J. Dairy Sci. 106:4158–4170
https://doi.org/10.3168/jds.2022-22598

A novel allelic donkey β-lactoglobulin I protein isoform generated by a non-AUG translation initiation codon is associated with a nonsynonymous SNP

G. Cosenza,[1]* P. Martin,[2] G. Garro,[1] D. Gallo,[1] B. Auzino,[3] R. Ciampolini,[3]* and A. Pauciullo[4]
[1]Department of Agricultural Sciences, University of Naples “Federico II,” 80055 Portici (Na), Italy
[2]Université Paris-Saclay, INRAe, AgroParisTech, GABI, 78350 Jouy-en-Josas, France
[3]Department of Veterinary Science, University of Pisa, 56100 Pisa, Italy
[4]Department of Agriculture, Forest and Food Sciences, University of Torino, 10095 Grugliasco (TO), Italy

ABSTRACT

β-lactoglobulin I (β-LG I) is one of the most important whey proteins in donkey milk. However, to our knowledge, there has been no study focusing on the full nucleotide sequences of this gene (*BLG* I). Current investigation of donkey *BLG* I gene is very limited with only 2 variants (A and B) characterized so far at the protein level. Recently, a new β-LG I variant, with a significantly higher mass (+1,915 Da) than known variants has been detected. In this study, we report the whole nucleotide sequence of the *BLG* I gene from 2 donkeys, whose milk samples are characterized by the β-LG I SDS-PAGE band with a normal electrophoretic mobility (18,514.25 Da, β-LG I B1 form) the first, and by the presence of a unique β-LG I band with a higher electrophoretic mobility (20,428.5 Da, β-LG I D form) the latter. A high genetic variability was found all over the 2 sequenced *BLG* I alleles. In particular, 16 polymorphic sites were found in introns, one in the 5′ flanking region, 3 SNPs in the 5′ untranslated region and one SNP in the coding region (g.1871G > A) located at the 40th nucleotide of exon 2 and responsible for the AA substitutions p.Asp28 > Asn in the mature protein. Two SNPs (g.920–922CAC > TGT and g.1871G/A) were genotyped in 93 donkeys of 2 Italian breeds (60 Ragusana and 33 Amiatina, respectively) and the overall frequencies of g.920–922CAC and g.1871A were 0.3065 and 0.043, respectively. Only the rare allele g.1871A was observed to be associated with the slower migrating β-LG I. Considering this genetic diversity and those found in the database, it was possible to deduce at least 5 different alleles (*BLG* I A, B, B1, C, D) responsible for 4 potential β-LG I translations. Among these alleles, B1 and D are those characterized in the present research, with the D allele of real novel identification. Haplotype data analysis suggests an evolutionary pathway of donkey *BLG* I gene and a possible phylogenetic map is proposed. Analyses of mRNA secondary structure showed relevant changes in the structures, as consequence of the g.1871G > A polymorphism, that might be responsible for the recognition of an alternative initiation site providing an additional signal peptide. The extension of 19 AA sequence to the mature protein, corresponding to the canonical signal peptide with an additional alanine residue, is sufficient to provide the observed molecular weight of the slower migrating β-LG I encoded by the *BLG* I D allele.

Key words: donkey, β-lactoglobulin I, polymorphisms, long protein isoform, alternative translation initiation

INTRODUCTION

Donkey (*Equus asinus*) milk is characterized by a lower protein content with respect to the ruminants' milk and a quantity of casein comparable to that of human milk (Cosenza et al., 2019). The 2 main fractions of donkey milk proteins are the caseins and whey proteins. Casein fraction in donkey milk (5.12 mg/mL) is mainly represented by α_{S1}-CN and β-CN and smaller amount of α_{S2}-CN and κ-CN (Chianese et al., 2010; Cosenza et al., 2019; Auzino et al., 2022).

Whey proteins content in donkey milk is about 4.9 to 9.6 mg/mL and is mainly made up of β-LG, α-LA, lysozyme, immunoglobulins, serum albumin, and lactoferrin. In particular, in donkey milk, the β-LG showed a mean content of 1.3 to 5.5 mg/mL, similar to the level observed in mare milk and lower than that of goat milk (Miranda et al., 2004; Brumini et al., 2016). In addition to being the major whey protein in the milk of most ruminants and equidae, β-LG is present in the milk of some monogastrics and marsupials but is absent in human, camel, lagomorph, and rodent milk (Wodas et al., 2020).

β-Lactoglobulin is a member of the ancient and widespread protein family of lipocalins, which determine nearly all major mammalian allergens (Jensen-Jarolim

Received July 29, 2022.
Accepted December 17, 2022.
*Corresponding authors: giacosen@unina.it and roberta.ciampolini@unipi.it

et al., 2016). To date, no clear physiological function has been defined for this protein, although several hypotheses have been advanced. Because β-LG is a lipocalin, a transport role has been suggested by analogy to other family members whose function is known. Among the several biological roles that have been proposed for β-LG, there are facilitators of vitamin A (retinol) uptake and an inhibitor, modifier, or promoter of enzyme activity. However, conclusive evidence for a specific biological function of β-LG is not available (Le Maux et al., 2014). The β-LG of ruminant species studied binds to other lipophilic vitamins and to fatty acids; however, equine and porcine β-LG do not bind to fatty acids. This different property indicates that the ability of ruminant β-LG to bind fatty acids is not shared by the homologous protein in nonruminant milk (Pérez et al., 1993; Pérez and Calvo, 1995; Sawyer and Kontopidis, 2000; Kontopidis et al., 2004; Mensi et al., 2013).

In equidae milk, β-LG exists predominantly as a monomer, whereas is dimeric in ruminant milk (Tidona et al., 2011). Two isoforms of β-LG exist in donkey milk; they are encoded by 2 paralogous genes, as also observed in horse, dog, and dolphin milk (Pervaiz and Brew, 1986; Godovac-Zimmermann et al., 1990), whereas 3 forms have been described in cat milk (Halliday et al., 1993; Pena et al., 1999). The 2 donkey β-LG have been named β-LG I and β-LG II. β- Lactoglobulin I is made of 162 AA residues and is similar to the horse β-LG I (from which it differs by 5 AA), whereas the β-LG II is 163 AA long. β-Lactoglobulin I is the major form representing 80% of the total donkey β-LG (Godovac-Zimmermann et al., 1990). The complete AA sequences of both donkey β-LG are known (Godovac-Zimmermann et al., 1988, 1990; Herrouin et al., 2000).

At present, the polymorphisms of donkey β-LG I and their effect on milk yield and composition are limited. To date, only 2 variants (A and B) have been characterized and only at the protein level. The 2 protein sequences show 98.1% similarity as a consequence of 3 AA substitutions: p.Ser36 > Glu, p.Thr97 > Ser and p.Ile150 > Val. As a result, the M_r (molecular mass) of the variant A is 18,528 Da, whereas the M_r of the variant B is of 18,514 Da (Godovac-Zimmermann et al., 1988, UniProtKB: P13613; Herrouin et al., 2000). Recently, a new β-LG I variant, with a significantly higher mass (20,428.5 Da) than known variants, has been detected by tandem MS analysis. The new variant showed an N-terminal extension likely related to the signal peptide sequence and apparently also characterized by the AA substitution p.Asp28 > Asn (Auzino et al., 2022).

Despite the interest in this species and the new attention surrounding it, thus far, this important gene has not been adequately investigated in donkeys. Thanks to the recently updated donkey genome assembly (BioProject: PRJNA688408), the complete β-LG I encoding gene (*BLG* I) sequence has been published and annotated on chromosome 10. Its total physical length is spread over 6 kb and it consists of 7 exons (NCBI Reference Sequence: NC_052186.1 from 8316754 to 8322198, Gene ID: 106829109).

Consequently, the identification of DNA polymorphisms at this locus and the knowledge of their impact on donkey milk composition are very limited. Only 4 SNPs were identified in the third intron of β-LG I in Turkish donkeys (Işık, 2019). However, unlike what has been done for horses (Wodas et al., 2020), no association studies with milk-related traits have been carried out.

This study had 2 goals. The first was to sequence and deeply annotate the whole *BLG* I gene of 2 donkeys, homozygotes for a β-LG I SDS-PAGE band with a normal electrophoretic mobility (18,514.25 Da) and for a higher electrophoretic mobility (20,428.5 Da), as determined by Auzino et al. (2022). The second was to compare the alleles in their complex genetic diversity and to identify the molecular event responsible for the different observed phenotype.

MATERIALS AND METHODS

Milk and DNA Samples

Individual milk and blood samples were collected from 93 female donkeys of 2 Italian breeds (60 Ragusana and 33 Amiatina) reared in Umbria and Toscana regions, respectively. All samples used in this study belong to collections of the University of Turin, University of Napoli Federico II, and University of Pisa. Individual blood samples were collected during the routine prophylaxis of the farm by an official veterinarian of ASL (Local Sanitary Unit) of the Ministry of Health. For this reason, Animal Care and Use Committee approval was not necessary.

The DNA was extracted from leucocytes according to Goossens and Kan (1981). The DNA concentrations and the ratio of absorbances at 260 and 280 nanometers (A260/280) were measured using a Nanodrop ND-2000 Spectrophotometer (Thermo Fisher Scientific Inc.). The DNA samples were used for sequencing and population analysis. Skim milk samples were analyzed by SDS-PAGE (Grosclaude et al., 1987).

