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

Pauciullo A.¹,²,* Giambra I.J.¹, Iannuzzi L.², Erhardt G.¹

¹Institute for Animal Breeding and Genetics, Justus Liebig University, Ludwigstraße 21 B, 35390 Gießen, Germany

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

*For correspondence: Alfredo Pauciullo, Ph.D
email: alfredo.pauciullo@agraro.uni-giessen.de; alfredo.pauciullo@cnr.it

Institute for Animal Breeding and Genetics, Justus Liebig University, Ludwigstraße 21 B, 35390 Gießen, Germany
Phone: +4906419937672
ABSTRACT

The β-casein is the most abundant protein in camel milk and its encoding gene (CSN2) is considered in other species a ‘major’ gene for the presence of alleles associated to different level of expression. We report for the first time the characterization of the nucleotide sequence of the whole β-casein-encoding gene (CSN2) plus 2141 bp at the 5’-flanking region in *Camelus dromedaries* and in addition the promoter region and the complete cDNA in *Camelus bactrianus*. The gene is spread over 7.8 kb and consists of 9 exons varying in length from 24 bp (exon 5) to 519 bp (exon 7), and 8 introns from 95 bp (intron 5) to 1950 bp (intron 1). The composite response element (CoRE) region was identified in the promoter, whereas the presence of mature microRNAs sequences improves the knowledge on the factors putatively involved in the gene regulation. A total of 46 polymorphic sites have been detected. The transition g.2126A>G falls within the TATA-box of dromedary CSN2 promoter with a putative influence on the transcription factor binding activity. The frequency of the G allele was 0.35 in a population of 180 she-camels belonging to 4 different ecotypes. In the same population, a conservative SNP (g.4175C>A) was found at the codon 7 of the signal peptide, whereas the comparative analysis with the cDNA available in EMBL evidenced a missense SNP (g.4180T>Leu>G>Arg) at the exon 2. Four SNPs were found in the bactrian camel. The SNP c.666G>A is responsible for the amino acid change Met$^{201}$→Ile and it represents the first missense allele at the β-casein in camels. Finally, five interspersed repeated elements were identified at the intronic level and putative bio-functional peptides belonging to ACE-inhibitor and anti-oxidative families were identified using in silico analysis.

Keywords: Camel, Casein, CSN2, Promoter analysis, Interspersed elements, bio-peptides, microRNA.
Abbreviations: CSN2, β-casein-encoding gene; SNP, single nucleotide polymorphism; CN, casein: cDNA, DNA complementary to RNA; EDTA, ethylene diamine tetra acetic acid; SDS, sodium dodecyl sulfate; PCR, polymerase chain reaction; MgCl₂, magnesium chloride; dNTP, deoxyribonucleoside triphosphate; TBE, tris boric acid EDTA; RFLP, restriction fragment length polymorphism; pI, isoelectric point; CoRE, composite response element; C/EBP, CCAAT enhancer bind protein; MGF, Mammary Gland Factor; GR, glucocorticoid receptor factor; YY1, Ying Yang 1; Oct-1, Octamer bind protein; TBP, Tata Binding Protein; NF-1, Nuclear factor-1; STR, short tandem repeat; SINE, short interspersed element; LINE, long interspersed element; MIR, mammalian-wide interspersed repeats; miR, microRNA.

1. INTRODUCTION

Camel’s milk has played and plays a key role for the outliving of the Bedouin, nomad and pastoral households in the hard arid and semi-arid regions of the world, supporting migrant populations since their domestication happened millennia ago. The Camelidae family consists of six species (C. dromedarius, C. bactrianus, L. glama, L. guanicoe, V. pacos, V. vicugna) among which the dromedary camel is certainly the most economically important domesticated one. According to FAO statistics (2011), most of the dromedary camels lives in the Horn of Africa mainly gathered in Somalia, Sudan, Ethiopia and Kenya, which together hold almost 60% of the entire population estimated to be about 24.5 million in the world. Despite their potential to survive on marginal resources in severe environment, camels have not been exploited as an important food source in the aforementioned countries. For instance, only 10% of the total milk produced in these rural regions is of camel origin (Faye and Konuspayeva, 2012). Conversely, in the countries of the Gulf, intensive camel milk production in high scale modernized unit has been already realized (Faye et al. 2002) and genetic improvement programmes for the milk production have been implemented (Hermas, 1988; Nagy et al. 2013).
The daily milk production of dromedary camels is estimated to vary between 3 and 10 kg during a lactation period of 12-18 months (Farah et al. 2007), depending on breed, stage of lactation, feeding and management conditions, with an average content of 2.9% and 3.1% of protein and fat respectively (Al haj and Al Kanhal, 2010).

