The Evolution of Milk Casein Genes from Tooth Genes before the Origin of Mammals

Kazuhiko Kawasaki,*[,1] Anne-Gaelle Lafont,[2] and Jean-Yves Sire[2]

[1]Department of Anthropology, Pennsylvania State University

[2]UMR 7138-Systématique-Adaptation-Evolution, Université Pierre et Marie Curie, Paris, France

***Corresponding author:** E-mail: kuk2@psu.edu.

Associate editor: Yoko Satta

Abstract

Caseins are among cardinal proteins that evolved in the lineage leading to mammals. In milk, caseins and calcium phosphate (CaP) form a huge complex called casein micelle. By forming the micelle, milk maintains high CaP concentrations, which help altricial mammalian neonates to grow bone and teeth. Two types of caseins are known. Ca-sensitive caseins (α_s- and β-caseins) bind Ca but precipitate at high Ca concentrations, whereas Ca-insensitive casein (κ-casein) does not usually interact with Ca but instead stabilizes the micelle. Thus, it is thought that these two types of caseins are both necessary for stable micelle formation. Both types of caseins show high substitution rates, which make it difficult to elucidate the evolution of caseins. Yet, recent studies have revealed that all casein genes belong to the secretory calcium-binding phosphoprotein (SCPP) gene family that arose by gene duplication. In the present study, we investigated exon–intron structures and phylogenetic distributions of casein and other SCPP genes, particularly the odontogenic ameloblast–associated (*ODAM*) gene, the SCPP-Pro-Gln-rich 1 (*SCPPPQ1*) gene, and the follicular dendritic cell secreted peptide (*FDCSP*) gene. The results suggest that contemporary Ca-sensitive casein genes arose from a putative common ancestor, which we refer to as *CSN1/2*. The six putative exons comprising *CSN1/2* are all found in *SCPPPQ1*, although *ODAM* also shares four of these exons. By contrast, the five exons of the Ca-insensitive casein gene are all reminiscent of *FDCSP*. The phylogenetic distribution of these genes suggests that both *SCPPPQ1* and *FDCSP* arose from *ODAM*. We thus argue that all casein genes evolved from *ODAM* via two different pathways; Ca-sensitive casein genes likely originated directly from *SCPPPQ1*, whereas the Ca-insensitive casein genes directly differentiated from *FDCSP*. Further, expression of *ODAM*, *SCPPPQ1*, and *FDCSP* was detected in dental tissues, supporting the idea that both types of caseins evolved as Ca-binding proteins. Based on these findings, we propose two alternative hypotheses for micelle formation in primitive milk. The conserved biochemical characteristics in caseins and their immediate ancestors also suggest that many slight genetic modifications have created modern caseins, proteins vital to the sustained success of mammals.

Key words: gene duplication, gene family, SCPP, lactation, casein micelle, enamel.

Introduction

Early amniotes split into synapsids and sauropsids in the Mid-Carboniferous. Sauropsids led to all modern reptiles and birds, whereas mammals arose in the synapsid lineage. Changes in bone and tooth morphology in early synapsids suggest gradual evolution of many mammalian characters (Benton 2005). It has been argued that lactation also evolved gradually in early synapsids and that the transformation of proto-lacteal fluid into nutritious milk was correlated with the evolution of other features that descended to modern mammals, such as an elevated metabolic rate, high aerobic capacity, rapid processing of nutrients, and fast growth rates (Oftedal 2002). Today, all extant mammals, monotremes, marsupials, and eutherians, rely entirely on milk to nourish the neonate.

Milk is a complex fluid consisting mainly of water, proteins, carbohydrates, lipids, salts, and vitamins (Fox 2009). In most mammals, most abundant milk proteins are caseins. In milk, caseins and calcium phosphate (CaP) form a huge complex called casein micelle (De Kruif and Holt 2003; Fox 2003). The caseins are a composite of proteins, Ca-sensitive and Ca-insensitive caseins (Swaisgood 2003). Ca-sensitive caseins bind CaPs but precipitate at high Ca concentrations. However, Ca-insensitive caseins stabilize the complex by forming micelles.

Ca-sensitive caseins have many Ser-Xaa-Glu/pSer (Xaa denotes any amino acid; pSer represents phospho-Ser; SXE) sequences in which the first Ser residue is usually phosphorylated (Holland 2009). Thus, a contiguous SXE sequence (typically SSSEE) forms a pSer cluster. Many pSer clusters interact with CaPs and assemble together into a nanometer-size cluster (De Kruif and Holt 2003). In addition to the protein–salt interaction, both types of caseins associate mainly through regions rich in Pro and/or Gln (P/Q), having been also referred to as hydrophobic regions (Horne 2009). Although Ca-sensitive caseins interact with CaP, this complex precipitates at high Ca concentrations. However, in casein micelles, Ca-insensitive caseins interact with Ca-sensitive caseins and stabilize the complex through a relatively hydrophilic Ser and/or Thr (S/T)-rich region that sits on the micelle surface (Schmidt 1982). Caseins thus form a huge micelle containing approximately 800 CaP nanoclusters (Smyth et al. 2004) and sequester a high