PCR Amplification, Sequencing, and Bioinformatics

Donkey genome sequences (GenBank acc. no. NC_052186.1, region: 8316254 to 8322698 and acc. no. PSZQ01002145.1, region: 1252372 to 1258209) were used to design primers for PCR amplification and sequencing. The *BLG* I gene spanning from the 5′- to the 3′-flanking regions of 2 Amiatina donkeys were amplified by iCycler (BioRad). One donkey produced milk characterized by a β-LG I SDS-PAGE band with a normal electrophoretic mobility (18,514.25 Da), whereas the milk of the other was characterized by the presence of a unique β-LG I band with a higher electrophoretic mobility (20,428.5 Da), as determined by Auzino et al. (2022).

A typical 25-μL PCR reaction mix included 100 ng of genomic DNA, 50 m*M* KCl, 10 m*M* Tris-HCl (pH 9.0), 0.1% Triton X-100, 3 m*M* $MgCl_2$, 200 nmol of each primer, dNTPs each at 400 μ*M*, 0.5 U of Taq DNA Polymerase (Promega), and 0.04% BSA. The thermal condition for the amplification consisted of an initial denaturation at 97°C for 4 min, followed by 29 cycles at 94°C for 45 s, 54.0 to 57.4°C for 45 s (according to the amplicon), and 72°C for 2 min. A final extension of 10 min was accomplished to end the reaction. All PCR products were analyzed directly by electrophoresis in 1.5% TBE (Tris-Borate-EDTA) agarose gel (Bio-Rad) in 0.5× TBE buffer and stained with SYBR green nucleic acid stain (Lonza Rockland Inc.).

The PCR products were purified with QIAquick columns (Qiagen) and sequenced in outsourcing on both strands by Eurofins Genomics (Ebersberg, Germany) by Sanger technology. BLAST analysis (www.ncbi.nlm.nih.gov/BLAST) was used to confirm the sequencing results. Homology searches, comparisons among nucleotide and AA sequences, and multiple alignments for polymorphism discovery were accomplished using Dnasis Pro (Hitachi Software Engineering Co.). The regulatory regions were analyzed for potential transcription factors (**TF**) by AliBaba v2.1 program (http://www.gene-regulation.com/pub/programs/alibaba2/index.html).

The haplotype structure was defined according to Gabriel et al. (2002) using Haploview software version 4.2 (http://www.broadinstitute.org/haploview/haploview). 95% confidence bounds on D′ are generated and each comparison is called strong LD, inconclusive, or strong recombination. A block is created if 95% of informative (i.e., noninconclusive) comparisons are strong LD. This method by default ignores markers with MAF <0.05.

The effects of g.1871G > A and g.920–922CAC > TGT polymorphisms on mRNA secondary structure were analyzed with RNAfold (http://rna.tbi.univie.ac.at//cgi-bin/RNAWebSuite/RNAfold.cgi; Gruber et al., 2008) under default parameters.

Genotyping and Data Analysis

Genotyping methods based on allele specific PCR (**AS-PCR**) were developed to screen the mutations g.1871G > A and g.920–922CAC > TGT in the population. Primer sequences are reported in Supplemental Table S1. Data are deposited into Harvard Dataverse (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DLFGKD; Cosenza, 2023). Reaction mix and thermal conditions were performed as reported above except for the annealing temperatures of AS-PCR protocols that were specifically defined according to the mutations targeted.

The entire panel of 93 donkey DNA samples was genotyped. Allelic frequencies and Hardy–Weinberg equilibrium (chi-squared test) were calculated (https://gene-calc.pl/hardy-weinberg-page).

RESULTS AND DISCUSSION

SDS-PAGE Analysis

The SDS-PAGE analysis of 93 individual milk samples obtained from donkeys of the 2 Italian breeds screened (60 Ragusana and 33 Amiatina, respectively), gave 3 different patterns in agreement to what has been described by Auzino et al. (2022). In detail, 86 samples showed the presence of a β-LG I band displaying an electrophoretic mobility consistent with a molecular weight of 18,514.25 Da (wild type), one sample (Amiatina breed) with a β-LG I band whose electrophoretic mobility was higher than the wild type, and 6 samples (one Ragusana and 5 Amiatina) with both electrophoretic bands. These results, consistent with those reported by Auzino et al. (2022), led to hypothesize the presence of a new donkey β-LG I isoform with a molecular weight of 20,428.50 Da.

BLG I Gene Structure in Donkeys

By using genomic DNA as template, we sequenced the whole gene encoding the β-LG I (*BLG* I) of 2 Amiatina donkeys, homozygotes for the electrophoretic bands of 18,514.25 Da (β-LG I wild type) and 20,428.50 Da (β-LG I new variant), respectively (GenBank accession no. ON886232 and ON886233).

The sequenced DNA region including the *BLG* I gene is about 5,900 bp long, and it includes 1,407 bp of exonic regions, about 4,000 bp of intronic regions, 289 nucleotides at the 5′ flanking region and 216 bp at the 3′ flanking region. The level of sequence similarity with the 2 investigated alleles is about 99.5%.

The *BLG* I gene is characterized by a relatively simple structure. Taking as reference the gene coding for the nonmutated form, it contains 7 exons ranging in size from 45 bp (exon 6) to 738 bp (exon 1), with exons 1 to 6 that contribute to the open reading frame. In detail, only the last 96 nucleotides of the first exon are coding. The whole highly conserved signal peptide (18 AA, MKCLLLALGLALMCGIQA) of the mature protein (162 AA) is encoded by the nucleotides 643 to 696 of exon 1 and the translation stop codon TAA is created by nucleotides 15 to 17 of exon 6. The AA sequence (from 16 to 150 of the mature protein) reveals the structurally conserved regions typical of a lipocalin core (Supplemental Figure S1; https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DLFGKD, Cosenza, 2023).

The deduced CDS (coding sequence) length of donkey *BLG* I gene is 543 bp long, with an average CG (Guanine-Cytosine) content value of 59.3% and with all the splice junctions following the 5′GT/3′AG splice rule. This structure is comparable to that described by Wodas et al. (2020) in horses (GenBank accession no. PJAA01000026.1 from 38880546 to 38874901; NM_001082493), which shows a homology of 97.82%.

Intronic and Exonic Polymorphisms Detection

The analysis and the alignment of the *BLG* I intronic sequences of the 2 donkeys used in this study have highlighted a discrete genetic variability. In detail, 17 polymorphic sites (6 transversions, 10 transitions, 1 deletion/insertion) were found between the 2 sequenced animals (Supplemental Figure S1; Table 1). None of the intronic polymorphisms is apparently located in the regulatory regions (splicing donor/acceptor site, enhancer/silencer) and, as a consequence, they likely do not affect the gene expression. In particular, taking as reference the sequence of the wild type (GenBank accession no. ON886232), 2 polymorphisms (g.1203–1204delTGGAGC and g.1339C > T) are detected within a mammalian long-terminal repeat retrotransposon (**MaLR**) located between the nt 161 and 443 of the first intron. The MaLR are retroviral sequences that integrated into germ line cells millions of years ago. Transcripts of these long-terminal repeat retrotransposons are present in several tissues, and their expression is modulated in pathological conditions, although their function remains often far from being understood. Over time, most of these elements have accumulated numerous mutations, often compromising their coding capability. An important function of MaLR insertions would be linked to species' evolution (Pisano et al., 2020). Finally, several microsatellite sequences are present in the donkey *BLG* I gene (Supplemental Figure S1).

As expected, the comparison of the coding regions showed a reduced level of polymorphism. Only one SNP was identified. It is a transversion (g.1871G > A) located at the 40th nucleotide of exon 2 and responsible of the AA substitution p.Asp28 > Asn in the mature protein (Supplemental Figures S1 and S2, https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DLFGKD, Cosenza, 2023; Table 1).

Instead, comparing the sequences analyzed in this study with those recorded in GenBank for the donkey *BLG* I gene, it is possible to identify other genetic variability. In particular, 10 single intronic polymorphisms (7 transitions, 1 transversion, 2 indel mutations), 2 exonic polymorphisms (g.458A > T and g.1863C > T), and 2 repeats as mono- (g.4934C[6]) and tetranucleotide (g.5524CCTC[2]; Table 1).

Regulatory Elements and Polymorphisms Detection at the Promoter of the Gene

Auzino et al. (2022) postulated that the increase of about 1,915 Da of the new β-LG I variant can be justified by the fact that the protein entering the secretion pathway in the endoplasmic reticulum may include a N-terminal sequence extension, corresponding to the canonical signal peptide (β-LG I precursor), with an additional Ala residue (<u>A</u>MKCLLLALGLALMCGIQA).

The analysis of the gene promoter and its regulatory regions may provide an answer to this hypothesis. In fact, variations in regulatory regions are known to affect the structure and expression of milk protein (Martin et al., 2002; Szymanowska et al., 2003; Cosenza et al., 2007, 2016, 2018). Therefore, in the present study, we decided to characterize and compare the proximal gene promoter and 5′ untranslated region (**UTR**; about 900 nt) of the 2 investigated alleles.

