ADDIS ABABA UNIVERSITY
SCIENCE FACULTY
BIOTECHNOLOGY PROGRAM

Molecular Characterization of the Bovine *CSN*3 and *LGB* Milk Protein Genes Using Sequencing and PCR-RFLP Markers in Ethiopian Indigenous Cattle Populations

---

A thesis submitted to the School of Graduate Studies in partial fulfillment of the requirements for the degree of Master of Science in Biotechnology

By

Robel Getachew

Supervisors:

Abiy Zegeye (PhD)

Tadelle Dessie (PhD)

January 2010

TABLE OF CONTENTS

ACKNOWLEDGMENTS

First of all, I would I like to owe my deepest gratitude to my supervisors Dr Abiy Zegeye and Dr Tadelle Dessie for their excellent guidance and mentorship which were indispensible for the fruitful execution of the project. They were always available tirelessly for open discussions and pertinent scientific questions without which this thesis would not have been possible.

I convey a special thanks to Dr Jean Hanson for providing me the lab space at the molecular laboratory facility (ILRI) where I had unreserved access to consumables and Biotech instruments. I am also grateful to Ato Temeselew Mamo for his assistance and support throughout the laboratory work. I am indebted to Dr Kassahun Tesfaye, W/o Adey Feleke and Ato Dawit W/Meskel for their continuous encouragement and valuable ideas. I am also thankful to Ato Eshetu Zerihun and Ato Michael Temesgen for their unreserved technical assistance during blood sample collection.

It gives me a great honor to thank Haramaya University for sponsoring my graduate study. Great tribute goes to the International Livestock Research Institute (ILRI) for accepting me as a graduate fellow and financially covering all the research expenses associated with this work through project No. BTO2. I thank Addis Ababa University for admitting me to the Biotechnology graduate Program. I am also indebted to the assistance conveyed by many research personnel at the DebreZeit Agricultural Research Center and Ghibe cattle breeding station.

Finally, I would like to extend my heartfelt appreciation to so many friends and colleagues who in one way or the other provided their unreserved support. Most of all, gratitude is to Aba, Mam, my brother (Sami) and sisters (Romi and Samri) for their continuous support and belief in me.

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

A	Adenine
A_e	Effective number of alleles
AFLP	Amplified Fragment Length Polymorphism
Ala	Alanine
ALA	Alpha-lactalbumin
AP	Mean number of alleles per polymorphic loci
Asp	Asparagine
bp	Base pair
C	Cytosine
CAPS	Cleaved Amplified Polymorphic Sequence
cDNA	Complimentary Deoxyribo Nucleic Acid
*CSN*3	Kappa Casein gene
CMP	Caseinomacropeptide
CCP	Colloidal calcium phosphate
D	Nei's genetic Distance
DAGRIS	Domestic Animal Genetic Resources Information System
DARC	DebreZeit Agricultural Research Center
dCAPS	Derived cleaved amplified polymorphic sequence
df	degrees of freedom
DNA	Deoxyribo Nucleic Acid
dNTP	deoxy Nucleotide Tri Phosphate
EARO	Ethiopian Agricultural Research organization
EIAR	Ethiopian Institute of Agricultural Research
EBV	Estimated Breeding Values
EDTA	Ethylene Diamine Tetraacetic Acid
EST	Expressed Sequence Tags
Gly	Glycine
HC	Holstein-Barka Cross breeds
H_e	Expected Heterozygosity
H_o	Observed Heterozygosity

HWE	Hardy-Weinberg Equilibrium
HWP	Hardy-Weinberg Proportion
I	Genetic Identity
IBC	Institute of Biodiversity Conservation
Ile	Isoleucine
ILRI	International Livestock Research Institute
ISSR	Inter Simple Sequence Repeats
KDa	Kilo Dalton
LD	Linkage Disequilibrium
LE	Linkage Equilibrium
LGB	Beta-Lactoglobulin gene
MAS	Marker Assisted Selection
mRNA	messenger Ribonucleic Acid
RAPD	Random Amplified Polymorphic DNA
NCBI	National Center of Biotechnology Information
Nm	Gene flow
OD	Optical density
PCR-RFLP	Polymerase Chain Reaction-Restriction Fragment Length Polymorphism
QTL	Quantitative Trait Loci
SDS	Sodium Dodecyl Sulfate
SNPs	Single Nucleotide Polymorphisms
SSCP	Single strand conformational polymorphism
SSR	Simple Sequence Repeats
T	Thiamine
Thr	Threonine
UTR	Untranslated Region
UV	Ultraviolet
Val	Valine

ABSTRACT

In the last few decades, genetic polymorphism of bovine milk protein genes such as caseins and whey proteins has achieved considerable research interest mainly due to their significant associations with fat and protein contents as well as the manufacturing properties of milk. In this study, genetic polymorphisms of two bovine milk protein genes namely CSN3 and LGB were analyzed in five indigenous and one crossbred cattle breeds using PCR-RFLP markers and sequencing. Moreover, the magnitude of genetic variability within and among these cattle populations at these two loci was estimated hence their population structure and phylogenetic relationships characterized. For this purpose, genomic DNA was extracted from 83 animals belonging to Abigar, Boran, Guraghe, Horro, Sheko and Holestin-Barka crossbred cattle populations. Both of the markers used to analyze polymorphisms in the CSN3 and LGB genes were informative and enabled us to detect the widely reported CSN3-A and B and LGB-A and B variants. Furthermore, sequencing revealed additional genotypes among which were the zebu-specific CSN3-A^1 *and H haplotypes observed in 50% of the sequences from the indigenous cattle and CSN3-E in one sequence derived from a Holstein-Barka crossbred cow. These CSN3 haplotypes were deduced analyzing a total of 7 point mutations found in the 633bp region sequenced of which 5 were base transitions and 2 transversions. Two of these sites were silent substitutions while the rest were non-synonymous. At LGB locus, 12 SNPs were detected in the amplified 529 bp region where only 3 resided in the coding region corresponding to 2 silent and one non-synonymous substitution. Analysis of these mutations yielded two new haplotypes (LGB-*B^1 *and B*) that where results of silent mutations. The PCR-RFLP assay revealed that CSN3-A and LGB-B were more prevalent than CSN3-B and LGB-A variants in the indigenous cattle population investigated. The frequency of allele CSN3-A (0.687) was more than twice that of B (0.313) where as LGB-B had a very high frequency of 0.861 compared to LGB-A (0.139). All of the breeds met HWE at both loci with the exception of Sheko that deviated significantly (*$P<0.05$*) at the LGB locus. The total population revealed a rather moderate level of genetic variability (0.34). The highest variability was observed in Sheko (0.414) while the lowest was in Guraghe (0.273). When the two loci were considered separately, CSN3 locus presented higher (43.04%) level of gene diversity compared with LGB locus (24.42%) which was rather more homogenous. A higher level of inbreeding was detected at LGB locus (*$F_{IS}=0.173$*) compared with CSN3 which instead showed excess heterozygosity with *F_{IS} *of -0.0282. When considering both loci, the mean *F_{IS} *and *F_{IT} *estimates were 0.044 and 0.07 respectively. The overall *F_{ST} *value (0.028) revealed little and insignificant (*$P>0.05$*) genetic differentiation among the populations studied. Most of the total genetic variation was attributed to the within population (97.23%) whereas little and insignificant level of variation was witnessed among the different cattle breeds (2.77%). Pair-wise genetic distances revealed very small distances among the breeds studied. However, Sheko had larger distances when paired with each breed where the maximum of which was with Abigar and the minimum that involving Guraghe indicating its relative departure from the rest. Similarly, the dendrogram generated clearly depicted Sheko forming a separate branch where the rest shared a second main branch.*

Key words: PCR-RFLP, Sequencing, diversity, Ethiopian cattle, SNPs, CSN3, LGB, HWE

I. INTRODUCTION

Ethiopia's livestock population is believed to be the largest in Africa with 41 million cattle and 50 million sheep and goats (IBC, 2004). Livestock is mostly kept by smallholder farmers and serves as an important source of meat, milk, cash, and hair/wool. This sector contributes 20 and 40% of the total and agricultural Gross Domestic Product (GDP), respectively, and provides livelihood for 65% of the population (IBC, 2004). Moreover, the genetic diversity among indigenous cattle breeds of Ethiopia is also considered to be very high based on the number of recognized breeds (IBC, 2004, DAGRIS). Despite this significant contribution of livestock to Ethiopia's economy, little attention has been given to identify, characterize, utilize and conserve the diversity of indigenous livestock types. Currently, genetic diversity of Ethiopian cattle populations is under constant threat mainly due to extensive unplanned and non-discriminate crossbreeding between adjacent indigenous breeds. Loss of genetic diversity in turn increases the risk of losing many unique attributes that are crucial for survival and productivity in the midst of continuing environmental and climatic challenges. Analysis of genetic variation among and within populations, therefore, allows the optimum utilization of cattle genetic resources and permits efficient genetic improvement for both production and conservation needs. The dairy production sector, among others, contributes immensely to income generation, poverty alleviation and food security. However, the national milk production still remains among the lowest in the world, even by African standards (Zegeye, 2003). Per capita consumption of milk is about 16 kg per person per year, which is much lower than the African and World averages of 27 kg/year and 100 kg/year, respectively (Saxena *et al*., 1997). Several factors are known to affect milk composition and yield in cattle among which are environmental and management practices like disease control, feeding and housing and equally importantly is the variability in the genetic make-up of cattle. This genetic variability offers an opportunity for enhancement of cattle milk production through implementation of genetic improvement programmes.

Recently, selection of superior animals has been made more effective through studies of major milk protein genes that are known to affect both milk yield and composition. Particularly, genotypes of the *CSN*3 and *LGB* genes have been shown to have significant effects on fat and protein contents as well as on the manufacturing properties of milk. Such genes that are correlated with performance parameters explain part of the genetic variance and can improve the estimation of breeding values during selection thereby supplementing conventional breeding programs. Moreover, milk protein genes have also proved to be useful tools for breed characterization, diversity and phylogenetic studies (Caroli *et al*., 2009). Unfortunately, there has been no effort to characterize these economically important milk protein genes and analyze their effects on milk production traits in our native Ethiopian cattle breeds. A lot of literature, however, has been accumulated so far dealing with these two milk protein genes mostly in cattle breeds of European origin hence cannot be directly translated to our native cattle breeds. This study, therefore, will aid in laying the foundation for designing association studies that will intern provide information for future cattle breeding programs aimed at enhancing various milk production traits.

With this background and understanding, this study was conducted to meet the following objectives:

GENERAL OBJECTIVES OF THE STUDY

- Molecular Characterization of the bovine *CSN*3 and *LGB* genes in Ethiopian indigenous cattle populations

SPECIFIC OBJECTIVES

- To analyze genetic polymorphisms of the two important milk protein genes: *CSN3* and *LGB* in five Ethiopian indigenous cattle breeds using sequencing and PCR-RFLP markers
- To determine the frequency distribution of *CSN*3 and *LGB* genetic variants detected across the cattle breeds examined
- To estimate the level of genetic diversity and differentiation within and between these breeds/subpopulations at these two loci
- To evaluate the existence of novel/new *CSN*3 and *LGB* genetic variants in our native cattle populations.

II. LITERATURE REVIEW

1. Bovine milk proteins

Milk is a complex biological fluid and serves as an essential source of energy, proteins, minerals, and vitamins for young mammals during the first period of their life. All species of mammals from man to whales produce milk for this purpose. Milk is a highly diverse fluid composed of mainly water, lactose, fat, proteins, organic acids and minerals. Throughout the following sections focus will be on the protein fraction of milk, a heterogeneous group of molecules where over 200 different types have been characterized.

On average bovine milk contains approximately 3.0-3.5% protein; the concentration of which usually changes during lactation. The natural function of milk proteins is to provide newborns with the essential amino acids needed for the development of muscular and other protein containing tissues (Fox and McSweeney, 1998). To best serve this purpose, milk proteins form complexes with a considerably large amount of calcium phosphate which subsequently coagulates in the stomach of neonates. The properties of milk and milk products primarily depend on milk proteins.

Both qualitative and quantitative differences are known to occur in milk of different species. However, milk proteins of all mammals can be generally divided into two major components: the caseins and the whey proteins.

Bovine milk contains four major caseins namely; αs_1 and αs_2 caseins, β-casein and kappa-casein in approximate proportions of 4:1:4:1 respectively (Hallen, 2008). These caseins (αs_1, αs_2, β and k) comprise the major protein component (80%) of ruminant milk and are secreted in the form of stable calcium phosphate micelles (Fig. 1). A heterogeneous group of proteins termed γ-caseins, degradation products of β-casein, also exist as minor components (Miller, 1990). The action of post-translational modifications such as phosphorylation, glycosylation, disulfide bonding, proteolysis along with genetic variations make caseins as one of the most diverse protein groups (Ng-Kwai-Hang, 2002). Because of high content of proline, caseins exhibit little rigid tertiary conformation or organized secondary structure (Fox and McSweeney, 1998). As there is very little structure to unfold, caseins are known for their high stability against heat denaturation whereas it is sensitive to pH and precipitates at its isoelectric pH of 4.6 (Walstra, 1999). Caseins undergo phosphorylation during which phosphate groups are esterified interacting with the hydroxyl group of serine resulting in phosphoserine residues which subsequently bind calcium and colloidal calcium phosphate (CCP). Such bonds are responsible for linking the caseins together to finally form micelles. From the overall calcium content in milk (≈1200 mg/l), around half of it is found bound to the casein fraction via CCP (Hallen, 2008).

Whey protein is the protein fraction remaining in solution after precipitating the casein micelle and fat globules at a pH of 4.6 and it makes up around 20% of the protein in bovine milk. Whey proteins are also termed as serum protein or non-casein nitrogen. The principal fraction of whey proteins are beta-lactoglobulin (*LGB*) and alpha-lactalbumin (ALA). However, bovine serum albumin (BSA) and immunoglobulins (IgGs) are also minor constituents of the total whey fraction. Beta lactoglobulin is the major whey protein in ruminant milk while alpha lactalbumin is part of the enzyme system involved in lactose synthesis. Both Beta lactoglobulin and alpha-lactalbumin are synthesized in the mammary gland whereas serum albumin is transported to the mammary gland via the blood stream. Many other proteins such as glycomacropeptides (GMP), lactoperoxidase, lysozymes, lactoferrin and plasmin are also present in trace amounts (Farrell *et al.*, 2004).

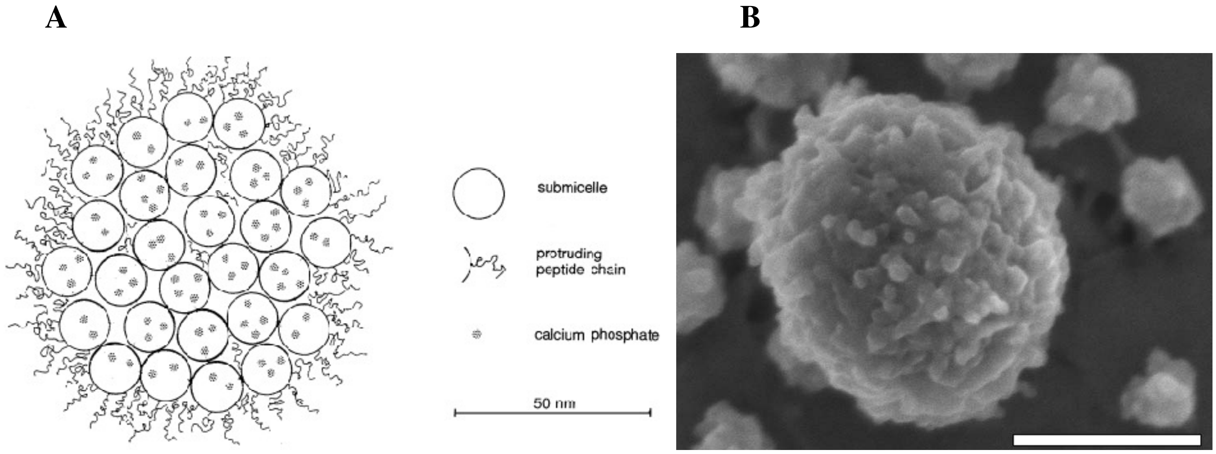


Figure 1. A) The casein submicelle model (Walstra, 1999), B) Electron micrograph of an individual casein micelle based on high-resolution field-emission scanning microscopy, Scale bar = 200 nm. (Courtesy of Dalgleish *et al.*, 2004)

1.1.Kappa casein (*CSN*3)

1.1.1. Protein structure and function

Each of the four major caseins (αs_1, αs_2, β and κ) manifests variability at the level of phosphorylation and glycosylation. Most αs and β-caseins are relatively more phosphorylated (5-13mol PO_4/mol) than κ-casein protein which is slightly phosphorylated (1-3mol PO_4/mol)

(Swaisgood, 1992). This is partly because of the fact that phosphorylation sites of κ-casein (at serine and threonine residues) are limited to the C-terminal region and exist as single sites unlike other caseins with clusters of such sites. Moreover, further phosphorylation of κ-casein may be halted because of early glycosylation at susceptible sites (Swaisgood, 1992). Casein kinase is responsible for the post-translational phosporylation of caseins in the Golgi apparatus (Bingham, 1979). κ-casein is the only casein which is glycosylated and contains galactose, galactosamine and N-aceytlneuraminic acid which are attached to theronine residues in the C-terminal region. As a result, κ-casein binds to calcium to a very limited extent compared to other caseins making it soluble over a broad range of calcium ion concentrations.

Glycosylation of κ-casein takes place within the Golgi apparatus of mammary epithelial cells during post-translational modification and is catalyzed by O-glycosyl transferases (Mepham *et al*., 1992). The carbohydrate moieties are attached to κ-casein via O-glycosidic linkages. Various genetic variants of bovine k-casein have been shown to induce significantly different degree of glycosylation. The level of glycosylation has been reported to affect the susceptibility of κ-casein to hydrolysis by chymosin, susceptibility decreasing with higher carbohydrate content (Fox, 1989). Reports indicated increased glycosylation during the colostral period, and in response to infection such as mastitis but decreased levels with successive periods of lactation (Dziuba and Minkiewicz, 1996).

The αs_1, αs_2 and β-caseins are termed as calcium-sensitive caseins since they are precipitated by calcium which binds to their phosphoserine residues (Phoebe, 2007). They aggregate in solution as micelles, sequestering up to 5% of their dry weight as Ca^{2+} (Koczan *et al*., 1991). The calcium-sensitive caseins are thought to be mainly situated within the micelle whereas k-casein is located on the surface, stabilizing the structure of the micelle (Horisberger and Vonlanthen, 1980). κ-casein interacts with the rest of the calcium sensitive caseins through its N-terminus while the C-terminus, with a net negative charge due to acidic amino acids, constitutes the integral part of the external layer of the casein micelle (Walstra, 1990). According to Horne (1998, 2002), steric and electrostatic repulsion of the polar C-terminal domain of the κ-casein protein are responsible for stabilizing the casein micelle.

κ-casein plays an important role during the chymosin-induced coagulation of milk occurring in the gut and during cheese making. Chymosin, the most active component of rennet, specifically cleaves the peptide bond of k-casein at Phe_{105}-Met_{106} (Jolles *et al*., 1968). This enzymatic hydrolysis of κ-casein produces the hydrophobic (insoluble) N-terminal portion or para k-casein and caseinomacropeptide (CMP), with the hydrophilic CMP part being released into the whey. Although the function of para k-casein is not known yet, CMP is responsible for clotting milk in gut which in turn allows more retention time and efficient digestion (Mercier *et al*., 1976). Following the release of the polar CMP portion at the surface of the micelles, electrostatic and steric stabilization of the micelle is disrupted leading to micelle association and subsequent coagulation (Swaisgood, 1992).

Ruminant κ-casein normally occurs as polymers of sizes ranging from 60 to 600 kDa (Wong *et al*., 1996). Two cysteine residues at positions 11 and 88 present in k-casein through disulphide bonding enable the formation of such a polymeric form which facilitates micellar surface coverage, thus stabilizing the micelle structure (Rasmussen *et al*., 1994). When milk is heat treated, κ-casein is linked with both αs_2-casein and whey proteins through disulphide bond, thus increasing the micelle surface and consequently affecting the manufacturing properties of milk (Fox and Brodkorb, 2008).

Aside from being the main source of amino acids, phosphate, and calcium; κ-casein and other caseins act as precursors for the biologically active opioid-like peptides (Miesel, 1997). These peptides are known to act as physiological modulators of different digestive and metabolic processes, the immune defense, nutrient uptake and the neuroendocrine information center (Miesel, 1997).

The bovine κ-casein protein was first isolated and its amino acid composition determined by Eigel *et al*. (1984) and Thompson *et al*. (1985). κ-casein is a single-chain polypeptide of 169 amino acids with a molecular weight of 19.2 KDa (Alexander *et al*., 1988).

1.1.2. The gene

The four major bovine caseins (αs_1, αs_2, β and k) are encoded by four tightly linked genes which are clustered in a 250 Kb of genomic DNA fragment (Fig. 2). The structure and genomic organization of the casein locus has been mapped on bovine chromosome 6 (BTA6) at q31-33 in the order of αs_1-β-αs_2-k (Threadgill and Womack, 1990). Except few variations in the number and distance separating casein genes, overall organization of the locus is fairly conserved between mammalian species. In the bovine genome, the κ-casein gene is located in a region 95-120 kb downstream of the αs_2-gene and about 200 kb from the αs_1/ β-casein region. The $\alpha s_1$and β-casein genes are found to be only 20 kb apart (Rijnkels *et al*., 1997). Casein genes exist as tightly linked clusters and the four casein loci are considered as one "genetic unit" where alleles are closely linked together and transmitted as a haplotype instead of individual alleles (Ginger and Grigor, 1999).

cDNA of *CSN*3 has been characterized in different species, including sheep (Furet *et al*., 1990), goat (Coll *et al*., 1993) and cattle (Gorodetskii and Kaledin, 1987). The bovine *CSN*3 mRNA encompasses an open reading frame of 573 bp coding for 21 amino acids of signal/leader peptide and 169 amino acids of mature protein (Alexander *et al*., 1988). When compared with the calcium sensitive genes, the signal peptide of *CSN*3 is different with respect to both amino acid sequence and length. The bovine *CSN*3 gene has been assigned to chromosome 6 (6q31) and spans 13.1 Kb that is arranged in five exons and four introns (Threadgill and Womack, 1990). The *CSN*3 gene is the only milk protein gene whose signal peptide is encoded by two exons (2 and 3). The 5' untranslated region (UTR) extends from exon 1(64 bp) to part of exon 2 (5 bp). The 3' UTR includes part of exon 4(34 bp), and the whole of exon 5 (173 bp). Introns 1-4 are 2.5 kb, 5.8 kb, 2.0 kb and 1.8 kb in length respectively (Fig. 2 and Table 1) (Alexander *et al*., 1988).

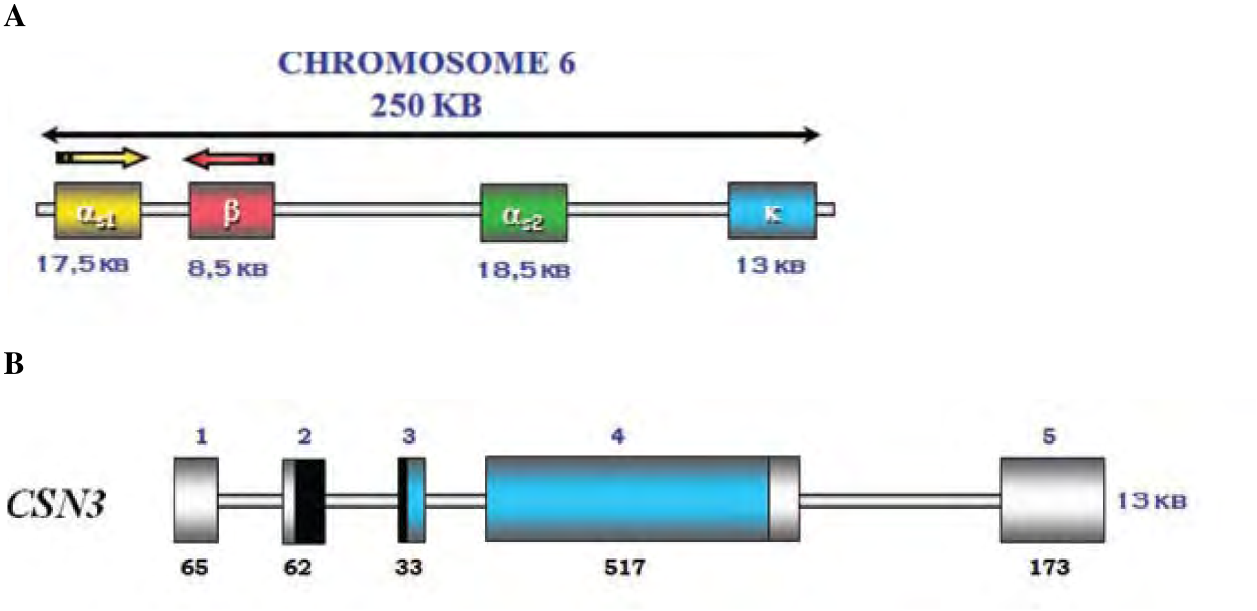


Figure 2. A) Overall genomic organization of the bovine casein locus B) Structure of the bovine kappa casein gene. Open bars represent introns; exons are depicted by large, gray (5′ and 3′ untranslated regions), black (part of exon encoding the signal peptide), and colored (exons and part of exons encoding matured proteins) boxes. Size of exons is given, in base pairs, under each exon with its number indicated on the top. (Courtesy of Martin *et al*., 2002)

Table 1. Exon and intron coordinates of the bovine *CSN3* gene. Courtesy of www.ncbi.nlm.nih.gov

Exon			Coding sequence		Intron	
No.	Coordinates	Length (bp)	Coordinates	Length (bp)	Coordinates	Length (bp)
1	1-64	64	-	-	65-2489	2425
2	2490-2551	62	2495-2551	57	2552-8465	5914
3	8466-8498	33	8466-8498	33	8499-10511	2013
4	10512-11028	517	10512-10994	483	11029-12892	1864
5	12893-13065	173	-	-	-	-

Exon information*:*

mRNA (NM 174294.1) Length: 849 bp, Number of Exon: 5

Protein (NP 776719.1) Length: 190 aa,

In the 5' flanking region of the *CSN*3 gene, several repetitive elements have been identified from different species. The bovine *CSN*3 gene contains a microsatellite repeat in intron 3 with six alleles (Lien and Rogne, 1993).

Several studies have indicated that the three genes encoding the calcium-sensitive caseins have originated from a common ancestral gene through events such as intra- and inter-genic duplication and exon shuffling (Groenen *et al*., 1993; Jones *et al*., 1985) and have also been known to share common regulatory motifs in the proximal 5' flanking region (Groenen *et al*., 1993). The *CSN*3 gene, however, does not appear to be evolutionarily related to these genes although its expression pattern is similar and its protein product is crucial for micelle formation and stability (Alexander *et al*., 1988). The *CSN*3 gene rather appears to be evolutionarily related to the fibrinogen (γ-chain) gene family which upon cleavage produces thrombin known to cause blood clotting (Jolles *et al*., 1978).

Despite its evolutionary divergence from the rest of the calcium sensitive caseins, *CSN*3 is still physically and functionally linked to them. However, all four casein genes are coordinately expressed at high levels in a tissue- and stage-specific manner (Rijnkels *et al*., 1997). Moreover, the proximal 5' flanking region of the *CSN*3 gene is organized quite differently from that of the other casein genes although its expression pattern seems to be similar with the rest (Alexander *et al*., 1988). When linked to the β-casein regulatory region, *CSN*3 gene has been shown to be expressed in the mammary gland of both transgenic mice and cattle (Persuy *et al*, 1999; Brophy *et al*., 2003). It has been suggested that regulatory elements probably located in the 5' proximal region of the casein cluster might be involved in the expression of the entire casein gene locus (Yahyaoui, 2003).

1.1.3. Polymorphism and association with production parameters

Owing to its importance in the chemistry and technological properties of milk, *CSN*3 gene polymorphism has been rigorously studied in most ruminants. The *CSN*3 gene has proved to be highly polymorphic in bovine and according to recent review of milk protein nomenclature (Caroli *et al*., 2009), eleven genetic variants have been characterized (Table. 2). These different variants where observed at varying frequencies in populations including *Bos taurus* (taurine ovine), *Bos indicus* (zebu), *Bos grunniens* (yak) and *Bos javanicus* (banteng of bali) (Ng-Kwai-Hang and Grosclaude, 1992). The most common and widely reported *CSN*3 variants are A and B, initially detected by paper electrophoresis (Neelin, 1964). The difference between variants A and B arises from two single base mutations in *CSN*3 gene which leads to two amino acid substitutions at positions 136 Thr (A**C**C)→Ile (A**T**C) and 148 (G**A**T) Asp→Ala (G**C**T), which both occur in the CMP region (Grosclaude *et al*., 1972). These two alleles can be distinguished by the presence or absence of a *Hind*III restriction endonuclease recognition site (Eggen and Fries, 1989). In addition, a change in amino acid position 148 destroys a *Hinf*I restriction site in the k-casein B allele (Medrano and Aquilar-Cordova, 1990b).

Characterization of the variants C and E was reported (Miranda *et al*., 1993) while variant H was described in zebu (Grosclaude *et al*., 1972). PCR-RFLP (CAPS) assays have already been developed to genotype k-casein variants (A, B, C, E, F, and G) using the endonucleases: *Hind*III (or *Hinf*I), *Hae*III, *Hha*I, and *Mae*II (Schielben *et al*., 1991; Prinzinberger *et al*., 1996).

Table 2. Bovine *CSN3* genetic variants (Yahyaoui, 2003)

Variants	**Amino acid positions**							**References**
	10	97	104	135	136	148	155	
A	Arg	Arg	Ser	Thr	Thr	Asp	Ser	Neelin (1964), Schmidt (1964), Woychik (1964)
B					Ile	Ala		
C		His			Ile	Ala		Di Stasio and Merlin(1979)
E							Gly	Erhardt (1989)
F	His							Prinzenberg *et al*. (1996)
G		Cys						Prinzenberg *et al*. (1996)
H				Ile				Prinzenberg *et al*. (1999)
I			Ala					Prinzenberg *et al*. (1999)
J					Ile	Ala	Arg	Mahe *et al*., 1999

Among the extensively studied and well characterized *CSN*3 genetic variants, most of the studies in the association of *CSN*3 with milk production traits limit their scope to two protein variants: A and B (Ng-Kwai-Hang *et al*., 1998). The majority of the researchers agree that milk from cows genotyped *CSN*3 BB is associated with a higher concentration of fat, protein and casein content (Ng-Kwai-Hang *et al*., 1998; Di Stasio and Mariani, 2000). Regarding association with desirable manufacturing properties of milk such as shorter rennet coagulation time, formation of a firmer curd and cheese yield, the *CSN*3-BB variant has been reported to be superior over the AA variant (Ng-Kwai Hang, 1986; Kubarsepp *et al*., 2005; Matejicek *et al*., 2008). The *CSN*3-BB has been shown to possess a more homogenous and small micellar pattern resulting in a larger micelle surface that allows formation of a firmer and consistent curd. Generally, the different genotypes of *CSN*3 were found to affect both total protein content and casein content in the order BB□AB□AA (Johann and Peter, 2000).

