BMC Genomics

RESEARCH

Open Access

Global analysis of bovine milk protein variants using multi-breed DNA sequence data

Ying Liu[1*], Etske Bijl[2], Junxin Gao[1], Rayner Gonzalez-Prendes[1], Martien A. M. Groenen[1], Juha Kantanen[3], Catarina Ginja[4], Nasser Ghanem[5], Donald Rugira Kugonza[6], Mahlako Makgahlela[7,8], Henk Bovenhuis[1] and Richard P. M. A. Crooijmans[1]

Abstract

Background To date, 63 variants of six major bovine milk proteins (α_{S1}-CN, β-CN, α_{S2}-CN, κ-CN, α-LA, and β-LG) have been described. These variants are caused by changes in the amino acid sequence of the mature protein, primarily resulting from missense variations in the exons of genes or splice sites. Several of these variants are known to be associated with milk production traits, cheese-processing properties, and the nutritional value of milk. In the past, milk protein variants have been identified in a limited number of breeds, especially in dairy cattle. The objective of this study was to investigate variation in milk proteins in a large number of cattle breeds based on whole genome sequencing data.

Results We investigated variants of the six major milk proteins in 3,824 cattle representing 113 breeds with wide geographical distribution using whole genome sequencing data. 59 missense variants of milk protein genes that can alter the amino acid sequence of the mature protein were detected. Notably, 10 out of 11 missense variants in *CSN3* were located within the region coding for the glyco-macropeptide, whereas the para-κ-casein region involved in micelle stabilization remained highly conserved with only one variant detected, suggesting functional constraint in this region. A total of 121 milk protein variants were identified based on different combinations of the 59 missense variants, of which 35 had been described previously. We detected 86 novel variants that had not been reported previously. These protein variants were likely missed in earlier studies due to technical limitations or the use of limited number of animals or breeds.

Conclusion This study provides a comprehensive overview of milk protein diversity across global cattle breeds, offering valuable insights for improving milk quality and properties, guiding selective breeding and prioritizing variants for future functional investigation in the dairy sector.

Keywords Milk protein, Genome sequencing, Genetic variation, Global cattle breeds

*Correspondence:
Ying Liu
ying1.liu@wur.nl
[1]Animal Breeding and Genomics, Wageningen University and Research, Wageningen 6700 AH, The Netherlands
[2]Food Quality and Design Group, Wageningen University and Research, Wageningen 6700 AA, The Netherlands
[3]Natural Resources Institute Finland, Jokioinen, Finland
[4]Faculty of Veterinary Medicine, CIISA, University of Lisbon, Lisbon, Portugal
[5]Animal Production Department, Faculty of Agriculture, Cairo University, Giza, Egypt
[6]Department of Agricultural Production, College of Agricultural and Environmental Sciences, Makerere University, Kampala, Uganda
[7]Agricultural Research CouncilAnimal Production Institute, Irene, South Africa
[8]Department of Animal, Wildlife and Grassland Sciences, University of the Free State, Bloemfontein, South Africa

Background

Cow milk proteins provide suckling calves with essential nutrients, including amino acids, calcium, phosphorus and potentially bioactive peptides. Moreover, milk proteins are the primary macromolecules responsible for the production of many dairy products, such as cheese and yoghurt. About 80% of the total bovine milk protein fraction is composed of four caseins (α_{S1}-CN, α_{S2}-CN, β-CN, and κ-CN) [1]. The remaining approximately 20% of the bovine milk proteins consists of two major whey proteins α-lactalbumin (α-LA) and β-lactoglobulin (β-LG), which are present in the serum phase of milk. These six major milk proteins are encoded by six genes. The 4 major casein genes *CSN1S1* (α_{S1}-CN), *CSN2* (β-CN), *CSN1S2* (α_{S2}-CN), and *CSN3* (κ-CN) which are located within a 250-kb genomic region on Bos taurus autosome 6 (BTA 6) (between 85.4 and 85.7 Mb on genome assembly ARS-UCD1.2). The genes coding for the main whey proteins, *LALBA* (α-LA) and *PAEP* (β-LG) are located on BTA5 and BTA11, respectively [2].

Bovine milk protein genes have been extensively investigated and characterized and all genes have been shown to be polymorphic [3]. Until now, a total of 63 protein variants have been reported [3–6]. Milk protein variants are caused by missense variations or splice sites variations, resulting in amino acid changes in the mature protein. Additional DNA variations occur in the regulatory regions of genes which can modulate (increase/decrease) the gene expression [7, 8]. Some milk protein polymorphisms have been found to lead to quantitative differences in protein composition or affect post-translational modification (PTM) [9]. Both could influence the nutritional value and technological properties of milk, which might affect the quality of dairy products [3, 10]. Thus there is a great deal of scientific interest in the identification of milk protein variants in cattle breeds.

Cattle breeds with different production characteristics have been shaped by divergent selective breeding histories. Dairy breeds have been intensively selected for high milk yield and composition to meet the demands of dairy production, whereas beef cattle have been selected mainly for growth and carcass traits. In contrast to dairy systems, where calves are separated early and fed milk replacer, calves in the beef production systems usually remain with their dams until weaning. As a result, maternal milk composition directly affects calf growth and weaning weight, which is an important breeding goal trait in beef cattle [11, 12]. κ-casein plays a critical role in milk coagulation, a process that influences the digestion rate and nutrient utilization in suckling calves. Variants that reduce renneting efficiency, such as κ-casein A and E, may negatively affect digestion and calf performance because milk passes through the abomasum too rapidly for an efficient extraction of nutrients [13]. We therefore hypothesize that these variants are under stronger purifying selection in beef than in dairy breeds.

In previous studies, two general approaches have been used to identify protein variants: separation of proteins and the DNA sequence analysis. Protein separation approaches including electrophoretic, chromatography- and mass spectrometry-based protein fractionation techniques have mainly been used for dairy cattle breeds where milk is readily available. However, these techniques separate protein variants in milk relying on differences in physicochemical characteristics such as charge, isoelectric point, hydrophobicity or molecular weight [14–17] but may fail to detect all variants. Most beef and indigenous breeds have been poorly studied on milk composition, as their main purpose is beef or draft rather than dairy products. DNA sequence analysis provides access to genomic data, which facilitate the accurate identification of milk protein variants in more variety of breeds for which milk samples are unavailable. With the development of genomic sequencing technologies, bovine whole genome sequencing (WGS) providing high quality DNA sequence data have been applied to characterize both known and novel milk protein variants accurately, while studies have so far been limited to a few breeds [4, 18, 19].

In the present study we utilize WGS data from global cattle breeds, including those from the 1000 Bull Genomes Project and the OPTIBOV Project, to give a comprehensive overview on variants of the six major milk proteins. The 1000 Bull Genomes Project (Run7.0) is a collection of WGS data from more than 3,000 individuals capturing a significant proportion of the world's cattle diversity, ranging from commercial breeds selected for high milk or meat production in temperate environments to numerous local breeds adapted to harsh conditions [20]. The OPTIBOV Project includes whole-genome sequence data of local bovine breeds from six countries, adapted to diverse environments in Europe and Africa (https://www.optibov.org/). This study provides a com prehensive diversity assessment of the major milk protein variants of global cattle breeds which may provide insights into positive selection on milk properties, breeding strategies and dairy product quality improvement.

Results

DNA sequence variations

The analysis of the six major bovine milk protein genes including the 2,000 bp flanking regions identified 3,404 genetic variants relative to the bovine reference sequence. Of these variants, 1,744 were known and 1,660 were novel (Supplementary Table S2). A total of 463 variants were located in the upstream and 579 in the downstream regions. Within the six milk protein genes, 2,362 genetic variants were identified, with polymorphic

sites accounting for 3.56% of the total base pairs. Most variants within the milk protein genes were located in intronic regions (89.9%), slightly lower than the genome-wide average (95.8%) (Table 1). Variants located in untranslated regions (UTRs) accounted for 5.75%, which is higher than the genome-wide average (1.05%). Missense variants accounted for 2.8% of all polymorphisms, nearly twice the whole bovine genome proportion (1.6%).

Missense variants

A total of 66 missense variants were identified within the coding regions of the six major milk protein genes. Detailed information including the amino acid changes, location, SIFT (Sorting Intolerant From Tolerant) score and allele frequencies of missense variants across breeds and the total population was provided in Supplementary Table S4. The missense variants for the six milk protein genes and their corresponding positions within the protein are illustrated in Fig. 1. Among the identified variants, eight missense variants were the most rare and each detected in only two individuals.

Of the in total 66 missense variations, 7 could not alter the mature protein. One missense variant, which initially was annotated to the first exon of *CSN2*, actually does not code for the β-CN protein. This suggests a misannotation in the reference genome. The other 6 variants (one each in *CSN1S1*, *CSN2*, *CSN1S2*, *PAEP*, and two in *LALBA*) that cannot alter the mature protein are located in the region coding for the signal peptide, and these have no effect on the mature protein. The remaining 59 missense variants change the amino acid sequence of the mature proteins.

The distribution of missense variants varied among genes (Fig. 1). In *CSN1S1*, eight variants were identified in exons 2, 4, 8, 10, 11 and 17, and exons 10, 11 and 17 appear to accumulate more variants. Sixteen missense variants in *CSN2* were distributed across the gene. In *CSN1S2*, eight missense variants were located in exons 2, 3, 6, 7, 8, 12, 13, and 16. *CSN3* displayed a strong localization pattern, with all eleven variants confined to exon 4. In *LALBA*, six missense variants were detected in exons 1 and 2, whereas no missense mutations were detected in exons 3 and 4. In *PAEP*, ten missense variants were found in exons 1, 2, 3, 4, and 6.

Table 1 Percentage of the polymorphisms in milk protein genes and the whole bovine genome annotated to the different DNA variant types

Variant types	[%] milk protein genes	[%] whole bovine genome[a]
5 prime UTR	3.04%	0.2%
Missense	2.84%	1.57%
Synonymous	1.53%	1.52%
Intron	89.89%	95.84%
3 prime UTR	2.71%	0.85%

[a]whole bovine genome: exclude intergenic region

Milk protein variant detection

The milk protein variants can be the result of combinations of two or more sequence variants and therefore were determined based on haplotypes that were constructed for each of the six milk protein genes. A total of 121 milk protein variants were identified based on different combinations of the 59 missense variants. Violin plots showing the distribution of sequencing read depth at missense variant sites were generated for each gene (Supplementary Figure S1-S6). The plots revealed sufficient coverage across sites with mean read depths above 8, indicating that protein variant identification was based on adequately covered sequencing data. In the past, the established systematic nomenclature in which almost all previously newly identified variants were named sequentially in alphabetical order. However, as the number of newly identified variants increases, existing nomenclature system becomes difficult to scale and provides limited traceability of relationships among closely related variants. Therefore, in this study, we proposed a hierarchical nomenclature system for the novel variants that represents an extension of the already known variant nomenclature. We suggest preliminary names for newly identified milk protein variants based on their origin and relationship. The name of the new variant is formed by adding a dot and a numerical digit to the base variant's name, reflecting successive layers of amino acid substitutions. For example, variants derived from a base variant named A1 are labeled A1.1, A1.2, A1.3, and so on. If additional variants are subsequently discovered based on A1.1, they are named A1.1.1, A1.1.2, A1.1.3, etc. This system allows for the traceable classification of variant lineages and the relationship between original and derived forms. The amino acid variations and positions of all variants detected in this study as well as known variants for each of the 6 major milk proteins were shown in Supplementary Table S5-S10. These tables present the correspondence between established variant names and the novel variants with newly proposed nomenclature. The relationship between all variants were visualized using network plots presented in Figs. 2, 3, 4, 5, 6 and 7. The frequencies of variants across different breeds are presented in Supplementary Tables S11–S16. A summary of the known milk protein variants detected and those not identified in this study is presented in Table 2.

