Genome-wide identification and expression analysis of the Collagen family during lactation in water buffalo (Bubalus Bubalis)

CURRENT STATUS: POSTED

Xingrong Lu
Buffalo research institute

Anqin Duan
Buffalo research institute

Shasha liang
buffalo research institute

Xiaoya Ma
Buffalo research institute

Chunying Pang
Buffalo research institute

Xianwei Liang
Buffalo research institute

Tingxian Deng
Buffalo Research Institute
dtx282000@163.com

Corresponding Author
ORCiD: https://orcid.org/0000-0002-7442-6739

DOI: 10.21203/rs.2.9247/v1

SUBJECT AREAS
Internal Medicine Specialties

KEYWORDS
buffalo, collagen family, evolutionary relationship, gene expression
Abstract

Background: Collagens, as extracellular matrix molecules, support cells for structural integrity and a variety of other functions, thereby contribute to support mammary basic structure and development. However, little information on the identification and expression profiles in response to the mammary gland of the collagen family in buffalo (Bubalus Bubalis) has been reported. Results: A total of 128 buffalo collagen protein sequences corresponding to 45 collagen genes were identified and classified into six clusters based on their phylogenetic relationships, conserved motifs, and gene structure analyses. A transcription factor binding sites (TFBS) analysis inferred that a total of 142 TFBS were predicted within the buffalo collagens, suggesting that different collagen subfamilies harbored different TFBS and played a variety of functions involved in the mammary gland development and lactation. The identified collagen sequences were unequally distributed on 17 chromosomes, 103 of which were determined to be tandem duplicated genes. Transcriptome data and qRT-PCR analysis revealed the expression diversity of buffalo collagen genes in various tissues. Most of the identified collagen genes were significantly up-regulated at the early lactation, 6 collagens upregulated at the peak lactation, and only COL24A1 was up-regulated at the late lactation. Conclusions: The present study provides significant insights into the potential functions of collagen family in dairy buffalo and helps in the functional characterization for collagen genes in further research.

Background

The mammary gland is a dynamic organ in mammals that produce milk to meet nourishment requirements for the offspring [1, 2]. It is also a unique glandular secretory organ composed of different cell types, including epithelial, adipose, fibroblasts, immune, lymphatic and vascular cells [2, 3]. Mammary epithelium consists of two cell layers: a luminal layer of secretory epithelial cells and a basal layer of myoepithelial cells. Luminal epithelial cell is responsible for the synthesis and secretion of milk components, while the myoepithelial cell mainly participates in the process of milk ejection. The myoepithelial cells reside in the basement membrane of the extracellular matrix (ECM). Accumulating evidence has demonstrated that collagen is the major component of ECM that play a crucial regulatory role in the epithelial architecture and function [4-6]. Schedin and Keely [7] also
highlighted ECM composition and organization are tightly regulated throughout the mammary gland development. Consequently, as the major component of ECM, the collagen family maybe play a vital role in the specific stages of mammary gland development.

Collagens represent the most abundant protein in mammals commonly possessing at least one triple-helical domain. Since collagen II was discovered by Miller and Matukas [8], twenty-eight members of the collagen family have recently been identified and characterized [9]. These identified collagen members are deposited in the ECM where most of them can form supramolecular aggregates. Notably, the knowledge about the molecular structure, biosynthesis, and function of the collagen family has emerged [10]. Twenty-eight collagen genes, for example, can be grouped into several clusters based on their structure [11]: Fibrillar collagens (types I, II, III, V, XI, XXIV, and XXVII), Fibril-associated collagens with interrupted triple helices (types IX, XII, XIV, XVI, XIX, XX, XXI, and XXII), Network-forming collagens (types IV, VIII, and X), Trans-membranous collagens (types XIII, XVII, XXIII, and XXV), Endostatin precursor collagens (types XV and XVIII), and others (types VI, VII, XXVI, and XXVIII). Moreover, evidence revealed that different collagen genes displayed different expression patterns and played the structural and functional roles [10]. Chen et al. [12] found the COL4A1 were significantly downregulated in the inflammation-associated fibroblasts extracted from bovine mammary glands with clinical mastitis compared with normal fibroblasts from a slaughtered dairy cow, implying that it might involve in the ECM remodeling. Crisà et al. [13] reported that COL4A1 was highly expressed in the mature milk compared with that of the colostrum milk in goat. Furthermore, collagens are demonstrated to participate in cell-matrix interactions via several receptor families [14-16]. Although the expression profiles of collagens have been previously reported, expression patterns of the collagen gene family in buffalo (Bubalus Bubalis) are poorly characterized.

Water buffalo is the second largest source of milk production that mainly distributed in the tropical and subtropical areas, providing more than 5% of the world’s milk supply [17]. Remarkably, buffalo milk is more abundant in fat and protein than cow’s milk [18], which has attracted widespread attention from the dairy industry. In the past several decades, numerous high-throughput data were
generated and utilized for identifying the candidate genes related to traits of interest in buffalo, such as the potential genes related to milk or productive traits [19, 20], transcriptome profiles of buffalo embryos with normal and retarded growth [21], and maternally-expressed proteins in buffalo oocyte [22]. All those data, along with the complete buffalo genome sequence [23], provide the possibility to perform the gene family analysis at a genome-wide level. In plants, numerous gene families have been identified and characterized, such as the sugar transporter gene family [24], KUP family [25], and heat shock factor gene family [26], which contribute to the understanding of gene structure, evolutionary patterns, and expression profiles. In animals, Körbes et al. [27] reported the molecular evolution of the lysophosphatidic acid acyltransferase (LPAAT) gene family, suggesting that the evolution of triacylglycerol biosynthesis shaped by diversification of the LPAAT genes. However, little information on the expression profiles of the collagen family in different tissues and developmental mammary gland of buffalo is available. Because of the vital functions of collagen in animals, it is of considerable importance to investigate the buffalo collagen family. In the present study, we identified collagen family from the buffalo genome and analyzed their evolutionary relationship, sequence features, chromosomal location, gene duplication, tissue-specific expression levels and dynamic expression patterns in response to different lactation points in mammary gland tissues. Our results provide some insight into the understanding of buffalo collagen family and present vital evidence for future functional studies.

Results
Identification and classification of the collagen family in buffalo

In the present study, an integrative analysis of BLAST and HMMER showed that a total of 128 putative collagen protein sequences were identified from the buffalo whole genome, belonging to the 45 buffalo collagen genes (Table 1). The open reading frames (ORFs) of the collagen proteins ranged from 1,317 to 9,654 bp in length encoding the protein of 438 to 3,217 residues, with the predicted MW from 107.46 to 786.60 kDa. The pi values of these proteins ranged from 4.54 to 4.92. Furthermore, the phylogenetic analysis revealed that the identified collagen genes could be classified into six
clusters (Figure 1). The top 3 clusters contained 8, 7, and 4 collagen genes, corresponding to the Cluster II, Cluster I, as well as Cluster IV and VI, respectively. The smallest group is the Cluster V with only two collagen members (collagens XV and XVIII). The constructed dendrogram further showed that buffalo collagen family was generally most closely evolutionary relationship with other five representative mammals.

Gene Structure, Conserved Motif Analysis, and Promoter Analysis

As shown in Figure 2A, the exon numbers of buffalo collagen family ranged from 2 to 118, and most of them harbored 40 to 60 exons. The analysis of intron numbers in the studied species showed that the intron number in the collagen family varied widely, ranging from 1 to 117 (Supplementary Figure S1). Most collagen sequences contained more than forty introns, whereas only the COL8A1, COL8A2, and COL10A1 contained fewer than three introns. The collagen members in the same group had similar numbers of exon and intron among the studied species. These findings suggest that different groups in the collagen family have different patterns of intron number, which verify our previous classification process.