A few studies, however, have shown no significant effect of *CSN*3 variants on either protein or casein content (Bobe *et al*., 1999; Lunden *et al*., 1997). A tendency for increased fat yield of animals having *CSN*3-AB genotype has been detected in Holstein Friesian cows. However, fat content and lactose yield were not found to be associated with the *CSN*3 genetic variants (Tsiaras *et al*., 2005). Such inconsistencies gave rise to a speculation that the *CSN*3 alleles might possibly be connected to different promoters or cis-acting regulatory sequences (Boettcher *et al*., 2004).

On the other hand, owing to high throughput SNP detection technologies through sequencing, numerous patterns of sequential SNPs termed as haploid genotypes (haplotypes) are continuously being deciphered among which are casein haplotypes. Instead of analyzing effects of single alleles, current studies are focusing on haplotypes that consider the whole *CSN*3 gene cluster (*CSN*1-*CSN*2-*CSN*3) simultaneously (Caroli *et al*., 2009). This approach is continuously being recommended for both research and breeding purposes since it is capable of giving a fuller picture of the casein gene complex. For instance, haplotype B-A^1-B (in the order: *CSN*S1-*CSN*2-*CSN*3) was found to be associated with increased percentage of both fat and protein in Finnish Ayrshire, Italian Holstein and Brown cattle (Ikonen *et al*., 2001) but had negative effects on milk yield (Boettcher *et al*., 2004). In general, haplotypes carrying the *CSN*3-B allele had positive effects on protein percentage relative to the corresponding haplotypes carrying *CSN*3-A (Boettcher *et al*., 2004). Moreover, recent investigation of the Dutch Holstein-Friesian (Heck *et al*., 2009) highlighted the superior performance of milk derived from *CSN*2-*CSN*3 haplotype A^2-B, together with *LGB*-B in cheese making.

1.2.Beta Lactoglobulin (*LGB*)

1.2.1. Protein structure and function

Among specific candidate genes which influence most of the economically important traits in cattle, the beta-lactoglobulin (*LGB*) locus has been extensively studied (Tsiaras *et al*., 2005). Beta-lactoglobulin (*LGB*) is the major whey protein in the milk of ruminant species. It has been found in most mammalian species with the exception of humans, rodents, and lagomorphs (Kontopidis *et al*., 2004). *LGB* is an amphiphatic and extremely acid stable protein which exist at the normal physiological conditions (pH 6.5) as a dimer presenting a globular and compact structure of size 36kDa comprising of 162 amino acid residues. However, it tends to dissociate into monomers (18kDa) at low and high pH (below 3.5 and above 7.5) (Kontopidis *et al*., 2004). The secondary and tertiary structure of *LGB* has been well characterized (Fig. 3) (Papiz *et al*., 1986; Monaco *et al*., 1987).

The amino acid sequences of bovine, ovine and caprine *LGB*s have revealed a higher degree of homology with inter-sequence identity greater than 95%. The ovine and caprine *LGB* differ from bovine *LGB* at only six positions (Yahyaoui, 2003).

LGB belongs to the lipocalin protein family which presents a diverse series of functions mainly through the process of ligand-binding. It has been speculated that *LGB* might play a similar physiological role hence present in significant quantities in bovine milk (2-3g/L) (Flower, 1996). Lipocalins are involved in several biological processes such as retinol and pheromone transport, synthesis of prostaglandins, immune response and cell homoeostasis (Kontopidis *et al*., 2004). Although various assumptions have been forwarded so far regarding the true physiological function of the *LGB*, it still remains unclear (Kontopidis *et al*., 2002). Since the protein has been found to be highly resistant to gastric proteases both *in vitro* and *in vivo*, it is unlikely to play a nutritive role (Reddy *et al*., 1988). Since *LGB* belongs to lipocalins, Puyol *et al*. (1991) suggested their possible role in transporting retinol and fatty acids. Perez and Calvo (1995) have proposed that *LGB* might play a role in lipid metabolism by activating lipases through free fatty acid uptake. Moreover, Ouwehand *et al*. (1997) suggested other possible function of *LGB* in passive immunity transfer from mother to offspring.

Figure 3. Structure of a β-lactoglobulin subunit. Ribbons denote the secondary structure, with arrows for beta strands and spirals for alpha-helices. Courtesy of Kontopidis *et al*., 2004

1.2.2. The gene

The cDNA encoding for *LGB* (486 bp coding for 162 amino acids) has been found to be highly conserved between bovine, ovine and caprine species. The complete sequence of *LGB* gene is described in bovine (Alexander *et al*., 1993) ovine (Ali and Clark, 1988), and caprine (Folch *et al*., 1994) species. The structural organization of the gene is highly conserved in the three species. The bovine *LGB* gene spans 4.7 Kb which is arranged in seven small exons and six introns (Fig. 4) and is located on chromosome 11. The first exon contains 5' UTR and the sequence encoding for the signal peptide and the next fourteen amino acids. The 3' UTR extends from part of exons 6 (25 bp) to exon 7 (117 bp) (Fig. 4 and Table 3).

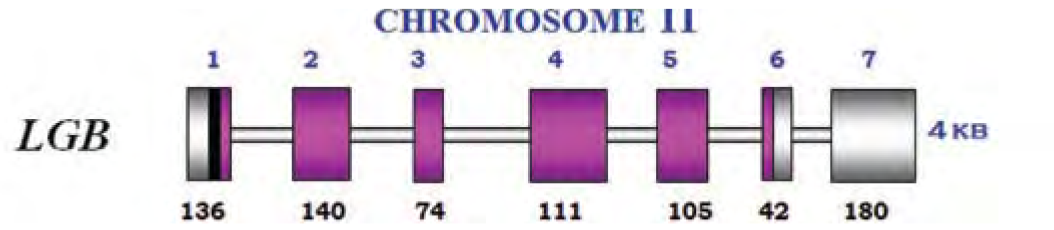


Figure 4. Structure of the bovine *LGB* gene. Open bars represent introns; exons are depicted by large, gray (5′ and 3′ untranslated regions), black (part of exon encoding the signal peptide), and colored (exons and part of exons encoding matured protein). (Courtesy of Martin *et al*., 2002)

Table 3. Exon and intron coordinates of the bovine *LGB* gene. Courtesy of www.ncbi.nlm.nih.gov

Exon			Coding sequence		Intron	
No.	Coordinates	Length (bp)	Coordinates	Length (bp)	Coordinates	Length (bp)
1	1-130	130	41-130	90	131-799	669
2	800-939	140	800-939	140	940-1803	864
3	1804-1877	74	1804-1877	74	1878-2997	1120
4	2998-3108	111	2998-3108	111	3109-3783	675
5	3784-3888	105	3784-3888	105	3889-4110	222
6	4111-4152	42	4111-4127	17	4153-4541	389
7	4542-4718	177	-	-	-	-

Exon Information:

mRNA (NM 173929.3) Length: 779 bp, Number of Exon: 7

Protein (NP 776354.2) Length: 178 aa,

1.2.3. Polymorphism and association with production parameters

Milk protein polymorphism was for the first time detected on bovine *LGB* (Aschaffenberg and Drewry, 1955). Since then 12 polymorphic variants of bovine *LGB* were identified of which variants A and B are the most frequent and are usually known to be associated with milk protein and composition (Martin *et al*., 2002). These two variants differ by two amino acid substitutions in the polypeptide chain arising from two single nucleotide substitutions in the *LGB* gene: 64 (G**A**T) Asp →Gly (G**G**T) and 118 (G**T**C) Val →Ala (G**C**C). The latter T→C substitution creates a *Hae*III restriction site enabling an RFLP analysis at the *LGB* locus (Medrano and Aguilar-Cordova, 1990). The rest of the alleles are less common and generally occur at low frequencies.

Several studies have described the association of the common *LGB* variants with milk composition and cheese making properties. Allele B of *LGB* has been associated with high casein and fat contents while milk derived from cows of AA genotype contained more whey protein than those of variants AB and BB (reviewed in Ng-Kwai Hang, 1998). Therefore, milk with *LGB* protein variant B results in higher cheese yield (Van der Berg *et al*., 1992).

A study conducted on Holstein Friesian cows revealed significant difference of *LGB* variants for milk yield (AB □ AA), total fat (BB and AB □ AA), percent fat (BB □ AA and AB), and lactose yield (AB □ AA) (Tsiaras *et al*., 2005).

It is unlikely that this difference in production parameters is caused by the amino acid substitutions, but rather by different levels of expression of the A and B alleles of *LGB* gene (Ganai *et al*., 2008). Ganai *et al*. (2008) suggested that this differential expression of *LGB* A and B alleles could be because of polymorphisms in regulatory sequences that are in linkage disequilibrium (LD) with the causal genetic polymorphisms of *LGB* protein variants A and B. For instance, 42 polymorphisms in complete LD with *LGB* variants A and B and associated with protein concentration were detected by Ganai *et al*. (2008).

Table 4. Bovine *LGB* genetic variants (Yahyaoui, 2003)

Variant	Amino acid positions											References
	28	45	50	56	59	64	78	108	118	129/130	158	
A						Asp			Val			Ascheffenburg and Drewry, 1955
B	Asp	Glu	Pro	Ile	Gln	Gly	Ile	Glu	Ala	Asp	Glu	
C					His							Bell, 1962
D		Gln										Grosclaede *et al.,*1976
Dr	Asn					Asp			Val			Bell, 1962
E											Gly	Grosclaude *et al.,* 1976
F			Ser							Tyr	Gly	Bell, 1981
G							Met				Gly	Bell, 1981
I								Gly				Baranyi *et al*., 1993

1. Overview of molecular markers and detection techniques

The concept of genetic markers is not new and dates back in the nineteenth century when Gregor Mendel used phenotype-based genetic markers in his experiment. Since the majority of phenotype-based genetic markers offered limited applications, more direct and highly informative DNA-based markers known as molecular markers later came into the picture.

A molecular marker is defined as a particular segment of DNA that is representative of the differences at the genome level. Molecular markers are located at specific locations of the genome and are used to flag the position of a particular gene or the inheritance of a particular trait; hence they are transmitted by the standard laws of inheritance from one generation to the next. Therefore, molecular markers may or may not correlate with phenotypic expression of a trait and can instead be thought of as constant land marks in the genome. Unlike phenotype-based conventional markers, molecular markers offer immense advantage since they prove to be stable and detectable in all tissues regardless of growth, differentiation, development, or defense status of the cell and are not confounded by the environment, pleiotropic and epistatic effects (Agarwal *et al*., 2008).

What is an ideal molecular marker?

A so called "ideal molecular marker" is known to posses the following desirable properties: be polymorphic and evenly distributed throughout the genome, provide adequate resolution of genetic differences, have high reproducibility and allow easy exchange of data between laboratories, have linkage to distinct phenotypes, show co-dominant inheritance (provide distinguishable homozygous and heterozygote states), not very expensive, and require no prior information about the genome of an organism.

However, no currently existing molecular marker is ideal for every situation at hand. The nature of each marker varies with respect to features such as: genomic abundance, level of polymorphism detected, locus specificity, reproducibility, technical requirements and cost. Depending on the type of study to be undertaken, a specific marker system can be identified fulfilling at least some of the characters mentioned above (Weising *et al.*, 1995).

Types of molecular markers

Various types of molecular markers have been utilized to evaluate genetic polymorphisms. These markers are generally classified into different groups based on: mode of transmission, mode of gene action (dominant or co-dominant markers) or method of analysis (Hybridization-based or PCR-based markers).

The following sections deal with the latter *i.e.* the classification of molecular markers based on the method of analysis employed. However, prime focus has been given to PCR-based marker known as Cleaved Amplified Polymorphic Sequence (CAPS) or also called (PCR-RFLP) since it is the method of choice to analyze genetic polymorphism of the two important bovine milk protein genes: *CSN*3 and L*GB*.

1.1.Hybridization-based molecular markers

In these types of markers, DNA profiles are visualized by hybridizing restriction enzyme digested DNA to a chemically labeled probe that is of known sequence origin. Among others, RFLP is the most frequently used hybridization-based molecular marker.

I. Restriction Fragment Length Polymorphism (RFLP)

RFLP markers were initially used in 1975 to identify DNA sequence polymorphisms in genetic mapping of a temperature-sensitive mutation of adeno-virus serotypes (Grodzicker *et al.*, 1975) and were later used for large-scale human genome mapping (Botstein *et al.*, 1980). The technique is primarily based on restriction enzymes that reveal a pattern difference of DNA fragments between individual organisms. Although individuals belonging to the same species posses nearly identical genomes, they usually manifest differences at a few nucleotide sequences due to one of the following causes: point mutation, insertion/deletion, translocation, inversion and duplication (Semagn *et al.*, 2006). Among such DNA sequence differences, some may lay at the restriction recognition sites of endonucleases and may trigger the gain, loss or relocation of a restriction site. Digestion of DNA with restriction endonucleases can subsequently provide several fragments varying in size and number among individuals, populations and species. The underlying procedures and principles of RFLP markers in conjunction with hybridization techniques are schematically depicted in Fig. 5.

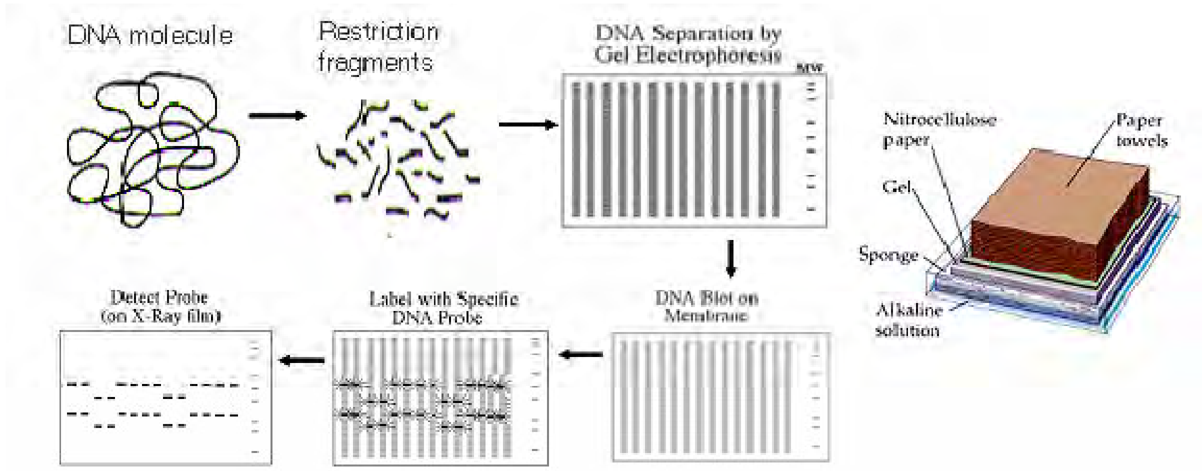


Figure 5. Outline of the different steps of restriction fragment length polymorphism (RFLP) markers. (Courtesy of Agarwal *et al.*, 2008)

Restriction endonucleases recognize specific four, six or eight base pair sequences and cleave whenever these sequences are encountered in DNA. The choice of using these enzymes depends on the resolution required. The four cutters (enzymes recognizing a four base pair sequence) produce the highest resolution since there are numerous such sites in the genome and thus higher chances of identifying single base alterations. The six cutter enzymes are mostly used for RFLP analysis since they are readily available, cheaper and usually produce fragments sizes in the range (200-20,000 bp) which are conveniently separated on agarose gels (Semagn *et al*., 2006). Following restriction enzyme digestion of DNA, the resulting mixture of linear double-stranded molecules with varying lengths are separated by electrophoresis through agarose or polyacrylamide gels.

DNA fragments which are separated by gel electrophoresis are then denatured and transferred onto a solid support (nitrocellulose membrane or filter) through a technique called 'southern blotting' (Southern, 1975). This technique basically transfers the DNA from the gel to a solid support maintaining fragment positions as they were in the gel, yet allowing hybridization reactions to take place. Filter immobilized DNA is then allowed to hybridize to labeled probe DNA, enabling visualization of specific DNA fragments. Most of the probes used for this purpose are single locus, species-specific sequences of about 0.5-3.0 kb bp in size, obtained from a cDNA or genomic library (Staub and Serquen, 1996). Although there are non-radioactive methods available now, labeling was mostly achieved by means of radioactive nucleotides where an autoradiography of membrane/filter reveals the set of fragments complementary to the probe upon exposure to an X-ray film (Agarwal *et al*., 2008).

The information derived from RFLP markers depend on the number of probes and restriction endonucleases employed since each type of probe hybridizes with a particular set of genomic DNA and each enzyme excises a segment of DNA at different points. The main advantages of RFLP markers include: high reproducibility, co-dominant inheritance, good transferability between laboratories, no sequence information required and their relative ease during scoring due to large fragment size variations (Kumar *et al*., 2009).

Application of RFLPs has been hampered, however, due to the following limitations during analysis: requires high quality and quantity of DNA, needs radioactively labeled probes that are relatively expensive and hazardous, low level of polymorphism and detection of few loci per assay. Moreover, it is labour-intensive and tedious.

Reduced variability observed in several domestic animals due to inbreeding made most RFLP sites less-informative (Teneva, 2009). These limitations led to the development of a new set of more informative and less technically challenging methods known as PCR-based markers.

1.2.PCR-based markers

PCR is a molecular biology technique of enzymatically replicating (amplifying) small quantities of DNA without using a living organism. It can be used to exponentially amplify a short (usually less than 10kb), well-defined part of a DNA strand from a single gene, part of a gene, or a non coding sequence (Semagn *et al*., 2006). After the first invention of PCR technology (Mullis *et al*., 1987), several PCR based markers were developed taking into account its apparent simplicity and higher probability of successful execution. Major advantages of PCR based markers over hybridization-based markers include: requirement of small amount of DNA, elimination of radioisotopes in most techniques, high polymorphism enabling generation of many markers in a short time and ability to screen many genes simultaneously.

PCR-based markers can further be classified into two subcategories depending on the nature of the primers used for amplification:

1. Arbitrarily primed PCR-based (sequence non-specific) markers are those developed without prior sequence information [RAPD, AFLP and ISSR]
2. Sequence targeted PCR-based markers are those developed from known DNA sequences [e.g., CAPS, SSR and EST]

2.1.1. Sequence targeted PCR-based markers

With the advent of high-throughput sequencing technology, vast information of DNA sequences for the genomes of several animal and plant species is being continuously generated. For correlating this vast DNA sequence information with particular phenotypes/traits of interest, sequence-specific molecular marker techniques have been developed. The following sections review some of the most commonly used sequence targeted PCR-based markers.

I. Cleaved amplified polymorphic sequence (CAPS) or PCR-RFLP

CAPS is a method developed combining the PCR and RFLP techniques, and it was originally called PCR-RFLP (Maeda *et al*., 1990). Due to its simplicity, CAPS is among one of the most frequently used methods for genetic polymorphism studies (Yahyaoui, 2003). CAPS involve amplification of a target DNA region encompassing the polymorphic restriction enzyme sites by PCR (using 20-25 bp primers) and subsequently digesting the amplicon with the respective restriction enzymes. Primers are designed using sequence information stored in molecular databases of genomic or cDNA sequences or cloned RAPD bands (Agarwal *et al*., 2008).

This technique has provided an alternative to exploit DNA sequences of mapped RFLP markers thereby eliminating the tedious southern blot hybridization and radioactive detection procedures (Komori and Nitta, 2005). CAPS detect the restriction fragment length polymorphisms caused by single base changes like SNPs, and insertions/deletions that modify restriction endonuclease recognition sites situated in PCR amplicons. Approximately 30-40% of SNPs has been reported to alter restriction endonuclease recognition sites which are commonly referred to as snip-SNPs (Wicks, 2001). This assay is performed through restriction digestion of locus-specific PCR products with one or more endonucleases, followed by separation on agarose or polyacrylamide gels (Fig. 6).

The CAPS markers are locus specific and are mostly inherited in a co-dominant manner enabling their widespread application to distinguish homozygotes from heterozygotes for specific alleles (Komori and Nitta, 2005). CAPS are thus affordable and practical options for genotyping, positional or map-based cloning and molecular identification studies where sequence-based identification is not necessary or impractical (Spaniolas *et al.* 2006; Weiland and Yu 2003). Most CAPS genotypes are scored and interpreted easily allowing data sharing between different laboratories. Moreover, since CAPS analysis is versatile; it has been possible to combine it with techniques such as single strand conformational polymorphism (SSCP), AFLP or RAPD to boost the possibility of finding DNA polymorphisms (Agarwal *et al*., 2008).

However, the application of CAPS has been limited since the development of these markers is only possible where mutations disrupt or create a restriction enzyme recognition site. To overcome such limitations, Michaels and Amasino (1998) proposed an alternative marker named dCAPS (derived cleaved amplified polymorphic sequence), which is a variant of the CAPS method. In dCAPS analysis mismatches are generated in a PCR primer, which are subsequently used to create a polymorphism based on the target mutation (Michaels and Amasino, 1998). The modified PCR product is then digested with restriction enzymes followed by determination of the absence or presence of the SNP through the resulting restriction pattern.

Another downside of CAPS when compared with RFLP analysis is the difficulty of finding polymorphisms within the limited size of the PCR amplified fragments (300-1800 bp) (Agarwal *et al*., 2008). Furthermore, the requirement of sequence data to synthesize primers adds to its limitations.

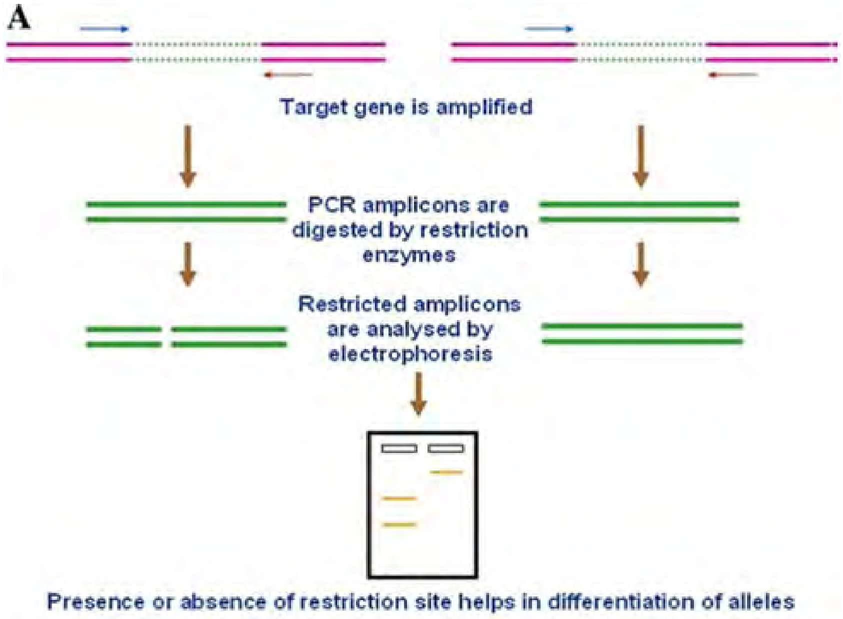


Figure 6. Schematic description of CAPS technique (To generate CAPS marker, the requisite DNA sequence is first amplified and then the amplicon is digested by the restriction enzyme. Based on the gel analysis the presence or absence of the restriction enzyme is elucidated. The presence or absence of the restriction helps to differentiate allelic differences). (Courtesy of Agarwal *et al*., 2008)

II. Sequencing and Single nucleotide polymorphism (SNPs)

In recent years, public access of genome sequences of several organisms has fueled the study of sequence variations within and among individuals. Such studies revealed the abundance of single nucleotide polymorphisms (SNPs), and insertion and deletion (InDels) in genomes of both plants and animals (Semagn, 2006). SNPs are single nucleotide variations in genome sequences of individuals of a population. SNPs are usually more prevalent in the non-coding region of the genome while in the coding regions, an SNP is either non-synonymous (those resulting in an amino acid sequence change) or synonymous (those that do not alter the amino acid sequences although might modify mRNA splicing resulting in phenotypic differences) (Sunyaev *et al*., 1999; Richard and Beckman, 1995). Most SNP genotyping assays are based on either of the following mechanisms: Allele specific hybridization, primer extension, oligonucleotide ligation or invasive cleavage (Semagn, 2006). When targeting a defined region such as those containing candidate genes, the simplest approach is to perform direct sequencing of genomic PCR products obtained in different individuals in a population.

2. Application of molecular markers in livestock improvement

The vast majority of economically important traits in livestock production are quantitative in nature and show continuous distributions due to the action (and interaction) of a large number of genes/loci that often have small contributions individually (Lande and Thompson, 1990; Hayes and Goddard 2001).

It is evident that the improvement of livestock regarding these quantitative traits of interest had benefited enormously from the conventional breed improvement programs implemented so far. Conventional animal breeding programs primarily rely on phenotypic selection of animals where traits are usually measured directly and those with superior performance are used as a breeding stock. For sex-linked traits such as milk production, progeny testing schemes (Estimated breeding values) have allowed estimation of genetic merit since males do not manifest the trait of interest. However, there are several limitations associated with phenotypic selection: influenced by environmental effects, limited to traits that are easily measured and involvement of high cost (Beuzen *et al*., 2000).

Recently, the use of molecular marker technologies has potentially offered a superior alternative that generates greater genetic gain due to the following properties: molecular information generated is hardly affected by environmental effects, selection of breeding animals at early age reduces generation intervals, and information can be obtained for sex-limited traits, traits that are expensive or difficult to record or carcass traits (Mitra *et al*., 2000).

Molecular markers are recently playing an immense role in livestock improvement programs through both conventional and transgenic breeding strategies. The various application of molecular markers can be divided into short-range (immediate) and long-range applications.

I. Short range (immediate) applications

Immediate applications of molecular markers includes: parentage determination, genetic distance estimation, determination of twin zygosity and freemartinism, sexing of pre-implantation embryos and identification of disease carrier.

II. Long range applications

The most common long range application of molecular markers in conventional breeding strategies involves mapping of several Quantitative Trait Loci (QTL) through linkage analysis. Such mapping information when available for loci playing significant role on performance traits and disease resistance can be used in breeding programs by either marker assisted selection (MAS) of young sires or in breed introgression programs (transgenics) (Naqvi, 2007).

2.1.Quantitative Trait Loci (QTL) mapping

Many studies have indicated that knowledge about genetic markers associated with genes affecting quantitative traits can increase the selection response of animal breeding programs. Specially for difficult to improve traits using the conventional selection methods (Meuwissen and Goddard,1996; Meuwissen and van Arendonk, 1992; Meuwissen *et al.*, 2002).

The major objective of QTL studies is to locate genes/markers that can be implemented in breeding programs via marker assisted selection (MAS). So far, two approaches have been employed to decipher QTL: the candidate gene approach and QTL mapping.

The candidate gene approach assumes that a particular gene responsible for the physiology of a trait could accumulate/harbor a mutation causing variation through time. Here the gene or part of the gene are first sequenced in a number of different animals and any of the DNA sequence variation observed are evaluated for a possible phenotypic association (Hayes, 2007).

In the QTL mapping approach, chromosome regions which are found to be associated with variations in phenotypic traits are identified. QTL mapping is conducted under the assumption that actual genes which influence quantitative traits are not known. This approach instead exploits the potential of neutral DNA markers and searches for association between allele variation at the marker and variation in quantitative traits (Hayes, 2007).

A DNA marker is an identifiable physical location on a chromosome whose inheritance can be monitored. It could be either an expressed region of DNA (gene) or usually certain segment of DNA with unknown coding function but with determined inheritance pattern. Available DNA markers can be used to determine whether variations at the molecular level are linked to variation in a quantitative trait. If such linkage is proved to exist, then the marker is linked to, or on the same chromosome as a QTL which has allelic variants causing variation in the quantitative traits.

1.2.Marker assisted selection (MAS)

In conventional breeding methods, selection of animals is mainly based on phenotypic and pedigree data which are then combined to generate estimated breeding values (EBVs). Currently, the development of DNA-based markers has offered a third alternative data for the selection of superior breeding stocks (Hayes, 2007).

The idea of MAS relies on the utilization of information from polymorphic loci as an aid to selection. Implementing MAS involves two steps: Identification of the marker loci that is known to be linked to QTL of economic importance and subsequent application of linkage association studies in genetic improvement programs (Naqvi, 2007). Marker assisted selection (MAS) usually relies on one of the following strategies: DNA in linkage equilibrium with a QTL (LE-MAS), molecular markers in linkage disequilibrium with a QTL (LD-MAS) or selection of the actual mutation causing the QTL effect (Gene-MAS). In LE markers, the linkage between the markers and QTL is not sufficiently close to ensure marker-QTL allele relationships. The LD-MAS is the most important genome scanning approach that uses anonymous molecular markers to identify QTL affecting most economic traits of the livestock industry (Hayes, 2007).

It is worthwhile noting that MAS is likely to compliment rather than replace the traditional breeding programs. MAS enables increased rate of genetic improvement through improved selection intensity, reduction of generation interval, accurate prediction and ease in selecting individuals early in life or for sex limited traits (Gholizadeh *et al.*, 2008).

According to Geldermann (1990), there are two possible approaches for identifying markers that affect QTL:

1. **Polymorphism in the coding sequences** – refers to polymorphisms occurring in and around the structural or regulatory sequences of a gene with physiological significance such as hormone and milk protein genes (Naqvi, 2007). Such variability may directly affect gene expression by changing mRNA splicing and stability, rate of gene transcription or the sequence of gene product. These changes have been known to contribute to the phenotypic variations among individuals both in terms of productivity and disease resistance/susceptibility (Beckmann and Soller, 1987). Evidently, DNA polymorphisms existing in such genes with *a priori* tendency to be closely linked with traits of importance can be used as markers (Naqvi, 2007).

Several studies have indicated that a number of single point mutations in structural genes that are inherited in a simple Mendelian manner are associated with quantitative traits of economic relevance. For instance, several milk protein genetic variants have been correlated with the differences in composition and processing qualities of milk (Rachagani and Gupta, 2007; Lunden *et al.*, 1997; Ng-Kwai-Hang, 1998; Robitaille *et al.*, 2002)

2. **Polymorphisms in the non-coding sequences**- here polymorphisms in non-coding sequences such as flanking regions or intergenic regions are indirectly employed as markers for linkage analysis. Presently, highly polymorphic markers like microsatellites are being utilized to locate QTL for economically important traits. Georges *et al.* (1995) detected the presence of QTL for milk production on five bovine chromosomes using 159 microsatellite markers in 14 Holstein half-sib families.