Overview of detected milk protein variants

Of the 121 detected milk protein variants, 35 corresponded to previously known variants and 86 were novel. The most common known variants included α_{S1}-CN B and C, β-CN A1 and A2, α_{S2}-CN A, κ-CN A and B, α-LA

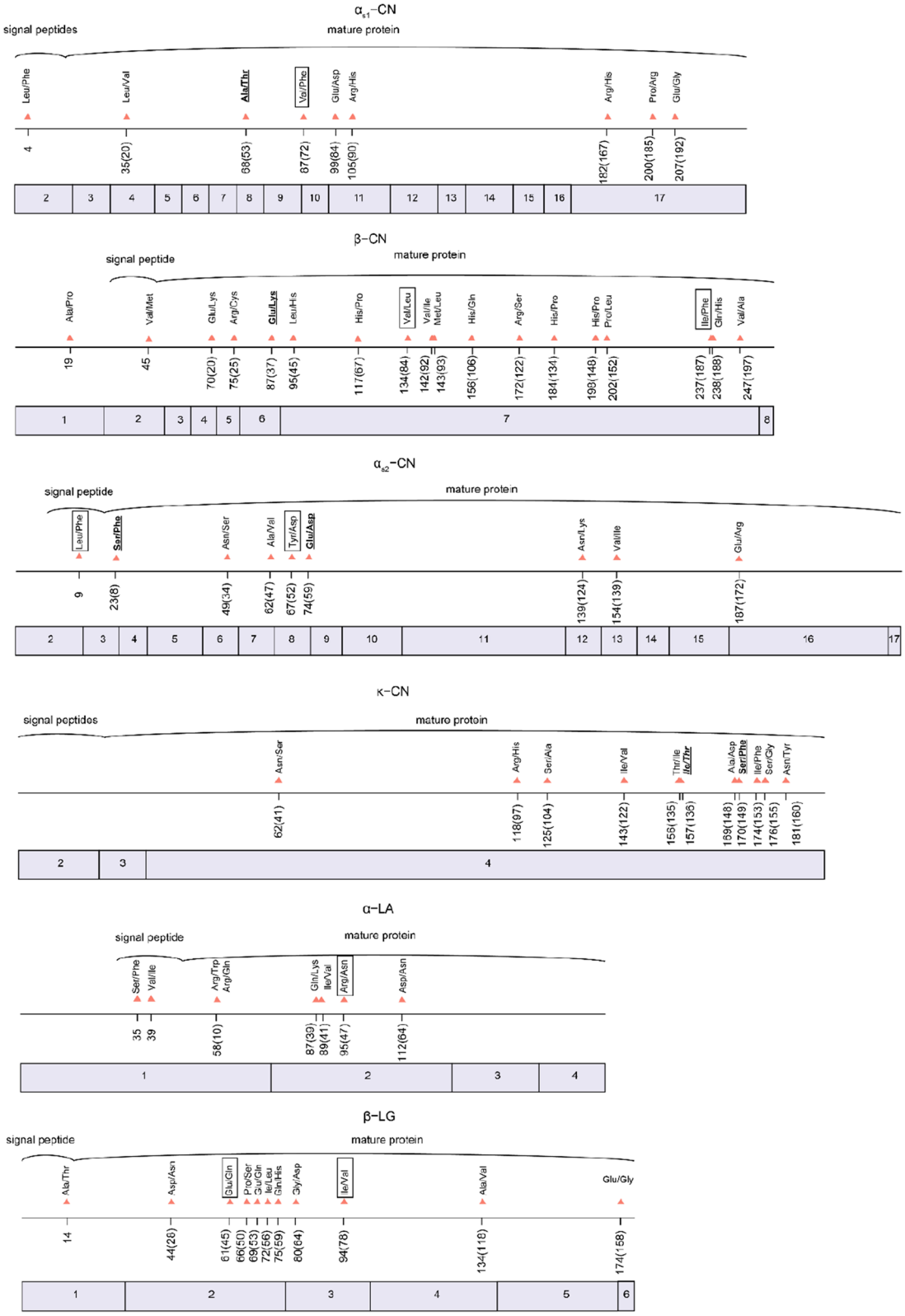

Fig. 1 Amino acid variations in the 6 major milk protein genes. Exon number and amino acid variations are numbered according to the reference protein sequence from the ARS-UCD1.2 assembly. Positions in the mature protein are given in parentheses. Protein variants coded by the reference genome sequence are: *CSN1S1*: α_{S1}-CN B; *CSN2*: β-CN B; *CSN1S2*: α_{S2}-CN A; *CSN3*: κ-CN B; *LALBA*: α-LA B; *PAEP*: β-LG B. Amino acid variations related to phosphorylation sites are in bold and underlined; glycosylation-related variants are in bold italic and underlined. The rarest variants (observed only twice) are in text boxes

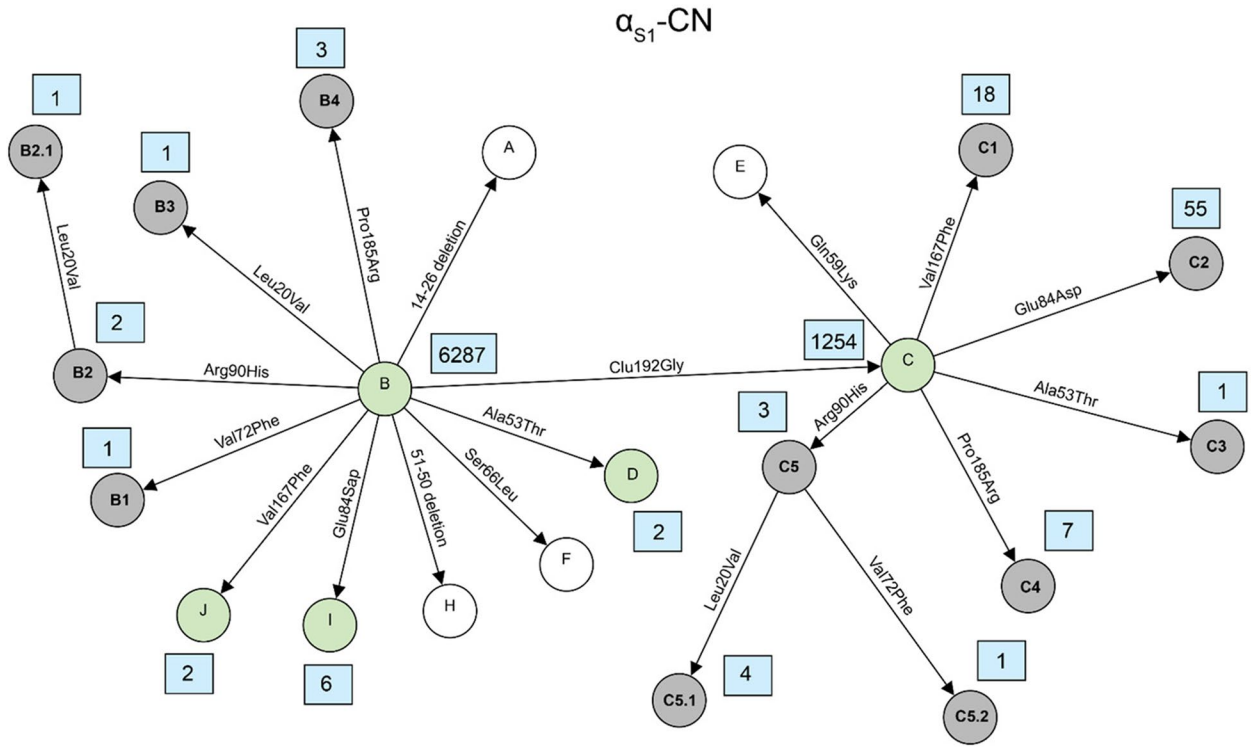

Fig. 2 Network diagram of amino acid changes among protein variants of α_{S1}-CN. Known variants which are detected in this study are shown in green circles, known variants but not detected in this study are shown in white circles and novel variants described for the first time in this study are shown in grey circles. Occurrences of each variant, which are defined here as the number of times each variant was observed in the dataset, are indicated in a blue rectangle next to its circle

B, and β-LG A and B. Several less common or rare variants displayed clear breed-specific distributions. Variants such as α_{S1}-CN D, β-CN A3, C, and I, κ-CN E, and β-LG C and D were detected exclusively in European taurine breeds. The α_{S2}-CN D variant, although predominantly present in European taurine breeds, was also identified in a few African and East Asian populations. In contrast, a number of variants including α_{S1}-CN I and J, α_{S2}-CN B, β-CN J and L, κ-CN H, α-LA A, E, and F, and β-LG W occurred mainly in Indian zebu, African, and East Asian native breeds with indicine introgression history. α_{S2}-CN B, κ-CN H, and α-LA E also occurred at low frequencies in a few European taurine breeds. The 86 newly identified milk protein variants showed diverse occurrence patterns. We highlighted three representative patterns: 1) relatively common variants detected with more than 10 occurrences; 2) geographical-specific variants restricted to particular populations; and 3) extremely rare variants with less than 4 occurrences met the setting quality criteria but were supported by read depth less than 6. Details of the newly identified variants in each representative occurrence pattern are provided in Supplementary Table S17.

Frequencies of functional variants

Frequencies of key functional variants including the β-CN A2 family referring to all β-CN variants carrying a proline residue at position 67, κ-CN A and B, and β-LG B variants in cattle breeds with at least 20 animals are illustrated in Figs. 8, 9a–b, and 10, respectively. β-CN A2 family showed intermediate to high frequencies in most breeds and appeared to be fixed in Mirandesa. It was also predominant in Brahman (0.976), Angus (0.942), and Barrosa (0.875). In contrast, four Dutch native breeds including Deep Red (0.25), Dutch Belted (0.413), Dutch Friesian (0.254) and Meuse Rhine Yssel (0.37) showed markedly lower A2 family frequencies, with the A1 variant being more common.