AliBaba v2.1 program was used to identify the potential binding sites of TF that could affect the gene expression revealed that the TATA box (TATATA) is located between the nucleotides −71/−66. In addition, we identified the following TF: CACCC elements flanked by clusters of overlapping specificity protein 1 sites; activating protein 2; transcription factor IID; CCAAT/enhancer-binding protein α; retinoic acid receptors subtypes α, β; Yin Yang transcription factor-1, nuclear factor I, and milk protein binding factor, enhancer box (E-box; Supplemental Figure S1).

The sequence comparison of the 5′ flanking region showed only the substitution g.269C > G, whereas the 5′ UTR is more variable due to the presence of 3 polymorphic sites: g.458G > A, g.496A > G and g.920–922CAC > TGT (Supplemental Figures S1 and S2, Table 1).

Table 1. Polymorphisms detected at donkey *BLG* I locus and comparison with horse and zebra gene counterpart sequences

	Equus asinus breed Amiatina[1]		*Equus asinus* breed Amiatina[2]		*Equus asinus* African ass[3]	*Equus asinus* breed Guanzhong[4]	*Equus asinus* breed Dezhou[5]	*Equus caballus* breed Mongolian[6]	*Equus quagga*[7]
Description	Mutation	Location	Mutation	Location	Mutation	Mutation	Mutation	Mutation	Mutation
5′ UTR	C	269	G	269	T	C	C	G	C
Exon 1	G	458	A	458	T	A	A	A	G
	A	496	G	496	G	A	G	G	A
	CAC	920–922	TGT	920–922	TGT	CAC	TGT	CAC	CAC
Intron 1	C	1067	C	1067	C	C	T	T	C
	G	1138	T	1138	G	G	G	G	G
	—	1203–1204	TGGAGC	1204–1209	TGGAGC	TGGAGC	TGGAGC	TGGAGC	TGGAGC
	C	1219	C	1225	T	T	C	C	C
	C	1339	T	1345	C	T	C	C	C
	T	1618	C	1624	C	T	C	T	T
	A	1716	C	1722	C	A	C	A	A
	T	1798	G	1804	G	G	G	T	T
Exon 2	C	1863	C	1869	T	C	C	C	C
	G	1871	A	1877	G	G	G	G	G
Intron 2	C	1992	T	1998	C	T	T	C	C
	T	2006	T	2012	C	T	T	T	T
	A	2166	A	2172	T	A	A	A	A
	G	2504	A	2510	G	G	G	A	G
	C	2588	T	2594	T	T	T	C	C
	—	2589–2590	—	2595–2596	—	—	T	T	T
Intron 3	—	3364–3365	—	3370–3371	—	—	A	—	—
	T	3390	T	3398	T	T	C	C	C
	A	3418	A	3424	A	A	G	G	G
	T	3613	C	3619	C	T	C	C	C
	G	3797	T	3803	T	G	T	G	G
Intron 5	A	4803	G	4809	G	G	G	G	G
	C	4853	T	4859	T		T	T	T
	CCCCCCC	4934–4940	CCCCCCC	4940–4946	CCCCCC		CCCCCCCC	CCCCCGCC	CCCCCCCCCC
Intron 6	C	5064	T	5070	T		C	C	C
	C	5164	C	5170	T		—	G	G
	C	5235	T	5241	C		T	C	T
	C	5236	G	5242	C		G	C	G
	C	5282	A	5288	A		C	A	A
Exon 7	CCTC	5524–5529	CCTC	5530–5535	CCTC		CCTCCCTC	CCTC	CCTC
	A	5679	A	5685	T		C	C	C

[1] Present work; GenBank ON886232.

[2] Present work; GenBank ON886233.

[3] GenBank PSZQ01002145.

[4] GenBank JREZ01001480, NW_014637182, XM_014838039.1.

[5] GenBank JADWZW010000011.

[6] GenBank PJAA01000026.1, ATDM01064188.1.

[7] GenBank JAKJSB010000001.1. UTR: untranslated region. The long dashes in the table indicate no mutation present (deletion/insertion). The absence of long dashes in the bottom of the Guanzhong column indicates no sequence available.

The SNPs in position 269 and 496 do not affect known TF and, consequently, no influence on gene expression was expected; whereas, the remaining 2 mutations may play a role in determining a differential gene expression. In fact, the presence of a G in position 458 would determine the disappearance of a binding site for the upstream stimulatory factor (**USF**). The mutation g.920-922CAC > TGT localized 4 nucleotides upstream the Kozak consensus sequence is responsible for the creation of a binding site for the eukaryotic initiation factor (**eIF**), whereas the repressor activator protein 1 (Rap1) and E-box sites disappear.

Such factors are known to be involved in gene transcriptional regulation. The USF binds to classical E-box elements and regulates the transcription through the recruitment of co-regulator complexes. The USF1 plays an important role in transcriptional regulation of a large number of seemingly unrelated genes (Corre and Galibert, 2005; Rada-Iglesias et al., 2008), consistent with the abundant distribution of E-box like elements in the genome. The USF1 binding signals strongly correlate with target gene expression levels, suggesting that USF1 plays an important role in transcription activation. The USF1 physically interacts with histone modifying enzymes, transcription preinitiation complex factors, coactivator, and corepressor proteins (Corre and Galibert, 2005; Huang et al., 2007; Corre et al., 2009). In addition, USF1 interacts with other TF to achieve cooperative transcriptional activation of individual genes (Corre and Galibert, 2005).

The Rap1 prevents initiation of divergent noncoding transcription near its binding sites and provides directionality toward productive transcription and its rapid depletion leads to widespread induction of divergent transcripts (Wu et al., 2018). In general, loss of Rap1 at Rap1 binding promoters more commonly leads to significantly decreased transcription (Yarragudi et al., 2007).

The eIF are protein complexes involved in the initiation phase of translation. Decreasing wild-type eIF1 abundance reduces initiation accuracy, whereas overexpressing eIF1 suppresses initiation at near cognates or AUG in poor context (Martin-Marcos et al., 2017). In particular, eIF1 plays a dual role in the scanning mechanism. Therefore, a deficiency in eIF1 decreases the fidelity of translation initiation (Visweswaraiah et al., 2015).

Genotyping

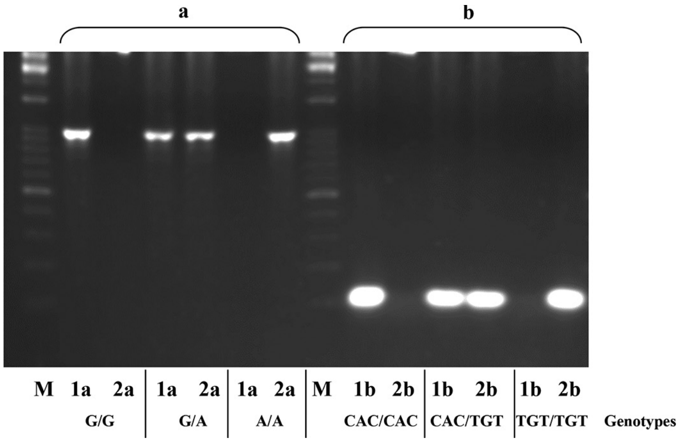

Figure 1. Identification by allele-specific PCR of the donkey carriers of (a) g.1871G > A and (b) g.920–922CAC > TGT mutations at *BLG* I locus. The marker (M) used is the 2-log DNA ladder (0.1–10 kb; Biolabs) 1a, "g.1871G" allele specific primer; 2a, "g.1871A" allele specific primer 1b, "g.920-922CAC" allele specific primer; 2b, "g.920-922TGT" allele specific primer.

To verify whether the SNPs at positions 269, 458, and 496 were associated with the different protein isoforms, the 5′ flanking region of the *BLG* I gene was sequenced for the 6 heterozygous donkeys carrying the new variant and 6 homozygous wild type donkeys randomly selected. The comparison of the sequences showed that the electrophoretic band of 20,428.50 Da was always in *cis* with the SNPs g.269C > G, g.458G > A and g.496A > G, but these mutations have been associated also with the electrophoretic band of 18,514.25 Da. Therefore, these polymorphisms do not seem to be involved in the generation of different donkey β-LG I isoforms.

To estimate the frequencies at the 2 polymorphic sites g.920–922CAC > TGT and g.1871G > A, and to determine the possible haplotypes, specific genotyping AS-PCR protocols have been developed (Figure 1). The genotype distributions and the allelic frequencies of the 2 polymorphisms, determined in all 93 investigated donkeys are reported in Table 2.

The g.920–922CAC was always in *cis* with the band of 20,428.50 Da, but also associated with the electrophoretic band of 18,514.25 Da. On the contrary, the subjects whose milk produced an SDS PAGE pattern with only the 20,428.50 Da band were always homozygous for g.1871A. All samples with a single electrophoretic band of 18,514.25 Da (27 Amiatina, 59 Ragusana samples) were homozygotes for g.1871G. The 6 donkeys with both bands in SDS PAGE pattern were heterozygous (g.1871A/G). The overall frequencies of g.920–922CAC and g.1871A were 0.3065 and 0.043 respectively, with slight differences between the 2 breeds and a genotype distribution consistent with Hardy–Weinberg's law at the level of significance of 0.05 (Table 2).

