Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Polymorphisms in the Kappa-Casein Gene Across Diverse Cattle Breeds: In silico Characterization, Genetic Variant and their Implications for Dairy Production

Nkosinathi Xulu

mthinayi123@gmail.com

University of KwaZulu-Natal - Westville Campus https://orcid.org/0000-0002-4855-1755

Isaac A. Sanusi

University of KwaZulu-Natal - Westville Campus

Moses Okpeku

University of KwaZulu-Natal - Westville Campus

Research Article

Keywords: kappa-casein, CSN3, cattle breeds, in silico analysis, nucleotide diversity, bovine genetics

Posted Date: November 10th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11406732/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Abstract

This study investigated polymorphism and evolutionary relationships at the *CSN3* locus across selected cattle breeds and related bovine taxa using a comparative in silico approach. Publicly available GenBank sequences representing *Bos taurus* alleles and pseudogene-like sequences, *Bos indicus*, *Bos gaurus*, *Bos javanicus*, *Bison bison*, a *Bos taurus* × *Bos indicus* hybrid, and *Bubalus* species were analysed to assess sequence variation, nucleotide composition, codon usage, genetic diversity, and phylogenetic relationships. Sequence alignment and pairwise comparisons indicated substantial conservation across the genus *Bos*, alongside notable variation among *B. taurus* sequences. Among the functional allelic forms examined, the A and B alleles exhibited comparatively greater sequence divergence, whereas the E allele showed high sequence similarity to the A allele. The sequences designated AA and BB displayed marked divergence from the functional allelic forms, although their non-functional status requires confirmation through additional evidence. Phylogenetic analysis revealed close relationships among the sampled *Bos* taxa, while *Bubalus* formed a more distinct lineage, consistent with broader evolutionary relationships among bovine taxa. Nucleotide diversity and the distribution of segregating sites further indicated substantial sequence variation within the sampled *B. taurus* sequences. These findings suggest that the *CSN3* locus is relatively conserved across the examined taxa but exhibits sufficient variation within *B. taurus* to warrant further investigation as a candidate marker for dairy cattle characterisation and breeding. The study provides a comparative sequence-based framework for investigating *CSN3* diversity and its potential associations with milk composition and processing properties. However, the practical utility of the identified variants for breeding decisions requires validation using phenotype-linked datasets and larger, representative cattle populations.

1. Introduction

Milk proteins are among the most economically important components of bovine milk because they determine both nutritional value and technological performance during dairy processing. The casein fraction accounts for approximately 80% of total milk protein and consists primarily of αs1-casein, αs2-casein, β-casein, and κ-casein, which together form the casein micelle structure responsible for milk stability and coagulation properties (Eigel et al., 1984; Huppertz et al., 2018; Kelly et al., 2025). Genetic variation within casein genes has therefore attracted considerable attention in dairy science, particularly in relation to milk composition, cheese-making efficiency, and the production of value-added dairy products. Among these genes, the kappa-casein gene (CSN3) has been widely recognized as a major candidate gene influencing milk coagulation and cheese yield.

The CSN3 gene is located within the casein gene cluster on bovine chromosome 6 (Amalfitano et al., 2022) and encodes κ-casein, a glycoprotein that stabilizes casein micelles and regulates the coagulation process during cheese production. During enzymatic coagulation, κ-casein is cleaved by chymosin, leading to micelle destabilization and curd formation (Waugh & von Hippel, 1956; Kayihura et al., 2024). Consequently, polymorphisms within CSN3 can significantly affect coagulation time, curd firmness, cheese yield, and other processing characteristics that are important for the dairy industry (Aleandri et

al., 1990; Ikonen et al., 1999; Lucey 2002; ; Comin et al., 2008; Mahmoudi et al., 2020). Numerous studies have demonstrated that specific allelic variants of CSN3, particularly the A and B alleles, are associated with differences in milk protein content and technological performance, with the B allele frequently linked to improved cheese-making properties (Bartonova et al., 2012).

More than a dozen genetic variants of bovine CSN3 have been reported (Fan et al., 2019), although the A and B alleles remain the most extensively studied in commercial dairy cattle populations. These variants arise from single nucleotide polymorphisms that alter the amino acid sequence of κ-casein and may influence protein structure and functionality (Grosclaude et al., 1972). In addition to the common functional alleles, several rare variants and pseudogene-like sequences have been identified, highlighting the evolutionary complexity of the locus. The distribution of CSN3 alleles varies among breeds, with marked differences observed between Bos taurus, Bos indicus, and indigenous cattle populations (Bangar et al., 2021; Carles and Martin 1985; Prinzenberg et al., 2005; Schopen et al., 2011). Such breed-specific variation provides valuable opportunities for genetic improvement through marker-assisted and genomic selection (Amalfitano et al., 2022).

Comparative analysis of CSN3 across bovine species and related taxa offers additional insight into the evolutionary conservation and diversification of milk protein genes. Species within the genus Bos, including Bos gaurus, Bos javanicus, and domestic cattle, share a common evolutionary origin and generally exhibit high sequence similarity in functional genes associated with lactation (Chessa et al., 2008). In contrast, buffalo species such as Bubalus arnee represent a more divergent lineage within the family Bovidae, making them useful for evaluating the extent of conservation at the CSN3 locus. Understanding these evolutionary relationships is important not only for molecular systematics but also for identifying conserved and variable regions that may influence protein function (Pinder et al., 1991; Hassanin et al., 2022)

Advances in molecular genetics and bioinformatics have greatly facilitated the investigation of candidate genes using publicly available genomic databases. In silico approaches enable researchers to compare nucleotide sequences, assess genetic diversity, examine codon usage patterns, and reconstruct phylogenetic relationships without the need for extensive laboratory experimentation. Such analyses are particularly valuable for genes like CSN3, where large numbers of sequences from different breeds and species are accessible through repositories such as GenBank (Benson et al., 2013). Comparative sequence analysis can reveal the degree of conservation among taxa, identify polymorphic sites, estimate nucleotide diversity, and infer evolutionary processes shaping the gene. These approaches complement traditional dairy genetics research and provide a cost-effective framework for preliminary characterization of functional genetic variation (Rohlf et al., 2004).

Despite the substantial body of research on CSN3 polymorphisms in dairy cattle, several knowledge gaps remain. Many previous studies have focused primarily on a limited number of commercial Bos taurus breeds and have concentrated on the association between the A and B alleles and milk production traits. Comparatively fewer investigations have examined CSN3 variation across a broader

range of bovine taxa, including related wild species, hybrids, and buffalo lineages. Furthermore, the evolutionary relationships among functional alleles, rare variants, and pseudogene-like sequences have not been comprehensively explored within a single comparative framework. Addressing these gaps could improve our understanding of the genetic architecture of CSN3 and its potential utility in diverse breeding systems (Huppertz et al., 2018; Kayihura (2024). Many previous studies have focused primarily on a limited number of commercial Bos taurus breeds and have concentrated on the association between the A and B alleles and milk production traits (Bangar et al., 2021; Mahmoudi et al., 2020).

Therefore, the present study aimed to investigate polymorphisms within the CSN3 gene across diverse cattle breeds and related bovine taxa using a comparative in silico approach. Publicly available GenBank sequences representing Bos taurus alleles and pseudogene-like forms, Bos indicus, Bos gaurus, Bos javanicus, Bison bison, a Bos taurus × Bos indicus, and Bubalus arnee / Bubalus bubalis were analysed to characterize sequence variation, phylogenetic relationships, nucleotide composition, codon usage, and genetic diversity. Also, by integrating comparative genomics with evolutionary analysis, this study sought to determine the extent of conservation and polymorphism at the CSN3 locus and to evaluate its continued relevance as a molecular marker for improving milk composition and dairy processing characteristics in cattle breeding programs.

2. Materials and methods

2.1 Sequence dataset

This study used a comparative in silico design to characterize polymorphisms in the bovine kappa-casein gene (CSN3) across diverse cattle breeds and related bovine taxa. A total of 11 publicly available NCBI nucleotide sequences that share substantial nucleotide homology within the CSN3 gene region, particularly the conserved region encompassing exon 4 and the corresponding partial coding sequence was selected. This high degree of sequence conservation reflected the common evolutionary origin of the CSN3 gene among the selected bovine and related taxa, while the observed nucleotide substitutions provide a basis for investigating interspecific and allelic variation in kappa-casein Table 1. The dataset included both functional alleles and pseudogene-like forms, with sequence lengths ranging from approximately 240 to more than 600 bp. Public sequence data were selected to capture both allelic variation within Bos taurus and broader interspecies divergence within bovids.

2.2 Sequence retrieval and curation

CSN3 nucleotide sequences were retrieved from the NCBI (https://www.ncbi.nlm.nih.gov/genbank/) in FASTA format. The accession number of each nucleotide is presented in Table 1. Retrieved records were inspected for annotation consistency, sequence length, and relevance to the CSN3 locus before inclusion in the analysis. Partial and longer sequences were retained when they contributed meaningfully to comparisons across taxa. The curated dataset was then prepared for downstream alignment, compositional analysis, and phylogenetic reconstruction.