κ-CN A and B variants were both common variants and detected in all breeds. The B variant showed particularly high frequencies in Jersey (0.938), Mirandesa (0.926), and Northern Finncattle (0.700). The A variant was more frequent in Groningen White Headed (0.857), Dutch Belted (0.826), Norwegian Red (0.810), and Dutch Friesian (0.790) whereas the B variant occurred at lower frequencies. In other taurine breeds, frequencies of A and B variants were more balanced. β-LG B variant was predominated in most breeds, with particularly high frequencies in Nganda (0.926), Bonsmara (0.925), and Dutch

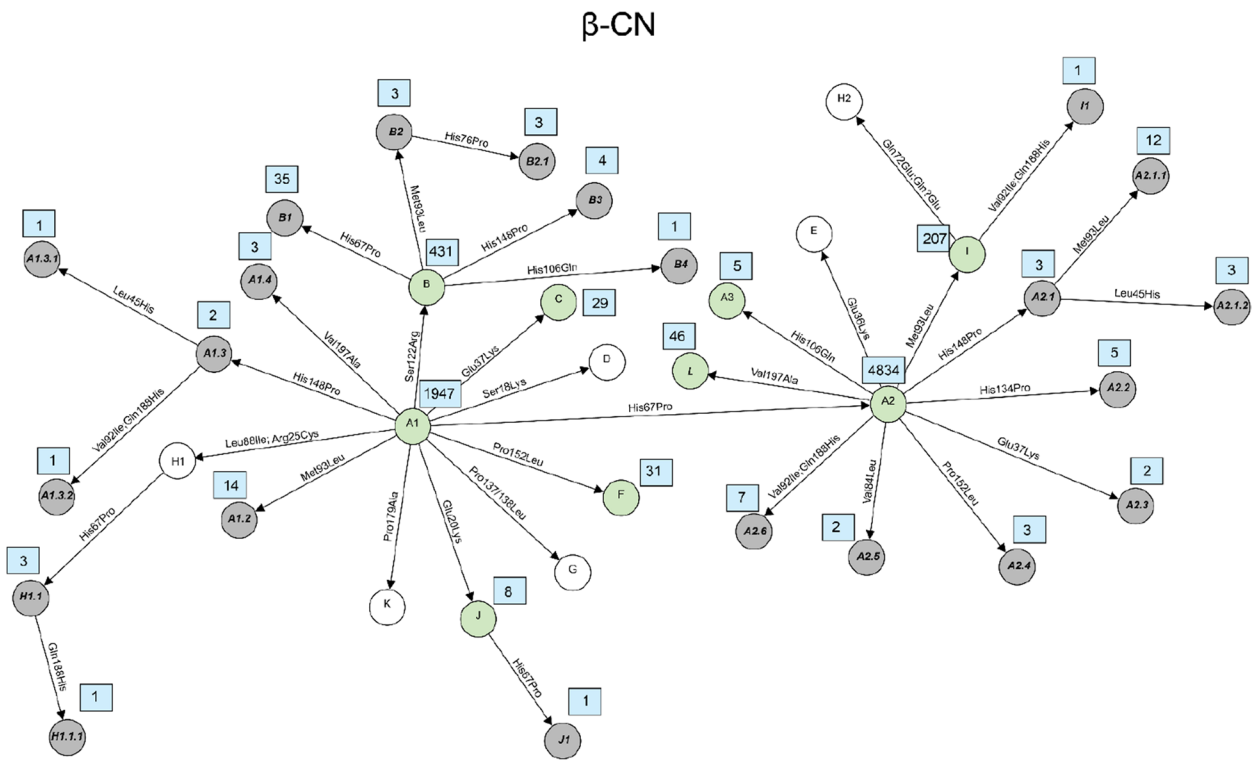

Fig. 3 Network diagram of amino acid changes among protein variants of β-CN. Known variants which are detected in this study are shown in green circles, known variants but not detected in this study are shown in white circles and novel variants described for the first time in this study are shown in grey circles. Occurrences of each variant, which are defined here as the number of times each variant was observed in the dataset, are indicated in a blue rectangle next to its circle

Belted (0.913). In contrast, it was relatively rare in Blonde d'Aquitaine (0.219) and Meuse Rhine Yssel (0.398).

Discussion

DNA sequencing variants

In this study, we analyzed sequences of six milk protein genes as well as 2,000 bp upstream and downstream regions. The variations detected in the upstream and downstream regions may include functional regulatory elements such as promoters or enhancers that modulate gene expression. Several variants in the upstream region have been reported to influence transcriptional activity and gene expression, and consequently affect milk composition traits [7, 8, 21, 22].

In our dataset, the level of polymorphism within milk protein genes is higher than the average observed across bovine genes, where approximately 2.1% of base pairs are polymorphic [20]. We observed that 97.16% of DNA variants were located in the UTR, synonymous sites, and splicing regions and did not alter the amino acid sequences. This proportion is consistent with previous reports, which ranged from 94.2% to 98.9% [4, 18, 19] and lower compared with the genome-wide average of 98.43% [20]. Although such type of sequence polymorphisms do not result in amino acid changes of proteins, they have been shown to contribute to the genetic variance of milk production traits in mammary gland [23]. Moreover, variants within UTRs may cause the gain or loss of microRNA binding sites or transcription factor motifs, thereby leading to expression differences of the genetic protein variants [3, 4, 19]. Together, these results emphasize that genetic variants although silent at the protein level, can play a critical role in modulating gene expression and thereby contributing to functional diversity.

Missense variant distribution

A total of 59 missense variants that change the amino acid sequence of the mature forms of six milk were detected. Variants located within predicted post-translational modification (PTM) sites may alter the phosphorylation or glycosylation levels, thereby influencing the technological properties of milk and dairy products [9, 24, 25]. Moreover, the distribution of missense variants highlights non-conserved and conserved regions across milk protein genes. We observe that the missense variants of *CSN3* gene almost exclusively located within the region coding for the hydrophilic κ-CN C-terminal (residues 113–169), which is part of the soluble casein macropeptide (CMP) after cleavage by chymosin between residues Phe^{105}–Met^{106} [26, 27]. In contrast, other regions of *CSN3* are highly conserved with few sequence variations. In this study, only one variation was

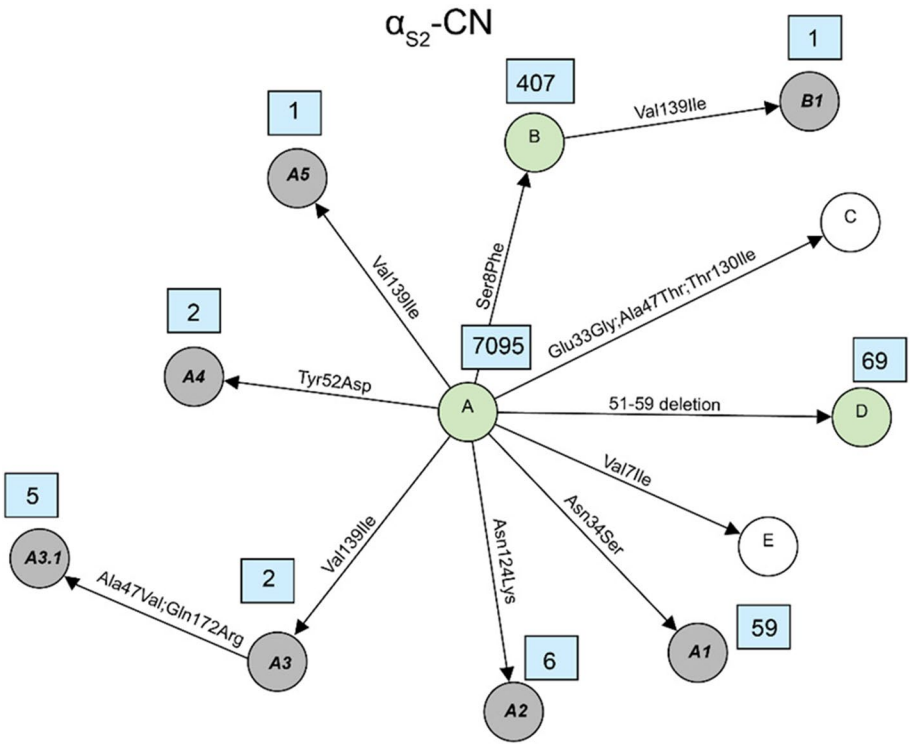

Fig. 4 Network diagram of amino acid changes among protein variants of α_{S2}-CN. Known variants which are detected in this study are shown in green circles, known variants but not detected in this study are shown in white circles and novel variants described for the first time in this study are shown in grey circles. Occurrences of each variant, which are defined here as the number of times each variant was observed in the dataset, are indicated in a blue rectangle next to its circle

found in the region of *CSN3*, which codes for the para-κ-CN of κ-CN (residues 1–95). This region is the hydrophobic part that remains in the micelle after cleavage by chymosin. In an evolutionary study on κ-CN, Manguy and Shields noted that the length of para-κ-CN is relatively constrained, which might be due to its critical role in stabilizing the casein micelle internal structure [28]. However, evolutionary large insertions were tolerated in the CMP [28], which is the part of κ-CN that is exposed on the exterior of the micelle. These results indicated a higher conservation level of para-κ-CN compared to the CMP, which is consistent with the results of our study. Overall, the conserved regions of milk protein genes reflect strong structural and functional constraints, particularly in casein genes that play crucial roles in micelle assembly and stability [29].

Known variants detection using WGS data

The frequencies of known variants across different breeds were largely consistent with those reported in earlier DNA-based studies, considering differences in sample size and geographic origin [18, 30–32]. Distinct geographic and ancestry-related distribution patterns were observed for several milk protein variants. Variants previously considered zebu-specific including α_{S1}-CN I and J, β-CN J and L, α_{S2}-CN B, κ-CN H, α-LA A, E, and F, and β-LG showed a wider distribution, particularly in Indian zebu, African and East Asian native breeds [4, 5, 33]. Moreover, α_{S2}-CN B, κ-CN H and α-LA A, E were also detected in several southern European cattle. This distribution pattern is consistent with previous studies reporting the introgression of indicine ancestry into African, East Asian cattle and limited in Southern European cattle [34, 35].

Compared with previous studies that characterized protein variants using protein separation techniques, some difference were observed. For example, we detected the β-CN I variant which co-elutes with the A2 variant when analyzed by capillary zone electrophoresis [36]. Similar chromatographic limitations were reported for the discrimination between κ-CN A and E, α_{S1}-CN B and C, and among different α_{S2}-CN variants of A, B and D [37]. The use of DNA sequencing in the present study allowed for accurate discrimination of such variants. Moreover, we identified the κ-CN G2 and β-LG B1 variant at low frequencies in more than 30 breeds, including mainstream dairy, beef and native breeds. κ-CN G2 was originally identified in Pinzgauer cattle but has not been reported since [38]. The B1 variant has been detected only in the Beninese cattle using DNA sequencing data [4]. The inconsistent detection of κ-CN G2 and β-LG B1 variants in breeds in the current and previous studies is

