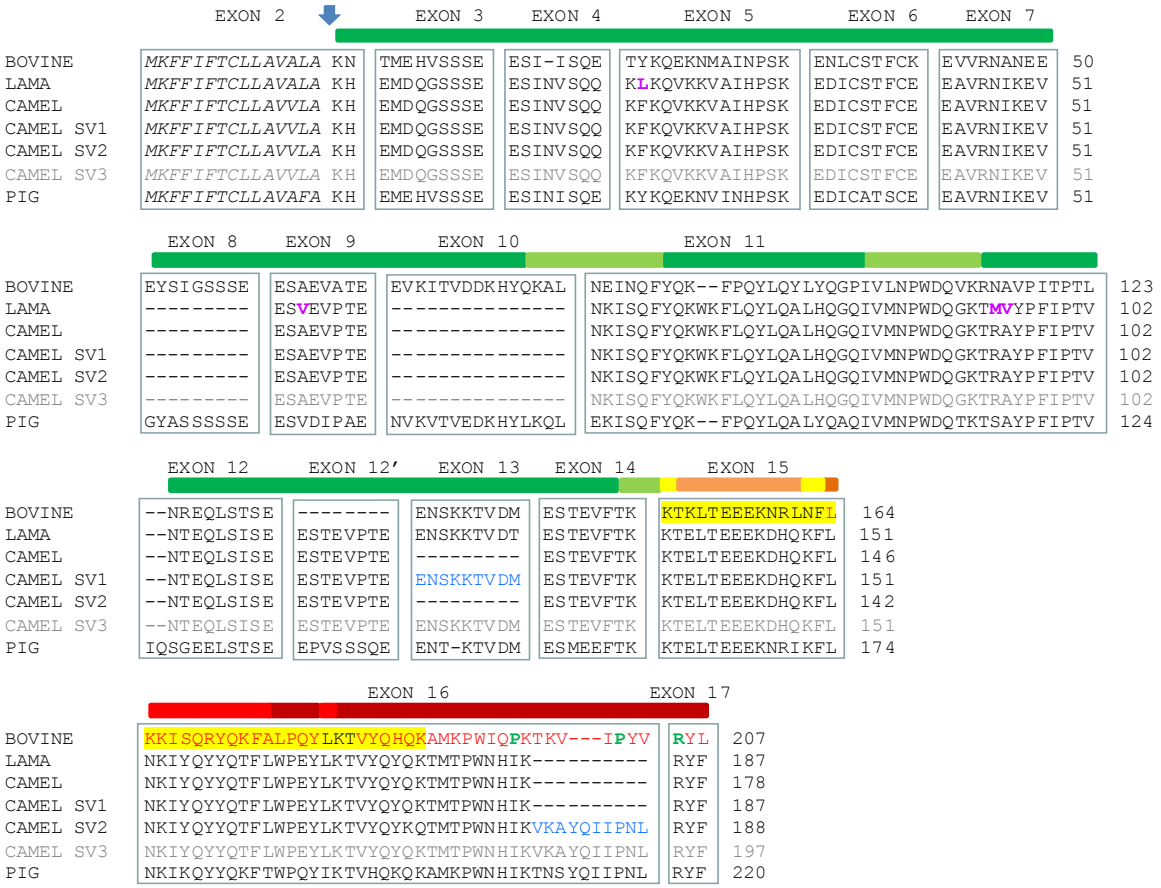


Figure 3.4. Multiple alignment of α_{s2}-CN protein sequences from different Artiodactyls species. *Bos taurus* (M16644), *C. dromedarius* (O97944 and splicing variants identified in the present study), *Lama glama* (A0A0D6DR01), and *Sus scrofa* (X54975) protein sequences are compared. Camel α_{s2}-CN putative isoform (α_{s2}-CNsv3) comprising both additional sequences, encoded by exon 13 and exon16 extension, is in grey. Sequences are split into blocks of amino acid residues to visualize the exon modular structure of the protein as deduced from known splice junctions of the bovine gene (Koczan, Hobom, & Seyfert, 1991). Exon numbering (top of blocks) is that of the bovine gene taken as reference for Artiodactyls. Amino-acid sequences characterizing UP1 (α_{s2}-CNsv1) and UP2 (α_{s2}-CNsv2) encoded by exon 13 and the extension of exon 16, respectively, are given in blue. Italics indicate the signal peptides, for which the vertical blue arrow points out the cleavage site. Dashes indicate missing aa residues. Amino acid mutations distinguishing camel and lama α_{s2}-CN are in fuchsia. The highest sequence antimicrobial peptide density is indicated by red on a heat map above the bovine protein sequence. The regions of Bioactive peptides encrypted in bovine α_{s2}-CN f(150-188) with antibacterial activities reported by Zucht *et al.* (1995) are highlighted in yellow, while two antibacterial domains f(164-179) and f(183-207) described by Recio and Visser (1999) are indicated in red. Amino acid residues increasing significantly antibacterial potency are in green. Full-length mature CSN1S2sv1 and CSN1S2sv2 aa sequences were submitted to Expasy UniProtKB database as splicing variants of *C. dromedarius* CSN1S2 with the following submission IDs: SPIN200013828 and SPIN200013835, respectively.

A deep and comprehensive analysis of the dromedary camel *CSN1S2* gene sequence available in GenBank (gi|742343530|ref|NW_011591251.1|), overlaying exon 12' (ESTEVPTE) to exon 14 (ESTEVFTK) allowed identifying a 27-nucleotide sequence corresponding to exon 13 (Figure 3.5). This sequence is flanked with consensus splice sites at the beginning (GTG/AAG) and end (polypyrimidine tract followed by XAG) of intron sequences. Therefore, this exon is included or not during the course of camel α_{s2}-CN pre-mRNA processing. This is possibly due to the weakness (presence of purine in the polypyrimidine tract at the 3'-end of the upstream intron) of the acceptor splice sequence. The short transcript (without exon 13) encodes the 193 aa residues (including the signal peptide) described by Kappeler *et al.*(1998) and the long transcript (with exon 13) codes for UP1 (202 aa including signal peptide). The mature protein corresponding to UP1 is named thereafter α_{s2}-CNsv1.

To confirm such an additional exon 13 hypothesis, detection of α_{s2}-CN peptides after trypsin action was performed using liquid chromatography coupled to tandem mass spectrometry (LC-MS/MS). A tryptic peptide composed of 12 aa residues TVDMESTEVFTK (Figure 3.6), identified through the *Bubalus bubalis* α_{s2}-CN sequence (UniProt KB accession number E9NZN2), was attributed to two coherent arranged sequences (ENSKKTVDM and ESTEVFTK) encoded by exons 13 and 14, respectively. The sequence is identical to that of the *Bos taurus* (UniProt KB accession number P02663). The presence of a TVDM peptide sequence confirmed the existence of transcripts having included exon 13 during the course of pre-mRNA processing. Therefore, the existence of an exon 13 alternatively spliced in the camel *CSN1S2* gene was successfully confirmed both at the protein (LC-MS and LC-MS/MS) and at the nucleotide (cDNA sequencing and genome data) levels. The same cDNA sequences encoding α_{s2}-CN with and without a 27-nucleotide additional sequence (exon 13) were found in all individual samples analyzed, including *C. bactrianus, C. dromedarius*, and hybrids.

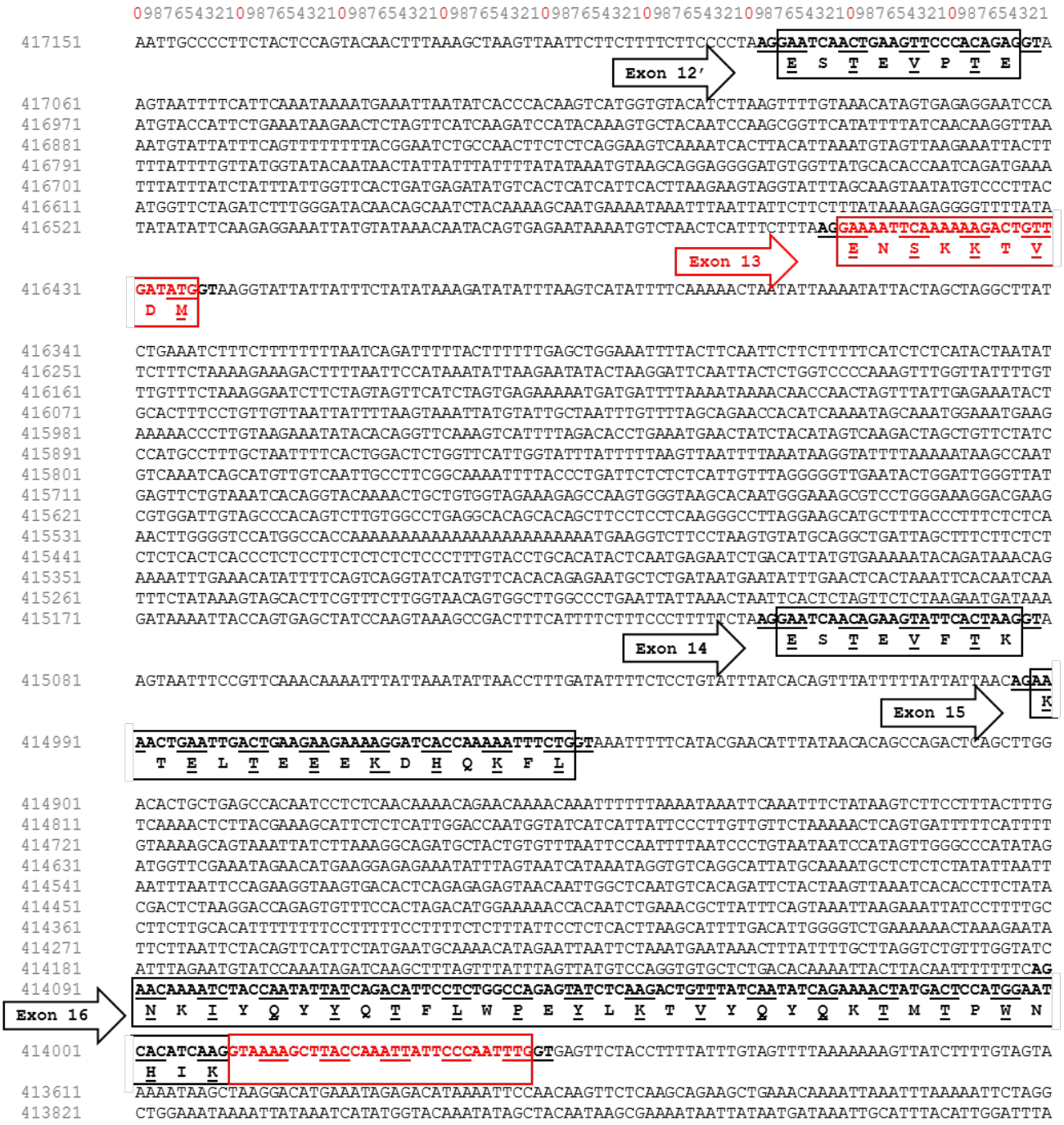


Figure 3.5. Nucleotide sequence view (from 417151 to 413731) of *C. dromedarius* (breed Arabia) taken from the unplaced genomic scaffold of CSN1S2 (LOC105090951). Already known exons 12', 14, 15 and 16 are given in black, and additional exon 13 and extension of 30 additional nucleotides of exon 16 are in red. Exon subdivisions are boxed with amino acid sequences beneath. Intron donor and acceptor splice sites are underlined.

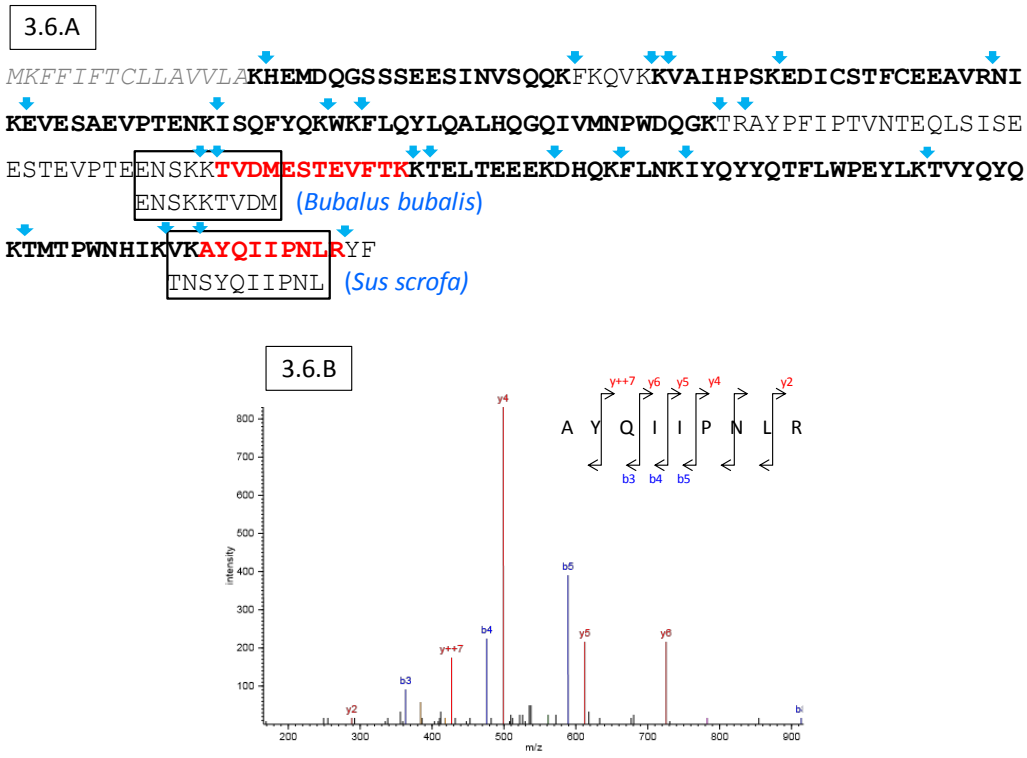


Figure 3.6. Identification and characterization of UP1 and UP2 as splicing variants of α_{s2}-CN by LC-MS/MS analysis. A. Camel α_{s2}-CN full-length sequence is given and its coverage (81%) from peptides identified by LC-MS/MS analysis is in bold. Blue arrows indicate a cleavage of camel α_{s2}-CN by trypsin. Tryptic peptides indicating the presence of exon 13 and extension of exon 16 are in red. Camel α_{s2}-CN peptide sequences encoded by exon 13 and by the extension of exon 16 matching with *Bubalus bubalis* (UniProt KB accession number E9NZN2) and *Sus scrofa* (UniProt KB accession number P39036) are framed. The signal peptide is in italics and in grey. B. Validation of the additional peptide sequence (AYQIIPNLR) with five and three ions from the "y" (including y7 double charged: y++7) and "b" series, respectively.

Concerning the second unknown protein detected (UP2) that showed molecular masses comprised between 23,046 Da and 23,286 Da with n and n+3 phosphate groups, in LC-ESI-MS, the mass difference observed was 1,140 Da, relative to the 8P-11P α_{s2}-CN protein reported by Kappeler and co-workers (1998). LC-MS/MS analysis revealed the occurrence of a 9 aa-long peptide (AYQIIPNLR) matching with the C-terminal sequence of *Sus scrofa* α_{s2}-CN (NP_001004030.1), strongly suggesting that mRNA described by Kappeler *et al.* (1998) was in fact the result of a cryptic splice site usage occurring in the antepenultimate exon of the camel *CSN1S2* gene (Figure 3.6).

Examination of intron sequence downstream of exon 16 (Figure 3.5) highlighted a 30-nucleotide segment: GTA AAA GCT TAC CAA ATT ATT CCC AAT TTG encoding 10 aa

residues (VKAYQIIPNL). The intron donor splice site following the previously considered ending sequence of exon 16 CACATCAAG│GTAAA was recognized by the spliceosome machinery to generate the protein described by Kappeler *et al.* (1998). Alternatively, a second downstream intron donor splice site (CCC AAT TTG│GTGAG), which also fulfils all requirements of a splicing recognition signal, may also be used as well (Figure 3.5). As a result, this alternative splicing event is responsible for the occurrence of two mature peptide chains, the first one made of 178 aa residues (21,906 Da with 8P), and the second one 10 aa residues longer (23,046 Da with 8P). The mature protein corresponding to UP2 is named thereafter α_{s2}-CNsv2. Interestingly, the 10 aa residue peptide (VKAYQIIPNL) included in the C-terminal part of the camel protein due to this alternative splicing event was highly similar with the porcine (TNSYQIIPNL) and donkey (TNSYQIIPVL) α_{s2}-CN sequences. Recently a shorter α_{s2}-CN isoform, in which a deletion of the heptapeptide YQIIPVL, was reported in donkey milk (Cunsolo et al., 2017; Saletti et al., 2012).

3.3.3 Cross-species comparison of the gene encoding α_{s2}-CN and primary transcript maturation

Comparative analysis of camel *CSN1S2* gene organization with orthologous bovine and pig genes is illustrated in Figure 3.7. The first camel α_{s2}-CN sequence published by Kappeler *et al.* (1998) lacks three peptide sequences encoded in cattle by exons 8 (EYSIGSSSE), 10 (EVKITVDDKHYQKAL), and 13 (ENSKKTVDM) composed of 27, 45 and 27 nucleotides, respectively. By contrast, exon 12' that encodes in camel and lama a peptide of 8 aa residues (ESTEVPTE), was believed to be missing in the bovine counterpart, while it was present in the porcine genome, coding for the EPVSSSQE peptide. Surprisingly, we succeed in finding a putative exon 12', encoding the octapeptide VSANSSQE, in intron 12 of the bovine gene. However, the downstream GTAAG donor splice site flanking this putative exon 12' is mutated in GCAAG, apparently preventing its recognition as such as an exon. On the other hand, we failed to find a putative exon 8 in intron 7 of the camel gene. Exon 10 is present both in bovine and pig *CSN1S2* genes. In addition, it is also present in intron 9 of the camel gene, being 9 nucleotides longer than in the other species (Figure 3.7), and bounded upstream and downstream by canonical intron consensus sequences. However, even though it seems to be perfectly eligible for splicing, we did not find any transcript nucleotide sequence, nor tryptic peptides at the protein level, signing its presence in multiple mRNA encoding α_{s2}-CN. By contrast, as demonstrated in the present study, exon 13 was actually present in some

camel *CSN1S2* transcripts, as well as the peptide sequence it is coding for in isoform α_{s2}-CNsv1. Finally, the camel *CSN1S2* gene, just as its lama counterpart (Pauciullo & Erhardt, 2015), is made up of at least 17 exons, since we have no objective demonstration of the usage of exon 10, whereas its bovine and porcine counterparts are made up of 18 and 19 exons, respectively. Since a further exon sequence (exon 7') occurs in the Equidaes *CSN1S2B* gene (not in *CSN1S2*-like *A*), we can hypothesize that the *CSN1S2* gene can comprise up to 20 exons with different combinatory splicing schemes across species. Interestingly, sequence alignments revealed that within the bovine intron 7, as well as in camels and pigs, the sequence corresponding to horse and donkey exon 7' is partially deleted.

Genomic and mRNA analyses carried out previously demonstrated that deletions of aa residues in CN across species occurred essentially by exon skipping during the processing of the primary transcripts (Johnsen, Rasmussen, Petersen, & Berglund, 1995; Leroux et al., 1992; P. Martin et al., 2013; Patrice Martin & Leroux, 1992; Matéos et al., 2009; Pauciullo & Erhardt, 2015). This event, leading to a shortening of the peptide chain length, is caused by weaknesses in the consensus sequences, either at the 5' and/or 3' splice junctions or at the branch point, or both (P. Martin et al., 2013). Therefore, alternative splicing has to be regarded as a frequent event, mainly in α_s-CN encoding genes, for which the coding region is divided into many short exons. Usage of cryptic splice sites is also responsible for the occurrence of multiple transcripts and finally for generating a protein molecular diversity. For example, the peptide sequence (VKAYQIIPNL) encoded by the "extension" of 30 nucleotides at the 3' end of exon 16, not previously detected in camel nor in lama α_{s2}-CN, was shown here to be alternatively included in camel *CSN1S2* transcripts. Extending the comparison to other species including ruminants, pigs and Equidaes, we show that the true donor splice site (GTGAG…) defining the end of exon 16 and common to the considered species (Figure 3.8), is located 30 nt downstream of that preferentially used in Camelidaes. In other words, the isoform corresponding to UP2/α_{s2}-CNsv2 is the genuine protein, whereas the isoform first described (Kappeler et al., 1998) corresponds to the protein arising from the usage of a cryptic splice site internal to an exon.

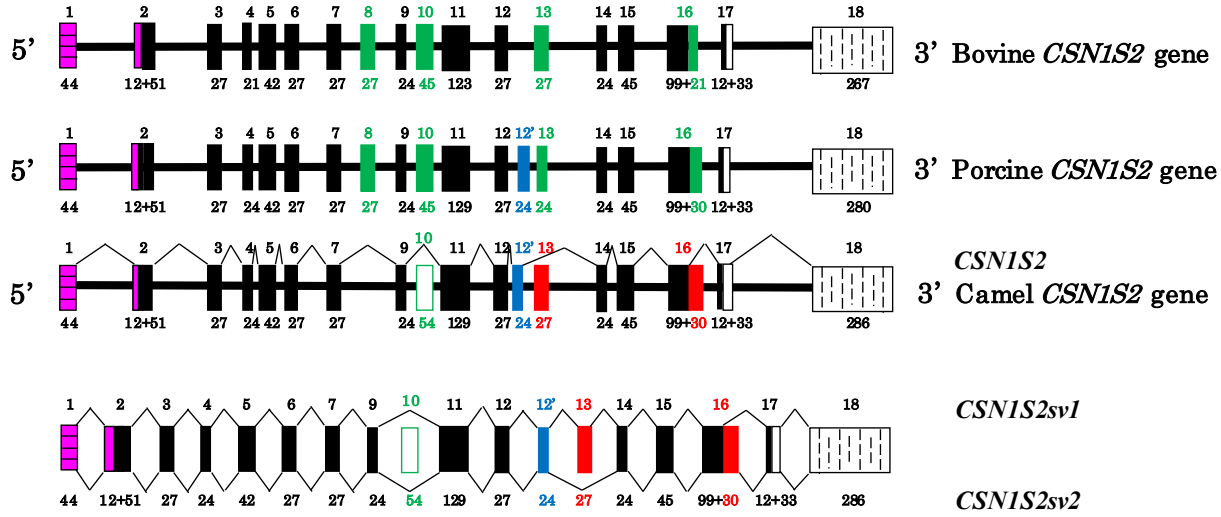


Figure 3.7. Structural organization of the bovine, porcine and camel *CSN1S2* transcription units and splicing patterns for camel (*CSN1S2, CSN1S2sv1* and *CSN1S2sv2*). *CSN1S2* corresponds to the splicing pattern characterized by Kappeler *et al.* (1998). Solid bars represent introns, and exons are depicted by blocks: 5' UTR and noncoding sequence are given in pink, leader peptide and coding frame are in black, exons absent from the camel protein are in green, exons absent from the bovine protein are in blue, exons found in our study are in red, and 3' UTR in white. Exons and exon sequences present in bovine and porcine *CSN1S2* but which were absent from the camel until now are highlighted in green, while exons present in the camel and pig are in blue. Exon 13 and the extension of exon 16 identified in this study are in red. Exon numbering (referring to bovine) and sizes (in bp) are indicated at the top, and at the bottom of the structures, respectively.

The combination of both splicing events such as exon skipping and cryptic splice site usage generates more transcript isoforms in the same species and is responsible for the differences across species in the aa sequences of α_{s2}-CN. However, regarding α_{s2}-CN in camels we were not able to detect any transcript in which both exon 13 and the extension of exon 16 were present (α_{s2}-CNsv3). That does not mean that this structure does not exist, even though the protein corresponding to both events was not detected in LC-MS profiling. Therefore, given that such an isoform is putatively present at a very low level, cloning PCR fragments and screening of a significant number of clones should probably make it possible to identify such a transcript.

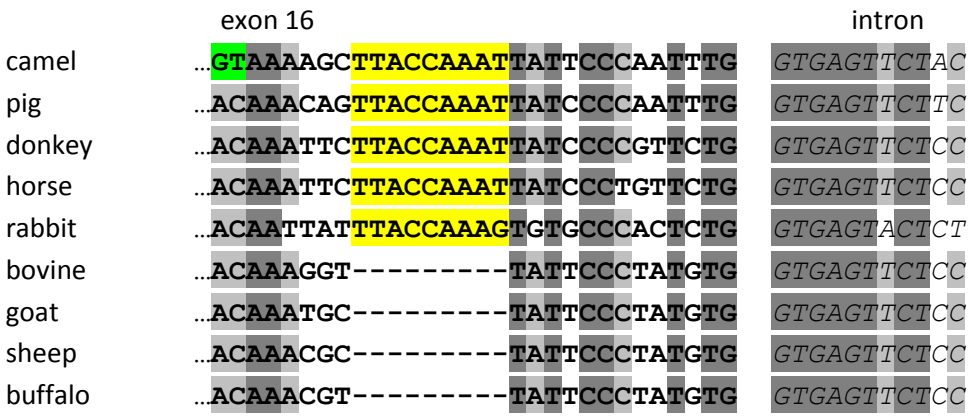


Figure 3.8. Alignment of nucleotide sequences of exon 16 3'-end and downstream intron across nine species. Accession numbers of different species are: camel (NCBI Gene ID: 105090951), pig (NCBI Gene ID: 445515), donkey (NCBI Gene ID: 106835119), horse (NCBI Gene ID: 100327035), rabbit (NCBI Gene ID: 100009288), bovine (NCBI Gene ID: 282209), goat (NCBI Gene ID: 100861229), sheep (NCBI Gene ID: 443383), and buffalo (NCBI Gene ID: 102395699). Exon sequences are in bold, intron sequences are in italics. Perfectly conserved nucleotides are dark-grey shaded. Nucleotides identical in more than eight animal species are light-grey shaded. Dashes in ruminants indicate missing nucleotides that are highlighted in yellow in the other species. The dinucleotide GT, highlighted in green in the camel sequence, generates the preferential site of splicing occurring within exon 16 that leads to the main α_{s2}-CN isoform first described by Kappeler *et al*. (1998).

3.3.4 Phosphorylation level enhances camel α_{s2}-CN isoform complexity

The non-phosphorylated peptide chain of the mature α_{s2}-CN protein, which comprises 178 aa residues, yields a molecular weight of 21,266 Da (Kappeler et al., 1998). Compared with other Ca-sensitive CNs, α_{s2}-CN is the most phosphorylated with 12 potential phosphorylation sites and it is therefore likely to be the major transporter of Ca-phosphate.

Structural characterization of the α_{s2}-CN fraction and relevant mRNA analyses has demonstrated that camel α_{s2}-CN should be theoretically present in milk as a mixture of at least 18 isoforms derived from three mature peptide chains comprising 178 (α_{s2}-CN), 187 (α_{s2}-CNsv1, UP1) and 188 (α_{s2}-CNsv2, UP2) aa residues originating from alternative splicing phenomena (Figure 3.4). Each splicing variant should display six phosphorylation levels ranging between 7 and 12 P groups. Based on LC-ESI-MS data, we identified 14 phosphorylation isoforms. Surprisingly, even though an additional peptide sequence does not provide further phosphorylation sites, the predominant phosphorylation level of each peptide

isoform was not the same: 8P for α_{s2}-CN, 8P for α_{s2}-CNsv1, and 10P for α_{s2}-CNsv2. The addition of 10 aa residues in the C-terminal part of α_{s2}-CNsv2 might induce conformational changes in the protein facilitating the modification of definite phosphorylable sites. Multiple non-allelic variants produced from at least three different mRNA were shown to occur in all thirty Kazakh individuals analyzed, apparently indicating a stabilized mechanism for the production of protein isoforms of different lengths, structures and possibly biological activities.

