ACADEMIC PRESS

Available online at www.sciencedirect.com

SCIENCE DIRECT®

GENOMICS

Genomics 82 (2003) 417–432

www.elsevier.com/locate/ygeno

Multispecies comparative analysis of a mammalian-specific genomic domain encoding secretory proteins

Monique Rijnkels,[a,*] Laura Elnitski,[b,c] Webb Miller,[b,d] and Jeffrey M. Rosen[a]

[a] *Department of Molecular and Cellular Biology, Baylor College of Medicine, One Baylor Plaza, Houston, TX 77030, USA*
[b] *Department of Computer Science and Engineering, Pennsylvania State University, University Park, PA 16802, USA*
[c] *Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, PA 16802, USA*
[d] *Department of Biology, Pennsylvania State University, University Park, PA 16802, USA*

Received 11 November 2002; accepted 7 April 2003

Abstract

The mammalian-specific casein gene cluster comprises 3 or 4 evolutionarily related genes and 1 physically linked gene with a functional association. To gain a better understanding of the mechanisms regulating the entire casein cluster at the genomic level we initiated a multispecies comparative sequence analysis. Despite the high level of divergence at the coding level, these studies have identified uncharacterized family members within two species and the presence at orthologous positions of previously uncharacterized genes. Also the previous suggestion that the histatin/statherin gene family, located in this region, was primate specific was ruled out. All 11 genes identified in this region appear to encode secretory proteins. Conservation of a number of noncoding regions was observed; one coincides with an element previously suggested to be important for β-casein gene expression in human and cow. The conserved regions might have biological importance for the regulation of genes in this genomic "neighborhood."

Keywords: Multispecies comparative sequence analysis; Casein gene; Gene cluster; Mammary gland; Salivary gland; Secretory protein; Milk protein; Sequence conservation

The increasing availability of extensive genomic sequences from different species allows new opportunities for dissecting gene regulation and molecular evolution. Comparative genomics is frequently used to discover new genes and aid in the identification of functional elements, enabling new insights into evolutionary processes [1–3]. Traits that are restricted to a group of organisms can be used to define the subset of species that are used in a comparison. For example, mammals are distinguished from other vertebrates by the act of nursing their young with mother's milk. Milk, the primary source of nutrients for suckling infants, provides all required components for normal growth and development. Caseins constitute the major nutritional proteins in milk and supply basic amino acids, calcium, phosphates, and bioactive peptides (e.g., antimicrobial and opioid) [4]. Understanding the mechanisms underlying production and composition of milk will contribute to development of strategies to improve infant formula, change milk composition, and produce biologically important proteins in milk of transgenic animals. Because milk proteins are produced in differentiated mammary epithelial cells, insight into their regulation will heighten understanding of mammary gland development, which could impact breast cancer treatment and prevention.

The "calcium-sensitive" caseins are encoded by a cluster of three or four evolutionarily related genes depending on the species: one α-s1-like casein (*CSN1S1* in human and cow, *Csn1s1* in mouse and rat), one β-like casein (*CSN2* in human and cow, *Csn2* in mouse and rat), and one or two α-s2-like caseins [*CSN1S2* in cow, *CSN1S2A* in human with its mouse and rat ortholog *Csn1s2a* (α-s2A or γ-casein), and the other α-s2-like casein, *CSN1S2B* in human and *Csn1s2b* (α-s2B or δ-casein) in mouse and rat]. In addition there is

* Corresponding author. Fax: +1-713-798-8012.
E-mail address: rijnkel@bcm.tmc.edu (M. Rijnkels).

0888-7543/03/$ – see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/S0888-7543(03)00114-9

the physically linked and functionally associated κ-like casein gene [human *CSN3* (or *CSN10*), cow *CSN3,* and mouse and rat *Csn3*]. In human, the casein region is located on HSA4q13, in mouse on chromosome 5 (MMU5E2), in rat on chromosome 14, and in cow on chromosome 6 (BTA6q31). Earlier studies have shown that the overall organization of the casein gene cluster and the structures of the individual genes are largely conserved in human, mouse, and cow although the coding regions have significantly diverged both within and between species [5–8].