2. Quantification of genetic diversity at polymorphic loci

2.1.Diversity and differentiation: measures of genetic variability

In most population genetics studies, diversity and differentiation serve as estimators of genetic variation. Diversity is a measure of individual variation within a population taking into account the number and frequencies of different variants/types while differentiation measures variation between two or more populations or subpopulations (Gregorious, 1987). Based on specific models and probabilistic concepts, several parameters have been developed as diversity estimators. Among them, Heterozygosity (H) is the most widely used measure of genetic variation within a population. Natural selection and genetic drift which are associated with reduced genetic diversity and high endogamy are known to favor genetic differentiation between populations. Conversely, gene flow enhances homogenization of genetic diversity between populations resulting in low differentiation and high genetic diversity. Various methods have been used to quantify the level of differentiation in subdivisions of a population among which is Wright's F-statistics. Wright (1951, 1965) developed three correlation coefficients which allocated the genetic variability to total population level (T), subpopulations (S) and individuals (I).

There are various standard population genetics parameters that measure genetic diversity for quantitative traits such as milk, protein and fat yield. These parameters can be estimated at two levels: within populations and among populations. Genetic diversity within populations is a function of total genetic diversity within the species and the proportion of the total genetic diversity found among its populations. Among some of the most common genetic parameters are percentage of polymorphic loci (P), mean number of alleles per locus (A), mean number of alleles per polymorphic locus (AP), effective number of alleles (A_e), expected heterozygosity (H_e) and observed heterzygosity (H_o). The usefulness of the parameters P, A and AP is dependent on the number of loci surveyed rendering them less informative when few loci are considered. In the coming sections the parameters A_e, H_e and H_o are described since only two loci were studied in this paper i.e. *CSN*3 and *LGB*.

2.2. Within-population versus among-population statistics

2.2.1. Estimators of heterozygosity/diversity

I. The effective number of alleles (A_e)

The effective number of alleles (A_e) is calculated as the reciprocal of the sum of squared allele frequencies:

$$\mathbf{A_e = 1/\sum_{i=1}^{n} p_i^2}$$

$$\mathbf{= 1/1\text{-}H_e,} \quad 1 < Ae < \infty$$

Where p_i is the allele frequency of the i^{th} allele and H_e is the expected heterozygosity

A_e measures the number of alleles at a locus and the evenness of allele frequencies at that locus. It should be noted, however, that equal values of A_e could be realized for populations with different distributions of allele frequencies. Usually mean A_e over all loci is reported and its value typically ranges from 1.1 to 1.4. Since A_e and H_e are positively correlated, these measures exhibit similar aspects of genetic diversity.

II. Expected Heterozygosity (H_e)

Expected heterozygousity (H_e), also known as gene diversity, is a composite measure that summarizes genetic variation at the allele level. It is a concept introduced by Nei (1978) to explain the probability that two randomly selected alleles are different. It is calculated as:

$$H_e = 1-\sum_{i=1}^{n} p_{ik}^2$$

Where P_i is the frequency of allele i at locus k; n is the number of alleles at locus k

III. Observed heterozygosity [H_o]

Observed heterozygosity [H_o] is calculated directly from observed genotype frequencies for each locus and population. H_o is less informative as a comparative measure of diversity since it is affected by inbreeding and other evolutionary processes violating the assumptions of Hardy-Weinberg's equilibrium.

$$H_o = \sum_{i=1}^{n} n_i/N$$

Where n_i is the number of heterozygous individuals at the ith locus; N is the total number of individuals analyzed.

2.2.2. Wright's F-statistics

Wright's fixation indices, or F-statistics, are one of the most widely used genetic parameters to describe population structure. Wright (1965) developed three F-statistics to summarize the genetic structure of a population and its subpopulations. He defined these statistics for one locus with two alleles, in terms of correlations between gametes uniting in an individual. F-statistics are basically measures of the fixation of alleles at various levels of organization i.e. individual, subpopulation and population. Using the more intuitive concepts of observed and expected heterozygosities; Nei (1977) reformulated the F-statistics for multiple alleles. The three coefficients or fixation rates are described as follows.

- **Inbreeding coefficient (F_{IS})**: is the proportional deviation of observed from expected heterozygosity within populations. It is the mean reduction in heterozygosity of an individual due to non-random mating within a sub-population. In another word, F_{IS} measures the extent of genetic inbreeding within sub-populations. The value of F_{IS} ranges from -1.0 when all individuals are heterozygous to 1.0 when there are no heterozygotes observed.

$\mathbf{F_{IS} = (H_S\text{-}H_o)/H_S}$, $\mathbf{-1 \leq F_{IS} \leq 1}$

Where H_S is the mean expected heterozygosity within random mating subpopulations ($2p_iq_i$), H_0 is mean observed heterozygosity per individual within sub-populations and H_T is expected heterozygosity in random mating total population (2pq)

- **Fixation index (F_{ST})**: is the mean reduction in heterozygosity of a sub-population (relative to the total population) due to genetic drift among sub-populations. It measures the extent of genetic differentiation among subpopulations. Its value can range from 0.0 when there is no differentiation to 1.0 whenever there is complete differentiation i.e. populations being fixed for different alleles.

  $\mathbf{F_{ST} = H_T\text{-}H_S/H_T}$, $\mathbf{0 \leq F_{ST} \leq 1}$

- **Overall fixation index (F_{IT})**: is the proportional deviation of observed from expected heterozygosity in the pooled population. In another word, it is the mean reduction in heterozygosity of an individual relative to the total population. F_{IT} combines contributions from non-random mating within sub-populations (F_{IS}) and effects of random drift among sub-populations (F_{ST}).

  $\mathbf{F_{IT} = (H_T\text{-}H_o)/H_T}$, $\mathbf{-1 \leq F_{IT} \leq 1}$

  F_{IS} and F_{IT} are measures of the deviation from Hardy-Weinberg equilibrium within subpopulations and in the total population respectively. The relationship between the three F-statistics is:

  $\mathbf{F_{ST} = F_{IT}\text{-}F_{IS}/(1\text{-}F_{IS})}$

2.2.3. Nei's diversity index and genetic distance

Expected heterozygosity also called genetic diversity; following Nei (1973, 1978) can be partitioned into within-population and among population components. The total genetic diversity at a locus ($H_T = 1\text{-}\sum p_i^2$, where pi is the mean frequency of the i^{th} allele in the pooled sample) is the sum of the mean genetic diversity within populations (Hs) plus the among-population genetic diversity (D_{ST}).

$$\mathbf{H_T = Hs + D_{ST}}$$

Nei's (1973) statistic, G_{ST}, was used to measure among-population differentiation, relative to the total diversity:

$$\mathbf{G_{ST} = D_{ST}/H_T, \qquad 0 \leq G_{ST} \leq 1}$$

$$\mathbf{G_{ST} = 1\text{-}H_S/H_T}$$

G_{ST} values are often averaged over all polymorphic loci to estimate population divergence. Numerically, a G_{ST} value of 0.05 means that 95% of total genetic diversity resides within populations and the remaining 5% resides among populations. All alleles are used in the calculation of H_T and Hs and, thus G_{ST} may be considered as a multiallelic form of Wright's F_{ST}, which was originally formulated for loci with two alleles. In case of only two alleles, G_{ST} is identical to F_{ST}. According to Schoen and Brown (1991), species with high interpoulation variation in allele frequencies with high G_{ST} are also likely to have high population to population heterogeneity (H_e). Thus inbreeding species which generally have high G_{ST} values tend to have low amounts of genetic diversity in most of their populations. Conversely, out-crossing species have lower G_{ST} values with less variation among populations.

Nei's genetic distance (D)

Nei's (1972) genetic distance has found the most widespread use and was designed to quantify the difference or distance between individuals or populations. Primarily, the normalized genetic identity (I) for a given locus is computed as:

$$\mathbf{I=J_{xy}/}\sqrt{\text{JxxJyy}} \qquad \mathbf{0 \leq I \leq 1}$$

J_{xx} is the probability that two alleles chosen at random from a given population *X* are identical *i.e.* $J_{xx} = \sum P_i{}^2{}_{ix}$ (sample gene identity in population *X)*; whereas J_{yy} is the analogous probability of population Y (sample gene identity in population *Y)*, and $J_{xy} = \sum P_{ix}P_{iy}$ (sample gene identity in both population *X* and *Y*). The genetic distance is then obtained by taking the negative of the natural log of the genetic identity (I):

$$\mathbf{D = -ln(I), \qquad 0 \leq D \leq \infty}$$

Two populations with equal allele frequencies have thus a genetic distance of 0 and a genetic identity of 1, even if they might have different genotype proportions. For a population containing all homozygotes (F=1) or all heterozygotes (F= -1), the genetic distance value assumes a 0 value. In such populations, G_{ST} will also be 0, since G_{ST} (and F_{ST}) is calculated from the expected heterozygosity which is entirely based on allele frequencies.

Nei (1978) modified D with sample size correction and termed it Nei's unbiased genetic distance. The unbiased estimators for J_X ($J'x$) and J_Y ($J'y$) are given by:

$$J'x = \frac{1}{r}(\sum_{j}^{r}(n_{xj} \sum_{i}^{mj} x_{ij}^{2} - 1)/(n_{xj} - 1)$$

$$J'y = \frac{1}{r}(\sum_{j}^{r}(n_{yj} \sum_{i}^{mj} y_{ij}^{2} - 1)/(n_{yj} - 1)$$

Where, n_{xj} and n_{yj} are the number of chromosomes examined at the j^{th} locus for populations X and Y. Thus the unbiased estimate of D_{ST} is calculated by substituting J_X and J_Y by J'_X and J'_Y, respectively.

2.3.Allelic and Genotypic frequencies

Allele frequencies are considered to be the basic genetic parameters of a population hence evolution has been defined as a change in allele frequencies through time (Mayr, 1963). Genotypic frequencies are relatively less informative since populations with different genotypic frequencies can have identical allele frequencies. Allele frequency is the number of copies of a particular allele divided by the number of copies of all alleles at a locus in a population. In population genetics, allele frequencies are widely used to depict the amount of genetic diversity at the individual, population and species level. If there are n_i copies of the i^{th} allele in a sample population, the sample frequency, p_i of the i^{th} allele is calculated as:

$$\boldsymbol{p_i = n_i/2n}$$

where n is the total number of sampled individuals

Genotypic frequencies are additional basic genetic parameters of a population. Molecular patterns of DNA bands on a gel reveal the number of homozygotes and heterozygotes which in turn enable to compute the observed genotypic frequencies. The frequency of a particular genotype can be computed by dividing the number of individuals possessing that genotype (n) to the total population sampled (N) i.e. n/N.

2.4.Hardy-Weinberg equilibrium/proportions

Based on Mendel's principles of inheritance, G.H. Hardy and Wilhelm Weinberg independently developed the concept that is known today as the 'Hardy-Weinberg equilibrium' (also known as the Hardy-Weinberg proportion/principle).

The Hardy-Weinberg equilibrium states that both allele and genotype frequencies in a population remain constant or are in equilibrium from one generation to the next given that there are no disturbing influences such as nonrandom mating, mutation, selection, limited population size, overlapping generations, random genetic drift and gene flow. However, it is crucial to note that outside the lab one or more of these influences are in effect. Therefore, no real/natural population exactly satisfies all of the above stringent criteria simultaneously. The concept of genetic equilibrium is thus, an ideal state that provides a baseline to measure genetic change in populations across generations. This equilibrium can be mathematically expressed based on simple binomial (for two alleles) or multinomial (for multiple alleles) distribution of the allele frequencies.

In the simplest case of a single locus with two alleles; the most familiar format considers a locus denoted A with alleles A_1 and A_2. The respective frequencies of alleles A_1 and A_2 are depicted as p_1 and p_2 respectively with ($p_1+p_2 =1$). If the population is in Hardy-Weinberg equilibrium, the proportions/frequencies of the different genotypes are as follows:

A_1A_1: p_1^2, A_1A_2: $2p_1p_2$, A_2A_2: p_2^2

For a locus with k distinct alleles and allele frequencies p_i ($\sum p_i =1$), the expected HWP are: p_i^2 for homozygotes (A_iA_i), and $2p_ip_j$ for heterozygotes (A_iA_j), where $i<j$, and $i, j =1, 2,\ldots k$.

These expectations are derived from Mendel's first law of independent segregation of the two parental alleles when applied to population level variation.

2.4.1. Testing for fit to HWE [The chi-square test of HWE]

A population is said to be at Hardy-Weinberg equilibrium (HWE) for a particular locus (gene) when the observed genotype frequencies are not statistically different from the frequencies determined by the appropriate products of its allele frequencies *i.e.* expected genotype frequencies. In testing HWE, allele frequencies are estimated from the observed genotype counts and then the expected genotype counts under HWP are estimated from these allele frequencies. The most straightforward test of HWE is a simple chi-square (χ^2) test, comparing counts of observed genotypes with those expected under HWE. The χ^2 is calculated by taking an observed genotypic frequency (*O*), subtracting the expected frequency (*E*), and then squaring this difference. Each difference is divided by the expected frequency, and these standardized differences are finally summed. As with most test statistics, the larger the difference between observed and expected, the larger the test statistic becomes. Since χ^2 has nearly the theoretical chi-square distribution, once we know the χ^2 test statistic, we can calculate the probability of getting that value.

$$\chi 2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

O_i- Observed genotypic frequencies of genotype i

E_i- Expected genotypic frequencies based on HWE of genotype i

In the chi-square test, the null hypothesis is that genotype frequencies for a particular locus in a given population are in HWE and the alternate hypothesis (H_a) is that they are not in HWE. For instance, if the frequency of one allele in a population is p and the other allele is q, the null hypothesis is that expected frequencies of the three genotypes are $p2$, $2pq$, and $q2$. It should be noted that this test of HWE cannot be applied to traits with a recessive or dominant mode of inheritance. The trait must instead show either co-dominance or incomplete dominance enabling to discriminate between heterozygous and homozygotes.

2.4.2. Deviations from the HWE

Most of the statistical tests used to detect significant deviation from the HWE have usually low power and when relatively well defined populations are considered, the majority of genetic markers do not often manifest deviations from the HWE. However, whenever significant deviations are seen, genotyping errors or potential population stratification are the first considerations rather than the operation of natural selection.

Violations from HWE cause deviation from expectations which in turn affect populations depending on the assumptions violated. The HWE states that a population will have the Hardy–Weinberg genotype proportions after a single generation of random mating within the population. When this assumption of random mating is violated, the population will not have HWE. Violations of this provision include inbreeding, assortative mating and small population size and are known to increase homozygosity.

The remaining assumptions (selection, mutation and migration) affect the allele frequencies, but do not, in themselves affect random mating. If a population violates one of these, the population will continue to have HWE each generation, but the allele frequencies will change with these forces.

III. MATERIALS AND METHODS

1. Study population

Five indigenous and one cross-bred cattle populations were considered for the study. A total of 83 blood samples were collected from experimental cattle herds of Boran (11), and Barka-Holstein Friesian cross-bred (11) cattle maintained at the DebreZeit Agricultural Research Center (DARC) of the Ethiopian Institute of Agricultural Research (EIAR) and pure-breeding experimental cattle herds of Sheko (16), Horro (23), Abigar (11), and Guraghe (11) maintained at the Ghibe cattle breeding station of the International Livestock Research Institute (ILRI) and the Ethiopian Institute of Agricultural Research (EIAR) in western Ethiopia

Animals were selected randomly from each breed with intentional exclusion of very closely related ones.

1.1.Description of the breeds

The study included 5 indigenous cattle breeds of Ethiopia belonging to 4 of the 10 major cattle breed groups of Africa reported by Rege (1999): Abigar (Sanga), Horro (Zenga or Zebu X Sanga), Boran and Guraghe (Zebu) and Sheko (African taurine).

Horro: This breed was developed through interbreeding of various Abyssinian highland Zebu and Nilotic Sanga, particularly the Abigar. They are of medium to large size, with small and finely shaped head, a straight profile and medium to large horns. The breed inhabits the highlands of north-western Ethiopia, mainly in the Horro Gudru area of eastern Wollega as well as adjoining parts of western Shoa and Illubabor. Although the total number of population is not yet recorded, DAGRIS (2007) reported that the breed is not at risk of genetic erosion.

Sheko: The Sheko are believed to be the last remnants of the original humpless Shorthorn (Bos taurus) cattle in eastern Africa. It is the only known taurine type breed in Ethiopia. At present some of the Sheko manifest small humps inherited from zebu through introgression. The breed is now considered endangered through gradual interbreeding with local zebu and sanga (Abigar) with an estimated population size of 2,400 (Rege, 1999). The breed inhabits the humid parts of south-western Ethiopia with the Sheko tribe around Bench zone and it is believed to have some level of trypanotolerance (DAGRIS, 2007).

Boran: The breed originally descended from the first introduction of zebu into Africa from West Asia. It established its presence first in the semi-arid and arid pastoral Borana plateau of southern Ethiopia. Pastoral movements and migrations led to spread of the Ethiopian Boran to the eastern rangelands in Ethiopia as well as into northern Kenya and south-western Somalia. The main location of Ethiopian Boran is the southern rangelands of Ethiopia, around Liben, Mega and Arero plains with the Borana pastoralists and bordering area of northern Kenya (DAGRIS, 2007).

Guraghe: These cattle types are not well studied and characterized as a breed. They were classified into Abyssinian short horned zebu or Ethiopian highland zebu (Rege *et al.*, 2001). They are small-sized, usually with red, chestnut or roan coat color, and are mainly found in Guraghe, Hadiya, and the tsetse-infested valleys of the Ghibe tributaries areas (Rege *et al.*, 2001; DAGRIS, 2007).

Abigar: It is classified into Sanga type that evolved in Ethiopia from early migrations and spread across southwestern Ethiopia into southern Sudan. As a result, these cattle are found around the White Nile in the Sudan and adjacent lowlands of southwest Ethiopia where they are mainly kept by the Nuer people in the Akobo area of the Gambella region (DAGRIS, 2007).

2. Sample collection and DNA extraction

Blood samples were collected from a total of 83 animals belonging to the following breeds/groups: Abigar (n=11), Boran (n=11), Guraghe (n=11), Horro (n=23), Sheko (n=16) and Holestin-Barka (n=11) cross bred cattle. Blood was drawn from the jugular vein of each animal into 10ml vacuum tubes containing EDTA (Ethylene Diamine Tetraacetic Acid) as an anticoagulant and kept on ice until transferred to -20 ^{o}C freezers.

Genomic DNA was extracted from whole blood using salting out procedure (Bruford *et al.*, 1992). Blood samples were thawed to room temperature and 5 milliliter of blood was then mixed with 15 ml of cold EL buffer [0.155M NH_4Cl, 10mM KCO_3, and 1mM EDTA, pH 7.4] in a falcon tube and placed on ice for 15 minutes, mixing occasionally. After the tube was centrifuged for 10 minutes at 800 g, the supernatant was removed retaining the pellet. The pellet was washed twice with EL buffer until there was no sign of hemoglobin and re-suspended in 3ml lysis buffer [KL buffer: 10mM Tris, 2mM EDTA, 0.4 M NaCl, pH 8.2]. The resulting cell lysates were incubated overnight at 37 °C with 100 µl proteinase K [20 mg/ml] and 150 µl 20% SDS following gentle shaking. After digestion was complete, 1.5 ml of saturated NaCl [6M] was added to each tube and shaken vigorously for 15 seconds, followed by centrifugation at 1700 g for 15 minutes at room temperature. The precipitated protein pellet was left at the bottom of the tube and 600 µl of the supernatant containing the DNA was transferred to a 2ml microcentrifuge tube. Two volumes [1200 µl] of absolute ethanol maintained at room temperature were added and the tubes inverted several times until the DNA precipitated. The precipitated DNA was spooled out with a plastic pipette, rinsed in 70% ethanol a few times and finally transferred to a 1.5 ml microcentrifuge tube containing 0.5 ml TE buffer [10 mM Tris-HCl, 0.1 mM EDTA, pH 7.4]. The DNA was allowed to dissolve overnight at 4°C.

The quality and intactness of the DNA was checked on 0.6% agarose gel. The purity and concentration of DNA was determined by taking spectrometric readings using Gene Quant Pro RNA/DNA Calculator. Samples having optical density ratio [OD_{260}/OD_{280}] between1.6 - 2.0 were considered acceptable and used for Polymerase Chain Reaction [PCR].

3. Genotype determination by PCR-RFLP assays

3.1.PCR-RFLP assay for *CSN*3 genotypes

For the detection of *CSN*3 genotypes, a 633 bp DNA fragment which consists a part of intron III (4 bp), exon IV (516 bp) and a part of intron IV (113 bp) was amplified by polymerase chain reaction (PCR) using primers described by Rachagani and Gupta (2007) [Forward: 5’-CAG CGC TGT GAG AAA GAT GA -3’ and Reverse: 5’-CCC ATT TCG CCT TCT CTG TA -3’].

Figure 7. Schematic diagram of bovine *CSN*3 gene PCR amplification target and primer binding locations. Exons are schematically depicted as boxes, blue (5’ and 3’ UTRs), and red (coding regions). Courtesy of NCBI (www.ncbi.nlm.nih.gov)

The PCR amplification reactions contained: 100 ng of genomic DNA, 0.5 µM of each primer, 5X PCR buffer, 2.5 mM $MgCl_2$, 0.2 mM dNTPs, 1 U of *Taq* DNA polymerase and sterilized

ddH_2O to a final volume of 50 µl. The amplification was carried out in a thermal cycler (GenAmp PCR system 2400, Perkin Elmer Cetus corp., Norwalk, CT) with the following steps: pre-denaturation for 3 min at 94 °C followed by 30 cycles of 94 °C for 30 s, 58 °C for 1 min, 72 °C for 2 min, and a final extension of 10 min at 72 °C. To confirm amplification of the target region, PCR products were separated by agarose gel (1.7% w/v) electrophoresis using a 100 bp ladder size marker. The gel was run using 1X TBE buffer and stained with ethidium bromide (1.75µg/ml) for 30 minutes and rinsed with distilled water for about 20 minutes to remove excess ethidium bromide. Gel pictures were taken using a gel documentation system (Biometera BDA Digital).

Restriction digestion of the *CSN*3 PCR products was carried out using two restriction endonucleases: *Hinf* I and *Hind* III. The digestion reaction for *Hinf* I was set up with 4.65 µl of ddH_2O, 2.25 µl of restriction endonuclease buffer 2 (New England Biolabs), 0.6 units of *Hinf* I and 10 µl of PCR product incubated at 37 ^{o}C for about 3 hours. The digestion reaction for *Hind* III was set up with 2 µl ddH2O, 2.5 µl of restriction endonuclease buffer 2(New England Biolabs), 0.6 units of *Hind* III and 10 µl of PCR product incubated at 37 ^{o}C for about 3 hours. The digested PCR products were loaded and separated on 2% agarose gels using a 100 bp marker. The gels were stained with ethidium bromide and photographed using a gel documentation system (Biometra BDA Digital). The different fragments were then scored manually by direct counting.

3.2.PCR-RFLP assay for *LGB* genotypes

To detect *LGB* genotypes, the sequence (GenBank Accession # NC_007309) deposited by the Baylor College of Medicine, Human Genome Sequencing Center was used for primer pair designing using the Primer3 software (Rozen and Shaletsky, 2000). The sequence of the designed forward primer is (5'- CCT GCT GGA ACT CAC TTT CC-3') and that of the reverse primer is (5' ACC TGC CAT TTG TTT TCA GG -3'). The primers covered a target of 529 bp DNA fragment which consists of a part of intron III (41 bp), exon IV (111 bp) and a part of intron IV (377 bp).

Figure 8. Schematic diagram of the bovine *LGB* gene amplification site and primer binding locations. Exons are schematically depicted as boxes, blue (5' and 3' untranslated regions), and red (coding regions). Courtesy of NCBI (www.ncbi.nlm.nih.gov)

PCR amplification of the *LGB* gene, gel electrophoresis and band visualization were performed in the same manner as that of *CSN*3 gene mentioned above.

Restriction digestion of *LGB* PCR product was undertaken using two restriction enzymes: *Hae* III and *Dde* I. The digestion reaction for *Hae* III was set up with 4.89 µl of ddH2O, 2.25 µl of restriction endonuclease buffer 4 (New England Biolabs), 3.6 units (0.36 µl) of *Hae* III and 10 µl of PCR product incubated at 37 °C for about 3 hours. The digested PCR products were loaded and separated on 2% agarose gels in 1XTBE buffer using a 100 bp marker. The gels were stained with ethidium bromide and photographed using a gel documentation system (Biometra BDA Digital). The different fragments were then scored manually by direct counting.

3.3.DNA sequencing

After conducting the PCR-RFLP analysis, the PCR products of *CSN*3 and *LGB* which showed different restriction digest patterns/ genotypes were cleaned-up using PEG (Polyethylene Glycol) precipitation method (Rosenthal *et al.*, 1993). Since the PCR reaction mix contains primers, primer-dimers, salts, dNTPs and proteins, it was cleaned-up from these constituents to avoid their possible interference during sequencing. For this purpose, 50µl of the PCR product was first mixed with 50µl of the PEG solution and incubated at 37°C for 15 minutes. The supernatant was discarded after the product-PEG mix was centrifuged at 15,000 g for 15 minutes. The product was then precipitated with 80% ice cold ethanol and left to air dry until there was no sign of ethanol. Finally the purified PCR product was dissolved with 25µl sterile deionized water and stored at -20°C until sequencing. To check and roughly quantify recovery, 4µl of the cleaned PCR products were run on 1.7% agarose along with 100 bp DNA ladder.

Following the respective PCR-RFLP assays, a total of 19 purified PCR fragments belonging to the different *CSN*3 and *LGB* homozygous genetic variants were selected from each breed and sequenced to determine the precise nucleotide sequence. Sequencing was performed in both forward and reverse directions using the same primers used for PCR amplification.The purified PCR products were sequenced using the big dye termination technique, ABI Capillary system (genetic Analyzer 3730-48) at BecA-ILRI Hub, Segolip Unit (Nairobi campus).

3.4.Sequence data and Single Nucleotide Polymorphism (SNP) analysis

BioEdit version 5.0.6 (Hall, 2001) was used to edit the DNA sequences after visually inspecting the chromatograms of every sequence data. Sequence data obtained from the forward and reverse primers where crosschecked to confirm the polymorphic sites detected. Ambiguous sequences at the beginning and extreme end of each sequence were trimmed to avoid possible errors in base calling.

For comparing the sequences with the existing nucleotide sequence database, the Basic Local Alignment Search Tool (BLAST) of the National Center for Biotechnology Information (NCBI) was employed (http://www.ncbi.nlm.nih.gov/). To locate different SNPs and search for previously reported haplotypes, *CSN3* nucleotide sequence positions were compared with GeneBank No. AY380228.1 corresponding to *CSN3*-A allele and that of *LGB* was compared with GeneBank No.X14710 that corresponds to *LGB*-B allele. In addition, the location of the sequences relative to neighboring introns and exons were also deduced. The sequences representing specific haplotypes of the polymorphic sites were then submitted online to GeneBank and unique accession numbers were assigned for each sequence.

Multiple alignments of the sequences were performed using the program ClustalW2 available at EMBL-EBI (http://www.ebi.ac.uk/Tools/clustalw2/). This program calculated the best match of the sequences by lining them up so that the identities, similarities and differences are easily noted. DNAdot (http://www.vivo.colostate.edu/molkit/mapper/index.html) was used to generate textual and graphical location of restriction or cleavage sites across the sequences. These virtual restriction maps were in turn used to look for endonucleases that can act on restriction sites encompassing the SNPs. Following this, those enzymes capable of discriminating between different haplotypes were screened and selected for future genotyping applications. The exact nucleotide sizes of the restriction digests were then deduced from this analysis.

*CSN*3-A and B were only considered for genetic frequency and diversity inferences of the populations studied. The remaining haplotypes were used to study the nature of molecular evolutionary divergence in *CSN*3-alleles. The *CSN*3 sequences representing the different bovine haplotypes were compared with the *CSN*3 (exon 4) sequences of *Bos taurus* (AY380228.1), *Bubalus bubalis* (AM900443.1), *Ovis aries* (AY237637) and *Capra hircus* (AY350425). Molecular evolutionary history and phylogenetic analyses of these different sequences was conducted using MEGA version 4 (Tamura *et al.*, 2007). *Bubalis bubalis* (domestic water buffalo) was used as an out-group in the across-species *CSN*3 molecular network. The evolutionary history was inferred using the Neighbor-Joining method (Saitou and Nei, 1987) where the bootstrap test of phylogeny from 1000 replicates (Felsenstein, 1985) was taken to represent the consensus tree. The evolutionary distances which are in units of number of base substitutions per site were computed using the Maximum Composite Likelihood method (Tamura *et al.*, 2007).

4. Statistical analysis

4.1.Genotypic and allelic frequencies

To determine allelic frequencies, the number of copies of a particular allele was counted and divided by the number of copies of all alleles at a specific locus (*CSN3* or *LGB*) in the population. If there are n_i copies of the i^{th} allele in the sample, the sample frequency, p_i, of the i^{th} allele is:

$$p_i = n_i/2n$$

Where p_i - frequency of i^{th} allele, n_i -number of copies of the i^{th} allele in the sample, n- total number of sampled individuals

To calculate genotypic frequencies of both *CSN*3 and *LGB* genetic variants, direct counting was used. The frequency of a specific genotype was obtained by dividing the number of individuals possessing that genotype (n) to the total population sampled (N) i.e. n/N.

4.2.Computations on measures of genetic variability

Several standard population genetic parameters were analyzed to measure genetic diversity of the two milk protein genes under consideration i.e. *CSN*3 and *LGB*. The following parameters were estimated at two levels: within and among populations/ breeds. Genetic diversity within populations was considered as a function of total genetic diversity within the species and the proportion of the total genetic diversity found among its populations.

4.2.1. Heterozygosity/diversity estimators [A_e, H_o, H_e]

The effective number of allele [A_e]

The effective number of allele [A_e] measures the number of alleles at a locus, and the evenness of allele frequencies at that locus. A_e was calculated as the reciprocal of the sum of squared allele frequencies:

$$\mathbf{A_e = 1/\sum p_i^2}$$

$$\mathbf{= 1/1\text{-}H_e,\ _{1<Ae<\infty}}$$

Observed heterozygosity [H_o]

Observed heterozygosity [H_o] was calculated directly from observed genotype frequencies for each locus and population/breed.

$$\mathbf{H_o = \sum n_i/N}$$

Where n_i is the number of heterozygous individuals at the ith locus; N is the total number of individuals analyzed.