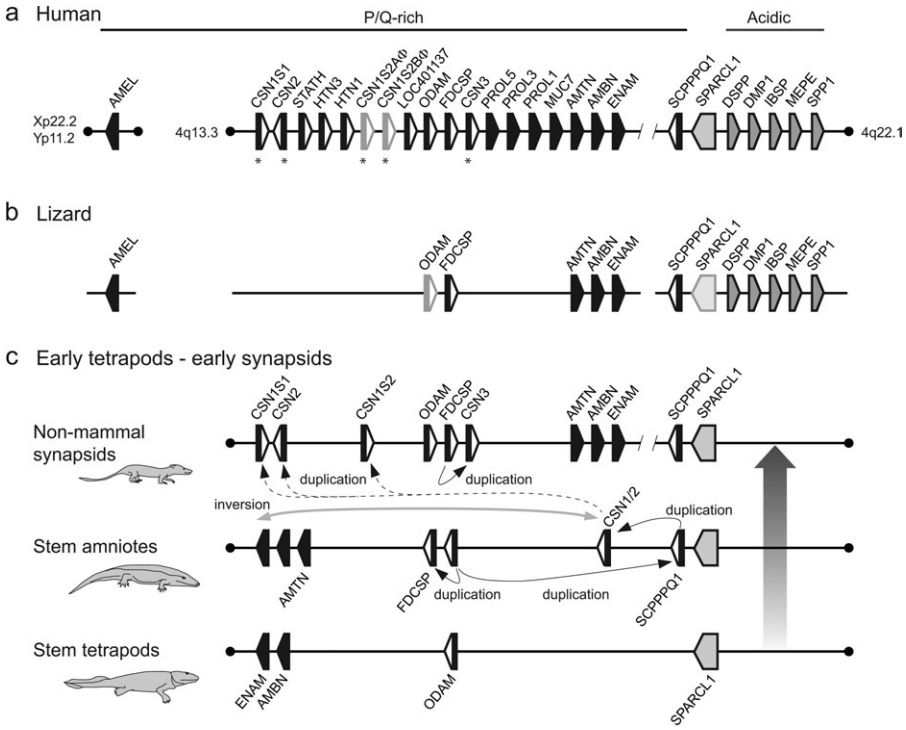

Fig. 1. Chromosomal location of SCPP genes in the genomes of humans (*a*), the lizard (*b*), and stem tetrapods to nonmammal synapsids (*c*). Each pentagon illustrates a gene and the transcriptional direction. P/Q-rich SCPP genes, acidic SCPP genes, and *SPARCL1* are shown in different gray scales. Among P/Q-rich SCPP genes, those possessing the entirely UT last exon (the termination codon resides within the penultimate exon) are shown with a white tail. Gene symbols not shown in the text were summarized previously (Kawasaki and Weiss 2008). (*a*) SCPP genes form two large clusters (4q13.3 and 4q22.1) with the exception of the amelogenin gene (*AMEL* on Xp22.2 and Yp11.2). Asterisks indicate casein genes. Two distinct *CSN1S2* pseudogenes (Φ) are represented with faded pentagons. (*b*) Neither *SPARCL1* nor *ODAM* has been confirmed in the lizard genome and is shown by a faded pentagon. (*c*) The duplication history from *CSN1/2* to modern Ca-sensitive casein genes was not resolved and is shown by dashed lines. In stem amniotes, gene symbols shown in stem tetrapods are omitted. Some P/Q-rich and acidic SCPP genes are not shown for clarity. Duplications of SCPP genes in early vertebrates were described previously (Kawasaki 2009).

content of CaP in milk, which helps altricial mammalian neonates to grow bone and teeth.

In all mammals studied to date, the Ca-sensitive casein is coded by two to four genes (Ginger and Grigor 1999; Rijnkels 2002; Rijnkels et al. 2003; Lefèvre et al. 2009). For example, the mouse genome has four different Ca-sensitive genes, encoding α_{s1}-casein (*CSN1S1*), β-casein (*CSN2*), and two distinct α_{s2}-caseins (*CSN1S2A* and *CSN1S2B*). However, in the human genome, both *CSN1S2* genes are nonfunctional (fig. 1*a*), and no *CSN1S2* has been found in marsupials (Lefèvre et al. 2007). By contrast, the Ca-insensitive casein is coded by a single κ-casein gene (*CSN3*).

Kawasaki and Weiss (2003, 2006) previously reported that casein genes were found only in mammalian genomes and that all casein genes are members of the secretory calcium-binding phosphoprotein (SCPP) gene family. The SCPP gene family initially arose from *SPARCL1* (SPARC-like 1) in an early vertebrate (Kawasaki et al. 2007), and, in modern vertebrates, many SCPPs are involved in mineralization of bone and teeth (Kawasaki et al. 2004; Kawasaki and Weiss 2008). In the human genome, we have identified 23 functional SCPP genes of which 22 genes form two large clusters (fig. 1*a*) (Kawasaki et al. 2009). These SCPP genes are characterized by two distinct types; one codes for an acidic protein and the other for a P/Q-rich protein. Due to these biased amino acid compositions, both types of SCPPs largely adopt flexible open structures (Kawasaki et al. 2007; Holt et al. 2009). It has been known that proteins with an open structure can tolerate more mutations than globular proteins (Holt and Sawyer 1993; Brown et al. 2002). Indeed, sequence similarities across different SCPPs are generally limited to the signal peptide (SP) that directs the mature protein into the extracellular space. However, SCPP genes have a well conserved and recognizable exon–intron structure. These and other similarities allowed us to identify many SCPP genes that evolved by gene duplication (Kawasaki et al. 2005).

All casein genes are among the P/Q-rich SCPP genes and are closely related to two tooth enamel matrix genes, ameloblastin (*AMBN*) and enamelin (*ENAM*) (Kawasaki and Weiss 2003). More recently, however, it was further demonstrated that some P/Q-rich SCPP genes have the entirely untranslated (UT) last exon. This exon was identified in all casein genes as well as the odontogenic ameloblast–associated (*ODAM*) gene, the SCPP-Pro-Gln-rich 1 (*SCPPPQ1*) gene, and the follicular dendritic cell secreted peptide (*FDCSP*, also called *C4ORF7*) gene (Kawasaki 2009). Phylogenetic distributions of these and other SCPP genes in tetrapod and teleost genomes suggested that *ODAM* was initially located close to acidic SCPP genes and that both

SCPPPQ1 and *FDCSP* arose from *ODAM* by tandem gene duplication. However, an intrachromosomal rearrangement split this original cluster into two isolated genomic regions before the divergence of synapsids and sauropsids, and, among P/Q-rich SCPP gens, only *SCPPPQ1* remains clustered with acidic SCPP genes in the lizard and mammalian genomes (fig. 1) (Kawasaki 2009). In this report, we argue that the Ca-sensitive casein genes originated from *SCPPPQ1*, whereas the Ca-insensitive casein gene differentiated from *FDCSP*. Thus, all casein genes share *ODAM* as a common ancestor.

Materials and Methods

Bioinformatic Analyses of Nucleotide and Amino Acid Sequences

We identified segments of lizard (*Anolis carolinensis*) *SCPPPQ1* and *FDCSP* using the gene prediction program "GENSCAN" (http://genes.mit.edu/GENSCAN.html) (Burge and Karlin 1997) in the genomic regions syntenic to their eutherian orthologs. The SP cleavage site was predicted using both "PSORT II" (http://psort.ims.u-tokyo.ac.jp/) (Nakai and Horton 1999) and "SignalP 3.0" (http://www.cbs.dtu.dk/services/SignalP/; hidden Markov model) (Emanuelsson et al. 2007).

Nucleotide sequences of casein and other genes and deduced amino acid sequences of their protein products were retrieved from GenBank, and their sequence similarities were studied through the National Center for Biotechnology Information (NCBI) website (http://www.ncbi.nlm.nih.gov/). Accession numbers of the sequences used in this study are shown in the legend of supplementary figure S1, Supplementary Material online. Amino acid sequences of various SCPPs were aligned using "Tcoffee" (http://www.tcoffee.org/) (Notredame et al. 2000) and manually corrected based on exon–intron borders. The exon–intron borders of various genes were determined by searching for sequence similarity between cDNA sequences and genomic sequences at the University of California, Santa Cruz website (http://genome.ucsc.edu/) or the NCBI website (whole-genome shotgun reads and trace archives). Likewise, the nucleotide sequences of gray short–tailed opossum (*Monodelphis domestica*) *CSN2* and elephant *CSN2* and *CSN3* were reconstructed from their genomic sequences by searching for similarities using already known cDNA sequences (supplementary fig. S1, Supplementary Material online). For all these exon–intron borders, the splice donor/acceptor sites and the polypyrimidine tract located upstream of the splice acceptor were confirmed. Versions of the genome sequences used in this study are shown in the legends of supplementary figures.

