Food Research International 160 (2022) 111611

Contents lists available at ScienceDirect

Food Research International

journal homepage: www.elsevier.com/locate/foodres

Top-Down proteomics based on LC-MS combined with cDNA sequencing to characterize multiple proteoforms of Amiata donkey milk proteins

Barbara Auzino [a,b,c], Guy Miranda [b], Céline Henry [d], Zuzana Krupova [b], Mina Martini [a], Federica Salari [a], Gianfranco Cosenza [c], Roberta Ciampolini [a,*], Patrice Martin [b]

[a] Department of Veterinary Science, University of Pisa, Italy
[b] Université Paris-Saclay, INRAe, AgroParisTech, GABI, 78350 Jouy-en-Josas, France
[c] Department of Agricultural Sciences, University of Naples "Federico II", Portici, Italy
[d] Université Paris-Saclay, INRAe, MICALIS Institute, PAPPSO, 78350 Jouy-en-Josas, France

ARTICLE INFO

Keywords:
Equus asinus
Milk proteins
Splicing variants
Exon skipping
High-performance liquid chromatography
Mass spectrometry, casein phosphorylation

ABSTRACT

An in-depth molecular characterization of the main milk proteins, caseins (CNs) and whey proteins, from Amiata donkey combining top-down proteomic analysis (LC-MS) and cDNA sequencing revealed multiple proteoforms arising from complex splicing patterns, including cryptic splice site usage and exon skipping events. Post-translational modifications, in particular phosphorylation, increased the variety and complexity of proteoforms. α_{s2}-CN perfectly exemplifies such a complexity. With 2 functional genes, *CSN1S2 I* and *CSN1S2 II*, made of 20 and 16 exons respectively, nearly 30 different molecules of this CN were detected in the milk of one Amiata donkey. A cryptic splice site usage, leading to a singular shift of the open reading frame and generating two α_{s2}-CN I isoforms with different C-terminal sequences, was brought to light. Twenty different α_{s1}-CN molecules with different phosphorylation levels ranging between 4 and 9P were identified in a single milk sample, most of them resulting from exon skipping events and cryptic splice site usage. Novel genetic polymorphisms were detected for CNs (β- and α_s-CN) as well as for whey proteins (lysozyme C and β-LG I). The probable new β-LG I variant, with a significantly higher mass than known variants, appears to display an N-terminal extension possibly related to the signal peptide sequence. This represents the most comprehensive report to date detailing the complexity of donkey milk protein micro-heterogeneity, a prerequisite for discovering new elements to objectify the original properties of donkey's milk.

1. Introduction

During the last twenty years, the milk from equids and especially donkey's milk has been the subject of a growing interest and this trend has increased even more in recent years (Kocic et al., 2020; Derdak et al., 2020; Massouras et al., 2020). This is mainly due to its well-known cosmetic and therapeutic properties but also to its nutritional qualities. Indeed, donkey's milk is generally considered to be a good substitute of cow's milk for the feeding of newborns with cow's milk allergy (Martini et al., 2021; Monti et al., 2012; Salimei & Fantuz, 2012; Sarti et al., 2019; Vincenzetti et al., 2014). Some studies have shown its relevance in the prevention of atherosclerosis (Tafaro et al., 2007) and its ability to up regulate the immune response of healthy elderly consumers (Jirillo et al., 2010). Its antimicrobial activity, both against Gram + and Gram- bacteria, is mainly due to whey proteins such as lysozyme C (Brumini et al., 2016) and, to a less extent, to lactoferrin. However, there is a growing body of work showing that peptides encrypted in caseins (CNs) from several species, including donkey, have anti-hypertensive and antimicrobial activities (Meisel, 2004; Nagpal et al., 2011; Weimann et al., 2009). With a low protein content (11 g/L) and a high lactose content (66 g/L) donkey's milk composition is very similar to that of human breast milk, except for its fat content (8 g/L) which is about 5 times lower (Cunsolo et al., 2017; Altomonte et al., 2019). However, as human breast milk, it displays a high content of polyunsaturated fatty acids (Ragona et al., 2016) and is rich in secretory immunoglobulin A (sIgA), a major glycoprotein in milk (Gnanesh Kumar & Rawal, 2020) that plays a key role in mediating immune protection of the gut mucosa (Huang et al., 2015a). Donkey's milk is easily digested (Polidori and Vincenzetti, 2010). This is due to its low CN content, since in donkey's milk whey proteins account, as in breast milk, for up to 58%

* Corresponding author.
E-mail address: roberta.ciampolini@unipi.it (R. Ciampolini).

https://doi.org/10.1016/j.foodres.2022.111611
Received 3 March 2022; Received in revised form 8 June 2022; Accepted 29 June 2022
Available online 6 July 2022

of the nitrogen fraction (Malacarne et al., 2019). However, the protein fraction of donkey's milk appears much more complex than that of human milk (Chianese et al., 2010; Herrouin et al., 2000). Several investigations performed during the last decade, essentially using MS-based methods, recently reviewed by Cunsolo et al. (2017), revealed that, beyond the well-known variability of CN phosphorylation, the high complexity of the donkey's milk protein fraction is mainly due to the occurrence of multiple isoforms (proteoforms) resulting from differential splicing events of primary transcripts encoding CNs, as reported previously in several species including donkey (Cosenza et al., 2010)[22], goat (Leroux et al., 1992; Ramunno et al., 2005), horse (Miranda et al., 2004), pig (Suteu et al., 2011) and more recently in camelids (Pauciullo & Erhardt, 2015; Ryskaliyeva et al., 2018; Ryskaliyeva et al., 2019a). Thus, αs2-CN which is a minor component of donkey's milk and shows a great heterogeneity due to variable degrees of phosphorylation (Chianese et al., 2013), displays multiple internally deleted isoforms generated by improper splicing events (Saletti et al., 2012). Two isoforms of β-CN have been described which differ in the presence/absence of the peptide sequence ESITHINK encoded by the fifth exon of the gene, whereas four non-allelic isoforms of αs1-CN have been characterized (Cunsolo et al., 2009a). Different isoelectric focusing (IEF) patterns of which one was characterized by the absence of αs1-CN bands were observed in Ragusana breed donkeys (Criscione et al., 2009). This result was confirmed by Reversed Phase Chromatography (RP-HPLC) coupled with Mass Spectrometry (ESI-MS) analysis. However, even though our knowledge of the protein fraction of donkey's milk is beginning to become consistent, it remains relatively limited as compared to milk proteins from ruminants, especially regarding genetic polymorphisms (Caroli et al., 2009; Martin et al., 2012; Selvaggi et al., 2014a, 2014b). Indeed, few studies have been devoted to this issue and mainly concerned whey proteins (Herrouin et al., 2000; Selvaggi et al., 2015; Criscione et al., 2018). Moreover, most of these studies have been done on donkey females belonging to the Italian Ragusana breed (Cunsolo et al., 2017; Cosenza et al., 2010; Saletti et al., 2012; Criscione et al., 2018). Our knowledge of donkey genes encoding milk proteins, particularly CN genes, remains limited and scarce are the studies performed at the genome level. The first study performed at the nucleotide level, and to date the only one, was carried out once more on Ragusana donkey and concerned αs2-CN (Cosenza et al., 2010). This study revealed the existence of 2 mRNA arising from the expression of two distinct genes (*CSN1S2 I* and *CSN1S2 II*), very likely duplicated at the *CSN* locus. The first one comprises 19 exons, of which 17 are coding for a full-length αs2-CN A or I of 236 amino acid residues including the signal peptide (GenBank: CAV00691.1), corresponding to the mature protein described by Saletti et al. (2012) and related deleted isoforms. The second consists in 16 exons and encodes a 160 amino-acid predicted pre-protein named αs2-CN B or II (GenBank: CAX65660.2). Contrary to what is happening in humans, the *CSN1S2 II* gene seems to be functional in donkey, probably leading to an effective translation of an mRNA into a αs2-CN II protein and possibly related deleted isoforms that were not detected so far in donkey's milk.

The genome of *Equus asinus* is now sequenced (Huang et al., 2015b; Renaud et al., 2018), and predicted mRNA have been derived from exon/intron structure by automated computational analysis for the genes encoding the 9 major donkey milk proteins: CSN1S1, CSN1S2 I and II, CSN2, and CSN3, LALBA, LGB-I, LGB-II/PAEP and LYZ-C, paving the way for genetic studies, as those performed in the horse (Milenkovic et al., 2002; Lenasi et al., 2003; Lenasi et al., 2005). Genetic variability of the equine CN genes was recently explored in 14 breeds (Brinkmann et al., 2016) and revealed nucleotide exchanges (SNP: single nucleotide polymorphism) leading to a total of 31 putative protein isoforms predicted at the DNA level for the 4 horse CNs, thus giving a comprehensive overview of the genetic variability at the *CSN* locus in horses. An equine αs2-CN variant due to a genomic deletion, spanning 2 coding exons, has been reported as occurring in the 8 breeds analysed (Brinkmann et al., 2015). The genetic variability in the first exon of the gene encoding κ-CN (CSN3) was assessed in four Italian horse populations and in a sample of Martina Franca donkey (Selvaggi et al., 2015). In other words, our current knowledge about genetic variability of donkey CNs remains extremely limited. By contrast, this aspect is rather well documented as far as donkey whey proteins are concerned. Whereas only two genetic variants of β-LG I (named A and B) have been detected and characterized so far (Herrouin et al., 2000; Godovac-Zimmermann et al., 1988), five genetic variants of β-LG II are known: A, B, C (Herrouin et al., 2000; Godovac-Zimmermann et al., 1990), D (Cunsolo et al., 2007), and E (Chianese et al., 2013). A sixth variant (F), associated with a null or severely reduced expression, was recently reported (Criscione et al., 2018). The complete primary structures of α-lactalbumin (α-LA) A and B have been determined (Giuffrida et al., 1992) and two genetic variants of calcium-binding lysozyme (LYSC1), A (Godovac-Zimmermann et al., 1988) and B (Herrouin et al., 2000), are known nowadays.

The present work was undertaken to get an in-depth protein profile of milk from the Amiata donkey, a native breed of Tuscany weakly explored until now (Licitra et al., 2019), combining molecular biology (cDNA sequencing) and proteomic (Liquid Chromatography coupled with Electro Spray Ionization - Mass Spectrometry: LC-ESI-MS) tools with the aim to detect genetic variability and identify novel milk protein variants and proteoforms, with the view of possibly discovering *in fine* new elements to objectify the original properties of donkey's milk.

2. Materials and methods

2.1. Milk samples collection and preparation

Raw milk samples were collected from thirty healthy donkeys of the Amiata breed during morning mechanical milking. The donkeys were all between the 6th and 7th month of lactation and were housed in the same farm, raised in the free range in a semi-intensive system. Their diet consisted of grass hay (ad libitum) supplemented with a commercial concentrate (2 kg/head/day). Whole-milk samples (50 mL) were centrifuged at 2,500 *g* for 20 min at 4 °C to separate fat from skimmed milk. Samples were quickly frozen and stored at −80 °C (fat) and −20 °C (skimmed milk) until analysis.

2.2. Liquid Chromatography-Electrospray Ionization-Mass spectrometric (LC-ESI-MS)

All chemicals used in the LC-ESI-MS analysis were of the highest purity commercially available and were used without further purification. Trifluoroacetic acid (TFA), urea, Bis-Tris, dithiothreitol, and acetonitrile were purchased from Sigma-Aldrich (St. Louis, MO). Ultrapure water (Milli-Q Plus System, >18.3 MΩcm) was produced in the laboratory.

Skim milk samples (40 μL) were then clarified by adding 160 μL of 0.1 M Bis-Tris buffer pH 8.0, containing 8 M urea, 1.3% trisodium citrate, and 0.3% dithiothreitol (Visser et al., 1991). Twenty μL of clarified milk samples were then injected onto a Discovery BIOWide Pore (Supelco) C5 column (150 × 2.10 mm, 300 Å). Reversed-phase HPLC was carried out with an Ultimate LC 3000 system (Thermo Fisher Scientific, Waltham, MA). During the analysis, the autosampler was kept at 10 °C, and the column was maintained at 42 °C. The column's mobile phase consisted of a gradient mixture of solvent A (0.025% TFA in ultrapure water, vol/vol) and solvent B (0.02% TFA in acetonitrile, vol/vol) at a flow rate of 0.2 mL/min. The elution conditions were successive linear gradients: from 29.5 to 34% B in 16 min, from 34 to 35.5% B in 0.1 min, from 35.5 to 37.5% B in 14.9 min, from 37.5 to 42% B in 14 min, from 42 to 95% B in 0.1 min, followed by an isocratic elution at 95% B for 5 min, and a linear return to 29.5% B in 0.1 min. The column was then re-equilibrated at 29.5% B as the starting condition for 10 min.

Protein elutes were detected by UV absorbance at 214 nm. The column was directly interfaced with an ESI-TOF mass spectrometer micrOTOF II focus (Bruker Daltonics, Wissembourg, France). The positive ion mode was used and mass scans were acquired over a range of 50 to 3,000 *m/z*. End plate offset voltage was set at − 500 V and capillary voltage to 4,500 V. Nebulizer gas (N2) pressure was maintained at 250 kPa and drying gas (N2) flow was set at 4.0 L/min at 200 °C. The LC-ESI-MS system was controled by Hystar software v.2.3 (Bruker Daltonics). The charge number of multicharged ions, the deconvoluted mass spectra, and the determination of average molecular mass (Mr) were obtained from Data Analysis v.3.4 software (Bruker Daltonics). A blank sample (clarification solution) was injected after every milk sample to avoid carryover effects. A reference milk sample was analyzed after every 10 milk samples to determine reproducibility.

We implemented the LC-ESI-MS method developed at INRAe (UMR1313 GABI, domaine de Vilvert, 78,350 Jouy-en-Josas) to simultaneously measure the relative concentrations of the major milk proteins and their isoforms, notably their phosphorylation isoforms (Miranda et al., 2013; Miranda et al., 2020). Protein variants and isoforms of the major milk proteins (αs1-CN, αs2-CN I, αs2-CN II, β-CN, κ-CN, α-LA, β-LG I and β-LG II) were identified by matching measured molecular masses with an in-house calculated/theoretical mass database built for donkey milk proteins.

2.3. SDS-PAGE analysis and Bottom-up identification

Both major and low-abundant proteins resolved by SDS-PAGE (1D sodium dodecylsulfate polyacrylamide gel electrophoresis) were identified after excision by mass analysis of the tryptic hydrolysate. The SDS-PAGE used was based on that devised by 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$^{\text{TM}}$ 10,000 XL, Sweden), resolved bands were excised from the gel and submitted to digestion by trypsin. Gel pieces were first washed for 15 min with an ACN/100 mM NH4HCO3 mixture (1:1). Digestion was performed in 50 mM NH4HCO3 pH 8.0 with 0.1 µg modified trypsin (Promega, sequencing grade) per sample, for 6 h at 37 °C. Tryptic peptides were then analyzed by tandem-mass spectrometry (LC-MS/MS), using UltiMate™ 3000 RSLC nano System (Thermo Fisher Scientific) coupled either to LTQ Orbitrap XL™ Discovery mass spectrometer or QExactive (Thermo Fischer Scientific), as described (Ryskaliyeva et al., 2018).

2.4. Milk fat globule collection and RNA extraction

Milk was centrifuged at 2,500 × *g* for 20 min to pellet somatic cells (SC) and to separate the upper milk fat globule (MFG) fraction. The MFG fraction was mixed with Trizol LS (Invitrogen) and heated briefly at 30̊C while shaking, to emulsify fat. Total RNA was extracted from milk fat following the protocol from the manufacturer, as described (Brenaut et al., 2012).

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) 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 37̊C to remove RNA from heteroduplexes. Single-strand cDNA thus obtained was stored at −20̊C. cDNA samples covering the entire coding regions of caseins were amplified. PCR was performed in an automated thermocycler GeneAmp1 PCR System 2,400 (Perkin-Elmer, Norwalk, USA) with GoTaq1 G2 Flexi DNA Polymerase Kit (Promega Corporation, USA). Reactions were carried out in 0.2 mL thin-walled PCR tubes with flat cap strips (Thermo Scientific, UK), in 5X Green or Colorless GoTag1 Flexi Buffer, $MgCl_2$ solution 25 mM, PCR Nucleotide Mix 10 mM each, GoTag1 G2 Flexi DNA Polymerase (5 U/µL), 10 mM each oligonucleotide primer, template DNA and nuclease-free water, up to a 50 µL final volume. Primer pairs, purchased from Eurofins (Eurofins genomics, Germany), were designed using published *Equus asinus* nucleotide (genome and cDNA) sequences (**Supplementary material, Table S1**). Sequencing of PCR fragments was performed with primers used for PCR from both strands, according to the Sanger method, by Eurofins MWG GmbH (Ebersberg, Germany). PCR products were cloned using the TOPO Cloning Reaction Kit (Invitrogen, Life Technologies Inc., Carlsbad, CA), following the manufacturer's instruction protocole.

3. Results

Thirty individual donkey milk samples 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 (mainly phosphorylation) isoforms were determined by deconvoluting multiple charged ion spectra in a true mass scale. By knowing their primary structures, it becomes possible to determine theoretical molecular masses of non-post-translationally modified proteins, and then we can precisely determine the mass of phosphorylation isoforms resulting from the addition of phosphate groups (+79.98 Da for each phosphate residue).