2.3 Alignment and polymorphism detection

Nucleotide sequences were aligned to identify substitutions and allelic differences across the CSN3 region. The aligned dataset was examined for single nucleotide polymorphisms (SNPs), sequence identity, and divergence among Bos taurus variants and among species. Within Bos taurus, pairwise comparisons highlighted the strong contrast between variants A and B, as well as the near identity of the E allele to variant A. Highly divergent pseudogene-like sequences were also identified on the basis of reduced similarity to the functional alleles.

2.4 Phylogenetic analysis

Evolutionary relationships among the CSN3 sequences were inferred using the UPGMA clustering method based on sequence similarity. The resulting tree was used to assess grouping patterns among Bos species and to evaluate the separation of Bubalus arnee from the Bos lineage. Sequence identity clusters were interpreted to distinguish closely related alleles, moderate within-genus relationships, and highly divergent outgroup sequences. The phylogenetic patterns were used to support evolutionary interpretation of the observed sequence variation (Rohlf et al., 2004).

2.5 Nucleotide composition and GC analysis

Base composition was calculated for each sequence to determine the proportion of adenine, thymine, guanine, and cytosine. GC content and AT content were compared across species and allele classes to evaluate compositional bias and lineage differentiation. The analysis showed an AT-rich pattern across Bos species, while Bubalus arnee exhibited a higher GC content than the Bos sequences. These compositional metrics were used as supplementary evidence of evolutionary divergence and constraint.

2.6 Codon usage analysis

Coding sequences were evaluated for codon usage patterns by enumerating observed codons and calculating their relative frequencies. The analysis identified frequently used codons such as AGA, TAC, and CAA in the Bos taurus and hybrid sequences. Codon frequency distribution was compared across single-sequence and multi-sequence taxa to assess the presence of codon bias. These results were interpreted in relation to sequence conservation and the maintenance of protein-coding function.

2.7 Statistical analysis

Genetic diversity was summarized using nucleotide diversity (π) and the number of segregating sites, calculated for taxa with more than one available CSN3 sequence, particularly Bos taurus. Pairwise sequence identity percentages were used to quantify divergence among alleles and species, while descriptive comparisons of GC content and codon frequency were used to summarize compositional differences. Because the study was exploratory and based on a limited number of publicly available sequences, results were presented primarily as descriptive statistics rather than inferential hypothesis testing. The combination of nucleotide diversity, segregating sites, identity values, and phylogenetic

clustering provided the basis for interpreting polymorphism patterns in relation to dairy production relevance.

Table 1
Publicly Available NCBI Nucleotide Sequences Analysed for Comparative CSN3 Analysis

No.	NCBI Accession	Species/Taxon	Variant / Allele	Sequence description
1	U84254.1	*Bison bison*	Variant 1	Kappa-casein gene, partial CDS
2	U84253.1	*Bos gaurus*	Not specified	Kappa-casein gene, partial CDS
3	KR149429.1	*Bos indicus*	CSN3-B	Exon 4 and partial CDS
4	KR149430.1	*Bos taurus × Bos indicus*	CSN3-A	Exon 4 and partial CDS
5	U84252.1	*Bos javanicus*	Not specified	Kappa-casein gene, partial CDS
6	U84250.1	*Bos taurus*	Variant A	Kappa-casein gene, partial CDS
7	U84251.1	*Bos taurus*	Variant B	Kappa-casein gene, partial CDS
8	AF041482.1	*Bos taurus*	CSN3-E	Kappa-casein gene, partial CDS
9	ON921315.1	*Bos taurus*	CSN3-AA	Partial sequence
10	ON921316.1	*Bos taurus*	CSN3-BB	Partial sequence
11	U96662.1	*Bubalus arnee/bubalis*	Not specified	Partial exon 4 and partial CDS

3. Results

3.1 Dataset characteristics and sequence

A total of 11 CSN3 nucleotide sequences representing diverse bovine taxa were retrieved from GenBank for comparative analysis. The dataset comprised five Bos taurus variants (A, B, E, AA and BB), together with sequences from Bos indicus, Bos gaurus, Bos javanicus, Bison bison, a Bos taurus × Bos indicus hybrid, and Bubalus arnee. Sequence lengths ranged from 257 bp to 633 bp, with nine sequences measuring approximately 300 bp and two partial genomic sequences exceeding 600 bp, resulting in an average sequence length of approximately 379 bp. The distribution of sequence lengths Table 2 showed that most sequences represented comparable coding regions, whereas the Bos indicus and Bos taurus × Bos indicus hybrid sequences were considerably longer. Bos taurus contributed the largest proportion of sequences (n = 5), enabling subsequent analyses of within-species polymorphism, while all remaining taxa were represented by single sequences.

Table 2
Comparative Sequence Characteristics and Genetic Diversity of the κ-Casein (CSN3) Gene Among Selected Bovine Species

Metric	Bos taurus	Bos indicus	Bubalus arnee
Sequence Length (bp)	298	633	257
GC Content (%)	38.93	40.44	42.02
A + T Content (%)	61.07	59.56	57.98
Codon Count	44	55	43
Unique Codons	~ 40	~ 50	~ 40
Segregating Sites†	226	N/A	N/A
Nucleotide Diversity (π)†	0.5863	N/A	N/A
Sequence Identity to *Bos gaurus* (%)	53.70	N/A	N/A
Sequence Identity to *Bubalus arnee* (%)	23.00	N/A	25.68

3.2 Nucleotide composition and GC content

The nucleotide composition of the representative Bos taurus variant A sequence was characterized by an adenine-rich profile, consisting of 122 adenine (40.9%), 60 thymine (20.1%), 51 guanine (17.1%), and 65 cytosine (21.8%) nucleotides (Fig. 1). Overall, AT content exceeded GC content, indicating an AT-biased coding sequence.

Across taxa, GC content remained highly conserved within the genus Bos, varying only slightly between approximately 39.1% and 39.5%. In contrast, Bubalus arnee exhibited the highest GC content (42.02%), while the hybrid sequence showed an intermediate GC content of approximately 40.5% (Fig. 1D). These results indicate relatively stable nucleotide composition among Bos species, with greater compositional divergence observed in the buffalo lineage.

3.3 Cross-species sequence conservation

Pairwise sequence comparisons demonstrated extensive conservation of the CSN3 locus across the genus Bos (Fig. 2). Using Bos taurus variant A as the reference sequence, Bison bison and Bos javanicus each exhibited 98.7% sequence identity, whereas Bos gaurus displayed 98.3% identity. These high similarity values indicate strong evolutionary conservation of the coding region among closely related bovine species.

By contrast, Bubalus arnee exhibited only 29.3% sequence identity relative to the Bos taurus reference, representing the greatest divergence among all taxa examined and clearly separating the buffalo lineage from members of the genus Bos. The sequence identity heatmap further confirmed these relationships

by identifying a highly conserved cluster comprising Bos taurus, Bos javanicus, Bos gaurus, and Bison bison, whereas Bubalus arnee consistently formed the most divergent sequence cluster.

3.4 Intra-species polymorphism in Bos taurus

Comparison of the five Bos taurus CSN3 variants revealed substantial allelic diversity (Fig. 3). Variant A served as the reference sequence and showed no nucleotide differences from itself. The B allele differed from variant A by 139 SNPs, representing approximately 46.6% of comparable nucleotide positions. In contrast, the E allele differed by only one nucleotide substitution (0.3%), indicating almost complete sequence conservation between variants A and E.

Markedly greater divergence was observed for the AA and BB sequences, which differed from variant A by 183 SNPs (74.1%) and 178 SNPs (71.8%), respectively. These highly divergent sequences exhibited substantially reduced similarity to the functional alleles, consistent with pseudogene-like characteristics. The logarithmic SNP distribution (Fig. 3) further emphasizes the marked contrast between the minimal divergence of the E allele and the extensive divergence of the AA and BB sequences.

3.5 Phylogenetic relationships

The UPGMA phylogenetic analysis resolved the CSN3 sequences into distinct evolutionary clusters (Fig. 4). The shortest branch length occurred between Bos taurus variant A and the E allele, confirming their extremely close evolutionary relationship. A second highly conserved cluster included Bison bison, Bos javanicus, Bos gaurus, and Bos taurus variant B, indicating minimal evolutionary divergence among these taxa.

The AA and BB sequences formed an independent cluster separated from the functional alleles by comparatively long branch lengths, reflecting their extensive sequence divergence. The greatest phylogenetic separation was observed for Bubalus arnee, which occupied the basal branch of the tree and exhibited the largest evolutionary distance from all Bos species. Overall, evolutionary distances were considerably lower within the genus Bos than between Bos and Bubalus, reflecting strong conservation within cattle and greater divergence between genera.

3.6 Sequence identity matrix

The pairwise sequence identity matrix (Fig. 5) supported the phylogenetic relationships observed in the dendrogram. The highest pairwise identity (100%) was observed between Bos taurus variant A and the E allele, confirming their near-identical nucleotide sequences. Similarly, identities exceeding 98% were observed among Bos taurus variant B, Bison bison, and Bos javanicus, indicating strong conservation within this group.