Using Haploview software version 4.2, only 3 different allelic combinations (out of the 4 expected) were observed: haplotypes 1 (920–922TGT/1871G), 2

Table 2. Genotyping data, allele frequency, relative frequencies of the polymorphisms g.920–922CAC > TGT and g.1871G > A of the *BLG* I gene in the Amiatina and Ragusana donkey population[1]

		Genotype distribution							Allelic frequency			
		g.920_922TGT > CAC							g.1871		g.920_922	
Polymorphism	Item	CAC/CAC	CAC/TGT	TGT/TGT	Obs.	Exp.	SDS-PAGE[2]	χ^2	G	A	CAC	TGT
g.1871G > A	G/G	10	35	41	86	85.172	+ +	4.3508	0.957	0.043	0.3065	0.6935
	G/A	—	2	4	6	7.6559	+ −					
	A/A	—	—	1	1	0.172	− −					
	Obs.	10	37	46	93							
	Exp.	11.7097	42.5806	38.7097								
	χ^2	0.3816										

[1] Yate's chi-squared value (χ^2) = 0.80095, Yate's *P*-value = 0.67. Obs. = observed; Exp. = expected.

[2] + + samples showed the presence of a β-LG I band displaying an electrophoretic mobility consistent with a molecular weight of 18,514.25 Da; − − sample with only a band with an electrophoretic mobility higher than regular band (20,428.50 Da); + − samples with both electrophoretic bands.

(920–922CAC/1871G), and 3 (920–922TGT/1871A). The first haplotype was the most represented with a frequency of 0.651, followed by the haplotypes 2 (0.306) and 3 (0.043). Because only the SNP g.1871A > G is associated with the different observed phenotypes, likely this mutation is involved in gene regulation processes.

Allele Nomenclature and Phylogenetic Relationship Among the Markers

Considering all 5 AA changes (gene markers in exons 2, 4, and 5) as revealed by the database analysis, the literature, the newly determined in the present study and the 3 observed haplotypes, it is possible to deduce at least 5 different alleles (*BLG* I A, B, B1, C, D) responsible for 4 potential β-LG I translations (Figure 2).

In detail, the variant A and B have already been well characterized at the protein level respectively by Godovac-Zimmermann et al. (1988) for a Sardinian donkey (Italy) and by Herrouin et al. (2000) for French donkeys of the "Commune" breed. The 2 sequences show a homology of 98.1% as a consequence of 3 AA substitutions: p.Ser36 > Glu, p.Thr97 > Ser and p.Ile150 > Val. Taken together, these 3 AA substitutions explained the mass difference of about 14 Da observed for the entire protein (18,514.25 vs. 18,528.23 Da; Herrouin et al., 2000).

The *BLG* I B1 (haplotype 1) is a silent allele because it differs from the *BLG* I B (haplotype 2) for the substitution g.920–922CAC > TGT in the 5′ untranslated region. The presence of $\text{CAC}^{920–922}$ should represent the ancestral condition of the gene as constitutive of *BLG* I and *BLG* II sequences in mares (GenBank accession no. PJAA01000026.1, region: 38874907 to 38880546, complement; AF107201.1) and zebras (GenBank accession no. JAKJSB010000001.1, region: 96008038 to 96002141; HM012800.1).

The *BLG* I B1 allele, wild type allele characterized in the present study, is considered as the ancestral form according to nucleotide and AA sequences available for equidae. No cDNA or gene sequence is available for the variant B (Herrouin et al., 2000). Therefore, currently it is not possible to establish the link with the variant B1.

The further allele that we named *BLG* I C derived by the subsequent comparison between the nucleotide sequences of this study and that of an African ass reared in the Copenhagen zoo (Denmark; PSZQ01002145.1, region: 1252372 to 1258209). This allele is characterized by the SNP g.1863C > T located at the 32th nucleotide of exon 2 (Table 1) responsible for the AA substitutions p.Ala25 > Val in the mature protein (deduced *Mr*: 18,542.30 Da). The allele C carries the same haplotype 1 of the allele B; therefore, it is likely

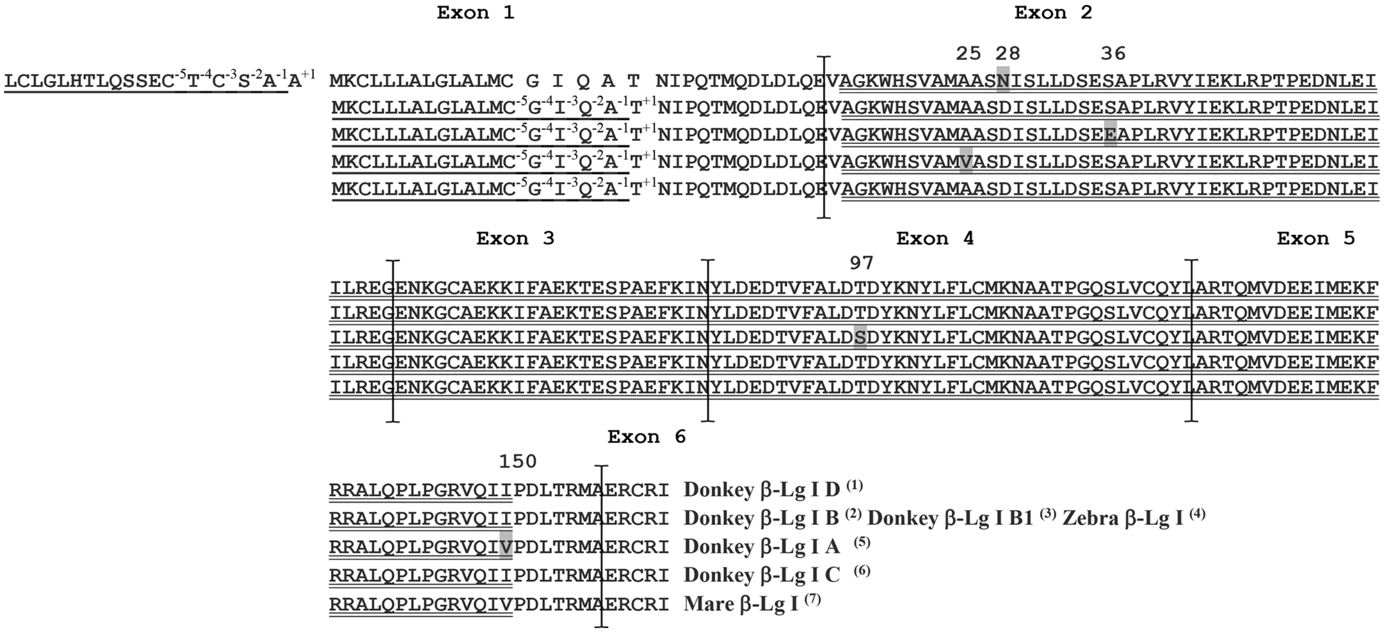
Exon 1
Exon 2
25 28 36
LCLGLHTLQSSEC^{-5}T^{-4}C^{-3}S^{-2}A^{-1}A^{+1} MKCLLLALGLALMC G I Q A T NIPQTMQDLDLQEVAGKWHSVAMAASNISLLDSESAPLRVYIEKLRPTPEDNLEI
MKCLLLALGLALMC^{-5}G^{-4}I^{-3}Q^{-2}A^{-1}T^{+1}NIPQTMQDLDLQEVAGKWHSVAMAASDISLLDSESAPLRVYIEKLRPTPEDNLEI
MKCLLLALGLALMC^{-5}G^{-4}I^{-3}Q^{-2}A^{-1}T^{+1}NIPQTMQDLDLQEVAGKWHSVAMAASDISLLDSEEAPLRVYIEKLRPTPEDNLEI
MKCLLLALGLALMC^{-5}G^{-4}I^{-3}Q^{-2}A^{-1}T^{+1}NIPQTMQDLDLQEVAGKWHSVAMVASDISLLDSESAPLRVYIEKLRPTPEDNLEI
MKCLLLALGLALMC^{-5}G^{-4}I^{-3}Q^{-2}A^{-1}T^{+1}NIPQTMQDLDLQEVAGKWHSVAMAASDISLLDSESAPLRVYIEKLRPTPEDNLEI
Exon 3
Exon 4
Exon 5
97
ILREGENKGCAEKKIFAEKTESPAEFKINYLDEDTVFALDTDYKNYLFLCMKNAATPGQSLVCQYLARTQMVDEEIMEKF
ILREGENKGCAEKKIFAEKTESPAEFKINYLDEDTVFALDTDYKNYLFLCMKNAATPGQSLVCQYLARTQMVDEEIMEKF
ILREGENKGCAEKKIFAEKTESPAEFKINYLDEDTVFALDSDYKNYLFLCMKNAATPGQSLVCQYLARTQMVDEEIMEKF
ILREGENKGCAEKKIFAEKTESPAEFKINYLDEDTVFALDTDYKNYLFLCMKNAATPGQSLVCQYLARTQMVDEEIMEKF
ILREGENKGCAEKKIFAEKTESPAEFKINYLDEDTVFALDTDYKNYLFLCMKNAATPGQSLVCQYLARTQMVDEEIMEKF
Exon 6
150
RRALQPLPGRVQIIPDLTRMAERCRI Donkey β-Lg I D (1)
RRALQPLPGRVQIIPDLTRMAERCRI Donkey β-Lg I B (2) Donkey β-Lg I B1 (3) Zebra β-Lg I (4)
RRALQPLPGRVQIVPDLTRMAERCRI Donkey β-Lg I A (5)
RRALQPLPGRVQIIPDLTRMAERCRI Donkey β-Lg I C (6)
RRALQPLPGRVQIVPDLTRMAERCRI Mare β-Lg I (7)