Likewise, masses of isoforms arising from cryptic splice site usage, frequently leading to the loss of the first codon of an exon (for example CAG specifying a glutamine (Q) residue: −128.13 Da), are easily deduced. Sequencing of messenger RNAs isolated from milk fat globules confirmed such events as well as exon skipping and much complex splicing events thus allowing to resolve the molecular complexity of Amiata donkey milk proteins. A donkey mass reference database was then created for the main milk proteins using the data available in UniProtKB (ExPASy SIB Bioinformatics Resource Portal) and the National Center for Biotechnology Information (NCBI). To illustrate the efficiency and limits of such an approach, a typical protein profile obtained with a donkey milk is shown in Fig. 1, and the identification of proteins from observed molecular masses is given in Table 1.

In fraction I (peak 4) we found a molecular mass of 14,638.39 Da (Fig. 2), with a signal intensity of 24,877, that we identified as the calcium-binding lysozyme C (milk isozyme or LYZ-C1). This molecular mass is ascribable to a new variant, named C (M_r = 14,638.56 Da), revealed from cDNA sequences and characterized by two amino acid substitutions at positions 52 (Y52S) and 61 (S61N) in the peptide chain, as compared to variant A (M_r = 14,687.63 Da) initially reported (Godovac-Zimmermann et al., 1988). Variant C differs from the B variant (Herrouin et al., 2000) by the D/N amino acid substitution in position 49 (Fig. 3). Variant C appears to be very predominant in the Amiata population, since the 30 donkeys analysed were homozygous C/C.

Molecular masses detected in fraction II (peaks 7 and 8), were assigned to an αs2-CN encoded by the second *CSN1S2* gene (alternatively named A, a, like or II, according to the species), since observed masses did not match with any mass derived from the protein recorded in the UniProtKB database (C1L3G3). To achieve this, we cloned and sequenced the cDNA coding for this protein. Twenty eight of the thirty milk samples analysed by LC-ESI-MS showed in peak 7 the presence of the same two proteins with molecular masses of 16,956.68 Da and 17,036.72 Da. From cDNA sequencing experiments (Fig. 4) we deduced that these masses originated in the translation of an mRNA in which exons 3 and 12′ (numbering of bovine *CSN1S2* gene exons, adopted for peptide sequence comparison purposes; additional exon, as compared with the bovine gene, are numbered with a') were lacking - very likely skipped during the course of primary transcripts maturation - and in which a supplementary exon (presence confirmed at the donkey genome level, 91 nucleotides downstream exon 9; NCBI Gene ID: 106828076), corresponding to exon 10, was observed. This exon was absent from the

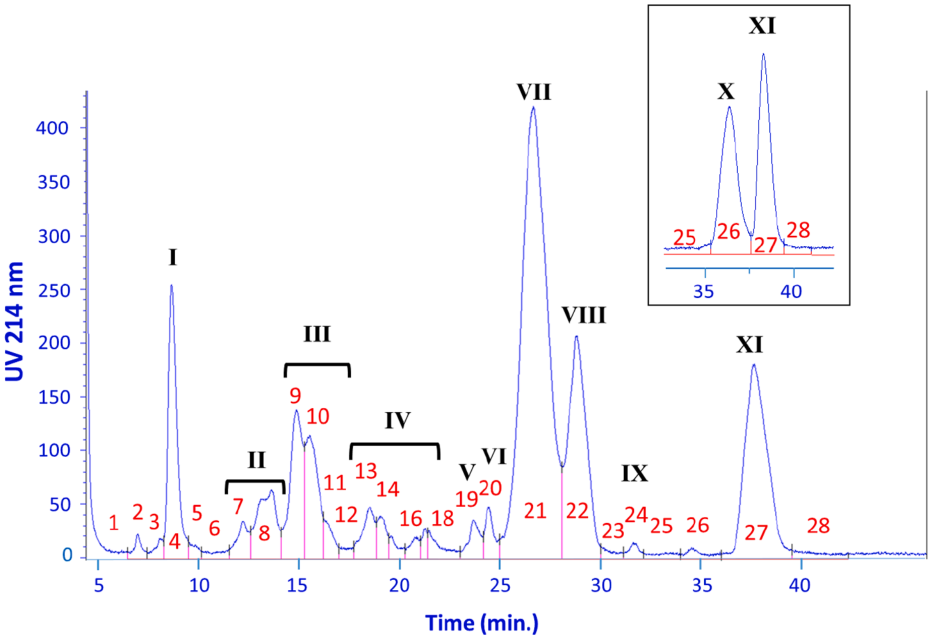


Fig. 1. Typical RP-HPLC profile of an individual Amiata donkey milk. Milk proteins were resolved into 28 peaks (in red) corresponding to 11 major milk protein fractions (Roman numbering), identified from MS data (Table 1) in the following order: Fraction I (peak 4) contained Lysozyme C; Fraction II (peaks 7 and 8) contained α_{s2}-CN II; Fraction III (peaks 9–11) contained α_{s1}-CN; Fraction IV (peaks 13–16) contained α_{s2}-CN I; Fraction V (peak 19) contained γ2-CN f [105–226] and [112–226]; Fraction VI (peak 20) contained β-lactoglobulin II,]; Fraction VII (peak 21) contained β-CN; Fraction VIII (peak 22) contained α-lactalbumin; Fraction IX (peak 24) contained γ-CN f [56–226] and Fraction XI (peak 27) contained β-lactoglobulin I. In some milk samples an additional fraction (X, peak 26) putatively ascribed to a novel variant of β-lactoglobulin I was observed (framed profile top right).

cDNA sequences previously reported by Cosenza et al. (2010). On the other hand, we were not able to detect in the 10 clones sequenced (3 clones from each of the first 2 donkeys and 4 from the third one, randomly taken), the duplication AAA CAG TTG coding for the tripeptide KQL found by these authors, at the beginning of exon 11. In addition, all cDNA sequences showed a transversion G/T at position 86 in exon 16, responsible for the amino acid substitution S143I (numbering of the full-length protein) when compared with the mRNA recorded in the EMBL database (FN298386). Finally, these two proteins, in which peptides EIKHVSSSE and EFTSISQE (encoded by exons 3 and 12′, respectively) are lacking, and displaying the additional sequence KIELTKEEKLYLKQL, encoded by exon 10 (quite perfectly duplicated as exon 15: EIELSDEEKNYLKQL), have the same peptide chain (140 amino-acid residues) with 8 and 9 phosphate groups, respectively. In some milk samples, we also found a mass of 17,117 Da, likely corresponding to the same αs2-CN II isoform, with 10 phosphate groups.

In peak 8, we found molecular masses of 18,199.54 Da, 18,279.35 Da and 18,358.85 Da, with successive increments of 80 Da (mass of one phosphate group: 79.98 Da), signing a membership to the phosphoprotein (casein) family (Fig. 5). Once again, from cDNA sequences, these molecular masses were attributed to αs2-CN II molecules (148 amino acid residues: aa), with three successive phosphorylation levels (12P, 13P and 14P), in which the peptide EIKHVSSSE (encoded by exon 3) was absent. A fourth molecule was detected in peak 8 with a lower molecular mass (16,016.42 Da) that should very likely correspond to another αs2-CN II isoform of 134 aa in which the peptide encoded by exon 3 would be present whereas peptides encoded by exons 10 and 12′ would be absent, with a relatively low phosphorylation level (7 phosphate groups, theoretical mass 16,015.78). However, we did not find any cDNA clone corresponding to such an isoform, no more than a full-length mature molecule (157 aa), derived from a transcript comprised of all the coding exons, including exon 10. On the other hand, the smallest molecule found (14,422.18 Da) was shown, to correspond to the translation of a messenger RNA in which exons 12′, 14, 15 and the first 33 nucleotides of exon 16 were missing (Fig. 4), with a high phosphorylation level (13 phosphate residues) leading to a theoretical molecular mass of 14,423.48 Da (Table 1). Such a phosphorylation level suggests however the full phosphorylation of the 11 Seryl residues occurring in a S-X-A sequence within this 115-amino acid long peptide chain and that 2 of the 3 Threonyl residues involved in a T-X-A sequence, in which A is an acidic residue should be phosphorylated too. Compilation of mass signal intensities observed for the different isoforms of this protein confirms that αs2-CN II is actually a minor component of donkey milk (less than 2% of caseins) and that the second *CSN1S2* gene (*CSN1S2 II*) that codes for this protein is nevertheless active, conversely to its human counterpart (*CSN1S2A*; Rinjkels, 2002), albeit weakly expressed.

Molecular masses detected in fraction III, composed of peaks 9, 10 and 11 of the chromatogram (Fig. 1, Table 1), were assigned to the αs1-CN family. Peak 9 contained αs1-CN molecules, corresponding to the dephosphorylated A variant (24,406 Da) reported by Cunsolo et al. (2009a), with 6 (24,886.46 Da) and 5 (24,806.98 Da) phosphate groups, respectively, in which the octapeptide DTSNESTE, encoded by exon 7, is lacking. In addition, we found molecular masses that fit perfectly with that of the dephosphorylated isoform, named A1 (24,278 Da) by Cunsolo et al. (2009a), in which Q88 (Q96 when exon 7 is included in the relative mRNA) encoded by the first codon (CAG) of exon 11 is skipped out, with 5 (24,678.76 Da) and 6 (24,758.32 Da) phosphate residues. The mass differential observed between 24,630.03 Da (a fifth molecule detected in peak 9) and 24,758.32 Da (αs1-CN var. A [-e7] [-Q] 6P) strongly suggests that the removal of a second glutamine (Q) residue (-128.13 Da) could occurred and therefore likely implies the use of another cryptic splice site that could concern the CAG encoding Q47 at the beginning of exon 6′. The mass 24790.78 Da, which is 160 Da more, could logically correspond to the same peptide chain (αs1-CN var. A [-e7] [-2Q], with 2 additional phosphate residues, *i.e.* 8P. The ultimate mass (25,001.07 Da) found in this peak was ascribed to a splicing isoform of αs1-CN A in which exon 5 (encoding the pentapeptide HTPRE) is lacking and having lost the Q88/96 residue with 6P (25,001.28 Da). In peak 10 were detected two masses (24,265.77 and 24,137.67 Da) corresponding to αs1-CN A in which exons 5 and 7 were missing with 6P (Table 1), differing in the presence of a glutamine residue at position Q83/91, *i.e.* the B1 isoform with 6P described by Cunsolo et al. (2009a) Besides, for the first time, two masses (25,910.43 and 25,993.23 Da) ascribable to full-length (including exon 7) αs1-CN were identified, with 8P and 9P, respectively. Masses of 25,290.03 Da (8P) and 25,369.16 Da (9P) detected in peak 11, were assigned to splicing variants of αs1-CN in which the pentapeptide encoded by exon 5 is lacking, as well as a second set of masses (25,161.68 Da and 25,241.97 Da) corresponding to the same peptide chain in which a glutamine residue (very likely Q83/91) is lost.

Table 1
Identification of donkey milk proteins from observed molecular masses using LC-ESI-MS.

Fraction	Peak	Intensity	Observed M_r Da	Theoretical M_r Da	Protein description	UniProt/NCBI GeneBank Accession number
I	4	24,877	14,638.9	14,638.56	Lysozyme C Milk var C	P17897
II	7	474	16,956.68	16,956.94	αs2-CN II [-e3] [-e12′] var C 8P	
		212	17,036.72	17,036.92	αs2-CN II [-e3] [-e12′] var C 9P	
	8	1371	18,358.80	18,358.78	αs2-CN II [-e3] var C 14P	
		1280	18,279.22	18,278.80	αs2-CN II [-e3] var C 13P	
		1036	16,016.42	16,015.78	αs2-CN II [-e10] [-e12′] var C 7P	
		1031	18,199.31	18,198.82	αs2-CN II [-e3] var C 12P	
		896	14,422.18	14,423.49	αs2-CN II - [-e12′] [-e14] [-e15] [-e16/3′] var C 13P	
III	9	15,008	24,886.46	24,886.29	αs1-CN A [-e7] 6P	P86272
		4056	24,758.32	24,758.16	αs1-CN A [-e7][-Q] 6P	
		2131	24,630.03	24,630.03	αs1-CN A [-e7][-2Q] 6P	
		1289	24,806.98	24,806.31	αs1-CN A [-e7] 5P	
		1110	24,678.76	24,678.18	αs1-CN A [-e7][-Q] 5P	
		860	25,001.07	25,001.28	αs1-CN A [-e5][-Q] 6P	
	10	8408	24,265.77	24,265.63	αs1-CN A [-e5][-e7] 6P	
		7746	24,137.67	24,137.49	αs1-CN A [-e5][-e7][-Q] 6P	
		4293	25,910.43	25,910.04	αs1-CN A 8P	
		2015	24,009.41	24,009.35	α αs1-CN A [-e5][-e7][-2Q] 6P	
		1373	25,991.01	25,990.02	αs1-CN A 9P	
	11	2059	25,290.03	25,289.38	αs1-CN A [-e5] 8P	
		972	25,161.68	25,161.24	αs1-CN A [-e5][-Q] 8P	
		693	25,241.97	25,241.22	αs1-CN A [-e5][-Q] 9P	
		552	25,369.16	25,369.36	αs1-CN A [-e5] 9P	
IV	13	2103	26,163.77	26,163.09	αs2-CN I [-e16/3′] 12P	B7VGF9
		2027	26,083.95	26,083.11	αs2-CN I [-e16/3′] 11P	
		900	25,843.29	25,843.17	αs2-CN I [-e16/3′] 8P	
		895	26,243.58	26,243.07	αs2-CN I [-e16/3′] 13P	
		686	26,003.39	26,003.13	αs2-CN I [-e16/3′] 10P	
		622	25,763.45	25,763.09	αs2-CN I [-e16/3′] 7P	
	14	2560	22,593.03	22,593.20	αs2-CN I [-e4][-e5][-e6][-e16/3′] 11P	
		1185	26,003.80	26,003.13	αs2-CN I [-e16/3′] 10P	
		731	26,082.79	26,083.11	αs2-CN I [-e16/3′] 11P	
		417	22,352.96	22,353.26	αs2-CN I [-e4][-e5][-e6][-e16/3′] 8P	
		167	25,923.63	25,923.15	αs2-CN I [-e16/3′] 9P	
	15	1324	22,513.09	22,513.22	αs2-CN I [-e4][-e5][-e6][-e16/3′] 10P	
		156	26,081.25	26,083.11	αs2-CN I [-e16/3′] 11P	
		127	22,593.51	22,593.20	αs2-CN I [-e4][-e5][-e6][-e16/3′] 11P	
	16	553	26,990.40	26,989.95	αs2-CN I 12P	
		454	26,909.40	26,909.97	αs2-CN I 11P	
		364	23,420.78	23,420.06	αs2-CN I [-e4][-e5][-e6] 11P	
		357	26,830.56	26,829.99	αs2-CN I 10P	
		347	23,338.74	23,340.08	αs2-CN I [-e4][-e5][-e6] 10P	
		314	27,071.58	27,069.93	αs2-CN I 13P	
		208	26,750.14	26,750.01	α αs2-CN I 9P	
V	19	2597	12,884.13	12,884.13	γ2-CN f [112–226] var. B	
		1348	13,567.13	13,565.96	γ2-CN f [105–226] var. B	
VI	20	3394	18,226.87	18,226.56	β-lactoglobulin II var B	ADG21873.1/P19647
VII	21	65,553	26,019.75	26,019.24	β-CN var. B* 6P	D2EC27 XP_014708641.1
		62,906	26,099.81	26,099.22	β-CN var. B 7P	
		24,996	25,939.76	25,939.26	β-CN var. B 5P	
		13,531	25,015.66	25,016.26	β-CN var. B [-e5] 5P	
		4158	25,860.15	25,859.28	β-CN var. B 4P	
VIII	22	49,371	14,222.24	14,222.30	α-lactalbumin	P28546
IX	24	3823	19,147.73	19,147.57	γ-CN f [56–226] var. B	
XI	27	111,985	18,514.44	18,514.25	β-lactoglobulin I var.B	P13613

*The variant observed here is different from variant A first described by Cunsolo et al. (2009a) in Ragusana breed. We propose to call it variant B. It corresponds to the product of the genomic sequence XP_014708641.1 (NCBI) published by a Chinese consortium (Dehzou breed).

