

AperTO - Archivio Istituzionale Open Access dell'Università di Torino

The β -casein in camels: molecular characterization of the CSN2 gene, promoter analysis and genetic variability.

This is the author's manuscript

Original Citation:

Availability:

This version is available <http://hdl.handle.net/2318/150208> since 2015-11-26T10:42:23Z

Published version:

DOI:10.1016/j.gene.2014.06.055

Terms of use:

Open Access

Anyone can freely access the full text of works made available as "Open Access". Works made available under a Creative Commons license can be used according to the terms and conditions of said license. Use of all other works requires consent of the right holder (author or publisher) if not exempted from copyright protection by the applicable law.

(Article begins on next page)



UNIVERSITÀ DEGLI STUDI DI TORINO

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22

This Accepted Author Manuscript (AAM) is copyrighted and published by Elsevier. It is posted here by agreement between Elsevier and the University of Turin. Changes resulting from the publishing process - such as editing, corrections, structural formatting, and other quality control mechanisms - may not be reflected in this version of the text. The definitive version of the text was subsequently published in [*Pauciullo A., Giambra I.J., Iannuzzi L., Erhardt G. (2014). The β -casein in camels: molecular characterization of the CSN2 gene, promoter analysis and genetic variability. Gene. 547,159-168 - DOI: 10.1016/j.gene.2014.06.055*].

You may download, copy and otherwise use the AAM for non-commercial purposes provided that your license is limited by the following restrictions:

- (1) You may use this AAM for non-commercial purposes only under the terms of the CC-BY-NC-ND license.
- (2) The integrity of the work and identification of the author, copyright owner, and publisher must be preserved in any copy.
- (3) You must attribute this AAM in the following format: Creative Commons BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/deed.en>), [+ *Digital Object Identifier link to the published journal article on Elsevier's ScienceDirect® platform: 10.1016/j.gene.2014.06.055*]

1 **The β -casein in camels: molecular characterization of the *CSN2* gene, promoter analysis and**
2 **genetic variability**

3
4 Pauciullo A.^{1,2,*}, Giambra I.J.¹, Iannuzzi L.², Erhardt G.¹

5
6
7
8
9
10 ¹*Institute for Animal Breeding and Genetics, Justus Liebig University, Ludwigstraße 21 B, 35390*
11 *Gießen, Germany*

12 ²*ISPAAM, Laboratory of Animal Cytogenetics and Gene Mapping, National Research Council, via*
13 *Argine 1085, 80147 Naples, Italy*

14
15
16
17
18 *For correspondence: Alfredo Pauciullo, Ph.D

19 email: alfredo.pauciullo@agrar.uni-giessen.de; alfredo.pauciullo@cnr.it

20 Institute for Animal Breeding and Genetics, Justus Liebig University, Ludwigstraße 21 B, 35390
21 Gießen, Germany

22 Phone: +4906419937672

1 **ABSTRACT**

2 The β -casein is the most abundant protein in camel milk and its encoding gene (*CSN2*) is
3 considered in other species a ‘major’ gene for the presence of alleles associated to different level of
4 expression. We report for the first time the characterization of the nucleotide sequence of the whole
5 β -casein-encoding gene (*CSN2*) plus 2141 bp at the 5’-flanking region in *Camelus dromedaries* and
6 in addition the promoter region and the complete cDNA in *Camelus bactrianus*. The gene is spread
7 over 7.8 kb and consists of 9 exons varying in length from 24 bp (exon 5) to 519 bp (exon 7), and 8
8 introns from 95 bp (intron 5) to 1950 bp (intron 1). The composite response element (CoRE) region
9 was identified in the promoter, whereas the presence of mature microRNAs sequences improves the
10 knowledge on the factors putatively involved in the gene regulation. A total of 46 polymorphic sites
11 have been detected. The transition g.2126A>G falls within the TATA-box of dromedary *CSN2*
12 promoter with a putative influence on the transcription factor binding activity. The frequency of the
13 G allele was 0.35 in a population of 180 she-camels belonging to 4 different ecotypes. In the same
14 population, a conservative SNP (g.4175C>A) was found at the codon 7 of the signal peptide,
15 whereas the comparative analysis with the cDNA available in EMBL evidenced a missense SNP
16 (g.4180T^{Leu}>G^{Arg}) at the exon 2. Four SNPs were found in the bactrian camel. The SNP c.666G>A
17 is responsible for the amino acid change Met²⁰¹→Ile and it represents the first missense allele at the
18 β -casein in camels. Finally, five interspersed repeated elements were identified at the intronic level
19 and putative bio-functional peptides belonging to ACE-inhibitor and anti-oxidative families were
20 identified using in *silico* analysis.

21

22 **Keywords:** Camel, Casein, *CSN2*, Promoter analysis, Interspersed elements, bio-peptides,
23 microRNA.

24

1 **Abbreviations:** *CSN2*, β -casein-encoding gene; SNP, single nucleotide polymorphism; CN,
2 casein: cDNA, DNA complementary to RNA; EDTA, ethylene diamine tetra acetic acid; SDS,
3 sodium dodecyl sulfate; PCR, polymerase chain reaction; $MgCl_2$, magnesium chloride; dNTP,
4 deoxyribonucleoside triphosphate; TBE, tris boric acid EDTA; RFLP, restriction fragment length
5 polymorphism; pI, isoelectric point; CoRE, composite response element; C/EBP, CCAAT enhancer
6 bind protein; MGF, Mammary Gland Factor; GR, glucocorticoid receptor factor; YY1, Ying Yang
7 1; Oct-1, Octamer bind protein; TBP, Tata Binding Protein; NF-1, Nuclear factor-1; STR, short
8 tandem repeat; SINE, short interspersed element; LINE, long interspersed element; MIR,
9 mammalian-wide interspersed repeats; miR, microRNA.

10

11 **1. INTRODUCTION**

12 Camel's milk has played and plays a key role for the outliving of the Bedouin, nomad and
13 pastoral households in the hard arid and semi-arid regions of the world, supporting migrant
14 populations since their domestication happened millennia ago. The Camelidae family consists of six
15 species (*C. dromedarius*, *C. bactrianus*, *L. glama*, *L. guanicoe*, *V. pacos*, *V. vicugna*) among which
16 the dromedary camel is certainly the most economically important domesticated one. According to
17 FAO statistics (2011), most of the dromedary camels lives in the Horn of Africa mainly gathered in
18 Somalia, Sudan, Ethiopia and Kenya, which together hold almost 60% of the entire population
19 estimated to be about 24.5 million in the world. Despite their potential to survive on marginal
20 resources in severe environment, camels have not been exploited as an important food source in the
21 aforementioned countries. For instance, only 10% of the total milk produced in these rural regions
22 is of camel origin (Faye and Konuspayeva, 2012). Conversely, in the countries of the Gulf,
23 intensive camel milk production in high scale modernized unit has been already realized (Faye et al.
24 2002) and genetic improvement programmes for the milk production have been implemented
25 (Hermas, 1988; Nagy et al. 2013).

1 The daily milk production of dromedary camels is estimated to vary between 3 and 10 kg during
2 a lactation period of 12-18 months (Farah et al. 2007), depending on breed, stage of lactation,
3 feeding and management conditions, with an average content of 2.9% and 3.1% of protein and fat
4 respectively (Al haj and Al Kanhal, 2010).

5 As for the other mammals, the main component of camel milk proteins are caseins, among
6 which the β -casein is the most abundant component (~65%) (Kappeler et al. 2003; El-Agamy,
7 2006). This amount is definitively higher than the 45% reported in bovine milk (Farrell et al. 2004)
8 and, since β -CN is more sensitive to peptic hydrolysis than α s-CN (El-Agamy et al. 2009), the
9 percentage of this casein fraction could reflect its higher digestibility rate and the lower incidence of
10 allergy in the gut in infants nourished with camel milk (El-Agamy et al. 2009).

11 The occurrence of different levels of phosphorylation of the β -casein are reported to affect the
12 availability and distribution of calcium and to influence the micelle stability (Amigo et al. 2000),
13 therefore the β -CN plays an essential role both for the nutrition aspects and for the impact on the
14 technological properties of the milk and dairy products.

15 The β -casein fraction has been deeply studied in ruminants and well characterized both at the
16 protein and DNA level. Many alleles associated with different rates of protein synthesis have been
17 identified in the corresponding coding gene (*CSN2*). At least 6 genetic variants have been reported
18 in sheep (Chianese et al. 1997, Ceriotti et al. 2004; Chessa et al. 2010), whereas at least 8 alleles
19 corresponding to 7 β -CN variants have been identified in goat (Cosenza et al. 2005; Caroli et al.
20 2006; Cosenza et al. 2007), and at least 17 alleles corresponding to 12 β -CN variants have been
21 identified so far in cattle (Caroli et al. 2009). Conversely, in camels no genetic variants have been
22 reported so far for the β -casein, and all the information we have -to date- (HPLC migration,
23 molecular mass, 2D electrophoresis pattern, etc...) have been determined at the protein level (Farah
24 and Farah-Riesen, 1985; Kappeler et al. 1998; El-Agamy et al. 2009; Hinz et al. 2012).

1 Genetic variants in camel milk are known for the α s1-casein fraction (Shuiep et al. 2013),
2 whereas Kappeler et al. (1998) reported on non-allelic variants of the same casein fraction and
3 described an amino acid exchange in position 30 (p.Glu > Asp) of the mature protein, which is not
4 mentioned in UniProt-database [O97943](#). The κ -casein gene (*CSN3*) has been also investigated in
5 camels and 17 polymorphic sites have been recently detected by Pauciullo et al. (2013).

6 The investigation at the DNA level of camel β -casein gene (*CSN2*) is limited to the cDNA
7 sequence (Kappeler et al. 1998) and to a comparison of the 5' flanking regions (Kappeler et al.
8 2003). Therefore, the aim of the present study was a full characterization and an extensive
9 annotation of the *CSN2* gene and promoter region. Furthermore, considering the growing interest
10 that camel milk is globally receiving especially for the potential health benefits obtainable through a
11 number of bioactive components of camel milk (for a review Al hay and Al Kanhal, 2010), a deep
12 investigation was undertaken to explore genetic variability at the *Camelus dromedaries* and
13 *Camelus bactrianus CSN2*.