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

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

The β-casein fraction has been deeply studied in ruminants and well characterized both at the protein and DNA level. Many alleles associated with different rates of protein synthesis have been identified in the corresponding coding gene (CSN2). At least 6 genetic variants have been reported in sheep (Chianese et al. 1997, Ceriotti et al. 2004; Chessa et al. 2010), whereas at least 8 alleles corresponding to 7 β-CN variants have been identified in goat (Cosenza et al. 2005; Caroli et al. 2006; Cosenza et al. 2007), and at least 17 alleles corresponding to 12 β-CN variants have been identified so far in cattle (Caroli et al. 2009). Conversely, in camels no genetic variants have been reported so far for the β-casein, and all the information we have-to date- (HPLC migration, molecular mass, 2D electrophoresis pattern, etc…) have been determined at the protein level (Farah and Farah-Riesen, 1985; Kappeler et al. 1998; El-Agamy et al. 2009; Hinz et al. 2012).
Genetic variants in camel milk are known for the αs1-casein fraction (Shuiep et al. 2013), whereas Kappeler et al. (1998) reported on non-allelic variants of the same casein fraction and described an amino acid exchange in position 30 (p.Glu > Asp) of the mature protein, which is not mentioned in UniProt-database O97943. The κ-casein gene (CSN3) has been also investigated in camels and 17 polymorphic sites have been recently detected by Pauciullo et al. (2013).

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

2. MATERIALS AND METHODS

2.1. Animals

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

Additional blood samples were collected from 3 Camelus bactrianus (bactrian camel) belonging to Wilhelma Zoo (Stuttgart, Germany), whereas hair samples were collected from further 5 bactrian
camels at the Frankfurt Zoo (Frankfurt, Germany). These samples were treated according to Spin Blood Mini Kit (Invitek, Germany) and Gen-ial All Tissue DNA Kit (Gen-ial, Germany), respectively.

2.2. DNA isolation

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

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

2.3. PCR amplification conditions and sequencing

Four test samples (one for each ecotype) were chosen for the sequencing of the whole camel CSN2 gene. DNA regions spanning from nucleotide -2141 to +7845 were amplified by means of a iCycler (Bio-Rad). A set of 31 primers for amplification and sequencing were designed by means of DNAsis-Max ver. 3.0 software (Hitachi), using as preliminary template the complete sequence of camel cDNA available in gene bank (EMBL acc. no. AJ012630) and then the new sequences determined in the course of the research. A typical PCR reaction mix (50 µl) comprised: 100 ng of genomic DNA, 1X PCR Buffer (Promega), 2.5 mM MgCl₂, 5 pmol of each primer, dNTPs each at 200 µM, 2.5 U of Taq DNA Polymerase (Promega). In table 1 we report the 10 starting overlapping amplicons for the sequencing of the whole gene. PCR was performed under the following conditions: 95°C (4 min), 35 cycles at 95°C (60 s), annealing temperatures depending on amplicon (Table 1) (45 s), 72°C (90 s), final extension at 72°C (10 min).
PCR products were purified and sequenced. The purification was carried out using MSB®Spin PCRapace kit (Invitek, Germany). Sequence was accomplished by ABI 3130 Genetic Analyzer (Applied Biosystem). Targeted fragments were sequenced in both directions using BigDye chemistry (Applied Biosystems).

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

2.4. Genotyping by HphI PCR-RFLP

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

The entire panel of 180 animals was genotyped for the g.2126A>G SNP using a PCR-RFLP method. Digestion of 17µl of each PCR amplification was accomplished with 5 U of HphI endonuclease (5’…GGTGANG6↓…3’) (New England Biolabs) over-night at 37°C. The digestion products were analysed directly by electrophoresis in 2.5% agarose gel in 0.5X TBE buffer and stained with Midori Green Advance (Nippon Genetics).

2.5. Bioinformatics

The allele frequencies and Hardy-Weinberg equilibrium ($\chi^2$ test) were calculated for the 180 camels. SNP discovery, homology searches, comparison among sequences, and multiple alignments were accomplished using DNAsis-Max ver. 3.0 software (Hitachi Software), whereas the putative transcription factor binding sites were searched by Transfact 7.0 software. Interspersed elements were found by RepeatMasker Web Server (http://www.repeatmasker.org/cgi-
bin/WEBRepeatMasker). Introns were also analyzed for potential microRNA sequences by using the bovine miRBase database (http://www.mirbase.org/search.shtml). NetPhos server 2.0 (http://www.cbs.dtu.dk/services/NetPhos/) was used for the prediction of phosphorylation sites, whereas BIOPEP database (http://www.uwm.edu.pl/biochemia/index.php/en/biopep) was used for the detection of putative bio-peptides.

3. RESULTS AND DISCUSSION

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

The entire gene encoding the camel β-casein (CSN2) plus 2141 bp at the 5' flanking region were sequenced and the sequences was deposited in EMBL under the following ID number: HG969421. The gene extends over 7819 nucleotides including 1115 bp of exonic regions and 6704 bp of intronic regions. The comparison with the corresponding bovine sequence (EMBL ID: X147111) showed a total similarity of about 78.0%. Approximately the same level of homology was also found by the following comparison to the β-casein gene in other species: 78.0% vs river buffalo (EMBL ID: FN424088), goat (EMBL ID: AJ011018) and pig (EMBL ID: EU025876), 73.0% vs donkey (EMBL ID: FN598778) and 70.0% vs human (EMBL ID: AF027807).