Expected heterozygosity [H_e]

Expected heterozygosity [H_e] is the expected proportion of heterozygous loci per individual. It is a composite measure that summarizes genetic variation at the allele level. The formulae for calculating H_e is derived right from the HWE. H_e, which is often referred to as genetic diversity is calculated as (1-Homozygosity) *i.e.*

$$\mathbf{H_e = 1\text{-}\sum_{i=1}^{n} p_{ik}^2}$$

Where P_i is the frequency of allele i at locus k; n is the number of alleles at locus k

4.2.2. Nei's diversity index [H_T, H_S D_{ST}, G_{ST}]

In order to describe the genetic structure of the population/breeds studied and to quantify the genetic diversity, estimates were made of the Nei's diversity index (Nei, 1973). For this purpose, expected heterozygosity [H_e] was partitioned into within-population and among-population components. The total genetic diversity at a locus ($H_T = 1-\sum p_i^2$, where p_i is the mean frequency of the i^{th} allele in the pooled sample) is obtained by summing the mean genetic diversity within populations (H_S) and the among-population genetic diversity (D_{ST}).

$$\mathbf{H_T = H_S + D_{ST}}$$

Nei's (1973) statistic, G_{ST}, was used to measure among-population differentiation, relative to the total diversity:

$$\mathbf{G_{ST} = D_{ST}/H_T, \qquad 0 \leq G_{ST} \leq 1}$$

$$\mathbf{G_{ST} = 1-H_S/H_T}$$

4.2.3. Wright's F-statistics [F_{IS}, F_{IT}, F_{ST}]

Three F-statistics developed by Wright (1978) to summarize the genetic structure of a population and its subpopulations were used. The F-statistics (F_{IS}, F_{IT} and F_{ST}) as measures of the fixation of alleles at different levels of organization (individuals, subpopulation, and population) were defined as follows:

$$\mathbf{F_{IS} = H_S-H_o/H_S, \qquad -1 \leq F_{IS} \leq 1}$$

Where F_{IS} is the mean reduction in heterozygosity of an individual due to non-random mating within a sub-population.

$$\mathbf{F_{ST} = H_T-H_S/H_T, \qquad 0 \leq F_{ST} \leq 1}$$

Where $\mathbf{F_{ST}}$ is the mean reduction in heterozygosity of a sub-population (relative to the total population) due to genetic drift among sub-populations.

$$\mathbf{F_{IT} = H_T-H_o/H_T, \qquad -1 \leq F_{IT} \leq 1}$$

Where F_{IT} is the proportional deviation of observed from expected heterozygosity in the pooled population.

4.2.4. Nei’s genetic distance (D)

Both Nei’s (1972) original and Nei’s (1978) unbiased genetic distances (D) were used to quantify the difference or distance between populations at *CSN*3 and *LGB* loci. Primarily, the normalized genetic identity (I) for a given locus was computed as:

$$I = J_{xy} / \sqrt{JxxJyy} \qquad 0 \leq I \leq 1$$

The genetic distance was then obtained by taking the negative of the natural log of the genetic identity (I):

$$D = -\ln(I), \qquad 0 \leq D \leq \infty$$

4.3. Hardy-Weinberg equilibrium test (HWT)

The Pearson’s chi-square ($\chi2$) and Fisher’s exact tests were used to check whether the population is in Hardy-Weinberg equilibrium or not. ($\chi2$) was calculated using the observed genotype frequencies obtained from the data and the expected genotypic frequencies obtained using the Hardy-Weinberg principles based on the algorithm (Levene, 1947).

$$\chi2 = \sum \frac{(O_i - E_i)^2}{E_i}$$

O_i- Observed genotypic frequencies of genotype i

E_i-Expected genotypic frequencies based on Hardy-Weinberg proportions of genotype i

In these two tests the null hypothesis (H_o) was that genotype frequencies for a particular locus in a given population is in Hardy-Weinberg proportions and the alternate hypothesis (H_a) was that it is not in Hardy-Weinberg proportions. The tests were conducted at 5% significant level ($P < 0.05$).

H_o was rejected: $X^2_{cal} > X^2_{tab}$

H_o was not rejected: $X^2_{cal} < X^2_{tab}$

The Fisher’s exact test was used to test for fit to HWP whenever the expected number of observations in any of the genotype classes was found to be less than five since Chi-square test may give less accurate results for very small observations.

Computations for the estimation of all the above genetic parameters were carried out using the statistical programs PopGene32 Version 1.31 (Yeh *et al*., 2000) and TFPGA (Tools for Population Genetics Analysis version 1.3) (Miller, 1997). TFPGA was primarily employed to undergo exact test for fit to Hardy-Weinberg’s equilibrium and for generating dendrograms based on Nei’s (1972, 1978) genetic distances between the breeds/ types investigated.

IV. RESULTS

The DNA extraction protocol used (Bruford *et al*., 1992) yielded a reasonable quality and quantity of DNA with the majority of the samples having an optical density ratio (OD_{260}/OD_{280}) ranging 1.6-2.0. Upon gel electrophoresis on 0.6% agarose gel, intact and sharp high molecular weight bands suitable for PCR-RFLP analysis were observed. Moreover, gene Quant Pro DNA analyzer showed that the DNA concentration of all the samples ranged from 150-220 ng/µl.

PCR amplification of *CSN*3 and *LGB* loci using the respective primers revealed clear and intact PCR fragments (Fig. 9 and Fig. 10). The primer pairs used for *CSN*3 product amplification produced a 633 bp fragment while that for *LGB* exhibited a 529 bp segment.

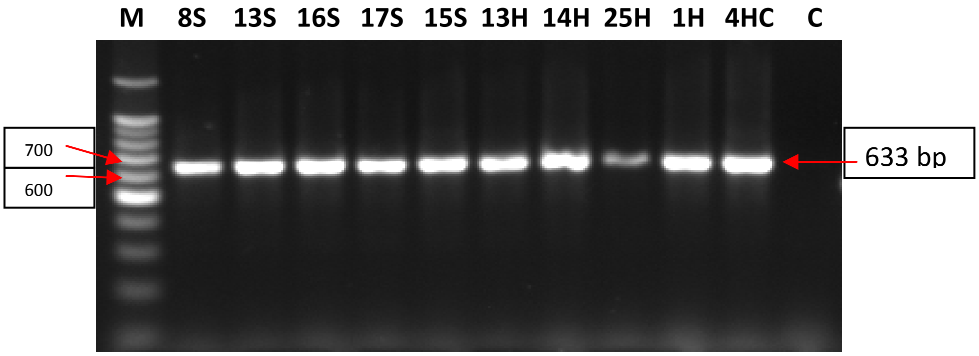


Figure 9. Electrophretogram of the 633 bp *CSN*3 fragments generated by PCR amplification of genomic DNA using *CSN*3 specific primers. Lane M is a 100 bp DNA marker. Lanes from 8S-15S are those taken from Sheko breed, lanes 13H-1H from Horro and lane 4HC from Holstein-Barka. Lane C is the negative control. Eight µl of the amplicons were loaded on each lane.

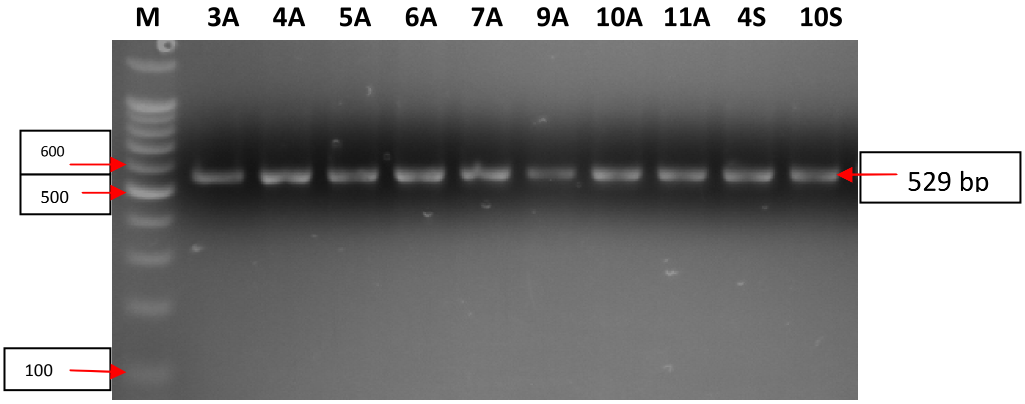


Figure 10. Electrophretogram of the 529 bp *LGB* fragments generated by PCR amplification of genomic DNA using *LGB* specific primers. Lane M is a 100 bp DNA marker. Lanes from 3A-11A are those taken from Abigar while 4S and 10S are from Sheko.

4.1. PCR-RFLP assay for *CSN*3

The restriction digestion analysis of the 633 bp *CSN*3 product using the endonucleases *Hinf*I and *Hind*III both indicated the presence of three types of restriction patterns.

Restriction digestion of the *CSN*3 product with enzyme *Hinf*I allowed us to identify three different band patterns/genotypes: AA (326, 131, 84 and 69 bp), AB (457, 326, 131, 84 and 69 bp) and BB (457, 84 and 69 bp). All these bands were clearly visible on 2% agarose gels except a 23 bp band which could not be detected since very small (Fig. 11). In addition, the restriction digestion of *CSN*3 with *Hind*III endonuclease has also revealed the same three distinct genotypes with the following fragment sizes: AA (uncut 633 bp), AB (633, 416, and 217 bp) and BB (416 and 217 bp) (Fig. 12). Genotype scoring conducted using both enzymes was in complete agreement and enabled us to discriminate between the *CSN*3 variants A and B.

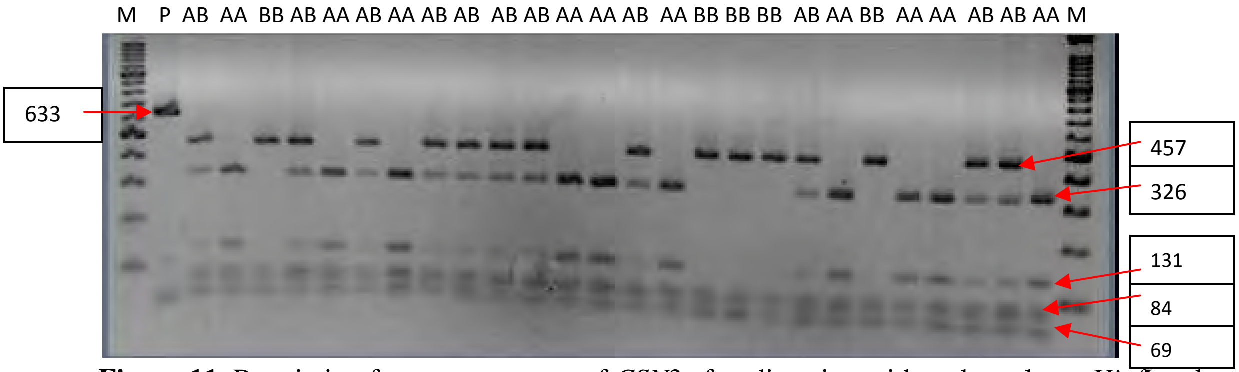


Figure 11. Restriction fragment patterns of *CSN*3 after digesting with endonuclease *Hinf*I and running on 2% agarose gel. The corresponding genotypes are shown at the top of each lane. Lane M is a 100 bp DNA ladder.

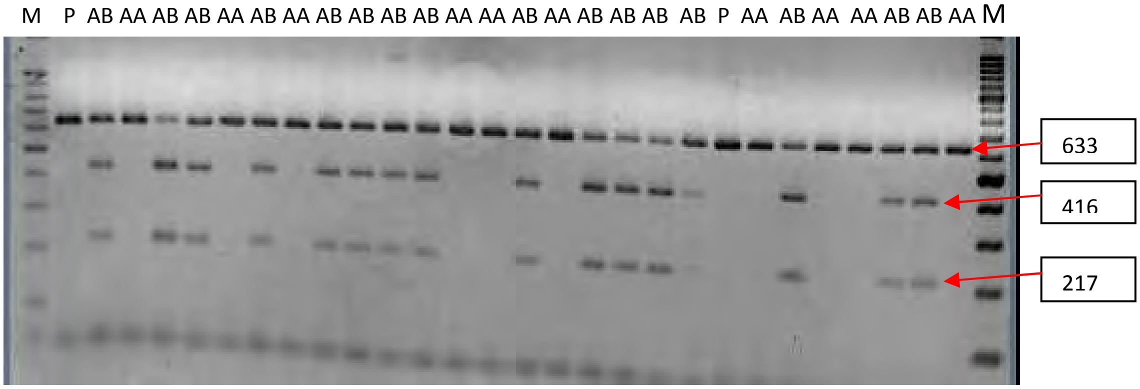


Figure 12. Restriction fragment patterns of *CSN*3 after digesting with endonuclease *Hind*III and running on 2% agarose gel. The corresponding genotypes are depicted at the top of each lane. Lane M is a 100 bp DNA ladder.

The genotypic frequencies of these three variants varied across the breeds examined. The variants *CSN*3-AA and AB were more predominant compared to variant BB in all the bovine breeds. The *CSN*3-BB variant was less frequent in most of the breeds and was even absent in individual animals investigated from the Abigar and Holstein-Barka breeds. The *CSN*3-BB was relatively more frequent in Sheko (0.25) followed by Guraghe (0.182), Boran (0.091) and Horro (0.044) breeds (Table 6 and Fig. 13). In Horro breed, the frequency of the AA and AB variants were found to be identical (0.4783). The overall genotypic frequencies of AA, AB and BB across all breeds were 0.469, 0.434 and 0.096 respectively (Table 5 and Fig. 13). These figures indicated that the BB variant was the least frequent whereas the AA and AB were nearly equally frequent and were the most common variants in the pooled cattle population.

As it can be inferred from the observed genotypes, the *CSN*3 marker was polymorphic and showed diallelic pattern (*i.e.* allele A and B). The frequencies of allele A and B showed a similar pattern in the majority of the breeds where allele A was found to be more frequent than allele B with the exception of Sheko breed which revealed nearly equal frequencies for both alleles *i.e.* A (0.531) and B (0.469) (Table 6). The highest frequency of the B allele (0.469) was found for the Sheko breed. The overall frequencies of allele A and B in the pooled population was 0.687 and 0.313 respectively indicating that allele A is the predominant one in the five local and one crossbred (HC) cattle populations investigated in the study (Table 6 and Fig. 13).

Table 5. Overall and across breed genotype frequencies of *CSN*3 genetic variants

Breed		*CSN*3	
	AA	AB	BB
Abigar (n=11)	0.5455	0.4545	0
Boran (n=11)	0.5455	0.3636	0.0909
Guraghe (n=11)	0.5455	0.2727	0.1818
Sheko (n=16)	0.3125	0.4375	0.25
Horro (n=23)	0.4783	0.4783	0.0435
HC (n=11)	0.4545	0.5455	0
Overall (83)	0.4699	0.4337	0.0964

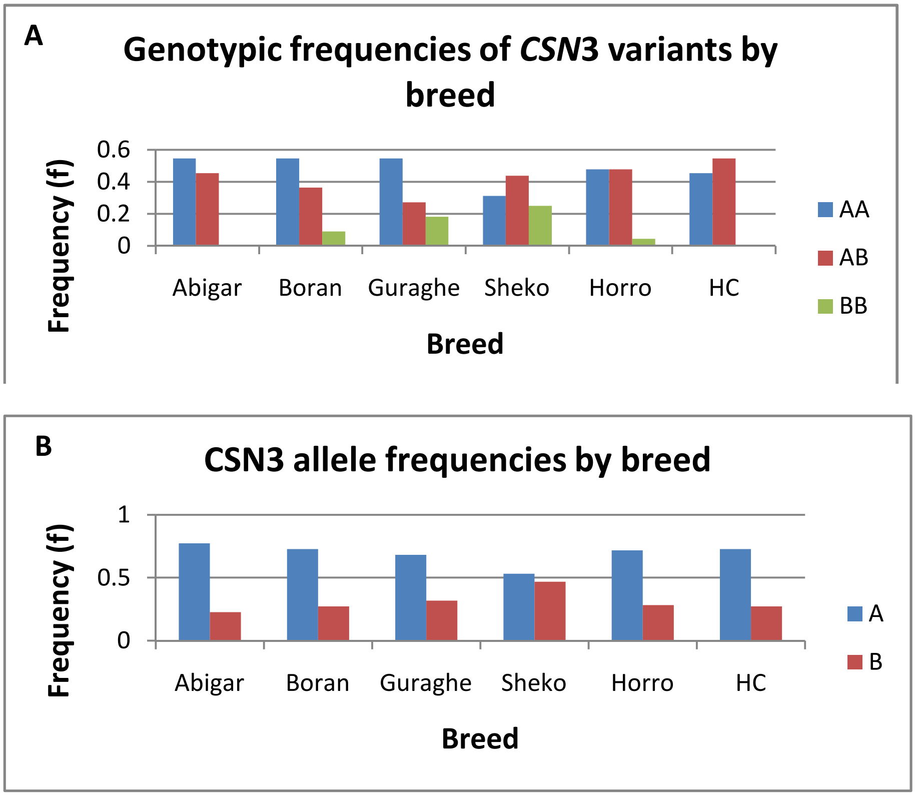


Figure 13. A) Genotypic frequencies of *CSN*3 variants by breed B) Allele frequencies of *CSN*3 variants by breed

To test genotypic frequencies for fit to Hardy-Weinberg's equilibrium (HWE), the Chi-square test was used which compares the observed and expected genotype counts. The *CSN*3 locus was found to be in HWE for all of the breeds examined with P-value ranging from 0.169 in Guraghe to 0.655 in Boran indicating that there were no significant deviations at 5% significance level (α=0.05) (Table 6). Out of a total of 83 animals genotyped, 39, 36 and 8 individuals belonged to AA, AB and BB variants respectively. The overall difference between the number of observed and that expected under HWE in each genotypic classes were very small resulting in a higher P-value (0.985) of the Chi-square test. This test indicated that the *CSN*3 locus is under HWE since the test revealed no significant deviation at 5% significance level and 1 degree of freedom.

Table 6. Allele frequencies and probability (P) of HWE tests (χ^2) for *CSN*3 locus across different cattle breeds

Locus	Allele	Abigar	Boran	Guraghe	Sheko	Horro	HC	Overall
CSN3	A	0.7727	0.7273	0.6818	0.5312	0.7174	0.7273	0.6867
	B	0.2273	0.2727	0.3182	0.4688	0.2826	0.2727	0.3133
P-value		***0.391173***	***0.654721***	***0.168584***	***0.538031***	***0.445635***	***0.263552***	***0.985167***^{ns}

*Statistically significant at α=5% since 0.01<P<0.05 [H_o is rejected]

** Highly significant at α=5% since P<0.01[H_o is rejected]

ns – Not significant at α=5% since P>0.1[H_o is not rejected]

4.2.Nucleotide sequence analysis of the *CSN*3 locus

After PCR-RFLP assays, a total of 10 PCR products which were genotyped as homozygous *CSN*3-AA and *CSN*3-BB were selected and sequenced in both forward and reverse direction from each breed with the exception of Abigar and HC which both manifested *CSN*3-AA variant only. Only homozygous genotypes were considered to avoid the possibility of ambiguous base calling during sequencing. Forward and reverse strand sequences were carefully inspected and crosschecked to come up with a single consensus sequence for each individual fragment.

For all of our *CSN*3 sequences, the BLAST search conducted using the reference genomic sequences database resulted in 99% match with the *Bos taurus* (Hereford breed) chromosome 6, Btau_4.2, whole genome shotgun sequence (Accession No. NC_007304.4). Moreover, BLAST search performed using the nucleotide collection (nr/nt) database produced close to 60 sequences with significant alignments to our query *CSN*3 sequences. Among these sequences the highest matches of identity scores (99%) were recorded with GeneBank accession No. AY380228.1 and AY380229 corresponding to bovine *CSN*3-A and *CSN*3-B alleles respectively. Similarly, EF378700 of *Bos taurus* and GU441771 of *Bos grunniens CSN*3 gene-exon 4 and partial cds had also high blast hits (99%).

To detect polymorphic sites/SNPs, all of our *CSN*3 sequences were aligned and compared with the reference sequence AY380228.1 corresponding with *CSN*3-A allele. The *CSN*3 fragment sequenced was confirmed to stretch from part of intron III (4 bp), the whole of exon IV (516 bp) and part of intron IV (113 bp). A total of 7 substitutions were identified within the amplified 633 bp *CSN*3 region where the majority (6) of which concentrated at coding exon IV sequence and only one was found at intron IV. All of the variations were point mutations corresponding to base transitions in 5 of the SNPs (T and G were substituted by C and A respectively) and 2 base transversions where C and T were substituted by A. Analysis of the deduced amino acid sequences showed that two of the SNPs were silent/synonymous base substitutions corresponding to amino acid residues proline (CC**G**→CC**A**) and alanine (GC**G**→GC**A**) at AA/nucleotide positions 150/13111 and 168/13165 respectively. The other four substitutions residing on the coding sequence (exon 4) were non-synonymous substitutions producing codon changes of isoleucine (A**T**C) to threonine (A**C**C) (at positions 135/13065 and 136/13068), alanine (G**C**T) to Aspartate (G**A**T) (at position 148/13104) and glycine (**G**GC) to serine (**A**GC) (at position 13124/155). All of the polymorphic sites were homozygous in all of the sequences analyzed except for a sequence derived from Abigar revealing heterozygous SNPs at all of the seven sites.

4.3.Molecular characterization of Kappa-casein (*CSN*3) haplotypes/variants

Out of 10 animals investigated through sequencing, a total of four distinct haplotypes were inferred from the 6 SNP sites analyzed in the *CSN*3 coding sequence (exon 4). The nucleotide sequences of these different *CSN*3-variants were submitted to GeneBank and deposited under the accession numbers (HQ589914, HQ589915, HQ589917, and HQ589916). Table 8 summarizes the detected *CSN*3 haplotype alleles (A^1, B, H and E) at the variable sites. The naming of *CSN3* haplotype alleles was based on the standard nomenclature of bovine milk proteins suggested recently by Caroli *et al*. (2009).

Table 7. Nucleotide and amino acid positions of detected polymorphisms in bovine *CSN*3 gene and predicted protein respectively

PCR-RFLP genotypes	Haplotypes	Hit	*CSN*3 Nucleotide/amino acid polymorphism positions [1]						
			13065/135	13068/136	13104/148	13111/150	13124/155	13165/168	13173
Abigar (BB)	BH	Q	ATC/**Ile**	ATC/**Ile**	GCT/**Ala**			*GCG/Ala*	T
		S	ACC/**Thr**	ACC/**Thr**	GAT/**Asp**			*GCA/Ala*	A
Boran 1 (AA)	A^1A^1	Q				CCG/Pro			
		S				CCA/Pro			
Boran 2 (BB)	BB	Q		ATC/**Ile**	GCT/**Ala**			*GCG/Ala*	T
		S		ACC/**Thr**	GAT/**Asp**			*GCA/Ala*	A
Guraghe1 (AA)	A^1A^1	Q				CCG/Pro			
		S				CCA/Pro			
Guraghe 2 (BB)	BB	Q		ATC/**Ile**	GCT/**Ala**			*GCG/Ala*	T
		S		ACC/**Thr**	GAT/**Asp**			*GCA/Ala*	A
Sheko 1 (AA)	HH	Q	ATC/**Ile**						
		S	ACC/**Thr**						
Sheko 2 (BB)	BB	Q		ATC/**Ile**	GCT/**Ala**			*GCG/Ala*	T
		S		ACC/**Thr**	GAT/**Asp**			*GCA/Ala*	A
Horro 1 (BB)	BB	Q		ATC/**Ile**	GCT/**Ala**			*GCG/Ala*	T
		S		ACC/**Thr**	GAT/**Asp**			*GCA/Ala*	A
Horro 2 (AA)	HH	Q	ATC/**Ile**						
		S	ACC/**Thr**						
HC (AA)	EE	Q					GGC/Gly		
		S					AGC/Ser		

[1] Polymorphic positions as per reference sequence GeneBank No. **AY380228.1**, corresponding to *CSN*3-A allele

- Q- Query sequence and S-reference sequence **AY380228.1**
- Synonymous SNPs are italicized while non-synonymous SNPs with their respective amino acids are indicated with bold letters

The sequences studied revealed no new mutations from those previously listed in *CSN*3 variants by Caroli *et al*. (2009). However, interestingly, variant *CSN*3-A[1] (Prinzenberg and Erhardt, 1998) which has been reported earlier in our study as *CSN*3-A using PCR-RFLP assay was detected in two sequences derived from Boran and Guraghe breeds. The difference between *CSN*3-A[1]and *CSN*3-A involved a single silent mutation (G→A transition) in the third codon position for proline (CC**G**→CC**A**) at AA/nt position 150/13111 (Table 7). In addition, two more sequences from Sheko and Horro breeds which were previously PCR-RFLP genotyped as *CSN*3-A were found to be *CSN*3-H (Prinzenberg *et al*., 1999). The *CSN*3-H variant deviated from that of *CSN*3-A by a Thr (A**C**C)→Ile (A**T**C) substitution at position 135/13065 (Table 8). Moreover, one sequence derived from a Holstein-Barka crossbred cow (previously identified as *CSN*3-A) revealed another variant called *CSN*3-E (Erhardt, 1989). The difference between *CSN*3-E and

*CSN*3-A involved a Gly(**G**GC)→Ser(**A**GC) substitution at position 155/13124 (Table 9). The rest of the sequences which were from Boran, Guraghe, Sheko and Horro previously genotyped *CSN*3-B using PCR-RFLP matched perfectly with CSN3-B sequence (AY380229) that varies from *CSN*3-A at three sites corresponding with two non-synonymous mutations (Ile136Thr and Ala148Asp) and one silent substitution 168(GC**G**→GC**A**) coding for Alanine (Table 7). Interestingly, we have also found one more transversion (T→A) at position 13173 located in intron IV when comparing *CSN*3-B and A indicating a possible linkage disequilibrium of this substitution with the other 3 SNPs residing in the coding region (exon 4).

Table 8. Bovine *CSN*3 haplotypes detected. Nucleotides present at polymorphic positions and corresponding amino acid positions (in parenthesis) in each variant are indicated. Nucleotides positions are after the reference sequence AY380228.1 (*CSN*3-A). The accession number for the sequences corresponding to *CSN*3 haplotypes A^{I}, B, E and H are HQ589914, HQ589915, HQ589917, and HQ589916 respectively.

Nucleotide position	Protein position	Haplotypes/Variants				
		A	A^{I}	B	E	H
13065	135	C (Thr)	C	C	C	**T (Ile)**
13068	136	C (Thr)	C	**T (Ile)**	C	C
13104	148	A (Asp)	A	**C (Ala)**	A	A
13111	150	A (Pro)	***G (Pro)***	A	A	A
13124	155	A (Ser)	A	A	**G (Gly)**	A
13165	168	A (Ala)	A	***G (Ala)***	A	A

After the sequence data analysis of the 10 *CSN*3 fragments, it became apparent that the endonucleases used in the PCR-RFLP assay failed to discriminate any of the 3 additional *CSN*3 haplotypes (A^{1}, H and E). The existence of other *CSN*3 haplotypes listed in the casein genotype nomenclature (Caroli et al., 2009) could not be examined since the primers used only covered a portion of the *CSN*3 coding sequence (exon 4 and immediate adjacent regions).

Another approach applied was analysis of different combinations of currently available restriction enzymes that are capable of detecting those SNP sites through virtual PCR-RFLP assays using sequence information. Such endonucleases were screened and selected through virtual restriction digestion of *CSN*3 fragments and subsequent generation of restriction maps using DNAdot software (New England Biolabs enzyme database). The patterns of the resulting fragments were used to discriminate between the *CSN*3 haplotypes. The presence or absence of a *BspM*II enzyme restriction site at nt/aa position 13111/150 enabled us to discriminate between the haplotypes *CSN*3-A (633 bp uncut) and A^1 (423 and 210 bp). Moreover, the endonucleases *Apa* I and *Hae* III were found to differentiate between *CSN*3-A and E through the presence or absence of their respective restriction sites covering the SNP at nt/aa position 13124/155. Digestion of *CSN*3 fragments using *Apa* I resulted in an uncut 633 bp segment in *CSN*3-A while it revealed two fragments of sizes 442 and 191 in *CSN*3-E haplotype. Digestion with *Hae* III resulted in two fragments (295 and 338 bp) in *CSN*3-A while it produced three fragments of sizes 295, 145 and 193 bp in haplotype *CSN*3-E. Virtual restriction maps of these digests are schematically depicted in Appendix 3. Unfortunately, endonucleases that are capable of discriminating *CSN*3-A and H haplotypes appear to be missing through this virtual restriction mapping analysis.

4.4.Phylogenetic relationship of the bovine *CSN*3 variants/haplotypes

The possible molecular evolution of the different *CSN*3 haplotypes was estimated through phylogentic analysis (MEGA4). A dendrogram (Fig. 14) was generated based on sequence comparison of the 520 bp long sequences of bovine *CSN*3 (exon 4) variants and the publicly available sequences (exon 4) of the *Bubalus* (AM900443)*, Caprine* (AY350425) and *Ovis* (AY237637) genera.

The dendrogram produced showed the very early departure of a common evolutionary line containing *Ovis aries* and *Capra hircus* and their subsequent split later into separate lineages which is supported by a high bootstrap value (100%). The second main branch contained the *Bubalus bubalis* and *Bos taurus CSN*3 sequences which subsequently diverged where *Bos taurus* further evolved into its different molecular forms/variants. In the process of the bovine *CSN*3 evolution, the B allele deviated from the rest earlier with high bootstrap value (100%) possibly making it the ancestral *CSN*3 haplotype. Eventually, the third node further evolved with bootstrap value of 81% into the *CSN*3-H allele and another group of *CSN*3 sequences containing the rest of the variants (*CSN*3-A, E and A^I) which appear to have very short evolutionary distances yet had fairly higher bootstrap values. Furthermore, the dendrogram shows CSN3-A^I

and E evolving shortly from their common ancestor, CSN3-A allele, with bootstrap value of 62%.

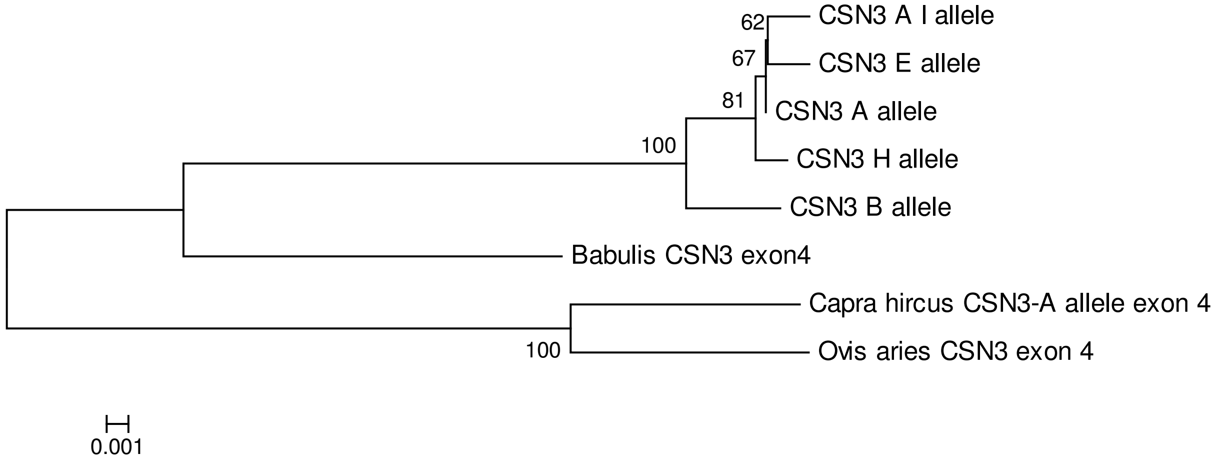


Figure 14. Dendrogram of the different bovine *CSN*3 sequences representing distinct haplotypes. Exon 4 of the *CSN*3 gene from *Bubalus bubalis* (AM900443*)*, *Capra hircus* (AY350425), and *Ovis aries* (AY237637) were also included in the analysis. Bootstrap values are expressed as percentage of times each node appeared in 1000 bootstrap resamplings.