To summarize, we were able to identify 20 different αs1-CN molecules, in this sample, most of them resulting from differential splicing events (exon skipping) in which exons 5 and 7 are involved. In addition, cryptic splice site usage involving CAG codons at the 5′ end of exons 11 and 6′ were also found as well as different phosphorylation levels ranging between 4 and 9P. In the overwhelming majority of samples, the predominant isoforms, namely αs1-CN A [-e7] 6P, αs1-CN A [-e5][-e7] 6P and αs1-CN A [-e5][-e7][-Q] 6P, correspond to splicing variants, with the same level of phosphorylation (6P). However, in some individuals, full-length isoforms (210 amino acid residues) with 8P can sometimes be part of the quantitative top trio. Isoforms arising from transcripts in which a deletion of exon 14 occurred, whether or not associated with another splicing abnormality, were also detected. While the A allele is clearly predominant, with a frequency reaching 73% in the population

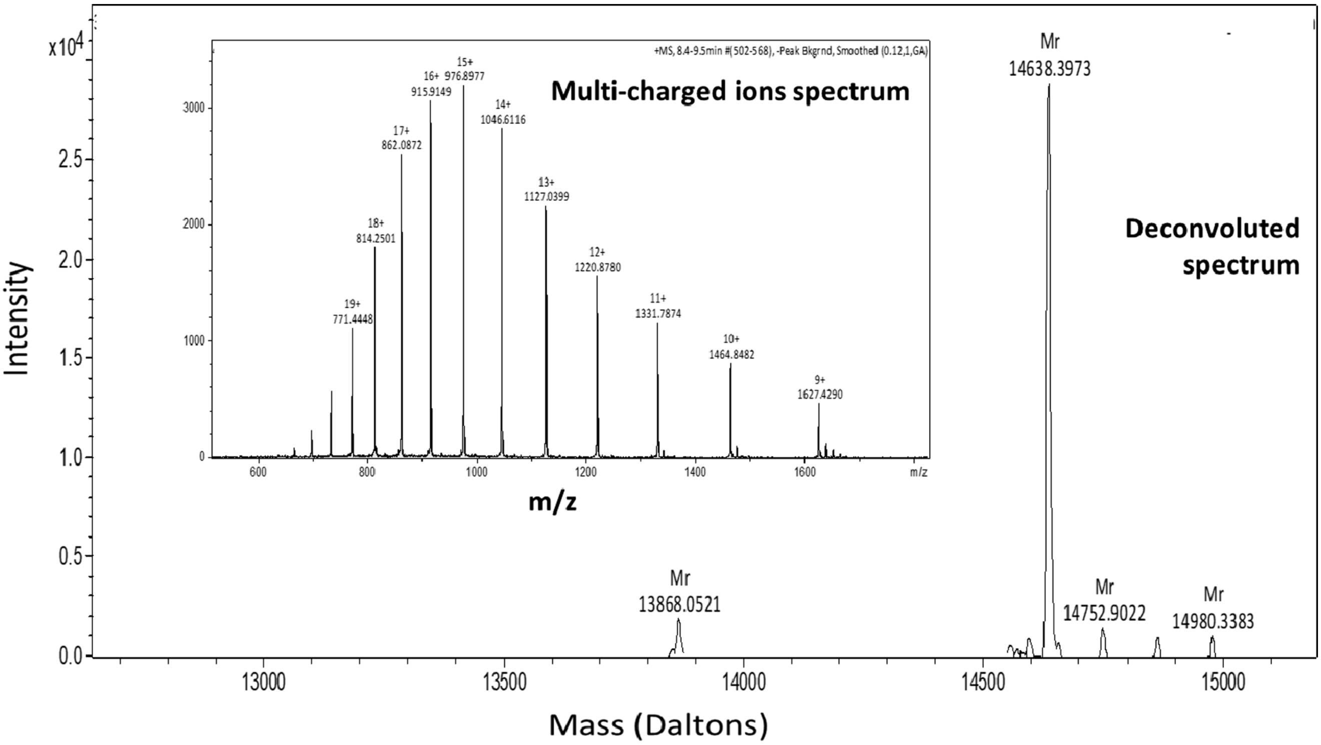


Fig. 2. ESI-TOF/MS spectral data (Multi-charged cluster ions and deconvoluted spectra) relative to peak 4 (corresponding to milk-derived Lysozyme C), resolved through RP-HPLC from an individual Amiata donkey milk sample (see Fig. 1).

analysed, we observed, in peak 9 from 12 milk samples, masses (24,872 and 24,913 Da) different from that of variant A (24,886 Da) in which exon 7 is skipped, with 6 phosphate groups, either at the hetero or homozygous state, suggesting the existence of two novel genetic variants that we propose to name D and E, respectively (Fig. 6). There is a strong argument in favour of such a hypothesis, given the existence of two series of parallel masses presenting the same mass differential, following the degree of phosphorylation. They are currently being characterized further (in progress).

Even more complex and challenging was fraction IV, gathering peaks 13 to 16 that all contained molecules derived from “true” αs2-CN (also named I or B) with different phosphorylation level and arising from several splicing events (Fig. 1, Table 1). The six masses found in peak 13 were shown to correspond to the same peptide chain, at different phosphorylation levels ranging between 13P (26,243.58 Da) and 7P (25,763.45 Da). This was deduced from cDNA sequencing (clone As2-D3, $\alpha s2^{\Delta 16\ \mathrm{part}}$) that revealed a cryptic splice site usage leading to a partial loss of exon 16 (35 nucleotides at its 3’ end) in the mature mRNA, during primary transcripts processing (Fig. 7). cDNA sequencing, in agreement with the gene sequence available (NCBI, Gene ID: 106835119), confirmed that the mRNA thus generated preserves an open reading frame in spite of a frame shift which ultimately results in the occurrence of a TAA stop codon 12 nucleotides downstream in exon 17, leading to the following C-terminal sequence: 206HKVLLRFLN214 (Fig. 7). Molecular masses of 22,593.03 Da and 22,512.09 Da, which represent respectively the major compounds of peaks 14 and 15, are ascribable to an isoform comprising 183 amino acid residues, with 11 and 10P, respectively. They result from the translation of a messenger RNA (clone As2-D2) displaying the same partial skipping of exon 16, in its 3’ end and in which exons 4, 5 and 6 are skipped “en bloc”. Similar, but not identical isoforms, also characterized by partial skipping events affecting exon 16, either on its 5’ or its 3’extremity, and giving rise to proteins in which peptides 176NKINQ180 or 212YQIIPVL218 were missing, together or not, with the absence of the peptide 12DSVNISQEKFKQEKYVVIPTSKESICSTSCE42, encoded by exons 4, 5 and 6, were previously reported. Surprisingly, the use of a cryptic splice site (GTAAG) downstream exon 16 that results in a modification of the C-terminal sequence of αs2-CN I, as we observed, gives a mass for the non-phosphorylated protein (25,203.33 Da) which is compatible with a putative deletion of the peptide 212YQIIPVL218, as proposed (Saletti et al., 2012). However, such a deletion is not consistent with the nucleotide sequences both of the cDNA (Fig. 7) and of the gene (NCBI, LOC106828076).

On the other hand, there is no obvious explanation, from the nucleotide sequence, to account for the deletion of the peptide 176NKINQ180 that the results of Saletti and co-workers seemed to suggest. Peak 16 contained the full-length αs2-CN (221 aa) with phosphorylation levels ranging between 9P (26,338.74 Da) and 13P (27,071.58 Da), in agreement with the sequence found by Cosenza et al. (2010). Moreover, in peak 16 are also present isoforms with molecular masses of 23,338.74 Da (10P) and 23,420.78 Da (11P) arising from the skipping of exons 4, 5 and 6 (“en bloc”), in which the peptide 12DSVNISQEKFKQEKYVVIPTSKESICSTSCE42 is missing. Finally, a minimum of 17 different αs2-CN I molecules from a single gene (*CSN1S2 I*) were identified. Isoforms derived from transcripts in which the last 35 nucleotides of exon 16 are missing are very largely the most abundant, particularly the isoform from transcripts comprising exons 4, 5 and 6 (αs2-CN I from clone As2-D3 with 11 and 12P). These molecules have a C-terminal sequence different from that of the full-length αs2-CN that represents between 15 and 25% of the total αs2-CN, according to individuals.

In fraction V (peak 19, Fig. 1, Table 1), two molecular masses (12,884.13 and 13,567.13 Da), corresponding to degradation products of β-CN [f(112–226) and f(105–226)], the so-called γ2-CN, were found. Other masses corresponding either to the complement of γ2-CN or to γ6-CN [f(56–226)] were found in peaks 17 and 24 respectively (see thereafter).

Regarding fraction VI (peak 20, Fig. 1, Table 1), the main mass found (18,226.87 Da) with an intensity of 3394, did not fit well with β-lactoglobulin II (β-LG II) variant A (UniProtKB - P19647), of which the

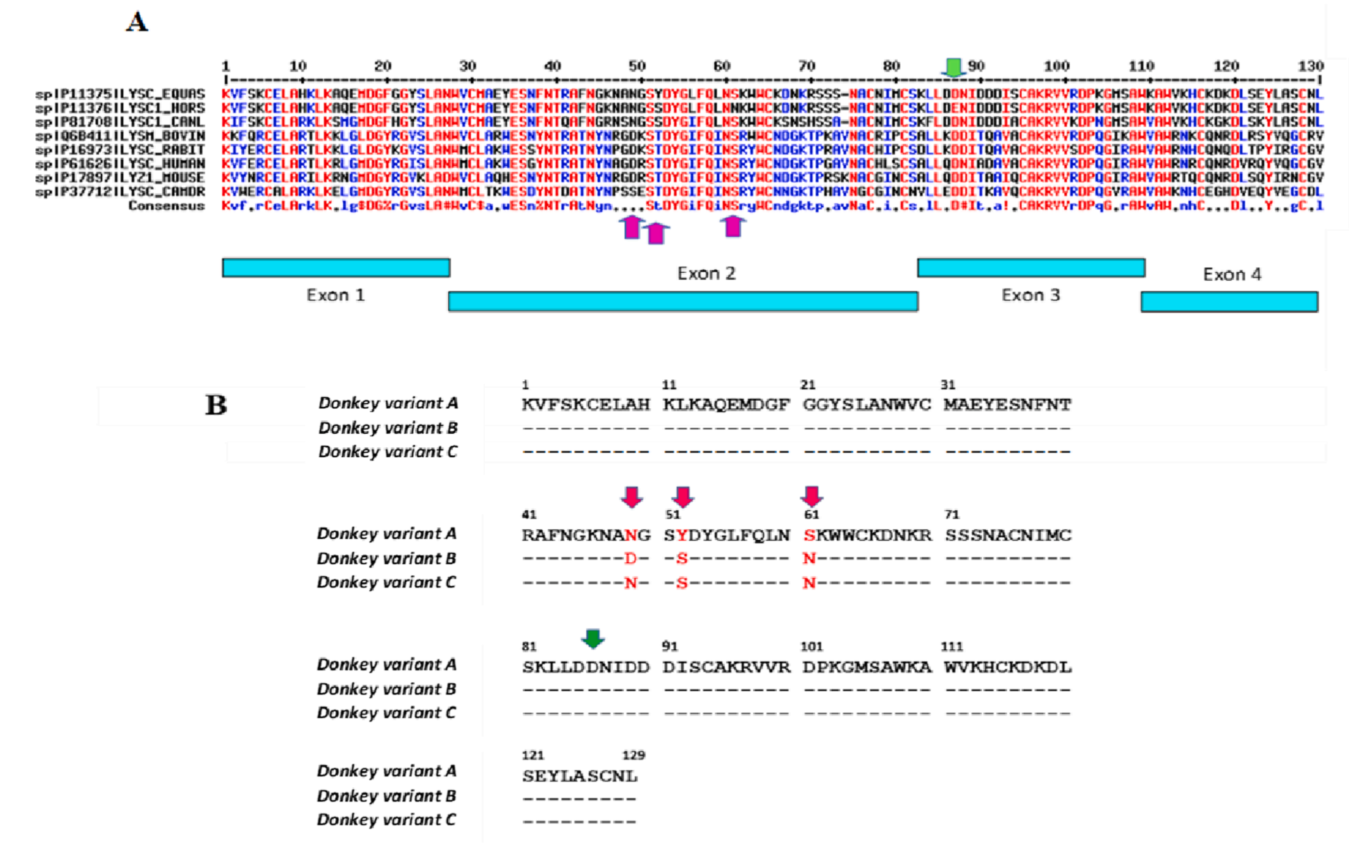


Fig. 3. Sequences of milk-derived Lysozyme across species and genetic polymorphisms in donkey. A. Multiple alignments of milk-derived isozyme of LYZ-C (calcium-binding lysozyme, Lyz-C1) across Mammalian species. Mutations differentiating donkey variants (pink vertical arrows) are found in the second exon of the gene of which the coding exon organization is given. B. Donkey LYS-C1 variant A (14,687.63 Da) designates the first milk-derived isozyme LYS-C1 variant reported by Godovac-Zimmermann et al. (1988). Donkey LYS-C variant B (14,639.54 Da) corresponds to the LYS-C variant reported by Herrouin et al. (2000) and LYS-C variant C designates the new genetic variant found in this study. Amino acid substitutions at positions 49, 52 and 61 are in bold (pink vertical arrows). The donkey variant C sequence is the closest to the horse sequence, with a single mutation D86E (green vertical arrow) that differentiates horse from all the other species.

molecular mass is 18,262.67 Da. However, it matches perfectly well with the B variant (18,226.56 Da) reported by Herrouin et al. (2000). The other 30 donkey milk profiles revealed the occurrence in peak 20 of proteins with such a molecular mass, as well as a molecular mass close to 18,241 Da, corresponding to the third genetic variant (C) of β-LG II, previously found in the commune species of donkeys from the south of France (Herrouin et al., 2000). Masses mostly found in the Amiata breed are those of variants B (18,226.56 Da), D (18314.68 Da) and C (18,240.59 Da) of which the allelic frequencies were 27.8, 27.8 and 44.4%, respectively. The milk samples from three of the 30 individuals analysed were apparently devoid of β-LG II. These 3 donkeys would possibly be homozygous for a “null” allele or a very weakly expressed allele, such as the F allele recently described (Criscione et al., 2018).

Fraction VII (peak 21, Fig. 1, Table 1) contained the theoretical molecular mass of a full-length β-CN (226 amino acid residues), with phosphorylation levels ranging between 4P (25,860.15 Da) and 7P (26,099.81 Da). These values did not match with the mass previously reported for the A variant (Cunsolo et al., 2009b). They are, in the other hand, in agreement with the molecular mass of the protein deduced from the genomic sequence (NCBI, XP_014708641.1), *i.e.* 10 Da higher than the mass reported by Cunsolo et al. (2009b). This β-CN variant was termed B. It is worth noting that this casein shows the highest intensity signal, with 65,553 and 62,906 for β-CN B with 6 and 7P, respectively. Besides, we detected a molecular mass of 25,015.66 Da, imputed to a splicing isoform displaying a mass differential of 923 Da relatively to the full-length protein, due to the loss in the mature mRNA of exon 5 encoding the octapeptide 27ESITHINK34, and termed B$^{\Delta5}$ by analogy to the nomenclature used by Cunsolo and co-workers.

Fraction VIII (peak 22, Fig. 1, Table 1) associated with a molecular mass of 14,222.24 Da, with a signal intensity of 49,371, was easily identified as α-lactalbumin given it matched perfectly with the mass (14,222.30 Da) recorded in the UniProtKB database (P28546) for *Equus asinus* LALBA.

A molecular mass of 19,147.73 Da with a mass signal intensity of 3,823 was found in peak 23 (fraction IX, Fig. 1, Table 1) of each of the 30 milk samples analysed. This molecule of which the identity remains to be confirmed is very likely ascribable to a further γ-CN-type degradation product corresponding to the f(56–226) of β-CN variant B whose theoretical mass is 19,147.57 Da.

Finally, the protein corresponding to peak 27 (Fraction XI, Fig. 1, Table 1) with a molecular mass of 18,514.44 Da was easily identified as being the B variant of β-LG I. However, 6 donkey milk samples displayed chromatographic profiles in which peak 27 was preceded by an additional peak (peak 26, see Fig. 1 framed profile top right). A single mass (20,429.24 Da) that did not match with any of the theoretical masses listed in the database built in the laboratory, was found in this peak, with an intensity *ca.* the half of the mass signal observed in peak 27 (Fig. 8). Thus, from this data we were unable to associate this mass to a known protein. Interestingly, in a seventh and single donkey milk sample, peak 26 was the only one, whereas peak 27 is apparently absent, suggesting that the molecule under peak 26 might correspond to a rare variant of β-LG I. The remaining 23 donkeys only have peak 27 (fraction XI). In order to identify this protein of 20,428.5 Da, present in peak 26, that did not correspond to any donkey milk protein identified so far, we implemented several additional and complementary approaches. First, since the protein exhibited a chromatographic behaviour similar to that of