With 11 potentially phosphorylated aa residues matching the S/T-X-A motif, camel α_{s2}-CN displays the highest phosphorylation level, as mentioned by Ryskaliyeva *et al.*[8]. To reach such a phosphorylation level, besides the nine SerP, two putative Threonine residues (T118 and T132) should be phosphorylated. However, in all the Kazakh milk samples analyzed in LC-ESI-MS we found α_{s2}-CN with up to 12 P groups. This means that at least another S/T residue that does not match the canonical sequence recognized by the mammary kinase(s), is potentially phosphorylated. According to Allende *et al*. (1995) the sequence S/T-X-X-A is in agreement with the minimum requirements for phosphorylation by the CN-kinase II (CK2). In this regard, it is critical to highlight that the A residue in this site, usually E or D, can be replaced by SerP or ThrP. Two T residues, namely T39 and T129 in the camel α_{s2}-CN fully meet the requirements of the above-mentioned motif and might be phosphorylated. Such an event is the only possible hypothesis to reach 12P for camel α_{s2}-CN. Since these two kinases are very likely secreted, the idea that phosphorylation at T39/T129 may occur in the extracellular environment cannot be excluded. This warrants further investigation. Fam20C, which is very likely the major secretory pathway protein kinase (Tagliabracci et al., 2015), might be responsible for the phosphorylation of S and T residues within the S/T-X-A motif, whereas a CK2-type kinase might be responsible for phosphorylation of the T residue within an S/T-X-X-A motif. This was in agreement with the hypothesis put forward by Bijl *et al*. (2014) and Fang *et al*. (2016), who suggest, from phenotypic correlations and hierarchical clustering, the existence of at least two regulatory systems for phosphorylation of α_s-CN. Interestingly, twelve phosphorylation sites were also predicted in llama α_{s2}-CN (Pauciullo & Erhardt, 2015), including two Threonine residues at T118 (instead of T114 as erroneously mentioned) and T141 (also T141 in camel α_{s2}-CNsv1). Phosphorylation sites matching the S/T-X-A motif in llama α_{s2}-CN are actually 12. Indeed, S122 (llama's numbering) that has been predicted as phosphorylated (Pauciullo & Erhardt, 2015) does not meet the criteria required by the S/T-X-A consensus motif and cannot be phosphorylated. By contrast, T128,

which is substituted by a methionine residue (M) in the camel α_{s2}-CNsv1, is potentially phosphorylated provided S130 has been phosphorylated before. On the contrary, sites potentially phosphorylated by a second kinase (CK2-type) identified in the camel sequence are also present in the llama sequence and therefore the phosphorylation level that could be reached in this species is potentially13P.

3.3.5 Alternate splicing isoforms of camel α_{s2}-CN increase its ability to generate potential bioactive peptides

A growing number of genes encoding milk proteins displays complex patterns of splicing, thus increasing their coding capacity to generate an extreme protein isoform diversity from a single gene. It is well established that milk proteins represent a reservoir of biologically active peptides (Meisel, 2004; Nagpal et al., 2011; Weimann, Meisel, & Erhardt, 2009), capable of modulating different functions. Therefore, beside genetic polymorphisms, the molecular diversity generated by differential splicing mechanisms can increase its content.

To evaluate this possibility, we undertook to search for bioactive peptides encrypted in the different camel α_{s2}-CN isoforms, using an *in silico* approach. Since alternative splicing events impact the C-terminal part of the molecule (f(150-197)) which seems, in addition, to be the most accessible domain of the bovine protein (Farrell et al., 2009; Tauzin, Miclo, Roth, Mollé, & Gaillard, 2003), we therefore focused our attention on this region. Previous studies performed on the bovine α_{s2}-CN have demonstrated that this casein is the least accessible in the micelles and that a limited number of tryptic peptides were released from its C-terminal part (Diaz, Gouldsworthy, & Leaver, 1996; Gagnaire & Léonil, 1998); of which some were subsequently shown to display antibacterial properties (Farrell et al., 2009). The first antibacterial peptide isolated from bovine α_{s2}-CN (f(150-188) of the mature protein), inhibiting the growth of *Escherichia coli* and *Staphylococcus carnosus*, was called casocidin-I (Zucht et al., 1995). Two distinct antibacterial domains f(164-179) and f(183-207), also located in the C-terminal part of the molecule, were subsequently isolated from a peptic hydrolysate of bovine α_{s2}-CN (Recio & Visser, 1999). It is worth noting that in our prediction analyses (Figure 3.9), the bovine peptide f(164-179) displays a rather high probability (0.685) to have an antimicrobial (AMP) activity; whereas peptide f(183-207) for which a probability of 0.312 was found, would not have such an activity. In contrast, peptide f(192-207) is by far the one with the highest probability (0.915) to exhibit an AMP activity.

The picture is less positive with regard to the corresponding camel sequences, since peptides f(179-197) and f(179-187), according to the splicing variant (α_{s2}-CN sv2 and α_{s2}-CN sv1, respectively), as well as f(151-166) from α_{s2}-CN sv1, compared with the bovine α_{s2}-CN f(164-179), gave more contrasted results (Figure 3.9). Given the magnitude of the splicing events occurring in the camel α_{s2}-CN pre-mRNA, it is not surprising that it would impact biological properties of α_{s2}-CN C-terminal peptides, including antimicrobial activity, since several aa residues of this region were shown to be essential regarding AMP activity (Recio & Visser, 1999; Zucht et al., 1995). Indeed, the importance of specific amino acids (P and R residues) at the C-terminus of the bovine milk-derived α_{s2}-CN f(183-207) peptide for its antibacterial activity against the food-borne pathogens *Listeria monocytogenes* and *Cronbacter sakazakii*, was recently demonstrated (Alvarez-Ordóñez et al., 2013). Nevertheless, this *in silico* screening remains a predictive approach, aimed at identifying sequences that would be potentially bioactive. It is therefore necessary to confirm experimentally, and possible discordances may occur between *in silico* and *in vitro* results. It is not because the sequence of a peptide is predicted as potentially bioactive that it will be actually active *in vitro* and if it is active *in vitro*, this does not mean that even though it will be active *in vivo*. McCann *et al.* (2005) identified 5 peptides from chymosin digests of a bovine sodium caseinate, all being once again from the C-terminal end of α_{s2}-CN, including f(164–207), f(175–207) and f(181-207), and showing *in vitro* antibacterial activity against *Listeria innocua*. However, they stressed that it was not excluded that these cationic peptides may lose their antibacterial activity *in vivo*. From all these studies it appears, nevertheless, that the C-terminal part of α_{s2}-CN was predicted to yield peptides with defensin-like activity, which may aid the immune system in fighting bacteria (Farrell et al., 2009).

Interestingly, further bioactive peptides with different properties such as AHT (Anti Hyper Tensive) activity were identified from camel α_{s2}-CN (Figure 3.9). Indeed, according to the splicing patterns, including or not exon 16 extension, two peptide sequences (KTMTPWNHIKRYF and KTMTPWNHIKVKAYQIIPNLRYF) occur within the C-terminal part of the molecule (Figure 3.4), thus giving rise to different peptides after digestion by proteolytic enzymes from the digestive tract, including pepsin, trypsin and chymotrypsin (supplementary data S2). Several peptides, related to the inserted VKAYQIIPNL decapeptide characterizing camel α_{s2}-CNsv2, were *in silico* identified as AHT peptides involved in the angiotensin I-converting enzyme (ACE) inhibitory activity, with SVM (Support Vector Machine) scores >1 (Figure 3.9). Two ACE-inhibitory dipeptides (f(185-186): VK and f(187-

188): AY) were found exclusively in camel α_{s2}-CNsv2 (and in the putative camel α_{s2}-CNsv3). Interestingly, the AY dipeptide was also found in the B variant of the camel α_{s1}-CN (Erhardt et al., 2016). A novel ACE inhibitory peptide (YQK) exhibiting an IC_{50} of 11.1 µM was recently isolated from a pepsin and trypsin hydrolysate of bovine α_{s2}-CN (Xue et al., 2018). An oral administration, using a rodent hypertensive model, revealed a significant decrease of systolic blood pressure, thus demonstrating its AHT effects. Such a tripeptide sequence also occurs in the C-terminal part of the camel α_{s2}-CN.

Figure 3.9. *In silico* analyses of α_{s2}-CN peptides for antimicrobial (yellow) and antihypertensive (green) activities

			Anti Microbial Peptide (AMP)		Anti Hyper Tensive (AHT)	
Peptide location*	SeqID	Sequence	Probability **	Classification	SVM score	Prediction
f(187-197)	Camel α_{s2}-CNsv2	AYQIIPNLRYF	0.444	Non-AMP	2.05	AHT
f(187-195)	Camel α_{s2}-CNsv2	AYQIIPNLR	0.428	Non-AMP	1.40	AHT
f(187-194)	Camel α_{s2}-CNsv2	AYQIIPNL	0.378	Non-AMP	1.26	AHT
f(185-194)	Camel α_{s2}-CNsv2	VKAYQIIPNL	0.061	Non-AMP	1.75	AHT
f(169-175)	**Camel α_{s2}-CN**	**TVYQYQK**	**0.520**	**AMP**	0.37	AHT
f(176-184)	Camel α_{s2}-CN	TMTPWNHIK	0.140	Non-AMP	0.54	AHT
sv1 f(179-187)	**Camel α_{s2}-CNsv1**	**PWNHIKRYF**	**0.622**	**AMP**	1.05	AHT
sv2 f(179-197)	Camel α_{s2}-CNsv2	PWNHIKVKAYQIIPNLRYF	0.367	Non-AMP		
sv1 f(151-166)	Camel α_{s2}-CNsv1	LNKIYQYYQTFLWPEY	0.092	Non-AMP		
f(164-179)	**Bovine α_{s2}-CN**	**LKKISQRYQKFALPQY**	**0,685**	**AMP**		
f(192-207)	**Bovine α_{s2}-CN**	**PWIQPKTKVIPYVRYL**	**0.915**	**AMP**		
f(183-207)	Bovine α_{s2}-CN	VYQHQKAMKPWIQPKTKVIPYVRYL	0.312	Non-AMP		
f(176-180)	Camel α_{s2}-CN	TMTPW			-0,11	non-AHT
f(189-194)	Camel α_{s2}-CNsv2	QIIPNL			0,44	AHT
f(181-184)	Camel α_{s2}-CN	NHIK			-0,90	non-AHT
f(146-149)	Camel α_{s2}-CN	DHQK			-0,31	non-AHT
f(162-166)	Camel α_{s2}-CN	LWPEY			1,42	AHT
		di- and tripeptides			pIC50***	
f(151-153)	Camel α_{s2}-CN	LNK			3.96	predicted
f(169-171)	Camel α_{s2}-CN	TVY			4.82	predicted
f(187-188)	Camel α_{s2}-CNsv2	AY			4.85	actual
f(185-186)	Camel α_{s2}-CNsv2	VK			4.89	actual
f(154-155)	Camel α_{s2}-CN	IY			5.68	actual
f(174-175)	Camel α_{s2}-CN	QK			3.05	actual
f(171-173)	Bovine α_{s2}-CN	YQK			4,56/4.96	actual
f(190-192)	Bovine α_{s2}-CN	MKP			4.60/6.37	actual

** Peptide location is given in the longest camel amino acid sequence (putative a_{s2}-CNsv3).*

*** Probability > 0.5 = Predicted AMP.*

**** pIC50 = -logIC50 with IC50 = peptide concentration (µmol/L) necessary to inhibit the angiotensin converting enzyme (ACE) activity by 50%.*

SVM (support vector machine) score: threshold = 0 (Kumar et al., 2015).

Tripeptides YQK and MKP recently identified as an antihypertensive peptide (Xue et al., 2018; Yamada et al., 2015)*are bolded and in red.*

To summarize, the data reported here allowed identifying UP1 and UP2 detected in our previous study (Ryskaliyeva et al., 2018) as splicing isoforms of α_{s2}-CN (α_{s2}-CNsv1 and

α_{s2}-CNsv2, respectively). These isoforms arise from different processing of the *CSN1S2* primary transcript, giving rise to the insertion of exon 13 in α_{s2}-CNsv1 and a downstream extension of exon 16 in α_{s2}-CNsv2. Thus, α_{s2}-CN was shown to be a mixture of at least 16 isoforms differing in polypeptide chain length and phosphorylation levels, identified in both *Camelus* species (*C. bactrianus* and *C. dromedarius*), as well in hybrids. Such a situation is not specific to Camelids and is frequently observed in most of the mammalian species, particularly in small ruminants and Equidae. Little is known about the mechanisms identifying alternatively spliced exons. Do those deletions/insertions in camel α_{s2}-CN simply reflect the lack of accuracy of an intricate processing mechanism whenever mutations induce conformational modifications of pre-mRNA, preventing the normal progress of the splicing process? There are more and more evidences to support the hypothesis that *cis*-acting sequences, both in introns and exons, are involved in the control of this process.

Despite the extreme conservation of the organization of the "casein" locus during the course of evolution (Figure 3.1), the sequences of the proteins encoded by each of the genes that compose this locus have rapidly evolved. Given the exon modular structure of messenger RNAs, the real similarity between α_{s2}-CN across species is significantly higher than it appears at first whether the exon modular structure is taken into account (Figure 3.4). The apparent divergence is in fact largely due to a splicing combinatorial assembly of exons specific of each species, as previously suggested by Martin *et al.* (2007), as far as α_{s1}-CN is concerned. Therefore, differential splicing, as well as genetic polymorphisms as described with camel α_{s1}-CN (Erhardt et al., 2016), generate a molecular diversity of sequences increasing the ability of camel caseins to generate potentially bioactive encrypted peptides.

Acknowledgements

The study was carried out within the Bolashak International Scholarship of the first author, funded by the JSC «Center for International Programs» (Kazakhstan). The research was partly supported by a grant from the Ministry of Education and Science of the Republic of Kazakhstan under the name “Proteomic investigation of camel milk” #1729/GF4, which is duly appreciated. The authors thank all Kazakhstani camel milk farms and Moldir Nurseitova with Ali Totaev for rendering help in sample collection, as well as PAPPSO and @BRIDGe teams at INRA (Jouy-en-Josas, France) for providing necessary facilities and technical support. The authors would like also to thank warmly Wendy Brand-Williams for English

language editing.

Author Contributions

AR carried out the study, collected milk samples, performed the experiments, and interpreted the data. CH performed LC-MS/MS analysis and analyzed the data. GM performed LC-ESI-MS analysis and analyzed the data. BF and GK provided funding. PM conceived and supervised the research, interpreted the data. The manuscript was written by AR, revised and approved by PM. All authors reviewed and contributed to the final manuscript.

Additional Information

Competing Interests Statement

The authors declare no competing interests.

References

Allende, J. E., & Allende, C. C. (1995). Protein kinases. 4. Protein kinase CK2: an enzyme with multiple substrates and a puzzling regulation. *The FASEB Journal : Official Publication of the Federation of American Societies for Experimental Biology*, *9*(5), 313–323.

Alvarez-Ordóñez, A., Begley, M., Clifford, T., Deasy, T., Considine, K., & Hill, C. (2013). Structure-activity relationship of synthetic variants of the milk-derived antimicrobial peptide αs2-casein f(183-207). *Applied and Environmental Microbiology*, *79*(17), 5179–5185. https://doi.org/10.1128/AEM.01394-13

Bijl, E., van Valenberg, H. J. F., Huppertz, T., van Hooijdonk, A. C. M., & Bovenhuis, H. (2014). Phosphorylation of αS1-casein is regulated by different genes. *Journal of Dairy Science*, *97*(11), 7240–7246. https://doi.org/10.3168/jds.2014-8061

Brenaut, P., Bangera, R., Bevilacqua, C., Rebours, E., Cebo, C., & Martin, P. (2012).

Validation of RNA isolated from milk fat globules to profile mammary epithelial cell expression during lactation and transcriptional response to a bacterial infection. *Journal of Dairy Science*, *95*(10), 6130–6144. https://doi.org/10.1016/j.jmb.2011.04.044

Clare, D. A., & Swaisgood, H. E. (2000). Bioactive Milk Peptides: A Prospectus. *Journal of Dairy Science*. https://doi.org/10.3168/jds.S0022-0302(00)74983-6

Cunsolo, V., Saletti, R., Muccilli, V., Gallina, S., Di Francesco, A., & Foti, S. (2017). Proteins and bioactive peptides from donkey milk: The molecular basis for its reduced allergenic properties. *Food Research International*. https://doi.org/10.1016/j.foodres.2017.07.002

Diaz, O., Gouldsworthy, A. M., & Leaver, J. (1996). Identification of Peptides Released from Casein Micelles by Limited Trypsinolysis. *Journal of Agricultural and Food Chemistry*. https://doi.org/10.1021/jf950832u

Erhardt, G., Shuiep, E. T. S., Lisson, M., Weimann, C., Wang, Z., El Zubeir, I. E. Y. M., & Pauciullo, A. (2016). Alpha S1-casein polymorphisms in camel (Camelus dromedarius) and descriptions of biological active peptides and allergenic epitopes. *Tropical Animal Health and Production*, *48*(5), 879–887. https://doi.org/10.1007/s11250-016-0997-6

Fang, Z. H., Visker, M. H. P. W., Miranda, G., Delacroix-Buchet, A., Bovenhuis, H., & Martin, P. (2016). The relationships among bovine αS-casein phosphorylation isoforms suggest different phosphorylation pathways. *Journal of Dairy Science*, *99*(10), 8168–8177. https://doi.org/10.3168/jds.2016-11250

Farrell, H. M., Malin, E. L., Brown, E. M., & Mora-Gutierrez, A. (2009). Review of the chemistry of αS2-casein and the generation of a homologous molecular model to explain its properties. *Journal of Dairy Science*, *92*(4), 1338–1353. https://doi.org/10.3168/jds.2008-1711

Gagnaire, V., & Léonil, J. (1998). Preferential sites of tryptic cleavage on the major bovine caseins within the micelle. *Lait*, *78*, 471–489. Retrieved from http://lait.dairy-journal.org/articles/lait/abs/1998/05/lait_78_1998_5_45/lait_78_1998_5_45.html

Groenen, M. A. M., Dijkhof, R. J. M., Verstege, A. J. M., & van der Poel, J. J. (1993). The complete sequence of the gene encoding bovine α2-casein. *Gene*, *123*(2), 187–193.

https://doi.org/10.1016/0378-1119(93)90123-K

Hayes, H., Petit, E., Bouniol, C., & Popescu, P. (1993). Localization of the αS2-casein gene (CASAS2) to the homoeologous cattle, sheep, and goat chromosomes 4 by in situ hybridization. *Cytogenetic and Genome Research*, *64*(3–4), 281–285. https://doi.org/10.1159/000133593

Henry, C., Saadaoui, B., Bouvier, F., & Cebo, C. (2015). Phosphoproteomics of the goat milk fat globule membrane: New insights into lipid droplet secretion from the mammary epithelial cell. *Proteomics*. https://doi.org/10.1002/pmic.201400245

Johnsen, L. B., Rasmussen, L. K., Petersen, T. E., & Berglund, L. (1995). Characterization of three types of human alpha s1-casein mRNA transcripts. *The Biochemical Journal*. https://doi.org/10.1042/bj3090237

Kappeler, S., Farah, Z., & Puhan, Z. (1998). Sequence analysis of Camelus dromedarius milk caseins. *The Journal of Dairy Research*, *65*(2), 209–222. https://doi.org/10.1017/S0022029997002847

Koczan, D., Hobom, G., & Seyfert, H. M. (1991). Genomic organisation of the bovine alpha-S1 casein gene. *Nucleic Acids Research*, *19*(20), 5591–5596. https://doi.org/10.1093/nar/19.20.5591

Kumar, R., Chaudhary, K., Singh Chauhan, J., Nagpal, G., Kumar, R., Sharma, M., & Raghava, G. P. S. (2015). An in silico platform for predicting, screening and designing of antihypertensive peptides. *Scientific Reports*. https://doi.org/10.1038/srep12512

Lefèvre, C. M., Sharp, J. A., & Nicholas, K. R. (2009). Characterisation of monotreme caseins reveals lineage-specific expansion of an ancestral casein locus in mammals. *Reproduction, Fertility and Development*. https://doi.org/10.1071/RD09083

Leroux, C., Mazure, N., & Martin, P. (1992). Mutations away from splice site recognition sequences might cis-modulate alternative splicing of goat α(s1)-casein transcripts. Structural organization of the relevant gene. *Journal of Biological Chemistry*, *267*(9), 6147–6157.

Marcone, S., Belton, O., & Fitzgerald, D. J. (2017). Milk-derived bioactive peptides and their

health promoting effects: a potential role in atherosclerosis. *British Journal of Clinical Pharmacology*. https://doi.org/10.1111/bcp.13002

Martin, P., Brignon, G., Furet, J. P., & Leroux, C. (2007). The gene encoding α s1 -casein is expressed in human mammary epithelial cells during lactation . *Le Lait*. https://doi.org/10.1051/lait:1996641

Martin, P., Cebo, C., & Miranda, G. (2013). Interspecies comparison of milk proteins: Quantitative variability and molecular diversity. In *Advanced Dairy Chemistry: Volume 1A: Proteins: Basic Aspects, 4th Edition* (pp. 387–429). https://doi.org/10.1007/978-1-4614-4714-6_13

Martin, P., & Leroux, C. (1992). Exon-skipping is responsible for the 9 amino acid residue deletion occurring near the N-terminal of human β-casein. *Biochemical and Biophysical Research Communications*, *183*(2), 750–757. https://doi.org/10.1016/0006-291X(92)90547-X

Matéos, a, Miclo, L., Mollé, D., Dary, a, Girardet, J.-M., & Gaillard, J.-L. (2009). Equine alpha S1-casein: characterization of alternative splicing isoforms and determination of phosphorylation levels. *Journal of Dairy Science*, *92*(8), 3604–3615. https://doi.org/10.3168/jds.2009-2125

Mati, A., Senoussi-Ghezali, C., Si Ahmed Zennia, S., Almi-Sebbane, D., El-Hatmi, H., & Girardet, J. M. (2017). Dromedary camel milk proteins, a source of peptides having biological activities – A review. *International Dairy Journal*. https://doi.org/10.1016/j.idairyj.2016.12.001

McCann, K. B., Shiell, B. J., Michalski, W. P., Lee, A., Wan, J., Roginski, H., & Coventry, M. J. (2005). Isolation and characterisation of antibacterial peptides derived from the f(164-207) region of bovine αs2-casein. *International Dairy Journal*, *15*(2), 133–143. https://doi.org/10.1016/j.idairyj.2004.06.008

Meisel, H. (2004). Multifunctional peptides encrypted in milk proteins. In *BioFactors*. https://doi.org/10.1002/biof.552210111

Menon, R. S., Chang, Y. F., Jeffers, K. F., Jones, C., & Ham, R. G. (1992). Regional localization of human β-casein gene (CSN2) to 4pter-q21. *Genomics*, *13*(1), 225–226.

https://doi.org/10.1016/0888-7543(92)90227-J

Minkiewicz, P., Dziuba, J., Iwaniak, A., Dziuba, M., & Darewicz, M. (2008). BIOPEP database and other programs for processing bioactive peptide sequences. In *Journal of AOAC International*. https://doi.org/10.1093/bib/bbl035

Miranda, G., Mahé, M. F., Leroux, C., & Martin, P. (2004). Proteomic tools characterize the protein fraction of Equidae milk. In *Proteomics* (Vol. 4, pp. 2496–2509). https://doi.org/10.1002/pmic.200300765

Mohanty, D. P., Mohapatra, S., Misra, S., & Sahu, P. S. (2016). Milk derived bioactive peptides and their impact on human health – A review. *Saudi Journal of Biological Sciences*. https://doi.org/10.1016/j.sjbs.2015.06.005

Nagpal, R., Behare, P., Rana, R., Kumar, A., Kumar, M., Arora, S., … Yadav, H. (2011). Bioactive peptides derived from milk proteins and their health beneficial potentials: An update. *Food and Function*. https://doi.org/10.1039/c0fo00016g

Nielsen, S. D., Beverly, R. L., Qu, Y., & Dallas, D. C. (2017). Milk bioactive peptide database: A comprehensive database of milk protein-derived bioactive peptides and novel visualization. *Food Chemistry*. https://doi.org/10.1016/j.foodchem.2017.04.056

Pauciullo, A., & Erhardt, G. (2015). Molecular characterization of the llamas (Lama glama) casein cluster genes transcripts (CSN1S1, CSN2, CSN1S2, CSN3) and regulatory regions. *PLoS ONE*, *10*(4). https://doi.org/10.1371/journal.pone.0124963

Ramunno, L., Cosenza, G., Pappalardo, M., Longobardi, E., Gallo, D., Pastore, N., … Rando, A. (2001). Characterization of two new alleles at the goat CSN1S2 locus. *Animal Genetics*, *32*(5), 264–268. https://doi.org/10.1046/j.1365-2052.2001.00786.x

Ramunno, L., Cosenza, G., Rando, A., Pauciullo, A., Illario, R., Gallo, D., … Masina, P. (2005). Comparative analysis of gene sequence of goat CSN1S1 F and N alleles and characterization of CSN1S1 transcript variants in mammary gland. *Gene*, *345*(2), 289–299. https://doi.org/10.1016/j.gene.2004.12.003

Recio, I., & Visser, S. (1999). Identification of two distinct antibacterial domains within the sequence of bovine α(s2)-casein. *Biochimica et Biophysica Acta - General Subjects*,

1428(2–3), 314–326. https://doi.org/10.1016/S0304-4165(99)00079-3

Rijnkels, M. (2002). Multispecies comparison of the casein gene loci and evolution of casein gene family. *Journal of Mammary Gland Biology and Neoplasia*. https://doi.org/10.1023/A:1022808918013

Rijnkels, M., Elnitski, L., Miller, W., & Rosen, J. M. (2003). Multispecies comparative analysis of a mammalian-specific genomic domain encoding secretory proteins. *Genomics*, *82*(4), 417–432. https://doi.org/10.1016/S0888-7543(03)00114-9

Ryskaliyeva, A., Henry, C., Miranda, G., Faye, B., Konuspayeva, G., & Martin, P. (2018). Combining different proteomic approaches to resolve complexity of the milk protein fraction of dromedary, Bactrian camels and hybrids, from different regions of Kazakhstan. *PLoS ONE*, *13*(5). https://doi.org/10.1371/journal.pone.0197026

Saadaoui, B., Bianchi, L., Henry, C., Miranda, G., Martin, P., & Cebo, C. (2014). Combining proteomic tools to characterize the protein fraction of llama (Lama glama) milk. *Electrophoresis*, *35*(10), 1406–1418. https://doi.org/10.1002/elps.201300383

Saletti, R., Muccilli, V., Cunsolo, V., Fontanini, D., Capocchi, A., & Foti, S. (2012). MS-based characterization of αs2-casein isoforms in donkey's milk. In *Journal of Mass Spectrometry* (Vol. 47, pp. 1150–1159). https://doi.org/10.1002/jms.3031

Tagliabracci, V. S., Wiley, S. E., Guo, X., Kinch, L. N., Durrant, E., Wen, J., … Dixon, J. E. (2015). A Single Kinase Generates the Majority of the Secreted Phosphoproteome. *Cell*, *161*(7), 1619–1632. https://doi.org/10.1016/j.cell.2015.05.028

Tauzin, J., Miclo, L., Roth, S., Mollé, D., & Gaillard, J. L. (2003). Tryptic hydrolysis of bovine αs2-casein: Identification and release kinetics of peptides. *International Dairy Journal*. https://doi.org/10.1016/S0958-6946(02)00127-9

Théolier, J., Fliss, I., Jean, J., & Hammami, R. (2014). MilkAMP: A comprehensive database of antimicrobial peptides of dairy origin. *Dairy Science and Technology*. https://doi.org/10.1007/s13594-013-0153-2

Threadgill, D. W., & Womack, J. E. (1990). Genomic analysis of the major bovine milk protein genes. *Nucleic Acids Research*, *18*(23), 6935–6942.

https://doi.org/10.1093/nar/18.23.6935

Veltri, D., Kamath, U., & Shehu, A. (2018). Deep learning improves antimicrobial peptide recognition. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bty179

Weimann, C., Meisel, H., & Erhardt, G. (2009). Short communication: Bovine κ-casein variants result in different angiotensin I converting enzyme (ACE) inhibitory peptides. *Journal of Dairy Science*. https://doi.org/10.3168/jds.2008-1671

Xue, L., Wang, X., Hu, Z., Wu, Z., Wang, L., Wang, H., & Yang, M. (2018). Identification and characterization of an angiotensin-converting enzyme inhibitory peptide derived from bovine casein. *Peptides*, *99*(September 2017), 161–168. https://doi.org/10.1016/j.peptides.2017.09.021

Yamada, A., Sakurai, T., Ochi, D., Mitsuyama, E., Yamauchi, K., & Abe, F. (2015). Antihypertensive effect of the bovine casein-derived peptide Met-Lys-Pro. *Food Chemistry*. https://doi.org/10.1016/j.foodchem.2014.09.098

Zucht, H. D., Raida, M., Adermann, K., Mägert, H. J., & Forssmann, W. G. (1995). Casocidin-I: a casein-αs2 derived peptide exhibits antibacterial activity. *FEBS Letters*, *372*(2–3), 185–188. https://doi.org/10.1016/0014-5793(95)00974-E

Chapter 4

The main WAP isoform usually found in camel milk arises from the usage of an improbable intron cryptic splice site in the precursor to mRNA in which a GC-AG intron occurs

Alma Ryskaliyeva[1], Céline Henry[2], Guy Miranda[1], Bernard Faye[3], Gaukhar Konuspayeva[4] and Patrice Martin[1]

[1]INRA, UMR GABI, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France
[2]INRA, MICALIS Institute, Plateforme d'Analyse Protéomique Paris Sud-Ouest (PAPPSO), Université Paris-Saclay, 78350 Jouy-en-Josas, France
[3]CIRAD, UMR SELMET, 34398 Montpellier, France
[4]Al-Farabi Kazakh State University, Biotechnology department, 050040 Almaty, Kazakhstan

BMC Genetics (2019) 20:14 https://doi.org/10.1186/s12863-018-0704-x

Abstract

Background: Whey acidic protein (WAP) is a major protein identified in the milk of several mammalian species with cysteine-rich domains known as four-disulfide cores (4-DSC). The organization of the eutherian WAP genes is highly conserved through evolution. It has been proposed that WAP could play an important role in regulating the proliferation of mammary epithelial cells. A bacteriostatic activity was also reported. Conversely to the other mammalian species expressing WAP in their milk, camel WAP contains 4 additional amino acid residues at the beginning of the second 4-DSC domain, introducing a phosphorylation site. The aim of this study was to elucidate the origin of this specificity, which possibly impacts its physiological functions.