4.5.PCR-RFLP assay for *LGB*

The PCR amplified *LGB* product was observed as a 529 bp fragment when run on a 2% agarose gel. The restriction digestion analysis of this product was conducted with endonucleases *Hae*III and *Dde*I where only *Hae*III revealed meaningful restriction patterns/ genotypes.

Digestion of the 529 bp *LGB* product with *Hae*III uncovered three restriction patterns/genotypes: AA (211, 181 and137 bp), AB (211, 181, 137, 125 and 86) and BB (181, 137, 125 and 86) (Fig. 15). All of these bands were clearly detected when run on 2% agarose gel. However, digestion of the *LGB* product with the enzyme *Dde*I failed to produce restriction patterns capable of discriminating between the different *LGB* genotypes. All of the *Dde*I digests resulted in similar band patterns (386, 93 and 50 bp) revealing no polymorphism in any of the breeds studied (Fig. 16).

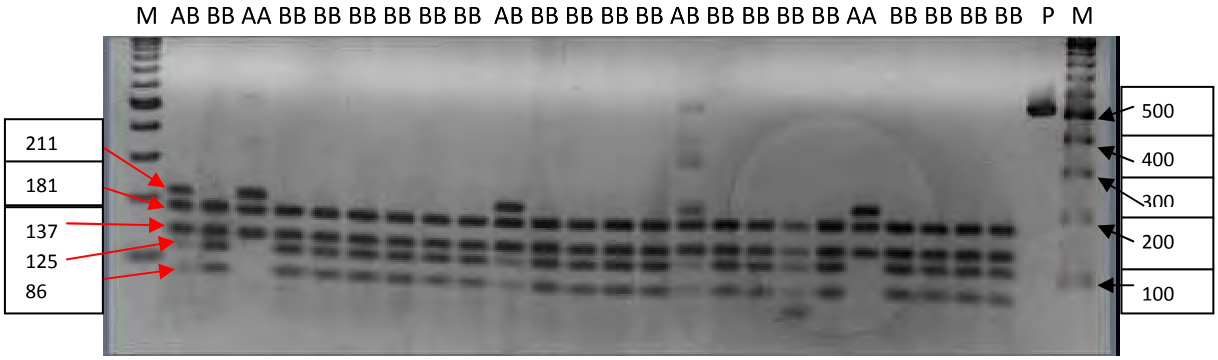


Figure 15. Restriction fragment patterns of *LGB* PCR product after digesting with *Hae*III and running on 2% agarose gel. The corresponding genotypes are shown at the top of each lane. Lane M is a 100 bp DNA marker.

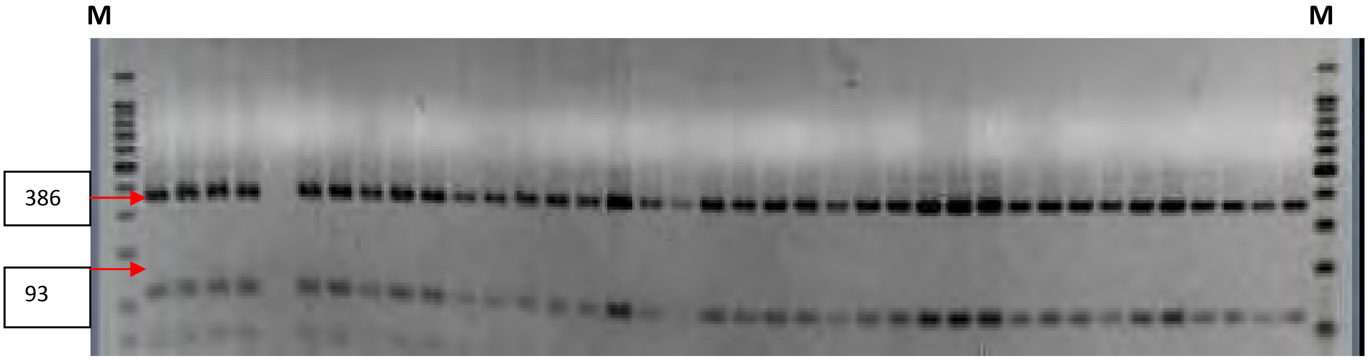


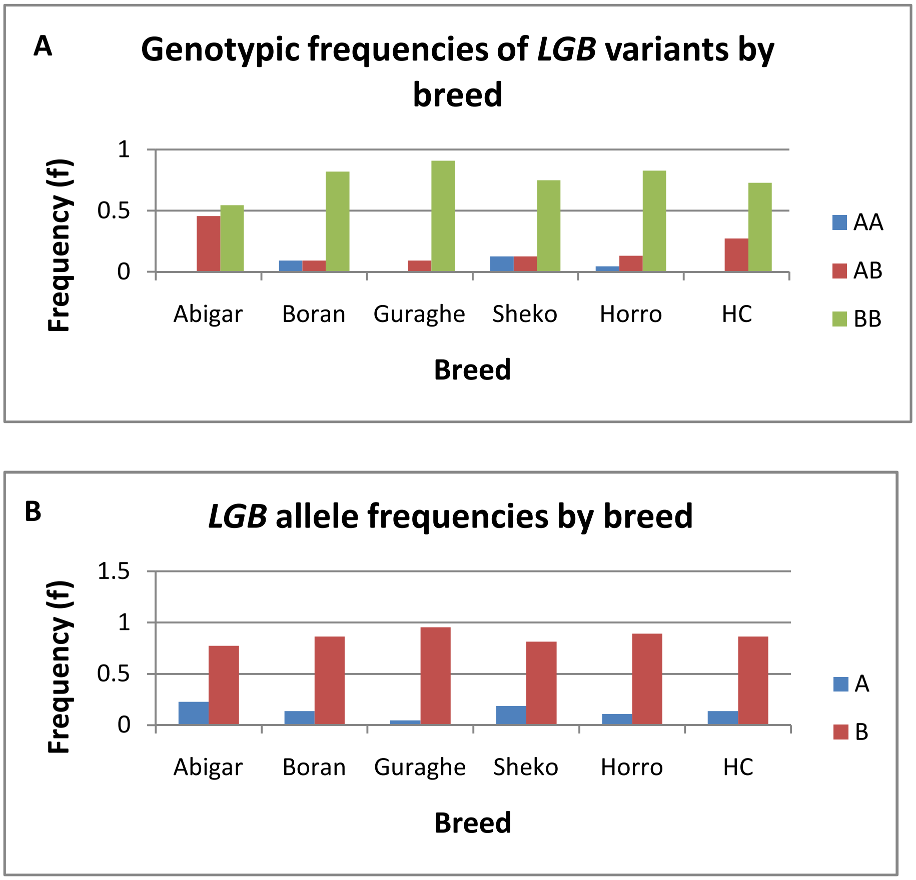
Figure 16. Restriction fragment pattern of *LGB* PCR products after digesting with *Dde*I and running on 2% agarose gel. Lane M is a 100 bp DNA marker.

The genotypic frequency of the *LGB* variants varied among the different breeds studied. In the majority of the breeds the BB variant was consistently predominant compared to both AB and AA. The AA genotype had very low frequency in most of the breeds and was even absent in Abigar, Guraghe and HC. Similarly, the AB variant was found to be less common than BB in most breeds except in Abigar which had nearly equally frequent AB and BB variants (0.454, 0.546 respectively). The highest frequency of the BB variant was observed in Guraghe (0.909) followed by Horro (0.826), Boran (0.818), Sheko (0.75), HC (0.727) and Abigar (0.546) (Table 9 and Fig. 17). The average genotypic frequency of the variants AA, AB and BB in the pooled population was 0.048, 0.180 and 0.771 respectively (Table 9).

The frequency of *LGB* A and B alleles showed a similar pattern across the breeds where B was found to be highly frequent than A. The frequency of B ranged from 0.955 in Guraghe to 0.773 in Abigar. The highest frequency of allele A was 0.227 for Abigar whereas the least was in Guraghe (0.046). Gene frequencies in the whole population were obtained by pooling the observed gene frequencies from each breed/subpopulation. Generally, allele B is the most predominant allele in the pooled population with an average allelic frequency of 0.861 (Table 10 and Fig. 17).

Table 9. Overall and across breed genotype frequencies of LGB genetic variants

Breed/Type	*LGB*		
	AA	AB	BB
Abigar (n=11)	0	0.4545	0.5455
Boran (n=11)	0.0909	0.0909	0.8182
Guraghe (n=11)	0	0.0909	0.9091
Sheko (n=16)	0.125	0.125	0.75
Horro (n=23)	0.0435	0.1304	0.8261
HC (n=11)	0	0.2727	0.7273
Overall (n=83)	0.0482	0.1807	0.7711



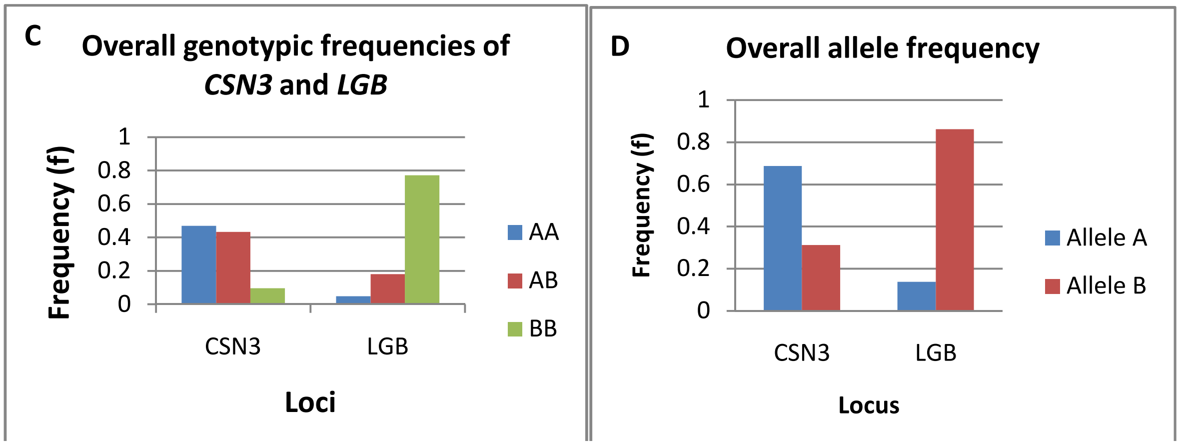


Figure 17. A) Genotypic frequencies of LGB variants by breed. B) Allele frequencies of *LGB* by breed. C) Overall genotypic frequencies of *CSN*3 and *LGB* D) Overall allele frequencies of *CSN*3 and *LGB*

The Chi-square test for Hardy-Weinberg's equilibrium at the *LGB* locus indicated that most of the breeds studied were under Hardy-Weinberg's equilibrium with the exception of Boran and Sheko breeds which presented significant deviations (P<0.05) with P-value of 0.012 and 0.009 respectively. The rest of the breeds showed no significant deviations from HWE at 5% significance level (α=0.05) (Table 9). From the total of 83 animals genotyped, 4, 15 and 64 individuals belonged to AA, AB and BB variants respectively (Table 9). The difference between the observed and the expected individuals in all genotype classes were relatively larger (Fig. 18). These deviations have significantly contributed to the very low P-value (0.021*) observed at the overall *LGB* locus (Table 11). Therefore, in the overall population, the *LGB* locus significantly deviated from the Hardy-Weinberg's equilibrium at 5% significance level (α=0.05) and 1 degree of freedom.

Table 10. Allele frequencies and probability (P) of HWE tests (χ^2) for *LGB* locus across different cattle breeds

Locus	Allele	Abigar	Boran	Guraghe	Sheko	Horro	HC	Overall
LGB	A	0.2273	0.1364	0.0455	0.1875	0.1087	0.1364	0.1386
	B	0.7727	0.8636	0.9545	0.8125	0.8913	0.8636	0.8614
P-value		***0.391173***	***0.011967****	***1.000000***	***0.009132*****	***0.069848***	***0.675323***	***0.021473****

*Statistically significant at α=5% since 0.01<P<0.05 [H_o is rejected]

** Highly significant at α=5% since P<0.01[H_o is rejected]

ns – Not significant at α=5% since P>0.1[H_o is not rejected]

The Fisher's exact test of HWE is more appropriate whenever the expected number of observations in any genotype class are too small (<5) which is the case at *LGB* locus i.e. 4 individuals belonged to genotype class AA. The exact test conducted indicated that all of the breeds examined were at HWE with the exception of Sheko which showed significant deviations at (α=0.05) with exact-p of 0.049*. In the overall population, similar to the chi-square test, the exact test showed that the *LGB* locus significantly deviated from HWE with an exact-p of 0.042* at 5% significance level.

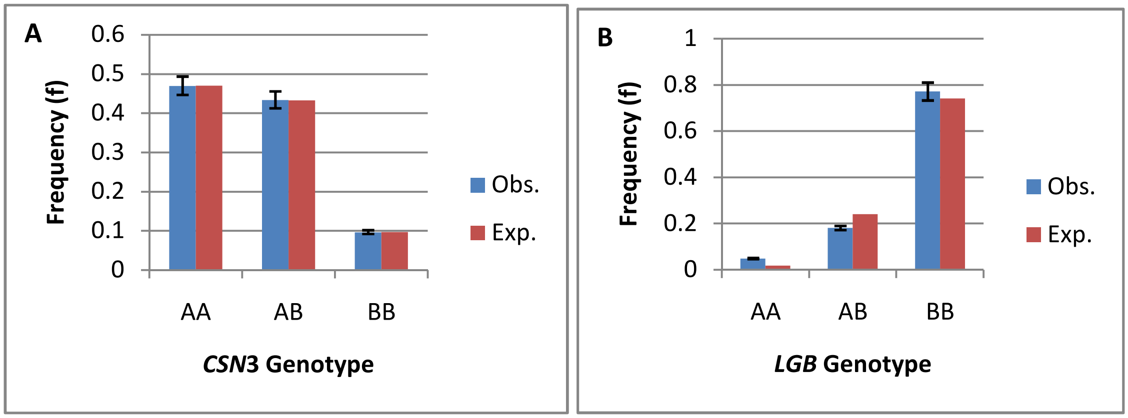


Figure 18. A) Overall genotype proportions at the *CSN*3 locus B) Overall genotype proportions at the *LGB* locus. Blue columns are observed proportions, with 5% error bars; red bars are expected proportions under HWE

In addition, the frequencies of *CSN*3-*LGB* genotype combinations were also computed in the pooled population. All of the 9 possible genotype combinations of *CSN*3 and *LGB* genes were observed in the total population sampled (83). The most predominant genotype combination was AABB (43.4%) followed by ABBB (28.9%) and ABAB (12%). The least frequent genotypes were ABAA and BBAA being observed in only one individual each (1.2%) of the breeds Sheko and Horro respectively (Fig. 19)

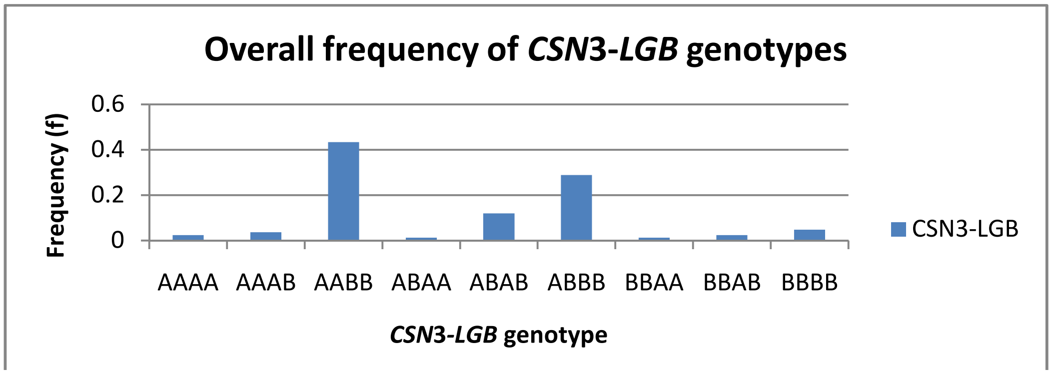


Figure 19. Overall frequency of the combinations of the CSN3-LGB genotypes

4.6.Nucleotide sequence analysis of the *LGB* locus

Following PCR-RFLP analysis, a total of 9 PCR products genotyped as homozygous *LGB*-AA (3) and *LGB*-BB (6) were selected and sequenced from each breed with the exception of Abigar, Guraghe and HC which manifested only one type of homozygous (*LGB*-BB) variant.

BLAST search of our *LGB* sequences conducted using the reference genomic sequence database revealed 99% match with the *Bos taurus* (Hereford breed) chromosome 11, Btau_4.2, whole genome shotgun sequence (NC-007309.4). In addition, BLAST search conducted employing the nucleotide collection (nr/nt) database resulted in close to 35 sequence hits with significant alignments with our query *LGB* sequences. Among these, the largest sequence identity of 99% was observed with three hits: X14710.1, DQ489319.1 and Z48305 corresponding with *Bos taurus LGB* gene, *LGB*-B variant and *LGB* gene B variant precursor respectively. The next two BLAST hits were *Capra hircus LGB* encoding gene (Z33881.1) and *Ovis aries* (X12817.1) *LGB* gene with identity scores of 91 and 89% respectively.

To look for polymorphisms, all of our *LGB* sequences were aligned and compared with the sequence X14710.1 which corresponds with *LGB*-B allele. The *LGB* fragment sequenced was confirmed to cover a part of intron III (41 bp), the whole exon IV (111 bp) and part of intron IV (377 bp). A total of 12 substitutions were detected in the amplified 529 bp *LGB* region where the majority of which (9) were located in the intronic regions and only 3 of them in the coding exon 4 region of *LGB* gene. All of the variations were single nucleotide substitutions corresponding to base transitions in 9 of the SNPs (T and A substituted by C and G respectively) and three base transversions (A and C substituted by T and G) (Table 12). Translation analysis of nucleotide sequences in the coding *LGB* region revealed that two of the SNPs were silent base substitutions corresponding to amino acid residues asparagine (AAT→AAC) and threonine (ACT→ACC) at positions 88/5174 and 97/5201 respectively and one SNP represented a non-synonymous mutation producing amino acid change of valine (GTC) to alanine (GCC) at 118/5263 (Table 11). Some of the samples had heterozygous SNPs at positions (5162, 5174, 5201, 5339, 5498 and 5530) (Table 11).

4.7. Molecular characterization of *LGB* haplotypes/variants

From 9 *LGB* sequences analyzed, a total of 4 unique haplotypes were inferred considering the 3 SNP sites in the coding *LGB* exon 4 sequence. The nucleotide sequences of these haplotypes together with their appropriate annotations were submitted to GeneBank and deposited in NCBI database with the following accession numbers (HQ589923, HQ589924 and HQ877769) (Table 12). The naming of *LGB* haplotype alleles was based on the standard nomenclature of bovine milk proteins suggested recently by Caroli *et al.* (2009). Two of the sequences from Boran and Horro breeds revealed a new haplotype allele which deviated from *LGB*-B allele through a single silent mutation for amino acid residue asparagine (AAT→AAC) at position 88/5174 (Table 11). Since this new haplotype is a close relative of *LGB*-B, it has been named *LGB*-B^1 allele (Table 12). More interestingly, a sequence derived from Abigar uncovered one new heterozygous (T/C) mutation site located at the third base of the codon for threonine (ACT→ACC) (97/5201) hence representing a silent substitution (Table 11 and Table 13). This was confirmed with the characterstics double chromatogram peaks of a new heterozygous SNP shown in Fig. 20. This new haplotype allele was also accompanied by three more heterozygous SNPs (A/G, C/G, and A/T) located at *LGB* intron IV region and were unique to this particular sequence (Table 11). This enabled us to suggest a possible new haplotype designated as *LGB*-B* allele since very similar to the normal B-allele in the rest of the SNP sites. These new haplotype alleles were previously genotyped as *LGB*-B allele after the PCR-RFLP assays. The remaining six sequences where genotyped either *LGB*-A or B and were in perfect concordance with the once reported by the PCR-RFLP method. The difference between *LGB*-A and B involves one silent mutation for amino acid residue asparagine (AAC→AAT) at 88/5174 and one non-synonymous mutation changing amino acid valine (GTC) to alanine (GCC) at 118/5263 (Table 11 Table 13). The sequence data analysis revealed two new haplotypes (*LGB*-B^1 and B*) which the PCR-RFLP assay failed to discriminate. Another interesting analysis was that two SNPs in intron III (T→C and A→T) at positions 5149 and 5162 respectively appear to be in linkage disequilibrium with the rest of the SNPs in the coding region distinguishing *LGB*-A from *LGB*-B (Table 11).

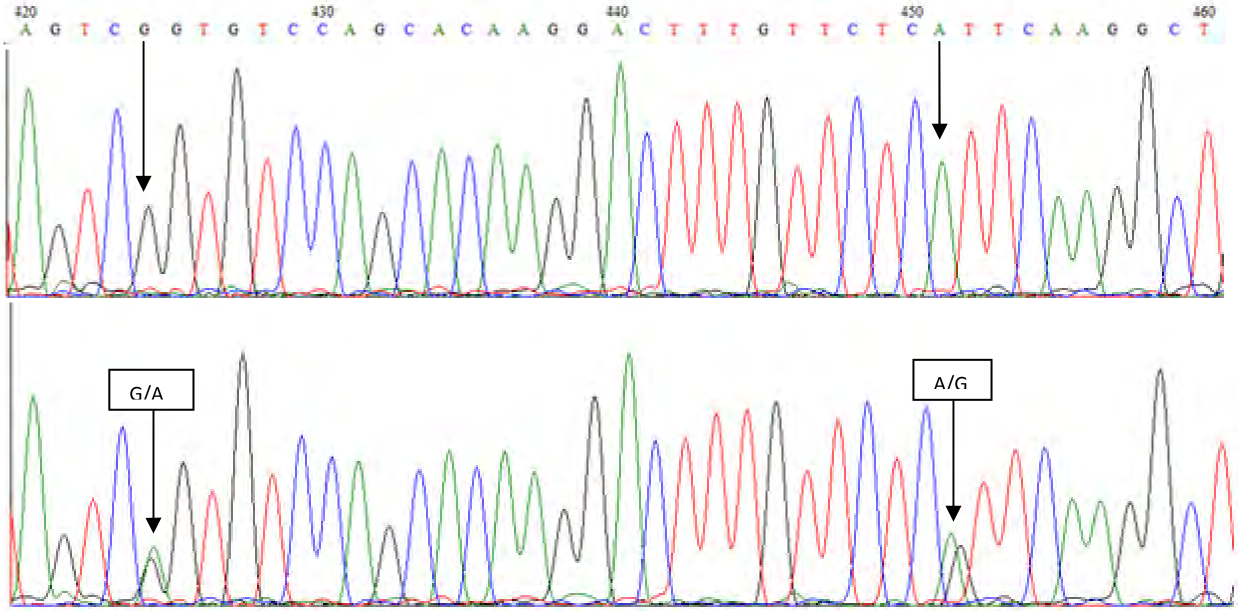


Figure 20. Portion of *LGB* sequence electropherograms showing the difference between the newly detected *LGB*-B* and the normal *LGB*-B alleles at positions 5174/88 and 5201/97. The pointing arrows indicate the heterozygous SNP sites which appear as two nearly equal peaks of different colors.

Table 11. Nucleotide and amino acid positions of detected polymorphisms in bovine *LGB* gene and the predicted protein respectively

PCR-RFLP genotypes	Haplotype	Hit	*LGB* Nucleotide/amino acid polymorphism positions ^{2}											
			5149	5162	5174/88	5201/97	5263/118	5339	5437	5448	5450	5498	5509	5530
Abigar (BB)	B*	Q			*AAC/Asn*	*AC**T**/Thr*		A/G				C/G		A/T
		S			*AAT/Asn*	*AC**C**/Thr*		G				G		T
Boran (AA)	A	Q	T	A	*AAC/Asn*		G**T**C/**Val**		A	A				
		S	C	T	*AAT/Asn*		G**C**C/**Ala**		G	G				
Boran (BB)	B^1	Q			*AAC/Asn*									
		S			*AAT/Asn*									
Guraghe (BB)	B	Q		A/T	*AAC/Asn*									
		S		T	*AAT/Asn*									
Sheko (BB)	B	Q		A/T	*AAC/Asn*									
		S		T	*AAT/Asn*									
Sheko (AA)	A	Q	T	A	*AAC/Asn*		G**T**C/**Val**		A	A				
		S	C	T	*AAT/Asn*		G**C**C/**Ala**		G	G				
Horro (AA)	A	Q	T	A	*AAC/Asn*		G**T**C/**Val**				T			
		S	C	T	*AAT/Asn*		G**C**C/**Ala**				C			
Horro (BB)	B^1	Q			*AAC/Asn*								A/G	
		S			*AAT/Asn*								G	
HC (BB)	B	Q	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩
		S	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩

[2] Polymorphic positions as per reference sequence Gene Bank No. **X14710.1** corresponding to *LGB* *B allele (Length = 7877 bp)

- Q- Query sequence and S-reference sequence of **X14710.1**
- The green filled row indicates a 100% match between the query sequence (9L) and the reference **X14710.1**
- Synonymous SNPs are italicized while non-synonymous SNPs with their respective amino acids are indicated with bold letters
- The blue filled boxes in the haplotype column show the newly detected LGB variants/haplotypes while the yellow filled boxes signifies heterozygous SNPs
- The first column indicate the samples sequenced with their respective PCR-RFLP genotypes

Table 12. Bovine *LGB* gene variants/haplotypes detected. Nucleotides present at polymorphic positions and corresponding amino acid positions (in parenthesis) in each variant are indicated. Positions of nucleotides are after the sequence X14710.1 corresponding to *LGB*-B allele. The Genebank accession numbers assigned for the sequences corresponding to the *LGB* haplotypes A, B, B^1, and B* are HQ589923, ____, HQ589924 and HQ877769 respectively.

Nucleotide position	Protein position	*LGB* Haplotypes/Variants			
		B	B^1	B*	A
5174	88	T (Asn)	T (Asn)	C (Asn)	C (Asn)
5201	97	C (Thr)	T (Thr)	C (Thr)	C (Thr)
5263	118	C (Ala)	C (Ala)	C (Ala)	T (Val)

4.8. Measures of Heterozygosity (H_o, H_e and A_e)

Effective number of alleles (A_e) was estimated for both loci as a measure of evenness of the most common alleles. A_e was found to be higher for *CSN*3 (1.7552) than *LGB* (1.3136) indicating a more even allele frequency at *CSN*3. Moreover, the mean A_e over both loci was found to be high (1.5344) hence the loci were regarded as highly informative (Fig. 21).

Observed (H_O) and expected (H_e) heterozygosity were calculated in order to estimate the level/ extent of genetic variability within the breeds and in the overall population. The mean H_o and H_e for the overall population both loci combined were 0.3097 and 0.3373 respectively. When the two loci were evaluated separately, the H_o and H_e for the *CSN*3 were 0.4253and 0.4304 and that for *LGB* locus were 0.1941 and 0.2442 respectively. These indicated a larger difference between H_o and H_e in *LGB* than *CSN*3 which had nearly equal estimates for these parameters (Table 13). The comparison of mean H_o and H_e across the different breeds revealed that most had higher H_e than H_O for the two loci considered. However, Abigar and HC had higher H_O than H_e whereas Horro had nearly equal ones. Generally, *LGB* locus presented low level of genetic variability (19.41%) when compared with *CSN*3 locus (42.52%) which was found to be highly variable (Table. 13).

Table 13. Observed hererozygosity (H_o) and expected heterozygosity (H_e) of *CSN*3 and *LGB* loci across breeds/types

Breed	Loci				Mean ± SD	
	*CSN*3		*LGB*		H_o	H_e
	H_o	H_e*	H_o	H_e*		
Abigar	0.4545	0.3680	0.4545	0.3680	0.4545±0	0.3680±0
Boran	0.3636	0.4156	0.0909	0.2468	0.2273±0.19	0.3312±0.12
Guraghe	0.2727	0.4545	0.0909	0.0909	0.1818±0.13	0.2727±0.26
Sheko	0.4375	0.5141	0.1250	0.3145	0.2812±0.22	0.4143±0.14
Horro	0.4783	0.4145	0.1304	0.1981	0.3043±0.25	0.3063±0.15
HC	0.5455	0.4156	0.2727	0.2468	0.4091±0.19	0.3312±0.12
Mean± SD	0.4253±0.09	0.4304±0.05	0.1941±0.14	0.2442±0.09	0.3097±0.16	0.3373±0.13

*Nei's (1978) unbiased heterozygosity

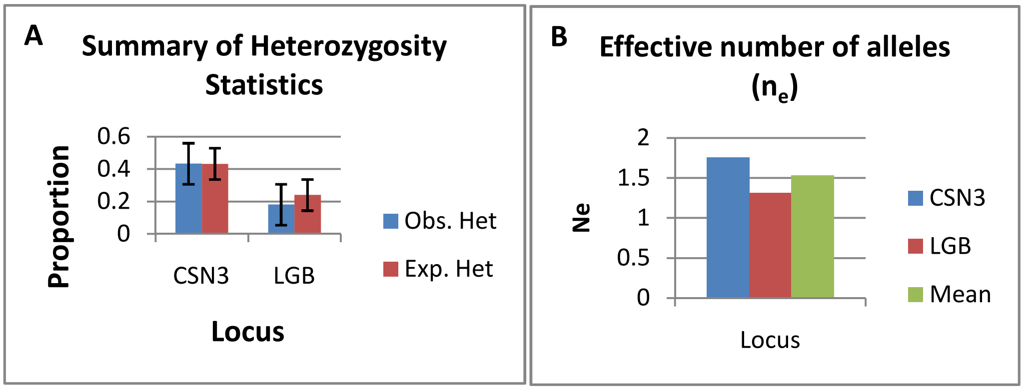


Figure 21. A) Overall H_o and H_e at *CSN*3 and *LGB* loci B) Effective number of alleles (ne) at *CSN*3 and *LGB* loci

4.9.Estimates of genetic diversity and fixation rates

I. Nei's diversity index [H_T, H_S, D_{ST}, G_{ST}]

Based on differences on observed allelic frequencies, Nei's (1973) statistics was calculated to quantify the level of genetic diversity and the distribution of total genetic variation (H_T) within (H_S) and between the breeds/populations (G_{ST}) for *CSN*3 and *LGB* loci. The mean H_T over both loci was 33.45% of which the vast majority of the total genetic variation was within the breeds (H_S) and was estimated to be 32.415% (Table 14). In other words, only 1.04% of the diversity was accounted for the among breed variability indicating very low level of differentiation between the six breeds. When the two loci were evaluated separately, *CSN*3 had a much higher total heterozygosity or diversity (43.03%), nearly twice that of the *LGB* locus (23.88%). For both loci, the within heterozygosity (H_S) was very high compared to the interbreed variation (G_{ST}) that hardly contributed to the total variation with 1.66 and 0.42% for *CSN*3 and *LGB* respectively (Table 14).