Intermediate sequence identities (approximately 50–75%) characterized comparisons involving Bos gaurus, while identities involving the AA and BB sequences were consistently below 45%, demonstrating

their marked divergence from functional CSN3 alleles. The lowest pairwise identities (23–29%) occurred in comparisons involving Bubalus arnee, confirming its distinct evolutionary position.

3.7 Genetic diversity

Genetic diversity analyses were possible only for Bos taurus, as all other taxa were represented by a single sequence. The Bos taurus dataset exhibited a nucleotide diversity of $\pi = 0.5863$, together with 226 segregating sites, indicating substantial sequence variation among alleles (Fig. 6D). These values demonstrate extensive allelic polymorphism at the CSN3 locus within cattle. No within-species diversity estimates were calculated for the remaining taxa because only one representative sequence was available for each species.

4. Discussion

The present study investigated sequence characteristics, nucleotide composition, sequence variation, pairwise sequence identity, nucleotide diversity and phylogenetic relationships among publicly available κ-casein (CSN3) sequences representing Bos taurus, Bos indicus, Bos javanicus, Bos gaurus, Bison bison and Bubalus arnee. The analysis provides a comparative assessment of sequence characteristics that may contribute to understanding the evolutionary conservation and genetic variation of CSN3 in domesticated cattle and related bovids. The biological importance of CSN3 arises from the role of κ-casein in the structure and stability of the casein micelle and from the documented effects of κ-casein genetic variation on milk composition and technological properties. Consequently, sequence-level comparisons of CSN3 can provide useful information for understanding the conservation of the gene among related taxa and for identifying variants that may be relevant to milk production and processing characteristics (Mahmoudi et al., 2020).

The sequence characteristics presented in Table 1 demonstrate differences among the analysed taxa in sequence length and nucleotide composition. The Bos taurus, Bos indicus and Bubalus arnee sequences were reported as 298, 633 and 257 bp, respectively, while the corresponding GC contents were 38.93%, 40.44% and 42.02%. The complementary A + T proportions were 61.07%, 59.56% and 57.98%, respectively. The nucleotide composition calculated for the B. taurus sequence is internally consistent, because the reported nucleotide counts sum to 298 bp and the calculated GC content of 38.93% corresponds to the combined G and C nucleotides. These results indicate that the analysed B. taurus fragment is relatively A + T-rich. However, the differences in sequence length among the taxa require careful interpretation because sequence length variation can influence pairwise identity, SNP counts and estimates of nucleotide diversity if comparisons are not restricted to homologous aligned positions. Therefore, the reported differences in sequence composition should initially be interpreted as characteristics of the particular sequence fragments analysed rather than as definitive characteristics of the complete CSN3 gene (Fan et al., 2019).

The reported codon counts require particular attention. The table gives codon counts of 44, 55 and 43 for B. taurus, B. indicus and B. arnee, respectively, despite sequence lengths of 298, 633 and 257 bp. If the complete nucleotide sequences were translated, the expected numbers of complete codons would be approximately 99, 211 and 85, respectively, before consideration of start and stop codons. Consequently, the reported codon counts cannot represent the total number of codons in the respective nucleotide sequences unless only specific coding regions or a subset of codons were analysed. This distinction should be clarified in the methodology and results. Failure to clarify the definition of "codon count" could create uncertainty regarding whether the sequences were analysed as complete coding sequences, partial sequences, or selected aligned regions (Kelly et al., 2025).

The nucleotide composition results are nevertheless useful for describing the analysed B. taurus sequence. The sequence contains 122 adenine, 65 cytosine, 51 guanine and 60 thymine nucleotides, resulting in 40.9%, 21.8%, 17.1% and 20.1%, respectively. The combined GC content is therefore 38.93%, while the combined A + T content is 61.07%. This pattern demonstrates that adenine and thymine constitute the majority of nucleotides in the analysed fragment. Nevertheless, nucleotide composition alone does not provide evidence that the observed composition has a direct effect on milk production or protein function. Such interpretations would require analyses linking sequence composition to coding changes, amino-acid substitutions, regulatory elements or experimentally measured phenotypes (Mahmoudi et al., 2020).

The comparative sequence analysis indicates high similarity among several members of the genus Bos. The close clustering of Bos sequences is biologically plausible and is consistent with previous investigations of CSN3 evolution. For example, Chessa et al. investigated genetic variation in CSN3 and phylogenetic relationships within the genus Bos, demonstrating substantial conservation of the locus among bovine taxa while also identifying lineage-specific variation. Their findings support the use of CSN3 as a useful molecular marker for investigating relationships among cattle and closely related bovids. Similarly, comparative analysis of buffalo κ-casein demonstrated that buffalo CSN3 shares substantial nucleotide and amino-acid similarity with cattle, and phylogenetic analysis placed buffalo κ-casein with cattle rather than with more distantly related livestock species (Rout et al., 2006). The latter study reported approximately 95.5% nucleotide and 92.6% amino-acid similarity between buffalo and cattle κ-casein sequences, demonstrating that evolutionary separation between Bos and Bubalus does not imply extreme divergence of the CSN3 coding sequence (Fan et al., 2019).

This published evidence is particularly important when interpreting the relatively low Bos taurus Bubalus arnee sequence identities reported in the present analysis. The current table reports 23.00% identity between B. taurus and B. arnee, while the corresponding figure reports a value of approximately 29.3%. These values are substantially lower than the approximately 95.5% nucleotide similarity reported between buffalo and cattle κ-casein by Rout et al. (2006). Therefore, the low values in the present analysis should not be interpreted as demonstrating that buffalo CSN3 is only approximately 23–29% identical to bovine CSN3. A more scientifically defensible interpretation is that the comparison requires verification, particularly with respect to the accession numbers, sequence orientation, sequence type,

aligned region and treatment of gaps. Differences in the lengths and regions of the sequences can produce misleading pairwise identity estimates when non-homologous regions are compared. The most appropriate comparison would be based on homologous aligned nucleotide positions shared across the sequences (Fan et al., 2019).

The analysis of the bovine κ-casein A and B variants is also important because these are among the best-characterised CSN3 polymorphisms. The current analysis reports a substantially large number of nucleotide differences between some of the sequences labelled as CSN3 A and B. However, the established molecular distinction between the classical bovine κ-casein A and B variants is based on two point mutations affecting amino-acid positions 136 and 148. Mercier et al. (1973) demonstrated that the two common bovine κ-casein variants A and B originate from two-point mutations in the codons corresponding to amino-acid positions 136 and 148. A subsequent study also confirmed that the two amino-acid substitutions distinguishing κ-casein A and B occur at positions 136 and 148 (Caroli et al., 2010). Therefore, if the sequences labelled A and B in the present analysis are intended to represent conventional bovine CSN3A and CSN3B coding sequences, the reported very large numbers of differences should be treated as a quality-control issue rather than as evidence of genuine divergence between the two alleles.

This observation is particularly important because the current results report 226 segregating sites and a nucleotide diversity value of $\pi = 0.5863$ for B. taurus. Nucleotide diversity represents the average number of nucleotide differences per site between pairs of sequences. A value of 0.5863 would indicate an exceptionally high level of sequence variation within the analysed region. If the denominator were the reported 298-bp sequence, the value would correspond to approximately 175 differences per sequence pair on average. Such a level of variation would be difficult to reconcile with a dataset consisting principally of the well-characterised bovine CSN3 allelic variants, particularly because the classical A and B variants are separated by only two principal coding substitutions. The reported 226 segregating sites would likewise represent approximately three quarters of the 298-bp sequence. Consequently, these values should be recalculated from the original aligned sequences before they are interpreted as evidence of exceptionally high CSN3 diversity in cattle.

The results therefore illustrate an important distinction between computational output and biological interpretation. A sequence-analysis programme can produce mathematically valid statistics from the sequences supplied to it, but a numerical result is not necessarily biologically meaningful unless the input sequences are homologous and correctly annotated. In the present case, the high values of segregating sites and nucleotide diversity may indicate genuine sequence diversity only if the sequences represent the same homologous genomic region and have been correctly aligned. Alternatively, they could result from inclusion of non-homologous regions, untranslated regions, different portions of the gene, sequencing artefacts, incorrect accession assignment or an inappropriate alignment. The quality of the final biological interpretation is therefore dependent on the quality and comparability of the input sequences rather than on the numerical output alone.

The UPGMA analysis provides an additional perspective on the relationships among the sequences. The tree broadly separates the Bubalus sequence from the Bos and Bison sequences, while several Bos sequences cluster closely together. This general pattern is consistent with previous comparative CSN3 studies in which cattle and closely related bovids form relatively close evolutionary groups and buffalo occupies a distinct lineage while remaining substantially related to cattle. Rout et al. (2006), for example, reported that buffalo κ-casein clustered with cattle in phylogenetic analysis based on both nucleotide and amino-acid sequences. The broad correspondence between the present tree and previous findings therefore provides some support for the general evolutionary pattern observed in the analysis.