Results: Using LC-ESI-MS, we identified in *C. bactrianus* from Kazakhstan a phosphorylated whey protein, exhibiting a molecular mass (12,596 Da), 32 Da higher than the original WAP (12,564 Da) and co-eluting with WAP. cDNA sequencing revealed a transition G/A, which modifies an amino acid residue of the mature protein (V12M), accounting for the mass difference observed between WAP genetic variants. We also report the existence of two splicing variants of camel WAP precursors to mRNA, arising from an alternative usage of the canonical splice site recognized as such in the other mammalian species. However, the major camel WAP isoform results from the usage of an unlikely intron cryptic splice site, extending camel exon 3 upstream by 12-nucleotides encoding the 4 additional amino acid residues (VSSP) in which a potentially phosphorylable Serine residue occurs. Combining protein and cDNA sequences with genome data available (NCBI database), we report another feature of the camel WAP gene which displays a very rare GC-AG type intron. This result was confirmed by sequencing a genomic DNA fragment encompassing exon 3 to exon 4, suggesting for the GC donor site a compensatory effect in terms of consensus at the acceptor exon position.

Conclusions: Combining proteomic and molecular biology approaches we report: the characterization of a new genetic variant of camel WAP, the usage of an unlikely intron cryptic splice site, and the occurrence of an extremely rare GC-AG type of intron.

Key words: Camel, milk, whey protein, splicing, genetic polymorphism

4.1 Background

Camel milk is characterized by a high content of vitamin C (average content ranging between 50 and 250 mg/L), and endowed with a unique composition of protein components (Alhaider et al., 2013; Hinz, O'Connor, Huppertz, Ross, & Kelly, 2012; Ryskaliyeva et al., 2018). Its protein content (35-50 g/L) is rather high (Konuspayeva, Faye, & Loiseau, 2009), with *ca.* 80% are caseins and 20% whey proteins that are soluble at pH 4.6 whereas caseins precipitate close to this pH. The casein fraction comprises 4 caseins (α_{s1}-, α_{s2}-, β- and κ-casein) encoded by four autosomal genes (*CSN1S1, CSN1S2, CSN2* and *CSN3*, respectively) mapped on chromosome 6 in cattle and goat (Hayes, Petit, Bouniol, & Popescu, 1993; Threadgill & Womack, 1990). This fraction is rather complex with many splicing variants and post-translational modifications (Ryskaliyeva et al., 2018). Whey proteins of camel milk mainly consist of α-lactalbumin (α-LA), glycosylation-dependent cell adhesion molecule 1 (GlyCAM1) or lactophorin which is closely related to the bovine proteose peptone component 3 (PP3), the innate immunity Peptido Glycan Recognition Protein (PGRP) and the Whey Acidic Protein (WAP).

WAP is a major whey protein identified in the milk of several species from eutherians as well as marsupial and monotremes (Sharp, Lefèvre, & Nicholas, 2007). It was first shown to be secreted in rodent milks (Hennighausen & Sippel, 1982), and a whey protein, rich in half-cystine residues (n=16), showing strong similarities with rodents WAPs was characterized two years later in camel milk (Beg, von Bahr-Lindstrom, Zaidi, & Jornvall, 1986). Then, the WAP has been identified in rabbits (Devinoy, Hubert, Schaerer, Houdebine, & Kraehenbuhl, 1988), porcine (Ranganathan, Simpson, Shaw, & Nicholas, 1999), wallaby (Topcic et al., 2009), brushtail possum (Demmer, Stasiuk, Grigor, Simpson, & Nicholas, 2001) and more recently in canine (Seki et al., 2012) milks. Camel WAP reaches an average concentration (157 mg/L) 10-folds lower than that (1,500 mg/L) in rodents milk (Kappeler, Farah, & Puhan, 2003), whereas it is a hundred times higher (15 g/L) in rabbit milk (Grabowski et al., 1991). Whey acidic protein (WAP) is expressed in the mammary gland under an extracellular matrix and lactogenic hormones regulation (Lin, Dempsey, Coffey, & Bissell, 1995). WAP gene expression is induced by prolactin, inhibited by progesterone, and strongly amplified by glucocorticoids (Devinoy et al., 1994).

The overall organization of the eutherian WAP genes is highly conserved through evolution (P. Martin, Cebo, & Miranda, 2013; Sharp et al., 2007). It is composed of 4 exons: E1, E3, E4 and E6 (Figure 4.1). While the size of each exon remains rather conserved between species, intron size varies considerably. The first exon encodes the 5'-UTR, N-terminal signal peptide of 19 aa residues, and the first 8-10 aa residues of mature eutherian proteins. Exons 3 and 4 encode two cysteine-rich domains (DI and DIIA) known as four-disulfide cores (4-DSC) in eutherian species (Hennighausen & Sippel, 1982). A third domain (DIII) encoded by exon 2 (missing in eutherian genes) is found in Monotrema and Marsupial species (Sharp et al., 2007). Exon 6 encodes the last 5-9 aa residues and the 3'-UTR while exon 5 (DIIB) is only used in Platypus and Marsupial species (Sharp et al., 2007). Even though, the promoter region of WAP gene is similar to house-keeping genes with weak or absent TATA signal (Kappeler et al., 2003), WAP is not found in all eutherian milks. The functionality of the gene encoding WAP has been lost in ruminants and primates due to a frameshift mutation (Hajjoubi et al., 2006). Consequently, there is no WAP in the milk of ruminants and primates.

The presence of 4-DSC domains in cysteine-rich proteins led to their classification as the WAP gene family. Proteins containing WAP domains with a characteristic 4-DSC occur not only in mammals but also in birds, reptiles, amphibians and fish (Smith, 2011). Each domain comprises eight cysteine residues with a core of six spatially conserved, while the remaining two are positioned at variable distances amino terminal from the core (Simpson & Nicholas, 2002).

The sequence conservation of 4-DSC motifs across species is significant, and it seems likely that the region may be involved in the biological function of the molecule. WAPs share structural similarity with serine protease inhibitors containing WAP motif domains characterized by a four-disulfide core (4-DSC) (Hennighausen & Sippel, 1982). Possible physiological functions of WAP have been proposed, based on its similarity to protease inhibitor (Grabowski et al., 1991). Using *in vitro* and *in vivo* systems, Nukumi et al. (2007) suggested that WAP plays an important role in regulating the proliferation of mammary epithelial cells by preventing elastase-type serine proteases from carrying out extracellular matrix laminin degradation. In addition, the same authors report a bacteriostatic activity of rat WAP against *Staphylococcus aureus* (Iwamori et al., 2010). In marsupial, Sharp et al. (Sharp et al., 2007) suggest that WAP may play also a role in the development of the young. WFDC2, a second WAP-like protein, is differentially expressed in the mammary gland of the

tammar wallaby and provides immune protection to the mammary gland and the developing pouch young (Watt, Sharp, Lefevre, & Nicholas, 2012).

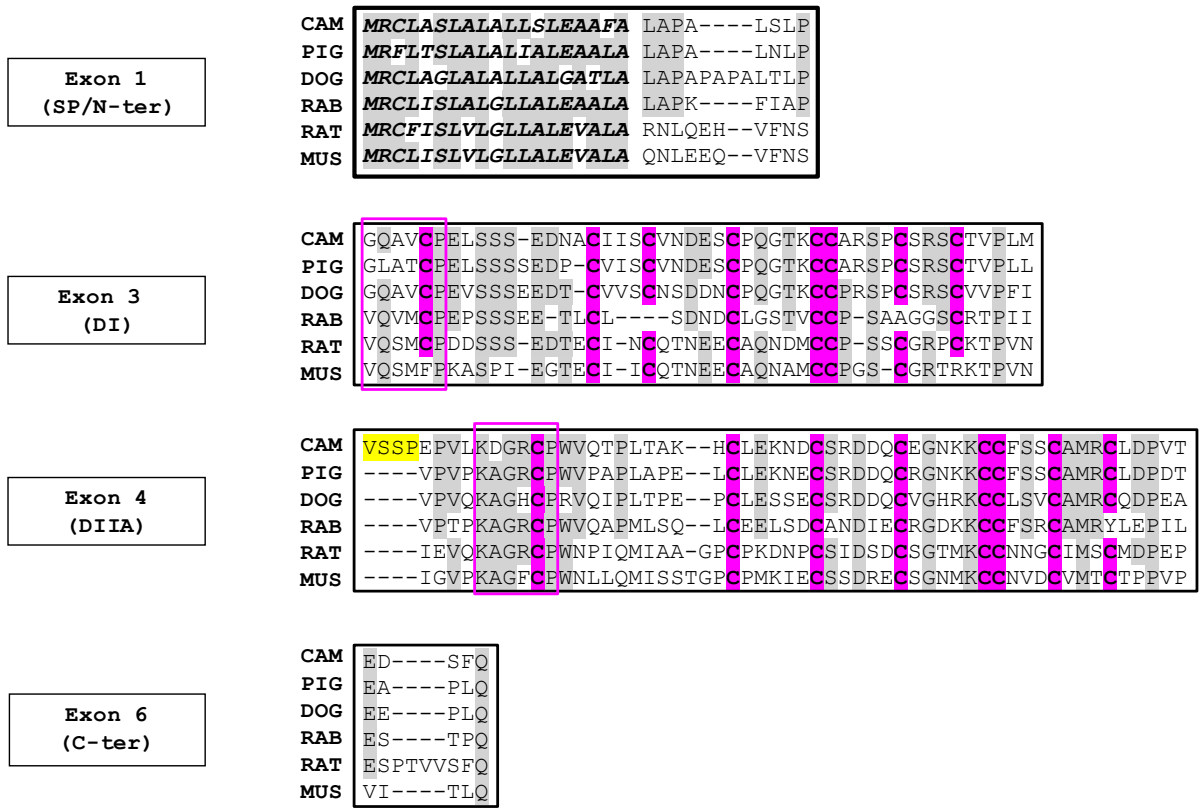


Figure 4.1. Multiple sequences alignment of WAP among Eutherian species including camel (NCBI, LOC105095719), pig (O46655), dog (GenBank AAEX02035361, positions 25,184-23,606), rabbit (P09412), rat (G3V718), and mouse (Q7M748). Four exons: E1, E3, E4 and E6 (numbering of the putative ancestral gene, proposed by Sharp et al. (2007), are given in black boxes. Exons 3 and 4 represent 4-DSC domains (DI and DIIA), while exon 1 and 6 indicate the signal peptide (SP) with the N-terminal part (N-ter) of the mature protein and the C-terminal part (C-ter) of the protein, respectively. WAP motifs are boxed in pink. Conserved Cysteine residues (C) in each 4-DSC domain are pink shaded. Residues identical in more than 3 animal species are grey shaded. Gaps are introduced to maximize similarities. Tetrapeptide VSSP, that is specific to camel WAP, is highlighted in yellow.

The present study was undertaken first to search for WAP genetic polymorphism in camel species (*Camelus bactrianus* and *Camelus dromedarius*) from Kazakhstan. The alignment of sequences of WAP from 5 Eutherian species in which the WAP gene is expressed reveals the occurrence of an additional sequence of 4 amino acid residues in the camel WAP (Figure 4.1). We show that this insertion is due to the usage of an intron cryptic

splice site. Finally, results reported here clarify discrepancies and erroneous data found in sequence databases (predicted sequence from genomic data) and literature. We also report that in camel, the gene encoding WAP comprises a rare GC-AG intron-type that represents less than 1% of annotated donor sites, which is at the origin of predicted sequence anomalies.

4.2 Methods

4.2.1 Milk Sample Collection and Preparation

Raw milk samples were collected during morning milking on healthy dairy camels (n=179) that belonged to two species: *C. bactrianus* (n=72) and *C. dromedarius* (n=65), and their hybrids (n=42) at different lactation stages, ranging between 30 and 90 days postpartum. Those milk samples were skimmed as previously described (Ryskaliyeva et al., 2018) and skimmed milks and fat were stored at -20°C and -80°C, respectively, until analysis.

4.2.2 Selection of Milk Samples for Analysis

A set of 58 milk samples, including individuals from each camel species and grazing regions, was selected, based on lactation stages and number of parities (from 2 to 14), for SDS-PAGE analysis. Then, eight (*C. bactrianus*, n=3, *C. dromedarius*, n=3, and hybrids, n=2) of those 58 milk samples from three different regions exhibiting the most representative profiles were analyzed by LC-MS/MS (LTQ-Orbitrap Discovery, Thermo Finnigan) after tryptic digestion of excised gel bands. Additionally, 30 milk samples (*C. bactrianus*, n=10; *C. dromedarius*, n=10; hybrids, n=10), taken from the 58 milks analyzed by SDS-PAGE, were analyzed by LC-ESI-MS (Bruker Daltonics). One camel milk sample (*C. bactrianus*) displaying a WAP genetic polymorphism in LC-ESI-MS was selected for amplification of WAP cDNA by RT-PCR and cDNA sequencing.

4.2.3 Milk Fat Globule Collection - RNA Extraction and Single-Strand cDNA Synthesis

Total RNA was extracted from milk fat globules (MFG) fraction stored at -80°C using LS Trizol (Invitrogen) following the protocol from the manufacturer, as described by Brenaut et al. (2012). Then, first-strand cDNA was synthesized as described (Ryskaliyeva et al.,

2018). One microliter of 2 U/μL RNase H (Invitrogen Life Technologies) was then added to remove RNA from heteroduplexes. Single-strand cDNA thus obtained was stored at -20°C.

4.2.4 Genomic DNA Isolation

Genomic DNA (gDNA) was isolated from fresh blood of *C. dromedarius* collected in EDTA using the Wizard® Genomic DNA Purification Kit (Promega Corporation, Madison, USA). Briefly, for 3 mL sample volume, 9 mL of Cell Lysis Solution was added and centrifuged at 2,000 x *g* for 10 min at room temperature (RT), after incubating the mixture for 10 min, at RT. The supernatant was removed and, 3 mL of Nuclei Lysis Solution was added to the resuspended white pellet containing red and white blood cells. Then, 1 mL of Protein Precipitation Solution was added to the nuclear lysate, and centrifuged at 2,000 x *g* for 10 min, at RT. The supernatant was transferred to a tube containing 3 mL isopropanol and centrifuged at 2,000 x *g* for 1 min, at RT. After decanting the supernatant, one sample volume of 70% ethanol was added to the DNA and centrifuged at 2,000 x *g* for 1 min, at RT. The ethanol was aspirated using a drawn Pasteur pipette and the pellet was air-dried for 10-15 min. DNA was rehydrated by adding 250 μL of DNA Rehydration Solution and incubated at 65^{o}C for 1 hour and stored at 4^{o}C.

4.2.5 PCR Amplification and DNA sequencing

cDNA and gDNA samples were amplified using primer pairs, of which sequences are given in Table 4.1, designed starting from the published *Camelus* gene sequence (NCBI, LOC105095719) and synthesized by Eurofins genomics (Ebersberg, Germany). PCR was performed in an automated thermocycler GeneAmp® PCR System 2,400 (Perkin-Elmer, Norwalk, USA) with GoTaq® G2 Flexi DNA Polymerase Kit (Promega Corporation, USA). Reactions were carried out in 0.2 mL thin-walled PCR tubes, as described by Ryskaliyeva et al. (2018), using the following PCR cycling conditions: denaturation of cDNA template at 94°C for 2 min, 35 cycles at 94°C for 45 s (denaturation), 58°C for 30 s (annealing) and 72°C for 1 min (extension), with a final extension step of 5 min at 72°C. Sequencing of PCR fragments was performed using PCR primers from both strands, according to the Sanger method by Eurofins MWG GmbH (Ebersberg, Germany).

Table 4.1. Primers used to amplify the cDNA and gDNA target of the WAP gene

	Position	Primer	Sequence 5'->3'	nt*	Amplicon sizes	Tm, ^{o}C
cDNA	5'-UTR	Forward	ATCTGTCACCTGCCTGCCACCTG	23	557	66
	3'-UTR	Reverse	TGAAGCTGAGTGGGTTTTTATTAGC	25		60
gDNA	intron 2	Forward	CAGCTGAGGCTGGCCCGCCTC	21	561	70
	intron 3	Reverse	GCTAGTCTGACACCCTCTCTCTA	23		62

**nucleotides*

4.2.6 1D Sodium Dodecyl Sulfate Polyacrylamide Gel Electrophoresis (SDS-PAGE) and protein identification by LC-MS/MS analysis

Both major and low-abundance proteins resolved by SDS-PAGE were identified by tandem mass analysis (LC-MS/MS) after excision of the relevant bands from the gel and trypsin digestion. The SDS-PAGE conditions was that from Laemmli (1970).

4.2.7 LC-ESI-MS

Fractionation of camel milk proteins and determination of their molecular masses were performed by coupling RP-HPLC to ESI-MS (micrOTOF^{TM} II focus ESI-TOF mass spectrometer; Bruker Daltonics) as described (Ryskaliyeva et al., 2018). Clarified milk samples (25 µL) were directly injected onto a Biodiscovery C5 reverse phase column (300 Å pore size, 3µm, 150x2.1mm; Supelco, France). Eluted proteins were detected by UV-absorbance at 214 nm and the effluent directly introduced to the mass spectrometer. Positive ion mode was used, and mass scans were acquired over a mass-to-charge ratio (m/z) ranging between 600 and 3000 Da (Ryskaliyeva et al., 2018).

4.3 Results

4.3.1 Nucleotide sequence of camel WAP cDNA

Using the rabbit and rodents WAP encoding gene sequences as references, we searched for the expected fourth exon of the WAP gene in the camel genome sequence (NCBI,
LOC105095719). A pair of primers (Table 4.1) was thus designed, in the first exon upstream the coding sequence (forward) and overlapping the downstream putative AATAAA polyA signal (reverse), to amplify a cDNA fragment that was subsequently sequenced. The cDNA sequence of camel WAP thus obtained and given at Figure 4.2, consists of 563 nucleotides encoding a 136 aa polypeptide of M_r 14,510.72 Da, including the signal peptide. The molecular mass of the corresponding mature protein is: 12,564.32 Da.

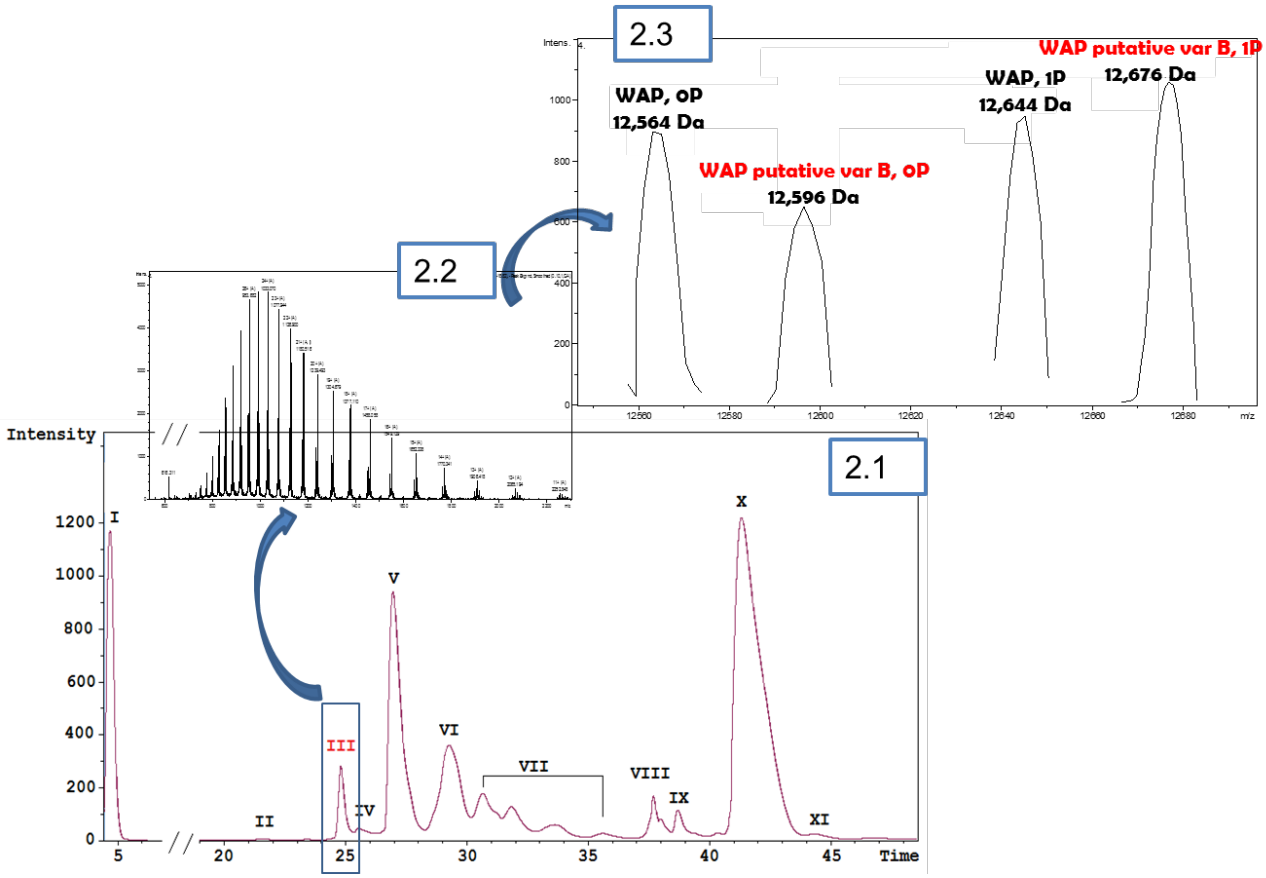


Figure 4.2. Nucleotide sequence of camel WAP cDNA. Primer pairs used for PCR and cDNA sequencing are highlighted in green. Start and stop codons are highlighted in fuchsia. AA residues encoded by triplet codons are bolded and in blue. The coding sequence and non-coding sequences are highlighted in cyan and grey, respectively.

4.3.2 Identification and characterization of a new WAP genetic variant in Bactrian camel milk

LC-ESI-MS analysis of a Bactrian camel milk from the Shymkent region (Figure 4.3), revealed the presence, in peak III (Table 4.2), of two molecular masses differing by 80 Da (one phosphate group) 12,596 Da and 12,676 Da, besides the cognate WAP (12,564 Da and 12, 644 Da). Such a result and the small mass difference (32 Da) strongly suggested the existence of a novel WAP genetic variant, which had not been described so far in camels.

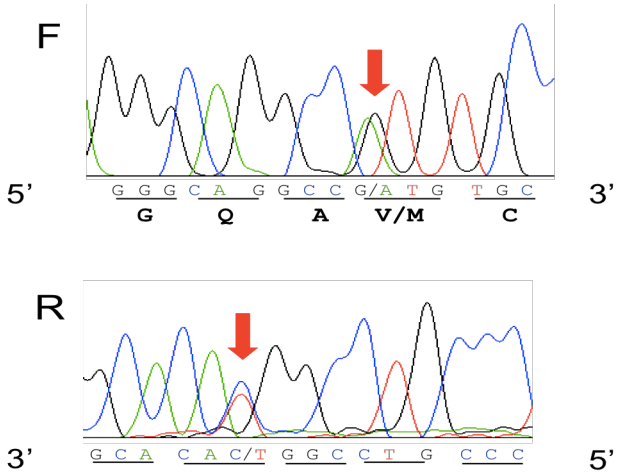


Figure 4.3. Milk protein profiling by LC-ESI-MS of a Bactrian camel milk from the Shymkent region. Eleven major milk protein fractions were identified from RP-HPLC profile (3.A) in the following order: glycosylated κ-CN A and B (I), non-glycosylated κ-CN A and B (II), WAP (III), shorter (Δex16 and 13') + short (Δex16) isoforms of α_{s1}-CN A and C (IV and V), α-LAC + α_{s1}-CN A and C (VI), α_{s2}-CN* (VII), PGRP + α_{s2}-CN* (VIII), LPO/CSA (IX), β-CN A and B (X) and γ_2-CN A and B (XI). Multicharged-ions spectrum from compounds contained in fraction III (3.B). After deconvolution (3.C) the spectrum shows the presence of cognate camel WAP A-0P (12,546 Da) and 1P (12,644 Da) indicated in black, and molecular masses corresponding to a new WAP variant (named B) without (12,596 Da) and with (12,676 Da) one phosphate group, indicated in red.

Table 4.2. Identification of WAP from molecular mass determination using LC-ESI-MS of a clarified Bactrian milk

Peak	Ret.Time, (min)	Observed M_r (Da)	Theoretical M_r (Da)	Protein description	UniProt accession	Intensity
III	24.31	12,564	12,564	WAP variant A, 0P	P09837	896
		12,596	n/a	WAP variant B, 0P	n/a	652
		12,644		WAP variant A, 1P		951
		12,677		WAP variant B, 1P		1,059

n/a - not applicable

Nucleotide sequences of two unrelated individuals (including the Bactrian camel from the Shymkent region) were compared across the complete coding sequence of the camel WAP cDNA, in both directions. PCR yielded a fragment of the expected length (*ca*. 560 bp) for a complete mRNA open reading frame of 408 bp, demonstrating that the primary transcript was correctly spliced. However, examining the nucleotide sequence manually, a transition G/A may be easily noticed (Figure 4.4), leading to the fourth codon change (GTG/ATG) of exon 2, confirmed by the reverse complement sequence. This single base substitution corresponds to a V/M amino acid substitution in position 12 of the mature protein, in agreement with the mass difference of 32 Da (V12M, 99 Da => 131 Da), found between WAP variants detected in LC-ESI-MS. We propose to name the camel WAP (V12) described by Beg et al. (1986) as variant A and the newly identified variant (M12) as variant B. Consequently, molecular masses observed by LC-ESI-MS (12,596 Da, 12,676 Da) precisely correspond to unphosphorylated and phosphorylated (1P) isoforms of WAP variant B, respectively. This B variant which was found in only one (Bactrian) of the 30 camel milk samples analyzed in LC-ESI-MS, at the heterozygous state, appeared therefore to be rare in the Kazakh population. As far as variant A is concerned, the unphosphorylated isoform seems to be prevalent, with relative proportions between unphosphorylated and phosphorylated protein ranging between 70/30 and 55/45, whereas the phosphorylated isoform is predominant (40/60) with the Bactrian camel heterozygous A/B.