Cloning and Sequencing of Lizard *SCPPPQ1* and *FDCSP*

Parts of lizard *SCPPPQ1* and *FDCSP* were amplified by reverse transcription-polymerase chain reaction (RT-PCR) using total RNA molecules extracted from the lower jaw, and their full-length nucleotide sequences were determined as described previously (Al-Hashimi et al. 2010). PCR primers used to amplify an internal region, a 5′-end, or a 3′-end were designed based on the gene prediction analysis described above. The nucleotide sequences of these primers are as follows: 5′-TGAAGCTGCTTGTTCTTCTTTC-3′ and 5′-GCAGGATTTACAGGAAATGGTC-3′ for the internal region of *SCPPPQ1*; 5′-CAGCTCAACCAAACGTTCCTCCACAGA-3′ and 5′-GATATCCTTTTCCGGGCGCTTATCCTG-3′ for the 5′-end of *SCPPPQ1*; 5′-CGTTTGGTTGAGCTGGAAACTGAGGTG-3′ and 5′-ACGCTCTTCACTGGAGCTTGCTGACCT-3′ for the 3′-end of *SCPPPQ1*; 5′-GAAGGCTCTACTTGTGCTTGC-3′ and 5′-TGAAGATGTGGAAAACAGCAC-3′ for the internal region of *FDCSP*; 5′-CAAGGAAAGAAGGGGGTACCGTCCAC-3′ for the 5′-end of *FDCSP*; and 5′-TCCTTGCTTCCACTTCTGCTGCCACT-3′ and 5′-GAGAACCAAAAGGCAGGCAAGCACAAG-3′ for the 3′-end of *FDCSP*. The nucleotide sequences of these two genes are available through GenBank (accession number, GU944675 for *SCPPPQ1* and GU944674 for *FDCSP*).

Results

Internal Duplications in Ca-Sensitive Casein Genes

All mammalian Ca-sensitive casein genes, *CSN1S1*, *CSN1S2*, and *CSN2*, share a characteristic exon–intron structure with the other SCPP genes (Kawasaki and Weiss 2003). Although these genes vary considerably in number of exons, the exons coding for most of the mature protein (excluding each two 5′- and 3′-exons) consist of only four distinct types: *a*, *b*, *c*, and P/Q-rich (fig. 2). Among these, type-*a*, -*b*, and -*c* exons are small in size, coding for 9 residues, including many charged amino acids and forming a Ca-binding pSer cluster at the 3′ end (type-*a*), 8 residues, including three or four charged amino acids (type-*b*), and 14 residues, including five or seven charged amino acids (type-*c*; see supplementary fig. S1, Supplementary Material online for deviations from these standards). Similar exons have been identified in bovine *CSN1S2* and *CSN2* (Groenen et al. 1993). However, our result demonstrates that all small exons found in the middle of the Ca-sensitive casein genes studied to date can be classified into only type-*a*, -*b*, or -*c*.

Among these small exons, the type-*a* exon is the smallest in number and is often followed by a type-*b* exon and then by a type-*c* exon, with type-*b* and -*c* exons being separated by a small intron (103 nucleotides or less; underlined in fig. 2). For example, platypus *CSN1S1* contains three tandem *a-b-c* exon units. The *a-b-c* exon unit is shared among different Ca-sensitive casein genes in eutherian, marsupial, and monotreme lineages, suggesting that the common ancestor of these genes had this exon unit. Subsequent duplications and deletions of genome segments led to more complicated, gene-specific and lineage-specific arrangements of these exons.

In addition to these small exons, Ca-sensitive casein genes have one (*CSN1S1* and *CSN2*) or two (*CSN1S2*) relatively large P/Q-rich (~25% P/Q) exon, which also codes for many aromatic residues, mostly Phe and/or Tyr (F/Y), in