Nevertheless, the UPGMA tree should be interpreted as a clustering representation rather than definitive evidence of phylogenetic history. UPGMA assumes an approximately constant evolutionary rate among the lineages being compared, effectively imposing a molecular-clock or ultrametric assumption. When evolutionary rates differ among lineages, UPGMA can produce inaccurate phylogenetic relationships. Previous methodological work has specifically noted that UPGMA performs poorly when a molecular-clock assumption does not apply. Consequently, the present UPGMA tree would be strengthened by comparison with a maximum-likelihood phylogenetic tree based on an appropriately selected nucleotide substitution model and supported by bootstrap analysis. Agreement between independent phylogenetic methods would provide greater confidence in the inferred clustering.

The pairwise sequence identity analysis provides quantitative support for the clustering pattern observed in the tree, with relatively high similarity among several bovine sequences and greater divergence involving buffalo. However, the unusually low identity values reported for some bovine comparisons should be treated cautiously because they appear inconsistent with the known molecular organisation of bovine CSN3. The established A and B variants are highly similar alleles distinguished by two major coding substitutions rather than by divergence over approximately half of the sequence. Consequently, pairwise identity should be recalculated after all sequences have been aligned across a common homologous region. This recalculation would also allow the SNP count, segregating-site estimate and nucleotide-diversity estimate to be evaluated using the same set of comparable positions.

The biological significance of CSN3 variation is nevertheless well established. κ-casein is an important component of the casein micelle, and genetic variation in CSN3 has been associated with differences in milk protein composition and technological characteristics. In a large meta-analysis involving 30,471 genotyped cattle, Mahmoudi et al. (2020) found that the CSN3 B allele, particularly the BB genotype, was significantly associated with lactation yield under some genetic models. They also identified small effects of CSN3 variants on fat percentage, protein percentage and daily milk yield, although no significant effects were observed for protein or fat yield. Importantly, the effects varied according to the genetic model and trait considered. This demonstrates that the relationship between CSN3 genotype and milk production is real but not uniform across all production traits.

Other studies provide further evidence for the potential functional relevance of CSN3 variation. Heck et al. (2009), using 1,912 Dutch Holstein-Friesian cows, reported associations between κ-casein genotype

and the relative concentrations of several milk proteins, including κ-casein, and found that the B variant was associated with higher protein percentage than the A variant. Their results also indicated that combinations of milk-protein genotypes could have greater relevance to milk composition and cheese-making suitability than a single genetic variant considered independently. Similarly, a study of Chinese Holstein cattle identified several CSN3 SNPs and reported significant associations with milk, protein and fat yields as well as protein and fat percentages, demonstrating that sequence variation within CSN3 can have measurable phenotypic associations in cattle populations. These findings support the potential application of CSN3 variation as a molecular marker, but they do not justify assuming that every sequence difference identified in the present dataset will affect milk production.

The practical application of the present analysis therefore lies primarily in the identification and characterisation of conserved and variable regions of CSN3. Once the sequence data have been verified and the genuine polymorphisms have been distinguished from alignment or sequence-record artefacts, the information could contribute to marker-assisted selection programmes targeting milk composition and processing characteristics. The classical CSN3 A and B variants are already of practical interest because the B allele has been associated in several studies with milk protein characteristics and cheese-making properties. Mercier et al. (1990) specifically noted the relevance of identifying the B allele for breeding programmes aimed at increasing its frequency because of its reported favourable cheese-making characteristics. More recent evidence indicates that CSN3 variation may contribute to milk production and composition, although the effects are generally trait- and population-dependent. Therefore, a validated CSN3 sequence dataset could potentially support breed-specific genotyping, marker-assisted selection, conservation of desirable alleles and further investigation of the molecular basis of milk-quality variation.

The results could also have applications in comparative genomics. The conservation of CSN3 among related bovids can provide information for identifying conserved functional regions, designing comparative molecular markers and investigating the evolutionary history of milk-protein genes. The presence of characteristic amino-acid substitutions among CSN3 variants can additionally provide a basis for studying whether particular substitutions are likely to influence protein structure or function. However, sequence variation alone cannot establish a causal relationship with milk production. Functional validation through genotype–phenotype association studies, protein-level analyses, milk-composition measurements and, where appropriate, structural or biochemical experiments would be required before a particular mutation could be described as functionally causal.

5. Reliability, limitations, research gap and novelty of the analytical model

The reliability of the analytical model should be evaluated in terms of both computational reproducibility and biological validity. At the computational level, several outputs, particularly the nucleotide counts and GC and A + T content calculations for the Bos taurus sequence, appear internally reproducible. However, the biological interpretation of the results remains uncertain because of apparent inconsistencies

among sequence lengths, codon counts, pairwise sequence identity, single-nucleotide polymorphism (SNP) counts, and nucleotide-diversity estimates. For example, the reported 23–29% sequence identity between cattle and buffalo appears substantially lower than expected for closely related Bovinae species; comparative analyses of casein sequences have reported approximately 96.25% similarity between cattle and water buffalo (Hassanin et al., 2022). Similarly, the reported high number of differences between bovine CSN3 A and B variants warrants re-examination, as these variants are conventionally distinguished by two nucleotide substitutions that result in amino-acid changes at positions 136 and 148, whereby variant A carries threonine and aspartic acid and variant B carries isoleucine and alanine, respectively (Li et al., 2025). These discrepancies do not necessarily invalidate the analytical approach, but they raise concerns about sequence selection, data processing, alignment, and parameter specification. Therefore, the current outputs should be interpreted as preliminary computational findings rather than definitive evidence of evolutionary divergence or genetic diversity.

5.1. Validation requirements and analytical improvements

Improving the reliability of the model requires systematic verification of the input sequences and analytical procedures. First, all NCBI accession numbers should be checked against their associated annotations to confirm that the records correspond to CSN3 and to establish whether each represents genomic DNA, messenger RNA (mRNA), a coding DNA sequence (CDS), or a partial sequence (Benson et al., 2013). Second, the sequences should be restricted to a common homologous region before comparative analyses are performed. Multiple-sequence alignment should then be used to identify corresponding nucleotide positions and account for gaps or ambiguous bases. Pairwise sequence identity, SNP counts, segregating sites, and nucleotide diversity should subsequently be recalculated using consistent alignment coordinates and clearly specified treatment of missing data and gaps.

Moreover, where coding sequences are available, translated amino-acid sequences should also be examined to verify the expected κ-casein protein sequence and distinguish synonymous from nonsynonymous substitutions. The established differences between bovine CSN3 A and B variants, particularly the substitutions associated with amino-acid positions 136 and 148, provide useful positive controls for assessing whether the sequence-processing and variant-detection pipeline performs as expected (Li et al., 2025). In variant A, threonine and aspartic acid occur at positions 136 and 148, respectively, whereas variant B carries isoleucine and alanine at these positions (Li et al., 2025). Furthermore, the reported results should be checked for consistency among nucleotide sequence length, codon number, and the number of bases included in each calculation. Given the high sequence similarity reported between cattle and water buffalo casein sequences, substantial discrepancies in pairwise identity would strongly suggest methodological rather than biological explanations (Hassanin et al., 2022). These validation steps would help determine whether the observed discrepancies arise from differences in the source records, analytical procedures, or the underlying biological sequences.

5.2. Methodological and interpretative limitations

Improving the reliability of the model requires systematic verification of the input sequences and analytical procedures. First, all NCBI accession numbers should be checked against their associated annotations to confirm that the records correspond to CSN3 and to establish whether each represents genomic DNA, messenger RNA (mRNA), a coding DNA sequence (CDS), or a partial sequence (Benson et al., 2013). Second, the sequences should be restricted to a common homologous region before comparative analyses are performed. Multiple-sequence alignment should then be used to identify corresponding nucleotide positions and account for gaps or ambiguous bases. Pairwise sequence identity, SNP counts, segregating sites, and nucleotide diversity should subsequently be recalculated using consistent alignment coordinates and clearly specified treatment of missing data and gaps.

Moreover, where coding sequences are available, translated amino-acid sequences should also be examined to verify the expected κ-casein protein sequence and distinguish synonymous from nonsynonymous substitutions. The established differences between bovine CSN3 A and B variants, particularly the substitutions associated with amino-acid positions 136 and 148, provide useful positive controls for assessing whether the sequence-processing and variant-detection pipeline performs as expected (Li et al., 2025). In variant A, threonine and aspartic acid occur at positions 136 and 148, respectively, whereas variant B carries isoleucine and alanine at these positions (Li et al., 2025). Furthermore, the reported results should be checked for consistency among nucleotide sequence length, codon number, and the number of bases included in each calculation. Given the high sequence similarity reported between cattle and water buffalo casein sequences, substantial discrepancies in pairwise identity would strongly suggest methodological rather than biological explanations (Hassanin et al., 2022). These validation steps would help determine whether the observed discrepancies arise from differences in the source records, analytical procedures, or the underlying biological sequences.

5.3. Research gap and potential novelty

A key research gap is the need for a consistently validated, sequence-level comparative framework for assessing CSN3 variation and evolutionary relationships among cattle, buffalo, and other relevant bovids. Comparisons based on heterogeneous public sequence records may produce misleading estimates when accession identity, sequence completeness, homologous coverage, and alignment procedures are not adequately harmonised (Benson et al., 2013). Moreover, nucleotide-level comparisons, variant identification, phylogenetic inference, and the interpretation of potential dairy-related associations are not necessarily integrated into a single, clearly validated analytical workflow.