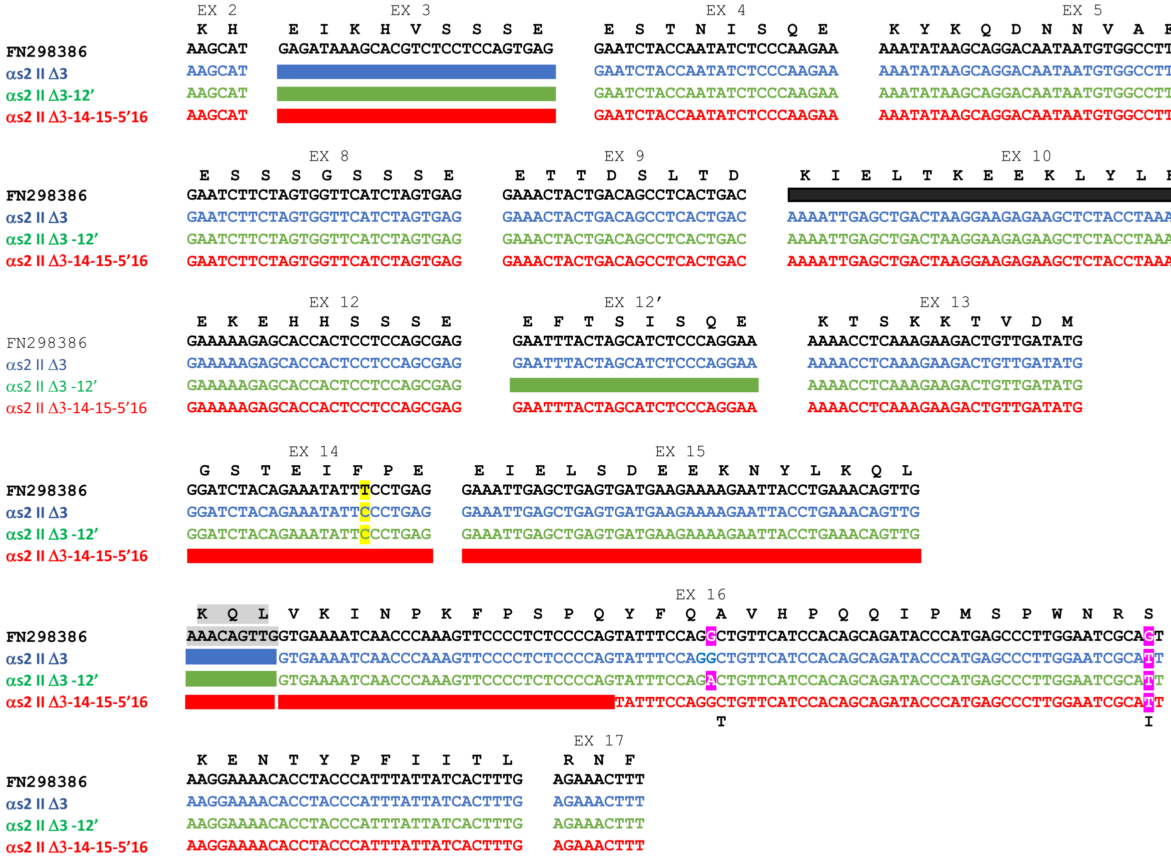

EX 2
K H
EX 3
E I K H V S S S E
EX 4
E S T N I S Q E
EX 5
K Y K Q D N N V A F Q T S Q
FN298386
AAGCAT GAGATAAAGCACGTCTCCTCCAGTGAG GAATCTACCAATATCTCCCAAGAA AAATATAAGCAGGACAATAATGTGGCCTTCCAAACCAGCCAG
αs2 II Δ3
AAGCAT GAATCTACCAATATCTCCCAAGAA AAATATAAGCAGGACAATAATGTGGCCTTCCAAACCAGCCAG
αs2 II Δ3-12′
AAGCAT GAATCTACCAATATCTCCCAAGAA AAATATAAGCAGGACAATAATGTGGCCTTCCAAACCAGCCAG
αs2 II Δ3-14-15-5′16
AAGCAT GAATCTACCAATATCTCCCAAGAA AAATATAAGCAGGACAATAATGTGGCCTTCCAAACCAGCCAG
EX 8
E S S S G S S S E
EX 9
E T T D S L T D
EX 10
K I E L T K E E K L Y L K Q L
FN298386
GAATCTTCTAGTGGTTCATCTAGTGAG GAAACTACTGACAGCCTCACTGAC
αs2 II Δ3
GAATCTTCTAGTGGTTCATCTAGTGAG GAAACTACTGACAGCCTCACTGAC AAAATTGAGCTGACTAAGGAAGAGAAGCTCTACCTAAAACAGCTG
αs2 II Δ3 -12′
GAATCTTCTAGTGGTTCATCTAGTGAG GAAACTACTGACAGCCTCACTGAC AAAATTGAGCTGACTAAGGAAGAGAAGCTCTACCTAAAACAGCTG
αs2 II Δ3-14-15-5′16
GAATCTTCTAGTGGTTCATCTAGTGAG GAAACTACTGACAGCCTCACTGAC AAAATTGAGCTGACTAAGGAAGAGAAGCTCTACCTAAAACAGCTG
EX 12
E K E H H S S S E
EX 12′
E F T S I S Q E
EX 13
K T S K K T V D M
FN298386
GAAAAAGAGCACCACTCCTCCAGCGAG GAATTTACTAGCATCTCCCAGGAA AAAACCTCAAAGAAGACTGTTGATATG
αs2 II Δ3
GAAAAAGAGCACCACTCCTCCAGCGAG GAATTTACTAGCATCTCCCAGGAA AAAACCTCAAAGAAGACTGTTGATATG
αs2 II Δ3 -12′
GAAAAAGAGCACCACTCCTCCAGCGAG AAAACCTCAAAGAAGACTGTTGATATG
αs2 II Δ3-14-15-5′16
GAAAAAGAGCACCACTCCTCCAGCGAG GAATTTACTAGCATCTCCCAGGAA AAAACCTCAAAGAAGACTGTTGATATG
EX 14
G S T E I F P E
EX 15
E I E L S D E E K N Y L K Q L
FN298386
GGATCTACAGAAATATTTCCTGAG GAAATTGAGCTGAGTGATGAAGAAAAGAATTACCTGAAACAGTTG
αs2 II Δ3
GGATCTACAGAAATATTCCCTGAG GAAATTGAGCTGAGTGATGAAGAAAAGAATTACCTGAAACAGTTG
αs2 II Δ3 -12′
GGATCTACAGAAATATTCCCTGAG GAAATTGAGCTGAGTGATGAAGAAAAGAATTACCTGAAACAGTTG
αs2 II Δ3-14-15-5′16
EX 16
K Q L V K I N P K F P S P Q Y F Q A V H P Q Q I P M S P W N R S
FN298386
AAACAGTTGGTGAAAATCAACCCAAAGTTCCCCTCTCCCCAGTATTTCCAGGCTGTTCATCCACAGCAGATACCCATGAGCCCTTGGAATCGCAGT
αs2 II Δ3
GTGAAAATCAACCCAAAGTTCCCCTCTCCCCAGTATTTCCAGGCTGTTCATCCACAGCAGATACCCATGAGCCCTTGGAATCGCATT
αs2 II Δ3 -12′
GTGAAAATCAACCCAAAGTTCCCCTCTCCCCAGTATTTCCAGACTGTTCATCCACAGCAGATACCCATGAGCCCTTGGAATCGCATT
αs2 II Δ3-14-15-5′16
TATTTCCAGGCTGTTCATCCACAGCAGATACCCATGAGCCCTTGGAATCGCATT
T
I
EX 17
K E N T Y P F I I T L
R N F
FN298386
AAGGAAAACACCTACCCATTTATTATCACTTTG AGAAACTTT
αs2 II Δ3
AAGGAAAACACCTACCCATTTATTATCACTTTG AGAAACTTT
αs2 II Δ3 -12′
AAGGAAAACACCTACCCATTTATTATCACTTTG AGAAACTTT
αs2 II Δ3-14-15-5′16
AAGGAAAACACCTACCCATTTATTATCACTTTG AGAAACTTT

Fig. 4. Nucleotide sequences of cDNA clones coding for α_{s2}-CN II compared to the cDNA sequence (FN298386) previously reported (Cosenza et al., 2010). Sequences are split into blocks of nucleotides to visualize the exonic modular structure of the mRNA as deduced from known splice junctions of the donkey *CSN1S2* gene. Exon numbering (above the blocks) is that of the bovine *CSN1S2* gene. Ex12′ is an additional exon as compared with the bovine gene; therefore it is numbered with'. Large black, blue, green and red boxes, depict alternatively skipped exons or sequences absent from the cDNA clones sequenced. For example, exon 3 is missing in the all clones sequenced in this study. The protein sequence given above each block is the sequence deduced from the previously reported cDNA sequence (black letters) by Cosenza et al. (2010). Silent mutations are highlighted in yellow wheras missense mutations are highlighted in pink and the amino acid change is given below the block.

β-LG I suggesting a possible common origin but with a significantly different mass (+1.914 Daltons), we assumed that it could be a variant of β-LG I. An anomaly in splicing (additional exon or intron cryptic splice sites) seemed very unlikely since the phase of exons that make up the β-LG-I gene is complex, with introns splitting the reading frame between codons (Phase 0) and within codons (Phases 1 and 2). Indeed, this presupposes an event that retains an open reading frame and the canonical sequence of the protein downstream, given the LC-MS/MS results (see thereafter). This makes it unlikely a putative additional exon or intronic sequence. On the other hand, the nucleotide sequence upstream from the initiation codon, in the first exon (NCBI Gene ID: 106829109), encodes a peptide sequence showing signal peptide characteristics. Thus, depending on whether the signal peptidase takes into account the first or the second signal peptide, the protein entering the secretion pathway in the endoplasmic reticulum, may or may not include an N-terminal sequence extension, corresponding to the canonical signal peptide, with an additional Alanine residue (Fig. 9).

To validate such a hypothesis, consistent with the mass observed for the protein eluted in peak 26, it was necessary to determine the N-terminal sequence of this protein or provide evidences of such an extension, to identify at the genome level the causal mutation(s) characterizing this highly probable new allele. Its frequency seems rather low (10%), with only one donkey homozygous at this locus. Therefore, three individual samples of donkey milk, displaying 3 different RP-HPLC profiles: with or without peak 26 (Fig. 1), were subjected to 1D SDS-PAGE fractionation for which we get a relatively good resolution. The three donkey milks gave 3 different SDS-PAGE patterns (Fig. 10**A**). Selected bands excised from the gel were submitted to LC-MS/MS analysis. Donkey milk A displayed, beside band 1 identified from LC-MS/MS data as being a mix of caseins, five main bands of which bands 5 and 6 were identified as lysozyme C and α-LA, respectively. The three remaining bands (2 to 4) were identified as β-LGs: bands 4 was ascribed from LC-MS/MS data to β-LG II whereas tryptic peptides from bands 2 and 3 covered quite the same part of the β-LG I sequence (Fig. 10**B**). Donkey milk B showed a single band (band 3) ascribed from LC-MS/MS data to β-LG I too. The third donkey milk sample (C) displayed only band 2 corresponding to a protein whose M_r is higher than that of β-LG I. Peptides identified by LC-MS/MS from band 2, both in samples A and C, were identical to those detected from band 3, leading to the identification of β-LG I. This strongly sustains our first hypothesis according to which the protein with the molecular mass of 20,428.50 Da found in sample C and A very likely derived from β-LG I. However, we were not able to find any tryptic peptide covering the canonical signal peptide. In parallel, we determined the nucleotide sequence of cDNA fragments obtained from reverse transcribed mRNAs and amplified using primers targeting β-LG I messengers. Sequences obtained from RNA extracted from milk fat globules appeared identical with the 3 individuals A, B and C all along the sequence encoding the mature β-LG I, except for the nonsynonymous G/A transition occurring at nucleotide 40 in exon 2 of the *BLG I* gene which results in the amino acid substitution D28N in the β-LG I polypeptide.

It is important to emphasize that no mass that could correspond to

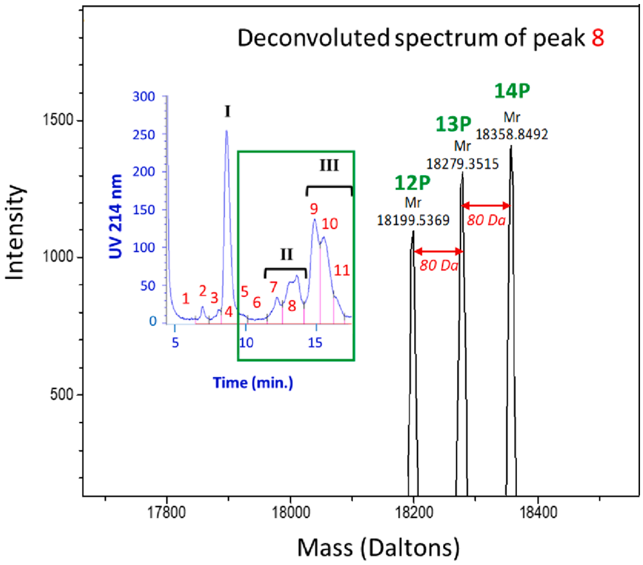


Fig. 5. ESI-TOF/MS deconvoluted spectrum of molecules present in peak 8 from the RP-HPLC profile (see Fig. 1). The initial part of the RP-HPLC chromatogram is given on the left to visualize the position of peak 8. The 3 main intact as2-CN II proteoforms detected in this peak displayed masses incrementing by 80 daltons, i.e. quite exactly the mass of a phosphate group (79.98 daltons). Identified as isoforms, in which the peptide EIKHVSSSE, encoded by exon 3, is lacking, their masses correspond to the same peptide chain (17239.06 Da) carrying 12, 13 and 14 phosphate residues respectively.

donkey κ-CN has been detected, possibly because of the chromatographic conditions used, the glycosylated isoforms are possibly eluted in the non-retained fractions. However, it is well known that in any case this CN is present at a very small amount in donkey milk and it is claimed that demonstrating the presence of κ-CN in donkey milk is very difficult and requires the use of specific techniques as 2D-electrophoresis and immunostaining (Chianese et al., 2010). It is worth noting that these authors reported also that no LC-eluted component was detected to confirm the presence of donkey κ-CN. However, from LC-MS/MS analysis we identified the donkey κ-CN with a *ca.* 30% peptide coverage of the mature protein. Interestingly, conversely to the donkey sequence recorded in Expasy (UniProtKB: F0V6V5), in which there is a Valine residue at position 28 of the mature protein we observed an Isoleucine residue, as in the predicted protein (NCBI, XP_014702750.1) deduced from the *Equus asinus* genome sequence and as reported in the *Equus caballus* protein sequence (UniProtKB: P82187).

4. Discussion

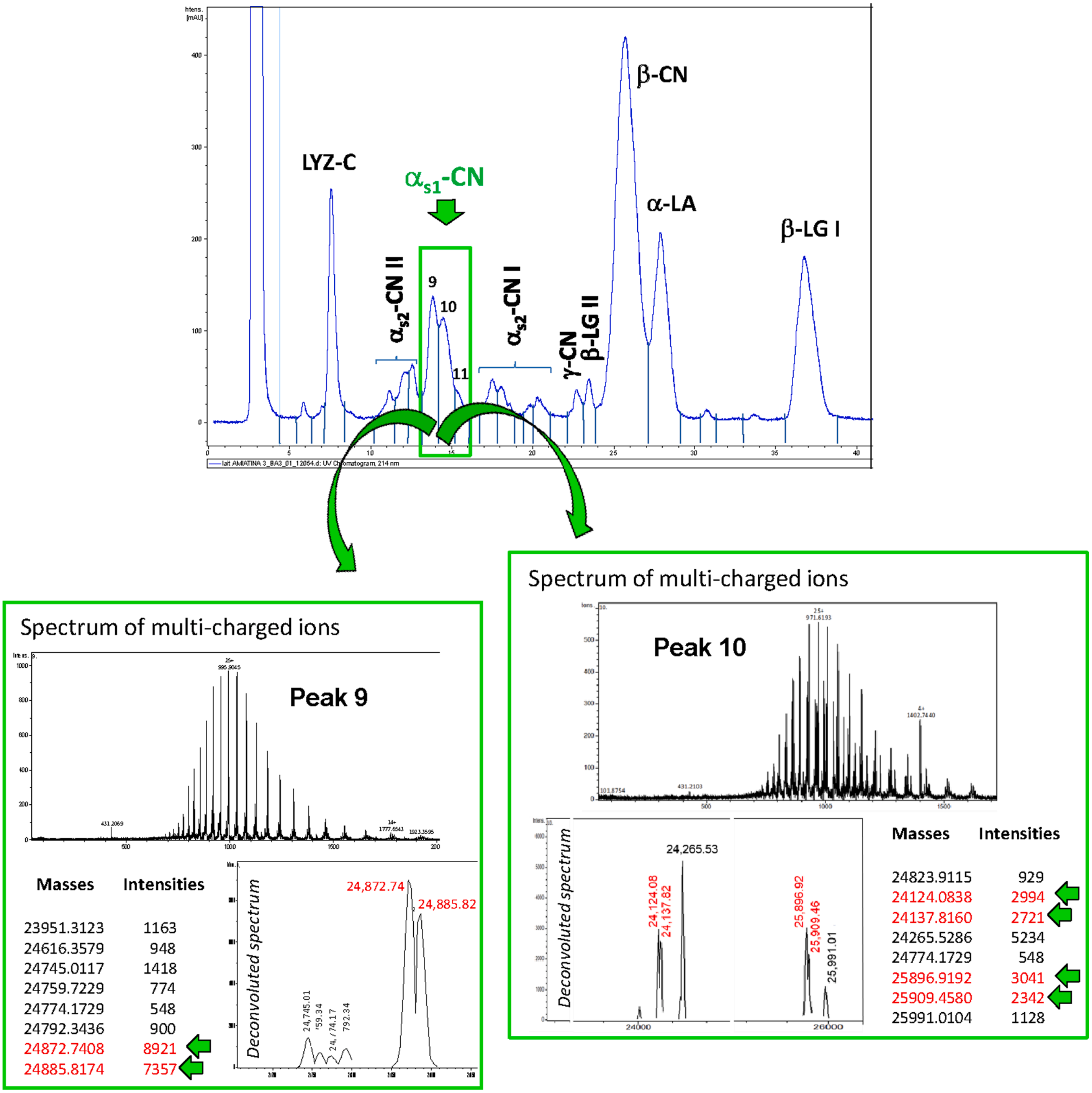
Whereas donkey milk, and particularly its protein fraction, remained poorly investigated as compared with ruminant milks at the beginning of the 21th century, a real leap has been made forward during the 10 last years, mainly by the Cunsolo group (University of Catania, Italy). The extensive analysis implemented and reported here, using LC-ESI-MS allowed a more thorough characterization of the less abundant casein components than previously reported. This approach shows that “Top-down” mass spectrometry targeting “native/intact” proteins, is sufficiently powerful to allow protein identification of most of the protein isoforms (proteoforms) from accurate experimentally determined masses and provides, in addition, information on post-translational modifications (PTM) such as phosphorylation, as well as on splicing variants and genetic polymorphisms.