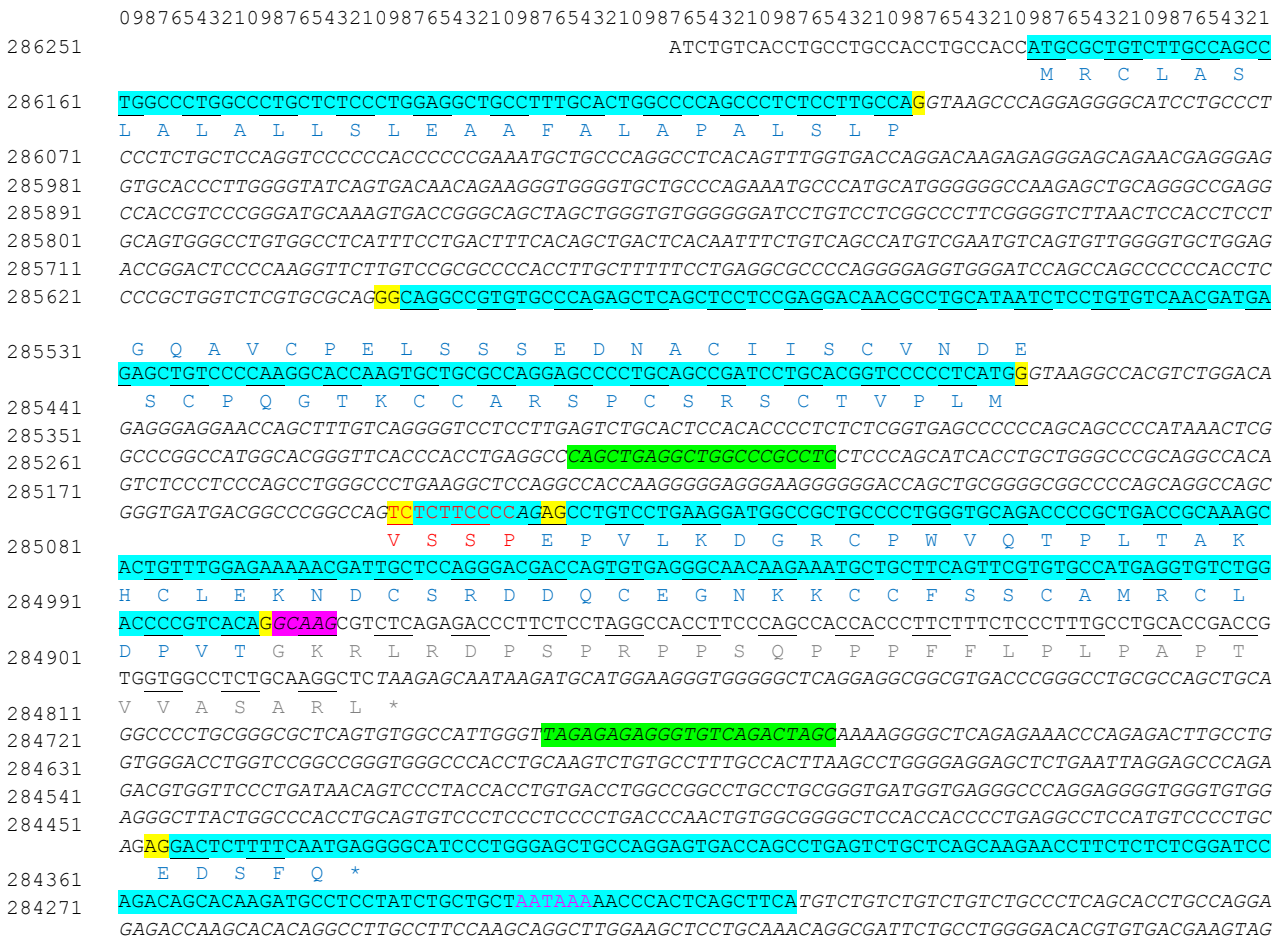


Figure 4.4. Partial sequence of WAP cDNA from *C. bactrianus* from Shymkent region. Nucleotides arising by manual reading, as G/A and C/T transitions, in forward (F) and reverse (R) sequences, respectively, are pointed out by vertical red arrows Nucleotides are clustered by three (codon) corresponding to aa residues (below codons in the forward sequence).

4.3.3 Camel WAP may exist as two isoforms differing in size

LC-MS/MS implemented to confirm the identification of major camel milk proteins (Ryskaliyeva et al., 2018) revealed the existence of two WAP isoforms. Indeed, two tryptic peptides: SCTVPLM**VSSP**EPVLK and SCTVPLMEPVLK identifying camel WAP and differentiating in the presence or absence of the VSSP tetrapeptide (Table 4.3) were found in four of the eight individuals analyzed, belonging to the three species. However, the amount of the isoform lacking the tetrapeptide VSSP appeared too weak (1 spectrum *vs* 10 spectra with VSSP) to be detected by LC-ESI-MS that is less sensitive than LC-MS/MS.

Table 4.3. Sequences of WAP tryptic peptides identified by LC-MS/MS in the milk of a *C. bactrianus*

ID	UniProt accession	Species	Peptide sequence	aa residue		M*r*	Spectra
				Start	Stop		
1	P09837	*C. dromedarius*	LAPALSLPGQAVCPELSSSEDN ACIISCVNDESCPQGTK	1	39	4,174.90	1
2	P09837	*C. dromedarius*	LSLPGQAVCPELSSSEDNACIIS CVNDESCPQGTK	5	39	3,822.69	1
3	P09837	*C. dromedarius*	IISCVNDESCPQGTK	25	39	1,707.77	1
4	P09837	*C. dromedarius*	VNDESCPQGTK	29	39	1,234.54	2
5	P09837	*C. dromedarius*	SCTVPLMVSSPEPVLK	49	64	1,743.90	6
6	P09837	*C. dromedarius*	SCTVPLMVSSPEPVLKDGR	49	67	2,072.05	3
7	P09837	*C. dromedarius*	PLMVSSPEPVLKDGR	53	67	1,624.87	1
8	P09837	*C. dromedarius*	DGRCPWVQTPLTAK	65	78	1,628.82	1
9	P09837	*C. dromedarius*	CPWVQTPLTAK	68	78	1,300.67	11
10	P09837	*C. dromedarius*	HCLEKNDCSR	79	88	1,318.56	2
11	P09837	*C. dromedarius*	HCLEKNDCSRDDQCEGNK	79	96	2,264.91	1
12	P09837	*C. dromedarius*	HCLEKNDCSRDDQCEGNKK	79	97	2,393.00	1
13	P09837	*C. dromedarius*	KCCFSSCAMR	97	106	1,322.51	1
14	P09837	*C. dromedarius*	CCFSSCAMR	98	106	1,161.39	1
15	P09837	*C. dromedarius*	CLDPVTEDSFQ	107	117	1,310.56	13
16	S9XKL5	*C. ferus*	SCTVPLMEPVLK	130	141	1,389.71	1

The table is classified by the start aa residues from the N-terminal sequence. Obtained data matches against UniprotKB taxonomy cetartiodactyla (SwissProt + Trembl) database. Molecular masses (M_r) of peptides are expressed in Da. Spectra indicates the number of spectra permitting the identification of peptides. Charge corresponds to the number of charges (z) of multi-charged ions precursors having given the MS/MS spectra. Peptide sequences including or not the tetrapeptide VSSP, which is highlighted in green, are in bold. Numbering of the *C. ferus* peptide sequence is from KB016488 Genomic DNA Translation.

4.4 Discussion

Here we provide the complete camel WAP mRNA sequence (408 nucleotides open reading frame) that encodes a N-terminal signal peptide of 19 aa residues and a mature protein of 117 aa residues, of which the molecular mass is 12,564 Da, and we report the occurrence in *C. bactrianus* from Kazakhstan of a WAP genetic variant, exhibiting a molecular mass of 12,596 Da (unphosphorylated isoform). cDNA sequencing revealed a transition G/A, which modifies an amino acid residue of the mature protein (V12M), accounting for the mass difference (32 Da) observed between this new genetic variant and the originally described variant (Beg et al., 1986).

4.4.1 Camel WAP is phosphorylated

Camel WAP contains five potential phosphorylation sites (S-X-A code) per molecule (S17, S18, S19, S58, and S87), meanwhile rat WAP has only three potential phosphorylation sites (Dandekar, Robinson, Appella, & Qasba, 1982). Whereas rat WAP is phosphorylated, it was reported that mouse WAP is apparently non-phosphorylated (Hennighausen & Sippel, 1982). From mass data, it is clear that only one site is phosphorylated in camel WAP. Given the extremely constrained and compact structure of the molecule with 8 S-S bridges, essential for folding and functionality of the protein, it is very likely that S58 which is located within the additional sequence connecting the two 4-DSC domains, is the only one seryl residue which is alternatively phosphorylated in camel. Therefore, the other potential phosphorylation sites, namely S17, S18, S19 and S87, should not be phosphorylated. Indeed, WAP contains 16 cysteinyl (C) residues, all of which being involved in disulfide bridges. C residues appear in unique arrangements, divided into two domains. Camel WAP consists of two 4-DSC domains, which are located between aa residues 9 and 55 (DI) and 64 and 111 (DIIA). Each domain begins with a six aa WAP motif (9GQAV**C**P14 and 64KDGR**C**P69), containing the first C residue of the 8 found in the domain.

4.4.2 The usage of an unlikely intron cryptic splice site is responsible for the insertion of 4 amino acid residues in the major camel WAP isoform

Camel WAP shows the higher sequence identity at the aa level (76%) to porcine WAP and much lower aa sequence identities to the WAP from dog (65%), rabbit (51%), rat (40%) and mouse (39%). The comparison of camel WAP sequence with that of the other 5 eutherian species in which the WAP gene is expressed *(Sus scrofa, Canis familiaris, Oryctolagus cuniculus, Rattus norvegicus* and *Mus musculus)*, shows a 4 aa residues (56VSSP59) insertion in the camel polypeptide chain at the beginning of the second 4-DSC domain (Figure 4.1). From the *Camel dromedarius* gene sequence (GenBank LOC105095719) this appears to be the consequence of the usage of an unlikely intron cryptic splice site extending camel exon 3 on its 5' side by 12-nucleotides, whereas in the other 5 species the canonic 3' end of intron 2 is used (Figure 4.5). Indeed, there are two potential intron donor splice sites responding to all requirements of splicing recognition signal: *CCCGGCCAG*|TCTCTTCCCC*AG*|AGCCTGTCCTG (vertical bars materializing possible

intron end recognition sites). Paradoxically it is the weakest site (*CCCGGCCA**AG***), with two purine residues (GG) interrupting the polypyrimidine tract, that is preferentially used by the splicing machinery. We confirmed this sequence, by sequencing a 580-bp fragment amplified from genomic DNA encompassing integrally exon 4 and flanking intron sequences (Figure 4.5).

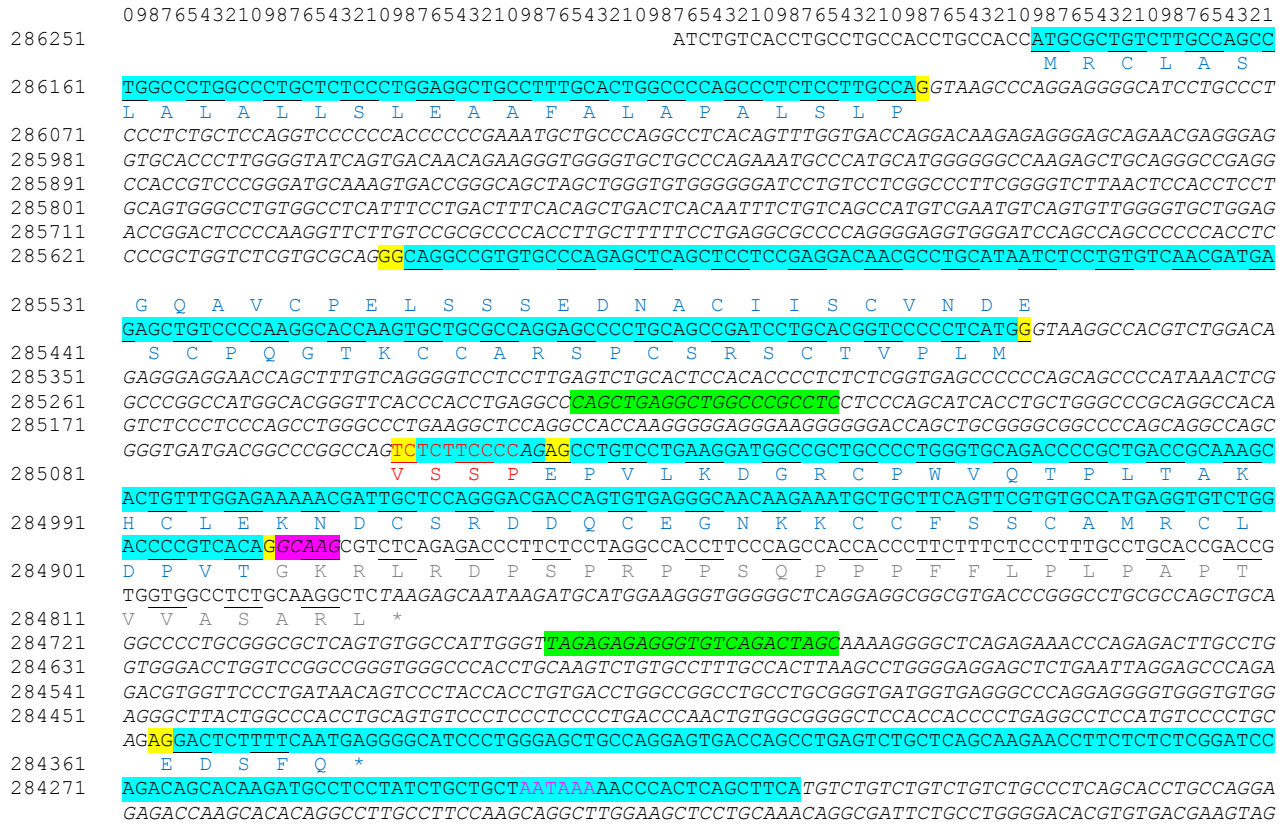


Figure 4.5. Complete sequence of *C. dromedarius* WAP gene available in GenBank (NCBI, LOC105095719). Primer pairs used for PCR and DNA sequencing are highlighted in green. Introns and intergenic sequences are in italics. Intron donor and acceptor sites are bolded. The effective intron donor site (GCAAG) is highlighted in fuchsia. The reading frame in WAP gene is preserved through a 2-1 exon phase (nucleotides involved at the 5'- and 3'-ends are highlighted in yellow). Triplet codons encoding aa are given in blue and encoding specific to camel WAP tetrapeptide VSSP in red. The wrong protein sequence is indicated in grey and polyA site (AATAAA) is in purple. Stop codons are indicated with *: the correct one is in blue and "algorithm" predicted is in grey.

Alternative pre-mRNA splicing generates multiple protein isoforms from a single gene. Leading to a shorter or longer peptide chain, a non-allelic exon-skipping event may occur during the course of the pre-mRNA splicing. It is thought to be provoked by the

weakness (unperfected complementarities between splice site signals and corresponding small ribonuclear protein particles that make up the spliceosome) in the consensus sequences, either at the 5' and/or 3' splice junctions or at the branch point, or both (Patrice Martin, Szymanowska, Zwierzchowski, & Leroux, 2002). As far as camel WAP is concerned, even though cDNA sequencing allowed characterizing a single transcript we could not exclude the existence of a non-allelic short isoform of camel WAP, encoded by a shorter mRNA arising from the usage, as in the other species, of the canonic 3' splice site (TCTCTTCCCC***AG***), which is indeed strongly suggested by the results of LC-MS/MS analyses.

4.4.3 Intron 3 of camel WAP gene is a GC-AG intron type

During the maturation process of pre-mRNA, introns are precisely removed by a large ribonucleoprotein complex: the spliceosome. This splicing step requires splice signals at the 5' and the 3' ends of the intron to be removed and a branch point (Burge, Tuschl, & Sharp, 1999). In their vast majority, introns begin with the standard form dinucleotide GT at the 5' splice site and terminate with the dinucleotide AG at the 3' splice site, so-called GT-AG introns. This rule hold in most cases, however some exceptions have been found (Wu & Krainer, 1999). For example, at the 5' terminus of a few introns, a dinucleotide GC can be occasionally found (Thanaraj T. A. & and Clark F., 2001). Based on the data sets derived from annotated gene structures, it has been reported that GC donor sites account for less than 1% of annotated donor sites and possess a strong consensus sequence (Thanaraj T. A. & and Clark F., 2001). GC-AG introns are processed by the same splicing pathway (U2-type spliceosome) as conventional GT-AG introns (Aebi, Hornig, & Weissmann, 1987). GC-AG introns works in balance with alternative GT splice donor and uses alternative donor and acceptor splice sites, and lack a reasonable poly pyrimidine tract (Thanaraj T. A. & and Clark F., 2001). In humans, about 0.7% of GC-AG introns are involved in regulated splicing (Farrer, Roller, Kent, & Zahler, 2002). In *Caenorhabditis elegans,* experiments indicate that the conserved C at the +2 position of the tenth intron of the let-2 gene is essential for developmentally regulated alternative splicing (Farrer et al., 2002). In camel WAP gene, the C might allow the splice donor to function as a very weak splice site that works in balance with an alternative GT splice donor. In this respect, the only possibility would be the use of the GTCAC site, 7 nt upstream of the GCGAG. Such an assumption would modify the 3' acceptor splice site of intron 3 to maintain a frame of reading in phase and to cause the loss of 3 aa residues: V and T (5' side of intron 3) and E (3' side of intron 3) in the camel WAP sequence.

The C-terminal sequence of the protein described by Beg and co-workers (1986) terminates with the peptide sequence DP**VTE**DSFQ. The protein sequence deduced from our cDNA sequencing, in accordance with the molecular mass determined from LC-MS, terminates with the identical DPVTEDSFQ peptide sequence. The usage of the postulated alternative GT donor splice site cannot be excluded. However, then to preserve the reading frame in phase, the upstream intron should end with the second AG (284 453/4) highlighted in yellow at Figure 4.5. However, we were unable to detect a DPDSFQ C-terminal, as well in LC-MS/MS as through cDNA sequencing. Surprisingly, in WAP gene, available in GenBank (NCBI, LOC105095719) exon 3 is prolonged with 99 nucleotides encoding 33 aa residues until the occurrence of a potential TAA stop codon, which would make exon 4 ineffective. From our results, in agreement with the protein structure first reported (Beg et al., 1986) and our results, the use of this GC donor site is more than likely. Especially since it was reported that alternative GC-AG introns show a compensatory effect in terms of a dramatic increase in consensus at the donor (AG-GCAAG) as well as at the polyYx(C/T)AG-G acceptor exon positions (Thanaraj T. A. & and Clark F., 2001).

4.5 Conclusions

In this study, combining proven proteomic and molecular biology approaches, three main findings in respect to camel WAP are provided. The first one is the identification of a new genetic variant (B), originating from a transition G => A, leading to a codon change (GTG/ATG) in the nucleotide sequence of a Bactrian cDNA, which modifies a single amino acid residue at position 12 of the mature protein (V12M). The second is the detection of two transcripts coding for camel WAP, of which the major one is arising from an improper and unusual processing of a unique pre-mRNA, due to a cryptic splice site usage. This phenomenon leads to the gain or loss of 4 amino acid residues (56VS**S**P59), of which one serine residue (in bold and underlined) is alternatively phosphorylated. Such a genetic polymorphism and splicing events generate a molecular sequence diversity that might account for physiochemical properties of camel WAP that would be quite different, and might contribute unique properties to camel milk. Finally, we report here the occurrence of a GC-AG intron-type (intron 3) in camel gene encoding WAP, showing a compensatory effect in terms of a dramatic increase in consensus at the acceptor exon position.

Abbreviations

aa: amino acid; bp: base pair; LC-MS/MS: liquid chromatography coupled to tandem mass spectrometry; LC-ESI-MS: liquid chromatography-electrospray ionization-tandem mass spectrometry; UTR: untranslated transcribed region; WAP: Whey acidic protein; 4-DSC: four-disulfide cores

Declarations

Ethics approval and consent to participate

Local ethics committee ruled out that no specific permission or formal ethics approval were required for this study with the exception of the rules of European Community regulations on animal experimentation (European Communities Council Directive 86/609/EEC). All animal studies were strictly carried out in compliance with European Community regulations on animal experimentation and with the authorization of the Kazakh Ministry of Agriculture. Milk and blood sampling were performed in appropriate conditions supervised by a veterinary accredited by the French Ethics National Committee for Experimentation on Living Animals.

A verbal informed consent, approved by the local ethics committee, was obtained from the camels owners. No endangered or protected animal species were involved in this study.

Consent to publish

The camels' owners are willing to publish the results of our study.

Availability of data and materials

All data generated or analyzed during this study are included in this published article. The additional datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

The authors declare that they have no competing interests.

Funding

The study was carried out within the Bolashak International Scholarship of the first author, funded by the JSC «Center for International Programs» (Kazakhstan). This work was supported in part by the grant for Scientific Research Project named "Proteomic investigation of exosomes from milk of *Camelus bactrianus* and *Camelus dromedarius*" #AP05134760 from the Ministry of Education and Science of the Republic of Kazakhstan, which is duly appreciated. The funding body had no role in the design of the study and collection, analysis, and interpretation of data and in writing the manuscript.

Author's Contributions

AR carried out the study, collected milk samples, performed the experiments, and interpreted the data. CH and GM have made substantial contribution to performing LC-MS/MS and LC-ESI-MS analysis, respectively, and to acquisition and interpretation of data. BF and GK have made substantial contribution to conception and design of the study, co-supervised the research and provided funding. PM has made substantial contribution to conception and design of the study, conceived and supervised the research, interpreted the data. AR and PM have been involved in writing and drafting the manuscript and revising it. Each author has contributed towards the article and given final approval of the version to be published.

Acknowledgements

The authors thank all Kazakhstani camel milk farms, as well as Moldir Nurseitova and Ali Totaev from the scientific and production company "Antigen" for rendering help in sample collection. We also acknowledge PAPPSO and @BRIDGe teams at INRA (Jouy-en-Josas, France) for providing necessary facilities and technical support.

References

Aebi, M., Hornig, H., & Weissmann, C. (1987). 5′ cleavage site in eukaryotic pre-mRNA splicing is determined by the overall 5′ splice region, not by the conserved 5′ GU. *Cell*, *50*(2), 237–246. https://doi.org/10.1016/0092-8674(87)90219-4

Alhaider, A., Abdelgader, A. G., Turjoman, A. A., Newell, K., Hunsucker, S. W., Shan, B., … Duncan, M. W. (2013). Through the eye of an electrospray needle: Mass spectrometric identification of the major peptides and proteins in the milk of the one-humped camel (Camelus dromedarius). *Journal of Mass Spectrometry*, *48*(7), 779–794. https://doi.org/10.1002/jms.3213

Beg, O. U., von Bahr-Lindstrom, H., Zaidi, Z. H., & Jornvall, H. (1986). A camel milk whey protein rich in half-cystine: Primary structure, assessment of variations, internal repeat patterns, and relationships with neurophysin and other active polypeptides. *European Journal of Biochemistry / FEBS*, *201*(1), 195–201.

Brenaut, P., Bangera, R., Bevilacqua, C., Rebours, E., Cebo, C., & Martin, P. (2012). Validation of RNA isolated from milk fat globules to profile mammary epithelial cell expression during lactation and transcriptional response to a bacterial infection. *Journal of Dairy Science*, *95*(10), 6130–6144. https://doi.org/10.1016/j.jmb.2011.04.044

Burge, C. B., Tuschl, T., & Sharp, P. A. (1999). Splicing of Precursors to mRNAs by the Spliceosomes. In *The RNA World*. https://doi.org/10.1101/087969589.37.525

Dandekar, a M., Robinson, E. a, Appella, E., & Qasba, P. K. (1982). Complete sequence analysis of cDNA clones encoding rat whey phosphoprotein: homology to a protease inhibitor. *Proceedings of the National Academy of Sciences of the United States of America*. https://doi.org/10.1073/pnas.79.13.3987

Demmer, J., Stasiuk, S. J., Grigor, M. R., Simpson, K. J., & Nicholas, K. R. (2001). Differential expression of the whey acidic protein gene during lactation in the brushtail possum (Trichosurus vulpecula). *Biochimica et Biophysica Acta - Gene Structure and Expression*, *1522*(3), 187–194. https://doi.org/10.1016/S0167-4781(01)00334-7

Devinoy, E., Hubert, C., Schaerer, E., Houdebine, L. M., & Kraehenbuhl, J. P. (1988). Sequence of the rabbit whey acidic protein cDNA. *Nucleic Acids Research*, *16*(16), 8180. https://doi.org/10.1093/nar/16.16.8180

Devinoy, E., Thepot, D., Stinnakre, M. G., Fontaine, M. L., Grabowski, H., Puissant, C., … Houdebine, L. M. (1994). High level production of human growth hormone in the milk of transgenic mice: the upstream region of the rabbit whey acidic protein (WAP) gene targets transgene expression to the mammary gland. *Transgenic Research*, *3*(2), 79–89.

https://doi.org/10.1007/BF01974085

Farrer, T., Roller, A. B., Kent, W. J., & Zahler, A. M. (2002). Analysis of the role of Caenorhabditis elegans GC-AG introns in regulated splicing. *Nucleic Acids Research*, *30*(15), 3360–3367. https://doi.org/10.1093/nar/gkf465

Grabowski, H., Le Bars, D., Chene, N., Attal, J., Malienou-Ngassa, R., Puissant, C., & Houdebine, L. M. (1991). Rabbit whey acidic protein concentration in milk, serum, mammary gland extract, and culture medium. *Journal of Dairy Science*, *74*(12), 4143–4150. https://doi.org/10.3168/jds.S0022-0302(91)78609-8

Hajjoubi, S., Rival-Gervier, S., Hayes, H., Floriot, S., Eggen, A., Piumi, F., … Thépot, D. (2006). Ruminants genome no longer contains Whey Acidic Protein gene but only a pseudogene. *Gene*, *370*(1–2), 104–112. https://doi.org/10.1016/j.gene.2005.11.025

Hayes, H., Petit, E., Bouniol, C., & Popescu, P. (1993). Localization of the αS2-casein gene (CASAS2) to the homoeologous cattle, sheep, and goat chromosomes 4 by in situ hybridization. *Cytogenetic and Genome Research*, *64*(3–4), 281–285. https://doi.org/10.1159/000133593

Hennighausen, L. G., & Sippel, A. E. (1982). Mouse whey acidic protein is a novel member of the family of "four-disulfide core" proteins. *Nucleic Acids Research*, *10*(8), 2677–2684. https://doi.org/10.1093/nar/10.8.2677

Hinz, K., O'Connor, P. M., Huppertz, T., Ross, R. P., & Kelly, A. L. (2012). Comparison of the principal proteins in bovine, caprine, buffalo, equine and camel milk. *Journal of Dairy Research*, *79*(02), 185–191. https://doi.org/10.1017/S0022029912000015

Iwamori, T., Nukumi, N., Itoh, K., Kano, K., Naito, K., Kurohmaru, M., … Tojo, H. (2010). Bacteriostatic activity of Whey Acidic Protein (WAP). *The Journal of Veterinary Medical Science / the Japanese Society of Veterinary Science*, *72*(5), 621–625. https://doi.org/10.1292/jvms.08-0331

Kappeler, S., Farah, Z., & Puhan, Z. (2003). 5′-Flanking Regions of Camel Milk Genes Are Highly Similar to Homologue Regions of Other Species and Can be Divided into Two Distinct Groups. *Journal of Dairy Science*, *86*(2), 498–508. https://doi.org/http://dx.doi.org/10.3168/jds.S0022-0302(03)73628-5

Konuspayeva, G., Faye, B., & Loiseau, G. (2009). The composition of camel milk: A meta-analysis of the literature data. *Journal of Food Composition and Analysis*. https://doi.org/10.1016/j.jfca.2008.09.008

Laemmli, U. K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature*. https://doi.org/10.1038/227680a0

Lin, C. Q., Dempsey, P. J., Coffey, R. J., & Bissell, M. J. (1995). Extracellular matrix regulates whey acidic protein gene expression by suppression of TGF-alpha in mouse mammary epithelial cells: studies in culture and in transgenic mice. *J Cell Biol*, *129*(4), 1115–1126. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7744960

Martin, P., Cebo, C., & Miranda, G. (2013). Interspecies comparison of milk proteins: Quantitative variability and molecular diversity. In *Advanced Dairy Chemistry: Volume 1A: Proteins: Basic Aspects, 4th Edition* (pp. 387–429). https://doi.org/10.1007/978-1-4614-4714-6_13

Martin, P., Szymanowska, M., Zwierzchowski, L., & Leroux, C. (2002). The impact of genetic polymorphisms on the protein composition of ruminant milks. *Reproduction Nutrition Development*, *42*(5), 433–459. https://doi.org/10.1051/rnd:2002036

Nukumi, N., Iwamori, T., Kano, K., Naito, K., & Tojo, H. (2007). Whey acidic protein (WAP) regulates the proliferation of mammary epithelial cells by preventing serine protease from degrading laminin. *Journal of Cellular Physiology*, *213*(3), 793–800. https://doi.org/10.1002/jcp.21155

Ranganathan, S., Simpson, K. J., Shaw, D. C., & Nicholas, K. R. (1999). The whey acidic protein family: a new signature motif and three-dimensional structure by comparative modeling. *Journal of Molecular Graphics & Modelling*, *17*(2), 106–113, 134–136. https://doi.org/10.1016/S1093-3263(99)00023-6

Ryskaliyeva, A., Henry, C., Miranda, G., Faye, B., Konuspayeva, G., & Martin, P. (2018). Combining different proteomic approaches to resolve complexity of the milk protein fraction of dromedary, Bactrian camels and hybrids, from different regions of Kazakhstan. *PLoS ONE*, *13*(5). https://doi.org/10.1371/journal.pone.0197026

Seki, M., Matsura, R., Iwamori, T., Nukumi, N., Yamanouchi, K., Kano, K., … Tojo, H.