On the whole, the camel CSN2 gene shares a similar organization with the bovine counterpart, with some differences both at the exonic and intronic level. The camel cDNA is in total 15 bp longer than the bovine one (1115 bp vs 1100 bp). The main differences in size were detected at the 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) and at the exon 7 for 7 triplets (519 bp vs 498 bp). Instead, the exon 9 corresponding to the 3' un-translated region is shorter in camel than in cows (317 bp vs 323 bp). The intron sizes of the β-casein gene are approximately comparable between these two species, however, the exon/intron size ratio is higher in camel (1:6.01) than that observed in cattle (1:6.74).
The camel CSN2 gene is characterized by high A/T content compared to G/C (66.5% vs 33.4%). This feature characterizes also the camel κ-casein gene (Pauciullo et al. 2013) and in general seems to be conserved among the species (Ward et al. 1997). It contains 9 exons, ranging in size from 24 bp (exon 5) to 519 bp (exon 7), and 8 introns from 95 bp (intron 5) to 1950 bp (intron 1). The first exon (51 bp) plus 12 bp of the exon 2 are not coding at all. The signal peptide (15 amino acids) is coded by the following 45 nucleotides of the exon 2, whereas the mature peptide (217 amino acids) is encoded by the last 2 triplets of the exon 2 and the remaining part of the cDNA. The translation stop codon TAA is realized between the 4\textsuperscript{th}-6\textsuperscript{th} nucleotides of the exon 8, whereas the polyadenylation site is located between the 296\textsuperscript{th} and 301\textsuperscript{st} nucleotide of the exon 9. Splice donor and acceptor consensus sequences conforming to the AG/TG rule were identified at the exon/ intron boundaries.

The translation initiation codon (AUG) and the DNA sequence immediately upstream and downstream are known to influence the translation process (Kozak, 1991). The consensus sequence (GCCRCCAUGG) was indicated as the optimal context for the initiation of the translation in mammals sequence (Kozak, 1987) and the nucleotides -6, -3, and +4 (taking the A nucleotide of the AUG codon as number +1) are reported as the most conserved positions in natural mRNA sequences. With the exception of the AUG (nucleotides: +1, +2 and +3), the dromedary CSN2 showed only 2 nucleotides homologous to Kozak sequence (positions: -3 and -1). On the contrary, the other species showed at least 3 conserved positions (table 2) for the β-casein. According to this comparison, the dromedary camel is characterized by the lowest homology with Kozak consensus site. Such condition should give a weaker translation efficiency, however, the camel β-casein shares 6 nucleotide positions out of 7 with ruminants (+4, -1, -3, -4, -5, -6) (Cosenza et al. 2011) and 5 out of 7 with mice (+4, -1, -3, -4, -6), rabbits (+4, -1, -2, -3, -4) and dogs (+4, -1, -3, -4, -6) (table 2). Therefore, it would be reasonable to achieve future translation efficiency studies to verify this statement in dromedary camels.
SNP discovery was accomplished by the re-sequencing of 40 she-camels mainly for the exonic regions and the gene promoter. The comparison of the sequences showed a total of 11 polymorphic sites (7 transitions and 4 transversions) (table 3). One nucleotide difference was detected at the exonic level. It is a transversion (g.4175C>A) which falls within the codon 7 of the signal peptide. Although this mutation is synonymous (GCC^{Ala}→GCA), it represents the first example of polymorphism in CSN2 coding region for such species.

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

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

Many genetic variants of β-casein have been identified so far at the DNA or protein level in domestic animals. For instance, cattle (Caroli et al. 2009) and goat (Caroli et al. 2006) are the most polymorphic species among ruminants, whereas pigs (Cieslak et al. 2012, Suteu et al. 2012) and horses (Miclo et al. 2007) are less polymorphic and no polymorphisms were detected in other investigated species like donkeys and rabbits. Camels were not investigated so far, but the limited number of polymorphisms found in coding regions and the number of mutations detected in the rest of the CSN2 gene suggests that the genetic variability of camel β-casein is closer to non-ruminants.
However, the β-casein is the most abundant protein component also in camel milk (Kappeler et al., 2003; El-Agamy, 2006) and any allele detected at CSN2 locus might have a potential influence on the quality and the technological properties of milk. For instance, the CSN2 B allele was found to be more favourable for rennet coagulation and the cheese making quality of cow milk (Di Stasio and Mariani, 2000), whereas the A1 and A2 alleles which differ only for a transversion (A>C) responsible for the His→Pro amino acid change at codon 67, show divergent bio-peptide activities as consequence of a conformational variation in the secondary structure of the expressed protein (Kaminski et al. 2007). Therefore, a larger investigation of β-casein should be also performed in camels, in order to detect nucleotide variability and evaluate it for the potential influence on the quality and properties of milk.