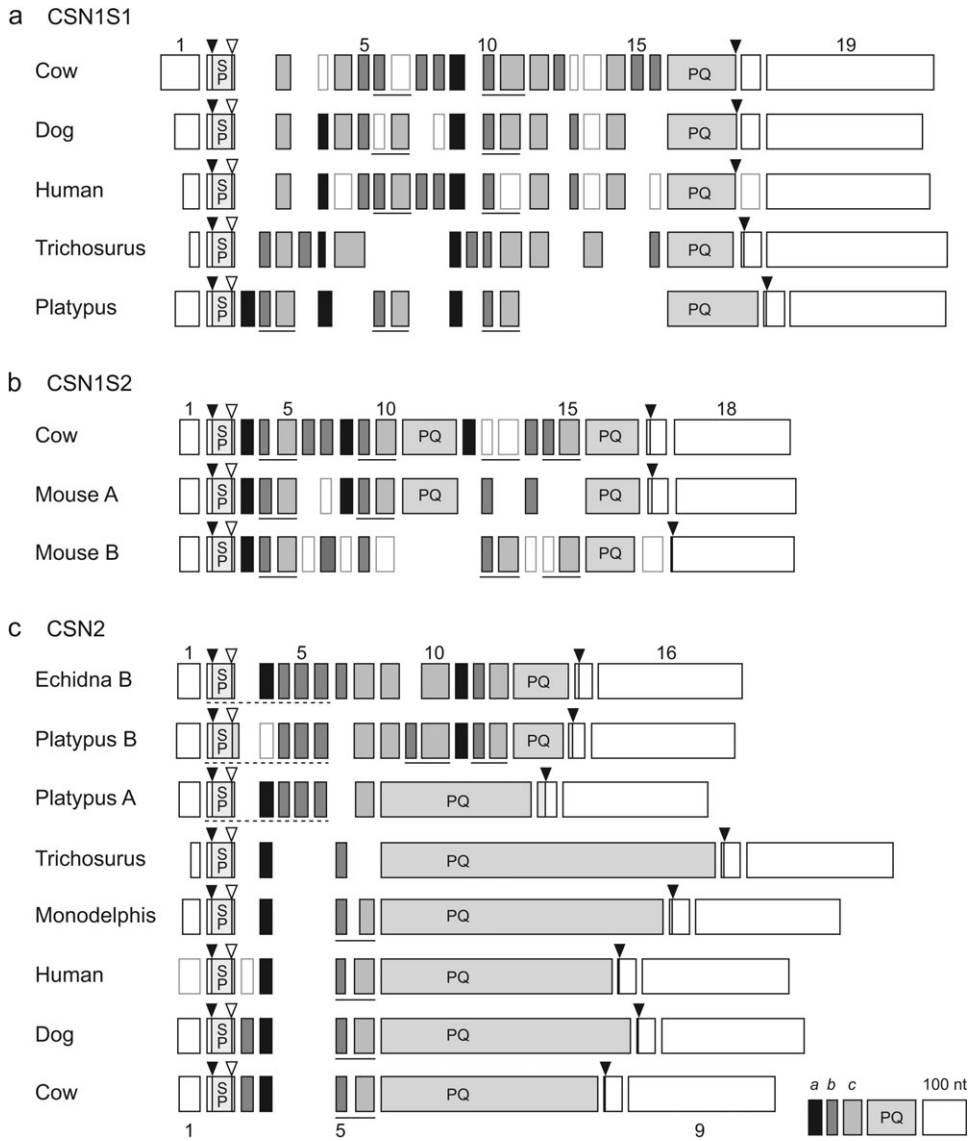

FIG. 2. Exon–intron structure of *CSN1S1* (*a*), *CSN1S2* (*b*), and *CSN2* (*c*) in representative mammals. Each separate box represents a single exon. Filled arrowheads represent the position of the initiation codon or the termination codon and delimit the translated region. An open arrowhead indicates the cleavage site of the SP. The entire SP is coded within exon 2 and shown with a light gray scale. The penultimate exon usually codes for the termination codon (the termination codon extends over two adjacent exons in eutherian *CSN1S1* genes). The 5′- or 3′-UT region is shown by a blank box. Type-*a*, -*b*, and -*c* exons and P/Q-rich exons (PQ) are illustrated in different gray scales (see the bottom). Skipped exons or pseudoexons are shown in gray outline. Some type-*b* and -*c* exons that are separated by a small intron (103 nt or less) are underlined. High sequence similarity is detected between monotreme *CSN2A* and *CSN2B* in exons 2–6, which are indicated with dashed underlines (*c*). Exon numbers for the cow casein genes and echidna *CSN2B* are shown at the top or bottom. The scale for exon length is shown at the bottom. Trichosurus is the brushtail possum. See supplementary fig. S1, Supplementary Material online for alignments of the amino acid sequences.

the 3′-half (supplementary fig. S1, Supplementary Material online). The P/Q-rich exon of *CSN2* varies in size but is larger than that in other Ca-sensitive casein genes (fig. 2*c*). This difference is partly due to different numbers and sizes of intraexonic duplication as shown for eutherian *CSN2* genes (Holt and Sawyer 1993). Intraexonic duplications are also detected even in the smallest P/Q-rich exon of monotreme *CSN2A* genes (*CSN2* ortholog; supplementary fig. S1*c*, Supplementary Material online), suggesting that the P/Q-rich exon of *CSN2* was originally small and hence shares a common ancestor with P/Q-rich exons of the other Ca-sensitive casein genes.

Recently, *CSN2B* was identified in monotremes (Warren et al. 2008). *CSN2B* shows a high sequence similarity to monotreme *CSN2A* in exons 2–6 but is more similar to eutherian *CSN1S2* in the chromosomal location and the small size of the P/Q-rich exon (fig. 2). These findings led to the suggestion that *CSN2B* is a chimera of *CSN2* and *CSN1S2* (Lefèvre et al. 2009). In our analysis, significant sequence similarity was not detected between monotreme *CSN2B* and eutherian *CSN1S2* in their possibly orthologous P/Q-rich sequences (supplementary fig. S1*b*, Supplementary Material online). In fact, no significant similarity has been identified across monotreme, marsupial, and eutherian

Ca-sensitive caseins in their orthologous P/Q-rich sequences. By contrast, within eutherian *CSN1S2* genes, two different P/Q-rich exons code for similar sequences (Stewart et al. 1987) (supplementary fig. S1*b*, Supplementary Material online). These findings demonstrate that the two P/Q-rich exons in *CSN1S2* arose by a relatively recent duplication in the eutherian lineage. This result is consistent with a previous hypothesis about the evolution of *CSN1S2* (Groenen et al. 1993).

Ca-Sensitive Casein Genes Likely Arose from *SCPPPQ1*

Above we suggested that an ancient Ca-sensitive casein gene had the *a-b-c* exon unit and a single P/Q-rich exon and that many redundant exons arose by duplication. In addition, different Ca-sensitive casein genes share similarly-sized 5′- and 3′-exons, similar positions of the initiation and termination codons, and similar cleavage sites for the SP (fig. 2 and supplementary fig. S1, Supplementary Material online). These findings suggest that all Ca-sensitive casein genes arose from a putative common ancestor that consisted of eight exons (referred to as *CSN1/2* in fig. 3*a*): a small entirely UT exon 1, SP-coding exon 2, the *a-b-c* unit in exons 3–5, P/Q-rich (partly F/Y-rich) exon 6, a small exon 7 coding for N-terminal one to three residues and the termination codon, and a large entirely UT exon 8. *CSN1/2* is thus reminiscent of *CSN2*, especially in Monodelphis, although *CSN2* has a larger P/Q-rich exon, as described above (fig. 3*a*). The *a-b-c* exon unit has been found only in Ca-sensitive casein genes among SCPP genes. Thus, similarities in the small size and coded amino acid composition (charged residues) across type-*a*, *-b*, and *-c* exons further suggest that these exons originated by duplication and that the earliest precursor of *CSN1/2* consisted of six exons (referred to as proto-*CSN1/2* in fig. 3*a*). This analysis reinforces previous hypotheses based on rodent casein genes (Hobbs and Rosen 1982; Jones et al. 1985).

Jones et al. also suggested that all exons in a primordial Ca-sensitive casein gene that corresponds to proto-*CSN1/2* were separately recruited from different genes. However, we found that exons showing similar characteristics are all present in both mammalian and lizard *SCPPPQ1* genes (fig. 3*a*; exons 1–4, 9, and 10). Lizard *SCPPPQ1* has type-*a*-like exon 3 coding for a potential pSer cluster and relatively large P/Q(F/Y)-rich exon 4, and hence, proto-*CSN1/2* is more similar to lizard *SCPPPQ1* rather than to mammalian *SCPPPQ1* (fig. 3*a*). As we describe below, *SCPPPQ1* was initially located adjacent to *ODAM* in some stem amniote, although these genes are separated today (fig. 1*c*) (Kawasaki 2009). Thus, our new finding supports the idea that proto-*CSN1/2* arose by tandem duplication from ancient *SCPPPQ1* that was similar to the contemporary lizard ortholog. *CSN1/2* (or proto-*CSN1/2*) was subsequently separated from *SCPPPQ1* by a chromosomal rearrangement before the divergence of synapsids and sauropsids, much earlier than the origin of lactation. Although we have been unable to find *CSN1/2* in the genomes of the lizard or birds, this gene could be retained in other sauropsids.