(2012). Identification of whey acidic protein (WAP) in dog milk. *Experimental Animals / Japanese Association for Laboratory Animal Science*, *61*(1), 67–70. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/22293674

Sharp, J. A., Lefèvre, C., & Nicholas, K. R. (2007). Molecular evolution of monotreme and marsupial whey acidic protein genes. *Evolution and Development*. https://doi.org/10.1111/j.1525-142X.2007.00175.x

Simpson, K. J., & Nicholas, K. R. (2002). The comparative biology of whey proteins. *Journal of Mammary Gland Biology and Neoplasia*. https://doi.org/10.1023/A:1022856801175

Smith, V. J. (2011). Phylogeny of whey acidic protein (WAP) four-disulfide core proteins and their role in lower vertebrates and invertebrates. *Biochemical Society Transactions*, *39*(5), 1403–1408. https://doi.org/10.1042/BST0391403

Thanaraj T. A., & and Clark F. (2001). Human GC-AG alternative intron isoforms with weak donor sites show enhanced consensus at acceptor exon positions. *Nucleic Acids Research*, *29*(12), 2581–2593. https://doi.org/10.1093/nar/29.12.2581

Threadgill, D. W., & Womack, J. E. (1990). Genomic analysis of the major bovine milk protein genes. *Nucleic Acids Research*, *18*(23), 6935–6942. https://doi.org/10.1093/nar/18.23.6935

Topcic, D., Auguste, A., De Leo, A. A., Lefevre, C. M., Digby, M. R., & Nicholas, K. R. (2009). Characterization of the tammar wallaby (Macropus eugenii) whey acidic protein gene: new insights into the function of the protein. *Evolution & Development*, *11*(4).

Watt, A. P., Sharp, J. A., Lefevre, C., & Nicholas, K. R. (2012). WFDC2 is differentially expressed in the mammary gland of the tammar wallaby and provides immune protection to the mammary gland and the developing pouch young. *Developmental and Comparative Immunology*. https://doi.org/10.1016/j.dci.2011.10.001

Wu, Q., & Krainer, a R. (1999). AT-AC pre-mRNA splicing mechanisms and conservation of minor introns in voltage-gated ion channel genes. *Molecular and Cellular Biology*, *19*(5), 3225–3236. https://doi.org/10.1128/MCB.19.5.3225

Chapter 5

Comprehensive Proteomic Analysis of Camel Milk-derived Extracellular Vesicles

Alma Ryskaliyeva[1], Zuzana Krupova[2], Céline Henry[3], Bernard Faye[4], Gaukhar Konuspayeva[5] and Patrice Martin[1]

[1]INRA, UMR GABI, AgroParisTech, Université Paris-Saclay, 78350 Jouy-en-Josas, France
[2]Excilone, 78990 Elancourt, France
[3]INRA, MICALIS Institute, Plateforme d'Analyse Protéomique Paris Sud-Ouest (PAPPSO), Université Paris-Saclay, 78350 Jouy-en-Josas, France
[4]CIRAD, UMR SELMET, 34398 Montpellier Cedex 5, France
[5]Al-Farabi Kazakh State University, Biotechnology department, 050040 Almaty, Kazakhstan

Manuscript in preparation

Abstract

Extracellular vesicles were recovered by optimized density gradient ultracentrifugation from milk of *C. dromedarius*, *C. bactrianus* and hybrids reared in Kazakhstan, visualized by transmission electron microscopy and characterized by nanoparticle tracking analysis. Purified extracellular vesicles had a heterogeneous size distribution with diameters varying between 25 and 170 nm, with average yield of $9.49 \times 10^8 – 4.18 \times 10^{10}$ particles per milliliter of milk. Using a comprehensive strategy combining classical and advanced proteomic approaches an extensive LC-MS/MS proteomic analysis was performed of EVs purified from 24 camel milks (*C. bactrianus*, n=8, *C. dromedarius*, n=10, and hybrids, n=6). A total of 1,010 unique proteins involved in different biological processes were thus identified, including most of the markers associated with small extracellular vesicles, such as CD9, CD63, CD81, HSP70, HSP90, TSG101 and ADAM10. Camel milk-derived EV proteins were classified according to biological processes, cellular components and molecular functions using gene-GO term enrichment analysis of DAVID 6.8 bioinformatics resource. Camel milk-derived EVs were mostly enriched with exosomal proteins. The most prevalent biological processes of camel milk-derived EV proteins were associated with exosome synthesis and its secretion processes (such as intracellular protein transport, translation, cell-cell adhesion and protein transport, and translational initiation) and were mostly engaged in molecular functions such as Poly(A) RNA and ATP binding, protein binding and structural constituent of ribosome.

Key words: milk, camel, exosomes, extracellular vesicles, MFGM, proteome, tetraspanins

5.1 Introduction

Milk is usually considered as a complex biological liquid in which supramolecular structures (casein micelles and milk fat globules) are found beside minerals, vitamins and soluble proteins (whey proteins) as well as cells. In addition to these components, it was recently shown that milk contains also extracellular vesicles that are released by cells as mediators of intercellular communication. Indeed, cell communicate with neighboring cells or with distant cells through the secretion of extracellular vesicles (Tkach & Théry, 2016). Phospholipid bilayer-enclosed extracellular vesicles (EVs) are naturally generated and released from several cell domains of life (*Bacteria, Archaea, Eukarya*) into the extracellular space under physiological and pathological conditions (Delcayre et al., 1996). EVs are commonly classified according to their sub-cellular origin into three major subtypes, such as microvesicles, exosomes, and apoptotic bodies. Contents of vesicles vary a with respect to mode of biogenesis, cell type, and physiologic conditions (Abels & Breakefield, 2016). Exosomes represent the smallest population among EVs ranging in size from 30 to 150 nm in diameter (Hromada et al., 2017). They are generated inside multivesicular bodies in the endosomal compartment during the maturation of early late endosomes and are secreted when these compartments fuse with the plasma membrane (van der Pol et al., 2012). Found in all biofluids exosomes harbor different cargos as a function of cell type and physiologic state (Abels & Breakefield, 2016).

Milk is the sole source of nutrients for the newborn and very young offspring, as well as being an important means to transfer immune components from the mother to the newborn of which the immune system is immature (Kelleher & Lonnerdal, 2001). Milk is therefore thought to play an important role in the development of the immune system of the offspring (Hanson, 2007). Milk is also a source of delivers molecules, via exosomes and/or microvesicles, acting on immune modulation of neonates due to their specific proteins, mRNA, long non-coding RNA and miRNA contents. Exosomes have come in the limelight as biological entities containing unique proteins, lipids, and genetic material. It was shown that the RNA contained in these vesicles could be transferred from one cell to another, through an emerging mode of cell-to-cell communication (Colombo et al., 2014; Simons & Raposo, 2009). RNAs conveyed by exosomes are translated into proteins within transformed cells (mRNA), and/or are involved in regulatory functions (miRNA). For this reason, exosomes are recognized as potent vehicles for intercellular communication, capable for transferring

messages of signaling molecules, nucleic acids, and pathogenic factors (Kabani & Melki, 2016).

Over the last decade, exosomes were widely explored as biological nanovesicles for the development of new diagnostic and therapeutic applications as a promising source for new biomarkers in various diseases (Kanada et al., 2015). For example, exosomes secreted by dendritic cells have been shown to carry MHC-peptide complexes allowing efficient activation of T lymphocytes, thus displaying immunotherapeutic potential as promoters of adaptive immune responses (Keller et al., 2006). Recently, cell culture studies showed that bovine milk-derived exosomes act as a carrier for chemotherapeutic/chemopreventive agents against lung tumor xenografts *in vivo* (Munagala et al., 2016). Nevertheless, their physiological relevance has been difficult to evaluate because their origin, biogenesis and secretion mechanisms remained enigmatic.

Despite a significant number of publications describing the molecular characteristics and investigating the potential biological functions of milk-derived exosomes (Reinhardt et al., 2013; van Herwijnen et al., 2016), there are only one dealing with exosomes derived from camel milk (Yassin et al., 2016). These authors report for the first time isolation and characterization using proteomic (SDS-PAGE and western blot analysis) and transcriptomic analyses exosomes from dromedary milk at different lactation stages. However, there is no comprehensive investigation on exosomal protein variations and variability in composition between individual camels. Milk-derived EVs from Bactrian and hybrid milks have never been explored before. Therefore, to gain insight into the protein diversity of camel milk-derived EVs, we herein provide results of isolation and in-depth morphological and protein characterization of milk-derived EVs from *C. dromedarius*, *C. bactrianus* and hybrids from Kazakhstan using a comprehensive strategy combining classical (SDS-PAGE) and advanced proteomic approaches (LC-MS/MS). Proteomic studies of camel milk and sub-fractions thereof, such as casein, whey, or the milk fat globule membrane (MFGM) have revealed a plethora of bioactive proteins and peptides beneficial for developing immune and metabolic systems (Casado et al., 2008; Kussmann & Van Bladeren, 2011). By contrast, camel milk-derived EVs are still a largely uncharted proteomic terrain, although we know that milk-derived EVs carry cell origin-specific cargo and transport both bioactivity and information between cells (de la Torre Gomez et al., 2018).

5.2 Materials and methods

5.2.1 Ethics statements

All animal studies were carried out in compliance with European Community regulations on animal experimentation (European Communities Council Directive 86/609/EEC) and with the authorization of the Kazakh Ministry of Agriculture. Milk sampling was performed in appropriate conditions supervised by a veterinary accredited by the French Ethics National Committee for Experimentation on Living Animals. No endangered or protected animal species were involved in this study. No specific permissions or approvals were required for this study with the exception of the rules of afore-mentioned European Community regulations on animal experimentation, which were strictly followed.

5.2.2 Milk sample collection and preparation

Raw milk samples were collected during morning milking on healthy dairy camels belonging to two species: *C. bactrianus* (n=72) and *C. dromedarius* (n=65), and their hybrids (n=42) at different lactation stages, ranging between 30 and 90 days postpartum. Camels grazed on four various natural pastures at extreme points of Kazakhstan: Almaty (AL), Shymkent (SH), Kyzylorda (KZ), and Atyrau (ZKO). Whole-milk samples were centrifuged at 3,000 *g* for 30 min at 4^{o}C (Allegra X-15R, Beckman Coulter, France) to separating fat from skimmed milk. Samples were quickly frozen and stored at -80°C (fat) and -20°C (skimmed milk) until analysis.

5.2.3 Selection of milk samples for analysis

In total 24 camel milk samples (*C. bactrianus*, n=8, *C. dromedarius*, n=10, and hybrids, n=6) were selected for isolation of camel milk-derived EVs, based on lactation stages and number of parities of each camel group composed by the species and grazing regions. It should be emphasized that data available on animals: breed, age, lactation stage and calving number, were estimated by a local veterinarian, since no registration of camels in farms is maintained. Six samples of camel milk-derived EVs (*C. bactrianus*, n=2, *C. dromedarius*, n=2, and hybrids, n=2) were selected randomly for transmission electron microscopy (TEM) with negative staining (uranyl acetate). Then, 15 milk samples including the 6 examined by TEM (*C. bactrianus*, n=5, *C. dromedarius*, n=5, and hybrids, n=5) were analyzed by SDS-

PAGE and LC-MS/MS analysis using a QExactive (Thermo Fischer Scientific) Mass Spectrometer after a tryptic digestion of excised gel bands.

5.2.4 Isolation of camel milk-derived EVs

First, skimmed milk samples (40-45 mL) were incubated at 37°C for 30 min in a water bath to enhance free β-casein adsorption to casein micelles. Then, acetic acid was added to the total volume of milk, to obtain a final concentration of 10% and thus acidified milk was incubated at 37°C for 5 min for precipitation of caseins. Finally, 1M sodium acetate was added to obtain a final concentration of 10% for salting out at RT for 5 min, followed by centrifugation at 1,500 *g* for 15 min at 20°C (Beckman Coulter, Allegra X-I5R Centrifuge). After being passed through sterilized vacuum-driven filtration system Millipore Steritop, 0.22 µm, the supernatant, namely the filtrated whey, was concentrated by centrifugation at 4,000 *g* and 20°C using Amicon 1,000K ultracentrifuge tube until to obtain 3 mL of concentrate remaining. The retentate thus obtained was ultra-centrifuged for pelleting the EVs at 33,000 *g* for 1h10 at 4°C (Beckman Coulter, Optima XPN-80, 50TI rotor). Next, the pellet was suspended in 500 µL of PBS and added to pre-prepared 11 mL of sucrose gradient 5-40% and ultra-centrifuged at 34,000 *g* for 18h at 4°C (Beckman Coulter, Optima XPN-80, SW41 rotor). In total, 12 fractions of 1 mL were collected. Fractions previously demonstrated to be enriched in exosomes (10, 11 and 12) were finally suspended into 50 µL of PBS and stored at -80°C, until further analyses.

5.2.5 Transmission electron microscopy (TEM)

The EVs were analyzed as whole-mounted vesicles deposited on EM copper/carbon grids during 5 min, and contrasted 10 sec in 1% uranyl acetate. Grids were examined with Hitachi HT7700 electron microscope operated at 80kV (Elexience – France), and images were acquired with a charge-coupled device camera (AMT).

5.2.6 Nanoparticle tracking analysis

The size distribution and concentration of EVs were measured by NanoSight (NS300) (Malvern Instruments Ltd, Malvern, Worcestershire, UK) according to manufacturer's instructions. A monochromatic laser beam at 405 nm was applied to the diluted suspension of vesicles. Sample temperature is fully programmable through the NTA software (version 3.2

Dev Build 3.2.16). A video of 30 sec was taken with a frame rate of 30 frames/s and particle movement was analyzed by NTA software.

5.2.7 Proteomic analysis

To estimate the concentration of total EVs, the Coomassie Blue Protein Assay was used (Bradford, 1976). Absorbance at 595 nm was measured using the UV-Vis spectrophotometer (UVmini-1240, Shimadzu). The reference standard curve was done with 1 mg/mL commercial bovine serum albumin (BSA, Thermo Fischer Scientific).

In order to identify proteins, mono dimensional electrophoresis (1D SDS-PAGE) followed by trypsin digestion and LC-MS/MS analysis, was used. Ten µg of each individual skimmed camel milk sample were loaded onto 4-15% Mini-PROTEAN® TGX™ Precast Gels (Bio-Rad, Marnes-la-Coquette, France) and subjected to electrophoresis. Samples were prepared with Laemmli Lysis-Buffer (Sigma-Aldrich) with β-mercapto ethanol and denatured at 100°C for 15 min. Separations were performed in a vertical electrophoresis apparatus (Bio-Rad, Marnes-la-Coquette, France). After a short migration (0.5 cm) of samples, gels were stored at -80°C until LC-MS/MS analysis.

Reduction of disulfide bridges of proteins was carried out by incubating at 37°C for one hour with dithiothreitol (DTT, 10 mM, Sigma), meanwhile the alkylation of free cysteinyl residues with iodoacetamide (IAM, 50 mM, Sigma) at RT for 45 min in total obscurity. After gel pieces were washed twice, first, with 100 µL 50% ACN/50 mM NH4HCO3 and then with 50 µL ACN, they were finally dried. The hydration was performed at 37°C overnight using digestion buffer 400 ng lys-C protease + trypsin. Hereby, peptides were extracted with 50% ACN/0.5% TFA and then with 100% ACN. Peptide solutions were dried in a concentrator and finally dissolved into 70 µL 2% ACN in 0.08% TFA.

The identification of peptides was obtained using UltiMate™ 3,000 RSLC nano System (Thermo Fisher Scientific) coupled to a QExactive (Thermo Fischer Scientific) mass spectrometer.

Four µL of each sample were injected at a flow rate of 20 µL/min on a precolumn cartridge (stationary phase: C18 PepMap 100, 5 µm; column: 300 µm x 5 mm) and desalted with a loading buffer 2% ACN and 0.08% TFA. After 4 min, the precolumn cartridge was connected to the separating RSLC PepMap C18 column (stationary phase: RSLC PepMap

100, 2 µm; column: 75 µm x 150 mm). Elution buffers were A: 2% ACN in 0.1% formic acid (HCOOH) and B: 80% ACN in 0.1% HCOOH. The peptide separation was achieved with a linear gradient from 0 to 35% B for 34 min at 300 nL/min. One run took 42 min, including the regeneration and the equilibration steps at 98% B.

Peptide ions were analyzed using Xcalibur 2.1 with the following machine set up in CID mode: 1) full MS scan in QExactive with a resolution of 15,000 (scan range [m/z] = 300-1,600) and 2) top 8 in MS/MS using CID (35% collision energy) in Ion Trap. Analyzed charge states were set to 2-3, the dynamic exclusion to 30 s and the intensity threshold was fixed at 5.0 x 102.

Raw data were converted to mzXML by MS convert (ProteoWizard version 3.0.4601). UniProtKB Cetartiodactyla database was used (157,113 protein entries, version 2015), in conjunction with contaminant databases were searched by algorithm X!TandemPiledriver (version 2015.04.01.1) with the software X!TandemPipeline (version 3.4) developed by the PAPPSO platform (http://pappso.inra.fr/bioinfo/). The protein identification was run with a precursor mass tolerance of 10 ppm and a fragment mass tolerance of 0.5 Da. Enzymatic cleavage rules were set to trypsin digestion ("after R and K, unless P follows directly after") and no semi-enzymatic cleavage rules were allowed. The fix modification was set to cysteine carbamido methylation and methionine oxidation was considered as a potential modification, Results were filtered using inbuilt X!TandemParser with peptide E-value of 0.05, a protein E-value of -2.6 and a minimum of two peptides.

5.2.8 Bioinformatics and functional enrichment analysis

Functional enrichment analyses on camel milk-derived EV protein was performed using online software for gene annotation "The Database for Annotation, Visualization and Integrated Discovery (DAVID)" version 6.8 (https://david.ncifcrf.gov/home.jsp/), as described (Huang et al., 2009).

5.3 Results and Discussion

5.3.1 Isolation of camel milk-derived EVs

EVs are complex and delicate systems requiring optimized isolation and characterization adapted to each fluid type of origin (de la Torre Gomez et al., 2018), which may be achieved by a variety of methods, including ultracentrifugation, filtration, immunoaffinity isolation and microfluidics techniques (Witwer et al., 2013). Choice of method should be guided by the required degree of EVs purity and concentration. General protocols to isolate EVs from cell culture supernatants and body fluids involve steps of differential ultracentrifugation and further purification on a sucrose density gradient (Zonneveld et al., 2014). Commercially produced kits for exosome isolation are nowadays available; however, they are not adapted to milk samples. Due to highly variable composition between different body fluids and even within milks of different species, special optimization steps are required. Isolating milk-derived EVs is complicated by milk composition that differs significantly across species, lactation stage, physiological and health status of individuals. In addition, the recovery of purified exosomes from milk for subsequent analysis requires, according to research objectives, to increase sample volume that is not compatible with classical protocols.

In our study, for milk-derived EVs isolation, the “gold standard method”, including differential ultracentrifugation with sucrose density gradient, was performed (Krupova et al., unpublished results). However, to achieve efficient and quantitative recovery of EVs from camel milk, commonly used protocol was modified. First, milk fat, cells and cellular debris were removed by differential ultracentrifugation. The resuspended pellet was loaded on top of a sucrose gradient and ultracentrifuged to allow for the separation and concentration of EVs. After ultracentrifugation, individual fractions were collected, and EVs enriched fractions (10 to 12) were pooled.

5.3.2 Morphology of isolated camel milk-derived EVs

The method comprising differential ultracentrifugation with density gradient ultracentrifugation was described as being suitable for efficient isolation and purification of higher quality EVs with native morphology intact (Yamada et al., 2012). To visualize and characterize the morphology and size distribution of camel milk-derived EVs, TEM and NTA

analyses was performed. In all 6 milk samples analyzed, we observed a high abundance of homogenous population of EVs enriched in spherical exosomes with average yield of 9.49 x 10^8 - 4.18 x 10^{10} particles per milliliter. The average sizes varied between 25 and 170 nm in diameter. A classical EV-like morphology has been noticed with no significant differences between *C. dromedarius*, *C. bactrianus* and hybrids samples (Figure 5.1). Thus, results confirmed that we have isolated both higher purity and higher quality EVs with intact morphological structures. Based on earlier observations described for dromedary milk (Yassin et al., 2016) and on milk of other species, such as bovine (Reinhardt et al., 2012), porcine (Chen et al., 2016), horse (Sedykh et al., 2017) and human (Admyre et al. 2007), obtained characteristics for *Camelus* milk appear common to EVs across species. Thus, we can conclude that the method of differential ultracentrifugation with sucrose density gradient ultracentrifugation resulted in efficient and reliable isolation of camel milk-derived EVs.

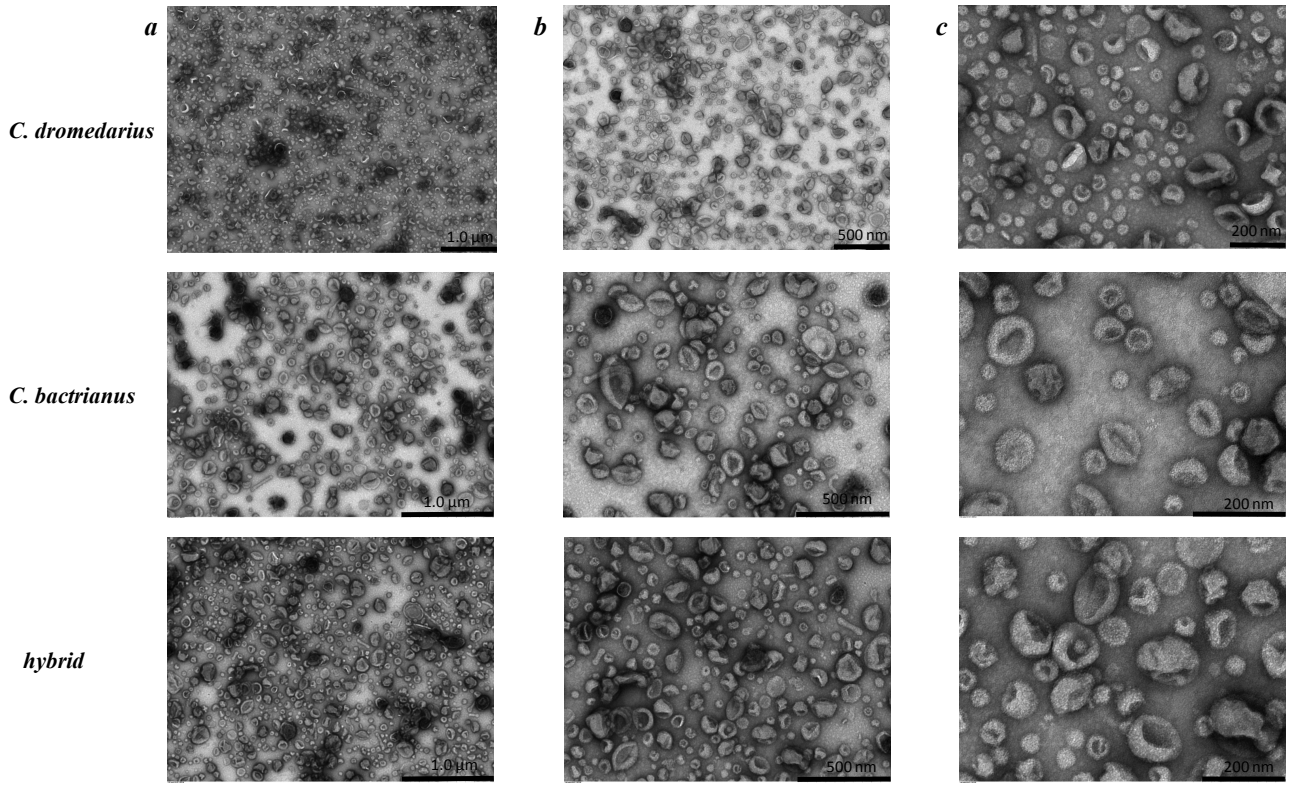


Figure 5.1. Representative electron micrographs of *C. dromedarius, C. bactrianus and hybrid* camel milk-derived EVs. Scale bar represents a) 1 µm, b) 500 nm, c) 200 nm.

5.3.3 In-depth proteomic analysis of camel milk-derived EVs

Apart from the morphology, specific protein composition enables to characterize EVs. To identify proteins present in camel milk-derived EVs extensive analyses involving trypsin digestion, LC-MS/MS (Q Exactive, Thermo Fisher Scientific) and database searches were performed. Recently, using a similar approach a total of 1,963 proteins were identified in human milk-derived EVs (van Herwijnen et al., 2016), and 2,107 unique proteins were described in bovine milk-derived EVs (Reinhardt et al., 2012). Here, from EV samples derived from 15 camel milks (*C. bactrianus,* n=5, *C. dromedarius,* n=5, and hybrids, n=5), a total of 1,010 functional groups of proteins (proteins belonging to a same group share common peptides) were detected (S1[1]). About 890 proteins were common between the three camel species as shown in Figure 5.2 (a), while there are several proteins indicated as unique to *C. bactrianus* (31)*, C. dromedarius* (5)*,* and hybrids (12). Using UniprotKB taxonomy cetartiodactyla (SwissProt + Trembl) database, proteins were identified as authentically matching with proteins in *Camelus* protein databases (*C. dromedarius, C. bactrianus,* and *C. ferus)*, and with the other mammalian species such as*, Lama glama, Lama guanicoe, Bos taurus, Bos mutus, Sus scrofa* and *Ovis aries* protein databases and others. Including the major exosomal protein markers identified, the higher number of low abundant and several differentially expressed proteins enhance the opportunity for revealing the crucial proteins, which can affect exosome synthesis and secretion pathways. By comparison, the proteome of camel milk-derived EVs identified in this study is relatively larger compared to the camel milk proteome reported in a previous study (Ryskaliyeva et al., 2018). A total of 391 functional groups of proteins have been identified from 8 camel milk samples using a less sensitive LC-MS/MS (LTQ Orbitrap XL^{TM} Discovery, Thermo Fisher Scientific), of which 235 proteins were observed as common across camel species. We cannot exclude that there may be several reasons for the significant difference in the number of proteins identified in camel milk-derived EVs, comparatively to previously published data on camel milk proteome. First and foremost the instruments (Q Exactive vs LTQ Orbitrap), since the Q Exactive analyzer was reported to provide significant improvement over the Orbitrap mass spectrometers (Michalski et al., 2011) in terms of sensitivity. Comparing the proteomes between camel milk and camel milk-derived EVs, identified 222 proteins as common (Figure 5.2 (b)), the list of which are provided as a supplementary data (S2[1]).