3.2. The β-casein cDNA in bactrian camel

Starting from the genomic DNA, all the exonic regions of the β-casein were amplified and sequenced also for Camelus bactrianus (EMBL ID: HG969423). This species resulted more polymorphic for the CSN2 locus than the dromedary camel. Four SNPs were detected in total (table 4), three of which in the coding region (exon 7) and the last one at the 3’ UTR (exon 9). The SNPs c.591G>A and c.732G>A are both conservative, whereas the SNP c.666G>A (corresponding to the nucleotides 8177 of C. dromedarius CSN2 gene) is responsible for the amino acid change Met→Ile at the position 201 of the mature protein. This mutation represents in absolute the first missense mutation described at the β-casein gene in this species and it identifies two variants that we name: CSN2 A (for the carriers of the G, which also characterizes the dromedary camel CSN2 cDNA sequence and it might be considered as the ancestral form of the gene), and CSN2 B (for the carriers of the A and it might be considered as the mutant form). The following comparison with the dromedary cDNAs showed an additional polymorphic site in the exon 9 (table 4). This mutation also falls within the 3’ UTR, therefore it has no influence at the protein level. Analyzing the same
cDNA fragment within the four main ruminant species (cow, buffalo, goat and sheep) we have found at least 32 SNPs and the insertion/deletion of 2 triplets (TATCC) in cattle and buffalo β-casein. These findings confirm again that the level of genetic variation detected within Camelidae is lower than the corresponding variability found in ruminants both at the cDNA and at gene level.

3.3. Phosphorylation prediction and active bio-peptide analysis

The complete cDNA of dromedary camel and the corresponding encoding protein is reported in figure 1. The analysis of the serine/threonine phosphorylation sites predicted by NetPhos server v2.0 indicated 10 putative sequence motifs. Previous studies have shown that the most frequent form of camel β-casein is constitutively phosphorylated at Ser\(^{15}\), Ser\(^{17}\), Ser\(^{18}\) of the mature protein (Kappeler et al. 1998). Recently, Saadaoui et al. (2014) showed the presence of an additional phosphate, most probably the Ser\(^{19}\) as also predicted by Kappeler et al. (1998). These data agree with the reports on bovine β-casein, which is characterized by the same phosphorylation sites and, in addition, it is phosphorylated also at Ser\(^{50}\) (Farrell et al. 2004). The last site is not conserved in camels being substituted by Ile at the same position. However recently in cattle Li et al. (2012) identified Thr\(^{56}\) as a new phosphorylation site. This amino acid falls within a stretch of 9 amino acids well conserved also in camels and, although the phosphorylation of threonine in the Thr-Xaa-Asp motif is relatively uncommon for caseins (Holland, 2009), we can’t exclude that such a phosphorylation occurs also for camels β-casein. A variable degree of phosphorylation has been already demonstrated in other species, for instance 7 for equine β-casein (Girardet et al. 2006), 6 for ovine β-casein (Mamone et al. 2003), and 5 for human β-casein (Poth et al. 2008). Moreover, in goat the β-casein is homologous to bovine β-casein concerning the 5P, but it presents a specific additional phosphorylation site (Neveu et al. 2002), which may indicate a different function of caseins between caprines and bovines. In light of these studies and considering contrasting information on camel β-casein (Kappeler et al. 1998; Saadaoui et al. 2014), further investigations
are necessary for elucidating the different phosphorylation level and the number of existing isoforms for this protein.

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

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

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

The analysis of 5’ flanking region provides an important contribution for the evaluation of the transcription factors involved in the regulation of the gene expression. Mutations detected in the promoter region have been reported to modify the transcription rates or the mRNA stability and, consequently, the content of a particular protein in the milk (Prinzenberg et al. 2003; Szymanowska et al. 2004; Kuss et al. 2005). We decided to extend the sequencing to the 5’-flanking region for SNP discovery purposes and to analyze the putative regulatory regions of the camel β-casein gene.

About 2150 bp of the CSN2 gene were amplified and sequenced for Camelus dromedarius (EMBL ID: HG969421) and Camelus bactrianus (EMBL ID: HG977711).

Homology level and similar locations of most putative transcription binding sites were already reported among bovine (X14711), caprine (AF409096), ovine (X79703), rabbit (X15735), human (AF027807), rat (M10936), murine (X13484) and porcine (AY452035) β-casein promoters (Malewski, 1998; Lee et al. 2008). Similar mechanisms of regulation are supposed for these species, and therefore camels β-casein promoter sequences were compared with the homologous sequences
of horse (AY579425S1) promoter - so far compared only to camel sequence cDNA (Lenasi et al. 2003) - and bovine, respectively as non-ruminant and ruminant counterpart.

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

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

Seven Octamer bind protein (Oct-1), two Tata Binding Protein (TBP) and one Nuclear factor-1 (NF-1) were found, all potentially involved in the activation of the camel β-casein transcription. For instance, Oct-1 is not known to be a strong transcriptional activator by itself, but in conjunction with other co-activators. Zhao et al. (2002) showed that Oct-1 and STAT5 are both involved in the hormonal induction of casein gene expression. They bind DNA cooperatively and interact in the
stabilization of the transcription complex. However, Oct-1 can also interact with the TBP components (Zwilling et al. 1994). Oct-1 and TBP are putative sites very representative of camel β-casein promoter being characterized by 7 Oct-1 and 2 TBP putative sites. Therefore, it would be reasonable to suggest that Oct-1 may also mediate stimulation of β-casein expression in mammary gland through this way.