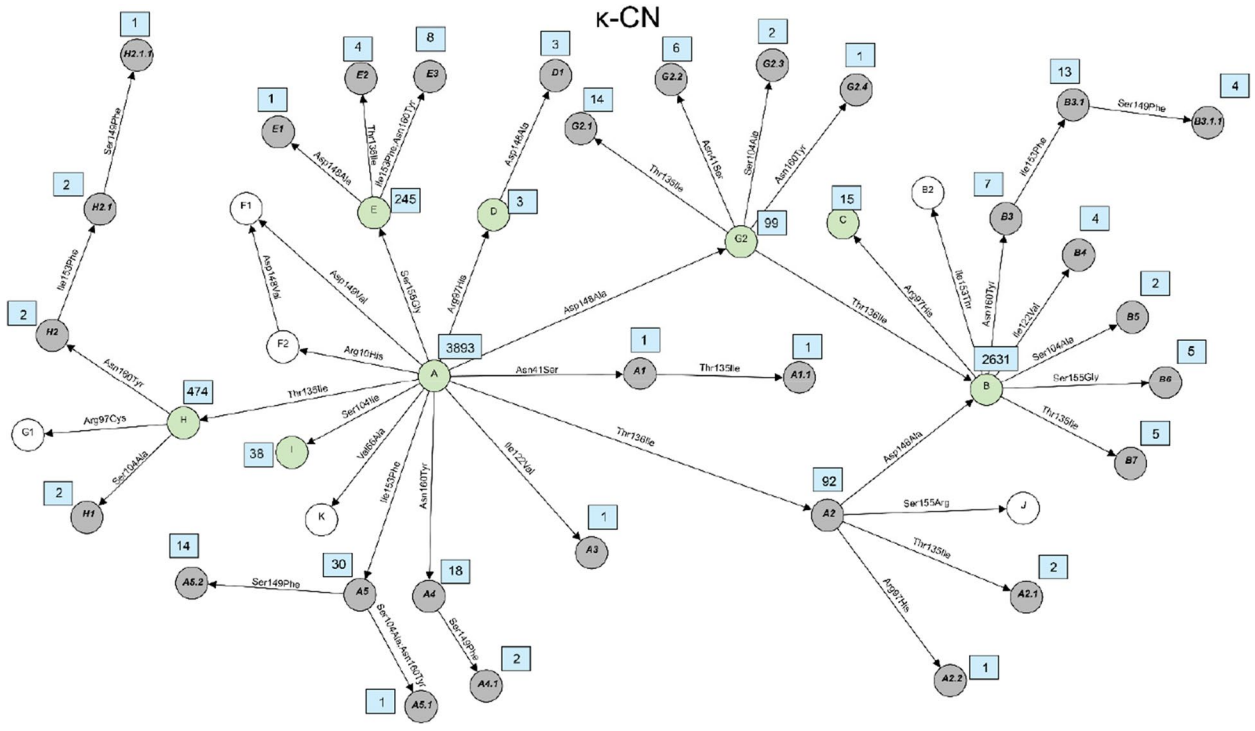

Fig. 5 Network diagram of amino acid changes among protein variants of κ-CN. Known variants which are detected in this study are shown in green circles, known variants but not detected in this study are shown in white circles and novel variants described for the first time in this study are shown in grey circles. Occurrences of each variant, which are defined here as the number of times each variant was observed in the dataset, are indicated in a blue rectangle next to its circle

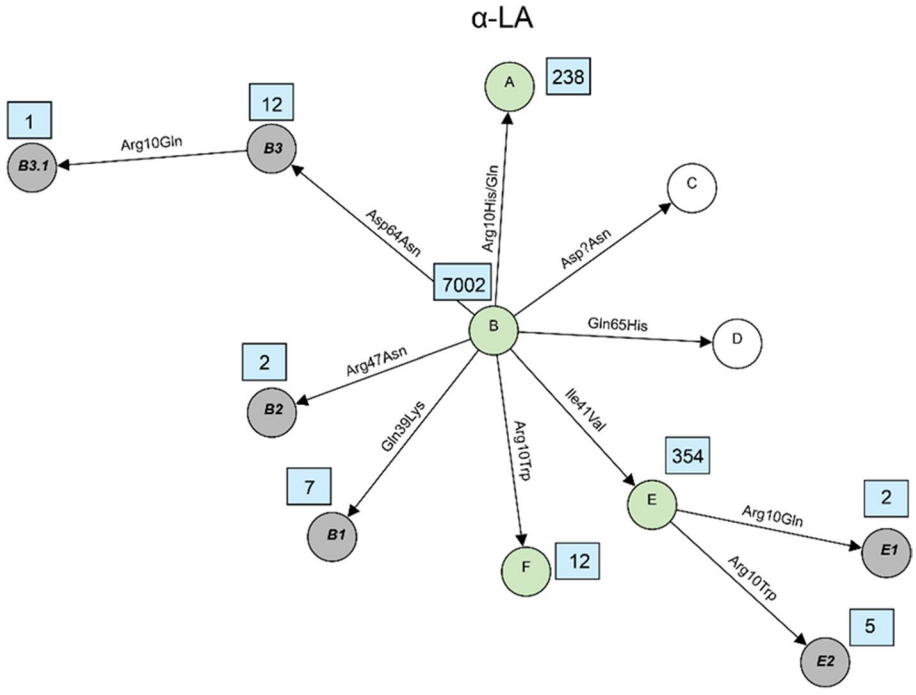

Fig. 6 Network diagram of amino acid changes among protein variants of α-LA. Known variants which are detected in this study are shown in green circles, known variants but not detected in this study are shown in white circles and novel variants described for the first time in this study are shown in grey circles. Occurrences of each variant, which are defined here as the number of times each variant was observed in the dataset, are indicated in a blue rectangle next to its circle

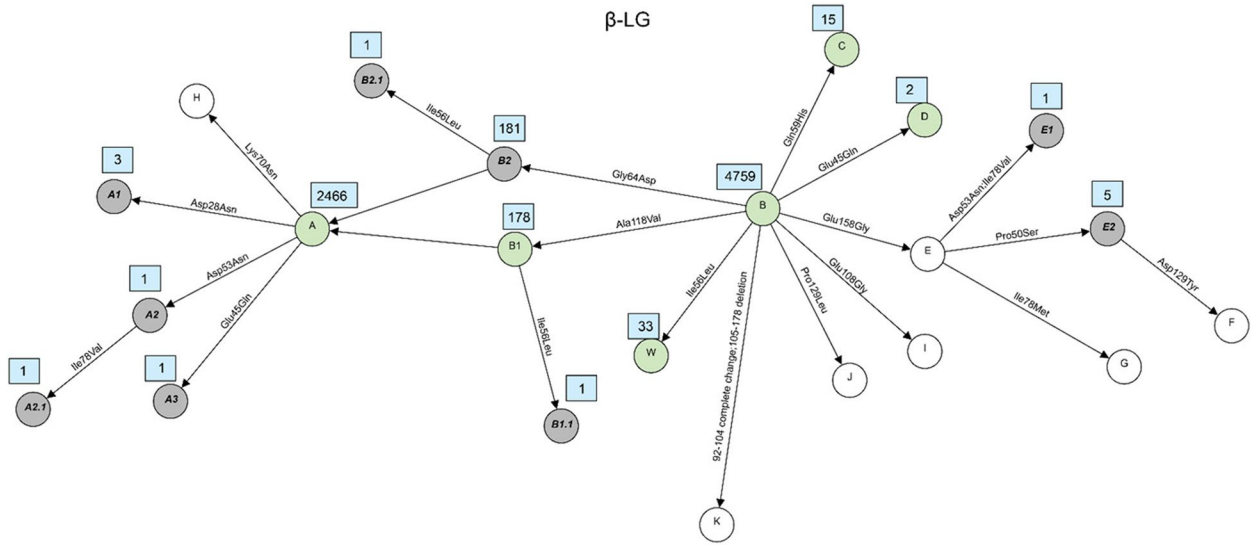

Fig. 7 Network diagram of amino acid changes among protein variants of β-LG. Known variants which are detected in this study are shown in green circles, known variants but not detected in this study are shown in white circles and novel variants described for the first time in this study are shown in grey circles. Occurrences of each variant, which are defined here as the number of times each variant was observed in the dataset, are indicated in a blue rectangle next to its circle

Table 2 Summary of known protein variants which are detected and not detected in this study. Variants in bold were detected in breeds not included in this study. Underlined variants were classified reported in literature as "rare"

Milk Protein	Known variants detected in current study	Known variants not-detected variants in current study
α_{S1}-CN	B, C, D, I, J	A, E, F, H
β-CN	A1, A2, A3, B, C, F, I,J, L	B2, D, E, G, H1, H2, K
α_{S2}-CN	A, B, D	C, E
κ-CN	A, B, C, D, E, H, G2, I	B2, F1, F2, G1, J, K
α-LA	A, B, E, F	C, D
β-LG	A, B, B1, C, D, W	E, F, G, H, I, J, K

most likely due to limitations of the protein separation techniques that were used. Our study identified such variants effectively using WGS data and provide the precise frequencies among breeds.

Undetected known variants

Some milk protein variants have been described in other studies but were not identified in the current study. The α_{S1}-CN G variant caused by a large insertion in the non-coding region of *CSN1S1* gene resulted in lower amounts of α_{S1}-casein protein in milk due to that reduces mRNA stability and transcription [39]. κ-CN AI variant is a new DNA variant of *CSN3* gene but results in a silent mutation at the amino acid sequence of κ-CN [40]. These two variants are not caused by amino acid changes and were therefore not considered in this study. α_{S1}-CN A variant was not detected in this study because it is caused by a distal splice site variation [41]. However, the genetic variation responsible for α_{S1}-CN A variant has not been experimentally validated yet. Thus appropriate milk samples are needed to clarify effects of potential splice site variants on the mature protein amino acid sequences [42].

Other variants appear to be specific to breeds or subspecies that were not represented in our analysis. Variants including α_{S1}-CN H, α_{S2}-CN E, κ-CN D, I, J, K were found to be specific for cattle breeds which were not included in this study [3]. Variants α_{S1}-CN E, α_{S2}-CN C, α-LA C, β-LG E, F and G were found to be specific for other *Bos* species such as *Bos grunniens* and *Bos javanicus*. Additional variants including α_{S1}-CN F, β-CN D, E, H1 and H2, κ-CN F1 and F2 have been reported in other studies, but at low frequencies (~0.01) [43–45]. The α-LA D variant was detected in 1 of the 1,948 Dutch Holstein Friesian cows [6]. β-CN G variant was identified in 7 of 158 Holstein Friesian cows [46]. Taken together, these variants are only present in specific breeds or herds, which likely explains why these variants were not detected in the current study.

Functional implication of known variants

Some variants that have been validated to be functional can be used in selective breeding in dairy sector due to their potential implication in human health. The β-CN A2 family variants, which have a proline residue at position 67, result in reduced release of β-casomorphin-7 (BCM-7), a peptide hypothesized to be associated with certain adverse health effects, although this remains controversial [47]. Regarding the manufacturing properties, κ-CN B is associated with improved milk rennet properties, increased cheese yield, and higher milk