[1] S1 and S2 will be available on request once the manuscript published

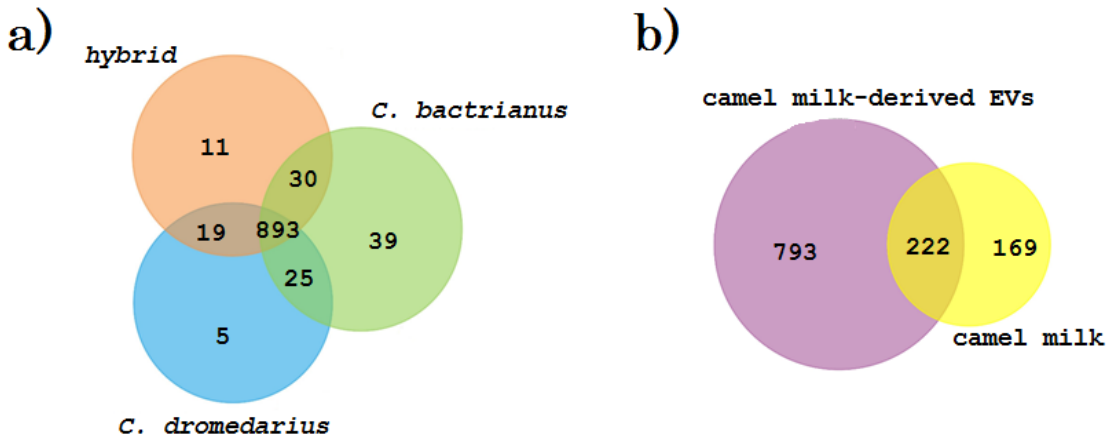


Figure 5.2. a) Venn diagram comparing proteins identified in *C. dromedarius, C. bactrianus* and hybrids milk-derived EVs. The diagram illustrates common and unique EV proteins between the three species **b)** Venn diagram comparing proteins identified in camel milk-derived EVs and proteins detected in camel milk reported in our previous study (Ryskaliyeva et al., 2018).

To get more insight into the subcellular origin of proteins identified, gene-GO term enrichment analysis was performed using DAVID bioinformatics resources 6.8. This analysis helps to understand the function of proteins and addresses them into different biological pathways (Lu et al., 2014). In total 890 and 235 common proteins expressed in camel milk-derived EVs and camel milk, respectively, have been classified according to cellular components. However, despite the limitations of the gene annotations not all camel proteins have been annotated, therefore only 517 exosomal and 96 milk proteins could to be converted to DAVID gene IDs. Thereby, 463 exosomal and 84 milk proteins matched to GO terms under the cellular components headings. As shown in Figure 5.3, both milk-derived EVs and milk samples were mostly enriched with extracellular exosomal proteins (31.09% vs 35.41%, respectively), the specific subset of cellular proteins that are targeted specifically to exosomes. These results coincide with data reported previously on human milk and milk-derived EVs, where a high percentage of proteins linked to GO terms like “exosomes” (van Herwijnen et al., 2016). The next biggest group represented a large number of cytoplasmic proteins (19.58% EVs vs 14.58% milk) found in milk-derived EVs and nucleus proteins (13.24% EVs vs 15.62% milk) in camel milk. Cytoplasmic proteins might originate from “cytoplasmic crescents”, which are trapped between the membrane layers of the MFGM during the budding process when the fat globule leaves the epithelial cell (McManaman & Neville, 2003). Thus, the MFGM can reflect dynamic changes within the MEC and may provide a “snapshot” of mammary gland biology, under particular patho-physiological

conditions. About 13.24% and 12.50% were reported to be membrane proteins identified in camel milk-derived EVs and milk samples. Membrane trafficking proteins represent Rab proteins, which belong to the Ras superfamily of small GTPase. Function of these proteins is central regulation of vesicle budding, motility and fusion. They play a role in endocytosis, transcytosis and exocytosis processes (Lu et al., 2014). In addition, some membrane proteins from intracellular organelles such as cytosol, mitochondrion and Golgi apparatus were highly expressed in camel milk-derived EVs.

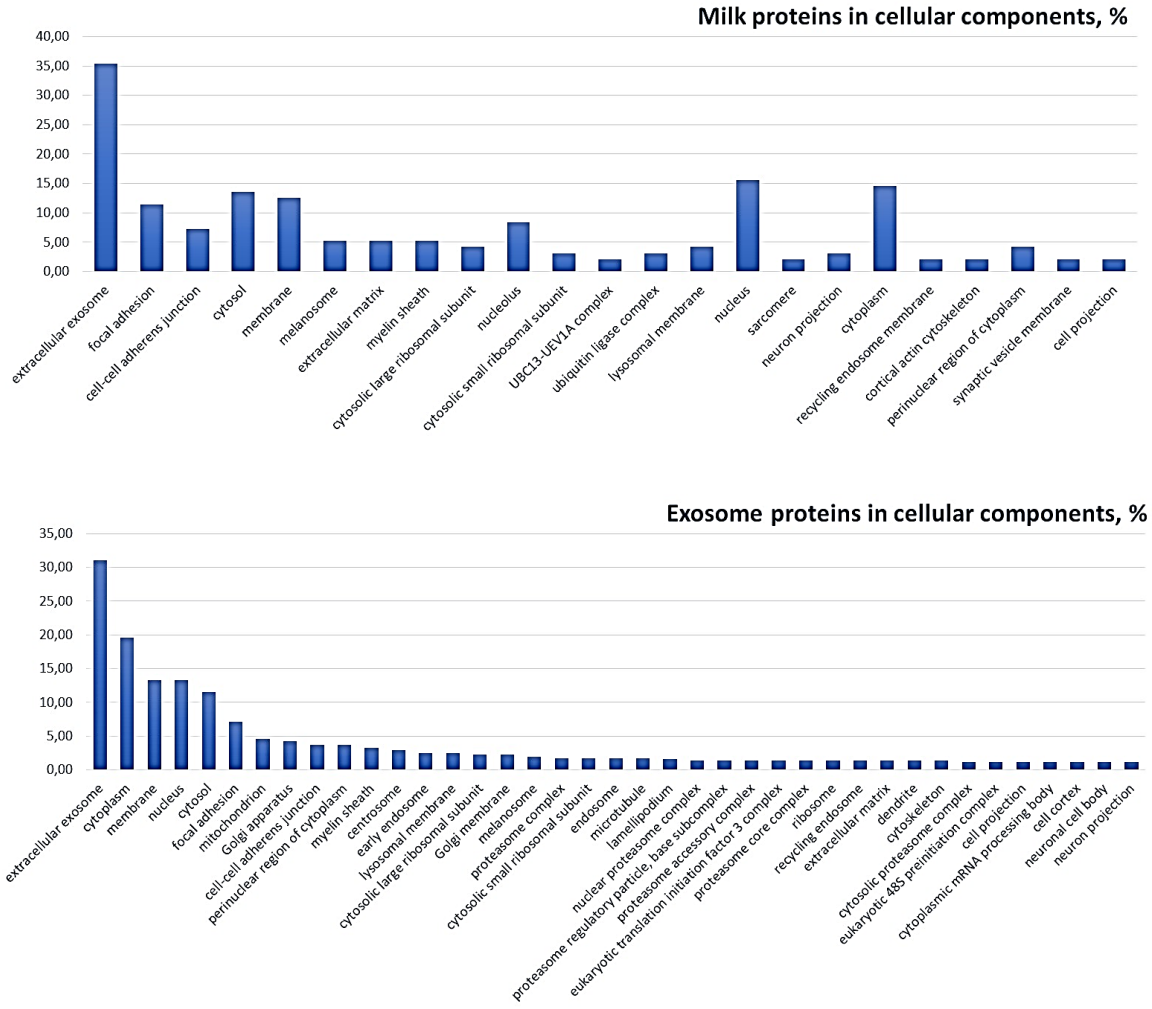
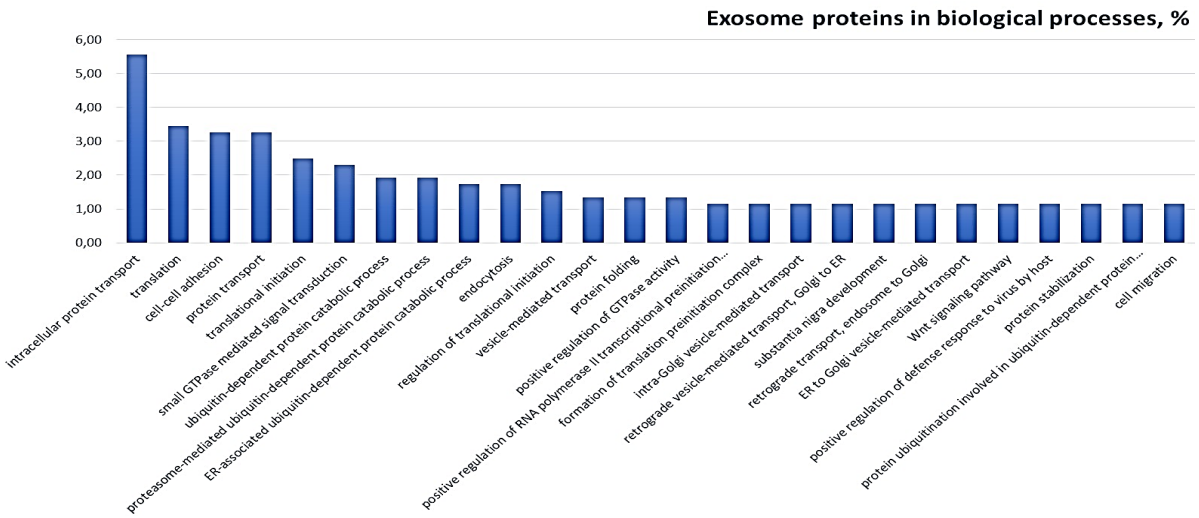


Figure 5.3. Functional annotations of camel milk and milk-derived EV proteins classified into cellular components using DAVID bioinformatics resources 6.8.

Next, we classified proteins expressed in camel milk-derived EVs according to biological processes, molecular functions, and KEGG pathways. Camel milk-derived EV proteins observed were involved in twenty-six GO biological process terms as shown in figure 5.4. The most prevalent biological processes of camel milk-derived EV proteins were associated with exosome synthesis and its secretion processes, such as intracellular protein transport (5.57%), translation (3.45%), cell-cell adhesion and protein transport (3.26%), and translational initiation. Exosomes are increasingly recognized as mediators of intercellular communication due to their capacity to merge with and transfer a repertoire of bioactive molecular content (cargo) to recipient cells (Keller et al., 2006). In addition, EV proteins were mostly engaged in cellular functions such as Poly(A) RNA and ATP (9.60%) binding, protein binding and structural component of ribosome (3.65%). About 3.84% proteins are considered to be associated with GTP binding function (Figure 5.4), regulating membrane-vesicle trafficking process. Proteins expressed in camel milk-derived EVs were categorized into 34 different KEGG pathways. As shown in Table 5.1, camel milk-derived EV proteins were mostly associated with endocytosis (5.57%), Epstein-Barr virus infection (4.03%), ribosome (3.84%), proteasome (3.45%), RNA transport and viral carcinogenesis (2.50%) KEGG pathways. It is known that exosomes display a wide variety of immuno-modulatory properties. This is highlighted by findings showing that exosomes secreted by Epstein-Barr virus (EBV)-transformed B cells are able to stimulate CD4+ T cells in an antigenic-specific manner (Keller et al., 2006).



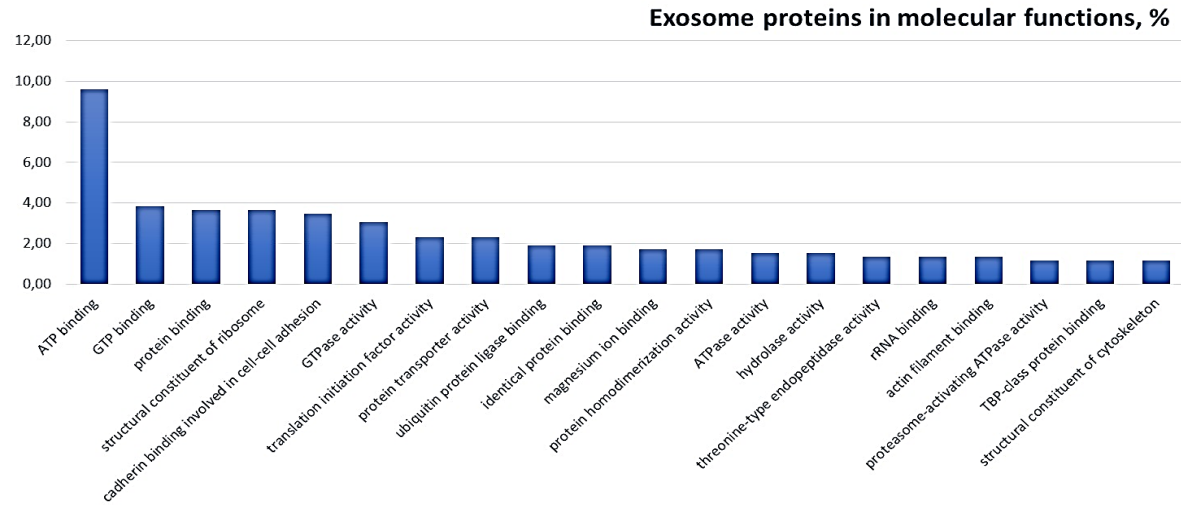


Figure 5.4. GO enrichment analysis of camel milk-derived EV proteins classified into biological processes and molecular functions using DAVID bioinformatics resources 6.8.

Table 5.1. KEGG pathway analysis of camel milk-derived EVs

KEGG pathway term	%	PValue	Fold enrichment
Endocytosis	5.57	8.0E-11	4.32
Epstein-Barr virus infection	4.03	8.7E-8	4.23
Ribosome	3.84	1.1E-9	5.77
Proteasome	3.45	4.1E-16	15.14
RNA transport	2.69	2.2E-4	3.41
Bacterial invasion of epithelial cells	2.11	2.5E-5	5.53
Tight junction	2.11	2.7E-3	3.11
Vasopressin-regulated water reabsorption	1.92	1.9E-6	8.41
Synaptic vesicle cycle	1.92	2.9E-5	6.14
Adherens junction	1.92	5.4E-5	5.69
Salmonella infection	1.73	1.3E-3	4.19
Fc gamma R-mediated phagocytosis	1.34	2.2E-2	3.19
Legionellosis	1.15	1.5E-2	4.07
mTOR signaling pathway	1.15	1.7E-2	3.93
Biosynthesis of amino acids	1.15	4.1E-2	3.14
Endocrine and other factor-regulated calcium reabsorption	0.96	2.2E-2	4.61
Collecting duct acid secretion	0.77	3.1E-2	5.73

5.3.4 Exosomes are a rich source of potential milk biomarkers

Isolation of EVs from milk is complicated by the high lipid content of milk (Witwer et al., 2013). Lipids are released in milk as fat globules (MFGs) by mammary epithelial cells. These MFGs are droplets of lipids surrounded by a complex phospholipid trilayer containing proteins and glycoproteins (Lopez & Ménard, 2011), and thus are a type of EVs. MFGs are largely heterogeneous in size, and their buoyant densities are different from those of EVs. Because of their plasma membrane origin, vesicular nature, and high abundance in milk, however, MFGs may be co-isolated with other EVs populations present in milk (Reinhardt et al., 2012; Witwer et al., 2013). As expected, camel milk-derived EVs analyzed were mostly enriched with MFGM-enriched proteins associated with milk, such as fatty acid synthase (FAS), MFG-E8 (also termed lactadherin), butyrophilin (BTN) and xanthine dehydrogenase. FAS, BTN and MFG-E8 are negative co-stimulatory molecules inhibiting anti-tumor immune responses, which have become novel target pathways for cancer- and immunotherapy development (Cubillos-Ruiz & Conejo-Garcia, 2011; Kuhajda, 2000; Neutzner et al., 2007).

Camel milk-derived EVs analyzed were highly enriched with ubiquitous, cell-specific and cytosolic proteins, including proteins associated with the endosomal pathway, involved in mechanisms responsible for exosome biogenesis. All populations of EVs analyzed expressed in abundance the small Rab GTPases, such as RAB1A, RAB11B, RAB5C, RAB18, RAB2A, RAB7A and RAB21. Rab GTPases are key regulators of intracellular membrane trafficking, from the formation of transport vesicles to their fusion with membranes. Additionally, exosomes derived from all camel milk analyzed were significantly enriched with certain multifunctional proteins, such as Alix (programmed cell death 6 interacting protein PDCD6IP) and TSG101 (tumor susceptibility gene 101). These Endosomal Sorting Complexes required for Transport (ESCRT) protein components of vesicular trafficking process are believed to be a specific exosome-segregated biomarker during its biogenesis (Samuel et al., 2017; Yassin et al., 2016). Recently, it was reported that syndecan-syntenin-ALIX is an important regulator of membrane trafficking and heparan sulphate-assisted signaling, which influences pathological processes, including cancer, the propagation of prions, inflammation, amyloid deposition and neurodegenerative disease (Baietti et al., 2012). Moreover, HSP70 and HSP90 proteins implicated in innate immune responses and antigen presentation (Srivastava, 2002), involved in signal transduction protein kinases and 14-3-3 proteins, and metabolic enzymes such as peroxidases, pyruvate kinases, and α-enolase were

also observed in camel milk-derived EVs. Cell membrane proteins, such as MHC I and MHC II, demonstrating vesical nature of the analyzed materials, were identified as well in all camel milk-derived EVs analyzed, as well as, cytosolic proteins such as tubulin, actin, and actin-binding proteins were highly expressed.

As a consequence of their endosomal origin, most of exosomes are composed of proteins involved in membrane transport and fusion, in multivesicular body biogenesis, in processes requiring heat shock proteins, integrins and tetraspanins (Simons & Raposo, 2009). While some of the proteins that are found in the proteome of many exosomal membrane preparations may merely reflect the cellular abundance of the protein, others are specifically enriched in exosomes and can therefore be defined as exosome-specific marker proteins. Apart from providing nourishment to the offspring, these proteins play a role in intercellular communication via transfer of biomolecules between cells. However, it is currently unknown whether exosomes found in milk originate from immune cells present in milk, from mammary epithelial cells, from circulating cells coming from elsewhere in the body or from bacterial species present in the mammary gland under mild permanent infection (sub-clinical mastitis).

Available proteomic studies define specific markers of the EVs (membrane and cytosolic proteins) and a specific subset of cellular proteins that are targeted specifically to exosomes, the functions of some of them still remain unknown (Théry et al., 2002). This is particularly interesting in relation to their possible involvement in human diseases. The knowledge of exosome proteomics can help not only in understanding their biological roles but also in supplying new biomarkers (Raimondo et al., 2011). Among the membrane proteins most enriched in exosomes are tetraspanins, which play a critical role in exosome formation and are involved in morphogenesis, fission and fusion processes (Rana & Zöller, 2011). Recently CD9, CD63, and CD81 tetraspanins have been defined as novel markers characterizing heterogeneous populations of EVs subtypes (Kowal et al. 2016), the presence of which, including CD82 and TSPAN14 proteins, were confirmed in camel milk-derived EVs. However, some exosome samples analyzed were devoid of CD63. The absence of this tetraspanin in secreted exosomes by some cell types was previously reported, and the necessity of analyzing instead either CD81- or CD9-bearing EVs was reported (Kowal et al., 2016).

Even in the case of markers with strong evidence for EVs subtype specificity, the presence of such markers does not rule out that other types of vesicles are present in a preparation simultaneously (Witwer et al., 2013). Not only the desired populations must be

confirmed as present; contaminants must be demonstrated to be absent. The purity of the exosomes isolated is highly variable due to the presence of contaminating particles, vesicles and molecules such as proteins and/or nucleic acids as well as other cellular components (Vaswani et al., 2017), which may co-purify in vesicle preparations and confound analysis (Mathivanan, Ji, & Simpson, 2010; Witwer et al., 2013). Minimizing contamination in the isolation of exosomes is vital in providing reliable information upon which to base new paradigms (Vaswani et al., 2017). It was reported that exosomes isolated by differential ultracentrifugation with density gradient ultracentrifugation method can be used to examine the relationship of EV proteins to physiological or disease status of the host without any involvement or contamination of other free proteins in milk (Yamada et al., 2012). Density gradients add stringency by efficiently separating particles of different density, which allows removing contaminating non-vesicular particles. Thus, the purity of the camel milk-derived vesicles isolated from contaminations with other multivesicular bodies has been examined and confirmed by the absence of microvesicle surface markers such as p-selectin and CD40, an endoplasmic reticulum marker calnexin, mitochondrial protein mitofilin, and an ER-associated protein GP96. Even though, we have applied a filtration step of the milk supernatant prior to the EVs pelleting step, camel milk-derived EVs were contaminated with caseins, the expression of which have been also detected previously in dromedary (Yassin et al., 2016), human (van Herwijnen et al., 2016) and bovine milk exosomes (Reinhardt et al., 2012).

5.4 Conclusions

Using an optimized isolation protocol, we obtained milk-derived exosomes originating from 15 camel (*C. dromedarius, C. bactrianus* and hybrids) milk samples that satisfied the typical requirements for exosomal morphology, size and protein content. LC-MS/MS analyses allowed identifying a thousand of different proteins that represents to our knowledge, the first comprehensive proteome of camel milk-derived EVs that appears wider than the milk proteome. As mentioned previously in other species camel milk-derived EVs contain proteins also present in other milk components. This is particularly the case for lactadherin/MFG-E8, Ras-related proteins or CD9 that have been reported to occur in MFG. Our results strongly suggest that milk-derived exosomes have different cellular origin. Indeed, besides exosomes originating from mammary epithelial cells there are milk-derived exosomes from immune cells. If we consider that milk-derived exosomes also carry microRNAs, these vesicles have to

be recognized as another important bioactive component of milk that might be involved in transmitting signals from the mother to the newborn but also represents a source of factors potentially responsible for the properties attributed to camelids milk and its health value.

Acknowledgments

The study was carried out within the Bolashak International Scholarship of the first author, funded by the JSC «Center for International Programs» (Kazakhstan). This work was supported in part by the grant for Scientific Research project named "Proteomic investigation of exosomes from milk of *Camelus bactrianus* and *Camelus dromedarius*" #AP05134760 from the Ministry of Education and Science of the Republic of Kazakhstan, which is duly appreciated. The authors thank all Kazakhstani camel milk farms for rendering help in sample collection, as well as PAPPSO, @BRIDGe, and MIMA2 platforms at INRA (Jouy-en-Josas, France) for providing necessary facilities and technical support. We are grateful to Christine Longin for assistance with TEM analysis.

References

Abels, E. R., & Breakefield, O. (2016). Introduction to Extracellular Vesicles: Biogenesis, RNA Cargo Selection, Content, Release, and Uptake. *Cellular and Molecular Neurobiology*, *36*(3), 301–312. https://doi.org/10.1007/s10571-016-0366-z

Admyre, C., Johansson, S. M., Qazi, K. R., Filen, J.-J. ., Lahesmaa, R., Norman, M., … Gabrielsson, S. (2007). Exosomes with Immune Modulatory Features Are Present in Human Breast Milk. *The Journal of Immunology*, *179*(3), 1969–1978. https://doi.org/10.4049/jimmunol.179.3.1969

Baietti, M. F., Zhang, Z., Mortier, E., Melchior, A., Degeest, G., Geeraerts, A., … David, G. (2012). Syndecan-syntenin-ALIX regulates the biogenesis of exosomes. *Nature Cell Biology*, *14*(7), 677–685. https://doi.org/10.1038/ncb2502

Bradford, M. M. (1976). A rapid and sensitive method for the quantitation of microgram quantities of protein using the principle of protein dye binding. *Analytical Biochemistry*, *72*, 248–254. https://doi.org/10.1016/0003-2697(76)90527-3

Casado Dones, M. J., Cruz Martin, R. M., Moreno Gonzalez, C., Oya Luis, I., & Martin Rodriguez, M. (2008). [Children who are allergic to cow's milk. Nutritional treatment]. *Rev Enferm*.

Chen, T., Xi, Q. Y., Ye, R. S., Cheng, X., Qi, Q. E., Wang, S. B., … Zhang, Y. L. (2014). Exploration of microRNAs in porcine milk exosomes. *BMC Genomics*, *15*(1). https://doi.org/10.1186/1471-2164-15-100

Chen, T., Xie, M. Y., Sun, J. J., Ye, R. S., Cheng, X., Sun, R. P., … Zhang, Y. L. (2016). Porcine milk-derived exosomes promote proliferation of intestinal epithelial cells. *Scientific Reports*, *6*. https://doi.org/10.1038/srep33862

Colombo, M., Raposo, G., & Théry, C. (2014). Biogenesis, Secretion, and Intercellular Interactions of Exosomes and Other Extracellular Vesicles. *Annual Review of Cell and Developmental Biology*, *30*(1), 255–289. https://doi.org/10.1146/annurev-cellbio-101512-122326

Cubillos-Ruiz, J. R., & Conejo-Garcia, J. R. (2011). It never rains but it pours: Potential role of butyrophilins in inhibiting anti-tumor immune responses. *Cell Cycle*, *10*(3), 368–369. https://doi.org/10.4161/cc.10.3.14565

de la Torre Gomez, C., Goreham, R. V., Bech Serra, J. J., Nann, T., & Kussmann, M. (2018). "Exosomics"—A Review of Biophysics, Biology and Biochemistry of Exosomes With a Focus on Human Breast Milk. *Frontiers in Genetics*, *9*, 92. https://doi.org/10.3389/fgene.2018.00092

Delcayre, A., Shu, H., & Le Pecq, J. B. (2005). Dendritic cell-derived exosomes in cancer immunotherapy: Exploiting nature's antigen delivery pathway. *Expert Review of Anticancer Therapy*. https://doi.org/10.1586/14737140.5.3.537

Hanson, L. Å. (2007). Session 1: Feeding and infant development Breast-feeding and immune function - Symposium on "Nutrition in early life: New horizons in a new century." In *Proceedings of the Nutrition Society*. https://doi.org/10.1017/S0029665107005654

Hromada, C., Mühleder, S., Grillari, J., Redl, H., & Holnthoner, W. (2017). Endothelial extracellular vesicles-promises and challenges. *Frontiers in Physiology*. https://doi.org/10.3389/fphys.2017.00275

Huang, D. W., Sherman, B. T., & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, *4*(1), 44–57. https://doi.org/10.1038/nprot.2008.211

Kabani, M., & Melki, R. (2016). More than just trash bins? Potential roles for extracellular vesicles in the vertical and horizontal transmission of yeast prions. *Current Genetics*. https://doi.org/10.1007/s00294-015-0534-6

Kanada, M., Bachmann, M. H., Hardy, J. W., Frimannson, D. O., Bronsart, L., Wang, A., … Contag, C. H. (2015). Differential fates of biomolecules delivered to target cells via extracellular vesicles. *Proceedings of the National Academy of Sciences*, 201418401. https://doi.org/10.1073/pnas.1418401112

Kelleher, S. L., & Lonnerdal, B. (2001). Long-term marginal intakes of zinc and retinol affect retinol homeostasis without compromising circulating levels during lactation in rats. *The Journal of Nutrition*, *131*(12), 3237–3242.