15 2. MATERIALS AND METHODS

16 2.1. Animals

17 Blood samples were collected from 180 Sudanese she-camels (*Camelus dromedarius*, locally
18 known as *Naga*) reared in five regions of the country and belonging to four different ecotypes:
19 Shanbali, Kahli, Lahaoui and Arabi camels. Due to the lack of records, local experience was used to
20 determine ecotypes in different locations and herds according to typical conformational/physical
21 characteristics (weight, hump, legs, body, colour, etc...). Blood samples were immediately applied
22 to classic filter paper (FTA[®] Classic Card- Whatman[®] BioScience, Maidstone, UK), allowed to dry
23 at room temperature and stored until DNA isolation.

24 Additional blood samples were collected from 3 *Camelus bactrianus* (bactrian camel) belonging
25 to Wilhelma Zoo (Stuttgart, Germany), whereas hair samples were collected from further 5 bactrian

1 camels at the Frankfurt Zoo (Frankfurt, Germany). These samples were treated according to Spin
2 Blood Mini Kit (Invitex, Germany) and Gen-ial All Tissue DNA Kit (Gen-ial, Germany),
3 respectively.

4

5 **2.2.DNA isolation**

6 The filter paper containing blood samples was soaked (56°C, overnight) in 500 µl sodium-tris-
7 EDTA buffer with 10 µl proteinase K (10 mg/ml) in presence of sodium dodecyl sulfate (SDS).
8 Then DNA was isolated from the emerging lysis according to the procedure described by Sambrook
9 et al. (1989). The isolated DNA was resuspended in 100 µl TE buffer pH 7.6 (10 mM Tris, 1mM
10 EDTA).

11 DNA concentration and OD_{260/280} ratio of the samples were measured with the Nanodrop ND-
12 1000 Spectrophotometer (Thermo Scientific).

13

14 **2.3.PCR amplification conditions and sequencing**

15 Four test samples (one for each ecotype) were chosen for the sequencing of the whole camel
16 *CSN2* gene. DNA regions spanning from nucleotide -2141 to +7845 were amplified by means of a
17 iCycler (Bio-Rad). A set of 31 primers for amplification and sequencing were designed by means of
18 DNAsis-Max ver. 3.0 software (Hitachi), using as preliminary template the complete sequence of
19 camel cDNA available in gene bank (EMBL acc. no. [AJ012630](#)) and then the new sequences
20 determined in the course of the research. A typical PCR reaction mix (50 µl) comprised: 100 ng of
21 genomic DNA, 1X PCR Buffer (Promega), 2.5 mM MgCl₂, 5 pmol of each primer, dNTPs each at
22 200 µM, 2.5 U of *Taq* DNA Polymerase (Promega). In table 1 we report the 10 starting overlapping
23 amplicons for the sequencing of the whole gene. PCR was performed under the following
24 conditions: 95°C (4 min), 35 cycles at 95°C (60 s), annealing temperatures depending on amplicon
25 (Table 1) (45 s), 72°C (90 s), final extension at 72°C (10 min).

1 PCR products were purified and sequenced. The purification was carried out using MSB[®] Spin
2 PCRapace kit (Invitex, Germany). Sequence was accomplished by ABI 3130 Genetic Analyzer
3 (Applied Biosystem). Targeted fragments were sequenced in both directions using BigDye
4 chemistry (Applied Biosystems).

5 SNP discovery was accomplished by the re-sequencing of additional 40 DNA samples (10 for
6 each ecotype) for all the exons including the corresponding flanking regions and the gene promoter.

7

8 ***2.4. Genotyping by HphI PCR-RFLP***

9 PCR reaction mixture and thermal conditions for the amplification of the DNA fragment 659 bp
10 long and spanning from -428 bp of 5' flanking region to +231 bp of the camel CSN2 gene were
11 accomplished by using the following primers: *forward* 5'-GTTTCTCCATTACAGCATC-3' and
12 *reverse* 5'-TCAAATCTATACAGGCACTT-3', according to standard PCR conditions already
13 reported above. Product specificity was confirmed by ethidium-bromide-stained 1.5% agarose gel
14 electrophoresis.

15 The entire panel of 180 animals was genotyped for the g.2126A>G SNP using a PCR-RFLP
16 method. Digestion of 17µl of each PCR amplification was accomplished with 5 U of *HphI*
17 endonuclease (5'...GGTGAN₈↓...3') (New England Biolabs) over-night at 37°C. The digestion
18 products were analysed directly by electrophoresis in 2.5% agarose gel in 0.5X TBE buffer and
19 stained with Midori Green Advance (Nippon Genetics).

20

21 ***2.5. Bioinformatics***

22 The allele frequencies and Hardy-Weinberg equilibrium (χ^2 test) were calculated for the 180
23 camels. SNP discovery, homology searches, comparison among sequences, and multiple alignments
24 were accomplished using DNAsis-Max ver. 3.0 software (Hitachi Software), whereas the putative
25 transcription factor binding sites were searched by Transfact 7.0 software. Interspersed elements
26 were found by RepeatMasker Web Server (<http://www.repeatmasker.org/cgi->

1 [bin/WEBRepeatMasker](#)). Introns were also analyzed for potential microRNA sequences by using
2 the bovine miRBase database (<http://www.mirbase.org/search.shtml>). NetPhos server 2.0
3 (<http://www.cbs.dtu.dk/services/NetPhos/>) was used for the prediction of phosphorylation sites,
4 whereas BIOPEP database (<http://www.uwm.edu.pl/biochemia/index.php/en/biopep>) was used for
5 the detection of putative bio-peptides.

6

7 **3. RESULTS AND DISCUSSION**

8

9 ***3.1. Analysis of the gene structure and genetic variability in dromedary camel***

10 The entire gene encoding the camel β -casein (*CSN2*) plus 2141 bp at the 5' flanking region were
11 sequenced and the sequences was deposited in EMBL under the following ID number: HG969421.
12 The gene extends over 7819 nucleotides including 1115 bp of exonic regions and 6704 bp of
13 intronic regions. The comparison with the corresponding bovine sequence (EMBL ID: [X14711](#))
14 showed a total similarity of about 78.0%. Approximately the same level of homology was also
15 found by the following comparison to the β -casein gene in other species: 78.0% vs river buffalo
16 (EMBL ID: [FN424088](#)), goat (EMBL ID: [AJ011018](#)) and pig (EMBL ID: [EU025876](#)), 73.0% vs
17 donkey (EMBL ID: [FN598778](#)) and 70.0% vs human (EMBL ID: [AF027807](#)).

18 On the whole, the camel *CSN2* gene shares a similar organization with the bovine counterpart,
19 with some differences both at the exonic and intronic level. The camel cDNA is in total 15 bp
20 longer than the bovine one (1115 bp vs 1100 bp). The main differences in size were detected at the
21 exon 1 in the 5' un-translated region (51 bp vs 44 bp), at the exon 6 for one triplet (45 bp vs 42 bp)
22 and at the exon 7 for 7 triplets (519 bp vs 498 bp). Instead, the exon 9 corresponding to the 3' un-
23 translated region is shorter in camel than in cows (317 bp vs 323 bp). The intron sizes of the β -
24 casein gene are approximately comparable between these two species, however, the exon/intron size
25 ratio is higher in camel (1:6.01) than that observed in cattle (1:6.74).

1 The camel *CSN2* gene is characterized by high A/T content compared to G/C (66.5% vs 33.4%).
2 This feature characterizes also the camel κ -casein gene (Pauciullo et al. 2013) and in general seems
3 to be conserved among the species (Ward et al. 1997). It contains 9 exons, ranging in size from 24
4 bp (exon 5) to 519 bp (exon 7), and 8 introns from 95 bp (intron 5) to 1950 bp (intron 1). The first
5 exon (51 bp) plus 12 bp of the exon 2 are not coding at all. The signal peptide (15 amino acids) is
6 coded by the following 45 nucleotides of the exon 2, whereas the mature peptide (217 amino acids)
7 is encoded by the last 2 triplets of the exon 2 and the remaining part of the cDNA. The translation
8 stop codon TAA is realized between the 4th-6th nucleotides of the exon 8, whereas the
9 polyadenylation site is located between the 296th and 301st nucleotide of the exon 9. Splice donor
10 and acceptor consensus sequences conforming to the AG/TG rule were identified at the exon/ intron
11 boundaries.

12 The translation initiation codon (AUG) and the DNA sequence immediately upstream and
13 downstream are known to influence the translation process (Kozak, 1991). The consensus sequence
14 (GCCRCCAAUGG) was indicated as the optimal context for the initiation of the translation in
15 mammals sequence (Kozak, 1987) and the nucleotides -6, -3, and +4 (taking the A nucleotide of the
16 AUG codon as number +1) are reported as the most conserved positions in natural mRNA
17 sequences. With the exception of the AUG (nucleotides: +1, +2 and +3), the dromedary *CSN2*
18 showed only 2 nucleotides homologous to Kozak sequence (positions: -3 and -1). On the contrary,
19 the other species showed at least 3 conserved positions (table 2) for the β -casein. According to this
20 comparison, the dromedary camel is characterized by the lowest homology with Kozak consensus
21 site. Such condition should give a weaker translation efficiency, however, the camel β -casein shares
22 6 nucleotide positions out of 7 with ruminants (+4, -1, -3, -4, -5, -6) (Cosenza et al. 2011) and 5 out
23 of 7 with mice (+4, -1, -3, -4, -6), rabbits (+4, -1, -2, -3, -4) and dogs (+4, -1, -3, -4, -6) (table 2).
24 Therefore, it would be reasonable to achieve future translation efficiency studies to verify this
25 statement in dromedary camels.

1 SNP discovery was accomplished by the re-sequencing of 40 she-camels mainly for the exonic
2 regions and the gene promoter. The comparison of the sequences showed a total of 11 polymorphic
3 sites (7 transitions and 4 transversions) (table 3). One nucleotide difference was detected at the
4 exonic level. It is a transversion (g.4175C>A) which falls within the codon 7 of the signal peptide.
5 Although this mutation is synonymous (GCC^{Ala}→GCA), it represents the first example of
6 polymorphism in *CSN2* coding region for such species.