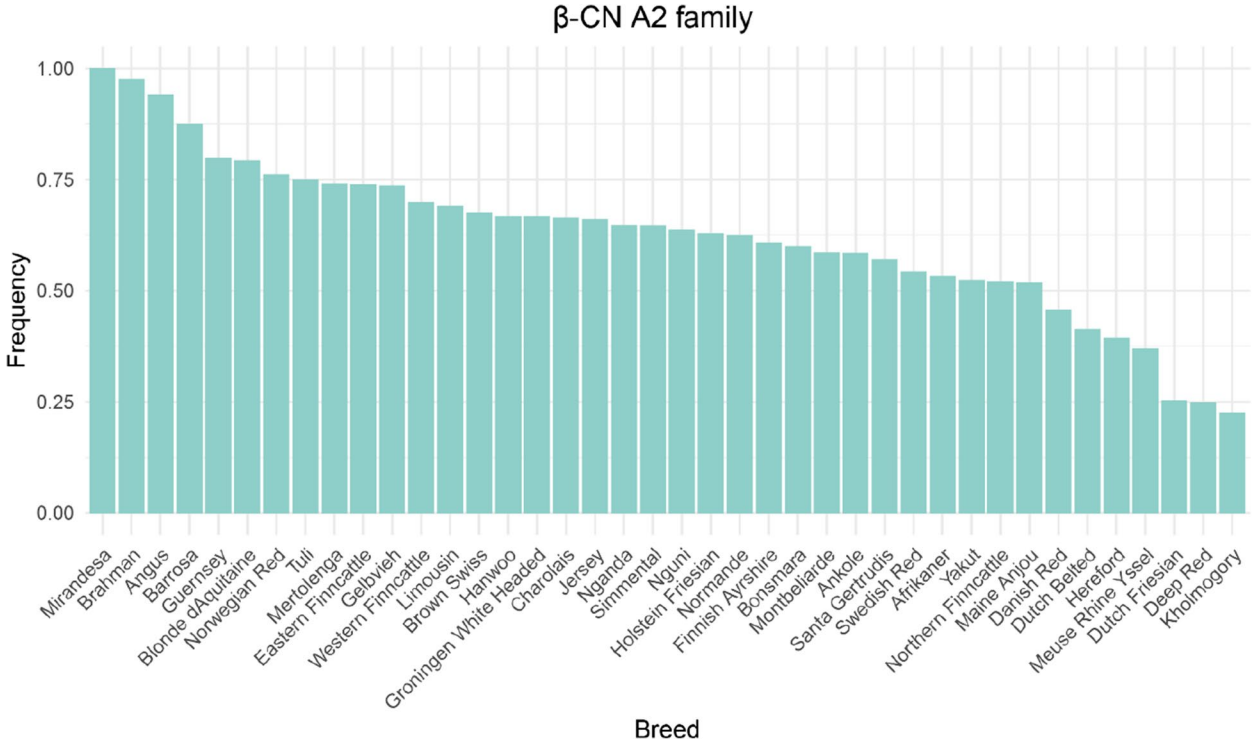

Fig. 8 Frequency of β-CN A2 family variant in cattle breeds with at least 20 animals

protein percentage [17, 48]. Moreover, milk with κ-CN B variant shows greater heat stability at natural pH [49]. κ-CN B is reported to be more resistant to gastric digestion compared with A variant [50]. β-LG B is associated with improved firmer curd and high dried weight cheese yield, and also show better foaming properties [49, 51]. In our dataset, the β-CN A2 family variants were fixed in Mirandesa and predominant in Brahman, and Angus. κ-CN B showed high frequencies in Jersey, Mirandesa, and Normande. The β-LG B variant was enriched in Nganda, Bonsmara, and Dutch Belted. The high frequencies distribution in breeds suggests that they could serve as valuable genetic resources for breeding programs aiming to improve specific processing properties of milk in the dairy sector. However, most Dutch native breeds displayed low frequencies of β-CN A2 family and κ-CN B variants. This may provide opportunities for breeders to increase the A2 milk production and cheese-making properties through selective breeding or genomic selection. In summary, the frequency distribution of functional variants across breeds provides valuable insight for designing breeding strategies aimed at improving the milk quality.

We observed a high frequency of the κ-casein A variant in both major commercial beef and dairy breeds. Among beef breeds, the frequency ranged from 0.425 in Limousin to 0.718 in Angus, with intermediate values in Charolais (0.427) and Hereford (0.635). In the main dairy breeds, the frequency was 0.606 in Holstein Friesian and 0.414 in Brown Swiss, but only 0.047 in Jersey. The E variant was a rare variant and occurred at low frequencies in both beef and dairy breeds. The comparable frequencies of these two variants between beef and dairy breeds indicate no clear evidence of differential selection, which contrasts with our initial hypothesis that stronger purifying selection might have acted against these variants in beef cattle, where milk composition directly influences calf growth and fitness. Our results suggest that these variants are not under strong functional constraint or that their effects are counterbalanced by other selective pressures across breeds. Remarkably, the κ-casein A variant is extremely rare in Jersey, reflecting long-term selection in favor of the κ-CN B allele, which improve milk quality and cheese-making properties [14, 52].

Novel variants characterization

Novel milk protein variants can be the result of novel missense variants or recombination events between existing polymorphisms that result in new haplotypes. We detected 86 previously unreported variants across six milk proteins in this study. One reason why novel variants have not been detected previously might be the difficulty to separate the protein variants based on the analytical methods used before [15, 30, 37, 53–55]. Moreover, our large-scale data including diverse origin resulted in a higher probability to identify new variants.

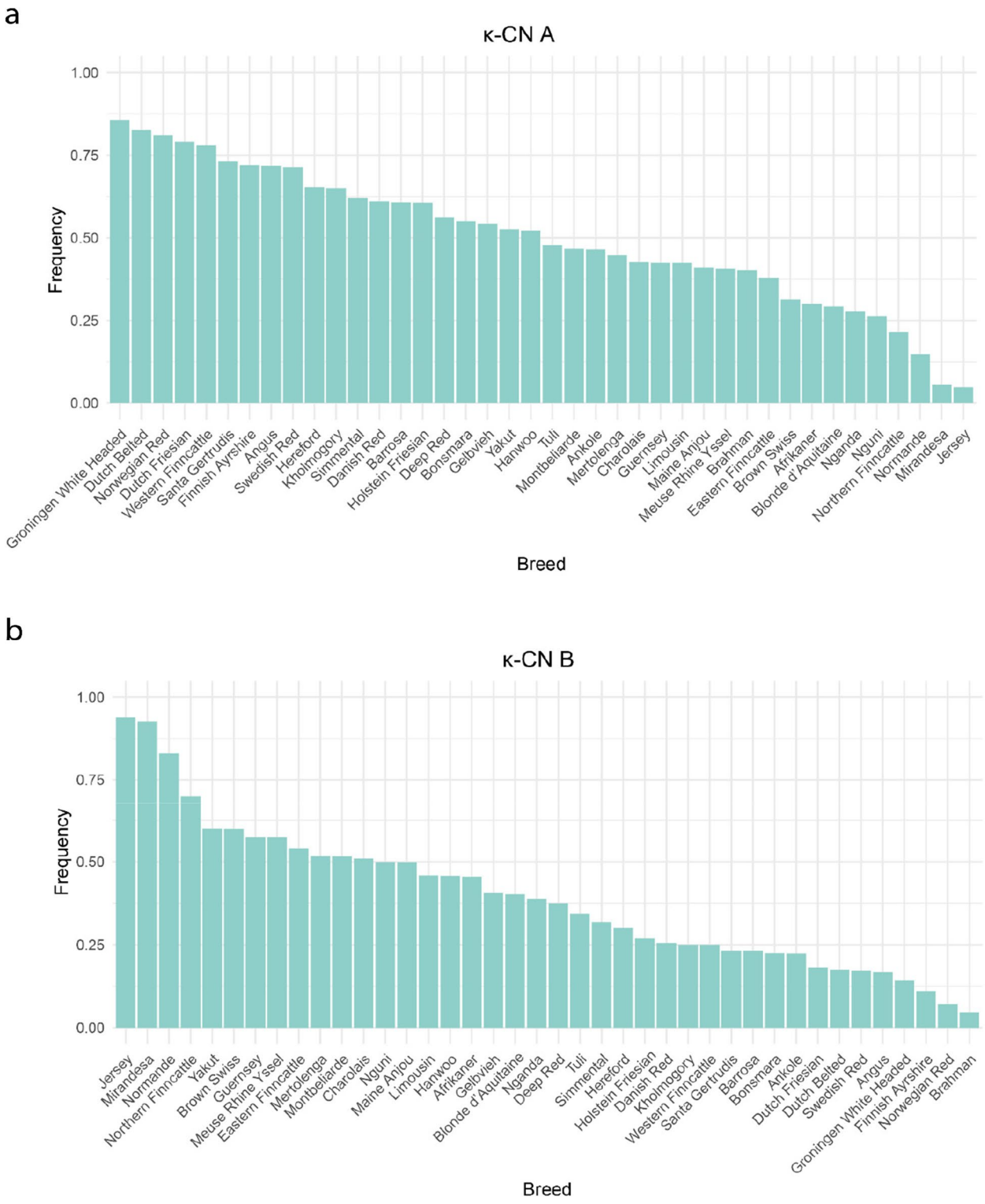

Fig. 9 Frequency of κ-CN variant in cattle breeds with at least 20 animals. **a** Frequency of κ-CN A variant. **b** Frequency of κ-CN B variant

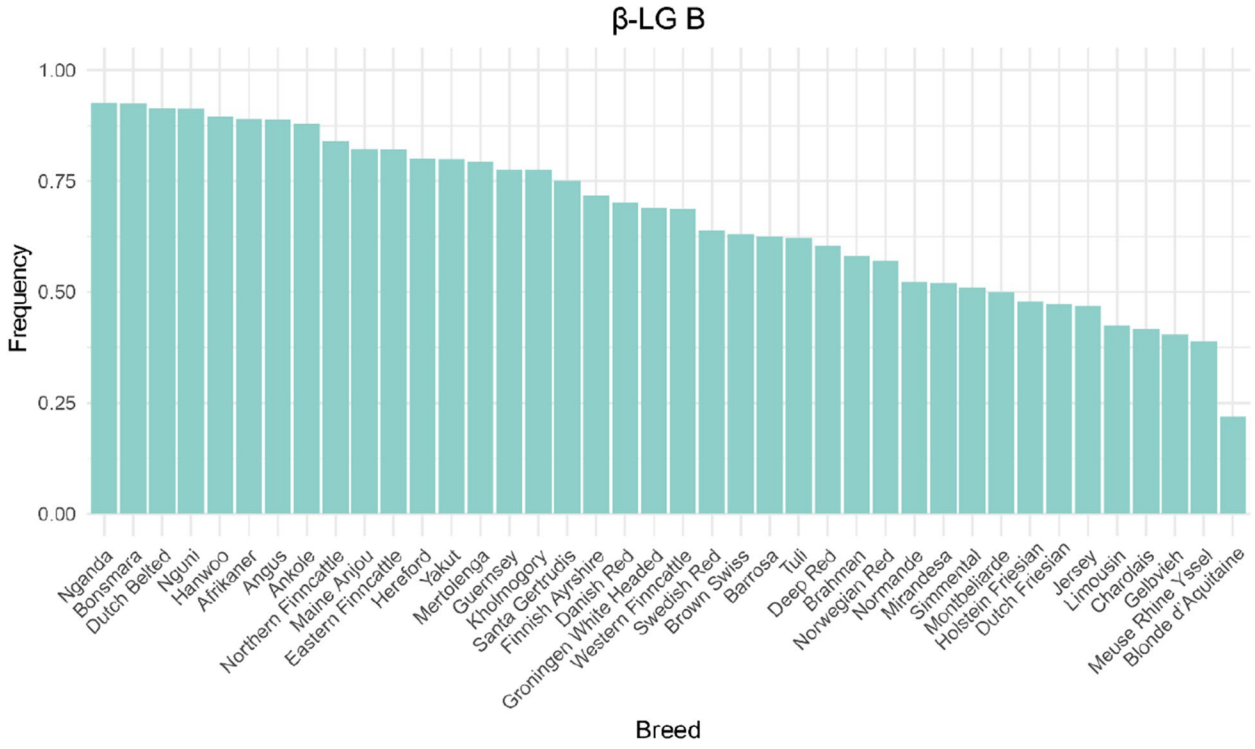

Fig. 10 Frequency of β-LG B variant in cattle breeds with at least 20 animals

Our newly proposed hierarchical nomenclature system for the novel variants builds upon and extends the existing variant nomenclature providing a scalable and traceable framework to organize an increasing number of milk protein variants. This nomenclature approach may serve as a practical framework for broader application in future studies.