The MEME analysis showed that a total of 10 conserved motifs were identified in all collagens (Figure 2B). After the Pfam search, the motifs 3 and 5 were annotated as the collagen domain (Table 2), suggesting that at least collagen domain existed in the identified collagens. The results were also supported by the identified collages blasted against the conserved domain database (CDD) from NCBI (Figure 2A). Interestingly, the Fibrillar collagen C-terminal domain (COLFI), Fibronectin type 3 domain (FN3), C-terminal tandem repeated domain in type 4 procollagen (C4), von Willebrand factor (vWF) type A domain (VWA), von Willebrand factor type C domain (VWC), and EMI domain also was determined in some collagen genes.

All identified buffalo collagens with the 1500 bp upstream sequences of coding sequence were selected and used to predict the TFBS. The results exhibited a total of 142 TFBS were predicted within
the buffalo collagen genes (Figure 3A). Among them, the largest TFBS detected in the identified collagen genes was the Myeloid Zinc Finger 1 (MZF1), with the numbers of 61 (7.96%), followed by the Homeobox protein Meis1 (MEIS1) (43; 5.61%), Zinc Finger Protein 354C (ZNF354C) (42; 5.48%), Rhox homeobox family member 1 (RHOXF1) (40; 5.22%), Krueppel-like factor 5 (KLF5) (33; 4.31%), Zinc finger protein 384 (ZNF384) (25; 3.26%), and Transcription factor Spi-B (SPIB) (21; 2.74%), while the remaining TFBS with the number of ≤ 20 (Supplementary Figure S2). For the buffalo collagen family, moreover, we found that the MZF1 element was mostly detected in the Clusters I, III, and V, the MEIS1 element was mostly predicted in the Clusters II and VI, and ZNF384 element was mainly discovered in the Cluster IV (Supplementary Figure S3).

Chromosomal location and Gene Duplication Analysis of buffalo collagen family

The 128 buffalo collagen sequences were distributed on 17 chromosomes randomly (Figure 4). Among them, chromosome 2, 4, 1, and 3 separately had 26, 18, 15, and 13 collagen each, chromosome 10 and X both contained 7 collagen each, chromosome 6 possessed 8 collagen each, chromosome 14 contained 6 collagen sequence, chromosome 7 and 15 both had 5 collagen each, while the remaining chromosome has less than 5 collagen each. Notably, the buffalo collagen genes were mainly located on the proximate or the distal ends of the chromosomes.

To investigate the evolution progress of putative gene family, we analyzed the duplication events of buffalo collagens. Here, a total of 103 (80.47%) collagen sequences were confirmed to be tandem duplicated genes. The majority of collagen tandem duplicated genes located on chromosome 1 ~ 4. Four separate pairs of collagen tandem duplicated genes located on chromosome 10 and X. Four groups of three tandem duplicated gene pairs located on chromosomes 6, 7, 14, and 15. Four chromosomes (9, 13, 23, and 24) had the tandem duplicated gene pairs with the numbers of ≤ 2. However, only one segmental duplication event (COL1A1 and COL1A2) was discovered in the buffalo collagen family. These results suggested that the tandem duplication played a predominant role in the expansion of buffalo collagen family.
To further explore whether the positive selection was involved in buffalo collagen genes, we analyzed the Ka/Ks ratios for each duplicated collagens pair (Table 3). The results showed that the COL14A1_2/COL14A1_1 duplicated collagen pair was identified with the Ka/Ks ratios > 1, implying that this gene pair might involve in the positive selection. Moreover, a total of 23 duplicated collagen gene pairs with the Ka/Ks ratios < 1 were identified, 8 of which might experience strong purifying selection pressures because they had the Ka/Ks ratios were less than 0.5.

Expression profiling of buffalo collagen genes

To investigate the tissue expression distribution of the buffalo collagen family, we examined the expression patterns of 45 collagen genes using the public RNA-seq data. The results showed that a total of 19 collagen genes were not expressed in any tissues of buffalo, including the COL10A1, COL11A1, COL13A1, COL19A1, COL1A2, COL20A1, COL21A1, COL23A1, COL24A1, COL25A1, COL26A1, COL4A4, COL4A6, COL5A1, COL6A3, COL6A4, COL6A5, COL6A6, and COL8A2 (Figure 5A). A total of 26 collagen genes was detected from at least one tissue and displayed differential expression in various buffalo tissues. Among them, 5 (COL9A3, COL6A1, COL6A2, COL3A1, and COL5A2) genes showed highest expression levels in the testis, while 3 (COL14A1, COL4A1, and COL4A2) genes had highest expression levels in the ovary follicle, indicating that these genes might be related to the reproductive traits in buffalo. Interestingly, the COL16A1 gene exhibited the highest expression levels in the mammary gland, suggesting that it might involve in the biological function of milk production trait in buffalo. Moreover, the result of qRT-PCR also showed a similar trend with the RNA-seq data (Figure 5B).

Using the RNA-seq data, we further dissect the transcript levels of buffalo collagen family in mammary gland tissues at four lactation points. The results showed that 29 out of 45 collagen genes were mainly up-regulated at the D7, six collagens (COL6A6, COL3A1, COL1A1, COL27A1, COL8A1, and COL14A1) upregulated at the D50, two collagen genes (COL4A3 and COL10A1) were mainly expressed
at the D50 and D140, and only the COL24A1 was up-regulated at the D280 (Figure 6). The results of qRT-PCR also demonstrated that four collagen genes in buffalo were up-regulated in the D50 and two collagens were highly expressed in the D7, implying that the RNA-seq data in the present study was available.

Discussions

Collagens, as extracellular matrix molecules, support cells for structural integrity and a variety of other functions [11], thus contribute to support mammary basic structure and development [28]. Currently, our understanding of the functional role of buffalo collagen family is limited. In the present study, we identified 128 collagen protein sequences in buffalo based on its complete genome sequence. The identified collagen protein sequences corresponded to the 45 collagen genes in buffalo, which was classified into six clusters based on their evolutionary relationships. The result was consistent with the previous classification of collagens described by S Ricardblum [9]. The conserved motif and gene structure analyses of buffalo collagen genes also supported this classification perspective. Conserved motif analysis indicated that all the identified collagens harbored at least collagen domain. Interestingly, the collagen cluster I genes contained a fibrillar collagen C-terminal domain (COLFI). For the gene structure analysis, most of the collagen genes contained 40 to 60 exons. The intron number analysis revealed that the majority of collagen genes contained more than forty introns, whereas only the COL8A1, COL8A2, and COL10A1 contained fewer than three introns. These results coincide with the exon-intron structure of collagen genes from other representative mammals, suggesting that collagen family had a conserved gene structure.

For the identified buffalo collagens, we further predicted the TFBS in their promoter regions. The results exhibited a total of 142 TFBS were detected within the buffalo collagen family. We found that the largest TFBS with the numbers of 61 was the MZF1 (7.96%), followed by the MEIS1 (5.61%), ZNF354C (5.48%), RHOXF1 (5.22%), KLF5 (4.31%), ZNF384 (3.26%), and SPIB (2.74%). Moreover importantly, we found that the MZF1 element was mostly detected in Clusters I, III, and V of buffalo collagen family, the MEIS1 element was mostly predicted in the Clusters II and VI, and ZNF384
element was mainly discovered in the Cluster IV. Previous studies have demonstrated that as an N-cadherin transcription factor, MZF1 could participate in the epithelial-mesenchymal transition that plays a critical role in cell invasion and metastasis [29]. It is noted that cell migration through the extracellular matrix was a key component of morphogenesis [30], suggesting that these genes in the Clusters I, III, and V of buffalo collagen family might be involved in the mammary gland morphogenesis. MEIS1 has been demonstrated to be a transcriptional regulator of PAX6 that is required for the normal development of the endocrine cells [31, 32]. This indicated that the MEIS1 along with PAX6 might be involved in the insulin secretion and synthesis pathway, thereby resulting in the changes of lactation function in mammary gland because the insulin plays a vital role for the mammary gland differentiation and subsequent lactation [33, 34]. Our data also revealed the MEIS1 element was mostly detected in Cluster II and VI, implying that the biological function of them might be related to the insulin secretion and synthesis. Moreover, evidence indicated that ZNF384 as a transcription factor could bind and regulate the promoters of the extracellular matrix genes that played critical roles in mammary gland development [35, 36]. Our finding predicted the ZNF384 element was mostly detected in the Cluster IV, indicating that these cluster genes had effects on the extracellular matrix remodeling in the mammary gland. These results indicated that the different collagen genes in different clusters play a variety of functions involved in the mammary gland development and lactation.