The potential novelty of the present study therefore lies not simply in comparing CSN3 sequences, but in establishing a reproducible and biologically interpretable framework that integrates accession verification, homologous sequence alignment, nucleotide-composition analysis, SNP identification, nucleotide-diversity estimation, amino-acid translation, and phylogenetic inference. Incorporating the established bovine CSN3 A/B substitutions at amino-acid positions 136 and 148 as positive controls would provide a practical means of verifying the accuracy of sequence processing and variant detection (Li et al., 2025). Similarly, comparing UPGMA-based clustering with a more flexible phylogenetic method, such as maximum-likelihood inference under an appropriate substitution model, could help assess the

robustness of the inferred relationships (Felsenstein, 2004). If supported by the validated results, this approach may help clarify the extent of sequence variation among the sampled taxa and identify candidate variants for subsequent investigation of their potential functional or dairy-related significance.

However, the novelty of this study should remain proportionate to the final analysis and a systematic assessment of previous studies. Until the sequence discrepancies have been resolved, the current findings are insufficient to establish previously unreported evolutionary relationships, novel genetic variants, or new associations with milk-production traits. This caution is particularly important because associations between CSN3 variants and dairy traits can vary across populations, phenotypes, and genetic models (Mahmoudi et al., 2020). The principal contribution at this stage is therefore the identification of a validation requirement and the development of a framework through which more reliable comparative and biological inferences can be obtained.

6. Conclusion

The comparative analysis in the present study suggests that CSN3 is a conserved yet polymorphic milk-protein gene among cattle and related bovids. The broad clustering of *Bos* sequences and their separation from *Bubalus* are generally consistent with established patterns of κ-casein sequence variation and evolutionary relationships. Furthermore, the biological relevance of CSN3 polymorphism is supported by previous studies reporting associations between specific variants and milk composition, lactation performance, and milk-processing properties. Nevertheless, several discrepancies in the present analysis warrant further investigation before the numerical results can be interpreted as reliable evidence of CSN3 diversity. Specifically, the exceptionally low sequence identities reported between cattle and buffalo, the unusually high number of segregating sites, the elevated nucleotide-diversity estimate, and the substantial number of apparent differences between bovine CSN3 A and B variants appear inconsistent with established sequence relationships and may reflect issues related to sequence selection, alignment, or data processing. The most scientifically appropriate conclusion is therefore that the current analysis provides a useful preliminary comparative framework, but that the sequence dataset should be re-aligned and re-analysed using verified homologous CSN3 regions before final biological interpretation. Following validation, the resulting sequence information could provide a robust basis for comparative genomics, identification of functionally relevant CSN3 polymorphisms and future marker-assisted selection studies aimed at improving milk composition and processing characteristics in cattle.

Declarations

Authorship contributions

Nkosinathi H. Xulu: Conceptualization, Methodology, Investigation, Data curation, Writing - original draft, Writing - review & editing. **Isaac A. Sanusi:** Methodology, Investigation, Writing - original draft, Writing - review & editing. **Moses Okpeku:** Conceptualization, Methodology & Supervision**.**

Competing interest statement

The authors declare that they have no known competing interests that could have influenced the work reported in this manuscript.

Data availability statement

Not applicable

Ethics approval

Not applicable

References

1. Aleandri, R., Buttazzoni, L. G., Schneider, J. C., & Caroli, A. (1990). The effects of milk protein polymorphisms on milk components and cheese-producing ability. *Journal of Dairy Science*, *73*(Suppl. 1), 1–19.
2. Amalfitano, N., Mota, L. F. M., Rosa, G. J. M., Cecchinato, A., & Bittante, G. (2022). Role of CSN2, CSN3, and BLG genes and the polygenic background in the cattle milk protein profile. Journal of Dairy Science, 105(7), 6001–6020. https://doi.org/10.3168/jds.2021-21421
3. Bangar, Y.C., Magotra, A., Chauhan, A. & Yadav, A.S. (2021). Genetic polymorphisms of kappa casein gene and its association with milk and composition traits in cows: An updated meta-analysis. Meta Gene, 30, 100948. https://doi.org/10.1016/j.mgene.2021.100948
4. Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2013). GenBank. *Nucleic Acids Research*, *41*(Database issue), D36–D42. https://doi.org/10.1093/nar/gks1195
5. Bonfatti, V., Di Martino, G., Cecchinato, A., Vicario, D., & Carnier, P. (2010). Effects of β-κ-casein (CSN2–CSN3) haplotypes and β-lactoglobulin (BLG) genotypes on milk production traits and detailed protein composition of individual milk of Simmental cows. *Journal of Dairy Science*, *93*(8), 3797–3808. https://doi.org/10.3168/jds.2009-2921pdfs.semanticscholar
6. Carles, C. & Martin, P. (1985). Kinetic study of the action of bovine chymosin and pepsin A on bovine kappa-casein. Archives of Biochemistry and Biophysics, 242(2), 411–416. https://doi.org/10.1016/0003-9861(85)90225-5
7. Caroli, A. M., Chessa, S., & Erhardt, G. J. (2009). Invited review: Milk protein polymorphisms in cattle: Effect on animal breeding and human nutrition. *Journal of Dairy Science*, *92*(11), 5335–5352. https://doi.org/10.3168/jds.2009-2438pdfs.semanticscholar
8. Chessa, S., [et al.]. (2008). Genetic variation in the K-casein gene (CSN3) of Chinese yak (*Bos grunniens*) and phylogenetic analysis of CSN3 sequences in the genus *Bos*. *Journal of Dairy Science*, *91*(3), 1198–1203. https://doi.org/10.3168/jds.2007-0589idosi

9. Comin, A., Cassandro, M., Chessa, S., Ojala, M., Dal Zotto, R., De Marchi, M., Gallo, L., Pagnacco, G. & Bittante, G. (2008). Effects of composite β- and κ-casein genotypes on milk coagulation, quality, and yield traits in Italian Holstein cows. Journal of Dairy Science, 91, 1274–1287.
10. Eigel, W. N., Butler, J. E., Ernstrom, C. A., Farrell, H. M., Harwalkar, V. R., Jenness, R., & Whitney, R. M. (1984). Nomenclature of proteins of cow's milk: Fifth revision. *Journal of Dairy Science*, *67*(7), 1599–1631. https://doi.org/10.3168/jds.S0022-0302(84)81485-Xpubmed.ncbi.nlm.nih+2
11. Erhardt, G. (2008). Genetic variation in the K-casein gene (CSN3) of Chinese yak (*Bos grunniens*) and phylogenetic analysis of CSN3 sequences in the genus *Bos*. *Journal of Dairy Science*, *91*(3), 1198–1203. https://doi.org/10.3168/jds.2007-0589idosi
12. Fan, X., Zhang, Z., Qiu, L., Zhang, Y. & Miao, Y. (2019). Polymorphisms of the kappa casein (CSN3) gene and inference of its variants in water buffalo (Bubalus bubalis). Archives Animal Breeding, 62, 585–596. https://doi.org/10.5194/aab-62-585-2019.
13. Farrell, H. M., Jr., Jimenez-Flores, R., Bleck, G. T., Brown, E. M., Butler, J. E., Creamer, L. K., Gordon, C. L., Groves, M. L., Klaarenbeek, H., Moll, P., Nakai, Y., & Swaisgood, H. E. (2004). Nomenclature of the proteins of cows' milk—Sixth revision. *Journal of Dairy Science*, *87*(6), 1641–1703. https://doi.org/10.3168/jds.S0022-0302(04)73319-6journalofdairyscience+1
14. Grosclaude, F., Mahé, M.-F., Mercier, J.-C., & Ribadeau-Dumas, B. (1972). Localisation des substitutions d'acides aminés différenciant les variants A et B de la caséine κ bovine. *Annales de Génétique et de Sélection Animale*, *4*(4), 515–521. https://doi.org/10.1186/1297-9686-4-4-515pubmed.ncbi.nlm.nih+5
15. Mahmoudi, P., Rostamzadeh, J., Rashidi, A., Zergani, E., & Razmkabir, M. (2020). A meta-analysis on association between CSN3 gene variants and milk yield and composition in cattle. Animal Genetics, 51(3), 369–381. https://doi.org/10.1111/age.12925
16. Hassanin, A. A., Osman, A., Atallah, O. O., El-Saadony, M. T., Abdelnour, S. A., Taha, H. S., Awad, M. F., Elkashef, H., Ahmed, A. E., Abd El-Rahim, I., Mohamed, A., & Eldomiaty, A. (2022). Phylogenetic comparative analysis: Chemical and biological features of caseins (alpha-S-1, alpha-S-2, beta- and kappa-) in domestic dairy animals. Frontiers in Veterinary Science, 9, 952319. https://doi.org/10.3389/fvets.2022.952319
17. Heck, J. M. L., Schennink, A., van Valenberg, H. J. F., Bovenhuis, H., & van Arendonk, J. A. M. (2009). Effects of milk protein variants on milk protein composition and cheese-making properties in Dutch Holstein-Friesian cows. *Journal of Dairy Science*, *92*(3), 1150–1159. https://doi.org/10.3168/jds.2008-1208
18. Huppertz, T., Fox, P.F. & Kelly, A.L. (2018). The caseins: Structure, properties and functionality. In Advanced Dairy Chemistry: Volume 1A: Proteins: Basic Aspects (4th ed.). Springer.
19. Ikonen, T., Ahlfors, K., Kempe, R., Ojala, M. & Ruottinen, O. (1999). Genetic parameters for the milk coagulation properties and prevalence of noncoagulating milk in Finnish dairy cows. Journal of Dairy Science, 82, 205–214.