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

A short tandem repeat (STR) was found between the nucleotides -1151/-1136. It is a polymorphic microsatellite (TA)$_n$ characterized by 8 and 9 repetitions for the two alleles found in Camelus dromedarius, whereas the bactrian camels were monomorph (TA)$_8$. So far, the total number of microsatellites used for parentage identification in camels were 17 in total (Spencer et al. 2010), and the knowledge in this field is still very limited. Therefore, the identification and the
characterization of this microsatellite in the promoter region of camel β-casein gene add new information for future parentage identification tests in this species. No further polymorphisms were evidenced by the direct sequencing of dromedary CSN2 promoter and no genetic variability was detected for the 8 available bactrian camels for the CSN2 5'-flanking region. However, the following comparison of the two homologous sequences evidenced a total of 12 SNP (table 3), which might be used as genetic marker to differentiate both dromedary and bactrian species or potentially useful for dairy product traceability. A total of 28 polymorphic sites were instead counted when the comparison was extended to the other available β-casein promoter sequences (table 3).

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

3.5. Genotyping

The transition g.2126A>G detected in the dromedary camel CSN2 promoter region removes a restriction site for the endonuclease HphI, thus a PCR-RFLP protocol was set up for the quick genotyping of 180 she-camels. The digestion of the PCR product (659 bp) allows the identification of both alleles (figure 2).
The restriction pattern is characterized by 2 fragments of 608 bp and 51 for the GG samples, whereas the band 608 bp long is further restricted into two fragments of 352 bp and 256 bp in the presence of the adenine. The restriction pattern of the heterozygous samples shows 4 fragments. The major allele had a relative frequency of about 0.65, with a variation among the ecotypes between 0.59 and 0.72 (table 5). The $\chi^2$ value showed that there was no evidence of departure from the Hardy-Weinberg equilibrium ($P \leq 0.05$). The lower frequency of the G allele is an indication that this genetic variant, responsible for the lower affinity of the TATA box binding site, might be less facilitate in the gene transcription process.

### 3.6. Interspersed elements and microRNA

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

Three elements (A, B, D) have been found in the promoter, intron 1 and intron 6 respectively and they appear to be LINEs. The B element situated between the nucleotides +3667/+3796 of the intron 1 shows a similarity of 71.5% with a LINE-2 family, whereas the A and D elements belong both to the family L1 and both are characterized by a complementary sequence. In particular, the A element is located in the promoter region approximately 1000 bp upstream the exon1 (nucleotides +925/+1143), it is 219 bp long and it contains a polymorphic microsatellites already described in the section 3.2. Conversely, the second element is located at the intron 6 between the nucleotides +7263/+7364. It is a truncated LINE (102 bp) which appears to be characteristic of the camel CSN2 sequence, since the bovine, buffalo, goat, sheep and donkey CSN2 sequences are characterized by the presence of a SINE, whereas pig, human and rabbit $\beta$-casein gene do not show any transposable element in that position.
The C element is also situated in the intron 6 between the nucleotides +6637/+6736. This is a SINE belonging to the class of mammalian-wide interspersed repeats (MIR/Mon1a7) and it seems to be very well conserved among the species. The same element was in fact found in all afore mentioned species with the exception of the rabbit which does not show repetitive elements and human which is characterized by the presence of a LINE.

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

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

The evolution of mammalian genome has been widely influenced by transposable elements which have played a prominent role in species diversification (Bowen and Jordan, 2002). The lower number of repetitive element in camel is an indication that *Tylopoda* diverged from *Ruminantia* before additional retrotransposition events took place at different times during the divergence of in such suborder (Nijman et al. 2002). This statement is also confirmed by the data already reported for the camel κ-casein gene (Pauciullo et al. 2013). Furthermore, the distinctive LINE (D element) found in camel CSN2 sequence and the truncated LINE (C element) found by Pauciullo et al. (2013)
belong both to L1MA family and most probably they have been originated by a similar replication event occurred MYA. The presence/absence of these elements and those characterizing the bovine sequence add interest in the knowledge of species evolution and they can be considered powerful phylogenetic markers for clustering studies.

17 putative mature sequences for microRNA were found in camel CSN2 (table 6) which in total show more than 2600 predicted target genes most of which related to immune system. It is interesting to notice that 6 out 17 are located in the intron 3 whose size is only 108 bp and they recognize almost the same sequence, a stretch of DNA of 20 bp between the nucleotides +5082/+5104. The biological functions of most miRNA are unknown, but it is estimated that >30% of protein-coding genes are regulated by miRNA (Lewis et al., 2005). For instance, one of the target genes for miR-190a and miR-2437 is C/EBP-α whose transcription factor binding site is present in the proximal promoter region of camel CSN2 and therefore these miRNAs might be regulator of the gene expression. Conversely, one of the miR-2345, miR-2361 and miR-2444 target genes are the casein kinase-1 (CSNK1-α1, alpha-1; CSNK1-γ1, gamma-1 and CSNK1-δ, delta) whose general activity is the selective phosphorylation of serine/threonine amino acids. As consequence, these miRNAs might have a post-transcriptional influence on β-casein.