Table 14. Estimates of total (H_T), intrapopulational genetic diversity (H_S) and the coefficient of interpopulational genetic differentiation (G_{ST}) for *CSN*3 and *LGB* loci

Markers	H_T	H_S	G_{ST}
*CSN*3	0.4303	0.4137	0.0386
LGB	0.2388	0.2346	0.0176
Means	0.33455	0.32415	0.0281

II. Wright's F-statistics [F_{IS}, F_{IT}, F_{ST}]

In addition to Nei's statistics, Wright's (1978) F-statistics was also used to analyze the rates of fixation or coefficients of inbreeding. Calculation of these hierarchical statistics relies on the mean reduction in observed heterozygosity to quantify the level of genetic differentiation within and among breeds/populations. The total reduction in heterozgosity or fixation rate (F_{IT}) is divided into the within breeds (F_{IS}) and between breeds (F_{ST}) fixation index. Average F_{IS} (within-population inbreeding coefficient) as a measure of heterozygote deficiency or excess was estimated for both loci across the different breeds and tested for significance using Fisher's exact test (TFPGA software) (Miller, 1997).

At *CSN*3 locus, HC (-0.3750) presented the lowest F_{IS} followed by Abigar (-0.2941) and Horro (-0.1795) indicating an inclination towards excess heterozygosity. The rest of the breeds had positive F_{IS} with Guraghe (0.371) having the largest estimate followed by Sheko (0.122) and Boran (0.083) pointing towards deficiency in heterozygosity. However, none of these breeds presented significant deviation from HWE ($P>0.05$) with p-values ranging from 0.232 to 1.00. Similarly, the overall F_{IS} at CSN3 was not significant ($p>0.05$); hence no significant inbreeding was detected (Table. 15).

At *LGB* locus, Sheko presented significant deviation from HWE ($P<0.05$) with p-value of 0.049 where its F_{IS} estimate (0.589) indicated a relatively larger level of inbreeding (heterozygote deficiency). The rest of the breeds, however, revealed no statistically significant deviations with p-values ranging from 0.143 to 1.000. The highest F_{IS} estimate was that of Boran (0.614) whereas only Abigar (-0.294) and Guraghe (-0.048) showed negative F_{IS} values. The overall F_{IS} (0.243) at LGB in the pooled population revealed a higher level of inbreeding and significantly deviated from HWE ($P<0.05$) with an exact p-value of 0.042* (Table. 15).

The overall F-statistic (Nei, 1987) and the corresponding gene flow levels (N*m*) were also estimated at both loci. The F_{IS} value for *CSN*3 and *LGB* loci was found to be -0.028 and 0.173 respectively. These indicated that there is higher degree of heterozygosity at the *CSN*3 loci compared to *LGB* which exhibited heterozygote deficiency or higher level of inbreeding. Moreover, the mean F_{IS} at both loci was 0.044 revealing low level of within-population inbreeding (Table 17). The overall fixation index or total inbreeding estimate (F_{IT}) averaged over both loci was low (0.07) where almost the entire fixation (reduction in heterozygosity) was contributed by *LGB* (0.195) as compared to the slight share by *CSN*3 (0.0004). The mean F_{ST} (estimate of population differentiation) over both loci was very low (0.028) indicating little (statistically insignificant i.e. $p>0.05$) genetic differentiation between the different breeds studied. Similarly, the exact test for all of the F-statistics conducted using GENEPOP version 4.0 software (Rousset, 2008) did not show any significant ($P>0.05$) fixation or inbreeding at all levels of the population studied. In addition, the extent of gene flow (N*m*) estimated by $0.25(1 - F_{ST})/F_{ST}$ was high with an average value of 8.7744. A slightly higher N*m* was observed for *LGB* (8.878) when compared to *CSN*3 (8.718) (Table 16).

Table 15. Wright's (1978) fixation index (F_{IS}) as a measure of heterozygote deficiency or excess and the corresponding p-values for HWE exact test

Locus/Breed	Average F_{IS} across breeds/types						
	Abigar	Boran	Guraghe	Sheko	Horro	HC	Overall
*CSN*3	-0.2941	0.0833	0.3714	0.1216	-0.1795	-0.3750	-0.0081
Exact p-value	1.000	1.000	0.2322	0.6376	0.6259	0.5046	1.0000
LGB	-0.2941	0.6140	-0.0476	0.5897	0.3268	-0.1579	0.2429
Exact p-value	1.000	0.1429	1.000	**0.0488***	0.2145	1.000	**0.0423***

* Significant ($P<0.05$) using exact test of the HWE, TFPGA software (Miller, 1997).

Table 16. Summary of Nei's (1987) F-Statistics with corresponding p-values of exact test for population differentiation (Fisher's Method) and gene flow (Nm) for each locus

Locus	F_{IS}	F_{IT}	F_{ST}	Exact p-value	Nm
*CSN*3	-0.0282	0.0004	0.0279	0.279286 ^{ns}	8.7178
LGB	0.1727	0.1953	0.0274	0.541876 ^{ns}	8.8781
Overall	0.0435	0.0700	0.0277	0.437101 ^{ns}	8.7744

Nm = Gene flow estimated from $F_{ST} = 0.25(1 - F_{ST})/F_{ST}$.
ns = Non significant (P>0.05) using Fisher's exact test, GENEPOP version 4.0 software (Rousset, 2008).

Nei's (1978) unbiased measure of genetic identity (above diagonal) and genetic distance (below diagonal) between each breed were estimated on a pairwise basis (Table 17). Moreover, to visualize the genetic distance between the breeds at the *CSN*3 and *LGB* loci, two dendrograms were generated based on Nei's (1978) unbiased estimate of minimum genetic distance and Nei's (1972) original genetic distance using the Un-weighted Pair-Group Method with Arithmetic Mean (UPGMA) with TFPGA software version 1.3 (Fig. 22 and Fig. 23). Nei's (1978) unbiased estimate of genetic distance replaces all negative values of D by zero unlike Nei's (1972) original genetic distance D.

Based on Nei's (1978) measures of unbiased genetic distances, most of the pairwise distance (D) estimates revealed negative values indicating a lower level of differentiation/classification among the breeds. However, Sheko appeared to uniformly have positive distance estimates when paired with all the other breeds revealing a relatively remarkable deviation from the rest. In addition, Abigar had positive distances when paired with Guraghe and Horro indicating a relatively larger difference between the breed pairs. Since Nei (1978) considered negative distances as zero, the dendrogram/ classification tree generated using this approach grouped the breeds Horro, HC, Guraghe and Boran in a single node with no differentiation. The tree, however, depicted Sheko breed forming a distinct branch separate from the others whereas the second main branch included all the other breeds with Abigar branching out subsequently into a separate group with a distance of 0.0015 (Fig. 22).

Furthermore, Nei's (1972) original measure of genetic identity and distance was also computed and revealed high level of pairwise homogeneity at the two loci considered. However, Sheko was once again found to significantly deviate with relatively larger distances from the rest of the breeds. The dendrogram generated using this approach showed Sheko branching out first forming a separate group with a distance of 0.033. The second main branch contained the rest of the breeds with Abigar and Guraghe subsequently branching out to form separate groups with distances of 0.005 and 0.014 respectively. Horro demonstrated a very slight deviation from Boran and HC which had revealed no differentiation with an identity of 1.00 (D=0) (Table 18 and Fig. 23).

Table 17. Nei's (1978) unbiased measures of genetic identity and genetic distance

Breed/Type	Abigar	Boran	Guraghe	Sheko	Horro	HC
Abigar	****	1.0043	0.9825	0.9640	0.9967	1.0043
Boran	-0.0043	****	1.0030	0.9806	1.0072	1.0109
Guraghe	0.0176	-0.0030	****	0.9827	1.0032	1.0030
Sheko	0.0367	0.0196	0.0175	****	0.9790	0.9806
Horro	0.0033	-0.0072	-0.0032	0.0212	****	1.0072
HC	-0.0043	-0.0108	-0.0030	0.0196	-0.0072	****

Nei's genetic identity (above diagonal) and genetic distance (below diagonal)

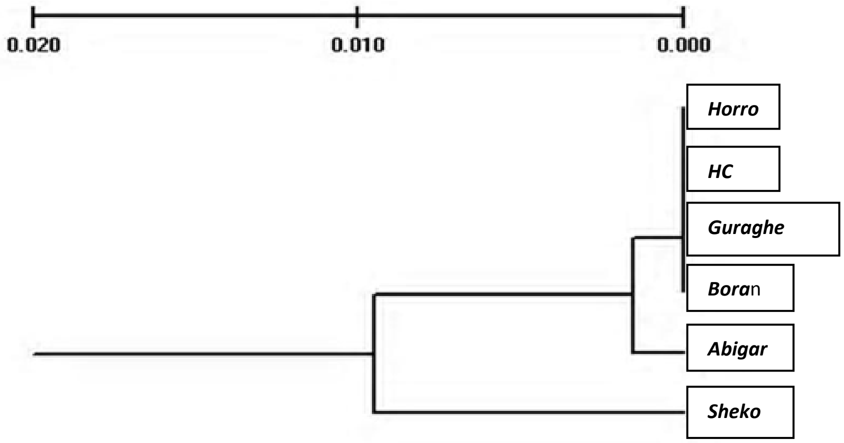


Figure 22. Dendrogram generated using UPGMA method for Nei's (1978) unbiased minimum genetic distance with TFPGA program version 1.3

Table 18. Nei's (1972) original measures of genetic identity and genetic distance

Breed/Type	Abigar	Boran	Guraghe	Sheko	Horro	HC
Abigar	****	0.9926	0.9723	0.9528	0.9880	0.9926
Boran	0.0074	****	0.9935	0.9701	0.9994	1.0000
Guraghe	0.0281	0.0065	****	0.9734	0.9967	0.9935
Sheko	0.0484	0.0303	0.0270	****	0.9715	0.9701
Horro	0.0120	0.0006	0.0033	0.0289	****	0.9994
HC	0.0074	0.0000	0.0065	0.0303	0.0006	****

Nei's genetic identity (above diagonal) and genetic distance (below diagonal)

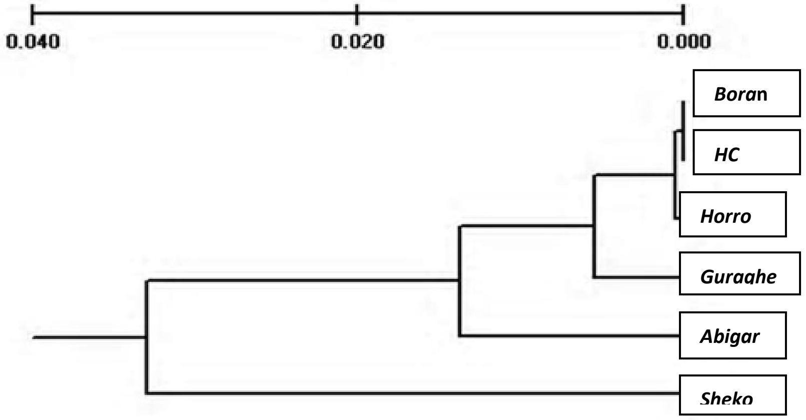


Figure 23. Dendrogram generated using UPGMA method for Nei's (1972) original genetic distance with TFPGA program version 1.3

In addition, the exact test for population differentiation was used to analyze the degree of genetic differentiation in the overall population using GENEPOP version 4.0 (Rousset, 2008). F_{ST} estimate for both loci did not show significant difference ($P>0.05$) in the overall populations examined with exact p-value of 0.279 for *CSN*3 and 0.542 for *LGB* (Table 19).

Pair-wise F_{ST} based on Weir and Cockerham (1984) and their corresponding p-values were also computed to test the level and significance of differentiation between each breed pair. The majority of the pair-wise F_{ST} estimates revealed negative values indicating very low level of genetic differentiation between the breed pairs. Positive F_{ST} were consistently found whenever Sheko was paired with each of the other breeds showing its relative departure from the rest. The pair- wise genetic differentiation of Sheko from each breed ranged from 0.23% with Boran to

9.34% with Guraghe. However, none of the pair-wise F_{ST} revealed significant differentiation between the breed pairs (exact-p > 0.05) and p-values ranged from 0.052 to 1.00 (Table 19).

Table 19. Pairwise F_{st}'s based on Weir and Cockerham (1984) and their corresponding p-values for exact test of Genic differentiation between each population pair over both loci (Fisher's method)

Population	Abigar	Boran	Guraghe	Sheko	Horro	HC
Abigar	****	0.948924	0.496142	0.244315	0.539676	0.949183
Boran	-0.0329	****	0.798079	0.374545	1.000000	1.000000
Guraghe	0.0179	-0.0307	****	0.052203	0.609313	0.798918
Sheko	0.0316	0.0023	0.0934	****	0.272572	0.381221
Horro	-0.0044	-0.0374	-0.0107	0.0266	****	1.000000
HC	-0.0195	-0.0497	-0.0134	0.0102	-0.0305	****

Pair wise Fst (below diagonal) and p-values for Fisher’s exact test of differentiation (above diagonal)

V. DISCUSSION

In the present study, genetic polymorphisms of two bovine milk protein genes namely *CSN*3 and *LGB* were analyzed in five indigenous and one Holstein-Barka crossbred cattle populations/ breeds using PCR-RFLP markers and sequencing. The polymorphisms were used to investigate the magnitude of genetic variability within and among these cattle populations at these two loci and further applied to study population structure and phylogenetic relationships. Moreover, results from sequencing of *CSN*3 and *LGB* PCR fragments were incorporated to confirm the PCR-RFLP assays, characterize additional haplotypes and suggest the most probable evolutionary history of the *CSN*3 alleles. To our knowledge, this study would be the first effort to characterize these two major milk protein genes in native Ethiopian cattle breeds hence would serve as a preliminary report for further detailed study.

In the following sections, the results are discussed focusing primarily on molecular characterization, gene and genotype frequency, level of heterozygosity, and genetic differentiation in the indigenous cattle populations investigated.

Molecular characterization of *CSN*3 and *LGB* loci

Sequence based and PCR-RFLP markers were used to analyze polymorphisms in the *CSN*3 and *LGB* genes. These markers were informative and able to discriminate and confirm the widely reported genetic variants of both genes. At both loci, PCR-RFLP analysis showed diallelic patterns revealing 3 genotype classes. Furthermore, sequencing of both loci revealed additional variants/haplotypes which were not detected using a conventional PCR-RFLP method.

*CSN*3

At *CSN*3 locus, the enzymes *Hind*III and *Hin*fI detected the widely reported *CSN*3-A and B genetic variants. The difference between these two variants (A→B) has been reported to arise from two linked nucleotide substitutions which lead to amino acid changes at positions 136 (Thr→Ile) and 148 (Asp→Ala) of the mature protein (Grosclaude *et al.*, 1972). As has been reported (Eggen *et al.*, 1989; Medrano and Aquilar-Cordova, 1990), the presence or absence of *Hind*III and *Hinf*I restriction sites at nucleotide/amino acid position 13104/148 (AY380228.1) revealed three different band patterns or genotypes when subjected to agarose gel electrophoresis *i.e.* AA, AB and BB. Therefore, either of these two enzymes could be used for genotyping these variants.

Sequencing of the *CSN*3 fragment uncovered three additional *CSN*3 variants namely: *CSN*3-A^I (Prinzenberg and Erhardt, 1998), *CSN*3-H (Prinzenberg *et al*., 1999) and *CSN3*-E (Erhardt, 1989). These haplotypes were deduced after analyzing mutations located at amino acid positions 150 (A^I), 135 (H) and 155 (E). These mutations were not detected through the PCR-RFLP assays conducted using the enzymes *Hind*III and *Hinf*I. This result, therefore, suggests that sequencing offers a superior advantage in distinguishing between the *CSN*3 haplotypes described so far. According to a recent review by Caroli *et al*. (2009) on the nomenclature of *CSN*3 variants; *CSN*3-A^I, E and H have been reported to be "rather less common" compared with the most frequent *CSN*3-A and B alleles.

On the other hand, virtual restriction mapping of the *CSN*3 sequences revealed endonuclease *Bsp*MII as being capable of discriminating between haplotypes *CSN*3-A and A^I while either of endonucleases *Apa*I or *Hae*III were found to differentiate between *CSN*3-A and E haplotypes. This suggested that these enzymes (*Bsp*MII and either *Apa*I or *Hae*III) could be used in combination with endonucleases *Hind*III and *Hinf*I for precise genotyping of *CSN*3 haplotypes (Appendix 3). Unfortunately, the distinction between *CSN*3-A and H haplotypes could not be detected by any of the currently available restriction enzymes (New England Biolabs database, DNAdot) making sequencing inevitable for discriminating these two haplotypes.

Another important result is the detection of an intronic SNP located at nucleotide position 13173 which appears to be in linkage disequilibrium with the 3 SNPs in the coding region that are known to distinguish *CSN*3-A from B. This SNP could be used as a potential *CSN*3 marker in the future.

Even more interesting result of this study was the detection of variants *CSN*3-A^I and H that were reported to be zebu-specific haplotypes (Caroli *et al*., 2009; Prinzenberg and Erhardt, 1998 and Farrell *et al*., 2004). These new zebu-specific haplotypes were found in 5 *i.e.* 50% of the representative sequences derived from the indigenous cattle sampled (Boran, Guraghe, Horro and Sheko). So far, these two *CSN*3 haplotypes have not been reported in pure taurine breeds (Caroli, 2009). The occurrence of such haplotypes in the indigenous populations indicated previous introgression of indicine genes into the original African taurine cattle which were thought to have intermingled and crossbred with zebu (*Bos indicus*) cattle to produce the initial foundation stock of the various types of cattle populations found in East Africa today (Pyne and Wilson, 1999). More importantly, the detection of *CSN*3-H haplotype in Sheko further confirmed the recently witnessed interbreeding of Sheko (African taurine) with thoracic-humped cattle such as Abigar around common borders of their habitats (DAGRIS, 2006). In addition, sequence analysis of Abigar revealed a heterozygous variant *CSN*3-BH perfectly agreeing with its previously suggested taurine and indicine background (Abigar is a Sanga being a cross of Zebu and taurine). Sanga in general and Abigar in particular were thought to be results of interbreeding between zebu and taurine cattle in Ethiopia (Epstein, 1971; Epstein and Masson, 1984 and Rege *et al.*, 2001).

A recent review by Caroli *et al.* (2009) has shown that the variant is rather less common and has only been detected so far in cattle breeds of taurine origin. In close agreement with this report, we have detected the *CSN*3-E haplotype in only one sequence derived from a Holstein-Barka cross bred cow.

Ikonen *et al.* (1997) has reported negative effects of *CSN*3-E allele on both milk production and milk quality parameters. Reports on the association of *CSN*3-A^{I} and H alleles with major milk yield and quality traits appear to be missing so far. However, the association of *CSN*3-A and B variants has long been established through numerous studies conducted in different cattle breeds mostly of taurine origin. The majority of such studies agree on the superior performance of *CSN*3-B allele over *CSN*3-A regarding fat, protein, and casein contents as well as desirable manufacturing properties of milk such as shorter rennet coagulation time, formation of a firmer curd and high cheese yield (Ng-Kwai Hang, 1986; Kubarsepp *et al.*, 2005; Matejicek *et al.*, 2008; Ng-Kwai-Hang *et al.*, 1998; Di Stasio and Mariani, 2000; Johann and Peter, 2000). Since the less favourable *CSN*3-A allele was more predominant in the studied population, future breed improvement programs could consider selecting for the superior *CSN*3-B allele to advance quality of milk and milk products. However, before any recommendation, more intensive studies should be conducted on the association of the most common *CSN*3 variants (yet to be confirmed in large scale sequencing studies of our indigenous cattle breeds) with major milk production and quality traits.

Molecular network analysis of the bovine *CSN*3 sequences representing the different variants identified showed early divergence of the *CSN*3-B allele making it phylogenetically older than the rest of the variants along the evolutionary line hence considered the ancestral haplotype. Moreover, the dendrogram indicated *CSN*3-H evolving next followed by the successive evolution of *CSN*3-A, E and A^I variants in a relatively shorter evolutionary time. All of the nodes separating the variants were supported by fairly high bootstrap values ranging from 62-100%. As expected, the out groups used in this analysis deviated early in the tree with *Bubalus bubalis* (Domestic water buffalo), a close relative to the *Bos* genera, relatively showing longer lineage than the more distantly related *Capra hircus* and *Ovis aries* (Domestic goat and sheep respectively). *Capra hircus* and *Ovies aries* are shown following an early and separate common lineage which later diverged with a very high bootstrap value of 100% (Fig. 14).

LGB

At *LGB* locus, the enzyme *Hae*III enabled us to differentiate between the two common and widely reported variants (A and B) whereas *Dde*I failed to reveal any meaningful polymorphism in any of the breeds investigated resulting in indifferent band patters in every individual sample. The difference between variants A and B arise from two base substitutios resulting in amino acid substitutions at positions 64 (Asp→Gly) and 118 (Val→Ala) (Eigel *et al*., 1984). The latter has been reported to create a *Hind*III restriction site (Medrano and Aguilar-Cordova, 1990) which can discriminate between the three different *LGB* genotypes (AA, AB and BB).

Moreover, sequencing analysis of *LGB* loci in Abigar, Boran and Horro breeds revealed two new *LGB* haplotypes that were not previously reported in earlier studies. These new variants are here now named *LGB*-B^1 and B* and were due to silent mutations at amino acid positions 88 and 97 respectively hence were considered to be iso-forms of the *LGB*-B allele. Although the PCR-RFLP assay failed to detect these isoforms, it precisely distinguished the widely reported *LGB*-A and B alleles.

Moreover, the detection of two intronic SNPs at positions 5149 and 5162 (X14710) that were in linkage disequilibrium with the rest of the SNPs in the coding region distinguishing *LGB*-A from *LGB*-B suggested their potential role as *LGB* markers for future applications. In addition, five other polymorphisms were detected in non-coding regions that are not in linkage disequilibrium with *LGB*-A and B variants and were located in introns III and IV at positions (5339, 5450, 5498, 5509 and 5530) (Table 12). The effects of these SNPs is yet to be thoroughly investigated since recent studies have shown some intronic mutations possibly affecting mRNA splice sites, consequently influencing their stability and eventually leading to truncated protein products or even lack of them (Caroli *et al*., 2009; Rando *et al*., 1996; Damiani *et al*., 2001 and Mohr *et al*., 1994). Recently, Ganai *et al*. (2008) detected eight polymorphisms in *LGB* gene which were not in complete linkage disequilibrium with *LGB*-A and B alleles where one of these SNPs had significantly affected *LGB* concentration. In addition, an SNP (C→A) located 215 bp upstream of translation initiation site for *LGB* gene was found to be associated with an abnormally reduced expression of *LGB* characterizing this variant (Braunschweig and Leeb, 2006).

Several studies on the association of the *LGB*-A and B alleles with milk casein and fat contents indicated the superior performance of the B allele over A while the reverse was found to be true regarding whey protein content (Ng-Kwai Hang, 1998 and Van der Berg *et al*., 1992).Therefore, seletion of the superior *LGB*-BB genotype could contribute its share in improving the manufacturing properties of milk. Generally, such molecular information could be incorporated in future breed improvement programs to supplement the existing convetional approaches. The effects of the new *LGB* haplotypes (B^I and B*) on milk composition and manufacturing properties is yet to be investigated in future studies.

Gene and genotype frequency and HWE

The allele and genotype frequencies of *CSN*3 variants deduced from the PCR-RFLP assay are summarized in (Table 7 and Fig. 13). The most predominant *CSN*3 genetic variants in the pooled population were AA (0.469) and AB (0.434) which together accounted for around 90% of the total individuals sampled. The BB variant had the lowest frequency and was observed in less than 10% of the population studied. This result is almost similar with that reported (Golijow *et* *al*., 1999) on Argentine Creole cattle belonging to Iberian breeds. Moreover, our figures are in close concordance with the results of earlier studies performed in *Bos taurus* (Ng-Kwai-Hang *et al*., 1983; Lin *et al*., 1986; Sabour *et al*., 1996) and *Bos indicus* (Rachagani and Gupta, 2007) indicating AA and AB as predominant variants compared to BB. When breeds were considered separately, BB was the least frequent in all of the breeds and was even absent in Abigar and HC breeds sampled in this study.

In the majority of the breeds, allele A was predominant over B with the exception of Sheko which presented almost equally frequent alleles. Sheko breed has also showed the highest frequency of *CSN*3-B allele (0.4688). This unique allelic distribution pattern of Sheko coincides with its suggested origin separating it from the rest of the breeds which are thought to have originated from crosses of the original African taurine and introduced zebu cattle from Asia. Sheko, however, is believed to be the original shorthorn taurine cattle of East Africa despite minor admixture of the modern day Sheko with zebu cattle due to uncontrolled interbreeding (Rege, 1999).

One unexpected result was the allele frequency of the Holstein-Barka (HC) crossbred cattle that had higher taurine blood level (75-93.75%). Allele frequency of HC was almost similar with the majority of the native breeds than with that of Sheko which is thought to have a taurine origin.
In the pooled population, the frequency of allele A (0.687) was more than twice that of B (0.313). This result is in close agreement with studies reported in breeds from both taurine and indicine origin. Lin *et al*., 1986 reported similar result for Holstein Friesian cows with frequencies of 0.688 and 0.312 for A and B alleles respectively. The predominance of allele A over B has also been reported in numerous studies; in Holstein dairy herds (Ng-Kwai-Hang *et al*., 1984), Argentine and Patagonian Creole cattle breeds (Liron *et al*., 2002), Canadian Holstein bulls (Sabour *et al*., 1996), Hungarian spotted and Grey cattle (Baranyi *et al*., 1993) and Dutch Black and White crossbred cows (Bovenhuis and van Arendonk, 1991).

On the contrary, very high frequency of *CSN*3-B allele has been reported in several Nordic cattle breeds where the largest of which was in Icelandic cattle (0.76) followed by Northern Finn and Swedish Mountain cattle (0.71) and Eastern Finncattle (0.57) (Lein *et al*., 1999). A fairly high frequency of the B allele (0.51) has also been reported in Brown Swiss cattle (Dogru and Ozdemir, 2009). Similarly, Liron *et al*. (2002) reported a highly frequent B allele with 0.645 and 0.574 in Bolivian Chaqueño and Yacumeño Creole cattle breeds respectively. This suggests that a clear distinction could not be estabilished between B*os taurus* and *Bos indicus* cattle breeds based on alleleic frequency information of these milk protein genes since each species does not seem to possess a characterstic genotype.

Since PCR-RFLP assay and sequencing identified some of the *CSN*3 variants, it is difficult to suggest that *CSN*3-A, thought to be the less favorable allele, is the most predominant one in the indigenous cattle studied. In the future, a more comprehensive and large scale studies would be appropriate to know the most common *CSN*3 alleles in the indigenous cattle breed and their possible effect on production and quality traits as well.

Usually farm animals are not Mendelian populations since several factors such as inbreeding, selection, breeding structure and the like can upset Hardy-Weinberg equilibrium. Inspite of such limitations, Chi-square test of *CSN*3 locus for fit to HWE did not reveal any significant deviation ($P>0.05$) in each of the breeds examined as well as in the pooled population. This result was further supported by Wright's F_{IS} (1978) as a measure of hetrozygote deficiency or excess where none of the breeds presented significant deviation (exact-$P>0.05$) from HWE. At *CSN*3 locus, it can be inferred that, no significant inbreeding or excess heterozygosity was witnessed at both breed and population levels.

In the pooled population, *LGB*-BB was the most common variant and was observed in more than 75% of the population followed by AB which accounted for around 18%. The AA genotype was the least frequent (only 5%) and was even absent in Abigar, Guraghe and HC samples. In all of the breeds studied, BB was the most predominant variant and was even approaching fixation in Guraghe with a frequency close to 91%.

The frequency of *LGB* alleles had similar pattern in all of the breeds where B was found to be the most predominant allele with very high frequency ranging from 95.45% in Guraghe to 77.27% in Abigar. Similarly, in the pooled population, B was the most prevalent allele with a frequency of

86.14%. These results are almost similar with the ones reported for; South Anatolian cattle in Turkey (Yardibi *et al.*, 2009), Estonian dairy cattle (Varv *et al.*, 2009), Aberdeen Angus cattle (Vasconcellos *et al.*, 2003), Ayrshire cattle (Lin *et al.*, 1986) and Saavedreño Creole cattle breed (Liron *et al.*, 2002). Moreover, the higher frequency of the B allele observed in our study was in close agreement to earlier works reported in *Bos taurus* (Ng-Kwai-Hang *et al.*, 1983; Lin *et al.*, 1985; Merlin and Distasio, 1982; Sabour *et al.*, 1996; Eenennaam and Medrano, 1991 and Aleandri *et al.*, 1990) and *Bos indicus* (Karimi *et al.*, 2009; Del Lama and Zago, 1996).

Conversely, a more frequent *LGB*-A allele has been reported in; Dutch Red and White dairy cattle (55.6%) (Bovenhuis and van Arendonk, 1991), Patagonian Creole cattle (84.5%) (Liron *et al.*, 2002), Czech Fleckvieh cattle breed (51%) (Kucerova *et al.*, 2006) and Brazilian Holstein cattle (51%) (Del Lama *et al.*, 1996).

Fisher's exact test of HWE, which was more appropriate for *LGB* locus, indicated that all breeds examined fulfilled HWE with the exception of Sheko that presented significant deviation ($P<0.05$). Moreover, in the pooled population a significant deviation ($P<0.05$) from this equilibrium was witnessed. These results were further supported by Wright's (1978) F_{IS} estimates which revealed very high degree of inbreeding (heterozygote deficiency) in Boran (0.614) and Sheko (0.589) breeds compared to the rest. In the overall population, there appears to be deficiency in heterozygosity mainly because of Sheko and Boran which contributed largely to this homogeneity at this locus.

Sheko cattle represent the only remnants of the original African *Bos taurus* cattle found in eastern Africa and has recently been designated as endangered with a population size of only 2,400 individuals (Rege, 1999). Therefore, it is likely that the significant level of inbreeding at *LGB* locus witnessed in Sheko could be possibly because of its reduced population size increasing the probability of mating between closely related individuals.

Little effort, if any, has been exerted to undertake systematic artificial selection in any of the indigenous Ethiopian cattle populations so far. Therefore, the failure of the pooled population to meet HWE is unlikely to be due to selection pressure. It seems more plausible that stochastic factors such as random genetic drift and group founder effects might have played a determinant role in the allelic distribution of *LGB* milk protein variants. These factors could have contributed to the prevalence of the homozygous BB variant in the majority of the individuals sampled which is close to 80%. Moreover, the limited number of sample taken from each breed might have also affected the allelic distribution since it was not exhaustively representative.

Since *CSN3* and *LGB* are located on separate chromosomes, the above mentioned factors do not put equal influence on each locus. This might explain the difference in fit to HWE witnessed in our study where the *CSN*3 locus was in equilibrium unlike *LGB*.