20. Kayihura, M., et al. (2024). Extent of κ-casein hydrolysis during renneting of bovine milk: A critical assessment of the analytical and estimation approaches. Food Science & Nutrition, 12, 1399–1412. https://doi.org/10.1002/fsn3.3868
21. Kelly, L.M., O'Mahony, J.A. & Tobin, J.T. (2025). Genetic variation in bovine milk proteins: Implications for functional and nutritional properties. International Journal of Dairy Technology, 78(1), e13152. https://doi.org/10.1111/1471-0307.13152.
22. Li, W., Wang, X., Xu, X., Wu, P., Fu, T., Zhang, L., & Gao, T. (2025). Identification of the bovine CSN3 core promoter region and the relationships between CSN3 promoter polymorphisms and the CSN3 A and B alleles. Animals, 15(2), 134. https://doi.org/10.3390/ani15020134
23. Lucey, J.A. (2002). Formation and physical properties of milk protein gels. Journal of Dairy Science, 85, 281–294.
24. Mahmoudi, P., Rostamzadeh, J., Rashidi, A., Zergani, E., & Razmkabir, M. (2020). A meta-analysis on association between CSN3 gene variants and milk yield and composition in cattle. *Animal Genetics*, *51*(3), 369–381. https://doi.org/10.1111/age.12920
25. Mercier, J.-C., Brignon, G., & Ribadeau-Dumas, B. (1973). Primary structure of a κ-casein B macropeptide. *European Journal of Biochemistry*, *27*(3), 535–547. https://doi.org/10.1111/j.1432-1033.1972.tb01870.xjournalofdairyscience+3
26. Mercier, J. C., Vilotte, J. L., & Provot, C. (1990). Structure and expression of milk protein genes. In Factors affecting the composition of milk from dairy cows (pp. 5–32).
27. NCBI. (2026). CSN3 casein kappa [Bos indicus (zebu cattle)]. National Center for Biotechnology Information, Gene ID 109560547.
28. NCBI. (2026). CSN3 casein kappa [Bos taurus (domestic cattle)]. National Center for Biotechnology Information, Gene ID 281728.
29. Pinder, A. M., Hines, M. E., & Broucek, J. (1991). κ-Casein genotypes in New Zealand dairy cattle. *Proceedings of the New Zealand Society of Animal Production*, *51*, 207–209.hal + 1
30. Prinzenberg, E. M., Gutscher, K., Chessa, S., Caroli, A., & Erhardt, G. (2005). Caprine κ-casein (CSN3) polymorphism: New developments in molecular knowledge. *Journal of Dairy Science*, *88*(4), 1559–1567. https://doi.org/10.3168/jds.S0022-0302(05)72817-4
31. Rohlf, F. J. Felsenstein, Inferring Phylogenies, Sinauer Assoc., 2004, pp. xx + 664.. Journal of Classification 22, 139–142 (2005). https://doi.org/10.1007/s00357-005-0009-4
32. Rout, P. K., Singh, L. W., & Mandal, A. B. (2006). Molecular characterization of κ-casein gene in Indian river buffalo (*Bubalus bubalis*). *Indian Journal of Biotechnology*, *5*, 509–513.
33. Schopen, G.C.B., Visker, M.H.P.W., Koks, P.D., Mullaart, E., van Arendonk, J.A.M. & Bovenhuis, H. (2011). Whole-genome association study for milk protein composition in dairy cattle. Journal of Dairy Science, 94, 3148–3158.
34. Waugh, D.F. & von Hippel, P.H. (1956). κ-Casein and the stabilization of casein micelles. Journal of the American Chemical Society, 78, 4576–4583.

Figures

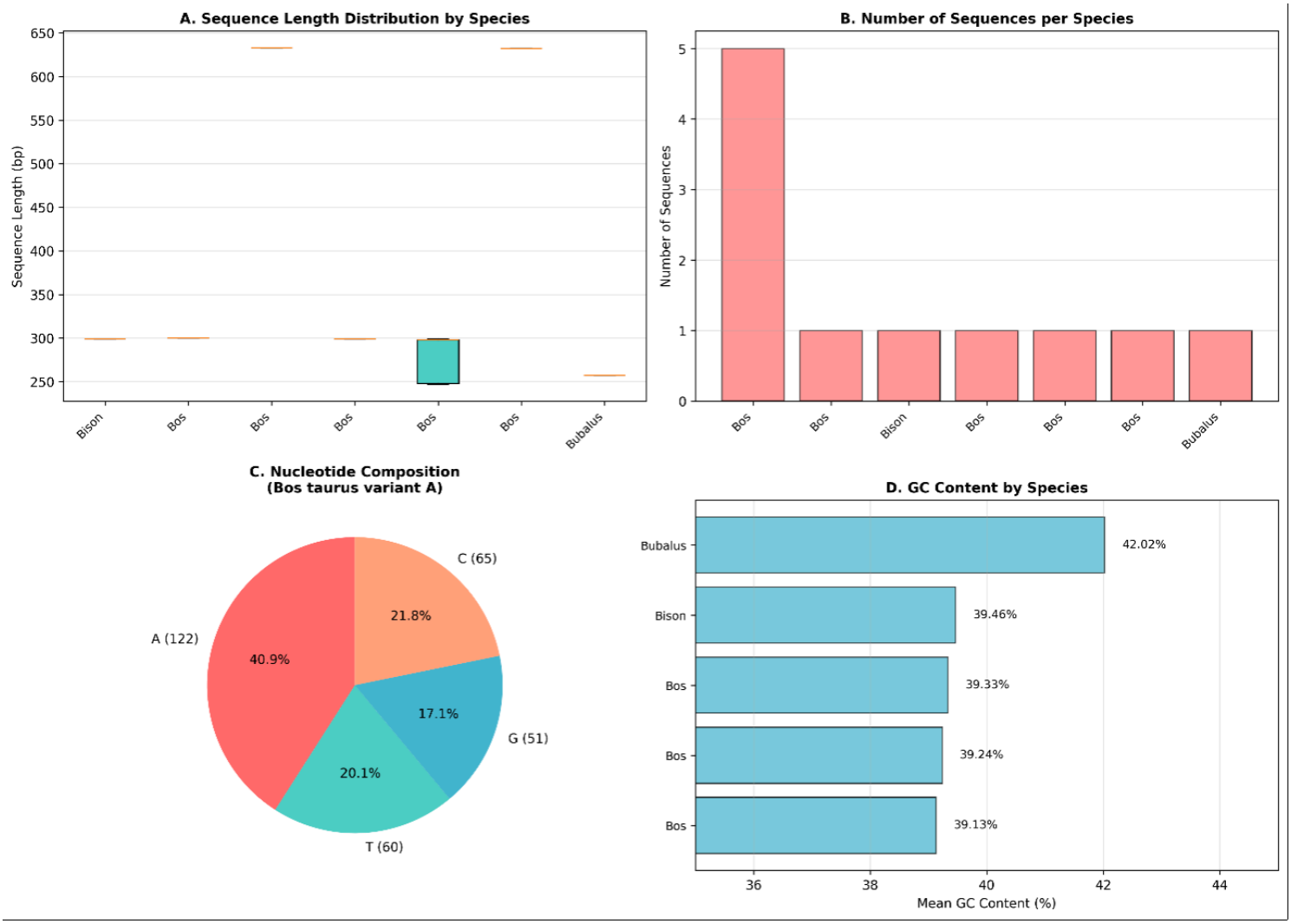


Figure 1

Summary of CSN3 sequence characteristics across the analysed taxa. (A) Sequence length, (B) number of sequences per species, (C) nucleotide composition of the Bos taurus variant A sequence, and (D) mean GC content by species.