7 The remaining 10 polymorphic sites were detected in the rest of the gene, in particular 2 SNPs
8 in the promoter region and 8 at the intronic level (table 3). The following comparison with the
9 published sequences of the camel cDNA (Kappeler et al. 1998) shows additional 4 polymorphic
10 sites (table 4) one of which results in an amino acid substitution at the 9th triplet of the exon 2
11 (CTG^{Leu}→CGG^{Arg}) within the signal peptide.

12 Leucine is an apolar amino acid whereas arginine is basic. Therefore the polarity of the signal
13 peptide changes and its predicted isoelectric point (*pI*) shifts from 8.78 to 9.93. Since the charges of
14 milk proteins are important for their interactions and properties (Khaldi and Shields, 2011), both
15 protein variants were submitted to the SignalP V4.1 software (www.cbs.dtu.dk/services/SignalP) to
16 determine potential performances alterations. A higher probability of cleavage (0.819 vs 0.534) and
17 a slight increase in all the output scores (raw, signal and combined cleavage site) were detected for
18 the leucin variant, however both signal peptides were characterized by the same predicted cleavage
19 site, suggesting that this mutation does not affect their function.

20 Many genetic variants of β -casein have been identified so far at the DNA or protein level in
21 domestic animals. For instance, cattle (Caroli et al. 2009) and goat (Caroli et al. 2006) are the most
22 polymorphic species among ruminants, whereas pigs (Cieslak et al. 2012, Suteu et al. 2012) and
23 horses (Miclo et al. 2007) are less polymorphic and no polymorphisms were detected in other
24 investigated species like donkeys and rabbits. Camels were not investigated so far, but the limited
25 number of polymorphisms found in coding regions and the number of mutations detected in the rest
26 of the *CSN2* gene suggests that the genetic variability of camel β -casein is closer to non-ruminants.

1 However, the β -casein is the most abundant protein component also in camel milk (Kappeler et al.,
2 2003; El-Agamy, 2006) and any allele detected at *CSN2* locus might have a potential influence on
3 the quality and the technological properties of milk. For instance, the *CSN2* B allele was found to be
4 more favourable for rennet coagulation and the cheese making quality of cow milk (Di Stasio and
5 Mariani, 2000), whereas the A1 and A2 alleles which differ only for a transversion (A>C)
6 responsible for the His→Pro amino acid change at codon 67, show divergent bio-peptide activities
7 as consequence of a conformational variation in the secondary structure of the expressed protein
8 (Kaminski et al. 2007). Therefore, a larger investigation of β -casein should be also performed in
9 camels, in order to detect nucleotide variability and evaluate it for the potential influence on the
10 quality and properties of milk.

11

12 ***3.2. The β -casein cDNA in bactrian camel***

13 Starting from the genomic DNA, all the exonic regions of the β -casein were amplified and
14 sequenced also for *Camelus bactrianus* (EMBL ID: HG969423). This species resulted more
15 polymorphic for the *CSN2* locus than the dromedary camel. Four SNPs were detected in total (table
16 4), three of which in the coding region (exon 7) and the last one at the 3' UTR (exon 9). The SNPs
17 c.591G>A and c.732G>A are both conservative, whereas the SNP c.666G>A (corresponding to the
18 nucleotides 8177 of *C. dromedarius CSN2* gene) is responsible for the amino acid change Met→Ile
19 at the position 201 of the mature protein. This mutation represents in absolute the first missense
20 mutation described at the β -casein gene in this species and it identifies two variants that we name:
21 *CSN2* A (for the carriers of the G, which also characterizes the dromedary camel *CSN2* cDNA
22 sequence and it might be considered as the ancestral form of the gene), and *CSN2* B (for the carriers
23 of the A and it might be considered as the mutant form). The following comparison with the
24 dromedary cDNAs showed an additional polymorphic site in the exon 9 (table 4). This mutation
25 also falls within the 3' UTR, therefore it has no influence at the protein level. Analyzing the same

1 cDNA fragment within the four main ruminant species (cow, buffalo, goat and sheep) we have
2 found at least 32 SNPs and the insertion/deletion of 2 triplets (TATCCC) in cattle and buffalo β -
3 casein. These findings confirm again that the level of genetic variation detected within *Camelidae* is
4 lower than the corresponding variability found in ruminants both at the cDNA and at gene level.

5

6 ***3.3. Phosphorylation prediction and active bio-peptide analysis***

7 The complete cDNA of dromedary camel and the corresponding encoding protein is reported in
8 figure 1. The analysis of the serine/threonine phosphorylation sites predicted by NetPhos server
9 v2.0 indicated 10 putative sequence motifs. Previous studies have shown that the most frequent
10 form of camel β -casein is constitutively phosphorylated at Ser¹⁵, Ser¹⁷, Ser¹⁸ of the mature protein
11 (Kappeler et al. 1998). Recently, Saadaoui et al. (2014) showed the presence of an additional
12 phosphate, most probably the Ser¹⁹ as also predicted by Kappeler et al. (1998). These data agree
13 with the reports on bovine β -casein, which is characterized by the same phosphorylation sites and,
14 in addition, it is phosphorylated also at Ser⁵⁰ (Farrell et al. 2004). The last site is not conserved in
15 camels being substituted by Ile at the same position. However recently in cattle Li et al. (2012)
16 identified Thr⁵⁶ as a new phosphorylation site. This amino acid falls within a stretch of 9 amino
17 acids well conserved also in camels and, although the phosphorylation of threonine in the Thr-Xaa-
18 Asp motif is relatively uncommon for caseins (Holland, 2009), we can't exclude that such a
19 phosphorylation occurs also for camels β -casein. A variable degree of phosphorylation has been
20 already demonstrated in other species, for instance 7 for equine β -casein (Girardet et al. 2006), 6 for
21 ovine β -casein (Mamone et al. 2003), and 5 for human β -casein (Poth et al. 2008). Moreover, in
22 goat the β -casein is homologous to bovine β -casein concerning the 5P, but it presents a specific
23 additional phosphorylation site (Neveu et al. 2002), which may indicate a different function of
24 caseins between caprines and bovines. In light of these studies and considering contrasting
25 information on camel β -casein (Kappeler et al. 1998; Saadaoui et al. 2014), further investigations

1 are necessary for elucidating the different phosphorylation level and the number of existing
2 isoforms for this protein.

3 The comparison of the camel β -casein with the list of active bio-peptides available in BIOPEP
4 databases evidenced the presence of at least 3 peptides: PFLQP (ID: 3503), QSLVY (ID: 3426) and
5 ENLHLPLP (ID: 2672) belonging to ACE-Inhibitor family (figure 1). Furthermore, two peptides:
6 KVLVPVQ (ID: 7876) in position 185-191 and VPYPQ (ID: 8278) at 194-198 of the camel β -
7 casein are two contiguous anti-oxidative elements (figure 1), which were already identified for their
8 properties by Rival et al. (2001) and Tsopmo et al. (2011) respectively.

9 Although in the present study these peptides were only putatively indicated, it's likely that they
10 might have a functional activity as already demonstrated in other species.

11

12 ***3.4. Analysis of the CSN2 gene promoter in dromedary and bactrian camel***

13 The analysis of 5' flanking region provides an important contribution for the evaluation of the
14 transcription factors involved in the regulation of the gene expression. Mutations detected in the
15 promoter region have been reported to modify the transcription rates or the mRNA stability and,
16 consequently, the content of a particular protein in the milk (Prinzenberg et al. 2003; Szymanowska
17 et al. 2004; Kuss et al. 2005). We decided to extend the sequencing to the 5'-flanking region for
18 SNP discovery purposes and to analyze the putative regulatory regions of the camel β -casein gene.
19 About 2150 bp of the *CSN2* gene were amplified and sequenced for *Camelus dromedarius* (EMBL
20 ID: HG969421) and *Camelus bactrianus* (EMBL ID: HG977711).

21 Homology level and similar locations of most putative transcription binding sites were already
22 reported among bovine ([X14711](#)), caprine ([AF409096](#)), ovine ([X79703](#)), rabbit ([X15735](#)), human
23 ([AF027807](#)), rat ([M10936](#)), murine ([X13484](#)) and porcine ([AY452035](#)) β -casein promoters
24 (Malewski, 1998; Lee et al. 2008). Similar mechanisms of regulation are supposed for these species,
25 and therefore camels β -casein promoter sequences were compared with the homologous sequences

1 of horse (AY579425S1) promoter - so far compared only to camel sequence cDNA (Lenasi et al.
2 2003) - and bovine, respectively as non-ruminant and ruminant counterpart.

3 A total of 277 high-scoring (85%-100%) putative binding sites were found by TFSEARCH tool.
4 Most of the consensus sequences related to protein and milk production (47 in total) were identified
5 in the proximal promoter region (Supplementary figure 1), which also resulted highly conserved
6 with an average homology of 78.7%. Over the first 500 bp upstream the exon 1, the sequences
7 became more species-specific reaching lower level of similarity (Supplementary figure 1). This
8 result is not surprising and it confirms the comparative analysis of Lenasi et al. (2005) among horse,
9 human, rabbit, pig, cow and mouse β -casein promoters.

10 The analysis of the putative transcription factor binding sites of camel β -casein gene promoter
11 revealed the presence of both consensus sequences for the crucial positive and negative regulators
12 of casein gene expression. In particular, 8 CCAAT enhancer bind protein (C/EBP), 3 Mammary
13 Gland Factor/STAT5 (MGF/STAT5) and 3 Glucocorticoid receptor factor (GR) were found in
14 cluster both the proximal and distal promoter region. These elements are known to be transducers
15 and activators of transcription, often acting in synergy through mutual transcription factor
16 interactions (Lechner et al. 1997; Wyszomierski and Rosen, 2001). The close proximity of STAT5,
17 GR and C/EBP- β between the nucleotides -900/-690 probably represent the camel composite
18 response element (CoRE). This region is in fact characterized also by the presence of two Yin
19 Yang 1 (YY1), which are negative regulatory elements and predominantly participate to the
20 repression of casein gene expression (Raught et al. 1994).