Four events during genetic evolution including the chromosome doubling, chromosome fragment insertion mutation, tandem duplication, and transposition provide a possibility for the novel gene function acquisition [37, 38]. Gene duplication including the tandem duplication and segmental duplication can mainly help to accelerate the gene family expansion and genome evolutionary mechanisms [39-41](Vision et al., 2000, Liu et al., 2009). In buffalo, all identified buffalo collagen genes were unevenly distributed on 17 chromosomes. Here, a total of 103 (80.47%) buffalo collagen sequences were confirmed to be tandem duplicated genes, but only one segmental duplication event to be discovered, which revealed that the tandem duplication had a predominant role in the expansion
of buffalo collagen family. This finding was supported by Liu et al. [39] who highlighted that segmental duplications in most mammalian lineages are organized in a tandem configuration. For the duplicated collagen pairs, interestingly, we found that the COL14A1_2/COL14A1_1 duplicated collagen pair underwent the positive selection during the evolution of buffalo collagen family. Furthermore, twenty-three pairs of duplicated collagens with the Ka/Ks ratios < 1 were identified, eight of them with the Ka/Ks ratios were less than 0.5 that might experience strong purifying selection pressures.

Tissue expression profiles from the RNA-seq data analysis were presented in this study, showing that a total of 26 collagen genes was detected from at least one tissue and displayed differential expression in various buffalo tissues. In particular, buffalo collagen genes in the same clusters had no consistent expression. For example, a total of five (COL9A3, COL6A1, COL6A2, COL3A1, and COL5A2) genes showed highest expression levels in the testis, while three (COL14A1, COL4A1, and COL4A2) genes had highest expression levels in the ovary follicle, indicating that these genes might be related to the reproductive traits in buffalo. The reason for the difference could be the limitation of available expression data for buffalo collagen. Importantly, some of the collagen genes exhibited tissue-specific expression, such as COL16A1 was mainly expressed in the mammary gland, while the COL2A1, COL8A1, and COL27A1 were explicitly expressed in the thymus, corpus luteum, and cerebellum, respectively. Apparently, the collagen family has been demonstrated to have various functions in multiple tissues because they contribute to maintaining the structural integrity and stability of tissues and organs [10]. Compared with the expression pattern represented by RNA-seq data, the expression profile generated by qRT-PCR was not wholly equal to that. We assumed that there existed multiple reasons to result in a discrepancy of expression profiles. For example, highly expressed genes can significantly affect the FPKM values, thereby resulting in the bias of FPKM [42]. In this regard, Zhao et al. [41] reported that the expression pattern of Hsp20 gene family in potato from the RNA-seq data and qRT-PCR was a little different.

Previously studies showed that the members of the collagen family are the most abundant proteins in ECM that are tightly regulated throughout the development of the mammary gland [28, 43, 44].
Therefore, to explore the expression pattern of collagen genes in the buffalo mammary gland at different lactation points is necessary, which contribute to dissect the potential roles of these genes in the mammary gland morphogenesis. In the present study, our data revealed that most of the collagen genes are mainly expressed in the mammary gland at early lactation. It is noted that a total of 6 buffalo collagens (COL6A6, COL3A1, COL1A1, COL27A1, COL8A1, and COL14A1) was upregulated at the peak lactation, two collagen genes (COL4A3 and COL10A1) were mainly expressed at the peak and mid-lactation, and only the COL24A1 was up-regulated at the late lactation. These results were also supported by the qRT-PCR. Interestingly, Dai et al. [45] found that COL8A1 and COL1A2 both were upregulated in the bovine mammary gland during lactation compared to the dry period. Our findings suggested that these buffalo collagen genes displayed the specific biology function at different stages of mammary gland development. However, these findings remain to be confirmed.

Conclusion
In this work, we performed a genome-wide identification of buffalo collagen family, with 128 collagen protein sequences confirmed. Next, we performed the analyses of buffalo collagen genes, including the phylogeny, gene structures, chromosomal location, gene duplication, and expression patterns in different tissues. Most of the collagen genes were expressed in the different buffalo tissues and were up-regulated in the mammary gland at the early lactation, indicating that these genes play vital roles in the milk production of buffalo. The study provides valuable information on the collagen family in buffalo and will helps in determining the collagen gene functions.

Methods
Animal source

A total of eleven Murrah buffalo was used as the animal material in the present study, which was kept at the Buffalo Research Institute, Chinese Academy of Agricultural Sciences (BRI-CAAS). A total of three buffaloes were slaughtered and used for the fresh tissue collection, including brain, heart, liver, spleen, lung, kidney, stomach, large intestine, ovary, oviduct, pituitary, uterus, hypothalamus, lymph, and mammary gland. The biopsy samples of mammary gland tissue from eight buffaloes were
collected on days 7 (D7), 50 (D50), 140 (D140), and 280 (D280) after calving, which was further used for the real-time quantitative PCR (qRT-PCR). The biopsy procedure was performed based on the method reported by Schmitz et al. [46]. All fresh samples were immediately preserved in the liquid nitrogen until use.

Identification of the buffalo collagen genes

The whole-genome data of six representative mammals including the human (GRCh38.p12), cattle (ARS-UCD1.2), buffalo (UOA_WB_1), goat (ARS1), sheep (Oar_rambouillet_v1.0), and horse (EquCab3.0) were downloaded from the NCBI Genome database (https://www.ncbi.nlm.nih.gov/genome/), aiming to identify the buffalo collagen genes. The HMM profile of the collagen domain (PF01391) from the Pfam database (http://pfam.xfam.org/family/Collagen) was employed to search the buffalo dataset using the HMMER [47, 48] software with an E-value cut-off of 1E-5. The identified buffalo collagen protein sequences were also validated by the BLAST with collagens from other five mammals as queries. Further, the ClustalW algorithm was used for the multiple sequence alignment of collagen with the full-length protein sequences. The aligned sequences were used for constructing the neighbor-joining (NJ) phylogenetic tree of collagen family using the MEGA7 [49] software with the Poisson model, pairwise deletion, and 1000 bootstrap resamplings.

Sequence Analysis

ExPASy proteomics server was utilized for the prediction of theoretical molecular weight (MW) and isoelectric points (pI) of buffalo collagen family (Gasteiger et al., 2003). The gene structure was assessed and visualized using the TBtools version. 0.6657 [50]. Exon/intron structure analysis was performed by the buffalo genome annotation file using the in-house scripts. We analyzed the conserved protein motifs of collagen proteins in buffalo using the MEME programs (http://meme-suite.org/tools/meme) at a maximum number of motifs, 10. Moreover, the transcription factor binding sites (TFBS) in the identified collagen promoter region with 1500 bp in length were predicted using the TFBSTools [51] R package. The main parameters of TFBS analysis were employed as the following:
JASPAR2018 was used to retrieve data from JASPAR database, the Homo sapiens was as the reference species, the minimum score for the hit was more than 0.98, and only the positive strand was considered.

Chromosome locations of each collagen gene were obtained from their genome resources. Collagen duplication events were identified using the Multiple Collinearity Scan Toolkit (MCScanX) previously described by Wang et al. [52]. Overall, a total of 58,532 buffalo protein sequences were analyzed using the BLAST search with E-value < 1E-5, and then the corresponding gene positions file were obtained from the buffalo genome annotation file with the GFF format, which can be set as the input files of MCScanX program.

For the duplicated collagen genes, we further performed the divergence estimates and diversity analysis. In brief, MEGA7 software with the MUSCLE algorithm was firstly utilized for the pairwise alignment of the homologous collagen gene pairs. Subsequently, the DnaSP v6.0 [53] software was employed to calculate the pairwise synonymous (Ks) and nonsynonymous (Ka) numbers of substitutions corrected for multiple hits.