Keller, S., Sanderson, M. P., Stoeck, A., & Altevogt, P. (2006). Exosomes: From biogenesis and secretion to biological function. *Immunology Letters*. https://doi.org/10.1016/j.imlet.2006.09.005

Kowal, J., Arras, G., Colombo, M., Jouve, M., Morath, J. P., Primdal-Bengtson, B., … Théry, C. (2016). Proteomic comparison defines novel markers to characterize heterogeneous populations of extracellular vesicle subtypes. *Proceedings of the National Academy of Sciences*, *113*(8), E968–E977. https://doi.org/10.1073/pnas.1521230113

Kuhajda, F. P. (2000). Fatty-acid synthase and human cancer: new perspectives on its role in tumor biology. *Nutrition (Burbank, Los Angeles County, Calif.)*, *16*(3), 202–208. https://doi.org/10.1016/S0899-9007(99)00266-X

Kussmann, M., & Van Bladeren, P. J. (2011). The extended nutrigenomics - understanding the interplay between the genomes of food, gut microbes, and human host. *Frontiers in Genetics*. https://doi.org/10.3389/fgene.2011.00021

Lopez, C., & Ménard, O. (2011). Human milk fat globules: Polar lipid composition and in situ structural investigations revealing the heterogeneous distribution of proteins and the lateral segregation of sphingomyelin in the biological membrane. *Colloids and Surfaces*

B: Biointerfaces. https://doi.org/10.1016/j.colsurfb.2010.10.039

Lu, J., van Hooijdonk, T., Boeren, S., Vervoort, J., & Hettinga, K. (2014). Identification of lipid synthesis and secretion proteins in bovine milk. *The Journal of Dairy Research*, *81*(1), 65–72. https://doi.org/10.1017/S0022029913000642

Mathivanan, S., Ji, H., & Simpson, R. J. (2010). Exosomes: Extracellular organelles important in intercellular communication. *Journal of Proteomics*. https://doi.org/10.1016/j.jprot.2010.06.006

McManaman, J. L., & Neville, M. C. (2003). Mammary physiology and milk secretion. *Advanced Drug Delivery Reviews*. https://doi.org/10.1016/S0169-409X(03)00033-4

Michalski, A., Damoc, E., Hauschild, J.-P., Lange, O., Wieghaus, A., Makarov, A., … Horning, S. (2011). Mass Spectrometry-based Proteomics Using Q Exactive, a High-performance Benchtop Quadrupole Orbitrap Mass Spectrometer. *Molecular & Cellular Proteomics : MCP*. https://doi.org/10.1074/mcp.M111.011015

Munagala, R., Aqil, F., Jeyabalan, J., & Gupta, R. C. (2016). Bovine milk-derived exosomes for drug delivery. *Cancer Letters*, *371*(1), 48–61. https://doi.org/10.1016/j.canlet.2015.10.020

Neutzner, M., Lopez, T., Feng, X., Bergmann-Leitner, E. S., Leitner, W. W., & Udey, M. C. (2007). MFG-E8/lactadherin promotes tumor growth in an angiogenesis-dependent transgenic mouse model of multistage carcinogenesis. *Cancer Research*, *67*(14), 6777–6785. https://doi.org/10.1158/0008-5472.CAN-07-0165

Raimondo, F., Morosi, L., Chinello, C., Magni, F., & Pitto, M. (2011). Advances in membranous vesicle and exosome proteomics improving biological understanding and biomarker discovery. *Proteomics*. https://doi.org/10.1002/pmic.201000422

Rana, S., & Zöller, M. (2011). Exosome target cell selection and the importance of exosomal tetraspanins: a hypothesis. *Biochemical Society Transactions*, *39*(2), 559–562. https://doi.org/10.1042/BST0390559

Raposo, G. (1996). B lymphocytes secrete antigen-presenting vesicles. *Journal of Experimental Medicine*, *183*(3), 1161–1172. https://doi.org/10.1084/jem.183.3.1161

Reinhardt, T. A., Lippolis, J. D., Nonnecke, B. J., & Sacco, R. E. (2012). Bovine milk exosome proteome. *Journal of Proteomics*, *75*(5), 1486–1492. https://doi.org/10.1016/j.jprot.2011.11.017

Reinhardt, T. A., Sacco, R. E., Nonnecke, B. J., & Lippolis, J. D. (2013). Bovine milk proteome: Quantitative changes in normal milk exosomes, milk fat globule membranes and whey proteomes resulting from Staphylococcus aureus mastitis. *Journal of Proteomics*, *82*, 141–154. https://doi.org/10.1016/j.jprot.2013.02.013

Ryskaliyeva, A., Henry, C., Miranda, G., Faye, B., Konuspayeva, G., & Martin, P. (2018). Combining different proteomic approaches to resolve complexity of the milk protein fraction of dromedary, Bactrian camels and hybrids, from different regions of Kazakhstan. *PLoS ONE*, *13*(5). https://doi.org/10.1371/journal.pone.0197026

Samuel, M., Chisanga, D., Liem, M., Keerthikumar, S., Anand, S., Ang, C. S., … Mathivanan, S. (2017). Bovine milk-derived exosomes from colostrum are enriched with proteins implicated in immune response and growth. *Scientific Reports*, *7*(1). https://doi.org/10.1038/s41598-017-06288-8

Sedykh, S. E., Purvinish, L. V., Monogarov, A. S., Burkova, E. E., Grigor'eva, A. E., Bulgakov, D. V., … Nevinsky, G. A. (2017). Purified horse milk exosomes contain an unpredictable small number of major proteins. *Biochimie Open*, *4*, 61–72. https://doi.org/10.1016/j.biopen.2017.02.004

Simons, M., & Raposo, G. (2009). Exosomes--vesicular carriers for intercellular communication. *Current Opinion in Cell Biology*, *21*, 575–581. https://doi.org/10.1016/j.ceb.2009.03.007

Srivastava, P. (2002). Interaction of heat shock proteins with peptides and antigen presenting cells: chaperoning of the innate and adaptive immune responses. *Annual Review of Immunology*, *20*(1), 395–425. https://doi.org/10.1146/annurev.immunol.20.100301.064801

Théry, C., Zitvogel, L., & Amigorena, S. (2002). Exosomes: composition, biogenesis and function. *Nature Reviews Immunology*, *2*(8), 569–579. https://doi.org/10.1038/nri855

Tkach, M., & Théry, C. (2016). Communication by Extracellular Vesicles: Where We Are

and Where We Need to Go. *Cell*. https://doi.org/10.1016/j.cell.2016.01.043

van der Pol, E., Boing, A. N., Harrison, P., Sturk, A., & Nieuwland, R. (2012). Classification, Functions, and Clinical Relevance of Extracellular Vesicles. *Pharmacological Reviews*, *64*(3), 676–705. https://doi.org/10.1124/pr.112.005983

van Herwijnen, M. J. C., Zonneveld, M. I., Goerdayal, S., Nolte – 't Hoen, E. N. M., Garssen, J., Stahl, B., … Wauben, M. H. M. (2016). Comprehensive Proteomic Analysis of Human Milk-derived Extracellular Vesicles Unveils a Novel Functional Proteome Distinct from Other Milk Components. *Molecular & Cellular Proteomics*. https://doi.org/10.1074/mcp.M116.060426

Vaswani, K., Koh, Y. Q., Almughlliq, F. B., Peiris, H. N., & Mitchell, M. D. (2017). A method for the isolation and enrichment of purified bovine milk exosomes. *Reproductive Biology*, *17*(4), 341–348. https://doi.org/10.1016/j.repbio.2017.09.007

Witwer, K. W., Buzás, E. I., Bemis, L. T., Bora, A., Lässer, C., Lötvall, J., … Hochberg, F. (2013). Standardization of sample collection, isolation and analysis methods in extracellular vesicle research. *Journal of Extracellular Vesicles*. https://doi.org/10.3402/jev.v2i0.20360

Yamada, T., Inoshima, Y., Matsuda, T., & Ishiguro, N. (2012). Comparison of Methods for Isolating Exosomes from Bovine Milk. *Journal of Veterinary Medical Science*, *74*(11), 1523–1525. https://doi.org/10.1292/jvms.12-0032

Yassin, A. M., Abdel Hamid, M. I., Farid, O. A., Amer, H., & Warda, M. (2016). Dromedary milk exosomes as mammary transcriptome nano-vehicle: Their isolation, vesicular and phospholipidomic characterizations. *Journal of Advanced Research*, *7*(5), 749–756. https://doi.org/10.1016/j.jare.2015.10.003

Zonneveld, M. I., Brisson, A. R., van Herwijnen, M. J. C., Tan, S., van de Lest, C. H. A., Redegeld, F. A., … Hoen, E. N. t. N. M. (2014). Recovery of extracellular vesicles from human breast milk is influenced by sample collection and vesicle isolation procedures. *Journal of Extracellular Vesicles*. https://doi.org/10.3402/jev.v3.24215

Chapter 6

General Discussion

The general objective of this thesis was to explore the protein fraction of camelid milks in order to identify original molecules (peptide, proteins) potentially responsible for the properties attributed to camel milk.

The analysis of the proteins of the milk fat globule membrane (MFGM) having been the subject of an in-depth study carried out previously (Saadaoui et al., 2013), we initially committed to analyze globally the composition of the milk protein fraction of camelids (proteomic approach), focusing mainly on caseins and the molecular diversity of caseins.

Regarding WPs, various options were possible, given the originality of camel milk in this regard. Detected in mammary secretions of porcine and camel (Kappeler et al., 2004), PGRP that binds to murein peptidoglycans (PGN) of Gram-positive bacteria, was a good candidate. This pattern receptor, involved in innate immunity, may kill Gram-positive bacteria by interfering with peptidoglycan biosynthesis. Lactophorin (GlyCAM1) which is highly expressed in camel milk was also an interesting WP since this phosphoglycoprotein, a component of the milk fat globule membrane, inhibits spontaneous lipolysis in milk by the lipoprotein lipase (Girardet et al., 1993). GlyCAM1 is also suspected to be a scaffold for carbohydrates that mediate functions such as epithelium protector in addition to cell adhesion (Dowbenko et al., 1993). Given the time we had, we made choices and we focused our efforts on the whey acidic protein (WAP) whose protease inhibitory properties are well established and which is an originality of camelids (only large mammals with the pig expressing this protein in milk).

Finally, we started to isolate extracellular vesicles from milk, which are known to carry genetic information (mRNA and microRNA) and proteins involved in the communication between cells and organisms, in order to characterize their proteome.

Variations observed in camel milk composition could be attributed to several environmental factors, such as geographical locations, seasonal variations, feeding conditions, and samples being taken from different breeds, in addition to other parameters including lactation stage, age and calving number (Al haj & Al Kanhal, 2010; Konuspayeva, Faye, et al., 2009). Therefore, for this study, we collected about 180 milk samples from two camel species (*C. bactrianus* and *C. dromedarius,* and their hybrids), at different lactation stages. Camels were grazed on four various natural pastures distant for more than 3,500 kms between the regions at extreme points of Kazakhstan: Almaty (AL) at the foot of Tien Shan Mountain,

Shymkent (SH) along deserts Kyzylkum and Betpak-Dala, Kyzylorda (KZ) on the edge of the steppe, and Atyrau (ZKO) at the mouth of the Caspian Sea.

These milks were submitted to different proven analytical techniques and proteomic approaches (SDS-PAGE, LC-MS/MS and LC-ESI-MS). This study also aimed to evaluate possible differences between species (genetic variability).

6.1 Global analysis: complexity of the camel milk proteome

To get an overview of the protein complexity of camel milk, and on its potential impact on milk characteristics, we provided a complete profiling of the milk protein fraction of Bactrian and dromedary camels from Kazakhstan, including a detailed characterization of camel CN and whey proteins, including variants related to genetic polymorphisms, splicing defects, phosphorylation levels. In addition, we introduce a reference point for further investigation on milk protein polymorphisms in the camel species. The main attractive point in searching for milk protein polymorphisms is to understand the biological significance of the genetic variations, which can be highlighted by evolutionary studies (Caroli et al., 2009). It has been already established clearly that mutations responsible for polymorphisms in milk proteins, occurring at the genomic level either alone or in combinations, might influence milk protein composition at the quantitative as well as at the qualitative levels (Martin et al., 2002).

A detailed characterization of 50 protein molecules, relating to genetic variants and isoforms arising from post-translational modifications and alternative splicing events, belonging to nine protein families (κ-, α_{s1}-, α_{s2}-, β-; and γ-CN, WAP, α-LAC, PGRP, CSA/LPO), was achieved by LC-ESI-MS. Against all expectations, peptides with sequence similarity with bovine *β*-lactoglobulin, the major allergen in bovine milk, were identified in the 8 camel milk samples (Bactrian, dromedary and hybrids), analyzed by LC-MS/MS. The coverage percentage ranged between 30 and 60% in individual milk samples, and reached 71% cumulating all the peptides found. Five peptides related to bovine *β*-lactoglobulin were also detected by Alhaider et al., (2013) in camel milk from Saudi Arabia and the United States. Youcef et al., (2009) revealed a weak cross reaction between dromedary WPs and IgG anti bovine *β*-lactoglobulin. Such findings disagree with the usually admitted notion that *β*-lactoglobulin is absent in camel milk (Hinz et al., 2012; Restani et al., 1999). Even though we cannot exclude a possible contamination by bovine milk (however this seems unlikely with the 8 camel milk samples analyzed by LC-MS/MS) or the presence in camel milk of a

Progesterone Associated Endometrial Protein (PAEP) displaying strong similarities with *β*-lactoglobulin. However, significant similarities between human PAEP and the peptides having allowed the identification of *β*-lactoglobulin in *C. bactrianus* milk were not found. So far, camel milk has never (or only rarely) been extensively studied with the cutting-edge proteomics.

Table 6.1. The nine major camel milk proteins including genetic and splicing variants

Protein	Var.	AA sequences	MW, Da	AA residues
α_{s1}-CN	A E30	RPKYPLRYPEVFQNEPDSIEEVLNKRKILELAVVSPIQFRQENIDELKDTRNEP TEDHIMEDTERKESGSSSSEEVVSSTTEQKDILKEDMPSQRYLEELHRLNKYKL LQLEAIRDQKLIPRVKLSSHPYLEQLYRINEDNHPQLGEPVKVVTQEQAYFHLE PFPQFFQLGASPYVAWYYPPQVMQYIAHPSSYDTPEGIASEDGGKTDVM PQWW	25,307	215
		RPKYPLRYPEVFQNEPDSIEEVLNKRKILELAVVSPIQFRQENIDELKDTRNEP TEDHIMEDTERKESGSSSSEEVVSSTTEQKDILKEDMPSQRYLEELHRLNKYKL LQLEAIRDQKLIPRVKLSSHPYLEQLYRINEDNHPQLGEPVKVVTQPFPQFFQL GASPYVAWYYPPQVMQYIAHPSSYDTPEGIASEDGGKTDVMPQWW	24,289	207
	C D30	RPKYPLRYPEVFQNEPDSIEEVLNKRKILDLAVVSPIQFRQENIDELKDTRNEP TEDHIMEDTERKESGSSSSEEVVSSTTEQKDILKEDMPSQRYLEELHRLNKYKL LQLEAIRDQKLIPRVKLSSHPYLEQLYRINEDNHPQLGEPVKVVTQEQAYFHLE PFPQFFQLGASPYVAWYYPPQVMQYIAHPSSYDTPEGIASEDGGKTDVMPQWW	25,293	215
		RPKYPLRYPEVFQNEPDSIEEVLNKRKILDLAVVSPIQFRQENIDELKDTRNEP TEDHIMEDTERKESGSSSSEEVVSSTTEQKDILKEDMPSQRYLEELHRLNKYKL LQLEAIRDQKLIPRVKLSSHPYLEQLYRINEDNHPQLGEPVKVVTQPFPQFFQL GASPYVAWYYPPQVMQYIAHPSSYDTPEGIASEDGGKTDVMPQWW	24,275	207
α_{s2}-CN	sv0	KHEMDQGSSSEESINVSQQKFKQVKKVAIHPSKEDICSTFCEEAVRNIKEVESA EVPTENKISQFYQKWKFLQYLQALHQGQIVMNPWDQGKTRAYPFIPTVNTEQLS ISEESTEVPTEESTEVFTKKTELTEEEKDHQKFLNKIYQYYQTFLWPEYLKTVY QYQKTMTPWNHIKRYF	21,266	178
	sv1	KHEMDQGSSSEESINVSQQKFKQVKKVAIHPSKEDICSTFCEEAVRNIKEVESA EVPTENKISQFYQKWKFLQYLQALHQGQIVMNPWDQGKTRAYPFIPTVNTEQLS ISEESTEVPTEENSKKTVDTESTEVFTKKTELTEEEKDHQKFLNKIYQYYQTFL WPEYLKTVYQYQKTMTPWNHIKRYF	22,268	187
	sv2	KHEMDQGSSSEESINVSQQKFKQVKKVAIHPSKEDICSTFCEEAVRNIKEVESA EVPTENKISQFYQKWKFLQYLQALHQGQIVMNPWDQGKTRAYPFIPTVNTEQLS ISEESTEVPTEESTEVFTKKTELTEEEKDHQKFLNKIYQYYQTFLWPEYLKTVY QYQKTMTPWNHIKVKAYQIIPNLRYF	22,406	188
β-CN	A I186	REKEEFKTAGEALESISSSEESITHINKQKIEKFKIEEQQQTEDEQQDKIYTFP QPQSLVYSHTEPIPYPILPQNFLPPLQPAVMVPFLQPKVMDVPKTKETIIPKRK EMPLLQSPVVPFTESQSLTLTDLENLHLPLPLLQSLMYQIPQPVPQTPMIPPQS LLSLSQFKVLPVPQQMVPYPQRAIPVQAVLPFQEPVPDPVRGLHPVPQPLVPVI A	24,632	217
		TAGEALESISSSEESITHINKQKIEKFKIEEQQQTEDEQQDKIYTFPQPQSLVY SHTEPIPYPILPQNFLPPLQPAVMVPFLQPKVMDVPKTKETIIPKRKEMPLLQS PVVPFTESQSLTLTDLENLHLPLPLLQSLMYQIPQPVPQTPMIPPQSLLSLSQF	23,685	210

		KVLPVPQQMVPYPQRAIPVQAVLPFQEPVPDPVRGLHPVPQPLVPVIA		
		RKEMPLLQSPVVPFTESQSLTLTDLENLHLPLPLLQSLMYQIPQPVPQTPMIPP QSLLSLSQFKVLPVPQQMVPYPQRAIPVQAVLPFQEPVPDPVRGLHPVPQPLVP VIA	12,357	111
	B M186	REKEEFKTAGEALESISSSEESITHINKQKIEKFKIEEQQQTEDEQQDKIYTFP QPQSLVYSHTEPIPYPILPQNFLPPLQPAVMVPFLQPKVMDVPKTKETIIPKRK EMPLLQSPVVPFTESQSLTLTDLENLHLPLPLLQSLMYQIPQPVPQTPMIPPQS LLSLSQFKVLPVPQQMVPYPQRAMPVQAVLPFQEPVPDPVRGLHPVPQPLVPVI A	24,650	217
		TAGEALESISSSEESITHINKQKIEKFKIEEQQQTEDEQQDKIYTFPQPQSLVY SHTEPIPYPILPQNFLPPLQPAVMVPFLQPKVMDVPKTKETIIPKRKEMPLLQS PVVPFTESQSLTLTDLENLHLPLPLLQSLMYQIPQPVPQTPMIPPQSLLSLSQF KVLPVPQQMVPYPQRAMPVQAVLPFQEPVPDPVRGLHPVPQPLVPVIA	23,703	210
		RKEMPLLQSPVVPFTESQSLTLTDLENLHLPLPLLQSLMYQIPQPVPQTPMIPP QSLLSLSQFKVLPVPQQMVPYPQRAMPVQAVLPFQEPVPDPVRGLHPVPQPLVP VIA	12,375	111
κ-CN	A F11	EVQNQEQPTCFEKVERLLNEKTVKYFPIQFVQSRYPSYGINYYQHRLAVPINNQ FIPYPNYAKPVAIRLHAQIPQCQALPNIDPPTVERRPRPRPSFIAIPPKKTQDK TVNPAINTVATVEPPVIPTAEPAVNTVVIAEASSEFITTSTPETTTVQITSTEI	18,254	162
	B C11	EVQNQEQPTCCEKVERLLNEKTVKYFPIQFVQSRYPSYGINYYQHRLAVPINNQ FIPYPNYAKPVAIRLHAQIPQCQALPNIDPPTVERRPRPRPSFIAIPPKKTQDK TVNPAINTVATVEPPVIPTAEPAVNTVVIAEASSEFITTSTPETTTVQITSTEI	18,210	162
	C C11, I111, V131	EVQNQEQPTCCEKVERLLNEKTVKYFPIQFVQSRYPSYGINYYQHRLAVPINNQ FIPYPNYAKPVAIRLHAQIPQCQALPNIDPPTVERRPRPRPSFIAIPPKKTQDK TVIPAINTVATVEPPVIPTAEPVVNTVVIAEASSEFITTSTPETTTVQITSTEI	18,236	162
α-LAC		KQFTKCKLSDELKDMNGHGGITLAEWICIIFHMSGYDTETVVSNNGNREYGLFQ INNKIWCRDNENLQSRNICDISCDKFLDDDLTDDKMCAKKILDKEGIDYWLAHK PLCSEKLEQWQCEKW	14,430	123
GlyCAM1		SLNEPKDEIYMESQPTDTSAQVIMSNHQVSSEDLSMEPSISREDLVSKDDVVIK SARRHQNQNPKLLHPVPQESSFRNTATQSEETKELTPGAATTLEGKLVELTHKI IKNLENTMRETMDFLKSLFPHASEVVKPQ	15,442	137
LTF		ASKKSVRWCTTSPAESSKCAQWQRRMKKVRGPSVTCVKKTSRFECIQAISTEKA DAVTLDGGLVYDAGLDPYKLRPIAAEVYGTENNPQTHYYAVAIAKKGTNFQLNQ LQGLKSCHTGLGRSAGWNIPMGLLRPFLDWTGPPEPLQKAVAKFFSASCVPCVD GKEYPNLCQLCAGTGENKCACSSQEPYFGYSGAFKCLQDGAGDVAFVKDSTVFE SLPAKADRDQYELLCPNNTRKPVDAFQECHLARVPSHAVVARSVNGKEDLIWKL LVKAQEKFGRGKPSGFQLFGSPAGQKDLLFKDSALGLLRISSKIDSGLYLGSNY ITAIRGLRETAAEVELRRAQVVWCAVGSDEQLKCQEWSRQSNQSVVCATASTTE DCIALVLKGEADALSLDGGYIYIAGKCGLVPVLAESQQSPESSGLDCVHRPVKG YLAVAVVRKANDKITWNSLRGKKSCHTAVDRTAGWNIPMGLLSKNTDSCRFDEF LSQSCAPGSDPRSKLCALCAGNEEGQNKCVPNSSERYYGYTGAFRCLAENVGDV AFVKDVTVLDNTDGKNTEQWAKDLKLGDFELLCLNGTRKPVTEAESCHLAVAPN HAVVSRIDKVAHLEQVLLRQQAHFGRNGRDCPGKFCLFQSKTKNLLFNDNTECL AKLQGKTTYEEYLGPQYVTAIAKLRRCSTSPLLEACAFLMR	75,250	689
WAP	A V12	LAPALSLPGQAVCPELSSSEDNACIISCVNDESCPQGTKCCARSPCSRSCTVPL MVSSPEPVLKDGRCPWVQTPLTAKHCLEKNDCSRDDQCEGNKKCCFSSCAMRCL DPVTEDSFQ	12,564	117
	B M12	LAPALSLPGQAMCPELSSSEDNACIISCVNDESCPQGTKCCARSPCSRSCTVPL MVSSPEPVLKDGRCPWVQTPLTAKHCLEKNDCSRDDQCEGNKKCCFSSCAMRCL DPVTEDSFQ	12,596	117

PGRP	REDPPACGSIVPRREWRALASECRERLTRPVRYVVVSHTAGSHCDTPASCAQQA QNVQSYHVRNLGWCDVGYNFLIGEDGLVYEGRGWNIKGAHAGPTWNPISIGISF MGNYMNRVPPPRALRAAQNLLACGVALGALRSNYEVKGHRDVQPTLSPGDRLYE IIQTWSHYRA	19,143	172

Amino acid sequences of mature proteins with potential phosphorylation sites. Potential phosphorylation sites are bolded: Seryl and Threonyl residues matching the S/T-X-A motif are in red and blue, respectively. Threonyl residues matching the S/T-X-X-A motif are in green. Aa residues generated by genetic polymorphism and alternative splicing are marked bold in red and highlighted in yellow.

6.2 Complexity of the "casein" fraction: the case of α_{s2}-CN and potential impact in terms of function

Up to now, the composition of camel casein fraction appeared to be relatively well established. However, analyzing milk of camelids originating from Kazakhstan, both in *C. dromedarius*, *C. bactrianus* and their hybrids, differences, lead us to consider more subtle composition though having obvious consequences at the technological and nutritional levels. Indeed, a great diversity of molecular species, originating in genetic variants, post-translational modifications but also in the processing of primary transcripts (splicing variants), was highlighted. This situation is particularly conspicuous regarding α_{s2}-CN for which three splicing variants were identified, including exon skipping and cryptic splice site usage. Camel α_{s2}-CN was shown to be a mixture of three splicing isoforms differing in polypeptide chain length. Isoform α_{s2}-CNsv0, initially reported in the literature, was the main isoform of α_{s2}-CN. Isoforms sv1 and sv2 were splicing isoforms of α_{s2}-CN arising from alternative processing of primary transcript and differing from α_{s2}-CNsv0 with the insertion of exon 13 (ENSKKTVDT) in sv1 and an extension of exon 16 (VKAYQIIPNL) in sv2, with phosphorylation levels for each of them ranging between between 7 and 12 Phosphate groups.

With 11 potentially phosphorylated aa residues matching the S/T-X-A motif, camel α_{s2}-CN displays the highest phosphorylation level of camel caseins. To reach such a phosphorylation level, besides the nine SerP, two putative ThrP (T118 and T132) have to be phosphorylated. In all the Kazakh milk samples analyzed in LC-ESI-MS we found α_{s2}-CN with 12 P groups, as the molecular mass of 22,226 Da observed corresponds to the mass of the peptide backbone (21,266 Da) increased by 960 Da, a mass increment which coincides with 12 P groups. That means that at least another S/T residue that does not match with the

canonic sequence (S/T-X-A) recognized by the mammary kinase(s), is potentially phosphorylated.

In the camel α_{s2}-CN, two threonine residues (T39 and T129) are located in a motif S/T-X-X-E/D/pS, which is a recognition motif for CN-kinase II (CK2). Indeed, albeit the consensus sequences of CK2 and the genuine casein kinase isolated from the Golgi apparatus of the lactating mammary gland (G-CK) are definitely distinct, they could be similar and sometimes overlapping (Tibaldi et al., 2015). The hypothesis that there are two different phosphorylation systems (kinases) suggested by Bijl et al., (2014) and then by Fang et al., (2016) takes therefore a little bit more consistency. This warrants further investigation. Fam20C which seems to be the major secretory pathway protein kinase (Tagliabracci et al., 2015) is very likely responsible for the phosphorylation of S and T residues within S/T-X-A motifs, whereas at least one T residue occurring in a S/T-X-X-A motif (T39 or T129) might be phosphorylated by a CK2-type kinase.

Such results provide useful novel information for the understanding of the evolution of the casein genes and their expression across Mammals. With the growing number of genes encoding milk proteins sequenced and displaying complex patterns of splicing, thus increasing the coding capacity of genes, the extreme protein isoform diversity generated from a single gene can no longer be considered as an epiphenomenon. Is it a fortuitous or a scheduled event to expand molecular diversity of milk caseins? Structural diversity and variability in expression level are both responsible for modifications in the organization and, consequently, changes in the physico-chemical properties of the casein micelle. A parsimonious vision of this issue addresses a major question: does this convey any biological significance? It has been established, that milk proteins represent a reservoir of biologically active peptides, capable of modulating different functions; the molecular diversity generated by differential splicing mechanisms can only increase their content. Thus, alternative splicing events produce novel potentially bioactive peptides. Important new insights are expected, in this field, in the near future.

6.3 WAP: originality of the protein and of the gene in the camel species

Camel WAP is a 117 aa residues protein (136 aa residues for the pre-protein) that shows the higher sequence identity at the aa level (76%) to porcine WAP (113 aa residues). It

contains five potential phosphorylation sites per molecule (S17, S18, S19, S58, and S87), whereas the rat WAP (118 aa) and the rabbit WAP (108 aa) have only three and two potential phosphorylation sites, respectively. It was reported, that mouse WAP (115 aa) is apparently not phosphorylated (Hennighausen & Sippel, 1982). From mass data (LC-MS) it appears that only one serine can be phosphorylated. Given the extremely constrained and compact structure of the molecule with 8 S-S bridges, essential for folding and functionality of the protein, we hypothesized that S58 which is located within the additional sequence connecting the two 4-DSC domains, is the seryl residue which is alternatively (*ca.* 50%) phosphorylated in camel.

The comparison of camel WAP sequence with that of the other 5 eutherian species in which the WAP gene is expressed (pig, dog, rabbit, rat and mouse), displays an insertion of 4 aa residues (56VSSP59) which extend the sequence inter 4-DSC domains. From the *Camel dromedarius* gene sequence (GenBank 105095719) this appears to be the consequence of the usage of an unlikely intron cryptic splice site extending camel exon 3 on its 5' side by 12-nucleotides, whereas in the other 5 species the canonic 3' end of intron 2 is used. There are actually two potential intron donor splice sites responding to all requirements of splicing recognition signal: CCCGGCCAG | TCTCTTCCCCAG | AGCCTGTCCTG, and paradoxically it is the weakest site, a polypyrimidic stretch interrupted by a GG doublet, which seems to be preferentially used by the splicing machinery. Indeed, the existence of a non-allelic short isoform of camel WAP, encoded by a shorter mRNA arising from the usage, as in the other species, of the canonic 3' splice site seems plausible. This assumption was supported by the occurrence of two tryptic peptides (SCTVPLMVSSPEPVLK and SCTVPLMEPVLK) identifying camel WAP in LC-MS/MS that differentiate by the presence or absence of the tetrapeptide 56VSSP59. Such a result confirms that the usage of a cryptic splice site during the splicing of precursors to WAP mRNA is responsible for the insertion of 4 amino acid residues between the two 4-DSC domains of the camel WAP.