4.1. Discrete phosphorylation of donkey calcium-sensitive caseins

The high degree of heterogeneity recorded with calcium-sensitive CNs in donkey milk is in part due to discrete phosphorylation. It is somewhat surprising to find such a variability given the crucial role of phosphate groups in the formation and stability of CN micelles. This is particularly significant with αs2-CN (I) and αs2-CN-like (II) of which the phosphorylation levels range from 7 to 13 and 8 to 15P/mole, respectively. However, whereas 18 and 21 potentially phosphorylated hydroxy-amino acid residues matching the S/T-X-A motif (with X as any aa and A as an acidic aa; Mercier, 1981) are present in both donkey αs2-CN and αs2-CN-like, respectively, the highest phosphorylation level recorded did not exceed 13P for αs2-CN and 15P in αs2-CN-like in Amiata donkey milk samples. This suggests that 5 and 6 potential phosphorylation sites are not accessible to the mammary kinase(s) or hardly phosphorylable in αs2-CN and αs2-CN-like, respectively. It is worth noting that amongst these potential phosphorylation sites, one is a T-X-E motif and two are T-X-D, in both αs2-CN, with decreasing potentiality relatively to the S-X-E motif. Moreover, regarding αs2-CN, two S-X-E sites are located within a short peptide sequence including the two cysteine residues present in αs2-CN and involved in the formation of intramolecular disulphide bridges that occurs in the endoplasmic reticulum and may prevent their phosphorylation that is occurring in the Golgi apparatus. We have to mention that the splicing isoforms derived from αs2-CN and αs2-CN-like follow in a consistent manner, the same “deficient” pattern of phosphorylation.

The full-length β-CN and its splicing variant (Δ exon 5) should be theoretically both equally phosphorylated with a maximum of 8P/mole with 6 Seryl and 2 Threonyl residues potentially phosphorylable. However, in agreement with Chianese et al. (2010), we only observed isoforms with up to 7P/mole (full-length), which appears, as reported for horse β-CN (Matéos et al., 2010), the higher level of phosphorylation found for this casein across species. This means that of the two Threonyl residues, it is very likely T12, which determines the phosphorylation of S10 that is phosphorylated, and not T207. Such an assumption is fully consistent with the results reported by Matéos and coll. who have shown that the *in vivo* phosphorylation of the equine β-CN is sequential and not randomly performed (Matéos et al., 2010). Mare’s β-CN isoform with 4P was found phosphorylated on residues S9, S23, S24 and S25, whereas the addition of phosphate groups on S18, T12, and S10 leading to the formation of the isoforms 5P-7P, respectively, occurs subsequently. Given isoforms with 6 and 7 phosphate residues are the most abundant in our study, this goes against the generally accepted idea that the sequence motif T-X-E is a poor substrate for the mammary gland kinase (s) and that the Threonyl residue in this sequence is usually not phosphorylated. In addition, it is highly likely that S15 and S17 are not phosphorylated, conversely to what was claimed by Cunsolo et al. (2009b).

As reported for the horse (Matéos et al., 2009), the maximum number of phosphate groups of donkey αs1-CN should be consistent with the existence of 8 potential phosphorylation sites involving Seryl residues (S18, S63, S66, S80, S82, S84, S85 and S86; numbering according to the full-length sequence, including the pentapeptide HTPRE encoded by exon 5) located in S-X-E/SP motifs. S81 becomes a possible phosphorylation site only if T83, located in a T-SP-SP motif, is itself phosphorylated, which brings then the maximum number of phosphate groups to 10P, if we exclude any possibility of phosphorylation of residues T67 and T109 that are notwithstanding located in a T-X-E motif. The highest level of phosphorylation observed in Amiata donkey was here 9P, which implies a different phosphorylation pattern, especially since the splice variants, in which exon 7 is skipped, reached a maximum level of 8P, consistent with S81 and T83 being phosphorylated. Alternatively, if T83 is not phosphorylated, then T67 and T109 should be. In a global way, it is important to note that the phosphorylation of CNs sensitive to calcium in donkey’s milk is clearly higher to what we know about human CNs, in part due to the fact that women’s milk does not seem to contain αs2-CN.



Masses	Intensities
23951.3123	1163
24616.3579	948
24745.0117	1418
24759.7229	774
24774.1729	548
24792.3436	900
24872.7408	8921
24885.8174	7357

Masses	Intensities
24823.9115	929
24124.0838	2994
24137.8160	2721
24265.5286	5234
24774.1729	548
25896.9192	3041
25909.4580	2342
25991.0104	1128

Fig. 6. Genetic polymorphism occurring at the *CSN1S1* locus, revealed by LC-MS analysis. The chromatogram profile (upper panel) was partitioned into 28 peaks whose contents were identified from the data generated by the mass spectrometer (MS). Peaks 9, 10 and 11, corresponding to the α_{s1}-CN (fraction III in Fig. 1, here framed in green), were subjected to a in-depth MS analysis to identify from deconvoluted mass signals, genetic variants and splicing isoforms with different phosphorylation levels. Masses and intensities corresponding to α_{s1}-CN allelic variants are in red and indicated with green arrows. Two masses detected in peak 9 (24,885.8174 and 24,872.8921 Da), with the highest intensity (8921 and 7357) and showing a differential of about 13 Da, correspond to genetic variants insofar as this same differential is reproduced in peak 10 for two series of masses (24,124 and 24,137 Da, on the one hand and 25,896 and 25,909 Da, on the other hand). The mass 25,909.458 Da was identified as the full-length α_{s1}-CNA with 8P, whereas the mass 24,137.816 Da corresponds to a splicing variant of α_{s1}-CNA with 6P in which exons 5 and 7 are lacking as well as a glutamine residue.

4.2. *New genetic variants of milk proteins identified in the Amiata donkey population*

Whereas αs2-CN I seemed to be monomorphic in the Amiata population studied, with the exception of 3 milk samples in which no trace of αs2-CN I could be detected, 6 SNPs of which 4 were responsible for amino acid changes, were reported in Ragusana donkeys (Cosenza et al., 2010; Cosenza unpublished results).

On the other hand, based on our results, there should theoretically be a minimum of 4 αs2-CN II variants. Indeed, given the sequence registered in the Expasy database (UniProtKB, C1L3G3) and the cDNA sequences determined in the present study, there must be at least four different variants arising from 4 alleles at the *CSN1S2 II* locus: variant A (A129- S143), B (A129-I143), C (T129-I143), and D (T129-S143). However, only two variants (B: A129-I143 and C: T129-I143) were detected in the panel of 30 donkeys of the Amiata breed analysed here, with a great prevalence for the B allele (*ca.* 82%) since 21 donkeys were homozygous B/B at this locus whereas only two were homozygous C/C. It is worth noting that the sequence reported in UniProtKB (C1L3G3), which corresponds to the A variant (A129-S143), differs from the

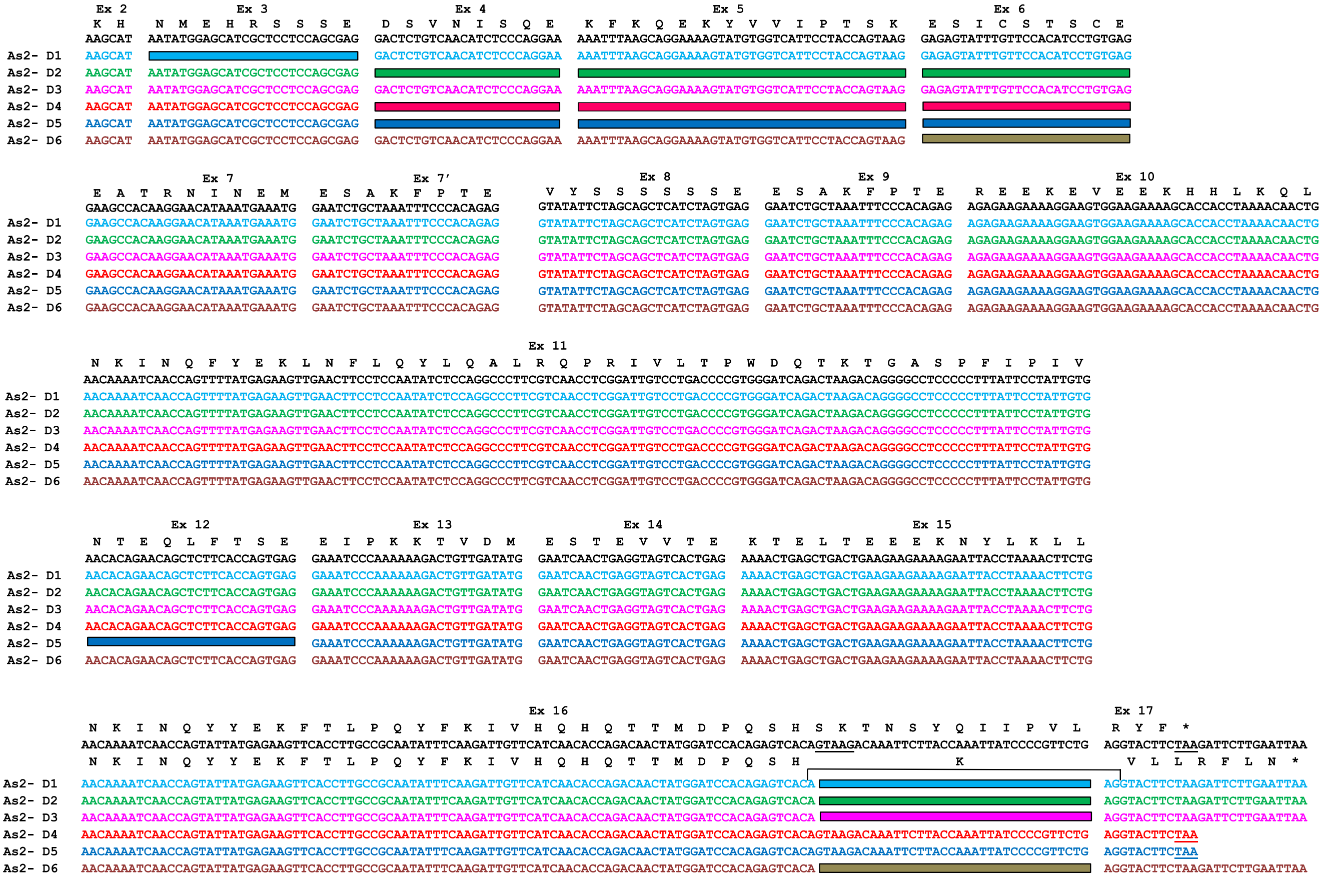


Fig. 7. Nucleotide sequences of cDNA clones coding for α_{s2}-CN I compared to the cDNA sequence (FM946022.1) previously reported(Cosenza et al., 2010). Sequences are split into blocks of nucleotides to visualize the exonic modular structure of the mRNA as deduced from known splice junctions of the donkey *CSN1S2* gene. Exon numbering (above the blocks) is that of the bovine *CSN1S2* gene. Ex 7′ that is an additional exon as compared with the bovine gene is numbered with'. Large green, pink, cyan and grey boxes, depict alternatively skipped exons or sequences absent from the cDNA clones sequenced. For example, sequences of exons 4, 5, 6 and 12 are absent from the clone As2-D5 sequenced in this study. The protein sequence given above each block is the sequence deduced from the previously reported cDNA sequence (black letters) by Cosenza et al. (2010). The removal of the 3′ part of exon 16, as a result of the use of the GTAAG site (highlighted in green) defining the beginning of an intronic sequence, leads to a frameshift inducing a modification in the C-terminal sequence of the protein encoded by the clones As2-D1, As2-D2, As2-D3 and As2-D6. This new C-terminal sequence is given below the reference nucleotide sequence (black letters). A silent transition (G/A) is highlighted in yellow. **Caution**! The sequences of the *CSN1S2* or *B (*or *I)* and *CSN1S2-like* or *A* (or *II*) genes are reversed in the NCBI database. The LOC106835119, identified as the *CSN1S2-like* or *A* gene is located close to the PRR27 locus, whereas in the other species the *CSN1S2-like* or *A* gene is usually closest to the *CSN2* locus and *CSN1S2* or *B* (LOC106828076 in the donkey genome) should be located close to the *ODAM* locus. Furthermore, from genome data these two genes appear to be convergently transcribed in the donkey, whereas they are transcribed in the same direction in the other species.

sequence we report here by the absence of exon 10 and the presence of a duplication of a KQL tripeptide. Interestingly, in a recent characterization of *CSN1S2 I* and *II* transcripts in the Ragusana donkey breed (Cosenza, unpublished results), the duplication AAA CAG TTG coding for the KQL tripeptide at the beginning of exon 16 previously reported (Cosenza et al., 2010), was not detected in all the *CSN1S2 II* mRNAs sequenced. Therefore, it seems that this duplication should be considered as an artifact. The fourth theoretical variant (D: T129-S143) was not found in this population. However, it is likely that the extent of the genetic polymorphism at this locus would be even more significant in the donkey since in the Ragusana breed, 4 possible further SNPs responsible for amino acid residue changes: L62F, D90N, H131Y, F157S, were found, beside silent SNPs (Cosenza, unpublished results). Genetic polymorphisms were also detected with αs1- and β-CNs. Two new αs1-CN genetic variants, named D and E (M_r 24,393 and 24,433 Da, for the main non-phosphorylated isoforms in which exon 7 is lacking, respectively), arising from the *CSN1S1* locus and displaying mass differences of *ca.* -13 Da and + 27 Da relatively to the A variant (M_r 24,406.41 Da), were identified (Fig. 6). The A variant remained the most frequent (68.3%), whereas D and E account for 10 and 21.7%, respectively. The mutations responsible for these polymorphisms are currently under investigation (in progress).

Regarding *CSN2*, beside the two variants (A: S37-V84 and B: V37-P84) known so far, a third variant named E with a mass (25,553.7 Da), *i.e.* 14.4 Da more than the mass of the B variant (25,539.36 Da for the non-phosphorylated full-length peptide chain), has been observed. This new variant is not very frequent (5 individuals carrying this allele, of which only 1 was homozygous), compared to the B variant which is highly represented, with an allelic frequency of 90% in the population analysed. Of note, none of the 30 individuals appeared to carry the A variant (S37- V84). The E variant has not been characterized so far but is objectively different from the theoretical putative C and D variants (V37-V84: 25,541.4 Da and S37-P84: 25,527.3 Da for the non-phosphorylated isoforms, respectively).

Amongst the whey proteins, α-lactalbumin (α-LA) and the milk isozyme of lysozyme C appeared both monomorphic in the set of Amiata milk samples analysed here. The genetic variant of α-LA observed corresponded to the only α-LA genetic variant reported so far, despite an apparent heterogeneity of the protein was observed in donkey milk (Giuffrida et al., 1992). Elsewhere, we report here the existence of a new genetic variant of lysozyme C, characterized from the cDNA sequence. It differs from the previously characterized A and B variants and displays

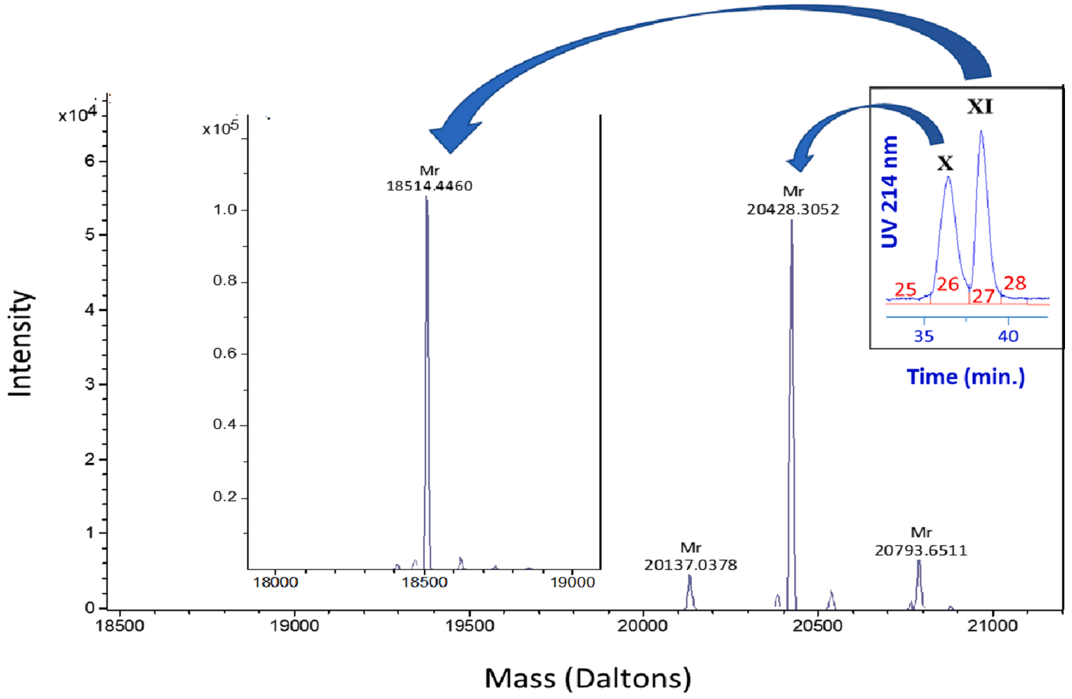
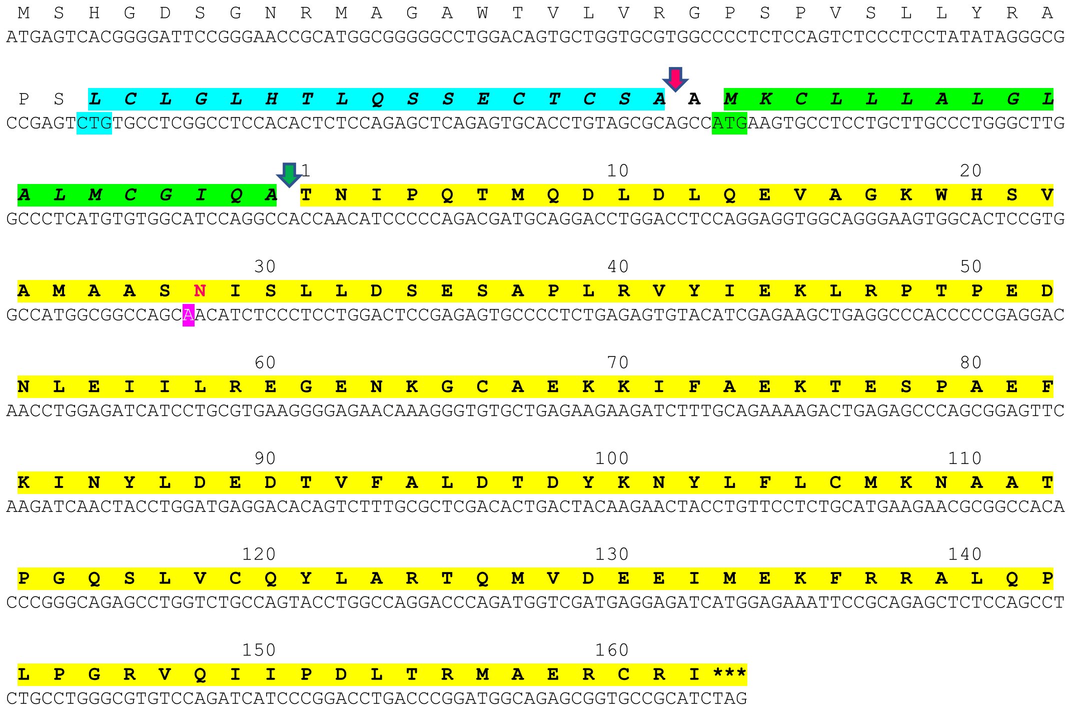
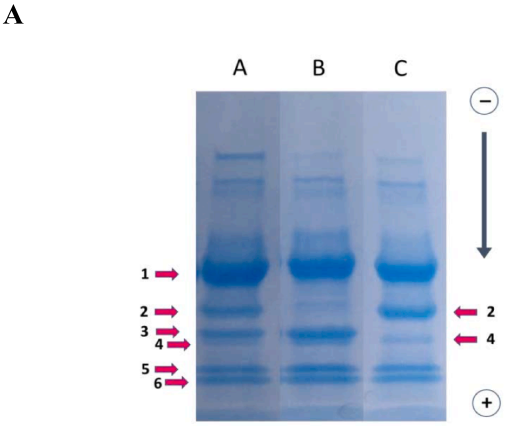