Expression pattern Analyses

In order to determine the tissue expression profiles of collagen family, the RNA sequencing (RNA-seq) data (BioProject: PRJNA207334) from 29 different Italian Mediterranean buffalo tissues was used to calculate the fragments-per-kilobase-per-million fragments mapped (FPKM) values of buffalo collagens. Briefly, FASTQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) program was used to remove the adapter sequence from the raw sequence reads. The clean reads were aligned to the buffalo reference genome using Hisat2 [54] with default parameters. FPKM values were calculated using the StringTie [55].

To explore the expression profiles of the buffalo collagen family in mammary gland tissues at different lactation, including the early lactation (D7), peak lactation (D50), mid-lactation (D140), and late lactation (D280), the raw data (BioProject ID: PRJNA480718) from buffalo mammary gland tissues submitted by our group was used. Hierarchical clustering was performed using the Pheatmap [56] R package with the normalized scale method.
qRT-PCR Analysis

Total RNA for the different buffalo tissues was isolated using the RNA Plus reagent (Tiangen, China). For the 2 μg total RNA, first-strand cDNA was synthesized using a reverse transcriptase kit (Takara, Dalian, China). SYBR® Premix Ex Taq™ (Takara, Dalian, China) was used for the qRT-PCR analysis, which was monitored on the LightCycler 480 (Roche, CH), and each reaction was performed in triplicate. The expression levels of buffalo collagen family were analyzed by the 2-ΔΔCt method [57] and normalized using the GAPDH expression analysis. The collagens primer pairs used for the qRT-PCR analysis are shown in the Supplementary Table S1.

Abbreviations

NCBI: National Center for Biotechnology Information; FPKM: fragments-per-kilobase-per-million fragments mapped; D7: day 7 after calving; D50: day 50 after calving; D140: day 140 after calving; and D280: day 280 after calving.

Declarations

Ethics approval and consent to participate

All experimental design and animal care were approved by the Animal Ethics Committee of the BRI-CAAS (approval code GXBRI-01-2019).

Consent for publication

Not applicable.

Availability of data and material

The data supporting our findings can be found in the additional supporting files.

Competing interests

The authors declare no competing financial interests

Funding

This study was supported by the Natural Science Foundation of Guangxi (2016GXNSFBA380226, 2017GXNSFBA198191, and 2017GXNSFBA198022), Major Science and Technology Projects in Guangxi (AA16450002), and the Central Guidance for Local Science and Technology Development Projects (ZY18164003). The funding bodies didn’t include in the design of the study nor in collection, analysis, and interpretation of data, and also not in writing of the manuscript.

Acknowledgments

We thank the anonymous for the smallholders for their supports on the buffalo samples. We also
would like to thank the anonymous reviewers for their constructive comments.

Authors' contributions
TXD designed the study and analyzed the data; TXD, XRL, and AQD wrote the drafted manuscript; SSL, XRL, XYM, and CYP performed the experiments and qRT-PCR analysis; WXL helped to draft the manuscript; All authors read and approved this paper.

Author details
1Guangxi provincial Key Laboratory of Buffalo Genetics, Breeding and Reproduction technology, Buffalo Research Institute, Chinese Academy of Agricultural Sciences, Nanning530001, China

Reference
1. Macias H, Hinck L: Mammary gland development. *Wiley Interdisciplinary Reviews Developmental Biology* 2012, 1(4):533-557.

2. Inman JL, Robertson C, Mott JD, Bissell MJ: Mammary gland development: cell fate specification, stem cells and the microenvironment. *Development* 2015, 142(6):1028.

3. Sharma N, Dong KJ: Stem Cell Research: A Novel Boulevard towards Improved Bovine Mastitis Management. *International Journal of Biological Sciences* 2013, 9(8):818-829.

4. Rozario T, Desimone DW: The extracellular matrix in development and morphogenesis: a dynamic view. *Developmental Biology* 2010, 341(1):126-140.

5. Morrissey MA, Sherwood DR: An active role for basement membrane assembly and modification in tissue sculpting. *Journal of Cell Science* 2015, 128(9):1661.

6. Bonnans C, Chou J, Werb Z: Remodelling the extracellular matrix in development and disease. *Nat Rev Mol Cell Biol* 2014, 15(12):786-801.

7. Schedin P, Keely PJ: Mammary Gland ECM Remodeling, Stiffness, and Mechanosignaling in Normal Development and Tumor Progression. *Cold Spring Harbor Perspectives in Biology* 2011, 3(1):a003228.

8. Miller Ej, Matukas VJ: Chick cartilage collagen: A new type of 1 chain not present in bone or skin of the species. *Proceedings of the National Academy of Sciences of the United States of America* 1969, 64(4):1264-1268.

9. Ricardblum S: The collagen family. *Cold Spring Harbor Perspectives in Biology* 2010, 3(1):a004978.

10. Gelse K, Pöschl E, Aigner T: Collagens—structure, function, and biosynthesis. *Advanced Drug
11. Gordon MK, Hahn RA: Collagens. *Cell & Tissue Research* 2010, 339(1):247.

12. Chen Q, He G, Zhang W, Xu T, Qi H, Li J, Zhang Y, Gao MQ: Stromal fibroblasts derived from mammary gland of bovine with mastitis display inflammation-specific changes. *Scientific Reports* 2016, 6:27462.

13. Crisà A, Ferrè F, Chillemi G, Moioli B: RNA-sequencing for profiling goat milk transcriptome in colostrum and mature milk. *Bmc Veterinary Research* 2016, 12(1):264.

14. Heino J, Huhtala M, Käpylä J, Johnson MS: Evolution of collagen-based adhesion systems. *International Journal of Biochemistry & Cell Biology* 2009, 41(2):341-348.

15. Humphries JD, Byron A, Humphries MJ: Integrin ligands at a glance. *Journal of Cell Science* 2006, 119(19):3901-3903.

16. Leitinger B, Hohenester E: Mammalian collagen receptors. *Matrix Biology* 2007, 26(3):146-155.

17. Wu JJ, Song LJ, Wu FJ, Liang XW, Yang BZ, Wathes DC, Pollott GE, Cheng Z, Shi DS, Liu QY: Investigation of transferability of BovineSNP50 BeadChip from cattle to water buffalo for genome wide association study. *Molecular Biology Reports* 2013, 40(2):743-750.

18. Khedkar CD, Kalyankar SD, Deosarkar SS: Buffalo Milk. *Encyclopedia of Food & Health* 2016:522-528.

19. Iamartino D, Nicolazzi EL, Van Tassell CP, Reecy JM, Fritz-Waters ER, Koltes JE, Biffani S, Sonstegard TS, Schroeder SG, Ajmone-Marsan P: Design and validation of a 90K SNP genotyping assay for the water buffalo (Bubalus bubalis). *PloS one* 2017, 12(10):e0185220.

20. Camargo GD, Aspilcuetaborquis RR, Fortes M, Portoneto R, Cardoso DF, Santos D, Lehnert SA, Reverter A, Moore SS, Tonhati H: Prospecting major genes in dairy buffaloes. *BMC Genomics* 2015, 16(1):1-14.

21. Strazzullo M, Gasparrini B, Neglia G, Balestrieri ML, Francioso R, Rossetti C, Nassa G, De MF, Weisz A, Di SF: Global transcriptome profiles of Italian Mediterranean buffalo embryos with normal and retarded growth. *Plos One* 2014, 9(2):e90027.

22. Chen F, Fu Q, Pu L, Zhang P, Huang Y, Hou Z, Xu Z, Chen D, Huang F, Deng T: Integrated analysis
of quantitative proteome and transcriptional profiles reveals the dynamic function of maternally-expressed proteins after parthenogenetic activation of buffalo oocyte. *Molecular & Cellular Proteomics* 2018:mcp. RA118. 000556.