CSN3 Arose from *FDCSP*

As shown in figure 3*b*, *CSN3* is comprised of entirely UT exons 1 and 5, SP-coding exon 2, a small exon 3 coding for 11 residues, and a relatively large P/Q-rich exon 4. All these five exons are also found in *FDCSP*, although a small exon 3 was not identified in lizard *FDCSP*. A notable structural difference between these two genes resides in the cleavage site of their SPs. The size of the SP is 20 residues in most κ-caseins, whereas it is 16 residues in FDCSP (supplementary fig. S3, Supplementary Material online). However, their first 20 amino acid sequences show similarities, suggesting that a slight change in the sequence shifted the end of the SP. Indeed, both SignalP and PSORT II predicted that a substitution of Ser to Pro within the SP in platypus κ-casein would shorten the length of SP to 16 amino acids, identical to that of FDCSP.

In addition to the exon–intron structure, κ-casein is also similar to FDCSP in the modular structure. Both proteins consist of three distinct modules: an N-terminal region possessing many charged amino acids, coded by exon 3 and the 5′-end of exon 4; an intermediate P/Q(F/Y)-rich region; and a C-terminal hydrophilic S/T-rich region (fig. 3*b*). The S/T-rich region of both proteins is also characterized by a low P/Q content and by almost no F/Y (or Trp) residues. In monotreme *CSN3* genes, the 5′-end of exon 4 codes for a potential pSer cluster (S** in fig. 3*b*). A similar pSer cluster is also present in FDCSP in mammals and the lizard, whereas no pSer residue has been found in marsupial or eutherian κ-caseins (supplementary fig. S3, Supplementary Material online). Consequently, in our position-specific iterated Blast search against GenBank, the N-terminus of platypus κ-casein (including the SP) showed a higher sequence similarity to the corresponding region of human FDCSP rather than to human κ-casein (score = 37.0 and 31.2 bits, respectively). These findings collectively suggest that *CSN3* was derived from ancient *FDCSP* (fig. 1*c*).

Two Types of Casein Genes Arose from *ODAM* Through Two Different Pathways

Among P/Q-rich SCPP genes, only *ODAM* has been identified in both teleost and tetrapod genomes, although this gene has not been found in the lizard genome (fig. 1) (Kawasaki 2009). *ODAM* has an entirely UT last exon in the zebrafish, frog, and rodents, but such an exon has not been found in the amelogenin (*AMEL*) gene, *AMBN*, *ENAM*, or the amelotin (*AMTN*) gene in any tetrapods studied to date (fig. 1). Although the entirely UT last exon could be secondarily lost, creation of such an exon would be a rare event. Indeed, no such exon has been identified in any acidic SCPP genes found to date. Thus, the phylogenetic distribution of P/Q-rich SCPP genes possessing the entirely UT last exon suggests that *ODAM* is the common ancestor of *SCPPPQ1* and *FDCSP*. *ODAM* is located immediately downstream of *SPARCL1* in the fugu and the frog genomes (Kawasaki 2009) and hence, probably also in

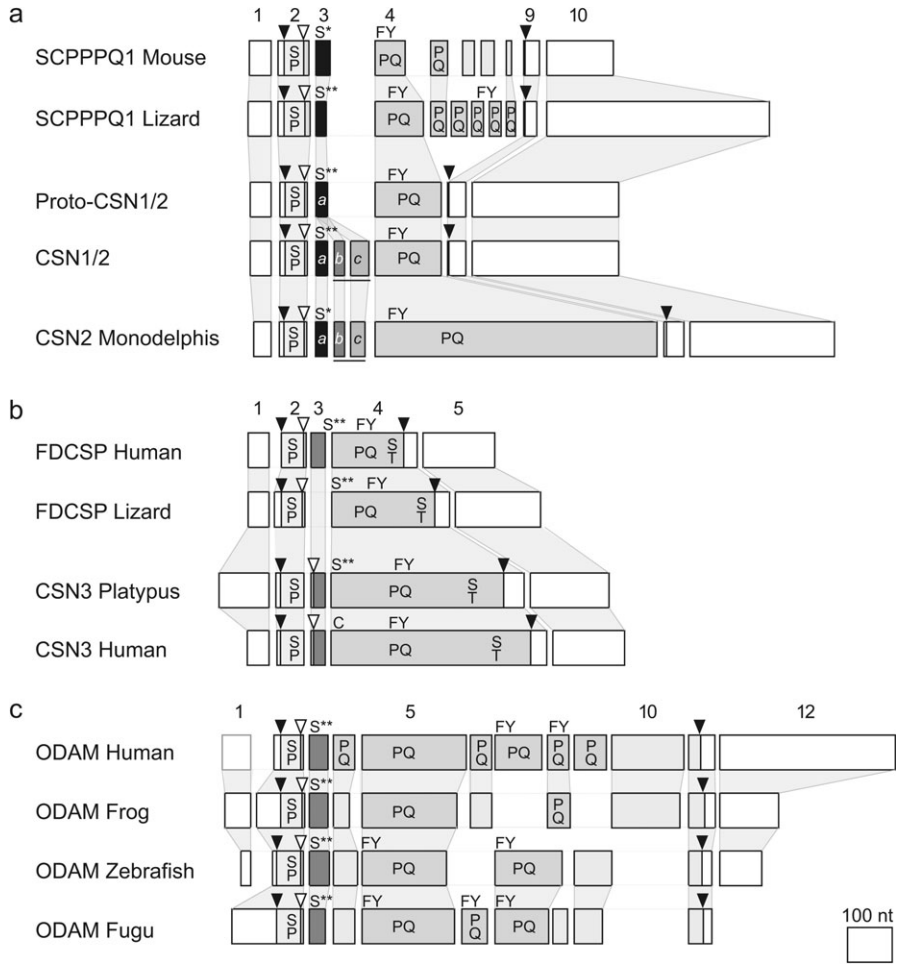

Fig. 3. Exon–intron structure of *SCPPPQ1*, *CSN1/2*, and *CSN2* (*a*), *FDCSP* and *CSN3* (*b*), and *ODAM* (*c*). See the legend of figure 2 for common annotations. Either orthologous or paralogous exons are linked with shadow. Regions coding for a high content of aromatic amino acids are shown by FY on the top and S/T residues by ST inside the exon. The position coding for a pSer cluster is represented by S**. (*a*) Exon 3 of mammalian *SCPPPQ1* codes for a single potential pSer (S*) but tentatively classified as type-*a*-like exon, because a SSSE sequence or a similar sequence is coded near the 3′-end. Exon 3 of Monodelphis *CSN2* also codes for a single potential pSer residue (S*) but an adjacent Thr residue (supplementary fig. S1*c*, Supplementary Material online) could be phosphorylated, as reported for some Thr residues forming a TXE sequence in other caseins (Holland 2009). In the lizard *SCPPPQ1*, exons 7 and 8 are both F/Y-rich (25% or more). *CSN1/2* was reconstructed as discussed in the text. The example shown in this figure is based on cow *CSN2* (exons 1, 2, 4–6, 8, and 9) and Trichosurus *CSN1S1* (P/Q-rich exon). See supplementary figure S2, Supplementary Material online for amino acid sequences of SCPPPQ1. (*b*) A Cys residue coded by the 5′-end of human *CSN3* is represented by C. The position coding for the pSer cluster in mammalian *FDCSP* extends from exon 3 to exon 4. See supplementary figure S3, Supplementary Material online for amino acid sequences of FDCSP and κ-casein. (*c*) The structure of *ODAM* is illustrated based on the nucleotide sequences of NM_017855.3 (human), EU642609.1 (frog), EU642608.1 (zebrafish), and NM_001037851.1 (fugu). Entirely UT exon 1 has not been reported in humans but found in the rat (NM_001044274.1). The corresponding exon, found in the human genome, is shown in gray outline. Some orthologous exons are P/Q-rich in humans but not in the frog or teleosts. The scale is shown at the bottom.