Fig. 8. ESI-TOF/MS deconvoluted spectra of molecules present in peak 26 and 27 from the RP-HPLC chromatogram (see Fig. 1 framed profile). The terminal part of the RP-HPLC chromatogram is given on the right to visualize the position of peaks 26 and 27, eluted from the column between 35 and 40 min.



```
 M  S  H  G  D  S  G  N  R  M  A  G  A  W  T  V  L  V  R  G  P  S  P  V  S  L  L  Y  R  A
ATGAGTCACGGGGATTCCGGGAACCGCATGGCGGGGGCCTGGACAGTGCTGGTGCGTGGCCCCTCTCCAGTCTCCCTCCTATATAGGGCG

 P  S  L  C  L  G  L  H  T  L  Q  S  S  E  C  T  C  S  A  A  M  K  C  L  L  L  A  L  G  L
CCGAGTCTGTGCCTCGGCCTCCACACTCTCCAGAGCTCAGAGTGCACCTGTAGCGCAGCCATGAAGTGCCTCCTGCTTGCCCTGGGCTTG

                         1                          10                         20
 A  L  M  C  G  I  Q  A  T  N  I  P  Q  T  M  Q  D  L  D  L  Q  E  V  A  G  K  W  H  S  V
GCCCTCATGTGTGGCATCCAGGCCACCAACATCCCCCAGACGATGCAGGACCTGGACCTCCAGGAGGTGGCAGGGAAGTGGCACTCCGTG

                    30                         40                         50
 A  M  A  A  S  N  I  S  L  L  D  S  E  S  A  P  L  R  V  Y  I  E  K  L  R  P  T  P  E  D
GCCATGGCGGCCAGCAACATCTCCCTCCTGGACTCCGAGAGTGCCCCTCTGAGAGTGTACATCGAGAAGCTGAGGCCCACCCCCGAGGAC

                    60                         70                         80
 N  L  E  I  I  L  R  E  G  E  N  K  G  C  A  E  K  K  I  F  A  E  K  T  E  S  P  A  E  F
AACCTGGAGATCATCCTGCGTGAAGGGGAGAACAAAGGGTGTGCTGAGAAGAAGATCTTTGCAGAAAAGACTGAGAGCCCAGCGGAGTTC

                    90                         100                        110
 K  I  N  Y  L  D  E  D  T  V  F  A  L  D  T  D  Y  K  N  Y  L  F  L  C  M  K  N  A  A  T
AAGATCAACTACCTGGATGAGGACACAGTCTTTGCGCTCGACACTGACTACAAGAACTACCTGTTCCTCTGCATGAAGAACGCGGCCACA

                   120                        130                        140
 P  G  Q  S  L  V  C  Q  Y  L  A  R  T  Q  M  V  D  E  E  I  M  E  K  F  R  R  A  L  Q  P
CCCGGGCAGAGCCTGGTCTGCCAGTACCTGGCCAGGACCCAGATGGTCGATGAGGAGATCATGGAGAAATTCCGCAGAGCTCTCCAGCCT

                   150                        160
 L  P  G  R  V  Q  I  I  P  D  L  T  R  M  A  E  R  C  R  I ***
CTGCCTGGGCGTGTCCAGATCATCCCGGACCTGACCCGGATGGCAGAGCGGTGCCGCATCTAG
```

Fig. 9. Open reading frame of the cDNA encoding the beta-LG I. The translation in amino acids is given above the nucleotide sequence and the numbering begins with the first amino acid of the mature protein. The canonical initiation codon (ATG) is highlighted in green as well as the signal peptide, cleaved by the signal peptidase at the green vertical arrow. The mature protein sequence (162 amino acid residues long with a M_r of 18,514.25 Da) is highlighted in yellow and the stop codon is symbolized by ***. An alternative initiation codon (CTG), located 54 nucleotides upstream, is highlighted in cyan defining accordingly a putative secondary signal peptide of 18 amino acid residues ahead of the canonical signal peptide. The putative cleavage site is indicated by the pink vertical arrow. The usage of such an initiation codon would lead to the release into the endoplasmic reticulum of a 20,428.75 Da protein whose N-terminal residue is the alanine just before the methionine encoded by the usual initiation codon and therefore including the canonical signal peptide. The ASN residue (position 28, in pink) is the only one amino acid substitution, due to a G/A transition (highlighted in pink) which was observed in the coding sequence of the 20,428.75 Da protein, comparatively to the 18,514.25 Da B variant.



B

Band 2 sample A

TNIPQTMQDLDLQEVAGKWHSVAMAASDISLLDSEEAPLRVYIEK**LRPTPEDNLEIILR**E
GENKGCAEK**KIFAEKTESPAEFK**INYLDEDTVFALDSDYK**NYLFLCMKNAATPGQSLVCQ**
YLARTQMVDEEIMEKFRRALQPLPGRVQIVPDLTRMAERCRI

Band 3 sample A

TNIPQTMQDLDLQEVAGKWHSVAMAASDISLLDSEEAPLRVYIEK**LRPTPEDNLEIILRE**
GENKGCAEK**KIFAEKTESPAEFK**INYLDEDTVFALDSDYK**NYLFLCMKNAATPGQSLVCQ**
YLARTQMVDEEIMEKFRRALQPLPGRVQIVPDLTRMAERCRI

Band 3 sample B

TNIPQTMQDLDLQEVAGKWHSVAMAASDISLLDSEEAPLRVYIEK**LRPTPEDNLEIILRE**
GENKGCAEK**KIFAEKTESPAEFK**INYLDEDTVFALDSDYK**NYLFLCMKNAATPGQSLVCQ**
YLARTQMVDEEIMEKFRRALQPLPGRVQIVPDLTRMAERCRI

Band 2 sample C

TNIPQTMQDLDLQEVAGKWHSVAMAASDISLLDSEEAPLRVYIEK**LRPTPEDNLEIILRE**
GENKGCAEKKIFAEKTESPAEFKINYLDEDTVFALDSDYK**NYLFLCMKNAATPGQSLVCQ**
YLARTQMVDEEIMEKFRRALQPLPGRVQIVPDLTRMAERCRI

Band 4 sample C

TDIPQTMQDLDLQEVAGRWHSVAMVASDISLLDSESAPLR**VYVEELRPTPEGNLEIILRE**
GANHVCVERNIVAQKTEDPAVFTVNYQGERKISVLDTDYAHYMFFCVGPPLPSAEHGTVC
QYLAR**TQKVDEEVMEKFSRALQPLPGHVQIIQDPSGGQER**CGF

Fig. 10. 1D-SDS-PAGE analysis of 3 individual donkey milks and identification of protein bands by LC-MS/MS. ***Panel A***: 1D-SDS-PAGE analysis of 3 individual donkey milk samples A, B and C. Band 1: caseins; band 2: unknown protein (20,428 Da); band 3: β-LG I (18,514 Da); 4: β-LG II (18,227 Da); 5: lysozyme C (14,638 Da), 6: α-lactalbumin (14,222 Da). ***Panel B***: Peptides identified by LC-MS/MS in bands 2, 3 and 4 isolated from 1D SDS-PAGE of donkey milk samples A, B and C. Sequences covering the donkey β-LG I sequence are highlighted in colors, whereas sequences covering the donkey β-LG II are highlighted in grey.

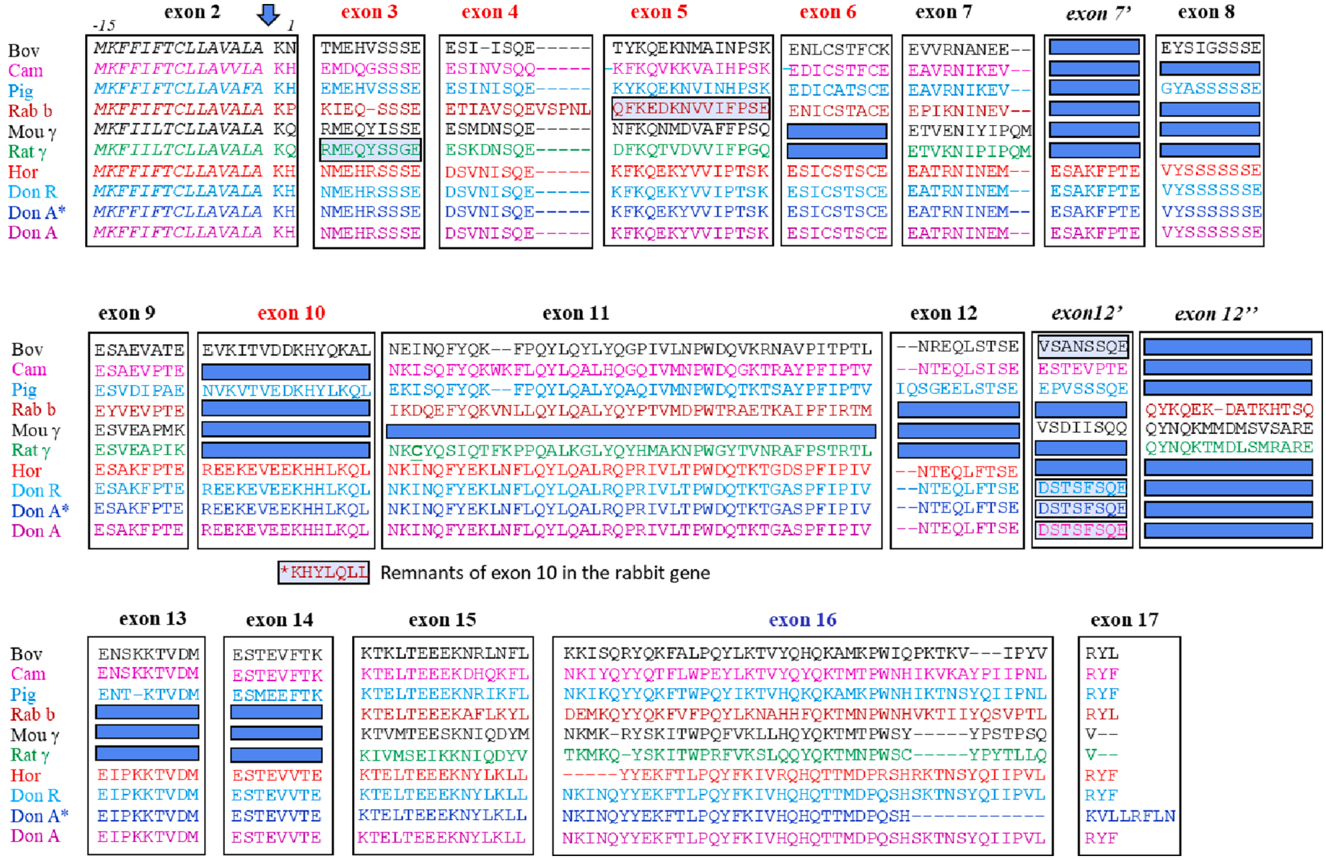
two amino acid substitutions (Y52S) and (S61N) in the peptide chain, as compared with variant A initially reported (Godovac-Zimmermann et al., 1988). On the other hand, regarding the other two main whey proteins, *i.e.* β-lactoglobulins I and II, genetic polymorphisms were recorded with β-LG II in the Amiata population. Three of the five variants identified and characterized, so far, named A (UniProtKB: P19647; Godovac-Zimmermann et al., 1990), B, C (Herrouin et al., 2000; Zuccaro et al., 2010), D (Cunsolo et al., 2007) and E Chianese et al., 2013) have been found (B, C and D) in the set of milk samples from the Amiata population analysed here. A null or weakly expressed allele, that could correspond to the F allele recently described (Criscione et al., 2018), was also observed. Variant C is the most frequent (44.4%) while variants B and D are equally represented with 27.8% if one disregards 3 possible F/F individuals. It is worth mentioning that the donkey β-LG II is the expression product of the improperly named *PAEP* gene that corresponds to the horse *LGB2* locus, showing a perfect similarity with its horse counterpart. Whereas to date only two variants (A and B) of β-LG I (UniProtKB: P13613) have been characterized at the protein level (Godovac-Zimmermann et al., 1988; Herrouin et al., 2000). The two sequences show 98.1% similarity as a result of three amino acid substitutions: E36S, S97T and V150I. Taken together these three amino acid substitutions at positions 36, 97, and 150 explained the mass difference of 14 Da observed for the entire protein (Herrouin et al., 2000). A third variant, possibly the ancestral β-LG I, differing from the B variant by a single amino acid substitution (A25V) was deduced from a genome sequence (PSZQ01002145.1). We propose to call it C whereas it seemed that the B variant (M_r = 18,514.25 Da), was the only one found in the Amiata population analysed. However, the additional peak observed (peak 26) with a mass of 20,428.5 Da (Fig. 8) which does not correspond to any donkey milk protein identified so far, is possibly a new genetic variant of β-lactoglobulin I. Given the first results we report here, this hypothesis seemed confirmed, even though we were not able to conclude definitely on the structure of this new β-LG I isoform. From LC-MS/MS data that allowed a quite total covering of the mature β-LG I peptide sequence, splicing anomalies do not seem to be involved in such a polymorphism. Further analyses aiming to analyse, at the genomic level, the regions upstream the forward primer used for cDNA amplification in the three donkeys are in progress.

4.3. *Splicing isoforms of CNs remain an amazing source of molecular diversity*

The process of intron removal, and exon joining (splicing), is a major function ensured, in the nucleus, by a large multicomponent complex, called spliceosome, assembled in a stepwise pathway. This accurate mechanism is governed by a set of rather strict rules to achieve high fidelity and efficiency in splicing. However, caseins spliced variants are widely spread across species (Leroux et al., 1992; Suteu et al., 2011; Ryskaliyeva et al., 2019a; Milenkovic et al., 2002; Matéos et al., 2009; Martin & Leroux, 1992; Miclo et al., 2007; Lenasi et al., 2006). A dysfunction of this machinery may have dramatic biological consequences by modifying the message and accordingly the primary structure of the protein. Our study revealed the existence of at least 20 splicing variants arising from 4 of the 5 *CSN* genes, involving different kind of event such as exon skipping, 5′ and 3′ cryptic splice site usage, with a singular shift of the open reading frame, generating two αs2-CN I isoforms with different C-terminal sequences. This amazing situation, which has reached proportions never seen before in any species studied so far, is partly due to the expression of a second gene encoding an αs2-CN-like, also expressed in rodents and lagomorphs.

Paradoxically, *CSN1S2 II* is more highly expressed than *CSN1S2 I* in the rabbit species (Hue-Beauvais et al., 2017). In order to compare CNs across species, the exon numbering adopted was that of the bovine genes, the first ones sequenced and which constitute the references. To compare αs2-CN derived from *CSN1S2 I* and *II* loci, we have adopted the same exon numbering for both genes compatible with the hypothesis that these 2 genes are, as in humans, the result of a duplication (Rinjkels, 2002). It should be noted that these 2 genes are by far the most impacted by splicing anomalies. Indeed, no less than 5 exons are concerned by exon skipping events in each of the 2 genes. For αs2-CN I, these are exons 3, 4, 5, 6 and 10 and for αs2-CN II, exons 3, 10, 12′, 14 and 15 (**Supplementary material**, **Fig. S1**). Interestingly, 2 skipped exons are common (exons 3 and 10). In addition, cryptic splice sites are used within exon 16, on the 3′ side for primary transcripts coding for αs2-CN I and on the 5′ side for primary transcripts encoding αs2-CN II. This obviously has consequences on the structure of the C-terminal part of this protein which is known to correspond to the region of highest density in potentially antimicrobial peptide sequences (Ryskaliyeva et al., 2019a).