21 Seven Octamer bind protein (Oct-1), two Tata Binding Protein (TBP) and one Nuclear factor-1
22 (NF-1) were found, all potentially involved in the activation of the camel β -casein transcription. For
23 instance, Oct-1 is not known to be a strong transcriptional activator by itself, but in conjunction
24 with other co-activators. Zhao et al. (2002) showed that Oct-1 and STAT5 are both involved in the
25 hormonal induction of casein gene expression. They bind DNA cooperatively and interact in the

1 stabilization of the transcription complex. However, Oct-1 can also interact with the TBP
2 components (Zwilling et al. 1994). Oct-1 and TBP are putative sites very representative of camel β -
3 casein promoter being characterized by 7 Oct-1 and 2 TBP putative sites. Therefore, it would be
4 reasonable to suggest that Oct-1 may also mediate stimulation of β -casein expression in mammary
5 gland through this way.

6 Two transitions were detected in the 5' flanking region of *Camelus dromedarius* within possible
7 trans-acting factor binding sites. In particular, the first SNP (g.1548C>T) creates a new putative
8 consensus for GATA-1 binding site (TFmatrix: M0075, binding affinity: 85.7%), whereas the
9 second SNP (g.2126A>G) falls 3 nucleotide downstream the TATA-box and modifies the putative
10 consensus site for a non-canonical TATA-box (TFmatrix: M00252), changing the binding affinity
11 from 87.7% to 86.2%. The importance of the TATA-box in the eukaryotic RNA polymerase
12 (RNAP) transcriptional machineries was already elucidated in details more than 20 years ago
13 (Sharp, 1992; Hernandez, 1993). Any changing in the consensus sequence results in a different
14 strength of RNA polymerase binding (Wang et al. 1996), and in particular G/C nucleotide
15 substitutions within this site might be highly deleterious for the transcription (Wang et al. 1996).
16 The weaker promoter activity of mutant TATA-boxes were previously reported for other genes like
17 vasopressin and β -globin (Xu et al. 1991; Ho and Murphy, 2002), and recently Lee et al. (2012)
18 described a lower luciferase activity in a mutant TATA-box found in the porcine β -casein promoter.
19 Therefore, in camels the present result opens the way to new investigations for expression analysis
20 studies -so far- never achieved in this species.

21 A short tandem repeat (STR) was found between the nucleotides -1151/-1136. It is a
22 polymorphic microsatellite (TA)_n characterized by 8 and 9 repetitions for the two alleles found in
23 *Camelus dromedarius*, whereas the bactrian camels were monomorph (TA)₈. So far, the total
24 number of microsatellites used for parentage identification in camels were 17 in total (Spencer et al.
25 2010), and the knowledge in this field is still very limited. Therefore, the identification and the

1 characterization of this microsatellite in the promoter region of camel β -casein gene add new
2 information for future parentage identification tests in this species. No further polymorphisms were
3 evidenced by the direct sequencing of dromedary *CSN2* promoter and no genetic variability was
4 detected for the 8 available bactrian camels for the *CSN2* 5'-flanking region. However, the
5 following comparison of the two homologous sequences evidenced a total of 12 SNP (table 3),
6 which might be used as genetic marker to differentiate both dromedary and bactrian species or
7 potentially useful for dairy product traceability. A total of 28 polymorphic sites were instead
8 counted when the comparison was extended to the other available β -casein promoter sequences
9 (table 3).

10 It is evident that the β -casein promoter region in camels resulted more polymorphic than the rest
11 of the *CSN2* gene (on average 11.6 vs 2.4 SNP/1000 bp). The reason for that should be deeply
12 investigated mainly for the potential effects on the gene transcription. For instance, in goat β -casein
13 the A (normal) and 01 (null) alleles differ only for intronic mutations (25 in total) and for two
14 transitions located in the 5' flanking region. In this study, the SNP g.1311T>C in the promoter
15 region is linked to the absence of β -casein in the milk (Cosenza et al. 2007), showing that the
16 differential expression of milk-protein alleles can be also the result of a linkage disequilibrium
17 condition between variants of coding regions and regulatory regions. Therefore, also in camel
18 further investigation is required to verify the influence of the identified polymorphisms in β -casein
19 gene regulation.

20

21 **3.5. Genotyping**

22 The transition g.2126A>G detected in the dromedary camel *CSN2* promoter region removes a
23 restriction site for the endonuclease *HphI*, thus a PCR-RFLP protocol was set up for the quick
24 genotyping of 180 she-camels. The digestion of the PCR product (659 bp) allows the identification
25 of both alleles (figure 2).

1 The restriction pattern is characterized by 2 fragments of 608 bp and 51 for the GG samples,
2 whereas the band 608 bp long is further restricted into two fragments of 352 bp and 256 bp in the
3 presence of the adenine. The restriction pattern of the heterozygous samples shows 4 fragments.
4 The major allele had a relative frequency of about 0.65, with a variation among the ecotypes
5 between 0.59 and 0.72 (table 5). The χ^2 value showed that there was no evidence of departure from
6 the Hardy-Weinberg equilibrium ($P \leq 0.05$). The lower frequency of the G allele is an indication that
7 this genetic variant, responsible for the lower affinity of the TATA box binding site, might be less
8 facilitate in the gene transcription process.

9

10 **3.6. Interspersed elements and microRNA**

11 Short INterspersed Elements (SINEs; Okada and Ohshima, 1995) and Long INterspersed
12 Elements (LINEs; Eickbush, 1992) are genetic elements propagated in the genome through a
13 process known as retroposition. The camel *CSN2* gene sequence is characterised by 5 DNA
14 fragments (boxes from A to E) showing high similarity (on average 80%) to interspersed elements
15 belonging to both groups (figure 3).

16 Three elements (A, B, D) have been found in the promoter, intron 1 and intron 6 respectively
17 and they appear to be LINEs. The B element situated between the nucleotides +3667/+3796 of the
18 intron 1 shows a similarity of 71.5% with a LINE-2 family, whereas the A and D elements belong
19 both to the family L1 and both are characterized by a complementary sequence. In particular, the A
20 element is located in the promoter region approximately 1000 bp upstream the exon1 (nucleotides
21 +925/+1143), it is 219 bp long and it contains a polymorphic microsatellites already described in
22 the section 3.2. Conversely, the second element is located at the intron 6 between the nucleotides
23 +7263/+7364. It is a truncated LINE (102 bp) which appears to be characteristic of the camel *CSN2*
24 sequence, since the bovine, buffalo, goat, sheep and donkey *CSN2* sequences are characterized by
25 the presence of a SINE, whereas pig, human and rabbit β -casein gene do not show any transposable
26 element in that position.

1 The C element is also situated in the intron 6 between the nucleotides +6637/+6736. This is a
2 SINE belonging to the class of mammalian-wide interspersed repeats (MIR/Mon1a7) and it seems
3 to be very well conserved among the species. The same element was in fact found in all afore
4 mentioned species with the exception of the rabbit which does not show repetitive elements and
5 human which is characterized by the presence of a LINE.

6 The last element (E) is located at the intron 7 in position +8399/+8506 and it has approximately
7 77% of similarity with the complementary sequence of hAT-Charlie/MER5B. This family of
8 transposon is widely spread in plant, animal and human genomes. With the exception of the little
9 brown bat (*Myotis lucifugus*) these elements seem to be not active in mammalian genomes
10 (Arensburger et al. 2011). This transposon is not present in ruminants (bovine, buffalo, goat and
11 sheep) and pig *CSN2* sequence, however it characterizes also the human and donkey β -casein gene.

12 The five repetitive elements observed at the camel *CSN2* gene represent the 6.6% of the
13 sequence deposited in the EMBL database. This feature raises to 13.2% in the bovine counterpart
14 which is characterized by similar insertions (B and C elements) and by different transposition
15 events (figure 3). A total of 949 bp corresponding to five extra repetitive elements (G, H, I, J, K)
16 also characterize the bovine *CSN2* sequence but they are not at all present in the camel counterpart
17 (figure 3). The camel *CSN2* is also characterized by several low complexity repeats. All of them
18 belong to AT rich regions and they cover a total of 92 bp, representing approximately the 1.2% of
19 the gene.

20 The evolution of mammalian genome has been widely influenced by transposable elements
21 which have played a prominent role in species diversification (Bowen and Jordan, 2002). The lower
22 number of repetitive element in camel is an indication that *Tylopoda* diverged from *Ruminantia*
23 before additional retrotransposition events took place at different times during the divergence of in
24 such suborder (Nijman et al. 2002). This statement is also confirmed by the data already reported
25 for the camel κ -casein gene (Pauciullo et al. 2013). Furthermore, the distinctive LINE (D element)
26 found in camel *CSN2* sequence and the truncated LINE (C element) found by Pauciullo et al. (2013)

1 belong both to L1MA family and most probably they have been originated by a similar replication
2 event occurred MYA. The presence/absence of these elements and those characterizing the bovine
3 sequence add interest in the knowledge of species evolution and they can be considered powerful
4 phylogenetic markers for clustering studies.

5 17 putative mature sequences for microRNA were found in camel *CSN2* (table 6) which in total
6 show more than 2600 predicted target genes most of which related to immune system. It is
7 interesting to notice that 6 out 17 are located in the intron 3 whose size is only 108 bp and they
8 recognize almost the same sequence, a stretch of DNA of 20 bp between the nucleotides
9 +5082/+5104. The biological functions of most miRNA are unknown, but it is estimated that >30%
10 of protein-coding genes are regulated by miRNA (Lewis et al., 2005). For instance, one of the target
11 genes for miR-190a and miR-2437 is C/EBP- α whose transcription factor binding site is present in
12 the proximal promoter region of camel *CSN2* and therefore these miRNAs might be regulator of the
13 gene expression. Conversely, one of the miR-2345, miR-2361 and miR-2444 target genes are the
14 casein kinase-1 (CSNK1- α 1, alpha-1; CSNK1- γ 1, gamma-1 and CSNK1- δ , delta) whose general
15 activity is the selective phosphorylation of serine/threonine amino acids. As consequence, these
16 miRNAs might have a post-transcriptional influence on β -casein.

17 Furthermore, considering that milk-specific miRNAs are also ideal biomarker for discriminating
18 poor-quality or “manipulated” milk from pure raw milk (Chen et al. 2010), the future investigation
19 in this direction will offer new opportunity for the quality control of commercial milk products also
20 in camels.