23. Low WY, Tearle R, Bickhart DM, Rosen BD, Kingan SB, Swale T, Thibaud-Nissen F, Murphy TD, Young R, Lefèvre L: Chromosome-level assembly of the water buffalo genome surpasses human and goat genomes in sequence contiguity. *Nature communications* 2019, 10(1):260.

24. Liu Q, Dang H, Chen Z, Wu J, Chen Y, Chen S, Luo L: Genome-Wide Identification, Expression, and Functional Analysis of the Sugar Transporter Gene Family in Cassava (*Manihot esculenta*). *International Journal of Molecular Sciences* 2018, 19(4):987.

25. Ou W, Mao X, Huang C, Tie W, Yan Y, Ding Z, Wu C, Xia Z, Wang W, Zhou S: Genome-Wide Identification and Expression Analysis of the KUP Family under Abiotic Stress in Cassava (*Manihot esculenta Crantz*). *Frontiers in Physiology* 2018, 9:17.

26. Wang X, Shi X, Chen S, Ma C, Xu S: Evolutionary Origin, Gradual Accumulation and Functional Divergence of Heat Shock Factor Gene Family with Plant Evolution. *Frontiers in Plant Science* 2018, 9.

27. Körbes AP, Kulcheski FR, Margis R, Margis-Pinheiro M, Turchetto-Zolet AC: Molecular evolution of the lysophosphatidic acid acyltransferase (LPAAT) gene family. *Molecular Phylogenetics & Evolution* 2016, 96:55-69.

28. Hu G, Li L, Xu W: Extracellular matrix in mammary gland development and breast cancer progression. *Frontiers in Laboratory Medicine* 2017, 1(1):36-39.

29. Ko H, Kim S, Yang K, Kim K: Phosphorylation-dependent stabilization of MZF1 upregulates N-cadherin expression during protein kinase CK2-mediated epithelial-mesenchymal transition. *Oncogenesis* 2018, 7(3):27.

30. Quaranta V: Cell migration through extracellular matrix: membrane-type metalloproteinases make the way. *The Journal of cell biology* 2000, 149(6):1167-1170.

31. Sander M, Neubüser A, Kalamaras J, Ee HC, Martin GR, German MS: Genetic analysis reveals that PAX6 is required for normal transcription of pancreatic hormone genes and islet development. *Genes & development* 1997, 11(13):1662-1673.
32. Zhang X, Rowan S, Yue Y, Heaney S, Pan Y, Brendolan A, Selleri L, Maas RL: Pax6 is regulated by Meis and Pbx homeoproteins during pancreatic development. Developmental biology 2006, 300(2):748-757.
33. Cohick WS: PHYSIOLOGY AND ENDOCRINOLOGY SYMPOSIUM: Effects of insulin on mammary gland differentiation during pregnancy and lactation. Journal of Animal Science 2016, 94(5):1812.
34. Laarveld B, Christensen DA, Brockman RP: The effect of insulin on net metabolism of glucose and amino acids by the bovine mammary gland. Endocrinology 1981, 108(6):2217.
35. Zhu J, Xiong G, Trinkle C, Xu R: Integrated extracellular matrix signaling in mammary gland development and breast cancer progression. Histology and histopathology 2014, 29(9):1083.
36. ZNF384 zinc finger protein 384 [ Homo sapiens (human) ] [https://www.ncbi.nlm.nih.gov/gene?Db=gene&Cmd=ShowDetailView&TermToSearch=171017]
37. Hughes AL: The Evolution of Functionally Novel Proteins after Gene Duplication. Proceedings Biological Sciences 1994, 256(1346):119-124.
38. Freeling M: Bias in plant gene content following different sorts of duplication: tandem, whole-genome, segmental, or by transposition. Annual Review of Plant Biology 2009, 60(60):433-453.
39. Liu GE, Ventura M, Cellamare A, Chen L, Cheng Z, Zhu B, Li C, Song J, Eichler EE: Analysis of recent segmental duplications in the bovine genome. BMC Genomics 2009, 10(1):571.
40. Feng X, Jiang J, Padhi A, Ning C, Fu J, Wang A, Mrode R, Liu JF: Characterization of genome-wide segmental duplications reveals a common genomic feature of association with immunity among domestic animals. Bmc Genomics 2017, 18(1):293.
41. Zhao P, Wang D, Wang R, Kong N, Zhang C, Yang C, Wu W, Ma H, Chen Q: Genome-wide analysis of the potato Hsp20 gene family: identification, genomic organization and expression profiles in response to heat stress. BMC Genomics 2018, 19(1):61.
42. Kukurba KR, Montgomery SB: RNA sequencing and analysis. Cold Spring Harbor Protocols 2015, 2015(11):951-969.
43. Ghajar CM, Bissell MJH, biology c: Extracellular matrix control of mammary gland morphogenesis and tumorigenesis: insights from imaging. 2008, 130(6):1105.
44. Zhu J, Xiong G, Trinkle C, Xu RJH, histopathology: Integrated extracellular matrix signaling in mammary gland development and breast cancer progression. 2014, 29(9):1083.

45. Dai W-t, Zou Y-x, White RR, Liu J-x, Liu H-y: Transcriptomic profiles of the bovine mammary gland during lactation and the dry period. *Functional & integrative genomics* 2018, 18(2):125-140.

46. Schmitz S, Pfaffl MW, Meyer HHD, Bruckmaier RM: Short-term changes of mRNA expression of various inflammatory factors and milk proteins in mammary tissue during LPS-induced mastitis. *Domestic Animal Endocrinology* 2004, 26(2):111-126.

47. Finn RD, Clements J, Arndt W, Miller BL, Wheeler TJ, Schreiber F, Bateman A, Eddy SR: HMMER web server: 2015 update. *Nucleic Acids Research* 2015, 43(1):30-38.

48. Finn RD, Clements J, Eddy SR: HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research* 2011, 39(Web Server issue):29-37.

49. Kumar S, Stecher G, Tamura K: MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular Biology & Evolution* 2016, 33(7):1870.

50. Chen C, Xia R, Chen H, He Y: TBtools, a Toolkit for Biologists integrating various HTS-data handling tools with a user-friendly interface. 2018.

51. Tan G, Lenhard B: TFBSTools: an R/bioconductor package for transcription factor binding site analysis. *Bioinformatics* 2016, 32(10):1555-1556.

52. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, Lee T, Jin H, Marler B, Guo H: MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Research* 2012, 40(7):e49-e49.

53. Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, Sánchez-González: DnaSP 6: DNA Sequence Polymorphism Analysis of Large Datasets. *Molecular Biology & Evolution* 2017, 34(12).

54. Kim D, Langmead B, Salzberg SL: HISAT: a fast spliced aligner with low memory requirements. *Nature Methods* 2015, 12(4):357-360.

55. Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL: StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology* 2015,
33(3):290-295.

56. Kolde R: pheatmap: Pretty Heatmaps. *R package version* 2012, 61:926.

57. Livak KJ, Schmittgen TDJm: Analysis of relative gene expression data using real-time quantitative PCR and the 2− ΔΔCT method. 2001, 25(4):402-408.

Additional File Legends

Additional file 1: Table S1. Primers of qRT-PCR for collagen genes in buffalo

Additional file 2: Figure S1. Information on intron numbers of the collagen family in studied species.

Additional file 3: Figure S2. Summary of the TFBS detected in the buffalo collagen genes.

Additional file 4: Figure S3. Summary of the TFBS detected in the buffalo collagen subfamilies.