Cunsolo group reported previously the existence, in the milk sample collected in Eastern Sicily from a single individual donkey belonging to the Ragusana breed, of 4 αs2-CN isoforms with experimentally determined M_r of 25,429, 21,939, 25,203 and 21,713 Da (Saletti et al., 2012). They demonstrated that the isoform with M_r 25,429 Da differed from the full-length αs2-CN (M_r 26,830 Da) for the deletion of the 176NKINQ180 pentapeptide, encoded by the first five codons of exon 16 (*Bos taurus*



	exon 2	exon 3	exon 4	exon 5	exon 6	exon 7	exon 7’	exon 8
	-15 1							
Bov	*MKFFIFTCLLAVALA* KN	TMEHVSSSE	ESI-ISQE-----	TYKQEKNMAINPSK	ENLCSTFCK	EVVRNANEE--	■	EYSIGSSSE
Cam	*MKFFIFTCLLAVVLA* KH	EMDQGSSSE	ESINVSQQ-----	-KFKQVKKVAIHPSK	-EDICSTFCE	EAVRNIKEV--	■	■
Pig	*MKFFIFTCLLAVAFA* KH	EMEHVSSSE	ESINISQE-----	KYKQEKNVINHPSK	EDICATSCE	EAVRNIKEV--	■	GYASSSSSE
Rab b	*MKFFIFTCLLAVALA* KP	KIEQ-SSSE	ETIAVSQEVSPNL	QFKEDKNVVIFPSE	ENICSTACE	EPIKNINEV--	■	■
Mou γ	*MKFIILTCLLAVALA* KQ	RMEQYISSE	ESMDNSQE-----	NFKQNMDVAFFPSQ	■	ETVENIYIPQM	■	■
Rat γ	*MKFIILTCLLAVALA* KQ	RMEQYSSGE	ESKDNSQE-----	DFKQTVDVVIFPGQ	■	ETVKNIPIPQM	■	■
Hor	*MKFFIFTCLLAVALA* KH	NMEHRSSSE	DSVNISQE-----	KFKQEKYVVIPTSK	ESICSTSCE	EATRNINEM--	ESAKFPTE	VYSSSSSSE
Don R	*MKFFIFTCLLAVALA* KH	NMEHRSSSE	DSVNISQE-----	KFKQEKYVVIPTSK	ESICSTSCE	EATRNINEM--	ESAKFPTE	VYSSSSSSE
Don A*	*MKFFIFTCLLAVALA* KH	NMEHRSSSE	DSVNISQE-----	KFKQEKYVVIPTSK	ESICSTSCE	EATRNINEM--	ESAKFPTE	VYSSSSSSE
Don A	*MKFFIFTCLLAVALA* KH	NMEHRSSSE	DSVNISQE-----	KFKQEKYVVIPTSK	ESICSTSCE	EATRNINEM--	ESAKFPTE	VYSSSSSSE

	exon 9	exon 10	exon 11	exon 12	exon12’	exon 12’’
Bov	ESAEVATE	EVKITVDDKHYQKAL	NEINQFYQK--FPQYLQYLYQGPIVLNPWDQVKRNAVPITPTL	--NREQLSTSE	VSANSSQE	■
Cam	ESAEVPTE	■	NKISQFYQKWKFLQYLQALHQGQIVMNPWDQGKTRAYPFIPTV	--NTEQLSISE	ESTEVPTE	■
Pig	ESVDIPAE	NVKVTVEDKHYLKQL	EKISQFYQK--FPQYLQALYQAQIVMNPWDQTKTSAYPFIPTV	IQSGEELSTSE	EPVSSSQE	■
Rab b	EYVEVPTE	■	IKDQEFYQKVNLLQYLQALYQYPTVMDPWTRAETKAIPFIRTM	■	■	QYKQEK-DATKHTSQ
Mou γ	ESVEAPMK	■	■	■	VSDIISQQ	QYNQKMMDMSVSARE
Rat γ	ESVEAPIK	■	NKCYQSIQTFKPPQALKGLYQYHMAKNPWGYTVNRAFPSTRTL	■	■	QYNQKTMDLSMRARE
Hor	ESAKFPTE	REEKEVEEKHHLKQL	NKINQFYEKLNFLQYLQALRQPRIVLTPWDQTKTGDSPFIPIV	--NTEQLFTSE	■	■
Don R	ESAKFPTE	REEKEVEEKHHLKQL	NKINQFYEKLNFLQYLQALRQPRIVLTPWDQTKTGASPFIPIV	--NTEQLFTSE	DSTSFSQE	■
Don A*	ESAKFPTE	REEKEVEEKHHLKQL	NKINQFYEKLNFLQYLQALRQPRIVLTPWDQTKTGASPFIPIV	--NTEQLFTSE	DSTSFSQE	■
Don A	ESAKFPTE	REEKEVEEKHHLKQL	NKINQFYEKLNFLQYLQALRQPRIVLTPWDQTKTGASPFIPIV	--NTEQLFTSE	DSTSFSQE	■

*KHYLQLL Remnants of exon 10 in the rabbit gene

	exon 13	exon 14	exon 15	exon 16	exon 17
Bov	ENSKKTVDM	ESTEVFTK	KTKLTEEEKNRLNFL	KKISQRYQKFALPQYLKTVYQHQKAMKPWIQPKTKV---IPYV	RYL
Cam	ENSKKTVDM	ESTEVFTK	KTELTEEEKDHQKFL	NKIYQYYQTFLWPEYLKTVYQYQKTMTPWNHIKVKAYPIIPNL	RYF
Pig	ENT-KTVDM	ESMEEFTK	KTELTEEEKNRIKFL	NKIKQYYQKFTWPQYIKTVHQKQKAMKPWNHIKTNSYQIIPNL	RYF
Rab b	■	■	KTELTEEEKAFLKYL	DEMKQYYQKFVFPQYLKNAHHFQKTMNPWNHVKTIIYQSVPTL	RYL
Mou γ	■	■	KTVMTEESKNIQDYM	NKMK-RYSKITWPQFVKLLHQYQKTMTPWSY-----YPSTPSQ	V--
Rat γ	■	■	KIVMSEIKKNIQDYV	TKMKQ-YSKITWPRFVKSLQQYQKTMNPWSC-----YPYTLLQ	V--
Hor	EIPKKTVDM	ESTEVVTE	KTELTEEEKNYLKLL	-----YYEKFTLPQYFKIVRQHQTTMDPRSHRKTNSYQIIPVL	RYF
Don R	EIPKKTVDM	ESTEVVTE	KTELTEEEKNYLKLL	NKINQYYEKFTLPQYFKIVHQHQTTMDPQSHSKTNSYQIIPVL	RYF
Don A*	EIPKKTVDM	ESTEVVTE	KTELTEEEKNYLKLL	NKINQYYEKFTLPQYFKIVHQHQTTMDPQSH------------	KVLLRFLN
Don A	EIPKKTVDM	ESTEVVTE	KTELTEEEKNYLKLL	NKINQYYEKFTLPQYFKIVHQHQTTMDPQSHSKTNSYQIIPVL	RYF

Fig. 11. Multiple alignment of the amino acid sequence of the αs2-casein I from 8 different species. Three donkey αs2-casein I sequences are given, corresponding to the sequences reported in the Ragusana breed (Don R, Saletti et al., 2012) and in the Amiatina breed (Don A* and Don A, present study). UniProtKB accession numbers are: Bov (P02663), Cam (O97944), Pig (P39036), Rab b (P50419), Rat γ (Q8CGR3), Mou γ (P02664), Hor (A0A0C5DH76), Don R (B7VGF9). Peptide sequences are split into blocks of amino acid residues to visualize the exonic modular structure of the protein. Numbering of the exons is that of the bovine gene and additional exons are numbered with ‘ and “ (in italics). Exons that may be subject to skipping events with donkey are indicated in red (above blocks). Exon 16 which is subject to the use of a cryptic splice site in the donkey (Don A*) is indicated in blue. Amino acid sequences framed and greyed for Rat γ (exon 3), Rab b (exon 5), Bov and the 3 donkey sequences (exon 12′) correspond to the translation of putative exons found in the genome of these species at the right place. Large blue boxes, within blocks, depict species-specific constitutively out spliced or absent exons. Amino acids in italics correspond to signal peptides, of which the cleavage site is indicated by the vertical blue arrow. Dashes (-) are inserted gaps introduced to maximize the alignment.

gene numbering). In the isoform of M_r 21,939 Da, the same pentapeptide is missing as well as the sequence D12-E42 encoded by exons 4, 5 and 6 (Fig. 11).

The third isoform reported by these authors (M_r 25,203 Da), was deduced from LC-MS/MS analysis as resulting from the deletion of the heptapeptide 212YQIIPVL218 in comparison with the full-length αs2-CN. This heptapeptide is encoded by the last seven 3′ codons of exon 16. Surprisingly, the use of a cryptic splice site (GTAAG) at the 3′ end of exon 16 that results in a modification of the C-terminal sequence of αs2-CN I, as we observed, gives a mass for the non-phosphorylated protein (25,203.33 Da) which is compatible with the hypothetical deletion of the heptapeptide 212YQIIPVL218, as proposed by Saletti et al. (2012). However, such a deletion is not consistent with the nucleotide sequences both of the cDNA (Fig. 7) and of the gene (NCBI, LOC106828076). On the other hand, there is no obvious explanation, from the nucleotide sequence, to account for the deletion of the peptide 176NKINQ180 that Saletti and co-workers suggested.

The only αs2-CN I genetic variant identified in the Amiata milk samples analysed in our study corresponds to the A variant (UniProtKB: B7VGF9) described by Saletti et al. (2012). Thanks to the cDNA sequencing results and given the genome sequence available we showed that this αs2-CN I isoform is due to an alternative splice site usage involving a donor GTAAG site located 35 nt upstream the canonical donor splice site (GTGAG) that is the 5′ end sequence of the penultimate intron. Moreover, we did not observe any αs2-CN I isoforms deleted from the pentapeptide 176NKINQ180 encoded by the first 15 nucleotides of the penultimate coding exon, in the Amiata breed. On the other hand, this deletion, which seems to be the norm in mare (Fig. 11), is very likely explained by a G->A transition abolishing the canonical acceptor splice site (**Supplementary material**, **Fig. S2**) in the horse genome (NCBI, Gene ID: 100327035). Such a mutation possibly occurred in the genome of the Ragusana donkey of which the milk has been analysed by Saletti and co-workers. Taking into account the predominant differential splicing events, the full-length donkey αs2-CN I and its three internally-deleted isoforms consist of 221, 214, 190 and 183 aa, respectively. This propensity to be the subject of exon-skipping, which characterizes the *CSN1S1* and *CSN1S2* genes and which is generally attributed to their fragmented structure and the small size of the coding exons that compose them, is not the only explanation here. Indeed, the “en bloc” skipping of exons 4, 5 and 6 in some αs2-CN I isoforms clearly depends on another factor, probably related to a specific secondary structure favouring this kind of event during the course of pre-mRNA processing, as previously reported for the goat αs1-CN (Leroux et al., 1992). From multiple alignments, exon 7’ appears to be equine specific in *CSN1S2 I* (Fig. 11), whereas exons 12 and 14 are only present in donkey and absent in all known *CSN1S2 II* sequences (**Supplementary material**, **Fig. S1**). In addition, exons 6, 7 and 7’ seem to be absent from all *CSN1S2 II.* Interestingly the genome segment encompassing exons 14 and 15 is an obvious duplication of the genome sequence which extends from exon 9 to exon 10 (**Suppementary material** Fig. S3).

One of the main biochemical features of αs2-CN, usually underlined as having an important functional role in the micelle structure, is its ability to form intramolecular disulphide bridges (Farrell et al., 2009), owing to the presence of two close cysteine residues, at positions 37 and 41 in the mature donkey αs2-CN I peptide chain. These residues are encoded by exon 6, the sequence of which is rather well conserved across species, when present, with both cysteine residues at the same position (Fig. 11). In contrast, in rat and mouse *CSN1S2 I* genes, which lack this exon, both these cysteine residues are obviously missing in the protein. Nevertheless, this deficit is covered with a single and two contiguous cysteine residues in the middle of the peptide sequence encoded by exon 11, in rat and mouse αs2-CN II, respectively. Interestingly, screening the genomic sequence of the donkey *CSN1S2 I* gene reveals the existence of a putative exon 12′ (**Supplementary material Fig. S4**). The same sequence is found in the *Equus caballus* gene, however the 5′ donor splice site G**T**AAA is mutated to G**C**AAA that can be sometimes functional, as shown for the gene encoding the Whey Acidic Protein in camel (Ryskaliyeva et al., 2019b). The same situation is observed in the bovine gene.

As in most species, donkey β-CN is less prone to splicing defects. Indeed, only one such event (skipping of exon 5) was recorded in our study, confirming previous results (Cunsolo et al., 2009b). This situation is probably due to the less split genomic organization of the gene encoding this protein, and even though exon skipping occurs less frequently, splicing-out of exon 3 was reported as constitutive in humans (Menon et al., 1992; Martin & Leroux, 1992) whereas splicing-out of exon 5 seems to be frequent in the horse (Milenkovic et al., 2002; Lenasi et al., 2003; Miranda et al., 2004; Miclo et al., 2007). Recently, skipping of exon 5 was also shown to occur at low noise (less than 4% of total β-CN) in cattle (Miranda et al., 2020). The use of cryptic splice sites also occurs and can have more dramatic consequences. This is well exemplified in mare's milk in which a low-molecular-weight β-CN variant, showing a large internal deletion (132 aa), has been characterized as arising from a cryptic splice site usage occurring within exon 7, during the course of primary transcripts processing (Miclo et al., 2007). Such a deviant splicing behaviour might be regulated by an intronic splicing enhancer located far upstream (first intron) from the concerned splice site (Lenasi et al., 2006). A similar low-molecular-weight (17 kDa) β-CN variant detected in donkey milk (Bertino et al., 2010) could be homologous to this equine β-CN splicing variant. On the other hand, probably because of its propensity to disassociate from the casein micelle at low temperatures, β-CN is the most susceptible CN to proteolysis by the endogenous milk protease (plasmin) which is responsible for the appearance of γ-CNs of which several were identified [γ2: f(112–226) and complement: f(1–111); γ4: f(120–226) and complement: f(1–119); γ6: f(56–226) and complement: f(1–55)] in this study.

Three αs1-CN isoforms were identified beside the full-length molecule, as a result of exons 5 and/or 7 skipping (**Supplementary material, Fig. S5**), as well as deletions of glutamine residues due to the loss of CAG codons at the 5′ head of exons 6′ (Q47) and 11 (Q96) which leads to nearly a dozen of different peptide sequences. Similar phenomena were reported in the horse (Matéos et al., 2009). αs1-CN, isolated from Haflinger mare's milk was shown to display, beside a great microheterogeneity due to variable level of phosphorylation, alternative splicing events involving exon 7, exon 14, or both, and the use of a cryptic splice site leading to the loss of the Q91 residue encoded by the first CAG codon of exon 11. This discrepancy regarding the numbering of the glutamine residue encoded by the first codon of exon 11 is due to the fact that exon 5, encoding the pentapeptide HTPRE, seemed to be constitutively missing in the protein sequence reported by Matéos and co-workers. However, exon 5 is clearly present at the horse genome level (**Supplementary material, Fig. S6**), in a context allowing theoretically its splicing-in in the mature transcript. Elsewhere, whereas the upstream and downstream environment of exon 14 appears to be very similar between the horse and the donkey genomes, we have no obvious explanation to the absence of skipping regarding exon 14 in the Amiata donkey milk samples. Intron sequences upstream and downstream exon 14 are identical. Interestingly, multiple alignments (**Supplementary material, Fig. S5**) shows that exon 3′ which is present across all species including *Equideas* is lacking in cattle. *A contrario* exon 13 is absent in all species except in ruminants, and we were not capable to find out any traces of this exon in the downstream intron sequence. By contrast, as reported previously (Martin et al., 1996), a virtual 18-nucleotide exon 13′, homologous to the sequence encoding the donkey's hexapeptide QAIYAQ, is present in the bovine genome, coding for EAIHDH. A G/A transition, at the first position of the 5′ donor splice site, known to recognize the upstream sequence as an exon, is very likely responsible for out-splicing of the virtual bovine exon 13′ which is nevertheless preceded by a typical polypyrimidine tract.

From a more general point of view, regarding the perspective to use LC-ESI-MS technology to analyse intact milk proteins, insofar as it is possible to quantify the different proteoforms present in milk, as has been shown for the bovine species (Miranda et al., 2020), this technology is undoubtedly, beyond the analysis of proteomes, a powerful tool to evaluate the actual impact of genetic diversity on the technological quality of milk at the phenotypic level. Furthermore, as reported for camelid milk (Ryskaliyeva et al., 2019a), the identification of splicing variants modifying the structure of the peptide chain provides elements that are still insufficiently exploited to understand and relate species specific peptides to the original properties of the milks of species wrongly considered as minor, such as camelids and donkey. For example, it was recently reported (Akan, 2021) that peptides released from digested casein fraction of camel and donkey milks have potent antioxidant and particularly antidiabetic properties. It remains now to identify the bioactive peptides involved.

5. Conclusions

From the present study, undertaken as a first attempt to explore the genetic variability of milk proteins in Amiata donkey, combining proven proteomic and molecular biology approaches, we highlighted original findings that allowed to address the extreme complexity of the donkey CN fraction, especially regarding αs-CN due to PTM (phosphorylation) and multiple splicing events (exon skipping and cryptic splice site usage). In addition, by demonstrating the effective presence of αs2-like-CN in donkey milk, we provide definite evidence of the expression of a second gene (*CSN1S2 II*) in the genome of this species. Finally, new genetic variants for both CNs and whey proteins were identified and we detect a so far unknown protein, very likely related to β-LG I. In-depth studies are currently underway to study this protein from a genomic point of view, in order to characterize the mutational events responsible for its occurrence.