The novel variants which are relatively common and were detected across multiple breeds or with multiple occurrences are considered as high reliability. We observe that some novel variants were exclusively present in specific native breeds. This may be attributed to the limited number of studies conducted on these breeds. Because most of them have traditionally been raised for beef or draft rather than dairy production, genetic diversity in milk protein genes has remained largely unexplored. The reliability of novel variants depends heavily on the DNA sequencing quality as well as on the accuracy of haplotype reconstruction. We have applied strict criteria for filtering of the sequencing data which is identical to previous study [18] and included the missense variant needs to been detected in at least two animals. However, changes in the read depth threshold directly influence genotype calling. When we change the criteria of read depth more than 6, 8 of variants (α_{S1}-CN B2, B2.1, B3, β-CN B2.1, B3, B4, β-LG E1, A3) were not present. This reflects that genotyping outcomes are sensitive to coverage stringency, and small changes in read depth thresholds can influence haplotype inference. Additionally, the accuracy of phasing by BEAGLE 5.4 depends mainly on the number of individuals, marker density, and relatedness [56]. As a result, genotypes were not 100% phased, and haplotype frequencies cannot be directly estimated by gene counting [57]. In this case, we applied EM algorithm implemented in the *haplo.stat* of R package to estimate haplotype likelihood frequencies, which resolving uncertainty in phase information and provides robust haplotype frequency estimates.

Potential implication of novel variants

We observed that some novel variants were detected exclusively in indigenous breeds from Europe, India, Africa and East Asia, respectively, suggesting distinct genetic backgrounds between these populations [34, 58]. Although the physiological and functional significance of these native breed-specific variants remains unknown, they may represent the unique milk protein characteristic across native breeds that is absent from commercial populations [59, 60]. Characterizing these unique variants and conserving the native breeds that harbor them can help prevent irreversible loss. Therefore, these native breed-specific variants should be prioritized for follow-up characterization, particularly aiming at expanding milk protein variant resources and physiological investigation. Some newly identified variants may be relevant for milk functionality, because amino acid substitutions

can influence PTMs or the structure of mature protein and thereby affect technological properties. Specifically, relative common novel variant κ-CN A5.2 and rare newly identified κ-CN variants including A4.1, B3.1.1, and H2.1.1 include a newly identified substitution (p.Ser149Phe) at phosphorylation site. This amino acid change may affect phosphorylation pattern of κ-CN and thus have potential effect on casein micelle formation [9]. Thus in the future, these functional variants could considered as prioritization to characterize and investigate biological or structural role, and further assess their potential suitability for the design or innovation of specific dairy products.

Conclusion

This study provides a comprehensive overview of milk protein variants across global cattle breeds, identifying 121 distinct protein forms, including 35 known and 86 novel variants. The conserved region within the *CSN3* para-κ-casein domain implies functional importance in micelle formation and stability, whereas casein macropeptide region display higher tolerance for mutation. The frequencies distribution of functional variants across breeds expands our understanding of milk protein diversity and provides a valuable genomic resource for dairy breeding programs aiming at improving milk quality and properties. The application of large-scale WGS data enables the detection of numerous novel variants, providing foundation of conservation and dairy products innovation with specific biological and technological characteristics, and offering new perspectives for functional validation in future studies.

Materials and methods

Animals and sequencing data

The raw sequence data were derived from two sources. We first used 544 animals from LEAP-Agri project OPTIBOV (https://www.optibov.org/) including 10 native cattle breeds originating from 3 European countries (Finland, Netherlands, Portugal), 12 native cattle breeds from 3 African countries (Egypt, Uganda, South Africa) and 30 Holstein Friesian cattle from these 6 countries. Blood samples were collected and DNA was extracted from EDTA-blood using the GENTRA Blood kit (Qiagen N.V.) and processed according to the protocol previously described by previous study [58]. To be specific, DNA was extracted from EDTA-blood samples using the GENTRA Blood kit (Qiagen N.V.). The quality and quantity of DNA were assessed using a Qubit fluorometer (Qiagen N.V.). DNA-sequence libraires were prepared by using the DNA Library Prep Kit (Illumina Inc., USA) and paired-end 150 bp sequenced on the Illumina NovaSeq6000 platform (Illumina Inc., USA). In addition, the raw sequence variation data of 1000 Bull Genome Project Run 7.0 were used in this study [20]. Crossbred individuals were removed and only breeds with at least 2 animals were selected for the analyses, so that the final dataset contained 96 different breeds and 3,280 animals. By combining information from the OPTIBOV and the 1000 Bull Genome Project, a total of 3,824 animals from 113 breeds were included in this study. Detailed information about breeds and number of animals per breed can be found in Supplementary Table S1.

Sequence data processing

Variants and genotypes of the OPTIBOV Project were called against reference genome assembly ARS-UCD1.2 using the haplotype-based method implemented in Freebayes [59]. Data of the 1000 Bull Genome Project, which was generated by aligning to the ARS-UCD1.2 reference genome, was used to filter for raw sequence variants [20]. Due to substantial variation in sequencing coverage across individuals, we applied filtering criteria based on thresholds established in a previous study [18]. SNP variants were only considered after aligning to the reference genome when at least three reads of the variant were detected. Moreover, SNPs with phred-scaled probability < 20, mapping quality < 20 and present in < 2 individuals were filtered out. Beagle 5.4 was used to infer haplotypes for genomic regions containing the 4 major casein genes on BTA6 and for the 2 major whey protein genes on BTA5 and BTA11 respectively using default settings [60].

Functional annotation of DNA sequence variants

Sequence variants located within the 6 major milk protein genes *CSN1S1*, *CSN1S2*, *CSN2*, *CSN3*, *LALBA*, *PAEP* (Table 3) including 2,000 bp upstream and downstream were selected for analyses. The positions of the identified sequence variants were inferred considering the gene sequence positions in the ARS-UCD1.2 genome assembly [61]. Variants were classified as novel when not found in the dbSNP and European Variation Archive—EVA database [62]. All the variants were categorized into variant types based on their genomic locations (2,000 bp upstream, 5′-UTR, intron, synonymous, missense, 3′-UTR, 2,000 bp downstream) using the Ensembl Variant Effect Predictor v111 (VEP) [63]. Additionally, the functional consequences of missense variations on protein were predicted using the VEP. The VEP results from two VCF files (OPTIBOV and 1000 Bull Genome Project) were combined to subsequently calculate the frequency of each variant type. Allele frequencies of missense variants in different breeds were calculated using BCFtools [64].

Table 3 Reference gene sequences used to investigate polymorphisms in milk protein genes

Gene Name	Gene ID	Transcript ID	BTA[a]	Start (bp)	End (bp)	Strand[b]	Size (bp)	Exon number
CSN1S1	ENSBTAG00000007695	ENSBTAT00000010119.3	6	85,411,118	85,429,268	+	18,151	19
CSN2	ENSBTAG00000002632	ENSBTAT00000003409.6	6	85,449,164	85,457,744	−	8,581	9
CSN1S2	ENSBTAG00000005005	ENSBTAT00000006590.6	6	85,529,905	85,548,556	+	18,652	18
CSN3	ENSBTAG00000039787	ENSBTAT00000028685.5	6	85,645,854	85,658,926	+	13,073	5
LALBA	ENSBTAG00000005859	ENSBTAT00000007701.2	5	31,183,432	31,186,209	+	2,778	4
PAEP	ENSBTAG00000014678	ENSBTAT00000019538.6	11	103,255,824	103,260,873	+	5,049	7

[a]BTA*: Bos taurus* chromosomes

[b]+ = forward strand, − = reverse strand

Milk protein variant identification

For each milk protein gene, haplotypes were constructed, as milk protein variants can be the result of combinations of two or more sequence variants in the coding regions. Haplotypes were constructed using the function *haplo.group* from R package *haplo.stats* with the default settings for the whole population and then within each breed using the EM algorithm [65]. DNA information obtained based on haplotypes were translated to amino acids and annotated to the milk protein variants following the existing nomenclature [3–6] by referring to the amino acid position in the mature proteins. Additionally, milk protein variants not found in literature are considered novel.

Abbreviations

α_{S1}-CN	Alpha-S1-casein
β-CN	Beta-casein
α_{S2}-CN	Alpha-S2-casein
κ-CN	Kappa-casein
α-LA	Alpha-lactalbumin
β-LG	Beta-lactoglobulin
WGS	Whole-genome Sequencing
BTA	Bos Taurus Autosome
CMP	Casein macropeptide
PTM	Post Translational Modification
VEP	Variant Effect Predictor
EM	Expectation–Maximization
UTR	Untranslated Region
SIFT	Sorting Intolerant From Tolerant
BCM-7	Beta-casomorphine7

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-026-12836-2.

Additional file 1: Supplementary Figure S1. Violin plot showing the distribution of sequencing read depth at genomic positions harboring missense variants in *CSN1S1* gene. Supplementary Figure S2. Violin plot showing the distribution of sequencing read depth at genomic positions harboring missense variants in *CSN2* gene. Supplementary Figure S3. Violin plot showing the distribution of sequencing read depth at genomic positions harboring missense variants in *CSN1S2* gene. Supplementary Figure S4. Violin plot showing the distribution of sequencing read depth at genomic positions harboring missense variants in *CSN3* gene. Supplementary Figure S5. Violin plot showing the distribution of sequencing read depth at genomic positions harboring missense variants in *LALBA* gene. Supplementary Figure S6. Violin plot showing the distribution of sequencing read depth at genomic positions harboring missense variants in *PAEP* gene.

Additional file 2: Supplementary Table S1. Origin of breeds or samples. Supplementary Table S2. List of the polymorphisms detected within six milk protein genes (*CSN1S1, CSN2, CSN1S2, CSN3, LALBA, LGB*) and their 2,000-bp upstream and downstream region. Supplementary Table S3. Number of variants (proportion in parentheses) found in the milk protein genes (*CSN1S1, CSN1S2, CSN2, CSN3, LALBA, PAEP*) and their 2,000-bp upstream and downstream region. Supplementary Table S4. Description (location, allele and ID) of the missense variations of the six milk protein genes (*CSN1S1, CSN2, CSN1S2, CSN3, LALBA, LGB*), their effects on the protein and allele frequencies in different breeds. Supplementary Table S5. Amino acid exchanges and position within the mature protein of α_{S1}-CN variants in Bos genus. Variants in green sheet were known variants detected in this study. Variants in grey sheet were novel variants detected in this study. Supplementary Table S6. Amino acid exchanges and position within the mature protein of β-CN variants in Bos genus. Variants in green sheet were known variants detected in this study. Variants in grey sheet were novel variants detected in this study. Supplementary Table S7. Amino acid exchanges and position within the mature protein of α_{S2}-CN variants in Bos genus. Variants in green sheet were known variants detected in this study. Variants in grey sheet were novel variants detected in this study. Supplementary Table S8. Amino acid exchanges and position within the mature protein of κ-CN variants in Bos genus. Variants in green sheet were known variants detected in this study. Variants in grey sheet were novel variants detected in this study Supplementary Table S9. Amino acid exchanges and position within the mature protein of α-LA variants in Bos genus. Variants in green sheet were known variants detected in this study. Variants in grey sheet were novel variants detected in this study. Supplementary Table S10. Amino acid exchanges and position within the mature protein of β-LG variants in Bos genus. Variants in green sheet were known variants detected in this study. Variants in grey sheet were novel variants detected in this study. Supplementary Table S11. The frequencies of α_{S1}-CN variants in different breeds. Supplementary Table S12. The frequencies of β-CN variants in different breeds. Supplementary Table S13. The frequencies of α_{S2}-CN variants in different breeds. Supplementary Table S14. The frequencies of κ-CN variants in different breeds. Supplementary Table S15. The frequencies of α-LA variants in different breeds. Supplementary Table S16. The frequencies of β-LG variants in different breeds. Supplementary Table S17. Novel milk protein variants grouped in different occurrence pattern.