However, this is not the only originality of this gene that exhibits in the camel species, extremely rare fact, an intron (intron 3) of the GC-AG type. The existence of variants to the standard (canonical) GT-AG introns is known but extremely rare (Burge et al., 1999; Burset, 2000). Burset and co-workers (2001) observed that GC splice sites account for 0.5% of annotate donor sites and that GC donor sites possess a strong consensus sequence. Since the maturation process is ensured by the same splicing machinery (U2-type spliceosome), whatever the intron is, GT-AG or GC-AG type, there is a mismatch between the donor site

sequence and the U1 snRNA. To compensate for such a weakening base pairing, the consensus sequence at the GC donor site has strengthened. It should be noted that such a rare event led to an erroneous annotation of this gene in genome database, since automated algorithms based on consensus sequences are used to predict donor and acceptor sites of introns.

6.4 EVs: Beyond their role in the communication between cells, what possible effects on the consumer (newborn or adult humans)

Using an optimized isolation protocol, we obtained milk-derived EVs from camel milk samples that satisfied the typical requirements for "exosomal" morphology, size and protein content. Thus, we provide, to our knowledge, a first comprehensive proteome of camel milk-derived exosomes that appears, with *ca*. one thousand different proteins identified, wider than the camel milk proteome (391 functional groups of proteins). As previously mentioned, camel milk-derived exosomes contain proteins, such as lactadherin/MFG-E8, Ras-related proteins or CD9 also present in MFG (Saadaoui et al., 2013). In addition, our results strongly suggest that EVs isolated from camel milks have different cellular origin, since besides EVs originating from mammary epithelial cells there are very likely EVs from immune cells.

Théry et al., (2002) established specific markers of the EVs (membrane and cytosolic proteins) and a specific subset of cellular proteins targeted specifically to exosomes, the function of which still remain unknown. This is particularly interesting in relation to their possible involvement in human diseases. Therefore, the knowledge of exosome proteomics can help not only in understanding their biological roles but also in supplying new biomarkers (Raimondo et al., 2011).

If we consider that milk-derived EVs also carry microRNAs, these vesicles have to be recognized as another important bioactive component of milk that might be involved in the transfer of immune components from the mother to the newborn, but also represents a source of factors potentially responsible for the properties attributed to camelids milk and its health value. Several publications suggest that EVs in foods, specifically bovine and human milk, carry a wide range of compounds with biological activities such as, lipids, proteins, noncoding RNAs (including microRNAs), and mRNAs. Exosomes seem to be a particularly important class of EVs since they protect labile cargos against degradation and provide a vehicle for cargo uptake through endocytosis of exosomes in virtually all tissues (Benmoussa

et al., 2016; Gu et al., 2012; Izumi et al., 2012; Izumi et al., 2015; Raposo & Stoorvogel, 2013; Wolf et al., 2015; Zempleni et al., 2017).

Humans of all ages consume worldwide milk from various sources and since microRNAs are highly conserved across mammals, one can expect that microRNA from milk-derived EVs may mediate the beneficial effects of dairy milk consumption in rheumatoid arthritis (Arntz et al., 2015) or in immune functions (Melnik et al., 2014). This is probably one of the greatest challenges facing milk science in the immediate future: to provide the food industry and consumers with the basis for health-promoting properties before their inclusion as ingredients into functional foods (Martin et al., 2013).

6.5 What should be implemented now?

One of the actions that should now be implemented as a priority would be to profile the RNA content (mRNA and microRNA) of extracellular vesicles isolated from camelids milk (Bactrian and camel) so that such profiles can be compared, in particular in microRNAs, with those produced from milks of other species including large and small ruminants. This requires the implementation of functional tests on animal models in vivo and/or ex vivo (cell cultures) to evaluate potential effects of the content of EVs from camelids milk and more generally milk on the consumers' physiology. As example, we can mention the case of a polyarthritis model in mice, used by Arntz et al., (2015) to highlight attenuation effects of bovine milk-derived EVs. We can still quote the work of Chen et al., (2016) who reported that porcine milk-derived exosomes can facilitate intestinal cell proliferation and intestinal tract development, thus giving a new insight for milk nutrition and newborn development and health.

Regarding the beneficial properties of camel milk, among them its antimicrobial activity, no doubt, that it can be attributed to the high content of protective proteins such as PGRP-1 and possibly WAP. This protein first described as an anti-protease able to limit tissue damage during inflammation, displays in fact a variety of other functions, including direct antimicrobial activity. However, several studies (Alvarez-Ordóñez et al., 2013; Farrell et al., 2009; McCann et al., 2005; Recio & Visser, 1999; Zucht et al., 1995) have demonstrated the antibacterial properties of bovine peptides derived from the C-terminal part of α_{s2}-CN. Given the possible extension of the repertoire of bioactive peptides of milk camelids in connection with splicing variants arising from camel α_{s2}-CN precursors to mRNA, it would be interesting

now to test biological activities and potentially the health value of peptides derived from camel α_{s2}-CN.

References

Al haj, O. A., & Al Kanhal, H. A. (2010). Compositional, technological and nutritional aspects of dromedary camel milk. *International Dairy Journal*. https://doi.org/10.1016/j.idairyj.2010.04.003

Alhaider, A., Abdelgader, A. G., Turjoman, A. A., Newell, K., Hunsucker, S. W., Shan, B., … Duncan, M. W. (2013). Through the eye of an electrospray needle: Mass spectrometric identification of the major peptides and proteins in the milk of the one-humped camel (Camelus dromedarius). *Journal of Mass Spectrometry*, *48*(7), 779–794. https://doi.org/10.1002/jms.3213

Alvarez-Ordóñez, A., Begley, M., Clifford, T., Deasy, T., Considine, K., & Hill, C. (2013). Structure-activity relationship of synthetic variants of the milk-derived antimicrobial peptide αs2-casein f(183-207). *Applied and Environmental Microbiology*, *79*(17), 5179–5185. https://doi.org/10.1128/AEM.01394-13

Arntz, O., Pieters, B., de Oliveira, M., Bennink, M., Plem, van L., van der Kraan, P., … van de Loo, F. (2015). A8.2 Oral administration of bovine milk-derived extracellular vesicles diminishes cartilage pathology in two arthritis models. *Annals of the Rheumatic Diseases*. https://doi.org/10.1136/annrheumdis-2015-207259.187

Benmoussa, A., Lee, C. H. C., Laffont, B., Savard, P., Laugier, J., Boilard, E., … Provost, P. (2016). Commercial Dairy Cow Milk microRNAs Resist Digestion under Simulated Gastrointestinal Tract Conditions. *Journal of Nutrition*. https://doi.org/10.3945/jn.116.237651

Bijl, E., van Valenberg, H. J. F., Huppertz, T., van Hooijdonk, A. C. M., & Bovenhuis, H. (2014). Phosphorylation of αS1-casein is regulated by different genes. *Journal of Dairy Science*, *97*(11), 7240–7246. https://doi.org/10.3168/jds.2014-8061

Burge, C. B., Tuschl, T., & Sharp, P. A. (1999). Splicing of Precursors to mRNAs by the

Spliceosomes. In *The RNA World*. https://doi.org/10.1101/087969589.37.525

Burset, M. (2000). Analysis of canonical and non-canonical splice sites in mammalian genomes. *Nucleic Acids Research*. https://doi.org/10.1093/nar/28.21.4364

Burset, M. (2001). SpliceDB: database of canonical and non-canonical mammalian splice sites. *Nucleic Acids Research*. https://doi.org/10.1093/nar/29.1.255

Caroli, A. M., Chessa, S., & Erhardt, G. J. (2009). Invited review: Milk protein polymorphisms in cattle: Effect on animal breeding and human nutrition. *Journal of Dairy Science*, *92*(11), 5335–5352. https://doi.org/10.3168/jds.2009-2461

Chen, T., Xie, M. Y., Sun, J. J., Ye, R. S., Cheng, X., Sun, R. P., … Zhang, Y. L. (2016). Porcine milk-derived exosomes promote proliferation of intestinal epithelial cells. *Scientific Reports*, *6*. https://doi.org/10.1038/srep33862

Dowbenko, D., Kikuta, A., Fennie, C., Gillett, N., & Lasky, L. A. (1993). Glycosylation-dependent cell adhesion molecule 1 (GlyCAM 1) mucin is expressed by lactating mammary gland epithelial cells and is present in milk. *Journal of Clinical Investigation*. https://doi.org/10.1172/JCI116671

Fang, Z. H., Visker, M. H. P. W., Miranda, G., Delacroix-Buchet, A., Bovenhuis, H., & Martin, P. (2016). The relationships among bovine αS-casein phosphorylation isoforms suggest different phosphorylation pathways. *Journal of Dairy Science*, *99*(10), 8168–8177. https://doi.org/10.3168/jds.2016-11250

Farrell, H. M., Malin, E. L., Brown, E. M., & Mora-Gutierrez, A. (2009). Review of the chemistry of αS2-casein and the generation of a homologous molecular model to explain its properties. *Journal of Dairy Science*, *92*(4), 1338–1353. https://doi.org/10.3168/jds.2008-1711

Girardet, J.-M., Linden, G., Loye, S., Courthaudon, J.-L., & Lorient, D. (1993). Study of Mechanism of lipolysis inhibition by bovine milk proteose peptone component 3. *Journal of Dairy Science*. https://doi.org/10.3168/jds.S0022-0302(93)77551-7

Gu, Y., Li, M., Wang, T., Liang, Y., Zhong, Z., Wang, X., … Lv, X. (2012). Lactation-related microRNA expression profiles of porcine breast milk exosomes. *PLoS ONE*, *7*(8).

https://doi.org/10.1371/journal.pone.0043691

Hennighausen, L. G., & Sippel, A. E. (1982). Mouse whey acidic protein is a novel member of the family of “four-disulfide core” proteins. *Nucleic Acids Research*, *10*(8), 2677–2684. https://doi.org/10.1093/nar/10.8.2677

Hinz, K., O’Connor, P. M., Huppertz, T., Ross, R. P., & Kelly, A. L. (2012). Comparison of the principal proteins in bovine, caprine, buffalo, equine and camel milk. *Journal of Dairy Research*, *79*(02), 185–191. https://doi.org/10.1017/S0022029912000015

Izumi, H., Kosaka, N., Shimizu, T., Sekine, K., Ochiya, T., & Takase, M. (2012). Bovine milk contains microRNA and messenger RNA that are stable under degradative conditions. *Journal of Dairy Science*, *95*(9), 4831–4841. https://doi.org/10.3168/jds.2012-5489

Izumi, H., Tsuda, M., Sato, Y., Kosaka, N., Ochiya, T., Iwamoto, H., … Takeda, Y. (2015). Bovine milk exosomes contain microRNA and mRNA and are taken up by human macrophages. *Journal of Dairy Science*, *98*(5), 2920–2933. https://doi.org/10.3168/jds.2014-9076

Kappeler, S., Heuberger, C., Farah, Z., & Puhan, Z. (2004). Expression of the peptidoglycan recognition protein, PGRP, in the lactating mammary gland. *Journal of Dairy Science*, *87*(8), 2660–8. https://doi.org/10.3168/jds.S0022-0302(04)73392-5

Konuspayeva, G., Faye, B., & Loiseau, G. (2009). The composition of camel milk: A meta-analysis of the literature data. *Journal of Food Composition and Analysis*. https://doi.org/10.1016/j.jfca.2008.09.008

Martin, P., Cebo, C., & Miranda, G. (2013). Interspecies comparison of milk proteins: Quantitative variability and molecular diversity. In *Advanced Dairy Chemistry: Volume 1A: Proteins: Basic Aspects, 4th Edition* (pp. 387–429). https://doi.org/10.1007/978-1-4614-4714-6_13

Martin, P., Szymanowska, M., Zwierzchowski, L., & Leroux, C. (2002). The impact of genetic polymorphisms on the protein composition of ruminant milks. *Reproduction Nutrition Development*, *42*(5), 433–459. https://doi.org/10.1051/rnd:2002036

McCann, K. B., Shiell, B. J., Michalski, W. P., Lee, A., Wan, J., Roginski, H., & Coventry,

M. J. (2005). Isolation and characterisation of antibacterial peptides derived from the f(164-207) region of bovine αs2-casein. *International Dairy Journal*, *15*(2), 133–143. https://doi.org/10.1016/j.idairyj.2004.06.008

Melnik, B. C., John, S. M., & Schmitz, G. (2014). Milk: An exosomal microRNA transmitter promoting thymic regulatory T cell maturation preventing the development of atopy? *Journal of Translational Medicine*. https://doi.org/10.1186/1479-5876-12-43

Raimondo, F., Morosi, L., Chinello, C., Magni, F., & Pitto, M. (2011). Advances in membranous vesicle and exosome proteomics improving biological understanding and biomarker discovery. *Proteomics*. https://doi.org/10.1002/pmic.201000422

Raposo, G., & Stoorvogel, W. (2013). Extracellular vesicles: Exosomes, microvesicles, and friends. *Journal of Cell Biology*. https://doi.org/10.1083/jcb.201211138

Recio, I., & Visser, S. (1999). Identification of two distinct antibacterial domains within the sequence of bovine α(s2)-casein. *Biochimica et Biophysica Acta - General Subjects*, *1428*(2–3), 314–326. https://doi.org/10.1016/S0304-4165(99)00079-3

Restani, P., Gaiaschi, a, Plebani, a, Beretta, B., Cavagni, G., Fiocchi, a, … Galli, C. L. (1999). Cross-reactivity between milk proteins from different animal species. *Clinical and Experimental Allergy : Journal of the British Society for Allergy and Clinical Immunology*, *29*(7), 997–1004. https://doi.org/cea563 [pii]

Saadaoui, B., Henry, C., Khorchani, T., Mars, M., Martin, P., & Cebo, C. (2013). Proteomics of the milk fat globule membrane from Camelus dromedarius. *Proteomics*, *13*(7), 1180–1184. https://doi.org/10.1002/pmic.201200113

Tagliabracci, V. S., Wiley, S. E., Guo, X., Kinch, L. N., Durrant, E., Wen, J., … Dixon, J. E. (2015). A Single Kinase Generates the Majority of the Secreted Phosphoproteome. *Cell*, *161*(7), 1619–1632. https://doi.org/10.1016/j.cell.2015.05.028

Théry, C., Zitvogel, L., & Amigorena, S. (2002). Exosomes: composition, biogenesis and function. *Nature Reviews Immunology*, *2*(8), 569–579. https://doi.org/10.1038/nri855

Tibaldi, E., Arrigoni, G., Cozza, G., Cesaro, L., & Pinna, L. A. (2015). “Genuine“ casein kinase: The false sister of CK2 that phosphorylates secreted proteins at S-x-E/pS motifs.

In *Protein Kinase CK2 Cellular Function in Normal and Disease States*. https://doi.org/10.1007/978-3-319-14544-0_13

Wolf, T., Baier, S. R., & Zempleni, J. (2015). The Intestinal Transport of Bovine Milk Exosomes Is Mediated by Endocytosis in Human Colon Carcinoma Caco-2 Cells and Rat Small Intestinal IEC-6 Cells. *Journal of Nutrition*. https://doi.org/10.3945/jn.115.218586

Youcef, N., Saidi, D., Mezemaze, F., El-Mecherfi, K. E., Kaddouri, H., Negaoui, H., … Kheroua, O. (2009). Cross reactivity between dromedary whey proteins and IgG anti bovine α-lactalbumin and anti bovine β-lactoglobulin. *American Journal of Applied Sciences*, *6*(8), 1448–1452. https://doi.org/10.3844/ajassp.2009.1448.1452

Zempleni, J., Aguilar-Lozano, A., Sadri, M., Sukreet, S., Manca, S., Wu, D., … Mutai, E. (2017). Biological Activities of Extracellular Vesicles and Their Cargos from Bovine and Human Milk in Humans and Implications for Infants. *The Journal of Nutrition*, *147*(1), 3–10. https://doi.org/10.3945/jn.116.238949

Zucht, H. D., Raida, M., Adermann, K., Mägert, H. J., & Forssmann, W. G. (1995). Casocidin-I: a casein-αs2 derived peptide exhibits antibacterial activity. *FEBS Letters*, *372*(2–3), 185–188. https://doi.org/10.1016/0014-5793(95)00974-E

Acknowledgments

I would like to express my sincere gratitude to all my supervisors for making this PhD happened: Gaukhar Konuspayeva, Bernard Faye, and Patrice Martin. Special thanks for giving me the opportunity to study in France, without your valuable input this thesis would not have existed. For me it is a great honor to be your apprentice! Much appreciation to Gaukhar and Bernard for recommending me as a potential candidate for performing this PhD. I admire your passion on Camelids and invaluable contribution in development of world camel dairy industry. It is inspiring! I would like to express my very profound gratitude to Patrice for supervision; it is not possible to express all my appreciation in a few words! With guidance, effort and support you could build and raise from me a “scientist”. I value the knowledge obtained, which is my treasure. Thank you for your comprehention and patience all over the thesis.

My sincere thanks are addressed to Dr. George Erhardt and Paola Sacchi, for your acceptance to be reporters of this thesis and for your valuable suggestions and comments on the entire work. I wish to thank all jury members for contributing time.

I gratefully acknowledge PAPPSO, @BRIDGe, and MIMA2 platforms at INRA (Jouy-en-Josas) for providing necessary facilities and technical support, namely Celine Henry for assistance with LC-MS/MS analysis and Christine Longin with TEM analysis.

Very special and warm thanks belong to my colleagues from LGS “Dream team” at INRA for providing a sociable working atmosphere and assisting me in numerous ways: Patrice and Claudia for hearty welcome in the team, Nicolas, Anne, and Maelle for sport coaching in pursuit of a perfect body, Maelle for sharing common office and apartment together. You are the best colocatrice ever, thank you for your hospitality! Thanks to Leonardo and Guy for a great patience on explaining proteomic knowledge, and special thanks to Guy for providing LC-ESI-MS analysis. I wish to thank Zuzana for discovering the world of extracellular vesicles for me. Results obtained on exosomes are your merit. Nicolas thanks for performing NTA analysis. It is a precious experience for me collaborating with all you!

I would like to thank all Kazakhstani camel milk farms for providing camel milk samples and Scientific-production laboratories “Antigen” (Almaty, Kazakhstan), namely to

Moldir Nurseitova and Ali Totaev, for the assistance in sample collection. Despite the sizzling sunshine (+60°C in Kyzylorda region), we could accomplish the distance in more than 3,500 kms between four regions for a short time.

I am in debt to the Bolashak International Scholarship funded by the JSC «Center for International Programs» (Kazakhstan) for being government-sponsored during three years of PhD.

Curriculum vitae

Alma Ryskaliyeva was born on 30 August in Uralsk, Kazakhstan. In 2009, she obtained B.Sc. and, in 2011, Master degree in chemical technology of organic compounds from Al-Farabi Kazakh National University, in Almaty, Kazakhstan. She did research on exploring the terpenoid content of *Mosses* growing in Kazakhstan with emphasis on medical properties. During her master, in 2010, she undertook an internship in Vorozhtsov Novosibirsk Institute of Organic Chemistry SB RAS, in Novosibirsk, Russia. In 2014, she was awarded a Bolashak International Scholarship established by President of the Republic of Kazakhstan and enrolled in the PhD project at AgroParisTech - Université Paris-Saclay - INRA. Her research focused on the exploring the fine composition of *Camelus* milk from Kazakhstan with emphasis on protective components, and the results are presented in this thesis.

List of Publications

Peer-reviewed publications

Ryskaliyeva, A., Henry, C., Miranda, G., Faye, B., Konuspayeva, G., and Martin P. 2019. Alternative splicing events expand molecular diversity of camel CSN1S2 increasing its ability to generate potentially bioactive peptides. *Scientific Reports*. 9:5243 10.1038/s41598-019-41649-5

Ryskaliyeva, A., Henry, C., Miranda, G., Faye, B., Konuspayeva, G., and Martin P. 2019. The main WAP isoform usually found in camel milk arises from the usage of an improbable intron cryptic splice site in the precursor to mRNA in which a GC-AG intron occurs *BMC Genetics,* 20:14 https://doi.org/10.1186/s12863-018-0704-x

Ryskaliyeva, A., Henry, C., Miranda, G., Faye, B., Konuspayeva, G., and Martin P. 2018. Combining different proteomic approaches to resolve complexity of the milk protein fraction of dromedary, Bactrian camels and hybrids, from different regions of Kazakhstan. *PLOS ONE, 13(5)*. https://doi.org/10.1371/journal.pone.0197026

Manuscript in preparation

Ryskaliyeva, A., Krupova, Z., Henry, C., Miranda, G., Faye, B., Konuspayeva, G., and Martin P. 2018. Comprehensive Proteomic Analysis of Camel Milk-derived Extracellular Vesicles.

Conference proceedings

Ryskaliyeva, A., Miranda, G., Henry, C., Faye B., Konuspayeva G. and Martin P. 2017. Alternative splicing a fortuitous or a scheduled event to expand molecular diversity of milk proteins: Camel CSN1S2, a relevant model to try to provide some response elements. Student Travel Award Winner in: 14th International Symposium on Milk Genomics and Human Health 2017, 26-28 September, Quebec, Canada.

Ryskaliyeva, A., Henry, C., Miranda, G., Faye, B., Konuspayeva, G., and Martin, P. 2016. Proteomic analysis of Camelus milks from Kazakhstan. In: The ICAR 2016 Satellite Meeting on Camelid Reproduction, 1st-3rd July, Tours, France.

Individual and Training Supervision Plan

Dissemination of Knowledge

International conferences

14th International Symposium on Milk Genomics and Human Health, Quebec, Canada (oral presentation)	2017
The FSEV Annual Meeting on Extracellular Vesicles, Paris, France	2017
The ICAR Satellite Meeting on Camelid Reproduction, Tours, France (oral presentation)	2016

Seminars and Workshops

Workshop Technique of International Symposium on Microgenomics, 31 May- 1 June, Jouy-en-Josas, France	2016
Annual meeting of the doctoral candidates of the Animal Genetics Division of INRA, Toulouse, France (oral presentation)	2016
Professionnal PhD networking forum of AgroParisTech & ABIES doctoral school	2016
Annual meeting of the doctoral candidates of the Animal Genetics Division of INRA, La Rochelle, France (poster presentation)	2015
Professionnal PhD networking forum of AgroParisTech & ABIES doctoral school	2015
My professional project in 180 seconds (oral presentation)	2015

Professional Skills Support Courses

Introduction to Research Ethics and Scientific Integrity	2018
Write Right – Writing and structuring of the scientific articles, Paris, France	2015
Building Your Base, Paris, France	2014
French as a foreign language, Jouy-en-Josas, France	2014-2016

Complied with the educational requirements set by the Graduate School of Agriculture, Food, Biology, Environment and Health of the Agricultural, Veterinary and Forest Institute of France

université PARIS-SACLAY
ÉCOLE DOCTORALE
Agriculture, alimentation, biologie, environnement, santé (ABIES)

Titre : Analyse de la composition fine du lait des Camelidés du Kazakhstan en ciblant plus spécifiquement la fraction protéique

Mots clés : caséines, WAP, polymorphisme génétique, variant d'épissage, vésicules extracellulaires

Résumé : Cette étude visait à explorer la fraction protéique des laits de camélidés de différentes régions du Kazakhstan, afin d'identifier des molécules originales (peptides, protéines) potentiellement responsables des « health-promoting properties » attribuées au lait de chamelle. Différentes techniques analytiques éprouvées et approches protéomiques (SDS-PAGE, LC-MS/MS et LC-ESI-MS), ont été mises en œuvre pour de caractériser la diversité moléculaire des lactoprotéines majeures, et particulièrement des caséines. Ainsi, deux isoformes, jusqu'ici inconnues, de la caséine αs2 de chamelle, résultant d'épissages alternatifs ont pu être caractérisées. Le séquençage des transcrits correspondants a révélé que l'une d'elles résulte de l'utilisation d'un exon additionnel dont la présence a été confirmée au niveau génomique, spécifiant le peptide ENSKKTVDM. La seconde isoforme, qui résulte de l'utilisation d'un site cryptique d'épissage intronique, comporte une séquence peptidique additionnelle (VKAYQIIPNL), spécifiée par une extension intronique en aval de l'exon 16 du gène. Nous rapportons par ailleurs, l'existence d'un variant de phosphorylation de cette caséine, suggérant l'existence probable dans la glande mammaire d'un second système enzymatique impliqué dans la phosphorylation des caséines.

S'agissant des protéines du lactosérum, un nouveau variant génétique de la Whey Acidic Protein (WAP) a été identifié chez *Camelus bactrianus*. L'isoforme prédominante de cette protéine originale, exprimée dans le lait de quelques mammifères, présente une insertion de 4 résidus d'acides aminés (56VSSP59) dans le segment peptidique reliant les deux domaines 4-DSC de la protéine. Cette insertion résulte de l'utilisation d'un site cryptique d'épissage conduisant à une extension intronique de 12 nucléotides en amont de l'exon 3 du gène. De plus, nous montrons que l'intron 3 du gène de la WAP est chez les camélidés, fait extrêmement rare, un intron de type GC-AG. Enfin, nous avons isolé des laits de camélidés analysés des vésicules extracellulaires, caractérisées morphologiquement, en Microscopie Electronique à Transmission et par « Nanoparticles Tracking Analysis ». Ces vésicules qui sont impliquées dans la communication entre cellules et organismes, sont vecteurs d'informations génétiques (ARNm et microARN) mais exportent aussi de nombreuses protéines. Un millier de ces protéines, impliquées dans de multiples fonctions cellulaires et notamment dans le transport et la sécrétion des vésicules et le transport de protéines intracellulaires, ont pu être identifiées par LC-MS/MS.

Title : Exploring the fine composition of Camelus milk from Kazakhstan with emphasis on protein components

Keywords : caseins, whey proteins, genetic polymorphism, splicing variant, extracellular vesicles

Abstract: This study aimed to explore the protein fraction of camelid milks from different regions of Kazakhstan, in order to identify original molecules (peptides, proteins) potentially responsible for the health-promoting properties attributed to camel milk. Various proven analytical techniques and proteomic approaches (SDS-PAGE, LC-MS / MS and LC-ESI-MS) have been implemented to characterize the molecular diversity of the major milk proteins, particularly caseins. Thus, two isoforms, hitherto unknown, of camel αs2 casein resulting from alternative splicing have been characterized. Sequencing of the corresponding transcripts revealed that one of these isoforms results from the use of an additional exon of which the presence has been confirmed at the genomic level, encoding the ENSKKTVDM peptide. The second isoform, which results from the use of a cryptic intronic splice site, has an additional peptide sequence (VKAYQIIPNL), encoded by an intron extension downstream from exon 16 of the gene. We also report the existence of a phosphorylation variant of this casein, suggesting very likely the existence in the mammary gland of a second enzymatic system involved in the phosphorylation of caseins.

With regard to whey proteins, a new genetic variant of the Whey Acidic Protein (WAP) has been identified in Camelus bactrianus. The predominant isoform of this original protein, expressed in the milk of a few mammals, has an insertion of 4 amino acid residues (56VSSP59) into the peptide segment connecting the two 4-DSC domains of the protein. This insertion results from the use of a cryptic splice site leading to an intron extension of 12 nucleotides upstream of exon 3 of the gene. Moreover, we show that intron 3 of the WAP gene is, in camelids, an extremely rare GC-AG type intron. Finally, camelids milks were analyzed for extracellular vesicles, morphologically characterized in Transmission Electron Microscopy and Nanoparticles Tracking Analysis. These vesicles, which are involved in the communication between cells and organisms, are vectors of genetic information (mRNA and microRNA) but also export many proteins. A thousand of these proteins, involved in multiple cellular functions, in particular in transport and secretion of vesicles and transport of intracellular proteins, were identified by LC-MS/MS.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France