the stem tetrapods (fig. 1*c*). In the lizard and mammalian genomes, *SCPPPQ1* is separated from all the other P/Q-rich SCPP genes and is located immediately downstream of *SPARCL1*, whereas *ODAM–FDCSP* is linked to *AMTN–AMBN–ENAM* (fig. 1*a* and *b*). Moreover, the arrangement of non-SCPP genes located immediately downstream of *SCPPPQ1* (*HSD17B11* and *KLHL8*) are common to the lizard and mammalian genomes. The arrangement of these genes suggests that the original SCPP gene cluster split by a chromosomal rearrangement before the divergence of synapsids and sauropsids and that *SCPPPQ1* originated from *ODAM* by tandem duplication before this rearrangement.

ODAM shares different types of exons with *SCPPPQ1*, putative proto-*CSN1/2*, *FDCSP*, and *CSN3* (fig. 3). However, *CSN3* is more similar to *FDCSP* than to *ODAM* in the organization of all five exons, as described above (fig. 3). In addition, a relatively large exon coding for charged residues at the 5′-end, followed by a P/Q(F/Y)-rich sequence and an S/T-rich sequence has been identified in *FDCSP* and *CSN3* but not in *ODAM* (fig. 3 and supplementary fig. S3, Supplementary Material online). It is unlikely that this type of complicated exon evolved as the result of convergence. Thus, it appears that *CSN3* originated from *FDCSP*, rather than from *ODAM*.

Similarly, we found that *SCPPPQ1* has all of the six distinct types of exons that are assumed to have constituted putative proto-*CSN1/2*. By contrast, exon 3 of *ODAM* is less similar to type-*a*, -*b*, or -*c* exons in size or the distribution of pSer

residues, and the penultimate exon of *ODAM* codes for larger numbers of amino acids than that of Ca-sensitive casein genes and *SCPPPQ1* (fig. 3). These characteristics of *ODAM* are conserved even in teleost orthologs. Thus, it is likely that proto-*CSN1/2* arose from *SCPPPQ1*, although, if all the similarities in exon 3 and in the penultimate exon of these two genes are the result of convergence, proto-*CSN1/2* could have arisen directly from *ODAM* more recently. Collectively, our findings suggest that all casein genes evolved from *ODAM* via two different pathways; Ca-sensitive casein genes likely originated through *SCPPPQ1*, whereas the Ca-insensitive casein gene differentiated from *FDCSP* (fig. 1*c*).

Discussion

The Origin of Ca-Sensitive and Ca-Insensitive Caseins

We have argued that the casein genes evolved from *ODAM* through two different pathways; Ca-sensitive casein genes likely originated directly from *SCPPPQ1*, whereas the Ca-insensitive casein genes directly differentiated from *FDCSP*. Expression of *ODAM* has been detected in epithelial cells that cover the tooth surface in both mammals and teleosts (Moffatt et al. 2008; Kawasaki 2009). A similar expression pattern has been also reported for *SCPPPQ1* in the rat (Moffatt et al. 2006). These studies suggest that both *ODAM* and *SCPPPQ1* are used in mineralization of the tooth surface. Further, FDCSP has been found in periodontal ligament, a soft connective tissue surrounding the roots of teeth, where it is thought to prevent spontaneous CaP precipitation (Nakamura et al. 2005). It thus appears that the immediate ancestors of caseins modulated mineralization of the tooth through association with Ca ions via a pSer residue or cluster at the N-terminus (fig. 3) and that both types of caseins have evolved as Ca-binding proteins.

This view is inconsistent with the hypothesis proposed decades ago (Jollès et al. 1978; Jollès and Henschen 1982) that κ-casein (known as a milk clotting factor; supplementary fig. S3, Supplementary Material online) was derived from γ-fibrinogen, a blood coagulation factor. This hypothesis largely depends on supposed amino acid sequence identities between human γ-fibrinogen and cow and sheep κ-caseins (31.4–34.1%). However, in a similar analysis (supplementary fig. S4, Supplementary Material online), no high sequence identity was detected for Monodelphis κ-casein (8.8%) and platypus κ-casein (4.3%). Moreover, the supposed similar regions are coded only by exon 4 in *CSN3*, whereas these regions are coded by six different exons in the γ-fibrinogen gene (supplementary fig. S4, Supplementary Material online). Given our new results, we conclude that κ-casein is evolutionarily distinct from γ-fibrinogen.

Conserved Modular Structures of Caseins and Micelle Formation in Primitive Milk

Milk of all modern mammals studied to date contains both α_{s1}- and β-caseins. However, *CSN1S1*-null goats and *CSN2*-deficient mice and goats all produce milk containing casein micelles (Kumar et al. 1994; Chanat et al. 1999). It is also known that β-casein is the principal Ca-sensitive casein in human milk (Nagasawa et al. 1970). These studies indicate that micelle formation does not require specific Ca-sensitive caseins. This conclusion is consistent with the fact that orthologous Ca-sensitive caseins have different numbers and sizes of pSer clusters and various lengths of P/Q-rich regions (fig. 2 and supplementary fig. S1, Supplementary Material online). It is thus possible that ancient CSN1/2, which is similar to β-casein (comprised of an N-terminal pSer cluster and a following P/Q-rich region; fig. 3*a*), already had the potential to form a primitive casein micelle in the presence of κ-casein.