Acknowledgements

The research presented in this paper was funded by the Long-term EU-Africa research and innovation Partnership on food and nutrition security and sustainable Agriculture (LEAP-Agri) as part of the OPTIBOV project (LEAP-Agri-326) and co-founded by the European Union's Horizon 2020 research and innovation program under grant agreement No 727715. We are deeply grateful to Bert Dibbits and Kimberley Laport from Animal Breeding and Genomics, Wageningen University & Research (WUR) for their technical assistance in the laboratory. We also gratefully acknowledge Tiina Reilas and Heli Lindberg from the Natural Resources Institute Finland for collecting Finnish samples for the OPTIBOV project. In Egypt, we extend our gratitude to Rania Agamy and Mohamed Hamada Elsawy from the Animal Production Department at Cairo University and the Department of Cattle from the Animal Production Research Institute for their efforts in collecting Egyptian samples for the OPTIBOV project. We also thank the director of Escola Profissional de

Agricultura e Desenvolvimento Rural de Vagos, Portugal, as well as Professor Filipe Ribeiro and veterinarian Ricardo Loureiro, for generously providing samples for the OPTIBOV project. We also greatly thank Daniel Gaspar and Carolina Bruno-de-Sousa for their collaboration in collecting samples. For South Africa, we gratefully thank Dr. Avhashoni Zwane and her team, Mr. Khanyisani Nxumalo and Mr. Maano Malima, for contributing to the OPTIBOV research samples. In Uganda, we are indebted to Dr. Morris Agaba for adapting the laboratory protocols and leading the sampling teams, and to Rodney Okwasimiire and Damian Munirwa for their extensive support in sample preparation and processing. We further acknowledge Dr. Barbara Mugwanya Zawedde and Ms. Christine Nakkazi from the National Agricultural Research Organisation (NARO) for facilitating sampling of the Nganda nucleus herd, as well as Mr. Emmanuel Tayebwa (Ankole Cattle Conservation Scheme), Mr. Bomera Asante (Butungama Multipurpose Cooperative, Ntoroko), and Mr. Charles Kasoro (Butuku Cattle Farmers Association, Ntoroko) for their valuable cooperation. Finally, we acknowledge the 1000 Bull Genomes Consortium for providing access to the sequence data used in this study.

Authors' contributions

YL, HB, RC, and EB conceived the study. YL drafted the manuscript. YL, JG and RG participated in the data analysis. RC, CG, FK, JK, NH and MM and collected the samples. HB, RC, and EB supervised the study. All the authors read and approved the manuscript.

Funding

This study was funded by the Long-term EU-Africa Research and Innovation Partnership on Food and Nutrition Security and Sustainable Agriculture (LEAP-Agri) as part of the OPTIBOV project (LEAP-Agri-326), and by the European Union's Horizon 2020 Research and Innovation Program under the grant agreement No. 727715. Additional national funding within the LEAP-Agri project was provided: for The Netherlands, from the Netherlands Organization for Scientistic Research (NWO-WOTRO), through grant number 2018/WOTRO/00488849; for Portugal from the Fundação Nacional para a Ciência e a Tecnologia (FCT), Portugal, through contract grant 2020.02754.CEECIND/CP1601/CP1649/CT0008; for Finland from the Research Council of Finland (decision number 319987) and the national organizations managing the LEAP-Agri funding; for Egypt, from the Science, Technology & Innovation Funding Authority (STDF), Egypt, grant number: LEAP-Agri 326; for Uganda from both the LEAP-Agri-326 grant of the European Union and the MoSTI/LEAP-11 grant of the Ministry of Science, Technology and Innovations, Uganda; for South Africa, from the National Research Foundation (NRF) of South Africa through the Leap Agri-326, Grant number 115577. Additional funding was obtained from the China Scholarship Council (CSC), China, Grant number 202209210016. The funding bodies had no role in the design of the study, the collection, analysis, interpretation of data, or the writing of the manuscript.

Data availability

All genome sequences are publicly available and accessible. The raw VCF file of OPTIBOV project is available in European Variation Archive (EVA) under the project accession number PRJEB101802.

Declarations

Ethics approval and consent to participate

All methods were performed in accordance with relevant guidelines and regulations. Blood samples were collected during the animals' annual health inspections, conducted by licensed veterinarians. Prior to sample collection, written informed consent was obtained from each animal's owner. In Finland, animal handling procedures and sample collections were performed in accordance with the legislation approved by Regional State Administrative Agency for Southern Finland (ESAVI/31854/2019). The Egyptian cattle blood sampling was done based on animal welfare guidelines of Institutional Animal Care and Use Committee, Cairo University (CU-IACUC) which approved this protocol under number CUIIF720. In South Africa, sampling of blood and hair was performed with the approval of the Animal Ethics Committee of the Agricultural Research Council (APAEC [2020/17]), according to guidelines for the proper handling of animals during sample collection.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

Received: 29 October 2025 / Accepted: 2 April 2026
Published online: 10 April 2026

References

1. Stelwagen K. Milk protein. Encyclopaedia of dairy science. Eds H Roginski, JW Fuquay, PF Fox. London: Academic Press; 2003;3:1835-42. https://doi.org/10.1016/B978-0-12-374407-4.00001-7.
2. Martin P, Szymanowska M, Zwierzchowski L, Leroux C. The impact of genetic polymorphisms on the protein composition of ruminant milks. Reprod Nutr Dev. 2002;42:433–59. https://doi.org/10.1051/rnd:2002036.
3. Caroli AM, Chessa S, Erhardt GJ. Invited review: milk protein polymorphisms in cattle: effect on animal breeding and human nutrition. J Dairy Sci. 2009;92:5335–52. https://doi.org/10.3168/jds.2009-2461.
4. Vanvanhossou SFU, Giambra IJ, Yin T, Brügemann K, Dossa LH, König S. First dna sequencing in beninese indigenous cattle breeds captures new milk protein variants. Genes (Basel). 2021;12. https://doi.org/10.3390/genes12111702.
5. Gallinat JL, Qanbari S, Drögemüller C, Pimentel ECG, Thaller G, Tetens J. DNA-based identification of novel bovine casein gene variants. J Dairy Sci. 2013;96:699–709. https://doi.org/10.3168/jds.2012-5908.
6. Visker MHPW, Heck JML, van Valenberg HJF, van Arendonk JAM, Bovenhuis H. Short communication: a new bovine milk-protein variant: α-lactalbumin variant D. J Dairy Sci. 2012;95:2165–9. https://doi.org/10.3168/jds.2011-4794.
7. Kuss AW, Gogol J, Geldermann H. Associations of a polymorphic AP-2 binding site in the 5'-flanking region of the bovine β-lactoglobulin gene with milk proteins. J Dairy Sci. 2003;86:2213–8. https://doi.org/10.3168/jds.S0022-0302(03)73811-9.
8. Ostrowska M, Zwierzchowski L, Brzozowska P, Kawecka-Grochocka E, Żelazowska B, Bagnicka E. The effect of single-nucleotide polymorphism in the promoter region of bovine alpha-lactalbumin (LALBA) gene on LALBA expression in milk cells and milk traits of cows. J Anim Sci. 2021;99:skab169. https://doi.org/10.1093/jas/skab169.
9. Bijl E, Holland JW, Boland M. Posttranslational modifications of caseins. Milk Proteins: From Expression to Food. 2020. pp. 173-211. https://doi.org/10.1016/B978-0-12-815251-5.00005-0.
10. Kelly LM, O'Mahony JA, Tobin JT. Genetic variation in bovine milk proteins: implications for functional and nutritional properties. Int J Dairy Technol. 2024. https://doi.org/10.1111/1471-0307.13152.
11. Corrêa MBB, Dionello NJL, Cardoso FF. Estimation of genetic parameters and (co) variance components for pre-weaning productive traits in Devon cattle in Rio Grande do Sul. Rev Bras Zootec. 2006;35:997–1004. https://doi.org/10.1590/S1516-35982006000400009.
12. Martins GA, Lima FdeAM, Lôbo RNB. Influence of genetic and environment factors on the growing traits of animals from Nellore breed at Maranhão State. R Bras Zootec. 2000;29:103–7. https://doi.org/10.1590/S1516-35982000000100014.
13. Mamo A, Balasubramanian N. Calf rennet production and its performance optimization. J Appl Nat Sci. 2018;10:247. https://doi.org/10.31018/jans.v10i1.1612.
14. Jensen HB, Poulsen NA, Andersen KK, Hammershøj M, Poulsen HD, Larsen LB. Distinct composition of bovine milk from Jersey and Holstein-Friesian cows with good, poor, or noncoagulation properties as reflected in protein genetic variants and isoforms. J Dairy Sci. 2012;95:6905–17. https://doi.org/10.3168/jds.2012-5675.
15. Miranda G, Bianchi L, Krupova Z, Trossat P, Martin P. An improved LC–MS method to profile molecular diversity and quantify the six main bovine milk proteins, including genetic and splicing variants as well as post-translationally modified isoforms. Food Chemistry: X. 2020;5:100080. https://doi.org/10.1016/j.fochx.2020.100080.
16. Poulsen NA, Rosengaard AK, Szekeres BD, Gregersen VR, Jensen HB, Larsen LB. Protein heterogeneity of bovine β-casein in Danish dairy breeds and association of rare β-casein F with milk coagulation properties. Acta Agriculturae Scandinavica, Section A — Anim Sci. 2016;66:190–8. https://doi.org/10.1080/09064702.2017.1342858.
17. Heck JML, Schennink A, Van Valenberg HJF, Bovenhuis H, Visker MHPW, Van Arendonk JAM, et al. Effects of milk protein variants on the protein