21

22 4. CONCLUSION

23 1. The *Camelus dromedarius CSN2* gene shares a similar organization compared to the
24 homologous gene in other species. The exon/intron ratio in camel *CSN2* was slightly higher
25 (1:6.01) than that one observed in bovine (1:6.74).

- 1 2. The different number and nature of the interspersed repeated elements (5 in camel vs 9 in
2 bovine) are responsible for the difference in size of the camel β -casein gene. One LINE
3 truncated is characteristic of the camel *CSN2* sequence. This element could be used as
4 genetic marker typical of the specie, as well as for evolution and clustering studies.
- 5 3. The camel composite response element (CoRE) region was identified and the relative
6 transcription factors (C/EBP MGF/STAT5, GR, NF-1 and YY1) were indicated in the 5'
7 flanking region. The interactions among the CoRE and other representative motifs like Oct-
8 1 and TBP suggests a possible way of regulation of the gene expression.
- 9 4. A total of 47 polymorphic sites have been described in the promoter as well as in the gene
10 and they represent a pre-requisite for the regulation of the β -casein gene expression. The
11 g.2126A>G SNP detected in dromedary *CSN2* promoter falls within the TATA-box binding
12 site. The transversion g.4175C>A represents the first synonymous allele for dromedary
13 camel *CSN2*. Conversely, the transition c.666G>A at the exon 7 of the bactrian *CSN2* cDNA
14 is responsible for the Met→Ile amino acid change at the position 201 of the mature protein.
15 It represents in absolute the first missense mutation at the camel β -casein and it identifies 2
16 variants named A and B.
- 17 5. The presence of putative bio-functional peptides belonging to ACE-inhibitor and anti-
18 oxidative families suggests a potential protective role of the camel milk for the human
19 nutrition, whereas the identification of microRNA sequences offers also in camels bio-
20 markers for the quality control of commercial milk products.

22 **5. ACKNOWLEDGEMENTS**

23 The authors thank Dr. El Tahir Salih Shuiep and the camel owners in Sudan for providing
24 *Camelus dromedarius* samples; Dr. Tobias Knauf-Witzens from Wilhelma Stuttgart Zoo and Dr.
25 Nicole Schauerte from Frankfurt Zoo (Germany) for providing *Camelus bactrianus* samples. The

1 authors would like also to thank Therese Kubetzki-Bauer, Anja Scheuermann and Zhaoxin Wang
2 for the excellent technical assistance.

3

4 **6. REFERENCES**

5 Al haj, O.A. and Al Kanhal, H.A., 2011. Compositional, technological and nutritional aspects of
6 dromedary camel milk. *Int. Dairy J.* 20, 811-821.

7

8 Amigo, L., Recio, I., Ramos, M., 2000. Genetic polymorphism of ovine milk proteins: its influence
9 on technological properties of milk - a review. *Int. Dairy J.* 10, 135-149.

10

11 Arensburger, P., Hice, R.H., Zhou, L., Smith, R.C., Tom, A.C., Wright, J.A., Knapp, J., O'Brochta,
12 D., Craig, N.L., Atkinson, P.W., 2011. Phylogenetic and functional characterization of the hAT
13 transposon superfamily. *Genetics* 188, 45-57.

14

15 Bowen, N.J. and Jordan, I.K., 2002. Transposable elements and the evolution of eukaryotic
16 complexity. *Curr. Issues Mol. Biol.* 4, 65-76.

17

18 Caroli, A., Chiatti, F., Chessa, S., Rignanese, D, Bolla P., Pagnacco G., 2006. Focusing on the goat
19 casein complex. *J. Dairy Sci.* 89, 3178-3187.

20

21 Caroli, A.M., Chessa, S., Erhardt, G.J., 2009. Milk protein polymorphisms in cattle: effect on
22 animal breeding and human nutrition. *J. Dairy Sci.* 92, 5335-5352.

23

24 Ceriotti, G., Chessa, S., Bolla, P., Budelli, E., Bianchi, L., Duranti, E., Caroli, A., 2004. Single
25 nucleotide polymorphisms in the ovine casein genes detected by Polymerase Chain Reaction-Single
26 Strand Conformation Polymorphism. *J. Dairy Sci.* 87, 2606-2613.

1

2 Chen, X., Gao, C., Li, H., Huang, L., Sun, Q., Dong, Y., Tian, C., Gao, S., Dong, H., Guan, D., Hu,
3 X., Zhao, S., Li, L., Zhu, L., Yan, Q., Zhang, J., Zen, K., Zhang, C.Y., 2010. Identification and
4 characterization of microRNAs in raw milk during different periods of lactation, commercial fluid,
5 and powdered milk products. *Cell Res.* 20, 1128-1137.

6

7 Chianese, L., 1997. The casein variants of ovine milk and the relationships between the α s1-casein
8 variants and milk composition, micellar size and cheese yield. In: *Caseins and caseinates:*
9 *structures, interactions, networks. Hannah Symposium.* Scotland, United Kingdom.

10

11 Chessa, S., Rignanese, D., Berbenni, M., Ceriotti, G., Martini, M., Pagnacco, G., Caroli, A., 2010.
12 New genetic polymorphisms within ovine β - and α s2-caseins. *Small Ruminant Res.* 88, 84-88.

13

14 Cieslak, J., Mankowska, M., Switonski, M., 2012. Between-breed variation in frequency of five
15 novel missense SNPs in porcine casein beta (*CSN2*) and casein kappa (*CSN3*) genes. *Anim. Genet.*
16 43, 363.

17

18 Cosenza, G., Pauciullo, A., Gallo, D., Di Berardino, D., Ramunno, L., 2005. A *SspI* PCR-RFLP
19 detecting a silent allele at the goat *CSN2* locus. *J. Dairy Res.* 72, 456-459.

20

21 Cosenza, G., Pauciullo, A., Colimoro, L., Mancusi, A., Rando, A., Di Berardino, D., Ramunno, L.,
22 2007. An SNP in the goat *CSN2* promoter region is associated with the absence of β -casein in milk.
23 *Anim. Genet.* 38, 655-658.

24

1 Cosenza, G., Pauciullo, A., Coletta, A. Di Francia, A., Feligini, M., Gallo, D., Di Bernardino, D.,
2 Ramunno, L., 2011. Translational efficiency of casein transcripts in Mediterranean river buffalo. J.
3 Dairy Sci. 94, 5691-5694.
4
5 Di Stasio, L. and Mariani, P., 2000. The role of protein polymorphism in the genetic improvement
6 of milk production. Zoot. Nutr. Anim. 26, 69-90.
7
8 Eickbush, T.H., 1992. Transposing without ends: the non-LTR retrotransposable elements. New
9 Biol. 4, 430-440.
10
11 El-Agamy, E.I., 2006. Camel milk. In: *Hand book of milk of non-bovine mammals*. Park, Haenlan
12 (Eds.), Blackwell, USA.
13
14 El-Agamy, E.I., Nawar, M., Shamsia, S.M., Awad, S., Haenlein, G.F.W., 2009. Are camel milk
15 proteins convenient to the nutrition of cow milk allergic children. Small Ruminant Res. 82, 1-6.
16
17 Farah, Z. and Farah-Riesen, M., 1985. Separation and characterization of major components of
18 camel milk casein. Milchwissenschaft. 40, 669-671.
19
20 Farah, Z., Mollet, M., Younan, M., Dahir, R., 2007. Camel dairy in Somalia: limiting factors and
21 development potential. Livest. Sci. 110, 187-191.
22
23 Farrell, H.M.Jr., Jimenez-Flores, R., Bleck, G.T., Brown, E.M., Butler, J.E., Creamer, L.K., Hicks
24 C.L., Hollar, C.M., Ng-Kwai-Hang, K.F., Swaisgood H.E., 2004. Nomenclature of the proteins of
25 cows' milk-Sixth revision. J. Dairy Sci. 87, 1641-1674.
26

1 Faye, B. and Konuspayeva, G., 2012. The sustainability challenge to the dairy sector e The growing
2 importance of non-cattle milk production worldwide. *Int. Dairy J.* 24, 50-56.
3
4 Faye, B., Grech, S., and Korchani, T., 2002. Le dromadaire, entre féralisation et intensification.
5 *Anthropozoologica.* 39, 7-13.
6
7 Girardet, J.M., Miclo, L., Florent, S., Mollé, D., Gaillard, J.L., 2006. Determination of the
8 phosphorylation level and deamidation susceptibility of equine b-casein. *Proteomics.* 6, 3707-3717.
9
10 Hermas, S.A., 1998. Genetic improvement and the future role of the camel in the Arab World:
11 problems and opportunities. In: *Proceedings of the Third Annual Meeting for Animal Production*
12 *Under Arid Conditions*, vol. 2, pp. 56–68.
13
14 Hernandez, N., 1993. TBP, a universal eukaryotic transcription factor? *Genes Dev.* 7, 1291-1308.
15
16 Hinz, K., O'Connor, P.M., Huppertz, T., Ross, R.P., Kelly, A.L., 2012. Comparison of the principal
17 proteins in bovine, caprine, buffalo, equine and camel milk. *J. Dairy Res.* 79, 185-191.
18
19 Ho, M.Y. and Murphy, D., 2002. The vasopressin gene non-canonical Hogness box: effect on
20 protein binding and promoter function. *Mol. Cell. Endocrinol.* 186, 17-25.
21
22 Holland, J.W., 2009. Post-translational modifications of caseins. In: *Milk protein from expression to*
23 *food*. Eds. Abby Thompson, Amsterdam, Netherlands 107-132.
24
25 Kaminski, S., Cieslinska, A., Kostyra, E., 2007. Polymorphism of bovine beta-casein and its
26 potential effect on human health. *J. Appl. Genet.* 48, 189-198.

1

2 Kappeler, S., Farah, Z., Puhan, Z., 1998. Sequence analysis of *Camelus dromedarius* milk caseins.
3 J. Dairy Res. 65, 209-222.

4

5 Kappeler, S.R., Farah, Z., Puhan, Z., 2003. 5'-Flanking regions of camel milk genes are highly
6 similar to homologue regions of other species and can be divided into two distinct groups. J. Dairy
7 Sci. 86, 498-508.