Similarities between monotreme κ-casein and FDCSP, as we described above, suggest that ancient κ-casein was similar to the monotreme ortholog, comprised of β-casein-like modules (N-terminal pSer cluster and P/Q-rich region) and a C-terminal S/T-rich region (fig. 3*b*). This modular structure suggests that ancient κ-casein interacted with CaP through the pSer cluster. By contrast, all marsupial and eutherian κ-caseins studied to date have one or two Cys residues, instead of the pSer cluster (C in fig. 3*b*). These Cys residues form disulfide bonds to different κ-caseins or Ca-sensitive caseins, and these crosslinks are thought to stabilize the casein micelle (Rasmussen et al. 1999). Hence, our result suggests that the interaction between κ-casein and CaP was originally important for micelle formation, but this interaction was replaced by the disulfide bond in marsupials and eutherians. That is, κ-casein became specialized to the micelle-stabilizing Ca-insensitive casein. It was experimentally shown that dephosphorylated β-casein stabilizes a complex of α_{s1}-casein and CaP more efficiently than native β-casein, especially at higher temperatures (Yoshikawa et al. 1975). We thus speculate that the loss of the pSer cluster in κ-casein was important for efficient micelle formation in marsupials and eutherians, which may be correlated with their increased body temperature.

Ancient κ-casein probably had all three modules, important for micelle formation. This structure implies that an ancient κ-casein alone had the potential to form a primitive casein micelle. It is known that κ-casein-rich milk has smaller micelles (Donnelly et al. 1984; Dalgleish et al. 1989); and these and other studies led to the notion that κ-casein on the micelle surface terminates the growth of the micelle (Schmidt 1982; Horne 2009). Given this idea and if a primitive micelle consisted of only κ-casein as a principal protein component, the micelle could have been small, perhaps formed around a single CaP nanocluster. Expression of *FDCSP* has been detected in the lactating mammary gland (Rijnkels et al. 2003), which supports an ancient origin of κ-casein in milk. CSN1/2 might have been co-opted later for micelle formation and contributed to increased CaP concentrations.

However, it is also possible that both ancient κ-casein and CSN1/2 initially inhibited spontaneous CaP precipitation in primitive milk (or proto-lacteal fluid) without forming micelles, similar to many other SCPPs in various tissues and biofluids (Kawasaki and Weiss 2006; Holt et al. 2009; Kawasaki et al. 2009). Subsequently, a primitive casein micelle was formed by these two caseins, when milk evolved to contain

higher concentrations of CaP and caseins. Evolution of larger P/Q-rich regions in both caseins (fig. 3*a* and *b*) may have enhanced casein–casein interactions and stabilized the micelle. The early origin of *CSN1/2* (fig. 1*c*) appears to corroborate this model. The potential for micelle formation by ancient κ-casein and CSN1/2 could be experimentally tested by synthesizing such proteins.

Early Origins of Caseins Before Lactation

Ancient proto-lacteal synapsids had only partially calcified parchment-shelled eggs, and Ca had to be supplied primarily from the egg yolk (Brawand et al. 2008). However, it was argued that additional Ca could be supplied from cutaneous glandular secretions through the eggshell and complement the needs of the hatchling in these synapsids (Oftedal 2002). This possibility would be supported if ancient casein(s) inhibited spontaneous precipitation of CaP in such proto-lacteal secretions. Oftedal also suggested that large casein micelles may block the passage of CaP through eggshell pores. Yet, ancient caseins may have formed small micelles or may not have formed micelles at all, as described above; and hence, these caseins could have facilitated the supply of CaP to the embryos through the eggshell surface. In addition, ancient caseins may have inhibited calcification of eggshell pores and maintained the integrity of the eggshell surface.

Extremely high substitution rates in casein genes and their immediate ancestors do not allow us to estimate the divergence dates of these genes. However, our results suggest the possibility that an unexpectedly early origin of *CSN1/2* before the divergence of synapsids and sauropsids. In addition, similarities between *CSN2* and *SCPPPQ1* as well as between *CSN3* and *FDCSP* illustrate that many slight genetic modifications of ancestral genes have created caseins, proteins vital to the sustained success of mammals.

Supplementary Material

Supplementary figures S1–S4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We are grateful to Prof. Kenneth M. Weiss and Dr Anne V. Buchanan at Penn State University for generous comments and encouragements and to Dr Pierre Moffatt at the Shriners Hospital for Children in Montréal for discussion. Financial support from the National Science Foundation, grant BCS0725227, BCS0343442, and the Penn State Evan Pugh Professors Research Fund to Prof. Kenneth Weiss is gratefully acknowledged. A-G.L. and J-Y.S. are financially supported by grants from the Centre National de la Recherche Scientifique and the Université Pierre et Marie Curie.

References

Al-Hashimi N, Lafont AG, Delgado S, Kawasaki K, Sire JY. 2010. The enamelin genes in lizard, crocodile and frog, and the pseudogene in the chicken provide new insights on enamelin evolution in tetrapods. *Mol Biol Evol*. 27:2078–2094.

Benton MJ. 2005. Vertebrate paleontology. Malden (MA): Blackwell.

Brawand D, Wahli W, Kaessmann H. 2008. Loss of egg yolk genes in mammals and the origin of lactation and placentation. *PLoS Biol*. 6:e63.

Brown CJ, Takayama S, Campen AM, Vise P, Marshall TW, Oldfield CJ, Williams CJ, Dunker AK. 2002. Evolutionary rate heterogeneity in proteins with long disordered regions. *J Mol Evol*. 55:104–110.

Burge C, Karlin S. 1997. Prediction of complete gene structures in human genomic DNA. *J Mol Biol*. 268:78–94.

Chanat E, Martin P, Ollivier-Bousquet M. 1999. α_{S1}-casein is required for the efficient transport of β and κ-casein from the endoplasmic reticulum to the Golgi apparatus of mammary epithelial cells. *J Cell Sci*. 112(Pt 19):3399–3412.

Dalgleish DG, Horne DS, Law AJR. 1989. Size-related differences in bovine casein micelles. *Biochim Biophys Acta*. 991:383–387.

De Kruif CG, Holt C. 2003. Casein micelle structure, functions and interactions. In: Fox PF, McSweeney PLH, editors. Advanced dairy chemistry. New York: Kluwer. p. 233–276.

Donnelly WJ, McNeill GP, Buchheim W, McGann TC. 1984. A comprehensive study of the relationship between size and protein composition in natural bovine casein micelles. *Biochim Biophys Acta*. 789:136–143.

Emanuelsson O, Brunak S, von Heijne G, Nielsen H. 2007. Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc*. 2:953–971.

Fox PF. 2003. Milk proteins: general and historical aspects. In: Fox PF, McSweeney PLH, editors. Advanced dairy chemistry. New York: Kluwer. p. 1–48.

Fox PF. 2009. Milk: an overview. In: Thompson A, Boland M, Singh H, editors. Milk proteins: from expression to food. San Diego (CA): Elsevier. p. 1–54.

Ginger MR, Grigor MR. 1999. Comparative aspects of milk caseins. *Comp Biochem Physiol B Biochem Mol Biol*. 124:133–145.

Groenen MA, Dijkhof RJ, Verstege AJ, van der Poel JJ. 1993. The complete sequence of the gene encoding bovine αs2-casein. *Gene* 123:187–193.