Genetic diversity, fixation rates and differentiation

In the pooled population, considering both loci simultaneously, the mean observed (H_o) and expected (H_e) heterozygosity estimates revealed a rather moderate level of genetic variability *i.e.* 0.3097 and 0.3373 respectively. The expected heterozygosity (H_e) across the different breeds ranged from 0.273 in Guraghe to 0.414 in Sheko. Using RAPD markers Hassen *et al.* (2007) reported a more or less similar gene diversity values ranging from 0.27 in Abigar to 0.35 in Sheko. A microsatellite analysis conducted (Dadi *et al.*, 2008) using 30 markers revealed very high gene diversity in 10 local cattle breeds with H_e ranging from 0.714 in Raya-Azebo to 0.735 in Sheko. Lower values of H_e are expected in biallelic markers (such as PCR-RFLP) compared with microsatellite markers which divulge multiallelic loci and are known to reveal hyper variability. Results of this study coincide with both reports in that Sheko consistently presented the highest H_e or gene diversity from the rest of the native Ethiopian cattle breeds. The relatively higher diversity in Sheko might be explained by its divergent taurine origin in its longer evolutionary line while the rest are known to share some common zebu background (Pyne and Wilson, 1999). The relatively higher diversity of Sheko signifies its potential role as a genetic reservoir to meet future conservation and production needs.

When the two loci were considered separately, *CSN*3 presented a higher level of genetic variability with H_o and H_e values of 42.53and 43.04% respectively. *LGB*, however, manifested very low genetic diversity with H_o and H_e of 19.41 and 24.42% respectively. This indicated that the reduced variability witnessed when both loci were considered simultaneously was greatly influenced by the less diverse *LGB* loci. A nearly similar level of heterozgosity (16 and 17.5% respectively) was reported by Karimi *et al.* (2009) for *LGB* locus in Iranian Najdi cattle (*Bos indicus*). Moreover, hetrozygosities at *LGB* locus, in close concordance with our values, have also been reported in Ayrshire cattle (Lin *et al.*, 1985), Guernsey cattle (Eenennaam and Medrano, 1991) and South Anatolian cattle (Yardibi *et al.*, 2009). The mean and overall H_o and H_e estimates per breed and per loci were summarized in Table 12. Generally, the majority of the breeds revealed a lower H_o than H_e with the exception of Abigar and HC which proved otherwise. Dadi *et al.* (2008) had similarly reported a lower H_o than H_e in 10 indigenous Ethiopian cattle populations including three of the breeds studied here.

Nei's (1973) diversity index was used to quantify the level of genetic diversity and evaluate the distribution of the total genetic variation (H_T) within (H_S) and between the breeds (G_{ST}) at the *CSN*3 and *LGB* loci. Considering both loci, the mean H_T was 33.45% indicating a moderate diversity level. The majority of the total genetic diversity was found to be within the breeds whereas only 1.04% was accounted for the interbreed variability revealing a very low level of breed differentiation. Marson *et al.* (2005) considered the *LHR* and *FSHR* loci and reported an elevated level of within breed variability (46.35%) compared with only 1.2% for the among breed variability in European-Zebu cattle.

Evaluating the two loci separately, *CSN*3 (43.03%) presented high genetic diversity which is almost twice that of *LGB* locus (23.88%).

The overall F-statistics (Nei's 1987) were also computed to estimate the total rates of fixation (F_{IT}) and rates of fixation (inbreeding) within breeds (F_{IS}) and between breeds (F_{ST}). A higher level of inbreeding was detected at *LGB* locus (F_{IS}=0.1727) compared with *CSN*3 which rather showed excess heterozygosity with F_{IS} of -0.0282. When considering both loci, the mean F_{IS} and F_{IT} estimates were 0.0435 and 0.07 respectively. The overall F_{ST} value (0.0277) in this report revealed little and insignificant ($P>0.05$) genetic differentiation among the populations studied. Agreeing with the low F_{ST} estimates, the corresponding gene flow levels (Nm) were very high (8.77). The F_{ST} estimate of this report was a bit higher than the one reported by Dadi *et al.* (2008) (0.013) using microsatellite markers in 10 indigenous cattle breeds. On the other hand, Hassen *et al.* (2007) reported a very large F_{ST} value of 0.119 using RAPD markers in 5 native Ethiopian cattle breeds. Studying *CSN*3 and *LGB* loci using PCR-RFLP markers, Varv *et al.* (2009) reported a fairly high F_{ST} of 0.066 in Estonian dairy cattle. Moreover, Golijow *et al.* (1999) reported an even larger and significant F_{ST} value of 0.071 in Argentine Creole cattle breeds using PCR-RFLP *CSN*3 marker.

F_{ST} values serve as estimates of population differentiation and in our study only 2.77% of the total genetic variation at these two loci was attributed to differences among breeds whereas the remaining 97.23% was accounted for variation within individuals. Generally, the F_{ST} and N_m estimates of this study revealed low level of population sub-structuring between the studied breeds. The most probable cause could be the high level of inter-population gene flow which is facilitated by long distance migrations within and across habitats, utilization of communal grazing lands and watering points as well as exchange of breeding animals. More importantly, the majority of present-day cattle populations in Ethiopia, with the exception of Sheko, have common historical origins and are thought to be founded as a consequence of previous admixture between *Bos indicus* and *Bos taurus* cattle manifesting a hybrid nature (Dadi *et al*., 2008). Moreover, traditionally practiced uncontrolled mating together with lack of breed improvement programmes might have also played their share in increasing gene flow among populations located in different production systems.

Pair-wise genetic distances computed based on Nei's (1972) original and (1978) unbiased measures of genetic distance and identity unanimously revealed very small distances among the breeds studied. An interesting result was, however, Sheko consistently showed the largest distances when paired with each breed where the maximum of which was with Abigar and the minimum that involving Guraghe indicating its relative departure from the rest. Similarly, the dendrogram generated clearly depicted Sheko forming a separate branch where the rest shared a second main branch probably because of their common zebu background. Interestingly, our phylogenetic analysis, using just two PCR-RFLP markers, exactly coincided with that reported by Dadi *et al*. (2008) and Hassen *et al*. (2007) which clearly revealed two groups of Ethiopian cattle where Sheko clustered separately from the other cattle populations consistently agreeing with their conjectured evolutionary history.

To test the level and significance of differentiation between each breed pair, pair-wise F_{ST} (Weir and Cockerham, 1984) were also computed. Agreeing with Nei's genetic distance estimates, majority of pair-wise F_{ST} estimates were negative indicating a very low yet insignificant ($P>0.05$) level of genetic differentiation between all of the breed pairs. However, a relatively larger and positive F_{ST} values were found when Sheko was paired with each of the other breeds confirming its relative departure once again. The pair-wise F_{ST} value of Sheko ranged from 0.23% with Boran to a nearly significant ($P=0.0522$) 9.34% with Guraghe. Similarly, Hassen *et al*. (2007) reported the largest genetic difference between Sheko and Guraghe studying 5 native cattle breeds with RAPD markers.

VI. CONCLUDING REMARKS AND FUTURE PROSPECTS

To our knowledge, this study is the first of its kind to deal with the molecular characterization of the two important milk protein genes *CSN*3 and *LGB* in the indigenous Ethiopian cattle populations including the currently endangered Sheko breed. Several literatures, however, has been produced so far dealing with these two loci mostly in cattle breeds of taurine (European) origin hence cannot be directly translated to our native cattle breeds. This study will play its part in aiding selection of superior animals with respect to traits of milk yield and quality. Moreover, it will provide information that could be used to monitor, conserve and manage the rich genetic diversity of our native cattle today. This work will serve as a preliminary report on the existing milk protein genetic resources of our local cattle breeds and pave the way for comprehensive and large scale studies that would serve as guidelines in devising future animal breeding strategies and policies.

Based on the findings, the following conclusions and relevant future research directions were suggested:

- Although the PCR-RFLP markers employed were informative in discriminating between the most widely reported *CSN*3 and *LGB* variants, sequencing of the PCR fragments revealed additional haplotypes which were not previously identified through restriction digest analysis. The three additional haplotypes of *CSN*3 (*CSN*3-A^I, H and E) were apparent after analyzing several SNP sites that were previously reported by different authors while those of *LGB* (*LGB*-B^1 and B*) were new haplotypes which have not been reported before. We, therefore, suggest sequencing of large number of samples representing each cattle breed to exhaustively check for existence of additional variants.
- The detection of the zebu-specific *CSN*3-A^I and *CSN*3-H haplotypes in almost all of the indigenous cattle breeds examined including the endangered Sheko confirmed the earlier suggestion of indicine gene flow into the African taurine cattle which were thought to crossbred with zebu (*Bos indicus*) producing the current cattle population in Ethiopia. It is more alarming that Sheko, thought to be the only remnant of the original African taurine cattle, is continuously being diluted through indiscriminate cross breeding with thoracic-humped cattle. Therefore, owing to such observations, we suggest that the Sheko

cattle should be the prime targets for future breed conservation programs through the most economically viable in-situ preservation of the pure live animals within their production environment.

- According to the PCR-RFLP genotyping, *CSN*3-A was more predominant than *CSN*3-B in the indigenous cattle populations examined. The predominance of the less favorable *CSN*3-A allele regarding fat and protein contents as well as desirable manufacturing properties of milk indicated that future breeding programs could consider strategies that select cattle with the superior *CSN*3-B allele. On the other hand, *LGB*-B was the most prevalent variant and has been reported by several authors to perform better than *CSN*3-A regarding fat percentage and cheese yield. Therefore, we suggest that the selection of sires genotyped BB-BB (*CSN*3-LGB) could be simultaneously considered to increase the quality of milk. The effect of the additional *CSN*3 and *LGB* variants is yet to be studied in the future with inclusion of more indigenous cattle breeds of Ethiopia.
- Since the different casein genes are tightly linked in a casein cluster, we suggest that future studies should focus on considering the whole casein complex (αs_1, αs_2, β and k) simultaneously where haplotype effects are studied rather than effects of single allele. Currently, this approach is being widely recommended for both research and breeding purposes since it sketches a wider picture suitable for MAS. Moreover, the possible role of several important polymorphisms in the non-coding regions (intronic SNPs) which are in linkage disequilibrium with known protein variants should be thoroughly investigated in the future.
- Regarding genetic diversity, considering both loci, the overall population revealed a rather moderate level of genetic variability (0.34). The highest variability was observed in Sheko (41.43) while the lowest was in Guraghe (0.2727). When the two loci were considered separately, *CSN*3 locus presented higher (43.04%) level of gene diversity compared with *LGB* locus (24.42%) which was rather more homogenous. The availability of such genetic variability would be of great relevance to meet future breed development needs in the country.
- The vast majority of the total genetic variation of the population was attributed to the within population variability where as little and insignificant level of differentiation (between breed variation) was witnessed among the different cattle breeds. This indicates

that highlighted risk of the current indigenous cattle population becoming genetically homogenous due to constant gene flow among the populations. We propose that effective and appropriate breed management practices be implemented to avert such threats.

VII. REFERENCES

1. Agarwal, M., Shirvastava, N. and Padh, H. (2008). Advances in molecular marker techniques and their applications in plant sciences. *Plant Cell Rep*. **27**: 617-631.
2. Aleandri, R., Butazzoni, L.G., Schneider, J.C., Caroli, A. and Davoli, R. (1990). The Effects of Milk Protein Polymorphisms on Milk Components and Cheese-Producing Ability. *J. Dairy. Sci*. **73**: 241-255.
3. Alexander, L.J., Stewart, A.F., Mackinlay, A.G., Kapelinskaya, T.V., Tkach, T.M., Gorodetsky, S.I. (1988). Isolation and characterization of the bovine kappa-casein gene. *Eur J Blochem.* **178**: 395-401.
4. Alexander, L.J., Hayes, G., Bawden, W., Stewart, A.F., and Mackinlay, A.G. (1993). Complete nucleotide sequence of bovine *β*-lactoglobulin gene. *Anim. Biotechnol*. 4:1–10.
5. Ali S and Clark AJ (1988). Characterization of the gene encoding ovine beta-lactoglobulin. *Journal of Molecular Biology.* **199**: 415-426.
6. Aschaffenburg, R. and Drewry, J. (1955). Occurrence of different beta-lactoglobulins in cow's milk. *Nature.* **176**: 218-219.
7. Baranyi, M., Bosze, Z.S., Buchberger, J. and Krause, I. (1993). Genetic Polymorphism of Milk Proteins in Hungarian Spotted and Hungarian Grey Cattle: A Possible New Genetic Variant of β-Lactoglobulin. *J. Dairy Sci*. **76**: 630-636.
8. Beckmann J.S., and Soller, M. (1987). Molecular markers in animal genetic improvement. *Bio/Technol*. 5: **573**–576.
9. Bell, K. (1962). One-dimensional gel electrophoresis of bovine skim milk. *Nature*. **195**: 705-706.
10. Bell, K., Hopper, K.E., and McKenzie, H.A. (1981). Bovine *α*-lactalbumin C and *α*s1-, *β*-, and *κ*-caseins of Bali (Banteng) cattle, *Bos* (Bibos) *javanicus*. *Aust. J. Biol. Sci*. **34**:149–159.
11. Beuzen, N.D., Stear, M.J. and Chang, K.C. (2000). Molecular markers and their use in animal breeding. *The veterinary Journal*. **160**: 42-52.
12. Bingham, E.W. (1979). Role of mammary casein kinase in the phosporylation of milk proteins. *Journal of Dairy Research*. **46**:181-185.

13. Bobe, G., Beita, D.C., Freeman, A.E. and Lindberg, G.L. (1999). Associations among individual proteins and fatty acids in bovine milk as determined by correlations and factor analysis. *J. Dairy Res*. **66**: 523-536.
14. Boettcher, P.J., Caroli, A., Stella, A., Chessa, S., Budelli, E., Canavesi, F., Ghiroldi, S., and Pagnacco, G. (2004). Effects of casein haplotypes on production traits in Italian Holstein and Brown cattle. *J. Dairy Sci*. **87**:4311–4317.
15. Botstein, D., White, R.L., Skolnick, M., Davis, R.W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet*. **32**: 314 331.
16. Bovenhuis, H. and Van Arendonk, J.A.M. (1991). Estimation of Milk Protein Gene Frequencies in Crossbred Cattle by Maximum Likelihood. *J. Dairy Sci*. **74**: 2728-2736.
17. Braunschweig, M.H. and Leeb, T. (2006). Aberrant low expression level of bovine *β-lactoglobulin* is associated with a C to A transversion in the BLG promoter region. *J. Dairy Sci.* **89**: 4414–4419.
18. Brophy, B., Smolenski, G., Wheeler, T., Wells, D., L'Huillier, P., and Laible, G. (2003). Cloned transgenic cattle produce milk with higher levels of β-casein and k-casein. *Nature Biotechnology.* **21**: 157-162.
19. Bruford, M.W., Hanotte, O., Brookfield, J.F.Y., Burke, T. (1992). Single-locus and multilocus DNA fingerprinting. In: Hoelzel AR (ed), Molecular Genetic Analysis of Populations: A practical Approach. 225-269.
20. Caroli, A.M., Chessa, A. and Erhardt, G.J. (2009). Invited review: Milk protein polymorphisms in cattle: Effect on animal breeding and human nutrition. *J. Dairy Sci.* **92**: 5335-5352.
21. Coll, A., Folch, J.M., and Sanchez, A. (1993). Nucleotide sequence of the goat kappa casein cDNA. *Journal of Animal Sciences.* **71**: 2833.
22. Dadi, H., Tibbo, M., Takahash, Y., Nomura, K., Hanada, H., Amano, T. (2008). Microsatellite analysis reveals high genetic diversity but low genetic structure in Ethiopian indigenous cattle populations. *Animal Genetics*. **39**: 425–431
23. DAGRIS (2007). Domestic Animal Genetic Resources Information System (DAGRIS). (*eds. S. Kemp, Y. Mamo, B. Asrat and T. Dessie*). International Livestock Research Institute, Addis Ababa, Ethiopia. http://dagris.ilri.cgiar.org.

24. Dalgleilsh, D.G., Spagnuolo, P.A. and Goff, H.D. (2004). A possible structure of the casein micelle based on high-resolution field-emission scanning electron microscopy. *International Dairy Journal*. **14**: 1025-1031.
25. Damiani, G., Caroli, A., Leone, P., Budelli, E., and Florio, S. (2001). Effect of Bov-A2 Sine elements on quantitative traits. Pages 49–51 in *Proc. XIV Natl. Congr. ASPA*, Florence, Italy. University of Florence Press, Florence, Italy.
26. Debeljak, M., Susnik, S., Marinsek-Logar, R., Medrano, J.F. and Dovc, P. (2000). Allelic differences in bovine kappa-CN gene which may regulate gene expression. *Pflagers Arch.* 439 R4–R6.
27. De Lama, S.N. and Zago, M.A. (1996). Identification for the kappa-casein and beta-lactoglobulin genotypes in Brazilian *Bos indicus* and *Bubalus bubalis* population. *Brazilian J. Genet*. **19**: 73-77.
28. Di Stasio, L., and Mariani, P. (2000). The role of protein polymorphism in the genetic improvement of milk production. *Zootecnica Nutrizione Animali.* 26: 69-90.
29. Di Stasio L and Merlin P (1979). Polimorfismi biochimici del latte nella razza bovina Grigio Alpina. *Riv. Zoot. Vet*. **2**:64–67.
30. Dogru, U. and Ozdemir, M. (2009). Genotyping of kappacasein locus by PCR-RFLP in brown Swiss cattle breed. *J Anim Vet Adv*. **8**: 779-781.
31. Dziuba, J. and Minkiewicz, P. (1996). Influence of glycosylation on micelle-stabilizing ability and biological properties of C-terminal fragments of cow's k-casein. *International Dairy Journal.* **6**: 1017-1044.
32. Eenennaam, A. and Medrano, J.F. (1991). Milk protein polymorphisms in California dairy cattle. *J. Dairy Sci*. ***74***:1730–1742.
33. Eggen, A. and Fries, R. (1989). Die Untersuchung von Kaseingenen mittels DNA-Analyse, ETH Landwirtschaft Schweiz Band. **2**: 231–235.
34. Eigel, W.M., Butter, J.E., Ernstrom, C.A., Farrel, H.M., Harwalkar, V.R., Jenness, R., and Whiteney, R.M. (1984). Nomenclature of proteins of cow's milk: fifth revision. *Journal of Dairy Science.* **67**: 1599-1631.
35. Erhardt, G. (1989). Evidence for a third allele at the beta lactoglobulin locus of sheep milk and its occurrence in different breeds. *Animal Genetics*. **20**: 197-204.

36. Epstein, H. (1971). The origin of the domestic animals of Africa. African Publishing Corporation (APC). Volume I. 101 Fifth Avenue, New York, N.Y. 10003, U.S.A.
37. Epstein, H., Mason, I.L. (1984). Cattle. In: Masson IL (ed), Evolution of Domesticated Animals. Longman Inc., New York, U.S.A. 6-28.
38. Farrell, H.M., Jimenez-Flores, Jr. R., Bleck, G.T., Brown, E.M., Butler, J.E., Creamer, L.K., Hicks, C.L., Hollar, C.M., Ng-Kwai-Hang, K.F., and Swaisgood, H.E. (2004). Nomenclature of the Proteins of Cows' Milk—Sixth Revision. *J. Dairy Sci*. **87**:1641–1674.
39. Felsenstein, J. (1985). Confidence limits on phylogenies: An approach using the bootstrap. *Evolution.* **39**:783-791.
40. Flower, D.R. (1996). The lipocalin protein family: structure and function. *Biochemistry Journal.* **318**: 1-14.
41. Folsh, J.M., Coll, A., and Sanchez, A. (1994). Complete sequence of the caprine beta lactoglobulin gene. *Journal of Dairy Science.* **77**: 3493-3497.
42. Fox, P.F. (1989). The milk protein system. In: Developments in Dairy Chemistry. Vol. 4. (Ed. P.F. Fox). London, Elsivier Applied Sciences. *Functional Milk Proteins*.1-53
43. Fox, P.F. and Brodkorb, A. (2008). The casein micelle: Historical aspects, current concepts and significance. *International Dairy Journal.* **18(7)**: 677-684.
44. Fox, P.F. and McSweeney, P.L.H. (1998). *Dairy chemistry and biochemistry* Springer - Verlag.
45. Furet, J.P., Mercier, J.C., Soulier, S., Gaye, P., Hue-Delahaie, D., and Vilotte, J.L. (1990). Nucleotide sequence of the ovine kappa casein cDNA. *Nucleic Acid Research.* **17**: 5286.
46. Ganai, N.A., Bovenhuis, H., van Arendonk, J.A.M. and Visker, M.H.P. (2008). Novel polymorphisms in the bovine b-lactoglobulin gene and their effects on b-lactoglobulin protein concentration in milk. *Animal Genetics*. **40**: 127–133.
47. Geldermann, H. (1990). Genome analysis in domestic animals. VCH, weinheim, *Genetics*. **134**: 943-951.
48. Georges, M., Nielsen, D., Mackinon, M., Mishra, A., Okimoto, R., Pasquino, A.T., Seargeant, L.S., Sorensen, A., Steele, M.R., Zhao, X., Womack, J.E. and Hoeschele, I.

(1995). Mapping quantitative trait loci controlling milk production in dairy cattle by exploiting progeny testing. *Genetics*. **139**: 907-920

49. Gholizadeh, M., Mianji, G.R., and Zadeh, H.S. (2008). Potential use of molecular markers in the genetic improvement of livestock. *Asian J. Anim. Vet Adv*. **3(3)**: 120-128
50. Ginger, M.R. and Grigor, M.R. (1999). Comparative aspects of milk caseins. Comparative biochemistry and physiology part B124: 133-145.
51. Golijow, C.D., Giovambattista, G., Rípoli, M.V., Dulout, F.N. and Lojo, M.M. (1999).Genetic variability and population structure in loci related to milk production traits in native Argentine Creole and commercial Argentine Holstein cattle. *Braz. J. Genet*. **22**: 395-398.
52. Gorodetskii, S.I. and Kaledin, A.S. (1987). Nucleotide sequence analysis of cow kappa casein cDNA. *Genetika.* **23**: 596-604.
53. Gregorius, H. (1987). The relationship between the concepts of genetic diversity and differentiation. *Theoretical and Applied Genetics*. **74**: 397-401.
54. Grodzicker, T., Williams, J., Sharp, P., Sambrook, J. (1975). Physical mapping of temperature sensitive *mutants of adenovirus*. Cold Spring Harbor. *Symp Quant Biol.* **39**: 439-446.
55. Groenen, M.A., Dijkhof, R.J., Verstege, A.J., van der Poel, J.J. (1993). The complete sequence of the gene encoding bovine alpha s2-casein. *Gene*. **123**: 187-193
56. Grosclaude, F., Mahe, M.F., Mercier, J.C. and Ribadeau-Dumas, B. (1972). Localization des substitutions des acides amines differenciant les variants A et B de la caseine kappa bovine. *Annales de Genetique et de selection Animale.* **4**: 515-521.
57. Grosclaude, F., Mahe, M.F ´., Mercier, J.C., Bonnemaire, J., and Teissier, J.H. (1976). Polymorphisme des lactoproteines de bovines Nepalais. Polymorphisme des caseine *α*s2-mineurs. *Ann. Genet. Sel. Evol.* **8**: 461-479.
58. Hall, T. (2001). BioEdit version 5.0.6. North Carolina State University, Department of Microbiology. http://www.mbio.ncsu.edu/BioEdit/biodoc.pdf.
59. Hallen, E. (2008). Coagulation properties of milk: Association with milk protein composition and genetic polymorphism. Doctoral thesis. Swedish University of Agricultural Sciences. Uppsala.

60. Hanotte, O., Tawah, C.L., Bradley, D.G., Okomo, M., Verjee, Y., Ochieng, J. and Rege, J.E.O. (2000). Geographic distribution and frequency of a taurine Bos taurus and an indicine *Bos indicus* Y specific allele amongst sub-Saharan African cattle breeds. *Molecular Ecology.* **9**: 387–396.
61. Hassen, F., Bekele, E., Ayalew, W. and Dessie, T. (2007). Genetic variability of five indigenous Ethiopian cattle breeds using RAPD markers. *African Journal of Biotechnology*. **6(19)**: 2274-2279.
62. Hayes, B. (2007). QTL mapping, MAS, and genomic selection. Iowa State University. Department of Animal Sciences.
63. Hayes, B. and Goddard, M.E. (2001). The distribution of the effects of genes affecting quantitative traits in livestock. *Genetics Selection Evolution.* **33**: 209– 229.
64. Heck, J.M.L., Schennink, A., van Valenberg, H.J.F., Bovenhuis, H., Visker, M.H.P.W., van Arendonk, J.A.M. and van Hooijdonk, A.C.M. (2009). Effects of milk protein variants on the protein composition of bovine milk. *J. Dairy Sci*. **92**:1192–1202.
65. Horisberger, M., and Vonlanthen, M. (1980). Localization of glycosylated kappa casein in bovine casein micelles by lectin-labeled gold granules. *Journal of Dairy Research*. **47**: 185-191.
66. Horne, D.S. (1998). Casein interactions: casting light on the black boxes, the structure in dairy products. *International Dairy Journal.* **8(3)**: 171-177.
67. Horne, D.S. (2002). Caseins, micellar structure. In: Roginski, H*., et al.* (Eds.) London: Academic Press.3. *Encyclopaedia of Dairy Sciences.* 1902-1909.
68. Ikonen, T., Ojala, M., Syva¨oja, E.L. (1997). Effects of composite casein and b-lactoglobulin genotypes on renneting properties and composition of bovine milk by assuming an animal model. *Agric. Food Sci. Finl*. **6**: 283–294.
69. Ikonen, T., Bovenhuis, H., Ojala, M., Ruottinen, O., and Georges, M. (2001). Associations between casein haplotypes and first lactation milk production traits in Finnish Ayrshire cows. *J. Dairy Sci*. **84**: 507–514.
70. Institute of Biodiversity Conservation (2004). The State of Ethiopia's Farm Animal Genetic Resources: Country Report. A Contribution to the First Report on the State of the World's Animal Genetic Resources. IBC. Addis Ababa, Ethiopia.

71. Johann, B. and Peter, D. (2000). Lactoprotein genetic variants in cattle and cheese making ability. *Food technol. Biotechnol*. **38(2)**: 91-98.
72. Jolles, J., Alais, C. and Jolles, P. (1968). The tryptic peptide with rennin-sensitive linkage of cow's kappa casein. *Biochemica Biophysica Acta.* **168**: 91-593.
73. Jolles, P., Loucheux-Lefebvre, M.H., Henschen, A. (1978). Structural relatedness of k-casein and fibrinogen γ-chain. *Journal of Molecular Evolution*. **11**: 271-277.
74. Jones, W.K., Yu-Lee, L.Y., Clift, S.M., Brown, T.L., and Rosen, J.M. (1985). The rat casein multigene family: Fine structure and evolution of the beta-casein gene. *Journal of Biological Chemistry*. **260**: 7042-7050.
75. Karimi, K., Nassiri, M.T.B., Mirzadeh, K., Ashayerizadeh, A., Roushanfekr, H. and Fayyazi, J. (2009). Polymorphism of the β-lactoglobulin gene and its association with milk production traits in Iranian Najdi cattle. *Iranian Journal of Biotechnology*. **7 (2).**
76. Koczan, D., Hobom, G., and Seyfert, H.M. (1991). Genetic organization of the bovine alpha-S1 casein gene. *Nucleic Acid Research*. **19(20)**: 5591-5596.
77. Komori, T. and Nitta, N. (2005). Utilization of CAPS/dCAPS method to convert rice SNPs into PCR-based markers. *Breed Sci.* **55**: 93-98.
78. Kontopidis, G., Holt, C. and Sawyer, L.M. (2004). Lactoglobulin: Binding Properties, Structure and Function. *Journal of Dairy Science*. **87**: 785-96.
79. Kontopidis, G., Holt, C., and Sawyer, L. (2002). The ligand-binding site of bovine *β*-lactoglobulin: Evidence for a function? *J. Mol. Bol*. **318**: 1043–1055.
80. Kubarsepp, I., Henno, M., Viinalass, H., and Sabre, D. (2005). Effect of k-casein and β-lactoglobulin genotypes on the milk rennet coagulation properties. *Agronomy Research.* **3 (1)**:55-64.
81. Kucerova, J., Matejicek, A., Jandurova, O.M., Sorensen, P., Nemcova, E., Stipcova, M., Kott, T., Bouska, J. and Frelich, J. (2006). Milk protein genes *CSN1S1*, *CSN2*, *CSN3*, *LGB* and their relation to genetic values of milk production parameters in Czech Fleckvieh. *Czech J. Anim. Sci.* **51(6)**: 241–247.
82. Kumar, P., Gupta, V.K., Misra, A.K., Modi, D.R., and Pandey, B.K. (2009). Potentia of molecular markers in plant biotechnology. *Plant omics Journal.* **2(4)**: 141-162
83. Lande, R. and Thompson, R. (1990). Efficiency of marker assisted selection in the improvement of quantitative traits. *Genetics.* **124**: 743-756.

84. Lemecha, H., Mulatu, W., Hussein, I., Rege, E., Tekle, T., Abdicho, S. and Ayalew, W. (2006). Response of four indigenous cattle breeds to natural tsetse and trypanosomosis challenge in the Ghibe valley of Ethiopia. *Veterinary Parasitology*. **141**: 165–176.
85. Lien, S., Kantanen, J., Olsaker, I., Holm. L.E., Eythorsdottir, E., Sandberg, K., Dalsgard, B, and Adalsteinsson, S. (1999). Comparison of milk protein allele frequencies in Nordic cattle breeds. *Animal Genetics*. **30**: 85-91.
86. Lien, S. and Rogne, S. (1993). Bovine casein haplotypes: number, frequency and applicability as genetic markers. *Animal Genetics.* **24**: 373-376.
87. Lin, C.Y., McAllister, A.J., Ng-Kwai-Hang, K.F. and Hayes, J.F. (1986). Effects of milk protein loci on first lactation production in dairy cattle. *J. Dairy Sci*. **69**: 704.
88. Liron, J.P., Ripoli, M.V., De Luca, J.C., Peral-Garcia, P. and Giovambattista, G. (2002). Analysis of genetic diversity and population structure in Argentine and Bolivian Creole cattle using five loci related to milk production. *Genetics and Molecular Biology*. **25(4)**: 413-419.
89. Lunden, A., Nisson, M. and Janson, L. (1997). Marked effect of β-lactoglobulin polymorphism on the ratio of casein to total protein in milk. *J Dairy Sci.* **80**: 2996-3005.
90. Maeda, M., Uryu, N., Murayama, N., Ishii, H., Ota, M., Tsuji, K., and Inoko, H. (1990). A simple and rapid method for HLA-DP genotyping by digestion of PCR amplified DNA with allele specific restriction endonucleases. *Hum. Immunol*. **27**: 111-121.
91. Mahe´, M.F., Miranda, G., Queral, R., Bado, A., Souvenir-Zafidrajaona, P., and Grosclaude, F. (1999). Genetic polymorphism of milk proteins in African *Bos taurus* and *Bos indicus* populations characterization of variants *α*S1-Cn H and *κ*-Cn J. *Genet. Sel. Evol*. **31**: 239–253.
92. Marson, E.P., Ferraz, J.B.S., Meirelles, F.V., Balieiro, J.C.C., Eler, J.P., Figueiredo, L.G.G. and Mourão, G.B. (2005). Genetic characterization of European-Zebu composite bovine using RFLP markers. *Genetics and Molecular Research*. **4(3)**: 496-505.
93. Martin, P., Szymanowska, M., Zwierzchowski, L., and Leroux, C. (2002). The impact of genetic polymorphisms on the protein composition of ruminant milks. *Reprd. Nutr. Dev*. **42**: 433-459.
94. Matejicek, A., Matejickova, J., Stipkova, M., Hanus, O., Gercurova, V., Kyselova, J., Nemcova, E., Kott, T., Sefrova, J., Krejcova, M., Melcova, S., Holzelova, I., Bouska, J.,

and Frelick, J. (2008). Joint effects of CSN3 and LGB genes on milk quality and coagulation properties in Czech Fleckvieh. *Czech J. Anim. Sci.*, **53(6)**: 24-252.