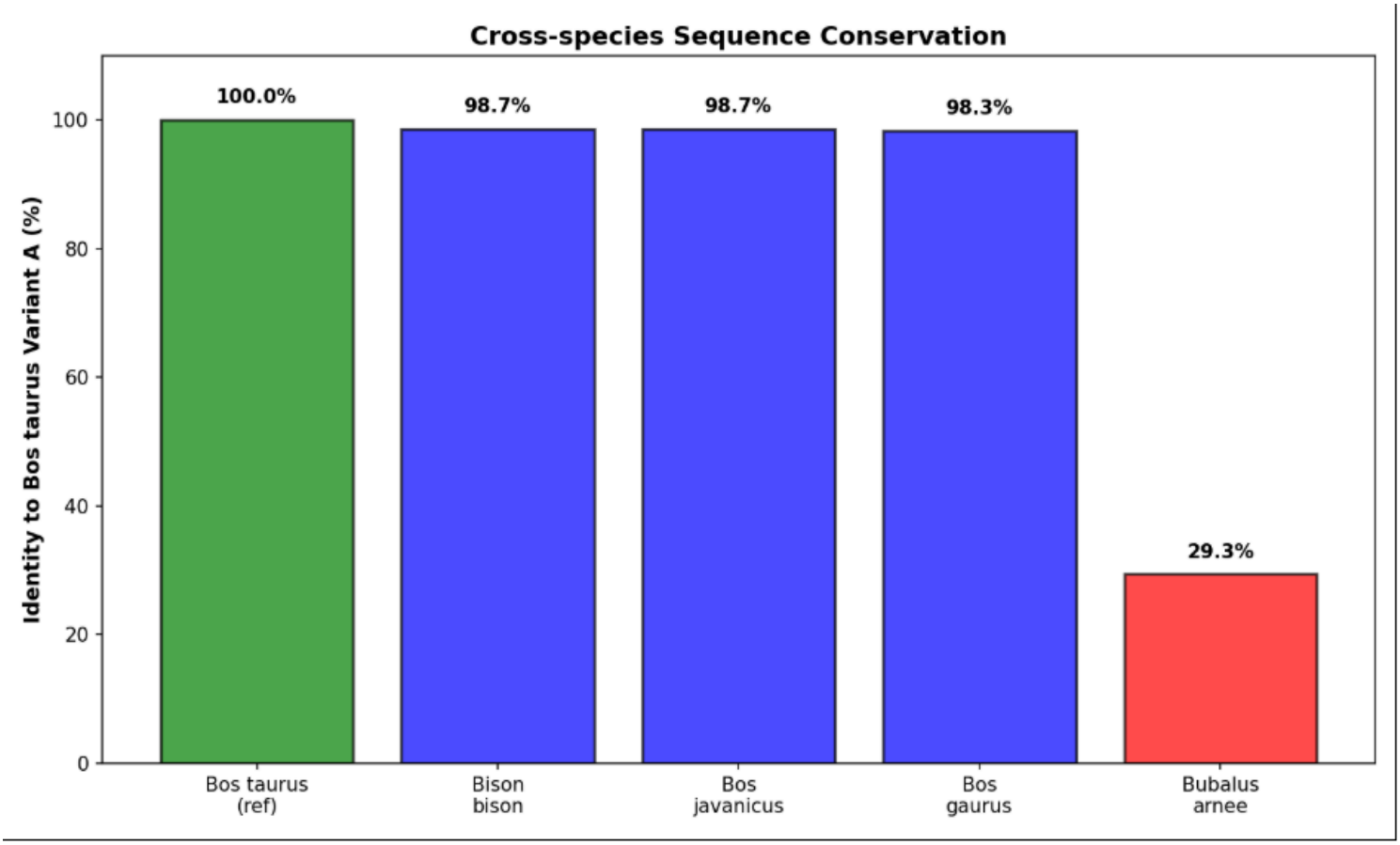


Figure 2

Cross-species sequence conservation of the Bos taurus CSN3 variant A relative to CSN3 sequences from Bison bison, Bos javanicus, Bos gaurus, and Bubalus arnee. Values represent percentage sequence identity.

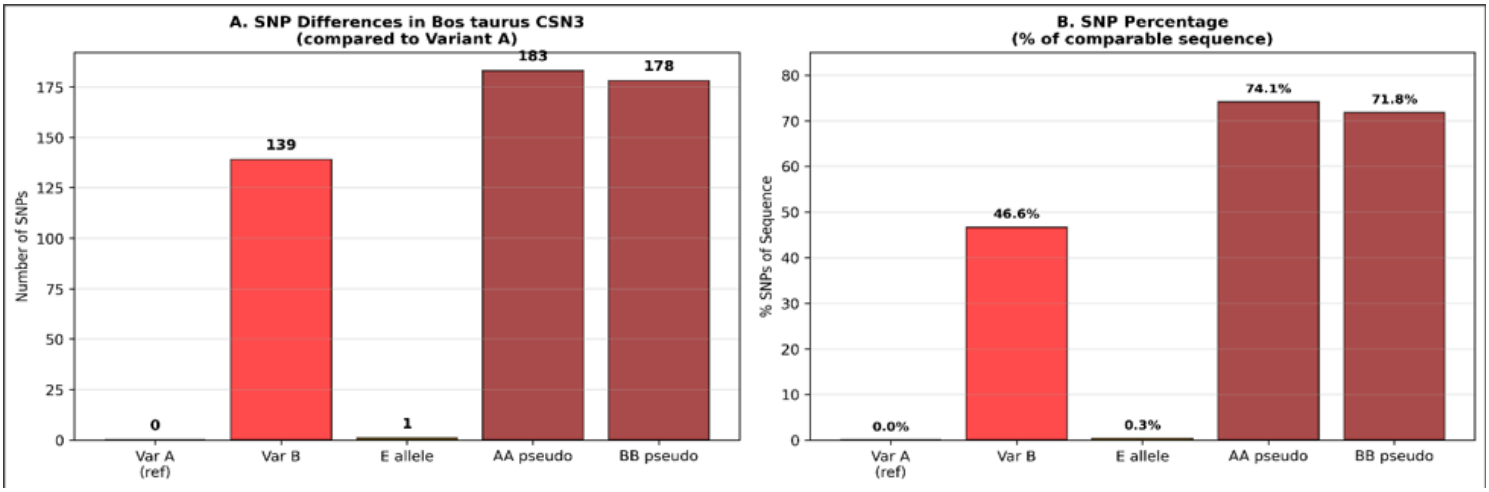


Figure 3

Comparative CSN3 sequence variation among Bos taurus variants relative to variant A. (A) Number of single-nucleotide polymorphism (SNP) differences and (B) SNP percentage relative to the comparable

sequence. The analysis includes variants A, B, E allele, AA pseudo, and BB pseudo.

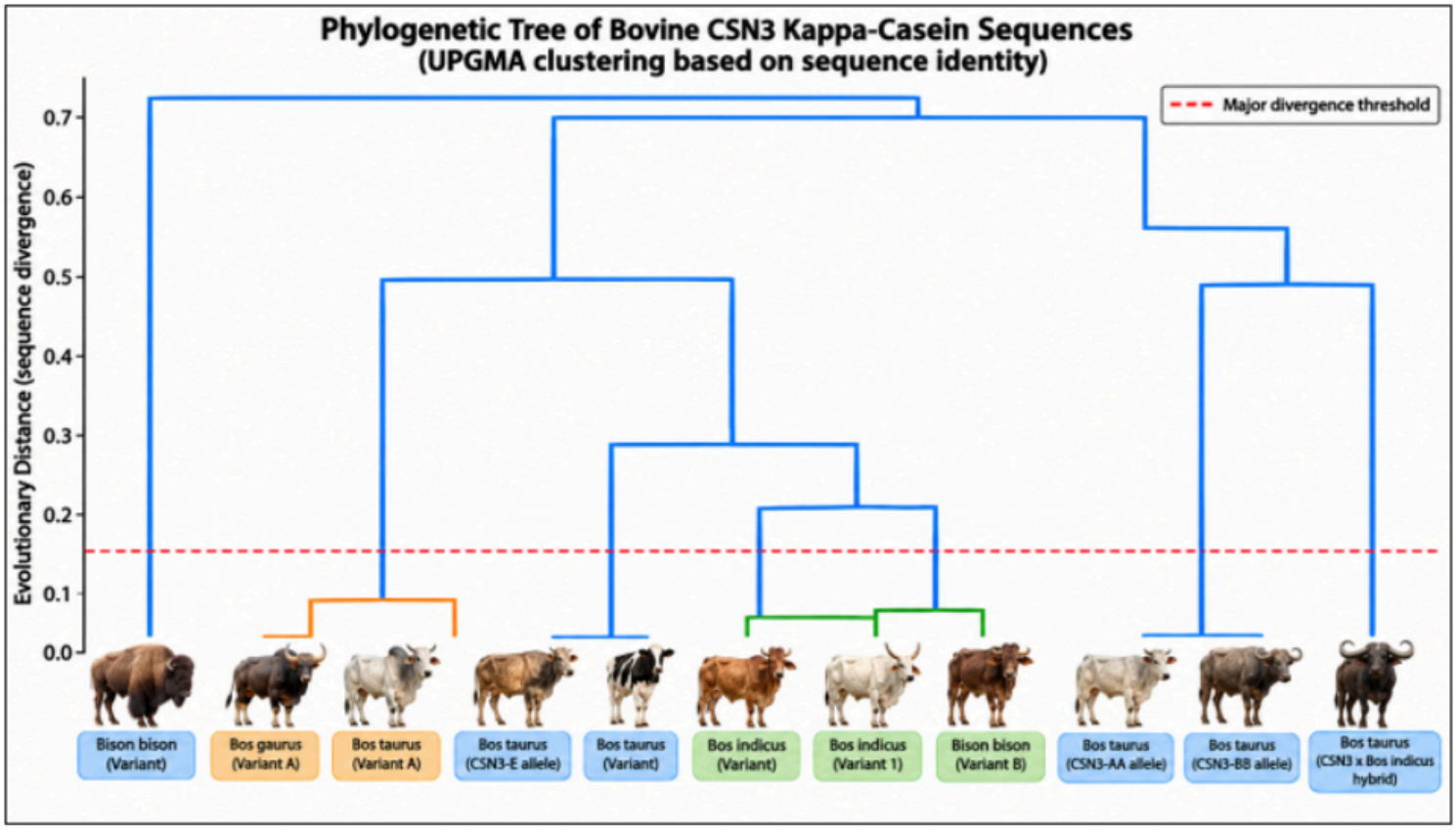


Figure 4

UPGMA dendrogram of bovine CSN3 kappa-casein sequences based on sequence identity. Branch heights represent evolutionary sequence distance, and the dashed horizontal line indicates the major divergence threshold used in the analysis.