Tables

Table 1. Features of the identified collagens protein sequences in buffalo

| Symbol  | Name       | GeneID     | ProteinID    | Protein(aa) | CDS(bp) | MW(kDa)  | pl    |
|---------|------------|------------|--------------|-------------|---------|----------|-------|
| COL10A1 | COL10A1    | 102389861  | XP 025150    | 2067        | 688     | 167.822  | 4.87  |
|         |            |            | 531.1        |             |         |          |       |
| COL11A1 | COL11A1    | 102411572  | XP 025143    | 5415        | 1804    | 435.856  | 4.71  |
|         |            |            | 983.1        |             |         |          |       |
| COL11A2 | COL11A2_1  | 102395601  | XP 025127    | 4029        | 1342    | 320.1372 | 4.71  |
|         |            |            | 671.1        |             |         |          |       |
| COL11A2 | COL11A2_2  | 102395601  | XP 025127    | 4029        | 1342    | 320.1372 | 4.71  |
|         |            |            | 672.1        |             |         |          |       |
| COL11A2 | COL11A2_3  | 102395601  | XP 025127    | 4029        | 1342    | 320.1372 | 4.71  |
|         |            |            | 673.1        |             |         |          |       |
| COL11A2 | COL11A2_4  | 102395601  | XP 025127    | 4968        | 1655    | 397.8534 | 4.66  |
|         |            |            | 670.1        |             |         |          |       |
| COL11A2 | COL11A2_5  | 102395601  | XP 006050    | 4890        | 1629    | 391.0736 | 4.66  |
|         |            |            | 365.1        |             |         |          |       |
| COL11A2 | COL11A2_6  | 102395601  | XP 006050    | 4953        | 1650    | 396.1965 | 4.66  |
|         |            |            | 364.1        |             |         |          |       |
| COL11A2 | COL11A2_7  | 102395601  | XP 025127    | 5031        | 1676    | 402.9763 | 4.66  |
|         |            |            | 669.1        |             |         |          |       |
| COL11A2 | COL11A2_8  | 102395601  | XP 006050    | 5211        | 1736    | 417.553  | 4.65  |
|         |            |            | 363.1        |             |         |          |       |
| COL11A2 | COL11A2_9  | 102395601  | XP 025127    | 5133        | 1710    | 410.7732 | 4.65  |
|         |            |            | 668.1        |             |         |          |       |
| COL12A1 | COL12A1_1  | 102397550  | XP 006062    | 5871        | 1956    | 481.867  | 4.66  |
|         |            |            | 066.1        |             |         |          |       |
| COL12A1 | COL12A1_2  | 102397550  | XP 006062    | 9198        | 3065    | 753.4744 | 4.58  |
|         |            |            | 065.1        |             |         |          |       |
| COL12A1 | COL12A1_3  | 102397550  | XP 006062    | 9363        | 3120    | 766.9291 | 4.58  |
|         |            |            | 064.1        |             |         |          |       |
| COL13A1 | COL13A1_1  | 102391074  | XP 025140    | 2163        | 720     | 169.9635 | 4.86  |
|         |            |            | 452.1        |             |         |          |       |
| COL13A1_1 | 102391074 | XP_025140 | 2349 | 782 | 187.4364 | 4.83 |
|---|---|---|---|---|---|---|
| COL13A1_1 | 102391074 | XP_025140 | 2076 | 691 | 163.4114 | 4.87 |
| COL13A1_1 | 102391074 | XP_025140 | 2091 | 696 | 164.273 | 4.87 |
| COL13A1_1 | 102391074 | XP_025140 | 2100 | 699 | 165.2032 | 4.87 |
| COL13A1_1 | 102391074 | XP_025140 | 187.4364 | 4.83 |
| COL13A1_1 | 102391074 | XP_025140 | 163.4114 | 4.87 |
| COL13A1_1 | 102391074 | XP_025140 | 164.273 | 4.87 |
| COL13A1_1 | 102391074 | XP_025140 | 165.2032 | 4.87 |
| COL13A1_1 | 102391074 | XP_025140 | 187.4364 | 4.83 |
| COL13A1_1 | 102391074 | XP_025140 | 163.4114 | 4.87 |
| COL13A1_1 | 102391074 | XP_025140 | 164.273 | 4.87 |
| COL13A1_1 | 102391074 | XP_025140 | 165.2032 | 4.87 |
| COL13A1_1 | 102391074 | XP_025140 | 187.4364 | 4.83 |
| COL13A1_1 | 102391074 | XP_025140 | 163.4114 | 4.87 |
| COL13A1_1 | 102391074 | XP_025140 | 164.273 | 4.87 |
| COL13A1_1 | 102391074 | XP_025140 | 165.2032 | 4.87 |
| COL13A1_1 | 102391074 | XP_025140 | 187.4364 | 4.83 |
| COL13A1_1 | 102391074 | XP_025140 | 163.4114 | 4.87 |
| COL13A1_1 | 102391074 | XP_025140 | 164.273 | 4.87 |
| COL13A1_1 | 102391074 | XP_025140 | 165.2032 | 4.87 |
| COL13A1_1 | 102391074 | XP_025140 | 187.4364 | 4.83 |
| COL13A1_1 | 102391074 | XP_025140 | 163.4114 | 4.87 |
| COL13A1_1 | 102391074 | XP_025140 | 164.273 | 4.87 |
| COL13A1_1 | 102391074 | XP_025140 | 165.2032 | 4.87 |
| COL13A1_1 | 102391074 | XP_025140 | 187.4364 | 4.83 |
| COL13A1_1 | 102391074 | XP_025140 | 163.4114 | 4.87 |
| COL13A1_1 | 102391074 | XP_025140 | 164.273 | 4.87 |
| COL13A1_1 | 102391074 | XP_025140 | 165.2032 | 4.87 |
| COL13A1_1 | 102391074 | XP_025140 | 187.4364 | 4.83 |
| COL13A1_1 | 102391074 | XP_025140 | 163.4114 | 4.87 |
| COL13A1_1 | 102391074 | XP_025140 | 164.273 | 4.87 |
| COL13A1_1 | 102391074 | XP_025140 | 165.2032 | 4.87 |
| COL13A1_1 | 102391074 | XP_025140 | 187.4364 | 4.83 |
| COL13A1_1 | 102391074 | XP_025140 | 163.4114 | 4.87 |
| COL13A1_1 | 102391074 | XP_025140 | 164.273 | 4.87 |
| COL13A1_1 | 102391074 | XP_025140 | 165.2032 | 4.87 |
| COL13A1_1 | 102391074 | XP_025140 | 187.4364 | 4.83 |
| COL13A1_1 | 102391074 | XP_025140 | 163.4114 | 4.87 |
| COL13A1_1 | 102391074 | XP_025140 | 164.273 | 4.87 |
| COL13A1_1 | 102391074 | XP_025140 | 165.2032 | 4.87 |
| COL13A1_1 | 102391074 | XP_025140 | 187.4364 | 4.83 |
| COL13A1_1 | 102391074 | XP_025140 | 163.4114 | 4.87 |
| COL13A1_1 | 102391074 | XP_025140 | 164.273 | 4.87 |
| COL13A1_1 | 102391074 | XP_025140 | 165.2032 | 4.87 |
| COL13A1_1 | 102391074 | XP_025140 | 187.4364 | 4.83 |
| COL13A1_1 | 102391074 | XP_025140 | 163.4114 | 4.87 |
| COL13A1_1 | 102391074 | XP_025140 | 164.273 | 4.87 |
| COL13A1_1 | 102391074 | XP_025140 | 165.2032 | 4.87 |
| COL13A1_1 | 102391074 | XP_025140 | 187.4364 | 4.83 |