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

4. CONCLUSION

1. The Camelus dromedarius CSN2 gene shares a similar organization compared to the homologous gene in other species. The exon/intron ratio in camel CSN2 was slightly higher (1:6.01) than that observed in bovine (1:6.74).
2. The different number and nature of the interspersed repeated elements (5 in camel vs 9 in bovine) are responsible for the difference in size of the camel β-casein gene. One LINE truncated is characteristic of the camel CSN2 sequence. This element could be used as a genetic marker typical of the specie, as well as for evolution and clustering studies.

3. The camel composite response element (CoRE) region was identified and the relative transcription factors (C/EBP MGF/STAT5, GR, NF-1 and YY1) were indicated in the 5’ flanking region. The interactions among the CoRE and other representative motifs like Oct-1 and TBP suggests a possible way of regulation of the gene expression.

4. A total of 47 polymorphic sites have been described in the promoter as well as in the gene and they represent a pre-requisite for the regulation of the β-casein gene expression. The g.2126A>G SNP detected in dromedary CSN2 promoter falls within the TATA-box binding site. The transversion g.4175C>A represents the first synonymous allele for dromedary camel CSN2. Conversely, the transition c.666G>A at the exon 7 of the bactrian CSN2 cDNA is responsible for the Met→Ile amino acid change at the position 201 of the mature protein. It represents in absolute the first missense mutation at the camel β-casein and it identifies 2 variants named A and B.

5. The presence of putative bio-functional peptides belonging to ACE-inhibitor and antioxidative families suggests a potential protective role of the camel milk for the human nutrition, whereas the identification of microRNA sequences offers also in camels biomarkers for the quality control of commercial milk products.

5. ACKNOWLEDGEMENTS

The authors thank Dr. El Tahir Salih Shuiep and the camel owners in Sudan for providing Camelus dromedarius samples; Dr. Tobias Knauf-Witzens from Wilhelma Stuttgart Zoo and Dr. Nicole Schauerte from Frankfurt Zoo (Germany) for providing Camelus bactrianus samples. The
authors would like also to thank Therese Kubetzki-Bauer, Anja Scheuermann and Zhaoxin Wang for the excellent technical assistance.

6. REFERENCES


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


### Table 1.

<table>
<thead>
<tr>
<th>Amplicon</th>
<th>Region covered</th>
<th>Primers</th>
<th>Sequence</th>
<th>Ta °C</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5' UTR</td>
<td>Cd CSN2 5' F*</td>
<td>5'-ATTCACATGGTTTCAACTATTAG-3'</td>
<td>60.0</td>
<td>1263 bp</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cd CSN2 5' R*</td>
<td>5'-GGTCTTGCCCAAGTATTA-3'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>5' UTR</td>
<td>Cd CSN2 prom F*</td>
<td>5'-TTGTAAAAATAAGTAGGTGT-3'</td>
<td>51.5</td>
<td>1093 bp</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cd CSN2 prom R*</td>
<td>5'-TAATTCAAAATCTGATAA-3'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>5' UTR/Ex1/Intr1</td>
<td>Cd CSN2 Intr1 R†</td>
<td>5'-TCAAATCTATACAGGCACCT-3'</td>
<td>59.5</td>
<td>659 bp</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cd CSN2 prom F2*</td>
<td>5'-GTTCCTCCATTACAGCATC-3'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Ex1/Intr1/Ex2</td>
<td>Cd CSN2 Ex1F</td>
<td>5'-GTTCCTCATCCTTCCCTCCAC-3'</td>
<td>61.5</td>
<td>2015 bp</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cd CSN2 Ex1R</td>
<td>5'-AGGCAAGGATGAGGAC-3'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Ex2/Intr2/Ex3/Intr3/Ex4</td>
<td>Cd CSN2 Ex2 F</td>
<td>5'-GAAGAATTCAAGACAGCGG-3'</td>
<td>61.5</td>
<td>1322 bp</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cd CSN2 Ex2 R</td>
<td>5'-CTGTTTGTGGCTGTTCC-3'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Ex3/Intr3/Ex4/Intr4/Ex5/Intr5/Ex6</td>
<td>Cd CSN2 Ex3 F</td>
<td>5'-GAATCTATTACACACATCA -3'</td>
<td>59.5</td>
<td>2005 bp</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cd CSN2 Ex3 R</td>
<td>5'-AACAGGCAGGACTT -3'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Ex5/Intr5/Ex6/Intr6/Ex7</td>
<td>Cd CSN2 Ex5 F</td>
<td>5'-GATGAAACACACAGATAAATC-3'</td>
<td>63.0</td>
<td>1159 bp</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cd CSN2 Ex5 R</td>
<td>5'-AAAGTGAGGAGGAGGATG-3'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Ex7/Intr7/Ex8</td>
<td>Cd CSN2 Ex7 F</td>
<td>5'-GCCTAAAGAGATTTCCAAGT-3'</td>
<td>62.5</td>
<td>1081 bp</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cd CSN2 Ex7 R</td>
<td>5'-AGATTAAGGCACAACATGG-3'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Ex8/Intr8/Ex9</td>
<td>Cd CSN2 Ex8 F</td>
<td>5'-TATAATAAAGACAGGAAAACTTAA-3'</td>
<td>56.0</td>
<td>476 bp</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Cd CSN2 Ex8 R</td>
<td>5'-TATAATAAAGACAGGAAAACTTAA-3'</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Intr8/Ex9/3'UTR</td>
<td>Ea CSN2 3end R</td>
<td>5'-TATAATAAAGACAGGAAAACTTAA-3'</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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).
<table>
<thead>
<tr>
<th>Species</th>
<th>EMBL ID</th>
<th>Position</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>-6</td>
</tr>
<tr>
<td>Kozak consensus sequence</td>
<td></td>
<td>G</td>
</tr>
<tr>
<td>C. dromedarius present work</td>
<td>HG969421</td>
<td>A</td>
</tr>
<tr>
<td>B. taurus</td>
<td>X14711</td>
<td>A</td>
</tr>
<tr>
<td>B. bubalis</td>
<td>FN424088</td>
<td>A</td>
</tr>
<tr>
<td>C. hircus</td>
<td>AJ011018</td>
<td>A</td>
</tr>
<tr>
<td>O. aries</td>
<td>X79703</td>
<td>A</td>
</tr>
<tr>
<td>E. caballus</td>
<td>AF214526</td>
<td>T</td>
</tr>
<tr>
<td>E. asinus</td>
<td>FN598778</td>
<td>T</td>
</tr>
<tr>
<td>S. scrofa</td>
<td>EU025876</td>
<td>A</td>
</tr>
<tr>
<td>M. musculus</td>
<td>X13484</td>
<td>A</td>
</tr>
<tr>
<td>O. conicus</td>
<td>M33582</td>
<td>-</td>
</tr>
<tr>
<td>C. lupus familiaris</td>
<td>AB035080</td>
<td>A</td>
</tr>
<tr>
<td>H. sapiens sapiens</td>
<td>AF027807</td>
<td>-</td>
</tr>
</tbody>
</table>