composition of bovine milk. J Dairy Sci. 2009;92:1192–202. https://doi.org/10.3168/jds.2008-1208.
18. Meier S, Korkuć P, Arends D, Brockmann GA. DNA Sequence Variants and Protein Haplotypes of Casein Genes in German Black Pied Cattle (DSN). Front Genet. 2019;10. https://doi.org/10.3389/fgene.2019.01129.
19. Lewerentz F, Vanhala TK, Buhelt Johansen L, Paulsson M, Glantz M, de Koning DJ. Re-sequencing of the casein genes in Swedish Red cattle giving milk with diverse protein profiles and extreme rennet coagulation properties. 2024. https://doi.org/10.3168/jdsc.2023-0412.
20. Hayes BJ, Daetwyler HD. 1000 bull genomes project to map simple and complex genetic traits in cattle: applications and outcomes. Annu Rev Anim Biosci. 2019;7:89–102. https://doi.org/10.1146/annurev-animal-020518-115024.
21. Szymanowska M, Siadkowska E, Lukaszewicz M, Zwierzchowski L. Association of nucleotide-sequence polymorphism in the 5′-flanking regions of bovine casein genes with casein content in cow's milk. Dairy Sci Technol. 2004;84:579–90.
22. Keating AF, Davoren P, Smith TJ, Ross RP, Cairns MT. Bovine κ-casein gene promoter haplotypes with potential implications for milk protein expression. J Dairy Sci. 2007;90:4092–9. https://doi.org/10.3168/jds.2006-687.
23. Cai W, Cole JB, Goddard ME, Li J, Zhang S, Song J. Mammary gland multi-omics data reveals new genetic insights into milk production traits in dairy cattle. PLoS Genet. 2025;21:e1011675.
24. Bijl E, Van Valenberg HJF, Huppertz T, Van Hooijdonk ACM, Bovenhuis H. Phosphorylation of αS1-casein is regulated by different genes. J Dairy Sci. 2014;97:7240–6. https://doi.org/10.3168/jds.2014-8061.
25. Fang Z-H, Bovenhuis H, van Valenberg HJF, Martin P, Duchemin SI, Huppertz T, et al. Genome-wide association study for αS1-and αS2-casein phosphorylation in Dutch Holstein Friesian. J Dairy Sci. 2019;102:1374–85. https://doi.org/10.3168/jds.2018-15593.
26. Horne DS. Casein micelle structure and stability. In: Milk proteins. Elsevier; 2020. pp. 213–50. https://doi.org/10.1016/B978-0-12-405171-3.00006-4.
27. McMahon DJ, Oommen BS. Casein micelle structure, functions, and interactions. In: Advanced Dairy Chemistry: Volume 1A: Proteins: Basic Aspects, 4th Edition. Springer; 2012. pp. 185–209. https://doi.org/10.1007/978-1-4614-4714-6_6.
28. Manguy J, Shields DC. Implications of kappa-casein evolutionary diversity for the self-assembly and aggregation of casein micelles. R Soc Open Sci. 2019;6:190939. https://doi.org/10.1098/rsos.190939.
29. Huppertz T, Fox PF, Kelly AL. The caseins: Structure, stability, and functionality. In: Proteins in Food Processing, Second Edition. Elsevier; 2017. pp. 49–92. https://doi.org/10.1016/B978-0-08-100722-8.00004-8.
30. Bovenhuis H, Van Arendonk JAM. Estimation of milk protein gene frequencies in crossbred cattle by maximum likelihood. J Dairy Sci. 1991;74:2728–36. https://doi.org/10.3168/jds.S0022-0302(91)78452-X.
31. Sanchez MP, Fritz S, Patry C, Delacroix-Buchet A, Boichard D. Frequencies of milk protein variants and haplotypes estimated from genotypes of more than 1 million bulls and cows of 12 French cattle breeds. J Dairy Sci. 2020;103:9124–41. https://doi.org/10.3168/jds.2020-18492.
32. Lien S, Kantanen J, Olsaker I, Holm LE, Eythorsdottir E, Sandberg K, et al. Comparison of milk protein allele frequencies in nordic cattle breeds. Anim Genet. 1999;30:85–91. https://doi.org/10.1046/j.1365-2052.1999.00434.x.
33. Ibeagha-Awemu EM, Prinzenberg EM, Jann OC, Lühken G, Ibeagha AE, Zhao X, et al. Molecular characterization of bovine CSN1S2*B and extensive Distribution of zebu-specific milk protein alleles in European cattle. J Dairy Sci. 2007;90:3522–9. https://doi.org/10.3168/jds.2006-679.
34. Chen N, Cai Y, Chen Q, Li R, Wang K, Huang Y, et al. Whole-genome resequencing reveals world-wide ancestry and adaptive introgression events of domesticated cattle in East Asia. Nat Commun. 2018;9. https://doi.org/10.1038/s41467-018-04737-0.
35. Kim K, Kim D, Hanotte O, Lee C, Kim H, Jeong C. Inference of admixture origins in indigenous African Cattle. Mol Biol Evol. 2023;40. https://doi.org/10.1093/molbev/msad257.
36. Visker MHPW, Dibbits BW, Kinders SM, van Valenberg HJF, van Arendonk JAM, Bovenhuis H. Association of bovine β-casein protein variant I with milk production and milk protein composition. Anim Genet. 2011;42:212–8. https://doi.org/10.1111/j.1365-2052.2010.02106.x.
37. Bonfatti V, Grigoletto L, Cecchinato A, Gallo L, Carnier P. Validation of a new reversed-phase high-performance liquid chromatography method for separation and quantification of bovine milk protein genetic variants. J Chromatogr A. 2008;1195:101–6. https://doi.org/10.1016/j.chroma.2008.04.075.
38. Erhardt G. Detection of a new k-casein variant in milk of Pinzgauer cattle. Anim Genet. 1996;27:105–8.
39. Rando A, Di Gregorio P, Ramunno L, Mariani P, Fiorella A, Senese C, et al. Characterization of the CSN1AG Allele of the Bovine αs1-Casein Locus by the Insertion of a Relict of a Long Interspersed Element1. J Dairy Sci. 1998;81:1735–42. https://doi.org/10.3168/jds.S0022-0302(98)75741-8.
40. Prinzenberg E-M, Erhardt G. A new CSN3 allele in *Bos indicus* cattle is characterised by MspI PCR-RFLP. Anim Genet. 1999;30:164.
41. Mohr U, Koczanb D, Linder D, Hobomb G, Erhardt G. A single point mutation results in A allele-specific exon skipping in the bovine aS1-casein mRNA. 1994. https://doi.org/10.1016/0378-1119(94)90095-7.
42. Stamm S, Ben-Ari S, Rafalska I, Tang Y, Zhang Z, Toiber D, et al. Function of alternative splicing. Gene. 2005;344:1–20. https://doi.org/10.1016/j.gene.2004.10.022.
43. Senocq D, Mollé D, Pochet S, Léonil J, Dupont D, Levieux D. A new bovine beta-casein genetic variant characterized by a Met93Leu substitution in the sequence A2. Lait. 2002;82:171–80. https://doi.org/10.1051/lait:2002002.
44. Han SK, Shin YC, Byun HD. Biochemical, molecular and physiological characterization of a new β-casein variant detected in Korean Cattle. Anim Genet. 2000;31:49–51. https://doi.org/10.1046/j.1365-2052.2000.00582.x.
45. Voglino G. A new beta-casein variant in Piedmont cattle. Anim Blood Groups Biochem Genet. 2009;3:61–2. https://doi.org/10.1111/j.1365-2052.1972.tb01233.x.
46. Dong C, Ng-Kwai-Hang KF. Characterization of a non-electrophoretic genetic variant of β-casein by peptide mapping and mass spectrometric analysis. Int Dairy J. 1998;8:967–72. https://doi.org/10.1016/S0958-6946(99)00019-9.
47. European Food Safety Authority. Review of the potential health impact of β-casomorphins and related peptides. EFSA J. 2009;7:231r. https://doi.org/10.2903/j.efsa.2009.231r.
48. Bonfatti V, Chiarot G, Carnier P. Glycosylation of κ-casein: genetic and nongenetic variation and effects on rennet coagulation properties of milk. J Dairy Sci. 2014;97:1961–9. https://doi.org/10.3168/jds.2013-7418.
49. Gai N, Uniacke-lowe T, O'regan J, Faulkner H, Kelly AL. Effect of protein genotypes on physicochemical properties and protein functionality of bovine milk: a review. Foods. 2021;10. https://doi.org/10.3390/foods10102409.
50. Fitzpatrick CJ, Freitas D, O'Callaghan TF, O'Mahony JA, Brodkorb A. Variations in bovine milk proteins and processing conditions and their effect on protein digestibility in humans: a review of in vivo and in vitro studies. Foods. 2024;13. https://doi.org/10.3390/foods13223683.
51. Meza-Nieto MA, González-Córdova AF, Piloni-Martini J, Vallejo-Cordoba B. Effect of β-lactoglobulin A and B whey protein variants on cheese yield potential of a model milk system. J Dairy Sci. 2013;96:6777–81. https://doi.org/10.3168/jds.2012-5961.
52. Bland JH, Grandison AS, Fagan CC. Effect of blending Jersey and Holstein-Friesian milk on Cheddar cheese processing, composition, and quality. J Dairy Sci. 2015;98:1–8. https://doi.org/10.3168/jds.2014-8433.
53. De Poi R, De Dominicis E, Gritti E, Fiorese F, Saner S, Polverino de Laureto P. Development of an LC-MS method for the identification of β-casein genetic variants in bovine milk. Food Anal Methods. 2020;13:2177–87. https://doi.org/10.1007/s12161-020-01817-0.
54. Buzás H, Székelyhidi R, Szafner G, Szabó K, Süle J, Bukovics S, et al. Developed rapid and simple RP-HPLC method for simultaneous separation and quantification of bovine milk protein fractions and their genetic variants. Anal Biochem. 2022;658. https://doi.org/10.1016/j.ab.2022.114939.
55. Heck JML, Olieman C, Schennink A, van Valenberg HJF, Visker MHPW, Meuldijk RCR, et al. Estimation of variation in concentration, phosphorylation and genetic polymorphism of milk proteins using capillary zone electrophoresis. Int Dairy J. 2008;18:548–55. https://doi.org/10.1016/j.idairyj.2007.11.004.
56. Browning SR, Browning BL. Haplotype phasing: existing methods and new developments. Nat Rev Genet. 2011;12:703–14. https://doi.org/10.1038/nrg3054.
57. Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. Score tests for association between traits and haplotypes when linkage phase Is ambiguous. Am J Hum Genet. 2002;70:425–34. https://doi.org/10.1086/338688.
58. Gonzalez-Prendes R, Ginja C, Kantanen J, Ghanem N, Kugonza DR, Makgahlela ML, et al. Integrative QTL mapping and selection signatures in Groningen White Headed cattle inferred from whole-genome sequences. PLoS One. 2022;17 . https://doi.org/10.1371/journal.pone.0276309.
59. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:12073907. 2012. https://doi.org/10.48550/arXiv.1207.3907.

60. Browning BL, Tian X, Zhou Y, Browning SR. Fast two-stage phasing of large-scale sequence data. Am J Hum Genet. 2021;108:1880–90. https://doi.org/10.1016/j.ajhg.2021.08.005.
61. Rosen BD, Bickhart DM, Schnabel RD, Koren S, Elsik CG, Tseng E, et al. De novo assembly of the cattle reference genome with single-molecule sequencing. Gigascience. 2020;9:giaa021. https://doi.org/10.1093/gigascience/giaa021.
62. Cezard T, Cunningham F, Hunt SE, Koylass B, Kumar N, Saunders G, et al. The European Variation Archive: a FAIR resource of genomic variation for all species. Nucleic Acids Res. 2022;50:D1216–20. https://doi.org/10.1093/nar/gkab960.
63. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The ensembl variant effect predictor. Genome Biol. 2016;17:1–14. https://doi.org/10.1186/s13059-016-0974-4.
64. Danecek P, McCarthy SA. BCFtools/csq: haplotype-aware variant consequences. Bioinformatics. 2017;33:2037–9. https://doi.org/10.1093/bioinformatics/btx100.
65. Sinnwell JP, Schaid DJ. Haplo Stats (version 1.7. 7) statistical methods for haplotypes when linkage phase is ambiguous. MN, USA: Mayo Clinic Division of Health Sciences Research Rochester; 2016.

Publisher's Note