Figure 2. Amino acid comparison of different donkey β-LG I variants (observed or deduced) and comparison with mare and zebra counterpart. Peptide leader encoding regions are underlined and numbered according to the rules proposed by von Heijne (1986). The double-underlined AA sequences correspond to the structurally conserved regions typical of a lipocalin core. Amino acid substitutions are highlighted in gray. (1) Present work, National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov/) reference sequence: ON886233; (2) NCBI reference sequence: XM_014838039.1; (3) Present work, NCBI reference sequence: ON886232; NCBI reference sequence: XM_044778355.1; (4) NCBI reference sequence: XP_046531205.1; (5) UniProtKB/Swiss-Prot: P13613.1; (6) NCBI reference sequence: PSZQ01002145.1, region: 1252372 to 1258209; (7) NCBI reference sequence: NP_001075962.1.

derived by the latter. The p.Val25 was found also in the sequences of monomeric β-LG as reported for donkey β-LG II (GenBank acc nos. HM012800.1; HM012799.1) and equine β-LG II (GenBank acc nos. AF107201.1; NM_001082494.1), while for all mammalian species the presence of alanine occurs at this position both in β-LG I and β-LG II.

Finally, the *BLG* I D allele corresponds to the variant described in the present study characterized by the Asn in position 28 (haplotype 3). Among the 5 alleles, the *BLG* I D is of real novel identification because this amino acid substitution was never reported earlier in databases. In fact, all monomeric and dimeric β-LG sequences are characterized by Asp at this position and, therefore, it has to be considered as the ancestral condition of these proteins.

However, the substitution p.Asp28 > Asn increases the protein mass of a single Da and does not justify the difference in the molecular weight (18,514.25 vs. 20,428.5 Da) observed in the SDS-PAGE of the present study and by Auzino et al. (2022).

In Figure 3 we proposed a possible phylogenetic map where X indicates a putative variant derived from the comparison with the sequence of mare β-LG I (GenBank accession no. NP_001075962.1; Conti et al., 1984) and considered as the ancestral form of this protein in donkey.

mRNA Secondary Structure

To understand the molecular mechanisms possibly responsible for the different protein isoform of the donkey *BLG* I D allele, we carried out simulations with RNAfold to investigate possible correlations between computed *BLG* I mRNA secondary structures, together with their corresponding minimum free energies.

RNAfold analysis was achieved using the deduced mRNA sequences starting at 319 nucleotides from the transcription start site and ending 13 nucleotides downstream of the poly-A site. Each sequence was 1,096 nucleotides long.

A first approach was to compare the minimum free energy structures of the polymorphic sequences. This comparison showed that only the SNP g.1871G > A is responsible for a change in the energy landscape. In fact, the presence of the adenine leads to a change in minimum free energy (least negative value) compared with the presence of the guanine: −460.10 kcal/mol for the haplotype 3 (920–922TGT/1871A) and −459.70 kcal/mol for haplotype 4 (920–922CAC/1871A, not

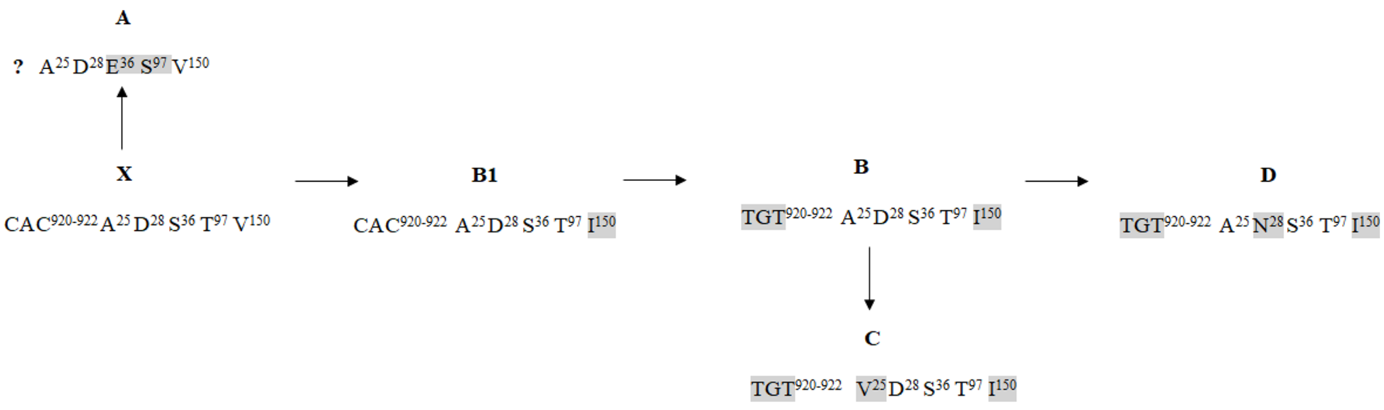
A
? A25 D28 E36 S97 V150
X
CAC920-922 A25 D28 S36 T97 V150
B1
CAC920-922 A25 D28 S36 T97 I150
B
TGT920-922 A25 D28 S36 T97 I150
D
TGT920-922 A25 N28 S36 T97 I150
C
TGT920-922 V25 D28 S36 T97 I150

Figure 3. Possible evolutionary pathway of donkey β-LG I encoding gene (*BLG* I) and differences between the corresponding variants.

observed) versus −461.90 kcal/mol for haplotypes 1 (920–922TGT/1871G) and 2 (920–922CAC/1871G; Figure 4).

These values would suggest that the G in position 1871 could provide a translational benefit. In fact, it is well known that among a series of same-sized RNAs, the one with the most negative value of free energy is considered the most structured, stable and associated with more efficient translation rates (McClements et al., 2021).

In addition, the local analyses of mRNA secondary structure showed a conformational change always as consequence of g.1871G > A polymorphism (Figure 4). Structural changes including extensively different stem-loops and bubbles were noticed in some regions of variant mRNA that may affect its stability. In particular, it is observed a conformational change in the region carrying the entire Kozak consensus sequence (GCAGCC<u>ATG</u>A) with a different secondary structure formation in the 5′ UTR of haplotypes 1 and 2 versus the haplotype 3 (Figure 4). The conservation of the consensus Kozak sequence (GCCRCCATGG) is crucial to maintain the rate of translation initiation and the disruption of this motif can lead to impairment in ribosome mRNA recognition and binding. In particular, within this motif, the purine in position −3 is the most highly conserved and functionally most important position. If the start codon does not have a purine at −3 position, then this sequence is in a weak context (Kozak, 2002).

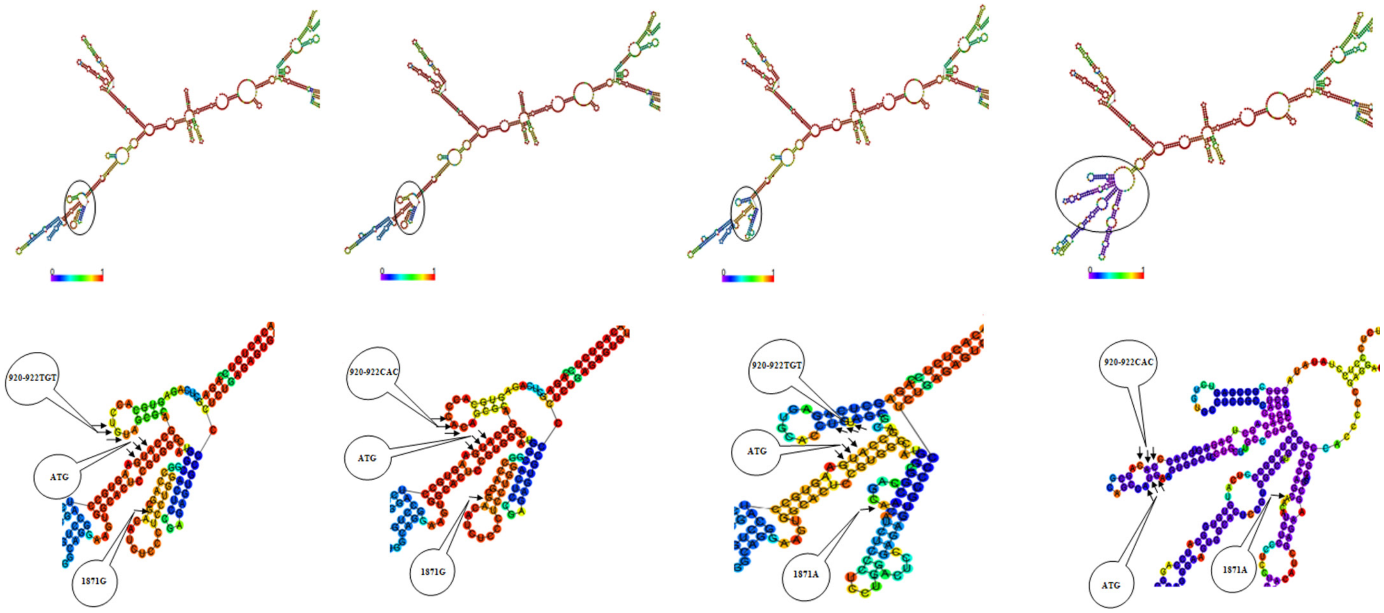

Figure 4. Secondary structure predictions (minimum free energy secondary structure) at the nucleotides immediately surrounding the start codon of donkey *BLG* I transcripts. Haplotype 4 was not observed.