8

9 Khaldi, N. and Shields, D., 2011. Shift in the isoelectric-point of milk proteins as a consequence of
10 adaptive divergence between the milks of mammalian species. Biol. Direct. 6, 40.

11

12 Kozak, M., 1987. An analysis of 5'-noncoding sequences from 699 vertebrate messenger RNAs.
13 Nucleic Acids Res. 15, 8125-8148.

14

15 Kozak, M., 1991. Structural features in eukaryotic mRNAs that modulate the initiation of
16 translation. J. Biol. Chem. 266, 19867-19870.

17

18 Kuss, A.W., Gogol, J., Bartenschlager, H., Geldermann, H., 2005. Polymorphic AP-1 binding site
19 in bovine *CSN1S1* shows quantitative differences in protein binding associated with milk protein
20 expression. J. Dairy Sci. 88, 2246-2252.

21

22 Lechner, J., Welte, T., Tomasi, J.K., Bruno, P., Cairns, C., Gustafsson, J.A., Doppler, W., 1997.
23 Promoter-dependent synergy between glucocorticoid receptor and Stat5 in the activation of β -casein
24 gene transcription. J. Biol. Chem. 272, 20954-20960.

25

1 Lee, P., Chung, H.K., Lee, H.G., Lee, H.C., Woo, J.S., Lee, S., Jo, S.J., Chang, W.K., Lee, H.T.,
2 Kwon, M., Park, J.K., 2008. Cloning and characterization of 5'-untranslated region of porcine beta
3 casein gene (CSN2). *Domest. Anim. Endocrinol.* 35, 245-253.
4
5 Lee, S.M., Kim, H.M., Moon, S.J., Kang, M.J., 2012. Cloning and molecular characterization of
6 porcine β -casein gene (CSN2). *Asian-Aust. J. Anim. Sci.* 25, 421-427.
7
8 Lenasi, T., Rogelj, I., Dovc, P., 2003. Characterization of equine cDNA sequences for α s1-, β - and
9 κ -casein. *J. Dairy Res.* 70, 29-36.
10
11 Lenasi, T., Kokalj-Vokac, N., Narat, M., Baldi, A., Dovc, P., 2005. Functional study of the equine
12 β -casein and κ -casein gene promoters. *J. Dairy Res.* 72, 34-43.
13
14 Lewis, B.P., Burge, C.B., Bartel, D.P., 2005. Conserved seed pairing, often flanked by adenosines,
15 indicates that thousands of human genes are microRNA targets. *Cell.* 120, 15-20.
16
17 Li, S., Wang, J., Wie, H., Yang, Y., Bu, D., Zhang, L., Zhou, L. 2012. Identification of bovine casein
18 phosphorylation using titanium dioxide enrichment in combination with nano electrospray
19 ionization tandem mass spectrometry. *J. Integr. Agr.* 11, 439-445.
20
21 Malewski T., 1998. Computer analysis of distribution of putative cis- and trans regulatory elements
22 in milk protein gene promoters. *BioSystems* 45, 29-44.
23

1 Mamone, G., Caira, S., Garro, G., Nicolai, A., Ferranti, P., Picariello, G., Malorni, A., Chianese, L.,
2 Addeo, F., 2003. Casein phosphoproteome: identification of phosphoproteins by combined mass
3 spectrometry and two-dimensional gel electrophoresis. *Electrophoresis* 24, 2824-2837
4
5 Miclo, L., Girardet, J.M., Egito, A.S., Mollé, D., Martin, P. Gaillard, J.L., 2007. The primary
6 structure of a low-M_r multiphosphorylated variant of β -casein in equine milk. *Proteomics* 7, 1327-
7 1335.
8
9 Nagy, P., Thomas, S., Marko, O., Juhasz, J., 2013. Milk production, raw milk quality and fertility of
10 dromedary camels (*Camelus dromedarius*) under intensive management. *Acta Vet. Hung.* 61, 71-84
11
12 Neveu, C., Molle, D., Moreno, J., Martin, P., Leonil, J., 2002. Heterogeneity of caprine beta-casein
13 elucidated by RP-HPLC/MS: genetic variants and phosphorylations. *J Protein Chem.* 21, 557-567.
14
15 Nijman, I.J., van Tessel, P., Lenstra, J.A., 2002. SINE retrotransposition during the evolution of the
16 Pecoran ruminants. *J. Mol. Evol.* 54, 9-16.
17
18 Okada, N. and Ohshima, K., 1995. Evolution of tRNA-derived SINEs, In: *Maraia, R.J., The impact*
19 *of short interspersed elements (SINEs) on the host genome*, Landes R.G. (Eds.), Austin (USA), pp.
20 61-79.
21
22 Pauciullo, A., Shuiep, E.S., Cosenza, G., Ramunno, L., Erhardt, G. 2013. Molecular
23 characterization and genetic variability at κ -casein gene (*CSN3*) in camels. *Gene* 513, 22-30.
24

1 Poth, A.G., Deeth, H.C., Alewood, P.F., Holland, J.W., 2008. Analysis of the human casein
2 phosphoproteome by 2-D electrophoresis and MALDI-TOF/TOF MS reveals new phosphoforms. *J.*
3 *Proteome Res.* 7, 5017-5027.

4

5 Prinzenberg, E.M., Weimann, C., Brandt, H., Bennewitz, J., Kalm, E., Schwerin, M., Erhardt, G.,
6 2003. Polymorphism of the bovine *CSN1S1* promoter: linkage mapping, intragenic haplotypes, and
7 effects on milk production traits. *J. Dairy Sci.* 86, 2696-2705.

8

9 Raught, B., Khursheed, B., Kazansky, A., Rosen, J., 1994. YY1 represses β -casein gene expression
10 by preventing the formation of a lactation-associated complex. *Mol. Cell. Biol.* 14, 1752–1763.

11

12 Rival, S.G., Fornaroli, S., Boeriu, C.G., Wichers, H.J., 2001. Caseins and casein hydrolysates. 1.
13 Lipxygenase inhibitory properties. *J. Agric. Food Chem.* 49, 287-294.

14

15 Saadaoui, B., Bianchi, L., Henry, C., Miranda, G., Martin, P., Cebo, C., 2014. Combining
16 proteomic tools to characterize the protein fraction of llama (*Lama glama*) milk. *Electrophoresis.*
17 35, 1406-1418.

18

19 Sambrook, J., Fritsch, E.F., Maniatis, T., 1989. *Molecular Cloning*, Cold Spring Harbor (Eds.), New
20 York (USA).

21

22 Sharp, P.A., 1992. TATA-binding protein is a classless factor. *Cell.* 68, 819-821.

23

24 Shuiep, E.S., Giambra, I.J., Yas Mohamed El Zubeir, I.E., Erhardt, G., 2013. Biochemical and
25 molecular characterization of polymorphisms of α s1-casein in Sudanese camel (*Camelus*
26 *dromedarius*) milk. *Int. Dairy J.* 28, 88-93.

1
2 Spencer, P.B.S., Wilson, K.J., Tinson, A., 2010. Parentage testing of racing camels (*Camelus*
3 *dromedarius*) using microsatellite DNA typing. *Anim. Genet.* 41, 662-665.
4
5 Suteu, M., Vlaic, A., Balteanu V.A., Wavreille, J., Renaville, R., 2012. Evidence of alternative
6 splicing of porcine β -casein (*CSN2*). *Anim. Genet.* 43, 474-475.
7
8 Szymanowska, M., Malewski, T., Zwierzchowski, L., 2004. Transcription factor binding to variable
9 nucleotide sequences in 5' flanking regions of bovine casein genes. *Int. Dairy J.* 14, 103-115.
10
11 Tsopmo, A., Romanowski, A., Banda, L., Lavoie, J.C., Jenssen, H., 2011. Novel anti-oxidative
12 peptides from enzymatic digestion of human milk. *Food Chem.* 126, 1138-1143.
13
14 Wang, Y., Jensen, R.C., Stumph, W:E., 1996. Role of TATA box sequence and orientation in
15 determining RNA polymerase II/III transcription specificity. *Nucleic Acids Res.* 24, 3100-3106
16
17 Ward, T.J., Honeycutt, R.L., Derr, J.N., 1997. Nucleotide sequence evolution at the k-casein locus:
18 evidence for positive selection within the family bovidae. *Genetics* 147, 1863-1872.
19
20 Wyszomierski, S.L., Rosen, J.M., 2001. Cooperative effects of STAT5 (signal transducer and
21 activator of transcription 5) and C/EBP beta (CCAAT/ enhancer-binding protein-beta) on beta-
22 casein gene transcription are mediated by the glucocorticoid receptor. *Mol. Endocrinol.* 15, 228-
23 240.
24
25 Xu, L.C., Thali, M., Schaffner, W., 1991. Upstream box/TATA box order is the major determinant
26 of the direction of transcription. *Nucleic Acids Res.* 19, 6699-6704.

1

2 Zhao, F.Q., Adachi, K., Oka, K., 2002. Involvement of Oct-1 in transcriptional regulation of β -
3 casein gene expression in mouse mammary gland. *Biochim. Biophys. Acta.* 1577, 27-37.

4

5 Zwilling, S., Annweiler, A., Wirth, T., 1994. The POU domains of the Oct-1 and Oct-2
6 transcription factors mediate specific interaction with TBP. *Nucleic Acids Res.* 22, 1655-1662.