CRediT authorship contribution statement

Barbara Auzino: Resources, Investigation, Writing – original draft. **Guy Miranda:** Methodology, Investigation, Visualization. **Céline Henry:** Methodology, Investigation, Visualization. **Zuzana Krupova:** Methodology, Investigation, Visualization. **Mina Martini:** Resources, Funding acquisition, Writing – review & editing. **Federica Salari:** Resources, Investigation. **Gianfranco Cosenza:** Conceptualization, Funding acquisition, Supervision, Writing – review & editing. **Roberta Ciampolini:** Conceptualization, Funding acquisition, Supervision, Project administration, Writing – review & editing. **Patrice Martin:** Conceptualization, Methodology, Investigation, Data curation, Visualization, Funding acquisition, Supervision, Writing – original draft, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to warmly thank the Regional Forestry Agricultural Complex Bandite di Scarlino (Grosseto, Italy) for help in milk sampling and the Doctoral School of Veterinary Sciences from the University of Pisa, for the scholarship granted to BA which allowed her to do an internship at the INRA Research Center of Jouy-en-Josas in the framework of her PhD. This work was partly supported by a grant PRA 2017 (Ateneo Research Project, University of Pisa) Improvement of productive performance of dairy donkey (Miglioramento delle performance produttive della specie asinina allevata per la produzione di latte) nI52F17000260005. The authors thank PAPPSO and @BRIDGe teams at INRAe (Jouy-en-Josas, France) for providing necessary facilities and technical support.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.foodres.2022.111611.

References

Akan, E. (2021). An evaluation of the in vitro antioxidant and antidiabetic potentials of camel and donkey milk peptides released from casein and whey proteins. *Journal of Food Science and Technology, 58*(10), 3743–3751.

Altomonte, I., Salaria, F., Licitra, R., & Martini, M. (2019). Donkey and human milk: Insights into their compositional similarities. *International Dairy Journal, 89*, 111–118.

Bertino, E., Gastaldi, D., Monti, G., Baro, C., Fortunato, D., Perono Garoffo, L., … Conti, A. (2010). Detailed proteomic analysis on DM: Insight into its hypoallergenicity. *Frontiers in Bioscience, 2*(2), 526–536.

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.

Brinkmann, J., Koudelka, T., Keppler, J. K., Tholey, A., Schwarz, K., Thaller, G., & Tetens, J. (2015). Characterization of an Equine αS2-Casein variant due to a 1.3 kb deletion spanning 2 coding exons. *PLoS One, 10*, Article e0139700.

Brinkmann, J., Jagannathan, V., Drögemüller, C., Rieder, S., Leeb, T., Thaller, G., & Tetens, J. (2016). Genetic variability of the equine casein genes. *Journal of Dairy Science, 99*, 5486–5497.

Brumini, D., Criscione, A., Bordonaro, S., Vegarud, G. E., & Marletta, D. (2016). Whey proteins and their antimicrobial properties in donkey milk: A brief review. *Dairy Science & Technology, 96*, 1–14.

Caroli, A. M., Chessa, S., & Erhardt, G. J. (2009). Milk protein polymorphisms in cattle: Effect on animal breeding and human nutrition. *Journal of Dairy Science, 92*, 5335–5352.

Chianese, L., Calabrese, M. G., Ferranti, P., Mauriello, R., Garro, G., De Simone, C., … Ramunno, L. (2010). Proteomic characterization of donkey milk "caseome". *Journal of Chromatography, 1217*, 4834–4840.

Chianese, L., De Simone, C., Ferranti, P., Mauriello, R., Costanzo, A., Quarto, M., & Ramunno, L. (2013). Occurrence of qualitative and quantitative polymorphism at donkey beta-Lactoglobulin II locus. *Food Research International, 54*, 1273–1279.

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 and polymorphisms detection. *Italian Journal of Animal Science, 9*, 206–211.

Criscione, A., Cunsolo, V., Bordonaro, S., Guastella, A.-M., Saletti, R., Zuccaro, A., … Marletta, D. (2009). Donkeys' milk protein fraction investigated by electrophoretic methods and mass spectrometric analysis. *International Dairy Journal, 19*, 190–197.

Criscione, A., Cunsolo, V., Tumino, S., Di Francesco, A., Bordonaro, S., Muccilli, V., … Marletta, D. (2018). Polymorphism at donkey β-lactoglobulin II locus: Identification and characterization of a new genetic variant with a very low expression. *Amino Acids, 50*, 735–746.

Cunsolo, V., Costa, A., Saletti, R., Muccilli, V., & Foti, S. (2007). Detection and sequence determination of a new variant beta-LG II from donkey. *Rapid Communication in Mass Spectrometry, 21*, 1438–1446.

Cunsolo, V., Cairone, E., Fontanini, D., Criscione, A., Muccilli, V., Saletti, R., & Foti, S. (2009a). Sequence determination of αs1-casein isoforms from donkey by mass spectrometric methods. *Journal of Mass Spectrometry, 44*, 1742–1753.

Cunsolo, V., Cairone, E., Saletti, R., Muccilli, V., & Foti, S. (2009b). Sequence and phosphorylation level determination of two donkey beta-caseins by mass spectrometry. *Rapid Communication in Mass Spectrometry, 23*, 1907–1916.

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, 99*, 41–47.

Derdak, R., Sakoui, S., Pop, O. L., Muresan, C. I., Vodnar, D. C., Addoum, B., … El Khalfi, B. (2020). Insights on Health and Food Applications of Equus asinus (Donkey) Milk Bioactive Proteins and Peptides - An Overview. *Foods., 9*, 1302.

Farrell, H. M., Jr, Malin, E. L., Brown, E. M., & Mora-Gutierrez, A. (2009). Review of the chemistry of alphaS2-casein and the generation of a homologous molecular model to explain its properties. *Journal Dairy Science, 92*, 1338–1353.

Gnanesh Kumar, B. S., & Rawal, A. (2020). Sequence Characterization and N-Glycoproteomics of Secretory Immunoglobulin A From Donkey Milk. *International Journal of Biological Macromolecules, 155*, 605–613.

Giuffrida, M. G., Cantisani, A., Napolitano, L., Conti, A., & Godovac-Zimmermann, J. (1992). The amino acid sequence of two isoforms of alpha-lactalbumin from donkey (Equus asinus) milk is identical. *Biological chemistry Hoppe-Seyler, 373*, 931–935.

Godovac-Zimmermann, J., Conti, A., & Napolitano, L. (1988). The primary structure of donkey (Equus asinus) lysozyme contains the Ca(II) binding site of alpha-lactalbumin. *Biological Chemistry Hoppe Seyler, 369*, 1109–1115.

Godovac-Zimmermann, J., Conti, A., Sheil, M., & Napolitano, L. (1990). Covalent structure of the minor monomeric beta-lactoglobulin II component from donkey milk. *Biological chemistry Hoppe-Seyler, 371*, 871–879.

Herrouin, M., Mollé, D., Fauquant, J., Ballestra, F., Maubois, J.-L., & Léonil, J. (2000). New genetic variants identified in donkey's milk whey proteins. *Journal of Protein Chemistry, 19*, 105–115.

Huang, J., Zhao, Y., Bai, D., Shiraigol, W., Li, B., Yang, L., … Dugarjaviin, M. (2015a). Donkey genome and insight into the imprinting of fast karyotype evolution. *Scientific Reports, 5*, 14106.

Huang, J., Guerrero, A., Parker, E., Strum, J. S., Smilowitz, J. T., German, J. B., & Lebrilla, C. B. (2015b). Site-specific glycosylation of secretory immunoglobulin A from human colostrum. *Journal of Proteome Research, 14*, 1335–1349.

Hue-Beauvais, C., Miranda, G., Aujean, E., Jaffrezic, F., Devinoy, E., Martin, P., & Charlier, M. (2017). Diet-induced modifications to milk composition have long-term effects on offspring growth in rabbits. *Journal of Animal Science, 95*, 761–770.

Jirillo, F., Jirillo, E., & Magrone, T. (2010). Donkey's and goat's milk consumption and benefits to human health with special reference to the inflammatory status. *Current Pharmaceutical Design, 16*, 859–863.

Kocic, H., Stankovic, M., Tirant, M., Lotti, T., & Arsic, I. (2020). Favorable effect of creams with skimmed donkey milk encapsulated in nanoliposomes on skin physiology. *Dermatologic Therapy, 33*(4), Article e13511.

Laemmli, U. K. (1970). Cleavage of structural proteins during the assembly of the head of the bacteriophage T4. *Nature, 227*, 680–685.

Lenasi, T., Rogelj, I., & Dovc, P. (2003). Characterization of equine cDNA sequences for αs1-, β- and κ-casein. *Journal of Dairy Research, 70*, 29–36.

Lenasi, T., Kokalj-Vokac, N., Narat, M., Baldi, A., & Dovc, P. (2005). Functional study of the equine betacasein and kappa-casein gene promoters. *Journal of Dairy Research, 72*, 34–43.

Lenasi, T., Peterlin, B. M., & Dovc, P. (2006). Distal regulation of alternative splicing by splicing enhancer in equine β-casein intron 1. *RNA, 12*, 498–507.

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

Licitra, R., Chessa, S., Salaria, F., Gattolinb, S., Bulgarid, O., Altomontea, I., & Martini, M. (2019). Milk protein polymorphism in Amiata donkey. *Livestock Science, 230*, Article 103845.

Malacarne, M., Criscione, A., Franceschi, P., Bordonaro, S., Formaggioni, P., Marletta, D., & Summer, A. (2019). New Insights into Chemical and Mineral Composition of Donkey Milk throughout Nine Months of Lactation. *Animals, 9*(12), 1161.

Martin, P., & Leroux, C. (1992). Exon-skipping is responsible for the 9 amino acid residue deletion occurring near the N-terminal of human beta-casein. *Biochemical and Biophysical Research Communications, 183*(2), 750–757.

Martin, P., Brignon, G., Furet, J.-P., & Leroux, C. (1996). The gene encoding αs1-casein is expressed in human mammary epithelial cells during lactation. *Lait, 76*, 523–535.

Martin, P., Bianchi. L., Cebo C. & Miranda G. (2012). *Genetic polymorphism of Milk Proteins. Advanced Dairy Chemistry, Volume 1A: Proteins, Basic Aspects*, 463-514, 4th Edition.

Martini, M., Altomonte, I., Tricò, D., Lapenta, R., & Salari, F. (2021). Current Knowledge on Functionality and Potential Therapeutic Uses of Donkey Milk. *Animals, 11*(5), 1382.

Massouras, T., Bitsi, N., Paramithiotis, S., Manolopoulou, E., Drosinos, E. H., & Triantaphyllopoulos, K. A. (2020). Microbial Profile Antibacterial Properties and Chemical Composition of Raw Donkey Milk. *Animals, 10*, 2001.

Matéos, A., Miclo, L., Mollé, D., Dary, A., Girardet, J.-M., & Gaillard, J.-L. (2009). Equine αs1- casein: Characterization of alternative splicing isoforms and determination of phosphorylation levels. *Journal of Dairy Science, 92*, 3604–3615.

Matéos, A., Girardet, J.-M., Mollé, D., Corbier, C., Gaillard, J.-L., & Miclo, L. (2010). Identification of phosphorylation sites of equine β-casein isoforms. *Rapid Communications in Mass Spectrometry, 24*, 1533–1542.

Meisel, H. (2004). Multifunctional peptides encrypted in milk proteins. *Biofactors., 21*, 55–61.

Menon, R. S., Chang, Y. F., Jeffers, K. F., & Ham, R. G. (1992). Exon skipping in human beta-casein. *Genomics, 12*, 13–17.

Mercier, J.-C. (1981). Phosphorylation of caseins, present evidence for an amino acid triplet code posttranslationally recognized by specific kinases. *Biochimie, 63*, 1–17.

Miclo, L., Girardet, J.-M., Egito, A. S., Mollé, D., Martin, P., & Gaillard, J.-L. (2007). The primary structure of a low-Mr multiphosphorylated variant of beta-casein in equine milk. *Proteomics, 7*, 1327–1335.

Milenkovic, D., Martin, P., Guérin, G., & Leroux, C. (2002). A specific pattern of splicing for the horse alphaS1-Casein mRNA and partial genomic characterization of the relevant locus. *Genetics Selection Evolution, 34*, 509–519.

Miranda, G., Mahé, M. F., Leroux, C., & Martin, P. (2004). Proteomic tools to characterize the protein fraction of Equidae milk. *Proteomics, 4*, 2496–2509.

Miranda, G., Krupova, Z., Bianchi, L. & Martin, P. (2013). A novel LC-MS protein profiling method to characterize and quantify individual milk proteins and multiple isoforms. In *10th Annual International Milk Genomics Consortium Symposium. October 1–3, 2013*. University of California-Davis Conference Center, Davis.

Miranda, G., Bianchi, L., Krupova, Z., Trossat, P., & Martin, P. (2020). An improved LC-MS method to profile molecular diversity and quantify the six main bovine milk proteins, including genetic and splicing variants as well as post-translationally modified isoforms. *Food Chemistry X, 5*, Article 100080.

Monti, G., Viola, S., Baro, C., Cresi, F., Tovo, P. A., Moro, G., … Bertino, E. (2012). Tolerability of donkey's milk in 92 highly-problematic cow's milk allergic children. *Journal of Biological Regulators and Homeostatic Agents, 26*, 75–82.

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, 2*, 18–27.

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*, Article e0124963.

Polidori, P., & Vincenzetti, S. (2010). Differences of Protein Fractions Among Fresh, Frozen and Powdered Donkey Milk. *Recent Patents on Food, Nutrition & Agriculture, 2*, 56–60.

Ragona, G., Corrias, F., Benedetti, M., Paladini, M., Salari, F., Altomonte, I., & Martini, M. (2016). Amiata Donkey Milk Chain: Animal Health Evaluation and Milk Quality. *Italian Journal of Food Safety, 5*(3), 5951.

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*, 289–299.

Renaud, G., Petersen, B., Seguin-Orlando, A., Bertelsen, M. F., Waller, A., Newton, R., … Orlando, L. (2018). Improved de novo genomic assembly for the domestic donkey. Science. *Advances, 4*, eaaq0392.

Rinjkels, M. (2002). Multispecies Comparison of the Casein Gene Loci and Evolution of Casein Gene Family. *Journal of Mammary Gland Biology and Neoplasia, 7*, 327–345.

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*, Article e0197026.

Ryskaliyeva, A., Henry, C., Miranda, G., Faye, B., Konuspayeva, G., & Martin, P. (2019a). Alternative splicing events expand molecular diversity of camel CSN1S2 increasing its ability to generate potentially bioactive peptides. *Scientific Reports, 9*, 5243.

Ryskaliyeva, A., Henry, C., Miranda, G., Faye, B., Konuspayeva, G., & Martin, P. (2019b). 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.

Saletti, R., Muccilli, V., Cunsolo, V., Fontanini, D., Capocchi, A., & Foti, S. (2012). MS-based characterization of α(s2)-casein isoforms in donkey's milk. *Journal of Mass Spectrometry, 47*, 1150–1159.

Salimei, E., & Fantuz, F. (2012). Equid milk for human consumption. *International Dairy Journal, 24*, 130–142.

Sarti, L., Martini, M., Brajon, G., Barni, S., Salari, F., Altomonte, I., … Novembre, E. (2019). Donkey's Milk in the Management of Children with Cow's Milk protein allergy: Nutritional and hygienic aspects. *Italian Journal of Pediatrics, 45*, 102.

Selvaggi, M., Laudadio, V., Dario, C., & Tufarelli, V. (2014a). Major proteins in goat milk: An updated overview on genetic variability. *Molecular Biology, 41*(2), 1035–1048.

Selvaggi, M., Laudadio, V., Dario, C., & Tufarelli, V. (2014b). Investigating the genetic polymorphism of sheep milk proteins: A useful tool for dairy production. *Journal of the Science of Food and Agriculture, 94*, 3090–3099.

Selvaggi, M., D'Alessandro, A. G., & Dario, C. (2015). Comparative characteristics of DNA polymorphisms of κ-casein gene (CSN3) in the horse and donkey. *Genetics and Molecular Research, 14*, 14567–14575.

Suteu, M., Vlaic, A., Bâlteanu, V. A., Wavreille, J., & Renaville, R. (2011). Evidence of alternative splicing of porcine β-casein (CSN2). *Animal Genetics, 43*, 474–475.

Tafaro, A., Magrone, T., Jirillo, F., Martemucci, G., D'Alessandro, A. G., Amati, L., & Jirillo, E. (2007). Immunological properties of donkey's milk: Its potential use in the prevention of atherosclerosis. *Current Pharmaceutical Design, 13*, 3711–3717.

Vincenzetti, S., Foghini, L., Pucciarelli, S., Polzonetti, V., Cammertoni, N., Beghelli, D., & Polidori, P. (2014). Hypoallergenic properties of donkey's milk: A preliminary study. *Veterinaria Italiana, 50*, 99–107.

Visser, S., Slangen, C. J., & Rollema, H. S. (1991). Phenotyping of bovine milk proteins by reversed-phase high-performance liquid chromatography. *Journal of Chromatography, 548*, 361–370.

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, 92*, 1885–1888.

Zuccaro, A., Guastella, A. M., Tidona, F., Bordonaro, S., Andrea, C., & Marletta, D. (2010). *Direct Submission in EMBL/GenBank/DDBJ Databases*.