When the AUG$^{\mathrm{START}}$ codon occurs in a strong context, all or almost all ribosomes stop scanning the mRNA sequence and initiate the translation at that point; on contrary, when the AUG resides in a very weak context few ribosomes initiate the translation at that point, but most continue scanning the mRNA and initiate the translation further downstream. This is known as leaky scanning (Kozak, 2002). By the same principle, alternative AUG codons upstream the optimal one or non-AUG codons (e.g., CUG, GUG, and UUG) allow to initiate translation although in a weaker way (Kearse and Wilusz, 2017). Alternative translation starts are common events perhaps because mRNA secondary structure slows scanning and gives more time for the mismatched codon to pair with Met-tRNAi (Kozak, 2002).

Thanks to recent developments of high-throughput technologies for translation genome-wide studies, alternative translation initiation and non-AUG initiation are known to be widespread in eukaryotes. The use of several translation initiation codons in a single mRNA, by expressing several proteins from a single gene, contributes to the generation of new proteins harboring different amino terminal domains that may potentially confer to these isoforms distinct functions (Touriol et al., 2003; Kearse and Wilusz, 2017; Xu and Zhang, 2020). Generally, the evolution tends to eliminate events of alternative translation initiation. However, many alternative ATG codons still appear near the beginning of the open reading frames. It is possible that these alternative start codons still exist because the fitness advantage of eliminating these codons is not high enough. An additional possibility is that at least some of the resultant alternative proteins from these alternative peptides are functional (Zur and Tuller, 2013).

Based on these considerations, the adenine at the 40th nucleotide of exon 2 of the donkey *BLG* I gene changes the mRNA secondary structures, increases the minimum free energies and, therefore, might be responsible for an alternative translation initiation and, consequently, for the higher molecular weight of the β-LG I encoded by the allele D.

Starting with the first in frame alternative start codon (CUG upstream the canonical AUG), the 5′ UTR sequence of the allele D provides additional 54 nt of coding sequence (all within the first exon) that would translate to further 18 AA (LCLGLHTLQSSECTCSAA) at the N terminus of the immature β-LG. The first 17 AA might lead to a new signal peptide of a newly synthesized protein that is direct toward the secretory pathway.

This assumption is supported by the fact that the region preceding the cleavage site of this new signal peptide, which constitutes the substrate recognition site for the signal peptidase enzyme, conforms to the rules proposed by von Heijne (1986). Indeed, it is now well known that in bacteria and eukaryotic cells, preproteins have a common pattern in the c-region of the signal peptide. The residue in position −1 must be small (i.e., Ala, Ser, Gly, Cys, Thr, or Gln); the residue in position −3 must not be aromatic (Phe, His, Tyr, Trp), charged (Asp, Glu, Lys, Arg), or large and polar (Asn, Gln), but residues tolerated at this position are Ala, Gly, Ser, Cys, Ile, Val, and Leu. Residues at the +1, −2, −4, and −5 positions were not generally critical, as almost any residue was tolerated. Further, it was suggested that Pro must be absent from positions −3 through +1 (Paetzel et al., 2002). Therefore, the recognition of a new signal peptide, the extension of 19 AA sequence of the mature protein, corresponding to the canonical signal peptide (Figure 2), with an additional Ala residue and the presence of Asn in position 28 is sufficient to provide the observed molecular weight (20,428.50 Da) of the slower migrating β-LG I D form (181 vs. 162 AA). These findings are a first hint that structural changes induced by a SNP may influence the donkey *BLG* I mRNA initiation efficiency and fidelity and, thereby, lead to proteins with different molecular weights.

It has been well documented that mutations in consensus sites in genes related to milk traits in ruminants and equines may alter or create essential sequence elements for splicing, processing, or translation of mRNA. These mutations are associated with altered length or steady-state level of cytoplasmic mRNA or different gene expression (Ramunno et al., 2005; Cosenza et al., 2009, 2016, 2017, 2018; Giambra et al., 2010; Balteanu et al., 2013; Gu et al., 2020). On the other hand, SNPs/indels that apparently do not affect RNA consensus could also lead to phenotypic effects, through mechanisms nonconsensus-dependent. Different studies documented that mutations in both coding and noncoding regions of DNA can potentially act in mildly deleterious and, in some cases, pathological fashion on pre- and post-translational levels through changes in RNA structure and stability, thus representing a powerful tool to study their effect on gene expression. For example, a considerable change in the secondary structure of the goat *SCD*1 mRNA was observed by an in silico analysis for samples carrying the c.*1902_1904delTGT polymorphism (Zidi et al., 2010). These authors hypothesized that such a mutation has causal effect on milk polyunsaturated and conjugated linoleic fatty acid levels by altering the amount of *SCD*1 transcripts in mammary epithelial cells.

Likewise, the lower amount of α_{S1}-CN observed for cattle *CSN1S1* G and goat *CSN1S1* E alleles, respectively, was explained by a reduced mRNA stability as consequence of a long interspersed nuclear element A+U rich insertion in their 3′ UTR region that affected the amount of mRNA free energy (Jansà Pérez et al., 1994; Rando et al., 1998). Furthermore, it has been also hypothesized that the low level of lactoferrin in bovine milk compared with that in human milk may be a direct consequence of the different mRNA nucleotide sequence responsible for a stronger secondary structure and a higher free energy in bovine (Schanbacher et al., 1993).

More recently, Erhardt et al. (2016) reported that an insertion of 11 bp at intron 17 negatively affects the secondary structure of the pre-mRNA of *CSN1S1* A and *CSN1S1* C alleles in camel. As consequence, these variants lack the exon 18 compared with *CSN1S1* B variant, thus changing the α_{S1}-CN peptide pattern, affecting the IgE-binding epitopes and altering the availability of bioactive peptides of the variants. These results are similar to what was previously reported for goat and sheep *CSN1S1* (Leroux et al., 1992; Passey et al., 1996) and human *CSN2* (coding for β-CN; Martin and Leroux 1992).

CONCLUSIONS

This study provides the first contribution to the full characterization of the genomic sequence of the donkey β-LG I encoding gene. Based on the detected genetic variability, 5 different alleles responsible for 4 potential β-LG I translations have been defined. In particular, the SNP g.1871G > A, responsible for the p.Asp28 > Asn AA substitution in the mature protein, has marked effects on the mRNA structural folds. This leads to the translation of a longer β-LG I form (181 vs. 162 AA) as a consequence of the recognition of an alternative initiation site leading to the extension of 19 AA of the mature protein, corresponding to the canonical signal peptide with an additional N-terminal alanine residue. To our knowledge, this study is the first to report the characterization of such an exceptional molecular event for a milk protein encoding gene, which should be investigated in further studies for its effects on donkey milk traits, including the related impact on human consumers.

ACKNOWLEDGMENTS

The authors gratefully acknowledge financial support provided by the Italian Ministry of Agricultural and Forestry Policies (SELMOL project). The author contributions included the following: G. Cosenza and P. Martin conceived and designed the experiments; Auzino, D. Gallo and G. Garro performed the experiments; G. Cosenza, A. Pauciullo and P. Martin analyzed the data; G. Garro, and R. Ciampolini contributed reagents, materials, and analysis tools; A. Pauciullo and R. Ciampolini provided milk and DNA samples; G. Cosenza, A. Pauciullo, and P. Martin wrote the paper; G. Cosenza, A. Pauciullo, and P. Martin revised the article critically for important intellectual content; G. Cosenza, A. Pauciullo, P. Martin, G. Garro, D. Gallo, R. Ciampolini, and B. Auzino gave final approval of the version to be published. The authors have not stated any conflicts of interest.

REFERENCES

Auzino, B., G. Miranda, C. Henry, Z. Krupova, M. Martini, F. Salari, G. Cosenza, R. Ciampolini, and P. Martin. 2022. Top-down proteomics based on LC-MS combined with cDNA sequencing to characterize multiple proteoforms of Amiata donkey milk proteins. Food Res. Int. 160:111611. https://doi.org/10.1016/j.foodres.2022.111611.

Balteanu, V. A., T. C. Carsai, and A. Vlaic. 2013. Identification of an intronic regulatory mutation at the buffalo α_{S1}-casein gene that triggers the skipping of exon 6. Mol. Biol. Rep. 40:4311–4316. https://doi.org/10.1007/s11033-013-2518-2.

Brumini, D., A. Criscione, S. Bordonaro, G. E. Vegarud, and D. Marletta. 2016. Whey proteins and their antimicrobial properties in donkey milk: A brief review. Dairy Sci. Technol. 96:1–14. https://doi.org/10.1007/s13594-015-0246-1.