---

| COL14A1_1 | 102389728 | XP_006074 | 5370 | 1789 | 438.6139 | 4.7 |
|---|---|---|---|---|---|---|
| COL14A1_2 | 102389728 | XP_006074 | 5418 | 1805 | 442.4683 | 4.69 |
| COL15A1_1 | 102398398 | XP_025137 | 4035 | 1344 | 331.1 | 4.72 |
| COL15A1_2 | 102398398 | XP_025137 | 4119 | 1372 | 337.7977 | 4.72 |
| COL15A1_3 | 102398398 | XP_025137 | 4131 | 1376 | 338.917 | 4.72 |
| COL15A1_4 | 102398398 | XP_025137 | 4131 | 1376 | 338.917 | 4.72 |
| COL15A1_5 | 102398398 | XP_025137 | 4065 | 1354 | 333.4969 | 4.72 |
| COL15A1_6 | 102398398 | XP_025137 | 4065 | 1354 | 333.1664 | 4.72 |
| COL16A1_1 | 102397619 | XP_025134 | 4878 | 1625 | 390.0826 | 4.68 |
| COL16A1_2 | 102397619 | XP_025134 | 4899 | 1632 | 391.9759 | 4.68 |
| COL16A1_3 | 102397619 | XP_025134 | 4851 | 1616 | 388.1235 | 4.68 |
| Gene    | Description | Chromosome | Gene ID  | Transcript ID  | Length | Start   | End     | ORF Length | ORF Score | Coverage |
|---------|-------------|------------|----------|---------------|--------|---------|---------|------------|-----------|----------|
| COL16A1_4 |            |            | COL16A1_4 | XP_025134     | 4833   | 1610    | 386.131 | 4.68       |           |          |
| COL16A1_5 |            |            | COL16A1_5 | XP_025134     | 4854   | 1617    | 388.024 | 4.68       |           |          |
| COL16A1_6 |            |            | COL16A1_6 | XP_025134     | 4809   | 1602    | 384.208 | 4.68       |           |          |
| COL17A1  |            |            | COL17A1_1 | XP_006067     | 4458   | 1485    | 365.688 | 4.69       |           |          |
| COL17A1  |            |            | COL17A1_2 | XP_006067     | 4347   | 1448    | 356.340 | 4.69       |           |          |
| COL18A1  |            |            | COL18A1_1 | XP_025148     | 3990   | 1329    | 323.492 | 4.69       |           |          |
| COL18A1  |            |            | COL18A1_2 | XP_006052     | 4614   | 1537    | 376.122 | 4.65       |           |          |
| COL18A1  |            |            | COL18A1_3 | XP_025148     | 4641   | 1546    | 378.307 | 4.65       |           |          |
| COL18A1  |            |            | COL18A1_4 | XP_025148     | 5355   | 1784    | 439.237 | 4.61       |           |          |
| COL18A1  |            |            | COL18A1_5 | XP_025148     | 5358   | 1785    | 439.468 | 4.61       |           |          |
| COL18A1  |            |            | COL18A1_6 | XP_025148     | 5382   | 1793    | 441.422 | 4.61       |           |          |
| COL19A1  |            |            | COL19A1_1 | XP_025150     | 3396   | 1131    | 271.476 | 4.8        |           |          |
| COL19A1  |            |            | COL19A1_2 | XP_025150     | 3342   | 1113    | 267.366 | 4.8        |           |          |
| COL20A1  |            |            | COL20A1   | XP_025119     | 4392   | 1463    | 363.036 | 4.67       |           |          |
| COL21A1  |            |            | COL21A1   | XP_025128     | 2871   | 956     | 231.076 | 4.86       |           |          |
| COL22A1  |            |            | COL22A1_1 | XP_025121     | 4845   | 1614    | 387.200 | 4.67       |           |          |
| COL22A1  |            |            | COL22A1_2 | XP_025121     | 4476   | 1491    | 357.010 | 4.69       |           |          |
| COL22A1  |            |            | COL22A1_3 | XP_025121     | 4350   | 1449    | 346.977 | 4.69       |           |          |
| COL23A1  |            |            | COL23A1   | XP_025148     | 1557   | 518     | 122.944 | 4.92       |           |          |
| COL24A1  |            |            | COL24A1_1 | XP_025144     | 5130   | 1709    | 412.848 | 4.74       |           |          |
| COL24A1  |            |            | COL24A1_2 | XP_025144     | 5277   | 1758    | 425.274 | 4.74       |           |          |
| COL24A1  |            |            | COL24A1_3 | XP_025144     | 5277   | 1758    | 425.274 | 4.74       |           |          |
| COL25A1  |            |            | COL25A1_1 | XP_025146     | 2001   | 666     | 157.810 | 4.91       |           |          |
| COL25A1  |            |            | COL25A1_2 | XP_025146     | 1974   | 657     | 155.769 | 4.91       |           |          |
| COL25A1_3 | 102390242 | XP 025146 097.1 | 1965 | 654 | 154.995 | 4.91 |
| COL25A1_4 | 102390242 | XP 025146 098.1 | 1956 | 651 | 154.1831 | 4.91 |
| COL25A1_5 | 102390242 | XP 025146 099.1 | 1938 | 645 | 152.9538 | 4.91 |
| COL26A1_1 | 102389806 | XP 006045 472.1 | 1317 | 438 | 107.4612 | 4.92 |
| COL26A1_2 | 102389806 | XP 006045 471.1 | 1323 | 440 | 107.9657 | 4.92 |
| COL27A1_1 | 102415572 | XP 025138 105.1 | 5469 | 1822 | 107.4612 | 4.92 |
| COL27A1_2 | 102415572 | XP 006061 277.1 | 5523 | 1840 | 448.1909 | 4.63 |
| COL27A1_3 | 102415572 | XP 025138 102.1 | 5526 | 1841 | 447.8854 | 4.63 |
| COL27A1_4 | 102415572 | XP 025138 099.1 | 5529 | 1842 | 448.1026 | 4.63 |
| COL27A1_5 | 102415572 | XP 025138 104.1 | 5475 | 1824 | 443.9739 | 4.63 |
| COL27A1_6 | 102415572 | XP 025138 103.1 | 5475 | 1824 | 443.8918 | 4.63 |
| COL28A1_1 | 102407731 | XP 006074 647.1 | 3444 | 1147 | 276.3516 | 4.81 |
| COL2A1_1 | 102403678 | XP 006059 500.1 | 4464 | 1487 | 364.1477 | 4.68 |
| COL2A1_2 | 102403678 | XP 006059 501.1 | 4257 | 1418 | 347.1815 | 4.69 |
| COL3A1_1 | 102394969 | XP 006073 767.2 | 4401 | 1466 | 357.2321 | 4.72 |
| COL4A1_1 | 102405654 | XP 025118 865.1 | 5010 | 1669 | 401.5158 | 4.66 |
| COL4A2_1 | 102403266 | XP 025118 864.1 | 5133 | 1710 | 411.3817 | 4.65 |
| COL4A3_1 | 102416276 | XP 006047 187.1 | 5013 | 1670 | 401.8 | 4.7 |
| COL4A3_2 | 102416276 | XP 025117 915.1 | 4296 | 1431 | 342.0154 | 4.73 |
| COL4A3_3 | 102416276 | XP 006047 188.1 | 4614 | 1537 | 369.6487 | 4.72 |
| COL4A4_1 | 102389047 | XP 025117 873.1 | 4950 | 1649 | 397.2704 | 4.69 |
| COL4A4_2 | 102389047 | XP 006047 189.1 | 5070 | 1689 | 406.8914 | 4.69 |
| COL4A4_3 | 102389047 | XP 006047 190.1 | 5070 | 1689 | 406.8914 | 4.69 |
| COL4A4_4 | 102389047 | XP 006047 191.1 | 5070 | 1689 | 406.8914 | 4.69 |
| COL4A4_5 | 102389047 | XP 006047 192.1 | 5070 | 1689 | 406.8914 | 4.69 |
| COL4A4_6 | 102389047 | XP 006047 193.1 | 5070 | 1689 | 406.8914 | 4.69 |
| COL4A4_7   | 102389047 | XP_006047 197.1 | 4254 | 1417 | 339.8506 | 4.73 |
| COL4A4_8   | 102389047 | XP_006047 198.1 | 4110 | 1369 | 328.1391 | 4.74 |
| COL4A5     | COL4A5_1  | 102399306 | XP_006072 403.1 | 5085 | 1694 | 414.9557 | 4.69 |
| COL4A5_2   | 102399306 | XP_006072 404.1 | 5076 | 1691 | 414.1878 | 4.69 |
| COL4A5_3   | 102399306 | XP_006072 408.1 | 5058 | 1685 | 412.6561 | 4.69 |
| COL4A5_4   | 102399306 | XP_006072 405.1 | 5076 | 1691 | 414.1878 | 4.69 |
| COL4A5_5   | 102399306 | XP_025131 953.1 | 5076 | 1691 | 414.1918 | 4.69 |
| COL4A5_6   | 102399306 | XP_006072 407.1 | 5067 | 1688 | 413.424  | 4.69 |
| COL4A6     | COL4A6    | 102410702 | XP_025131 955.1 | 5085 | 1694 | 412.8158 | 4.68 |
| COL5A1     | COL5A1    | 102414254 | XP_025118 027.1 | 5541 | 1846 | 447.3157 | 4.62 |
| COL5A2     | COL5A2    | 102395275 | XP_006073 768.1 | 4500 | 1499 | 361.1947 | 4.74 |
| COL5A3_1   | COL5A3_1  | 102407916 | XP_006044 492.1 | 5241 | 1746 | 420.8161 | 4.66 |
| COL5A3_2   | COL5A3_2  | 102407916 | XP_006044 493.1 | 5238 | 1745 | 420.5848 | 4.66 |
| COL5A3_3   | COL5A3_3  | 102407916 | XP_025149 345.1 | 4881 | 1626 | 392.9009 | 4.68 |
| COL6A1     | COL6A1    | 102389816 | NP_001277 797.1 | 3084 | 1027 | 248.0209 | 4.77 |
| COL6A2     | COL6A2_1  | 102397245 | XP_025148 825.1 | 2781 | 926 | 224.0028 | 4.79 |
| COL6A2_2   | COL6A2_2  | 102397245 | XP_006047 111.1 | 3084 | 1027 | 248.74  | 4.77 |
| COL6A2_3   | COL6A2_3  | 102397245 | XP_025148 817.1 | 3084 | 1027 | 248.74  | 4.77 |
| COL6A3     | COL6A3    | 102392465 | XP_025144 989.1 | 9654 | 3217 | 786.5997 | 4.54 |
| COL6A4     | COL6A4    | 102391810 | XP_025127 254.1 | 6504 | 2167 | 526.7032 | 4.66 |
| COL6A5     | COL6A5    | 102394387 | XP_025149 876.1 | 7821 | 2606 | 636.0357 | 4.64 |
| COL6A6_1   | COL6A6_1  | 102393741 | XP_006043 497.1 | 6813 | 2270 | 553.3681 | 4.64 |
| COL6A6_2   | COL6A6_2  | 102393741 | XP_025149 867.1 | 6813 | 2270 | 553.3681 | 4.64 |
| COL6A6_3   | COL6A6_3  | 102393741 | XP_006043 498.1 | 6801 | 2266 | 552.4591 | 4.64 |
| COL7A1     | COL7A1    | 102405985 | XP_006047 225.2 | 8793 | 2930 | 703.4311 | 4.54 |
| COL8A1     | COL8A1    | 102398637 | XP_006047 2238 | 745 | 181.1666 | 4.84 |
Table 2. Ten different motifs commonly observed in buffalo collagen family