95. Mayr, E. (1963). Animal species and evolution. Belknap Press, Cambridge, Mass.
96. Medrano, J.F. and Aguilar-Cordova, E. (a) (1990). PCR-amplification of β-lactoglobulin genomic sequences and identification of genetic variants by RFLP analysis. *Animal Biotechnology.* **8**: 144-146.
97. Medrano, J.F., Aquilar-Cordova, E. (b) (1990). Genotyping of bovine kappa-casein loci following DNA sequence amplification. *Biotechnology.* **8**: 144–146.
98. Mepham, T.B., Gaye, P., Marin, P., and Mercier, J.C. (1992). Biosynthesis of milk proteins. In: Advanced Dairy Chemistry. Vol.1. Proteins. (Ed. P.F. Fox). London, Elsivier Applied Sciences. 491-543.
99. Mercier, J.C., Chobert, J.M., Addeo, F. (1976). Comparative study of the amino acid sequences of the caseinomacropeptides from seven species. *FEBS Lett.* **72(2)**: 208-214.
100. Merlin, P. and Di Stasio, L. (1982). Study on milk protein loci in some decreasing Italian cattle breeds. *Ann. Genet. Sel. Anim.* **14(1)**: 17-28.
101. Meuwissen, T.H.E., Goddard, M.E. (1996). The use of marker haplotypes in animal breeding schemes. *Genet. Sel. Evol.* **28**: 161-176.
102. Meuwissen, T.H.E., Karlsen, A., Lien, S., Olsaker, I. and Goddard, M.E. (2002). Fine mapping of a quantitative trait locus for twinning rate using combined linkage and linkage disequilibrium mapping. *Genetics.* **161**: 373-379.
103. Meuwissen, T.H.E., van Arendonk, J.A.M. (1992). Potential improvements in rate of genetic gain from marker-assisted selection in dairy cattle breeding schemes. *J. Dairy Sci.* **75**: 1651-1659.
104. Michaels, S.D. and Amasino, R.M. (1998). A robust method for detecting single-nucleotide changes as polymorphic markers by PCR. *Plant J.* **14**: 381-385.
105. Miesel, H. (1997). Biochemical properties of bioactive peptides derived from milk proteins: potential nutraceuticals for food and pharmaceutical applications. *Livestock Production Science.* **50**: 125-138.
106. Miller, M.J.S., Witherly, S.A. and Clark, D.A. (1990). Casein: a milk protein with diverse biologic consequences. *Proc Soc Exp Biol Med.* **195**: 143–159.

107. Miller, M.P. (1997). Tools for population genetic analysis (TFPGA) (Version 1.3): a Windows program for the analysis of allozyme and molecular population genetic data. Computer software distributed by the author.

108. Miranda, G., Anglade, P., Mahe, M.F., and Erhardt, G. (1993). Biochemical characterization of the bovine genetic kappa casein C and E variants. *Animal Genetics.* **24**: 27-31.

109. Mitra, A., Yadav, B.R., Ganai, N.A., and Balakrishnan, C.R. (2000). Molecular markers and their applications in livestock improvement. www.ias.ac.in/currsci/oct25/articles19.html.

110. Mohr, U., Koczan, D., Linder, D., Hobom, G., and Erhardt, G. (1994). A single point mutation results in A allele-specific exon skipping in the bovine αS1-casein mRNA. *Gene.* **143**:187–192.

111. Monaco, H.L., Zanotti, G., Spadon, P., Bolognesi, M., Sawyer, L., Eliopoulos, E.E. (1987). Crystal structure of the trigonal form of bovine β-lactoglobulin and of its complex with retinol at 2.5A resolution. *Journal of Molecular Biology*. **197**: 695-706.

112. Mullis, K.B., Faloona, F.A., Scharf, S.J., Saiki, S.K., Horn, G.T., Erlich, H.A. (1986). Specific enzymatic amplification of DNA *in vitro*: the polymerase chain reaction. *Cold Spring Harbor. Symp. Quant. Biol.* **51**: 263-273.

113. Naqvi, A.N. (2007). Application of Molecular Genetic Technologies in Livestock Production: Potentials for Developing Countries. *Advances in Biological Research.* **1(3-4)**: 72-84.

114. Neelin, J.M. (1964). Variants of k-casein revealed by improved starch gel electrophoresis. *Journal of Dairy Science.* **47**: 506-509.

115. Nei, M. (1972). Genetic distances between populations. *Amer. Nat*. 106:283-292.

116. Nei M (1973). Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Science*. **70**: 3321-3323.

117. Nei, M. (1977). F-Statistics and analysis of gene diversity in subdivided populations. *Annals of Human Genetics*.**41**: 225-233.

118. Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics*. **89**: 583-590.

119. Ng-Kwai-Hang, K.F. (2002). Milk Proteins: Heterogeneity, Fractionation and Isolation. In: Roginski, H., *et al.* (Eds.). *Encyclopedia of Dairy Sciences*. 1881-1894.

120. Ng-Kwai Hang, K.F. (1998). Genetic polymorphism of milk proteins: relationship with production, milk composition and technological properties. *Canadian Journal of Animal Science*. **78** (Supplement): 131-147.

121. Ng-Kwai-Hang, K.F. and Grosclaude, F. (1992). Genetic polymorphism of milk proteins. Ed. Elsevier Applied Science, New York, NY. *Advanced Dairy Chemistry*. 405–455.

122. Ng-Kwai-Hang, K.F., Hayes, J.F., Moxley, J.E., and Monardes, H.G. (1986). Relationships between milk protein polymorphisms and major milk constituents in Hoistein-Friesian cows. *J. Dairy Sci*. **69**: 22–26.

123. Ng-Kwai-Hang, K.F., Hayes, J.F., Moxley, J.E., and Monardes, H.G. (1984). Associations of genetic variants of casein and milk serum proteins with milk, fat and protein production in dairy cattle. *J. Dairy Sci*. **67**: 835-840.

124. Ouwehand, A.C., Salminen, S.J., Skurnik, M., and Conway, P.L. (1997). Inhibition of pathogen adhesion by β-lactoglobulin. *International Dairy Journal.* **7**: 685-692.

125. Papiz, M.Z., Sawyer, L., Eliopoulos, E.E., North, A.C.T., Findlay, J.B.C., Sivaprosadarao, R., Jones, T.A., Newcomer, M.E., and Kraulis, P.J. (1986). The structure of β-lactoglobulin and its similarity to plasma and retinol-binding protein. *Nature.* **324**: 383-385.

126. Payne, W.J.A. and Wilson, T.R. (1999). An Introduction to Animal Husbandry in the Tropics, 5th edn. Blackwell Science Ltd, Oxford.

127. Perez, M.D. and Calvo, M. (1995). Interaction of β-lactoglobulin with retinol and fatty acids and its role as a possible biological function for this protein: a review. *Journal of Dairy Science*. **78**: 978-988.

128. Persuy, M.A., Printz, C., Medrano, J.F., and Mercier, J.C. (1999). A single nucleotide deletion resulting in a premature stop codon is associated with marked reduction of transcripts from a goat β-casein null allele. *Animal Genetics.* **30**: 444-451.

129. Phoebe, X.Q. (2007). Studies of casein micelle structure: the past and the present. *Lait*. **87**: 363-383.

130. Prinzenberg, E.M. and Erhardt, G. (1998). A new CSN3 allele in *Bos indicus* cattle is characterized by *Msp*I PCR-RFLP. *Animal Genetics.* **30**: 161-168.

131. Prinzenberg, E.M., Hiendleder, S., Ikonen, T., and Erhardt, G. (1996). Molecular genetic characterization of new bovine kappa casein alleles F and G and genotyping by PCR-RFLP. *Animal Genetics*. **27**: 347-349.

132. Prinzenberg, E.M., Krause, I., and Erhardt, G. (1999). SSCP analysis at the bovine *CSN3* locus discriminates six alleles corresponding to known protein variants (*A, B, C, E, F, G*) and 3 new DNA polymorphisms (*H, I, AI*). *Anim. Biotechnol*. **10**:49–62.

133. Puyol, P., Perez, M.D., Ena, J.M., and Calvo, M. (1991). Interaction of bovine beta lactoglobulin and other bovine and human proteins with retinol and fatty acids. *Agricultural Biological Chemistry*. **10**: 49-62.

134. Rachangani, S. and Gupta, I.D. (2008). Bovine K-casein gene polymorphism and its association with milk production traits. *Genetics and Molecular Biology.* **31(4)***:893-897.*

135. Rando, A., Ramunno, L., and Masina, P. (2000). Mutations in casein genes. *Zoot. Nutr. Anim*. **26**: 105–114.

136. Rando, A., Pappalardo, M., Capuano, M., Di Gregorio, P., Ramunno, L. (1996). Two mutations might be responsible for the absence of betacasein in goat milk. *Anim. Genet*. **27**: 31.

137. Rasmussen, L.K., Hojrup, P., Petersen, T.E. (1994). Disulphide arrangement in bovine caseins: localization of intrachain disulphide bridges in monomers of k- and αs2-casein in bovine milk. *Journal of Dairy Research.* **61**: 485-493.

138. Reddy, M., Kella, N.K.D., and Kinsella, J.E. (1988). Structural and conformational basis of the resistance of beta-lactoglobulin to peptic and chymotryptic digestion. *Journal of Agricultural Food Chemistry*. **36**: 737-742.

139. Rege, J.E.O. (1999). The state of African cattle genetic resources: I. Classification framework and identification of threatened and extinct breeds. *Anim. Genet. Resour. Inf. Bull.* **25**:1–25.

140. Rege, J.E.O., Kahi, A.K., Okomo-Adhiambo, M., Mwacharo, J. and Hanotte, O. (2001). Zebu cattle of Kenya: Uses performance, farmer preferences, measures of genetic

diversity and options for improved use. Animal Genetic Resources Research 1. International Livestock Research Institute (ILRI), Nairobi, Kenya. p. 103.

141. Reid, R.S., Kruska, R.L., Muthui, N., Taye, A., Watton, S., Wilson, C.J., Mulatu, W. (2000). Land-use and land-cover dynamics in response to changes in climate, biological and socio-political forces: the case of southwest Ethiopia. *Landscape Ecol*. **15 (4)**: 339–355.

142. Richard, I. and Beckman, J.S. (1995). How neutrals are synonymous codon mutations? *Nat Genet*. 10:259.

143. Rijnkels, M., Kooiman, P.M., de Boer, H.A., and Pieper, F.R. (1997). Organization of the bovine casein gene locus. *Mammalian genome.* **8**: 148-152.

144. Robitaille, G., Britten, M., Morisset, J. and Petitclere, D. (2002). Quantitative analysis of β-lactoglobulin A and B genetic variants in milk of cow's β-lactoglobulin AB throughout lactation. *J Dairy Sci.* **69**: 651-654.

145. Rousset, F. (2008). GENEPOP'007: a complete re-implementation of the GENEPOP software for Windows and Linux. *Molecular Ecology Resources*. **8(1)**: 103-106.

146. Rozen, S. and Skaletsky, H.J. (2000). Primer3 on the WWW for general users and for biologist programmers. Primer3 on the WWW for general users and for biologist programmers. In: Krawetz S and Misener S (eds) Bioinformatics Methods and Protocols: methods in molecular Biology. *Humana Press, Totowa*. 365-386.

147. Rosenthal, A., Coutelle, O. and Craxton, M. (1993). Large scale production of DNA sequencing templates by microtitre format PCR. *Nucleic Acids Res*. 11: **21(1)**: 1773-4.

148. Sabour, M.P., Lin, C.Y., Lee, A.Y., McAllister, A.J. (1996). Association between milk protein variants and genetic values of Canadian Holstein bulls for milk yield traits. *J. Dairy Sci.* **79**: 1050-1056.

149. Saxena, M.M., Katpatal, B.G. and Pandey, H.S. (1997). Study of milk constituents and their yield in Holstein–Friesian cows. *Indian Journal of Animal Production and Management.* **13(3)**:127–130.

150. Schielben, S., Erhardt, G., and Senft, B. (1991). Genotyping of bovine kappa casein A, B, C, and E following DNA sequence amplification and direct sequencing of kappa casein E product. *Animal Genetics.* **22**: 333-342.

151. Schmidt, D.J. (1964). Starch gel electrophoresis of K-casein. *Biochim. Biophys. Acta*. **90**: 411.

152. Schoen, D.J. and Brown, A.H. (1991). Intraspecific variation in population gene diversity and effective population size correlates with the mating system in plants. *Proc Natl Acad Sci*. **88(10)**: 4494-4497.

153. Semagn, K., Bjorstad, A., and Ndjiondjop, M.N. (2006). Review: An overview of molecular marker methods for plants. *African Journal of biotechnology.* **5(25)**: 2540-2568.

154. Southern, E.M. (1975). Detection of specific sequences among DNA fragments separated by gel electrophoresis. *J. Mol. BioI*. **98**: 503-517.

155. Spaniolas, S., May, S.T., Bennette, M.J., Tuker, G.A. (2006). Authentication of coffee by means of PCR-RFLP analysis and lab-on-a chip capillary electrophoresis. *J Agric Food Chem*. **54**: 7466-7470.

156. Staub, J.E. and Serquen, F.C. (1996). Genetic markers, map construction, and their application in plant breeding. *Hort Science*. **31**: 729-740.

157. Sunyaev, S., Hanke, J., Aydin, A., Wirkner, U., Zastrow, I., Reich, J., Bork, P. (1999). Prediction of nonsynonymous single nucleotide polymorphisms in human disease associated genes. *J Mol Med.* **77**: 754-760.

158. Swaisgood, H.E. (1992). Chemistry of the caseins. N: Advanced Dairy Chemistry. Vol.1 Proteins. (Ed. PF Fox). London, Elsivier. *Applied sciences*. 63-110.

159. Tamura, K., Dudley, J., Nei, M. & Kumar, S. (2007). *MEGA4*: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Molecular Biology and Evolution 24:1596-1599. (Publication PDF at http://www.kumarlab.net/publications).

160. Teneva, A. (2009). Molecular markers in animal genome analysis. *Biotechnology in Animal Husbandry*. **25(5-6)**: 1267-1284.

161. Thompson, M.D., Dave, R.J., and Nakhashi, H.L. (1985). Molecular cloning of mouse mammary gland k-casein: comparison with rat k-casein and rat and human γ-fibrinogen. *DNA*. **4**: 263-271.

162. Threadgill, D.W. and Womack, J.E. (1990). Genomic analysis of the major bovine milk protein genes. *Nucleic Acid Research.* **18**: 6935-6942.

163. Tsiaras, A.M., Bargouli, G.G., Banos, G. and Boscos, C.M. (2005). Effect of kappa-casein and beta-lactoglobulin loci on milk production traits and reproductive performance of Holstein cows. *J. Dairy Sci*. **88**: 327-334.

164. Ulek, S., Tekun, M.E. (2006). Calculation of Adjustment Factors for Standardizing Lactations to Mature Age and 305-Day and Estimation of Heritability and Repeatability of Standardized Milk Yield of Simmental Cattle Reared on Kazova State Farm. *Turk. J. Vet. Anim. Sci.* **30**: 293-289.

165. Van den Berg, G., Escher, J.T.M., De Koning, P.J. and Bovenhuis, H. (1992). Genetic polymorphism of κ-casein and β-lactoglobulin in relation to milk composition and processing properties. *Netherlands Milk and Dairy Journal*. **46**: 145–68.

166. Varv, S., Belousova, A., Sild, E., and Viinalass, H. (2009). Genetic diversity in milk proteins among Estonian dairy cattle. *Vet Med Zoot.* **48**: 70.

167. Vasconcellos, L.P.M., Tambasco-Talhari, D., Pereira, A.P., Coutinho, L.L. and Regitano, L.C.A. (2003). Genetic characterization of Aberdeen Angus cattle using molecular markers. *Genetics and Molecular Biology*. **26(2)**: 133-137.

168. Walstra, P. (1990). On the stability of casein micelles. *Journal of Diary Sciences.* **73**: 1965-1979.

169. Walstra, P. (1999). Casein sub-micelles: do they exist? *International Dairy Journal.* **9(3-9)**: 189- 192.

170. Weiland, J.J. and Yu, H. (2003). A cleaved amplified polymorphic sequence (CAPS) marker associated with root-knot nematode resistance in sugar beet. *Crop Sci.* **43**: 814-881.

171. Weir, B.S. and Cockerham, C.C. (1984). Estimating F-statistics for the analysis of herd structure. *Evolution*. **38**: 1358–1370.

172. Weising, K., Nybom, H., Wolff, K., Meyer, W. (1995). *DNA Fingerprinting in Plants and Fungi* (ed. Arbor, A.) CRC Press, Boca Raton. 1–3.

173. Wicks, S.R., Yeh, R.T., Gish, W.R., Waterson, R.H., and Plasterk, R.H. (2001). Rapid gene mapping in *Caenorhabditis elegans* using a high density polymorphism map. *Nat. Genet.* **28**: 160164.

174. Wong, D.W.S., Camirand, W.M., and Pavlath, A.E. (1996). Structure and functionalities of milk proteins. *Critical Reviews in Food Science and Nutrition*. **36**: 807-844.

175. Wright, S. (1965). The interpretation of population structure by *F*-statistics with special regard to systems of mating. *Evolution*. **19**: 395–420.

176. Wright, S. (1978). Evolution and the genetics of population, variability within and among natural populations. The University of Chicago Press, Chicago.

177. Woychik, J.H. (1964). Polymorphism in k-casein of cow's milk. *Biochem. Biophys*. *Res. Commun*. **16**: 267.

178. Yahyaoui, M.H., Coll, A., Sanchez, A., and Folch, J.M. (2001). Genetic polymorphism of the caprine kappa casein gene. *Journal of Dairy Research.* 68:209-216.

179. Yahyaoui MH, Angiolillo A, Pilla F, Sanchez A and Folch JM (2003). Characterization and Genotyping of the Caprine *κ*-Casein Variants. *J. Dairy Sci*. **86**: 2715–2720.

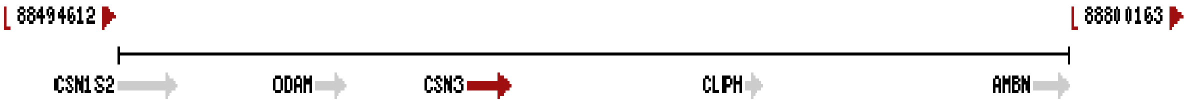
180. Yardibi, H., Turkay, G., Mengi, A., Kaygistz, F. and Oztabk, K. (2009). Genetic polymorphisms of α-lactalbumin and β-lactoglobulin in south Anatolian and east Anatolian red cattle. *Asian Journal of Animal and Veterinary Advances*. **4(5)**: 252-257.

181. Yeh, F.C., Yang, R. and Boyle, T. (1999). POPGENE Version 1.31: Microsoft Window-based Freeware for Population Genetic Analysis. ftp://ftp.microsoft.com/Softlib/MSLFILES/HPGL.EXE

182. Zegeye, Y. (2003). Imperative and challenges of dairy production, processing and marketing in Ethiopia. In: Jobre Y and Gebru G (eds), *Challenges and opportunities of livestock marketing in Ethiopia. Proceedings of the 10th annual conference of the Ethiopian Society of Animal Production (ESAP) held in Addis Ababa, Ethiopia, 22–24 August 2002*. ESAP, Addis Ababa, Ethiopia. pp. 61–67.

APPENDIX

1. Kappa casein (CSN3) Genomic Context



Location of the *CSN*3 gene on the chromosome in non-sequence coordinates. This schematic diagram also shows neighboring genes and indicates their orientations relative to CSN3 gene.Key to Neighboring genes: [CSN1S2-casein alpha S2], [ODAM-odontogenic, ameloblast associated], [CLPH- Casein-like phosphoprotein], [AMBN-ameloblastin, enamel matrix protein]

2. CSN3 amplified region

A 633 bp DNA fragment: a part of intron III (4 bp), exon IV (516 bp) and a part of intron IV (113 bp)

```
        5’-CAGCGCTGTGAGAAAGATGA-3’ (Primer CN-F)
     CAGCGCTGTGAGAAAGATGAAAGATTCTTCAGTGACAAAATAGCCAAATATATCC
CAATTCAGTATGTGCTGAGTAGGTATCCTAGTTATGGACTCAATTACTACCAACAGAAAC
CAGTTGCACTAATTAATAATCAATTTCTGCCATACCCATATTATGCAAAGCCAGCTGCAG
TTAGGTCACCTGCCCAAATTCTTCAATGGCAAGTTTTGTCAAATACTGTGCCTGCCAAGT
CCTGCCAAGCCCAGCCAACTACCATGGCACGTCACCCACACCCACATTTATCATTTATGG
CCATTCCACCAAAGAAAAATCAGGATAAAACAGAAATCCCTACCATCAATACCATTGCTA

                    Thr→Ile Ile→Thr
GTGGTGAGCCTACAAGTACACCTACCATCGAAGCAGTAGAGAGCACTGTAGCTACTCTAG
                        T  C

 Ala→Asp              Ser→Gly
AAGCTTCTCCAGAAGTTATTGAGAGCCCACCTGAGATCAACACAGTCCAAGTTACTTCAA
   A      G            G


CTGCGGTCTAAATACTCTAAGGAGACATCAAAGAAGACAACGCAGGTAAATAAGCAAAAT
    A       A

GAATAACAGCCAAGATTCATGGACTTATTAATAAAATCGTAACATCTAAACTAGCGTAGA
TGGATAAATTAAATCTGTTACAGAGAAGGCGAAATGGG
               3’-ATGTCTCTTCCGCTTTACCC-5’(Primer CN-R
```

3. Virtual restriction maps generated with DNAdot

Red highlighted: Restriction site for HinfI

Green highlighted: Restriction site for HindIII

Enzyme	Cuts	Base 5' to Cleave Sites

HinfI [G^ANTC]	3	21	90	547
HindIII [A^AGCTT]	1	414		
BspMII [T^CCGGA]	1	423		
ApaI [GGGCC^C]	1	295		
HaeIII [GG^CC]	2	295	440	

CSN3 with HindIII [fragment size]	Genotypes			
	AA	AB	BB	Uncut
633	▬	▬		▬
414		▬	▬	
219		▬	▬	

CSN3 with HinfI [Fragment size]	Genotypes			
	AA	AB	BB	Uncut
633				▬
474		▬	▬	
335	▬	▬		
139	▬	▬		
96	▬	▬	▬	
63	▬	▬	▬	

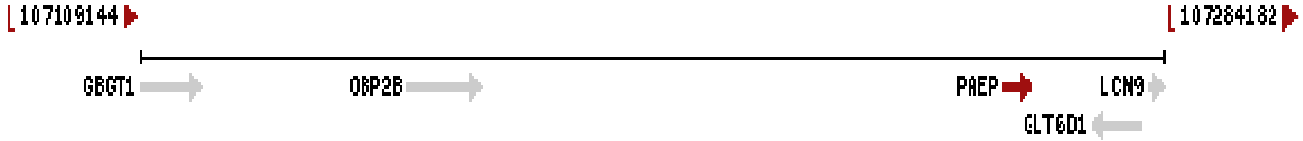
CSN3 with BspMII [Fragment size]	Haplotypes			
	AA	AA^1	A^{1}A^1	Uncut
633	▬	▬		▬
423		▬	▬	
210		▬	▬	

CSN3 with ApaI [Fragment size]	Haplotypes			
	AA	AE	EE	Uncut
633	▬	▬		▬
442		▬	▬	
191		▬	▬	

CSN3 with HaeIII [Fragment size]	Haplotypes			
	AA	AE	EE	Uncut
633				▬
338	▬	▬		
295	▬	▬	▬	
193		▬	▬	
145		▬	▬	

4. **Lactoglobulin Beta [LGB]**

chromosome: 11; **Location:** 11q28



Location of the *LGB* gene on the chromosome in non-sequence coordinates.
Key to Neighboring genes: [GBGT1-Globoside alpha-1, 3-N-acetylgalactosaminyl transferase I], [OBP2B-Odorant binding protein 2B], [GLT6D1glycocyl transferase 6 domain containing 1] and [LCN9-Lipocalin 9]

5. *LGB* amplified region

A 529 bp DNA fragment: a part of intron III (41 bp), exon IV (111 bp) and a part of intron IV (377 bp)

```
                  5′-CCTGCTGGAACTCACTTTCC-3′ (Primer LG-F)
2957                 CCTGCTGGAACTCACTTTCCTCCCGTCTTGATCTCTTCCAGCCT
                                            T            A

3001 TGAATGAGAACAAAGTCCTTGTGCTGGACACCGACTACAAAAAGTACCTGCTCTTCTGCA
         C                          T

                                    Ala→Val
3061 TGGAGAACAGTGCTGAGCCCGAGCAAAGCCTGGCCTGCCAGTGCCTGGGTGGGTGCCAAC
                                      T

3121 CCTGGCTGCCCAGGGAGACCAGCTGTGTGGTCCTCGCTGCAACGGGGCCGGGGGGGACGG
                                                      C

3181 TGGGAGCAGGGAGCTTGATTCCCAGGAGGAGGAGGGATGGGGGGTCCCCGAGTCCCGCCA
          A                    T

3241 GGAGAGGGTGGTCATATACCGGGAGCCGGTGTCCTGGGGGTCTGTGGGTGACTGGGGACG
                                A          A C  A

3301 GGGGCCAGACACACAGGCTGGGAGACGGGGGGCTGCAGCGCTCTGGTGTGACCATCACGA
3361 TGGAGCCGGCGGTCACTATGAATCTAACAGCCTTTGTTACCGGGGAGTTTCAATTATTTC
3421 ATCAAATAAGAACTCAGGCACAAAGCTGTCTTTCAACTGTCACGTCCTGAAAACAAATGG
                                               3′-GGACTTTTGTTTACC
     CAGGT
     GTCCA-5′(Primer LG-R)
```

Restriction map generated with DNAdot:

Blue highlighted: Restriction site for HaeIII

Enzyme	Cuts	Base 5' to Cleave Sites		
HaeIII [GG^CC]	3	137	211	348
DdeI[C^TNAG]	2	117	477	

LGB with HaeIII [Size in bp]	Genotypes			
	AA	AB	BB	Uncut
529				———
221	———	———		
177	———	———	———	
138	———	———	———	
115		———	———	
86	———	———	———	
13		———	———	

LGB with DdeI [Size in bp]	Genotypes			
	AB	AB	AB	Uncut
529				———
386	———	———	———	
93	———	———	———	
50	———	———	———	

6. *LGB and CSN3 Genotypes of the 5 local and 1 cross-breed cattle populations*

1. [Abigar]

Serial No. **[Abigar]**	*LGB* Genotype [HaeIII]	*CSN*3 Genotype [HindIII]
1A	AB	AA
2A	AB	AA
3A	BB**	AA
4A	BB	AA
5A	AB	AB
6A	BB	AB
7A	BB	AA**
8A	BB	AA
9A	AB	AB
10A	AB	AB
11A	BB	AB

2. [Boran]

Serial No. **[Boran]**	*LGB* Genotype [HaeIII]	*CSN*3 Genotype [HindIII]
B1	BB	AA**
B2	BB	AB
B3	BB	AA
B4	AB	AB
B5	AA**	AA
B6	BB**	AA
B7	BB	AA
B8	BB	AB
B9	BB	AB
B10	BB	AA
B11	BB	BB**

3. [Guraghe]

Serial No. **[Guraghe]**	*LGB* Genotype [HaeIII]	*CSN*3 Genotype [HindIII]

G1	BB**	BB
G2	BB	AA**
G3	BB	AA
G4	BB	AB
G5	BB	AB
G6	BB	AA
G7	AB	BB**
G8	BB	AB
G9	BB	AA
G10	BB	AA
G11	BB	AA

4. [Sheko]

Serial No. **[Sheko]**	*LGB* Genotype [HaeIII]	*CSN*3 Genotype [HindIII]
1S	AB	AB
2S	BB	AA**
3S	AA**	AB
4S	BB	BB**
6S	BB	AB
7S	BB**	AA
8S	BB	AB
9S	BB	AB
10S	BB	AA
11S	BB	AB
12S	BB	BB
13S	BB	AA
14S	AA	AA
15S	BB	AB
16S	AB	BB
17S	BB	BB

5. [Horro]

Serial No. **[Horro]**	*LGB* Genotype [HaeIII]	*CSN*3 Genotype [HindIII]
1H	BB	AA
2H	BB	AA
3H	AB	AB
4H	BB	AA
5H	BB	AB
6H	BB	AA**
7H	BB	AB
8H	BB	AA
9H	BB	AB
10H	BB	AA
11H	AB	AB
12H	BB	AB
13H	AA**	BB**
14H	BB	AA
15H	BB**	AA
16H	BB	AA
17H	BB	AB

19H	AB	AA
20H	BB	AB
21H	BB	AB
22H	BB	AB
23H	BB	AB
25H	BB	AA

6. [Holstein*Barka]- HC

Serial No. **[Holstein*Barka]**	*LGB* Genotype [HaeIII]	*CSN*3 Genotype [HindIII]
1HC	BB	AB
2HC	BB	AA**
3HC	BB**	AA
4HC	AB	AB
5HC	BB	AA
7HC	AB	AB
9HC	BB	AB
10HC	BB	AB
11HC	BB	AA
12HC	AB	AB
13HC	BB	AA

****These were the samples used for sequencing the *CSN*3 and *LGB* PCR products**

Declaration

I, the undersigned, declare that this thesis is my original work. It has never been submitted in any institution and that all sources of materials used for the thesis have been duely acknowledged.

Name: **Robel Getachew**

Signiture ____________________

Date ____________________

This thesis has been submitted for examination with our approval as:

Major Supervisor: **Abiy Zegeye (PhD)**

Signiture _________________

Date ____________________

Co-Supervisor: **Tadelle Dessie (PhD)**

Signiture ____________________

Date _______________________