Chianese, L., M. G. Calabrese, P. Ferranti, R. Mauriello, G. Garro, C. De Simone, M. Quarto, F. Addeo, G. Cosenza, and L. Ramunno. 2010. Proteomic characterization of donkey milk "caseome". J. Chromatogr. A 1217:4834–4840. https://doi.org/10.1016/j.chroma.2010.05.017.

Conti, A., J. Godovac-Zimmermann, J. Liberatori, G. Braunitzer, and D. Minori. 1984. The primary structure of monomeric β-lactoglobulin I from horse colostrum (*Equus caballus*, Perissodactyla). Hoppe Seylers Z. Physiol. Chem. 365:1393–1401. https://doi.org/10.1515/bchm2.1984.365.2.1393.

Corre, S., and M. D. Galibert. 2005. Upstream stimulating factors: highly versatile stress-responsive transcription factors. Pigment Cell Res. 18:337–348. https://doi.org/10.1111/j.1600-0749.2005.00262.x.

Corre, S., A. Primot, Y. Baron, J. Le Seyec, C. Goding, and M. D. Galibert. 2009. Target gene specificity of USF-1 is directed via p38-mediated phosphorylation-dependent acetylation. J. Biol. Chem. 284:18851–18862. https://doi.org/10.1074/jbc.M808605200.

Cosenza, G. 2023. Replication Data for: A novel allelic donkey b-Lg I protein isoform generated by a non-AUG translation initiation codon is associated with a non-synonymous SNP. Harvard Dataverse, V1. https://doi.org/10.7910/dvn/dlfgkd.

Cosenza, G., R. Ciampolini, M. Iannaccone, D. Gallo, B. Auzino, and A. Pauciullo. 2018. Sequence variation and detection of a functional promoter polymorphism in the lysozyme c-type gene from Ragusano and Grigio Siciliano donkeys. Anim. Genet. 49:270–271. https://doi.org/10.1111/age.12647.

Cosenza, G., M. Iannaccone, B. A. Pico, D. Gallo, R. Capparelli, and A. Pauciullo. 2017. Molecular characterisation, genetic variability and detection of a functional polymorphism influencing the promoter activity of OXT gene in goat and sheep. J. Dairy Res. 84:165–169. https://doi.org/10.1017/S0022029917000097.

Cosenza, G., M. Iannaccone, B. A. Pico, L. Ramunno, and R. Capparelli. 2016. The SNP g1311T>C associated with the absence of β-casein in goat milk influences CSN2 promoter activity. Anim. Genet. 47:615–617. https://doi.org/10.1111/age.12443.

Cosenza, G., R. Mauriello, G. Garro, B. Auzino, M. Iannaccone, A. Costanzo, L. Chianese, and A. Pauciullo. 2019. Casein composition and differential translational efficiency of casein transcripts in donkey's milk. J. Dairy Res. 86:201–207. https://doi.org/10.1017/S0022029919000256.

Cosenza, G., A. Pauciullo, L. Colimoro, A. Mancusi, A. Rando, D. Di Berardino, and L. Ramunno. 2007. An SNP in the goat CSN2 promoter region is associated with the absence of β-casein in the milk. Anim. Genet. 38:655–658. https://doi.org/10.1111/j.1365-2052.2007.01649.x.

Cosenza, G., A. Pauciullo, M. Feligini, A. Coletta, L. Colimoro, D. Di Berardino, and L. Ramunno. 2009. A point mutation in the splice donor site of intron 7 in the αs2-casein encoding gene of the Mediterranean River buffalo results in an allele-specific exon skipping. Anim. Genet. 40:791. https://doi.org/10.1111/j.1365-2052.2009.01897.x.

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

Gabriel, S. B., S. F. Schaffner, H. Nguyen, J. M. Moore, J. Roy, B. Blumenstiel, J. Higgins, M. DeFelice, A. Lochner, M. Faggart, S. N. Liu-Cordero, C. Rotimi, A. Adeyemo, R. Cooper, R. Ward, E. S. Lander, M. J. Daly, and D. Altshuler. 2002. The structure of haplotype blocks in the human genome. Science 296:2225–2229. https://doi.org/10.1126/science.1069424.

Giambra, I. J., L. Chianese, P. Ferranti, and G. Erhardt. 2010. Genomics and proteomics of deleted ovine CSN1S1*I. Int. Dairy J. 20:195–202. https://doi.org/10.1016/j.idairyj.2009.09.005.

Godovac-Zimmermann, J., A. Conti, L. James, and L. Napolitano. 1988. Microanalysis of the amino-acid sequence of monomeric beta-lactoglobulin I from donkey (*Equus asinus*) milk. The primary structure and its homology with a superfamily of hydrophobic molecule transporters. Biol. Chem. Hoppe Seyler 369:171–179. https://doi.org/10.1515/bchm3.1988.369.1.171.

Godovac-Zimmermann, J., A. Conti, M. Sheil, and L. Napolitano. 1990. Covalent structure of the minor monomeric beta-lactoglobulin II component from donkey milk. Biol. Chem. Hoppe Seyler 371:871–879. https://doi.org/10.1515/bchm3.1990.371.2.871.

Goossens, M., and Y. W. Kan. 1981. DNA analysis in the diagnosis of hemoglobin disorders. Methods Enzymol. 76:805–817. https://doi.org/10.1016/0076-6879(81)76159-7.

Grosclaude, F., M. F. Mahé, G. Brignon, L. Di Stasio, and R. Jeunet. 1987. A Mendelian polymorphism underlying quantitative variations of goat αs1-casein. Genet. Sel. Evol. (1983) 19:399–412. https://doi.org/10.1186/1297-9686-19-4-399.

Gruber, A. R., R. Lorenz, S. H. Bernhart, R. Neuböck, and I. L. Hofacker. 2008. The Vienna RNA websuite. Nucleic Acids Res. 36(Web Server):W70–W74. https://doi.org/10.1093/nar/gkn188.

Gu, M., G. Cosenza, G. Gaspa, M. Iannaccone, N. P. P. Macciotta, G. Chemello, L. Di Stasio, and A. Pauciullo. 2020. Sequencing of lipoprotein lipase gene in the Mediterranean river buffalo identified novel variants affecting gene expression. J. Dairy Sci. 103:6374–6382. https://doi.org/10.3168/jds.2019-17968.

Halliday, J. A., K. Bell, K. McAndrew, and D. C. Shaw. 1993. Feline beta-lactoglobulins I, II and III, and canine beta-lactoglobulins I and II: amino acid sequences provide evidence for the existence of more than one gene for beta-lactoglobulin in the cat and dog. Protein Seq. Data Anal. 5:201–205.

Herrouin, M., D. Mollé, J. Fauquant, F. Ballestra, J. L. Maubois, and J. Léonil. 2000. New genetic variants identified in donkey's milk whey proteins. J. Protein Chem. 19:105–116. https://doi.org/10.1023/A:1007078415595.

Huang, S., X. Li, T. M. Yusufzai, Y. Qiu, and G. Felsenfeld. 2007. USF1 recruits histone modification complexes and is critical for maintenance of a chromatin barrier. Mol. Cell. Biol. 27:7991–8002. https://doi.org/10.1128/MCB.01326-07.

Işık, R. 2019. The Identification of novel single-nucleotide polymorphisms of equine beta-lactoglobulin and lactotransferrin genes. J. Equine Vet. Sci. 75:60–64. https://doi.org/10.1016/j.jevs.2019.01.005.

Jensen-Jarolim, E., L. F. Pacios, R. Bianchini, G. Hofstetter, and F. Roth-Walter. 2016. Structural similarities of human and mammalian lipocalins, and their function in innate immunity and allergy. Allergy 71:286–294. https://doi.org/10.1111/all.12797.

Kearse, M. G., and J. E. Wilusz. 2017. Non-AUG translation: A new start for protein synthesis in eukaryotes. Genes Dev. 31:1717–1731. https://doi.org/10.1101/gad.305250.117.

Kontopidis, G., C. Holt, and L. Sawyer. 2004. Invited review: β-Lactoglobulin: Binding properties, structure, and function. J. Dairy Sci. 87:785–796. https://doi.org/10.3168/jds.S0022-0302(04)73222-1.

Kozak, M. 2002. Pushing the limits of the scanning mechanism for initiation of translation. Gene 299:1–34. https://doi.org/10.1016/S0378-1119(02)01056-9.

Le Maux, S., S. Bouhallab, L. Giblin, A. Brodkorb, and T. Croguennec. 2014. Bovine β-lactoglobulin/fatty acid complexes: binding, structural, and biological properties. Dairy Sci. Technol. 94:409–426. https://doi.org/10.1007/s13594-014-0160-y.

Leroux, C., N. Mazure, and P. Martin. 1992. Mutations away from splice site recognition sequences might cis-modulate alternative splicing of goat αs1-casein transcripts. Structural organization of the relevant gene. J. Biol. Chem. 267:6147–6157. https://doi.org/10.1016/S0021-9258(18)42674-9.

Martin, P., and C. Leroux. 1992. Exon-skipping is responsible for the 9 amino acid residue deletion occurring near the N-terminal of human beta-casein. Biochem. Biophys. Res. Commun. 183:750–757. https://doi.org/10.1016/0006-291X(92)90547-X.