Hobbs AA, Rosen JM. 1982. Sequence of rat α- and γ-casein mRNAs: evolutionary comparison of the calcium-dependent rat casein multigene family. *Nucleic Acids Res*. 10:8079–8098.

Holland JW. 2009. Post-translational modifications of caseins. In: Thompson A, Boland M, Singh H, editors. Milk proteins: from expression to food. San Diego (CA): Elsevier. p. 107–132.

Holt C, Sawyer L. 1993. Caseins as rheomorphic proteins: interpretation of primary and secondary structure of the α_{s1}-, β- and κ-caseins. *J Chem Soc Faraday Trans*. 89:2683–2692.

Holt C, Sørensen ES, Clegg RA. 2009. Role of calcium phosphate nanoclusters in the control of calcification. *Febs J*. 276: 2308–2323.

Horne DS. 2009. Casein micelle structure and stability. In: Thompson A, Boland M, Singh H, editors. Milk. San Diego (CA): Elsevier. p. 133–162.

Jollès P, Henschen A. 1982. Comparison between the clotting of blood and milk. *Trends Biochem Sci*. 7:325–328.

Jollès P, Loucheux-Lefebvre MH, Henschen A. 1978. Structural relatedness of κ-casein and fibrinogen γ-chain. *J Mol Evol*. 11:271–277.

Jones WK, Yu-Lee LY, Clift SM, Brown TL, Rosen JM. 1985. The rat casein multigene family. Fine structure and evolution of the β-casein gene. *J Biol Chem*. 260:7042–7050.

Kawasaki K. 2009. The SCPP gene repertoire in bony vertebrates and graded differences in mineralized tissues. *Dev Genes Evol*. 219: 147–157.

Kawasaki K, Buchanan AV, Weiss KM. 2007. Gene duplication and the evolution of vertebrate skeletal mineralization. *Cells Tissues Organs*. 186:7–24.

Kawasaki K, Buchanan AV, Weiss KM. 2009. Biomineralization in humans: making the hard choices in life. *Annu Rev Genet*. 43:119–142.

Kawasaki K, Suzuki T, Weiss KM. 2004. Genetic basis for the evolution of vertebrate mineralized tissue. *Proc Natl Acad Sci U S A*. 101:11356–11361.

Kawasaki K, Suzuki T, Weiss KM. 2005. Phenogenetic drift in evolution: the changing genetic basis of vertebrate teeth. *Proc Natl Acad Sci U S A*. 102:18063–18068.

Kawasaki K, Weiss KM. 2003. Mineralized tissue and vertebrate evolution: the secretory calcium-binding phosphoprotein gene cluster. *Proc Natl Acad Sci U S A*. 100:4060–4065.

Kawasaki K, Weiss KM. 2006. Evolutionary genetics of vertebrate tissue mineralization: the origin and evolution of the secretory calcium-binding phosphoprotein family. *J Exp Zool B Mol Dev Evol*. 306:295–316.

Kawasaki K, Weiss KM. 2008. SCPP gene evolution and the dental mineralization continuum. *J Dent Res*. 87:520–531.

Kumar S, Clarke AR, Hooper ML, Horne DS, Law AJ, Leaver J, Springbett A, Stevenson E, Simons JP. 1994. Milk composition and lactation of β-casein-deficient mice. *Proc Natl Acad Sci U S A*. 91:6138–6142.

Lefèvre CM, Digby MR, Whitley JC, Strahm Y, Nicholas KR. 2007. Lactation transcriptomics in the Australian marsupial, *Macropus eugenii*: transcript sequencing and quantification. *BMC Genomics* 8:417.

Lefèvre CM, Sharp JA, Nicholas KR. 2009. Characterisation of monotreme caseins reveals lineage-specific expansion of an ancestral casein locus in mammals. *Reprod Fertil Dev*. 21: 1015–1027.

Moffatt P, Smith CE, Sooknanan R, St-Arnaud R, Nanci A. 2006. Identification of secreted and membrane proteins in the rat incisor enamel organ using a signal-trap screening approach. *Eur J Oral Sci*. 114(1 Suppl):139–146 discussion 164–165, 380–381.

Moffatt P, Smith CE, St-Arnaud R, Nanci A. 2008. Characterization of Apin, a secreted protein highly expressed in tooth-associated epithelia. *J Cell Biochem*. 103:941–956.

Nagasawa T, Kiyosawa I, Kuwahara K. 1970. Human casein. II. Isolation of human β-casein fraction and human β-casein B. *J Dairy Sci*. 53:136–145.

Nakai K, Horton P. 1999. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci*. 24:34–36.

Nakamura S, Terashima T, Yoshida T, Iseki S, Takano Y, Ishikawa I, Shinomura T. 2005. Identification of genes preferentially expressed in periodontal ligament: specific expression of a novel secreted protein, FDC-SP. *Biochem Biophys Res Commun*. 338:1197–1203.

Notredame C, Higgins DG, Heringa J. 2000. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J Mol Biol*. 302:205–217.

Oftedal OT. 2002. The mammary gland and its origin during synapsid evolution. *J Mammary Gland Biol Neoplasia*. 7:225–252.

Rasmussen LK, Johnsen LB, Tsiora A, Sørensen ES, Thomsen JK, Nielsen NC, Jakobsen HJ, Petersen TE. 1999. Disulphide-linked caseins and casein micelles. *Int Dairy J*. 9:215–218.

Rijnkels M. 2002. Multispecies comparison of the casein gene loci and evolution of casein gene family. *J Mammary Gland Biol Neoplasia*. 7:327–345.

Rijnkels M, Elnitski L, Miller W, Rosen JM. 2003. Multispecies comparative analysis of a mammalian-specific genomic domain encoding secretory proteins. *Genomics* 82:417–432.

Schmidt DG. 1982. Association of caseins and casein micelle structure. In: Fox PF, editor. Developments in dairy chemistry. New York: Applied Science Publishers. p. 61–86.

Smyth E, Clegg RA, Holt C. 2004. A biological perspective on the structure and function of caseins and casein micelles. *Int J Dairy Technol*. 57:121–126.

Stewart AF, Bonsing J, Beattie CW, Shah F, Willis IM, Mackinlay AG. 1987. Complete nucleotide sequences of bovine α_{S2}- and β-casein cDNAs: comparisons with related sequences in other species. *Mol Biol Evol*. 4:231–241.

Swaisgood HE. 2003. Chemistry of the caseins. In: Fox PF, McSweeney PLH, editors. Advanced dairy chemistry. New York: Kluwer. p. 139–201.

Warren WC, Hillier LW, Marshall Graves JA, et al. (20 co-authors). 2008. Genome analysis of the platypus reveals unique signatures of evolution. *Nature* 453:175–183.

Yoshikawa M, Sugimoto E, Chiba H. 1975. Studies on the interaction between α_{S1}- and β-caseins. *Agric Biol Chem*. 39:1843–1849.