7

Amplicon	Region covered	Primers	Sequence	Ta °C	Size
1	5' UTR	Cd CSN2 5' F*	5'-ATTCACATGGTTTCAACTATTAG-3'	60.0	1263 bp
		Cd CSN2 5' R*	5'-GGTCTTGCCCAAGTATTA-3'		
2	5' UTR	Cd CSN2 prom F*	5'-TTGTAAAAATAAGTAGTGT-3'	51.5	1093 bp
		Cd CSN2 prom R*	5'-TAATTCACAATCTGATAA-3'		
3	5' UTR/Ex1/Intr1	Cd CSN2 prom F2*	5'-GTTTCTCCATTACAGCATC-3'	59.5	659 bp
		Cd CSN2 Intr1 R [†]	5'-TCAAATCTATACAGGCACTT-3'		
4	Ex1/Intr1/Ex2	Cd CSN2 Ex1F	5'-TTCACCTTCTTTTCCTCCAC-3'	61.5	2015 bp
		Cd CSN2 Ex2 R	5'-AGGCAAGGATGAGGAC-3'		
5	Ex2/Intr2/Ex3/Intr3/Ex4	Cd CSN2 Ex2 F	5'-GTCCTCATCCTTGCCT-3'	61.0	974 bp
		Cd CSN2 Ex4 R	5'-TCACTGCTTGAAATGCTTTC-3'		
6	Ex3/Intr3/Ex4/Intr4/Ex5/Intr5/Ex6	Cd CSN2 Ex3 F	5'-GAAGAATTCAAGACAGCCG-3'	61.5	1322 bp
		Cd CSN2 Ex6 R	5'-CTGTTTGCTGCTGTTCC-3'		
7	Ex5/Intr5/Ex6/Intr6/Ex7	Cd CSN2 Ex5 F	5'-GAATCTATTACACACATCA -3'	59.5	2005 bp
		Cd CSN2 Ex7 R	5'-AACAGGCAGGACTTT -3'		
8	Ex7/Intr7/Ex8	Cd CSN2 Ex7 F	5'-GATGAACAGCAGGATAAAAATC-3'	63.0	1159 bp
		Cd CSN2 Ex8 R	5'-AAAAGTGAGGAGGAGGTA-3'		
9	Ex8/Intr8/Ex9	Cd CSN2 Ex8 F	5'-GCCTAAGAAGATTTCAAAGT-3'	62.5	1081 bp
		Cd CSN2 Ex9 R	5'-TAGTTTATTGAAGTGACTGGT-3'		
10	Intr8/Ex9/3'UTR	Cd CSN2 Intr8 F [†]	5'-AGATAAGGCACAACATGG-3'	56.0	476 bp
		Ea CSN2 3end R	5'-TATAATAAAGACAGGAAAACCTAAA-3'		

Table 1. Primer sequences, annealing temperature and amplicon size used for the characterization of *Camelus dromedarius* CSN2. Asterisks refer to primers designed on camel CSN2 promoter sequence (EMBL acc. No. [AJ409279](#)); the crosses indicate primers designed on sequences obtained from this study, whereas the other primers were designed on cDNA sequence (EMBL acc. No. [AJ012630](#)). The last primer was designed on *Equus asinus* CSN2 sequence (EMBL acc. No. [FN598778](#)) after multi alignment with the homologous gene sequences of cow (EMBL acc. No. [X14711](#)), buffalo (EMBL acc. No. [FN424088](#)), goat (EMBL acc. No. [AJ011018](#)) and pig (EMBL acc. No. [EU025876](#)).

Species	EMBL ID	Position									
		-6	-5	-4	-3	-2	-1	+1	+2	+3	+4
Kozak consensus sequence		G	C	C	R	C	C	<u>A</u>	<u>U</u>	<u>G</u>	G
<i>C. dromedarius present work</i>	HG969421	A	G	A	-	T	-	-	-	-	A
<i>B. taurus</i>	X14711	A	G	A	-	-	-	-	-	-	A
<i>B. bubalis</i>	FN424088	A	G	A	-	-	-	-	-	-	A
<i>C. hircus</i>	AJ011018	A	G	A	-	-	-	-	-	-	A
<i>O. aries</i>	X79703	A	G	A	-	-	-	-	-	-	A
<i>E. caballus</i>	AF214526	T	-	A	-	-	-	-	-	-	A
<i>E. asinus</i>	FN598778	T	-	A	-	-	-	-	-	-	A
<i>S. scrofa</i>	EU025876	A	T	-	-	-	-	-	-	-	A
<i>M. musculus</i>	X13484	A	-	A	-	-	-	-	-	-	A
<i>O. coniculus</i>	M33582	-	-	A	-	T	-	-	-	-	A
<i>C. lupus familiaris</i>	AB035080	A	T	A	-	-	-	-	-	-	A
<i>H. sapiens sapiens</i>	AF027807	-	T	A	-	-	-	-	-	-	A

Table 2. Comparison of start codon flanking sequences of the *CSN2* transcript for 12 species. In the first line is reported the optimal context for initiation of translation in mammals (Kozak consensus sequence). The start codon (AUG) is underlined. Conserved nucleotides are reported as dash.

Position	Nucleotide	A	B	C	D
	110	G	G	*	A
	159	A	A	*	G
	380	G	G	A	G
	382	G	G	T	G
	483	T	T	C	T
	567	T	G	T	T
	635	A	A	G	A
	695	G	G	G	A
	697	A	A	A	G
	746	C	C	C	T
	749-750	-	C	-	-
	755-756	-	C	-	-
	775-776	-	C	-	-
promoter	817-818	-	C	-	-
	875	A	A	A	G
	903	C	C	C	T
	947-948	-	C	-	-
	987	G	C	G	G
	991	T	C	T	T
	1203	G	G	G	T
	1548	Y	T	A	T
	1566	T	T	T	C
	1721	C	C	T	C
	1875	A	A	G	A
	1887	T	T	T	G
	2013-2014	-	T	-	-
	2126	R	A	A	G
Exon 1 (5'UTR)	2177	C	C	A	C
	2447	A	-	*	*
	2453	T	C	*	*
Intron 1	2460	A	-	*	*
	2465	C	T	*	*
	2467	A	T	*	*
	4109	K	*	*	*
Exon 2	4175	M ^{Ala/Ala}	*	*	*
	4213	W	*	*	*
Intron 2	4321	K	*	*	*
	4931	Y	*	*	*
Intron 4	5901	R	*	*	*
Intron 6	6313-6314	YR	*	*	*
Intron 7	8312	R	*	*	*

A and D: Present study; B: Kappeller et al. (2003); C: EMBL ID: [AM259943](#)

Table 3. Polymorphisms detected by the comparison among *Camelus dromedarius* CSN2 gene sequence of the present study (A) with: the CSN2 promoter sequence reported by Kappeller et al. (2003) (B); the 5' flanking region submitted under EMBL acc. N. [AM259943](#) (C); the *Camelus bactrianus* CSN2 promoter sequences of the present study (D). Grey cells correspond only to mutations detected in our investigated samples (Y=C/T, R=A/G, K=G/T, M=A/C, W=A/T). In

italics the 12 nucleotide differences between the sequences A and D. Dashes indicate deleted nucleotides, asterisks show unavailable sequences.

Position	Nucleotide genomic DNA	Nucleotide cDNA	A	B	C
Exon 1	2186	45	A	A	-
(5'UTR)	2190-2191	49-50	AA	AA	TT
Exon 2	4175	84	M ^{Ala/Ala}	C ^{Ala}	C ^{Ala}
	4180	89	T ^{Leu}	T ^{Leu}	G ^{Arg}
Exon 7	8102	591	G ^{Gln}	R ^{Gln/Gln}	G ^{Gln}
	8177	666	G ^{Met}	R ^{Met/Ile*}	G ^{Met}
	8243	732	G ^{Val}	R ^{Val/Val}	G ^{Val}
Exon 9 (3'UTR)	9745	900	T	C	T
	9778	933	C	Y	C

Table 4. Comparative analysis of the dromedary (A) and bactrian (B) *CSN2* cDNA sequences of the present work with the cDNA sequence (C) reported by Kappeler et al. (1998). Grey cells correspond to polymorphic sites detected in our investigated samples (M=A/C, R=A/G, Y=C/T). The dash indicates a deleted nucleotide, whereas the asterisk shows the first non conservative allele detected for the bactrian camel β -casein.

Ecotypes	Observed Genotypes				Allelic frequency	
	AA	AG	GG	Total	A	G
<i>Shanbali</i>	11	19	2	32	0.64	0.36
<i>Kahli</i>	24	20	3	47	0.72	0.28
<i>Arabi</i>	13	12	2	27	0.70	0.30
<i>Lahaoi</i>	23	41	10	74	0.59	0.41
Total	71	92	17	180	0.65	0.35

$\chi^2=2.737$ - $p \leq 0,05$ - d.o.f.=1

Table 5. Genotyping data and allele frequency of the g.2126A>G SNP at the *CSN2* promoter gene in Sudanese *Camelus dromedarius* population.

Accession number	ID	Location	Position	Sequence	Strand
MIMAT0009251	bta-miR-190a	Intron 1	+2753/+2771	5'-TAATATTTCAAAAATATAA-3'	-
MIMAT0011880	bta-miR-2345		+4081/+4096	5'-GAAATAAGAAAATACA-3'	-
MIMAT0011794	bta-miR-2287	Intron 2	+4418/+4431	5'-CTGCCACCAGTCCC-3'	-
MIMAT0011805	bta-miR-2297		+4486/+4504	5'-AAGAACAACAAAAAAGAGT-3'	-
MIMAT0012011	bta-miR-2437	Intron 3	+5086/+5104	5'-TGATTTTTTTTTTCCTTT-3'	+
MIMAT0011946	bta-miR-2393		+5085/+5104	5'-TTGATTTTTTTTTTCCTTT-3'	+
MIMAT0011899	bta-miR-2361		+5082/+5098	5'-GTGTTGATTTTTTTTTT-3'	+
MIMAT0011934	bta-miR-2325c		+5087/+5104	5'-GATTTTTTTTTTCCTTT-3'	+
MIMAT0011897	bta-miR-2359		+5082/+5097	5'-GTGTTGATTTTTTTTTT-3'	+
MIMAT0024574	bta-miR-3613		+5083/+5103	5'-TGTTGATTTTTTTTTTCCTT-3'	+
MIMAT0009947	bta-miR-376c	Intron 6	+6468/+6488	5'-TAAACATATAATCAATTTCCA-3'	-
MIMAT0009237	bta-miR-147		+6569/+6584	5'-TAGTAGAAACATTTCC-3'	-
MIMAT0025545	bta-miR-411c-5p		+7394/+7413	5'-ATCTATGATCTATTATCAAC-3'	-
MIMAT0024594	bta-miR-6122-5p		+7467/+7485	5'-GCACAGGAAAAGTAAAAC-3'	-
MIMAT0011832	bta-miR-2284t-3p		+7500/+7519	5'-CCAAACATTAATTCCACTTT-3'	-
MIMAT0003830	bta-miR-361	Intron 7	+8476/+8495	5'-TACCCACGGAGATTCTGATA-3'	-
MIMAT0003797	bta-miR-215	Intron 8	+9326/+9346	5'-TGTCTTTCATTTCACAAGTCA-3'	-

Table 6. Mature sequences of putative microRNA with possible involvement in the regulation of milk protein- and immune-related gene expression found in camel *CSN2* by the comparison with bovine database. Accession number is related to miRBase database. Numbering is relative to the first nucleotide (+ 1) of the *C. dromedarius* β -casein gene and promoter sequence (EMBL ID: HG969421). + and – indicate the normal (5'→3') or the complementary strand of the homologous microRNA sequence identified on the camel *CSN2*.