Martin, P., M. Szymanowska, L. Zwierzchowski, and C. Leroux. 2002. The impact of genetic polymorphisms on the protein composition of ruminant milks. Reprod. Nutr. Dev. 42:433–459. https://doi.org/10.1051/rnd:2002036.

Martin-Marcos, P., F. Zhou, C. Karunasiri, F. Zhang, J. Dong, J. Nanda, S. D. Kulkarni, N. D. Sen, M. Tamame, M. Zeschnigk, J. R. Lorsch, and A. G. Hinnebusch. 2017. eIF1A residues implicated in cancer stabilize translation preinitiation complexes and favor suboptimal initiation sites in yeast. eLife 6:e31250. https://doi.org/10.7554/eLife.31250.

McClements, M. E., A. Butt, E. Piotter, C. F. Peddle, and R. E. MacLaren. 2021. An analysis of the Kozak consensus in retinal genes and its relevance to gene therapy. Mol. Vis. 27:233–242. http://www.molvis.org/molvis/v27/233.

Mensi, A., Y. Choiset, H. Rabesona, T. Haertle, P. Borel, and J.-M. Chobert. 2013. Interactions of β-lactoglobulin variants A and B with Vitamin A. Competitive binding of retinoids and carotenoids. J. Agric. Food Chem. 61:4114–4119. https://doi.org/10.1021/jf400711d.

Miranda, G., M. F. Mahé, C. Leroux, and P. Martin. 2004. Proteomic tools to characterize the protein fraction of Equidae milk. Proteomics 4:2496–2509. https://doi.org/10.1002/pmic.200300765.

Paetzel, M., A. Karla, N. C. J. Strynadka, and R. E. Dalbey. 2002. Signal Peptidases. Chem. Rev. 102:4549–4580. https://doi.org/10.1021/cr010166y.

Passey, R., W. Glenn, and A. Mackinlay. 1996. Exon skipping in the ovine α_{s1}-casein gene. Comp. Biochem. Physiol. B Biochem. Mol. Biol. 114:389–394. https://doi.org/10.1016/0305-0491(96)00075-2.

Pena, R. N., A. Sanchez, A. Coll, and J. M. Folch. 1999. Isolation, sequencing and relative quantitation by fluorescent-ratio PCR of feline beta-lactoglobulin I, II, and III cDNAs. Mamm. Genome 10:560–564. https://doi.org/10.1007/s003359901044.

Pérez, M. D., and M. Calvo. 1995. Interaction of β-lactoglobulin with retinol and fatty acids and its role as a possible biological function for this protein: A review. J. Dairy Sci. 78:978–988. https://doi.org/10.3168/jds.S0022-0302(95)76713-3.

Pérez, M. D., P. Puyol, J. M. Ena, and M. Calvo. 1993. Comparison of the ability to bind lipids of beta-lactoglobulin and serum albumin of milk from ruminant and non-ruminant species. J. Dairy Res. 60:55–63. https://doi.org/10.1017/S0022029900027345.

Pérez, M. J., C. Leroux, A. Sanchez Bonastre, and P. Martin. 1994. Occurrence of a LINE sequence in the 3′ UTR of the goat αs1-casein E-encoding allele associated with reduced protein synthesis level. Gene 147:179–187. https://doi.org/10.1016/0378-1119(94)90063-9.

Pervaiz, S., and K. Brew. 1986. Purification and characterization of the major whey proteins from the milks of the bottlenose dolphin

(*Tursiops truncatus*), the Florida manatee (*Trichechus manatus latirostris*), and the beagle (*Canis familiaris*). Arch. Biochem. Biophys. 246:846–854. https://doi.org/10.1016/0003-9861(86)90341-3.

Pisano, M. P., O. Tabone, M. Bodinier, N. Grandi, J. Textoris, F. Mallet, and E. Tramontano. 2020. RNA-Seq transcriptome analysis reveals long terminal repeat retrotransposon modulation in human peripheral blood mononuclear cells after in vivo lipopolysaccharide injection. J. Virol. 94:e00587-20. https://doi.org/10.1128/JVI.00587-20.

Rada-Iglesias, A., A. Ameur, P. Kapranov, S. Enroth, J. Komorowski, T. R. Gingeras, and C. Wadelius. 2008. Whole-genome maps of USF1 and USF2 binding and histone H3 acetylation reveal new aspects of promoter structure and candidate genes for common human disorders. Genome Res. 18:380–392. https://doi.org/10.1101/gr.6880908.

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

Rando, A., P. Di Gregorio, L. Ramunno, P. Mariani, A. Fiorella, C. Senese, D. Marletta, and P. Masina. 1998. Characterization of the CSN1AG allele of the bovine αs1-casein locus by the insertion of a relict of a long interspersed element. J. Dairy Sci. 81:1735–1742. https://doi.org/10.3168/jds.S0022-0302(98)75741-8.

Sawyer, L., and G. Kontopidis. 2000. The core lipocalin, bovine beta-lactoglobulin. Biochim. Biophys. Acta. 1482:136–148. https://doi.org/10.1016/S0167-4838(00)00160-6.

Schanbacher, F. L., R. E. Goodman, and R. S. Talhouk. 1993. Bovine mammary lactoferrin: Implications from messenger ribonucleic acid (mRNA) sequence and regulation contrary to other milk proteins. J. Dairy Sci. 76:3812–3831. https://doi.org/10.3168/jds.S0022-0302(93)77725-5.

Szymanowska, M., N. Strzalkowska, E. Siadkowska, J. Krzyzewski, Z. Ryniewicz, and L. Zwierzchowski. 2003. Effects of polymorphism at 5′-noncoding regions (promoters) of α_{S1}- and α_{S2}-casein genes on selected milk production traits in Polish Black-and-White cows. Anim. Sci. Pap. Rep. 21:97–108.

Tidona, F., C. Sekse, A. Criscione, M. Jacobsen, S. Bordonaro, D. Marletta, and G. E. Vegarud. 2011. Antimicrobial effect of donkeys' milk digested in vitro with human gastrointestinal enzymes. Int. Dairy J. 21:158–165. https://doi.org/10.1016/j.idairyj.2010.10.008.

Touriol, C., S. Bornes, S. Bonnal, S. Audigier, H. Prats, A. C. Prats, and S. Vagner. 2003. Generation of protein isoform diversity by alternative initiation of translation at non-AUG codons. Biol. Cell 95:169–178. https://doi.org/10.1016/S0248-4900(03)00033-9.

Visweswaraiah, J., Y. Pittman, T. E. Dever, and A. G. Hinnebusch. 2015. The β-hairpin of 40S exit channel protein Rps5/uS7 promotes efficient and accurate translation initiation in vivo. eLife 4:e07939. https://doi.org/10.7554/eLife.07939.

von Heijne, G. 1986. A new method for predicting signal sequence cleavage sites. Nucleic Acids Res. 14:4683–4690. https://doi.org/10.1093/nar/14.11.4683.

Wodas, L., M. Mackowski, A. Borowska, K. Puppel, B. Kuczynska, and J. Cieslak. 2020. Genes encoding equine β-lactoglobulin (LGB1 and LGB2): Polymorphism, expression, and impact on milk composition. PLoS One 15:e0232066. https://doi.org/10.1371/journal.pone.0232066.

Wu, A. C. K., H. Patel, M. Chia, F. Moretto, D. Frith, A. P. Snijders, and F. J. van Werven. 2018. Repression of divergent noncoding transcription by a sequence-specific transcription factor. Mol. Cell 72:942–954.e7. https://doi.org/10.1016/j.molcel.2018.10.018.

Xu, C., and J. Zhang. 2020. Mammalian alternative translation initiation is mostly nonadaptive. Mol. Biol. Evol. 37:2015–2028. https://doi.org/10.1093/molbev/msaa063.

Yarragudi, A., L. W. Parfrey, and R. H. Morse. 2007. Genome-wide analysis of transcriptional dependence and probable target sites for Abf1and Rap1 in *Saccharomyces cerevisiae*. Nucleic Acids Res. 35:193–202. https://doi.org/10.1093/nar/gkl1059.

Zur, H., and T. Tuller. 2013. New universal rules of eukaryotic translation initiation fidelity. PLOS Comput. Biol. 9:e1003136. https://doi.org/10.1371/journal.pcbi.1003136.

Zidi, A., V. M. Fernández-Cabanás, B. Urrutia, J. Carrizosa, O. Polvillo, P. González-Redondo, J. Jordana, D. Gallardo, M. Amills, and J. M. Serradilla. 2010. Association between the polymorphism of the goat stearoyl-CoA desaturase 1 (*SCD*1) gene and milk fatty acid composition in Murciano-Granadina goats. J. Dairy Sci. 93:4332–4339. https://doi.org/10.3168/jds.2009-2597.

ORCIDS

G. Cosenza https://orcid.org/0000-0001-6006-4987
P. Martin https://orcid.org/0000-0002-8296-7404
D. Gallo https://orcid.org/0000-0001-8803-8465
B. Auzino https://orcid.org/0000-0001-9555-9680
R. Ciampolini https://orcid.org/0000-0001-5676-1798
A. Pauciullo https://orcid.org/0000-0002-3140-9373