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.
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.
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.

<table>
<thead>
<tr>
<th>Position</th>
<th>Nucleotide genomic DNA</th>
<th>Nucleotide cDNA</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exon 1</td>
<td>2186</td>
<td>45</td>
<td>A</td>
<td>A</td>
<td>-</td>
</tr>
<tr>
<td>(5’UTR)</td>
<td>2190-2191</td>
<td>49-50</td>
<td></td>
<td>AA</td>
<td>AA</td>
</tr>
<tr>
<td>Exon 2</td>
<td>4175</td>
<td>84</td>
<td>M^Ala/Ala</td>
<td>C^Ala</td>
<td>C^Ala</td>
</tr>
<tr>
<td></td>
<td>4180</td>
<td>89</td>
<td>T^Leu</td>
<td>T^Leu</td>
<td>G^Arg</td>
</tr>
<tr>
<td>Exon 7</td>
<td>8102</td>
<td>591</td>
<td>G^Gln</td>
<td>R^Gln/Gln</td>
<td>G^Gln</td>
</tr>
<tr>
<td></td>
<td>8177</td>
<td>666</td>
<td>G^Met</td>
<td>R^Met/Val</td>
<td>G^Met</td>
</tr>
<tr>
<td></td>
<td>8243</td>
<td>732</td>
<td>G^Val</td>
<td>R^Val/Vul</td>
<td>G^Val</td>
</tr>
<tr>
<td>Exon 9</td>
<td>9745</td>
<td>900</td>
<td>T</td>
<td>C</td>
<td>T</td>
</tr>
<tr>
<td>(3’UTR)</td>
<td>9778</td>
<td>933</td>
<td>C</td>
<td>Y</td>
<td>C</td>
</tr>
<tr>
<td>Ecotypes</td>
<td>AA</td>
<td>AG</td>
<td>GG</td>
<td>Total</td>
<td>Allelic frequency</td>
</tr>
<tr>
<td>-----------</td>
<td>-----</td>
<td>-----</td>
<td>-----</td>
<td>-------</td>
<td>-------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>A</td>
</tr>
<tr>
<td>Shanbali</td>
<td>11</td>
<td>19</td>
<td>2</td>
<td>32</td>
<td>0.64</td>
</tr>
<tr>
<td>Kahli</td>
<td>24</td>
<td>20</td>
<td>3</td>
<td>47</td>
<td>0.72</td>
</tr>
<tr>
<td>Arabi</td>
<td>13</td>
<td>12</td>
<td>2</td>
<td>27</td>
<td>0.70</td>
</tr>
<tr>
<td>Lahaoi</td>
<td>23</td>
<td>41</td>
<td>10</td>
<td>74</td>
<td>0.59</td>
</tr>
<tr>
<td>Total</td>
<td>71</td>
<td>92</td>
<td>17</td>
<td>180</td>
<td>0.65</td>
</tr>
</tbody>
</table>