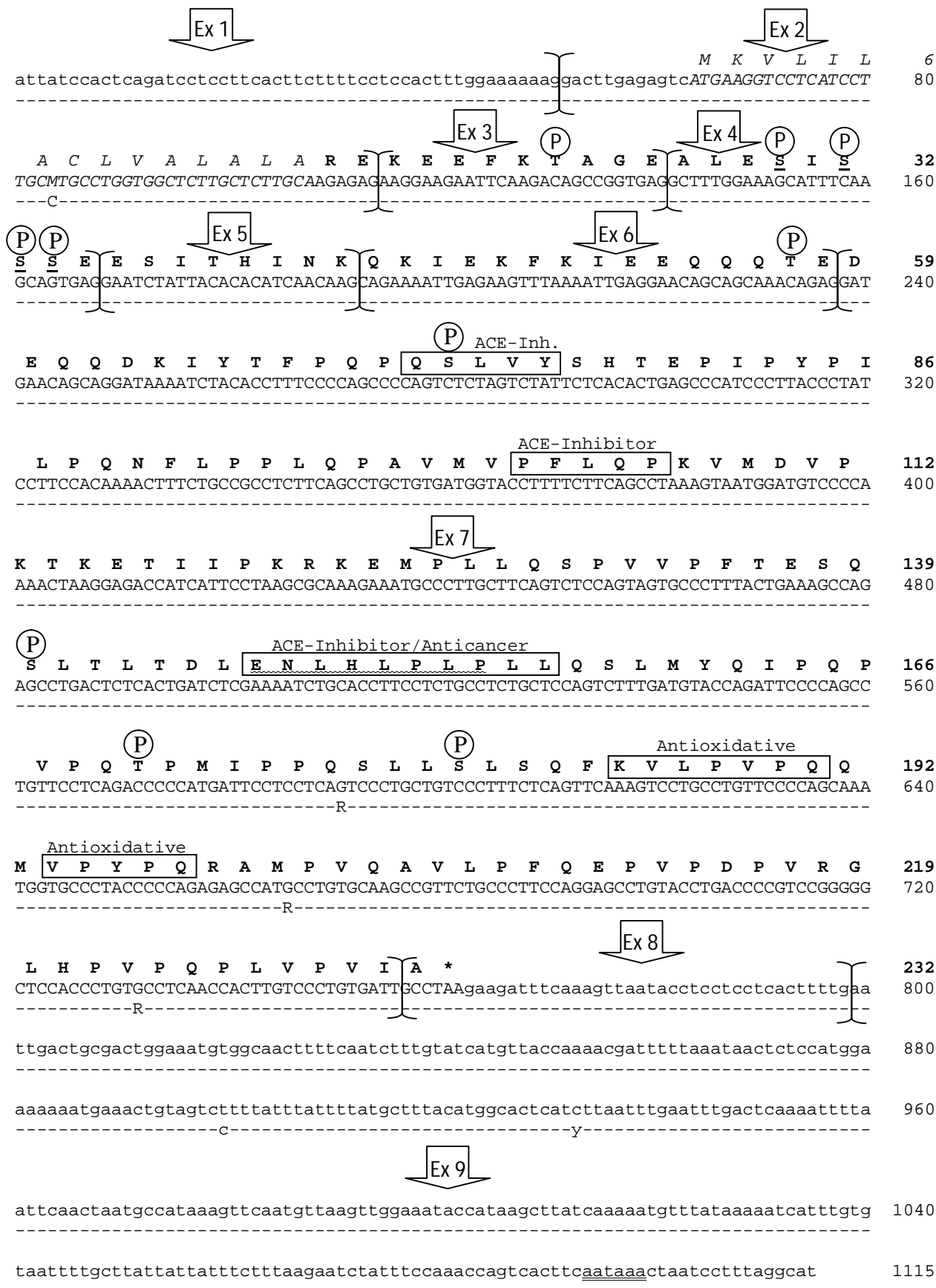


Fig 1. Complete cDNA sequence and exon subdivision of *C. dromedarius* CSN2 (upper line) and comparative alignment with the homologous β -casein cDNA of *C. bactrianus*. Dashes represent identical nucleotides to those in upper lines. In lower cases the 5'- and 3'- Un-Translated Regions (UTR), the polyadenylation signal is double underlined. The corresponding mature protein is reported in bold, whereas the signal peptide is in italics and asterisk represents the termination stop

codon. Putative phosphorylation sites are indicated with P. Phosphorylated serines reported by Kappeler et al. (1998) are underlined. Boxes represent potential active bio-peptides. Polymorphic sites within the investigated samples are indicated with R=A/G and Y=C/T.

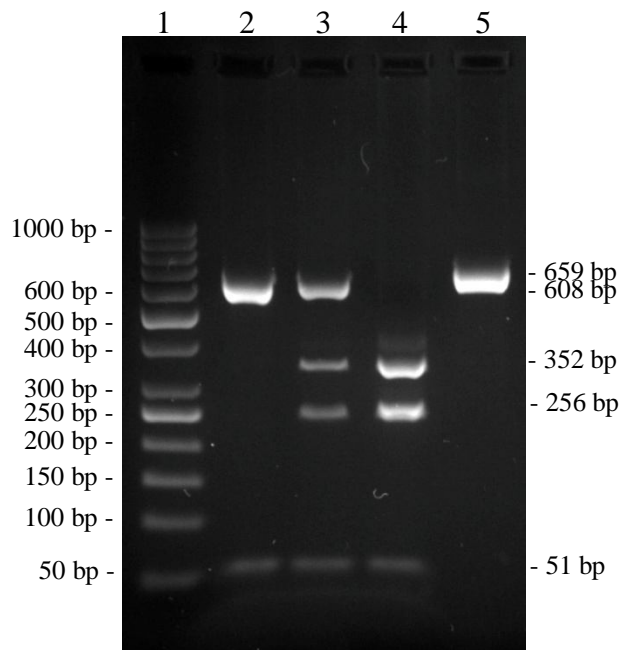


Fig. 2. Genotyping of *C. dromedarius* CSN2 g.2126A>G SNP by *Hph* I (5'...GGTGAN8↓...3') PCR-RFLP. Line 2: AA homozygous samples; line 4: GG homozygous samples; line 3: heterozygous samples; line 5: undigested sample. Line 1 is GeneRuler™ 50bp DNA ladder (Thermo Scientific).

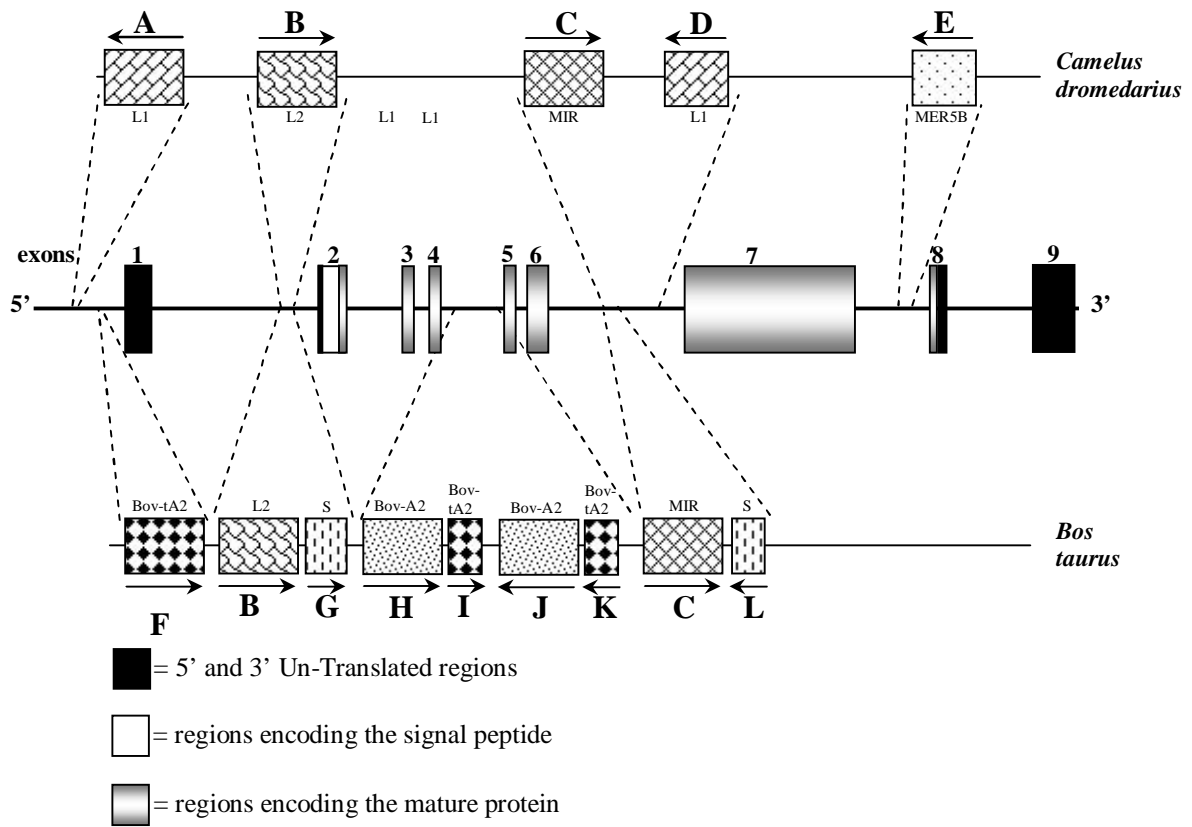


Fig 3. Schematic representation of the *CSN2* gene and of the interspersed elements observed in dromedary camel and cattle. Common interspersed elements are indicated with the same letter.