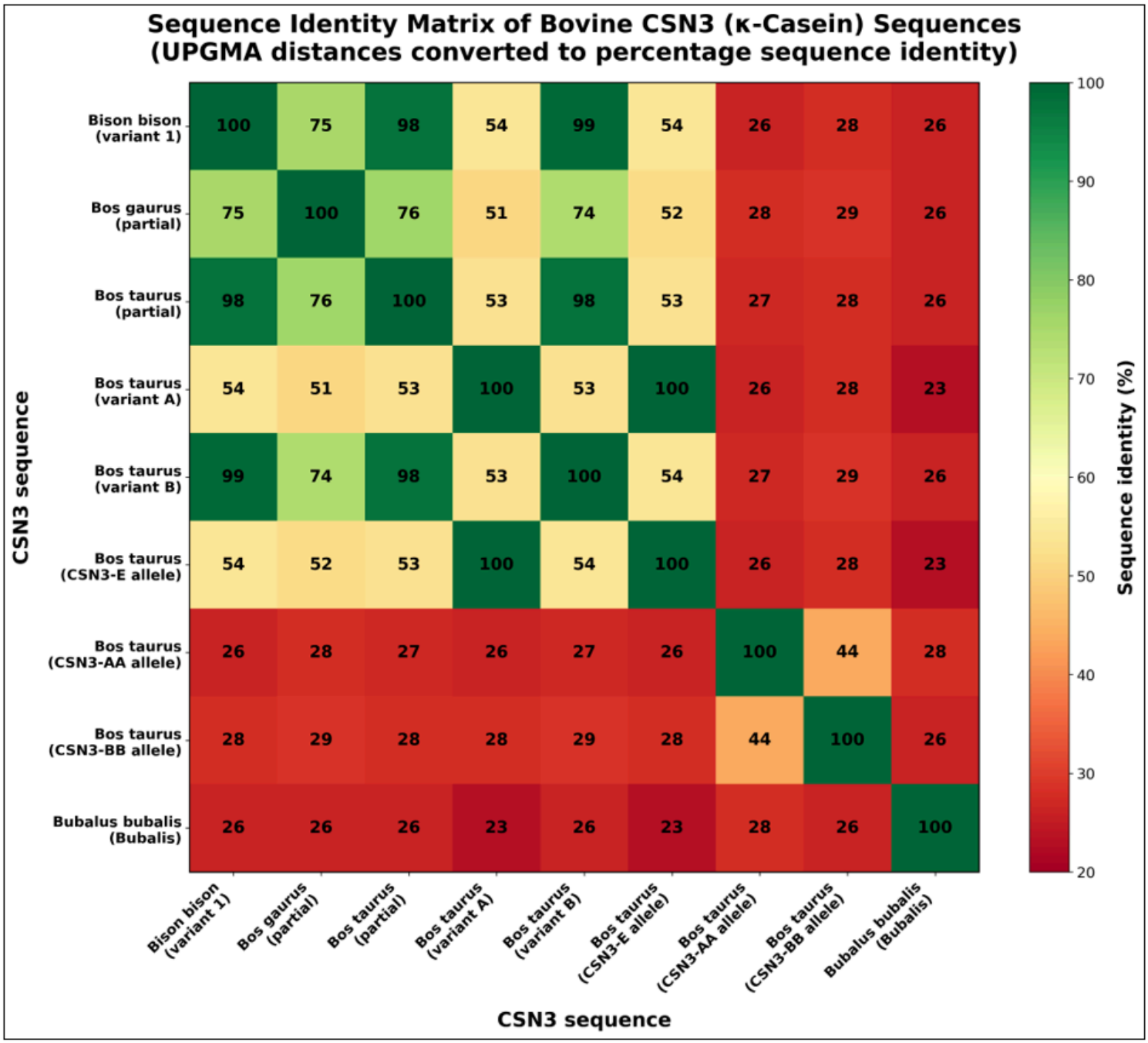


Figure 5

Pairwise sequence identity matrix for bovine CSN3 kappa-casein sequences. The matrix presents percentage sequence identity derived from UPGMA distance estimates, with values shown for each pair of sequences.

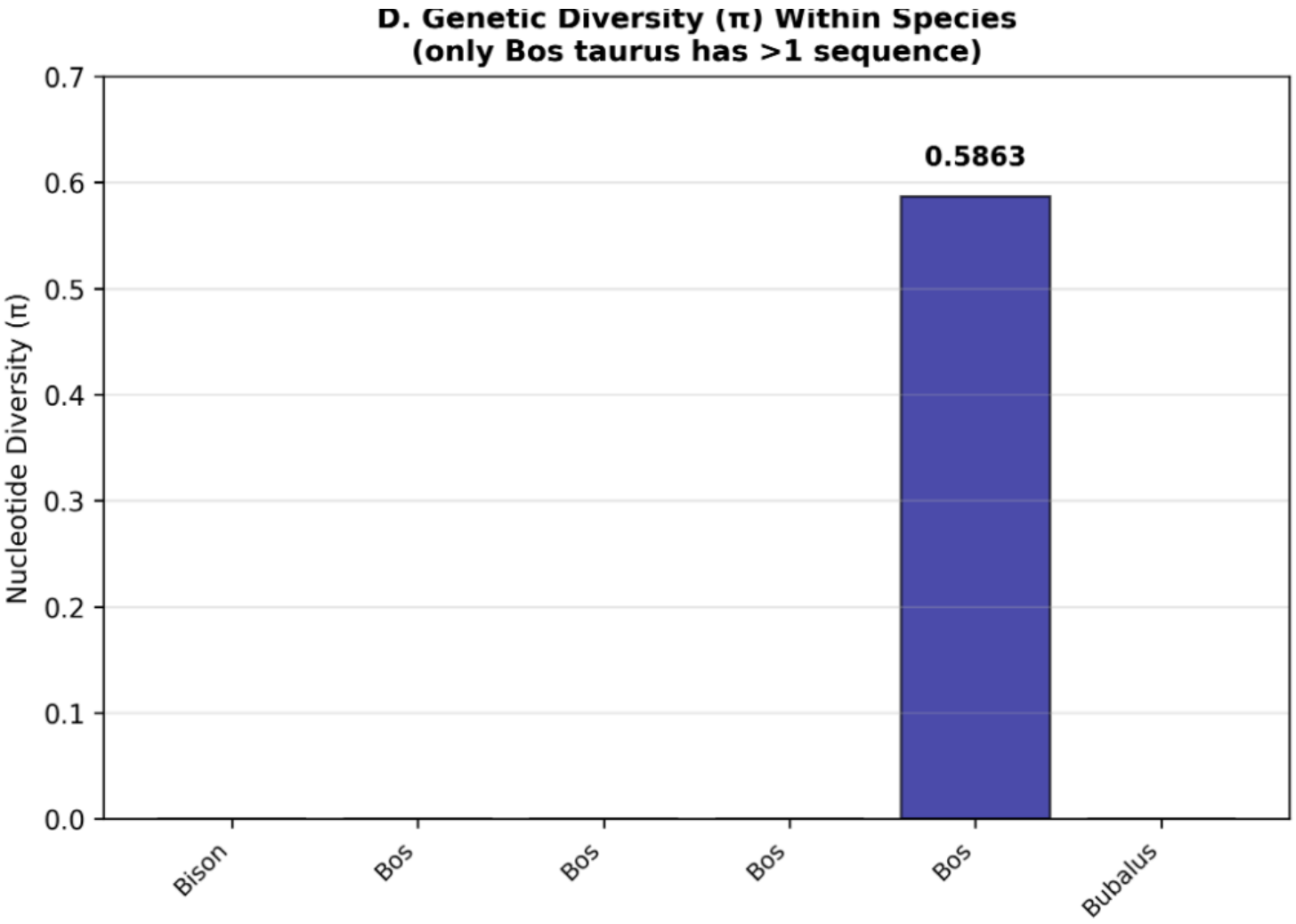


Figure 6

Nucleotide diversity (π) within the analysed CSN3 sequences. The estimate of π = 0.5863 was obtained for Bos taurus, for which more than one sequence was available; other taxa were not included in the within-species diversity estimate.