$\chi^2=2.737 - p\leq0.05 - d.o.f.=1$

**Table 5.** Genotyping data and allele frequency of the g.2126A>G SNP at the CSN2 promoter gene in Sudanese *Camelus dromedarius* population.
<table>
<thead>
<tr>
<th>Accession number</th>
<th>ID</th>
<th>Location</th>
<th>Position</th>
<th>Sequence</th>
<th>Strand</th>
</tr>
</thead>
<tbody>
<tr>
<td>MIMAT0009251</td>
<td>bta-miR-190a</td>
<td>Intron 1</td>
<td>+2753/+2771</td>
<td>5’-TAATATTTCAAAAAATATAA-3’</td>
<td>-</td>
</tr>
<tr>
<td>MIMAT0011880</td>
<td>bta-miR-2345</td>
<td></td>
<td>+4081/+4096</td>
<td>5’-GAAATAAGAAATACA-3’</td>
<td>-</td>
</tr>
<tr>
<td>MIMAT0011794</td>
<td>bta-miR-2287</td>
<td>Intron 2</td>
<td>+4418/+4431</td>
<td>5’-CTGCCACCAGTCCC-3’</td>
<td>-</td>
</tr>
<tr>
<td>MIMAT0011805</td>
<td>bta-miR-2297</td>
<td></td>
<td>+4486/+4504</td>
<td>5’-AAGAAACACAAAAAGAGT-3’</td>
<td>-</td>
</tr>
<tr>
<td>MIMAT0012011</td>
<td>bta-miR-2437</td>
<td>Intron 3</td>
<td>+5086/+5104</td>
<td>5’-TGATTTTTTTTTTCTTT-3’</td>
<td>+</td>
</tr>
<tr>
<td>MIMAT0011946</td>
<td>bta-miR-2393</td>
<td></td>
<td>+5085/+5104</td>
<td>5’-TTGATTTTTTTTTTCTTT-3’</td>
<td>+</td>
</tr>
<tr>
<td>MIMAT0011899</td>
<td>bta-miR-2361</td>
<td></td>
<td>+5082/+5098</td>
<td>5’-GTGTTGATTTTTTTTT-3’</td>
<td>+</td>
</tr>
<tr>
<td>MIMAT0011934</td>
<td>bta-miR-2325c</td>
<td></td>
<td>+5087/+5104</td>
<td>5’-GATTTTTTTTTTCTTT-3’</td>
<td>+</td>
</tr>
<tr>
<td>MIMAT0011897</td>
<td>bta-miR-2359</td>
<td></td>
<td>+5082/+5097</td>
<td>5’-GTGTTGATTTTTTTTT-3’</td>
<td>+</td>
</tr>
<tr>
<td>MIMAT0024574</td>
<td>bta-miR-3613</td>
<td></td>
<td>+5083/+5103</td>
<td>5’-TGTTGATTTTTTTTTCTT-3’</td>
<td>+</td>
</tr>
<tr>
<td>MIMAT0009947</td>
<td>bta-miR-376c</td>
<td>Intron 6</td>
<td>+6468/+6488</td>
<td>5’-TAACACATATAATCAATTTC-3’</td>
<td>-</td>
</tr>
<tr>
<td>MIMAT0009237</td>
<td>bta-miR-147</td>
<td></td>
<td>+6569/+6584</td>
<td>5’-TAGTAGAAACATTTCC-3’</td>
<td>-</td>
</tr>
<tr>
<td>MIMAT0025545</td>
<td>bta-miR-411c-5p</td>
<td></td>
<td>+7394/+7413</td>
<td>5’-ATCTATGATCTATTCAC-3’</td>
<td>-</td>
</tr>
<tr>
<td>MIMAT0024594</td>
<td>bta-miR-6122-5p</td>
<td></td>
<td>+7467/+7485</td>
<td>5’-GCACAGAAAAAGTAAAAC-3’</td>
<td>-</td>
</tr>
<tr>
<td>MIMAT0011832</td>
<td>bta-miR-2284t-3p</td>
<td></td>
<td>+7500/+7519</td>
<td>5’-CCAAACATTAATCCACTT-3’</td>
<td>-</td>
</tr>
<tr>
<td>MIMAT0003830</td>
<td>bta-miR-361</td>
<td>Intron 7</td>
<td>+8476/+8495</td>
<td>5’-TACCCACCGAGATTCTGATA-3’</td>
<td>-</td>
</tr>
<tr>
<td>MIMAT0003797</td>
<td>bta-miR-215</td>
<td>Intron 8</td>
<td>+9326/+9346</td>
<td>5’-TGTCTTTCCATTCACAAGTCA-3’</td>
<td>-</td>
</tr>
</tbody>
</table>

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’…GGTGAN↓…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.