| Motif   | Protein sequence                        | Length | Pfam domain |
|---------|-----------------------------------------|--------|-------------|
| MEME-1  | GLPGLKGEKGCAGLPGFPGKGEK                  | 29     | -           |
| MEME-2  | KGEDGLPGPLLPGKGEKGEK                     | 29     | -           |
| MEME-3  | GEKGERGLPGKKGKAKGEPGAKGEKPGGGPGPGE       | 41     | Collagen    |
| MEME-4  | GPPGPPPGPSPGPPGLPGGP                    | 30     | -           |
| MEME-5  | PGPPGPKGPGKGEKDPGSTGPGPGLPGPGEK         | 41     | Collagen    |
| MEME-6  | GPKERGPKGQKGEKGP                        | 21     | -           |
| MEME-7  | TGPPGPGPGLPGPKGEK                      | 21     | -           |
| MEME-8  | GEPGAPGKGEKPGGPGPGLPGKPGK              | 29     | -           |
| MEME-9  | GEQGERPKEKGEA                           | 15     | -           |
| MEME-10 | RGEPGLPGPGP                               | 15     | -           |

Table 3. Analysis of the Ka/Ks ratios for each pair of duplicated collagen genes in buffalo
| Gene pairs         | CHR | Ka   | Ks   | Ka/Ks ratio |
|--------------------|-----|------|------|-------------|
| COL11A2_4/COL11A_2  | 2   | 0.0093 | 0.0095 | 0.9805 |
| COL12A1_2/COL12A_1  | 10  | 0.0005 | 0.0014 | 0.3362 |
| COL12A1_3/COL12A_2  | 10  | 0.0003 | 0.0009 | 0.3327 |
| COL13A1_2/COL13A_1  | 4   | 0.0158 | 0.0182 | 0.8665 |
| COL13A1_3/COL13A_1  | 4   | 0     | 0.0017 | 0     |
| COL13A1_10/COL13A_1  | 4   | 0     | 0.0017 | 0     |
| COL14A1_2/COL14A_1  | 15  | 0.0035 | 0.0008 | 4.3743 |
| COL17A1_2/COL17A_1  | 23  | 0.0003 | 0.0018 | 0.1825 |
| COL22A1_1/COL22A_1  | 15  | 0.0112 | 0.0131 | 0.8518 |
| COL22A1_2/COL22A_1  | 15  | 0.0033 | 0.0038 | 0.8753 |
| COL27A1_5/COL27A_1  | 3   | 0     | 0.0021 | 0     |
| COL4A2/COL4A1_1     | X   | 0.5336 | 1.9459 | 0.2742 |
| COL4A3_1/COL4A3_1   | X   | 0.0029 | 0.0086 | 0.3357 |
| COL4A3_2/COL4A3_2   | X   | 0.0339 | 0.0586 | 0.5781 |
| COL4A4_2/COL4A4_2   | X   | 0.0061 | 0.0073 | 0.8406 |
| COL5A3_1/COL5A3_1   | 9   | 0     | 0.0007 | 0     |
| COL5A3_2/COL5A3_2   | 9   | 0     | 0.0007 | 0     |
| COL6A1/COL6A2_1     | 1   | 0.7068 | 1.0659 | 0.6631 |
| COL6A2_1/COL6A2_1   | 1   | 0.0547 | 0.0696 | 0.7861 |
| COL6A2_2/COL6A2_1   | 1   | 0.0547 | 0.0696 | 0.7861 |
| COL6A6_2/COL6A5_1   | 1   | 0.3759 | 1.4319 | 0.2625 |
| COL6A6_3/COL6A6_1   | 1   | 0.0002 | 0.0006 | 0.3146 |
| COL9A1/COL19A1_2    | 6   | 0.6218 | 2.6177 | 0.2375 |
| COL9A2_2/COL9A2_1   | 6   | 0.0239 | 0.0263 | 0.9098 |

Note. CHR=Chromosome of buffalo
Figures

Figure 1
Phylogenetic relationship of collagen proteins in 6 mammals. Different capitals represent different species: B means buffalo, C means cattle, G means goat, S means sheep, H means horse, and Hu mean human; Line with different colors indicates different subfamilies.

Figure 2
Phylogenetic relationships (A), gene structure (B), and conserved protein motifs (C) of buffalo collagen family. The different-colored boxes indicate different domains or motifs of the collagen family in buffalo.

Figure 3
TFBS analysis of the promoter region in buffalo collagen. (A) The phylogenetic tree of buffalo collagen genes was constructed, and different colors indicated different clusters. (B) The predicted TFBS was displayed in the promoter region of buffalo collagen genes. The different-colored boxes indicate different TFBS.

Figure 4
Chromosome location and gene duplication of buffalo collagen sequences. The tandem duplicated genes were marked by the red line, and segmentally duplicated genes are indicated by symbol #.

Figure 5
Tissue expression distribution of buffalo collagen using RNA-seq data and qRT-PCR. (A) tissue expression distribution of buffalo collagen genes based on the transcriptomic data. (B) tissue expression pattern of buffalo collagen genes based on the qRT-PCR.
Figure 6

Cluster analysis of buffalo collagen in mammary gland tissue at different lactation points.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

Supplementary material.pdf
NC3Rs ARRIVE Guidelines Checklist 2014.docx