The proximal control elements responsible for the hormonal, developmental, and cell-specific regulation of casein gene expression have been studied in cell culture and transgenic mouse models [9,10]. Clusters of *trans*-acting factor binding sites known as composite response elements are present in the proximal promoters of the casein genes and upstream enhancers of the β- and α-s1-like genes [9–15]. The transcription factors that recognize and bind to these sites are not exclusively expressed in the mammary gland, and their DNA binding sites are often not the high-affinity consensus binding sites identified for these factors. It appears that a combination of protein–DNA and protein–protein interactions is required to confer tissue- and developmental stage-specific expression of the milk protein genes.

To understand the structure of the entire casein gene cluster and the mechanisms regulating its expression, we analyzed the region in the genome that harbors the casein locus. For instance, the evolutionary conservation of sequences in regions other than the transcriptional unit and promoter of a gene may indicate the presence of functional elements. Such regions may be involved in the overall regulation of the casein genes. To identify and characterize conserved regions, we initiated a multispecies comparative sequence analysis of the casein gene cluster region using sequence data from human, cow, mouse, and rat. Humans, rodents, and cattle are at approximately equal evolutionary distances (80–100 Myr) [19,20]. The use of several distantly related species increases the analytical power of a comparative analysis [16–18] by limiting the identification of homologies occurring solely by chance. In the case of the casein gene region, i.e., a purely mammalian gene cluster, the choices of species are limited to mouse, human, and a few other species for which sequence or BAC libraries are available. Cattle are the principal farm animal used in generating dairy products and, therefore, an obvious choice. Furthermore, the casein genes have been studied extensively in this species, and BAC libraries are available. Inclusion of the rat sequence in this analysis provides additional support to the results obtained in the three-way comparison. In this analysis, we show conservation of the organization of the casein gene cluster region and known regulatory regions across all species examined. We identify several genes in this region, some of which are uncharacterized, and several distal regions of conservation between all species that might be involved in regulation. The use of a third species with the same evolutionary distance from human, but different evolutionary history, significantly increases the stringency of the comparative analysis.

Results

Isolation and characterization of BACs

Casein gene cDNAs were used to isolate clones from human, mouse, and cow BAC libraries. Restriction maps were determined for the isolated BACs and confirmed by comparison to previously established genomic and YAC-based physical maps [5–7]. Newly isolated overlapping BACs were sequenced along with a previously isolated bovine cosmid (AY154895). BAC, cosmid, and other sequences available in GenBank were used for comparative sequence analysis (see Materials and Methods).

Comparative analysis in human, mouse, and bovine

We used PipMaker [21] and MultiPipMaker (both at http://bio.cse.psu.edu) to compare the sequence of the human, mouse, and cow casein gene regions. These sequences contain up to 40 kb of flanking regions in the human, 12 kb upstream and 25 kb downstream in mouse, and roughly 40 kb upstream and 5 kb downstream in bovine (Fig. 1). Conserved linkage is seen between these human, mouse, and bovine genomic intervals (Fig. 1), and comparison of annotations in regions flanking the human and mouse intervals implies that it extends much further. Furthermore, these comparative analyses show more similarity between the human and the cow sequences than between human and mouse or mouse and cow. In fact, mouse and cow share the least sequence conservation; the overall percentage identity in alignments is 67% for human/mouse, 73% for human/cow, and 66% for the mouse/cow alignment. When measured as the percentage of nucleotides within an alignment, similarity in mouse/cow alignments is the lowest (Table 1). There is no apparent difference in conservation (% ID) between exons, introns, and intergenic sequences in most of the genes in this region (see Pip online).

In general, the casein orthologs are more similar to each other than to the paralogs (*CSN1S1, CSN2,* and *CSN1S2*). Nevertheless, even the casein orthologs are not conserved to a high degree. In fact, the most conserved parts are the 5′ and 3′ UTR (~60%) and the signal peptide-encoding exon 2 (70–80%), which is usually very short (~63 bp). The mature peptides have 30–40% identity. The molecular diversity of the casein genes is achieved through variable use of exons in different species and high evolutionary divergence.

The coding exons in the casein orthologs and paralogs, as well as the other noncasein genes in this region, are phase class 0 (no coding triplet is interrupted). The nonsynonymous and synonymous substitutions are almost equal in all

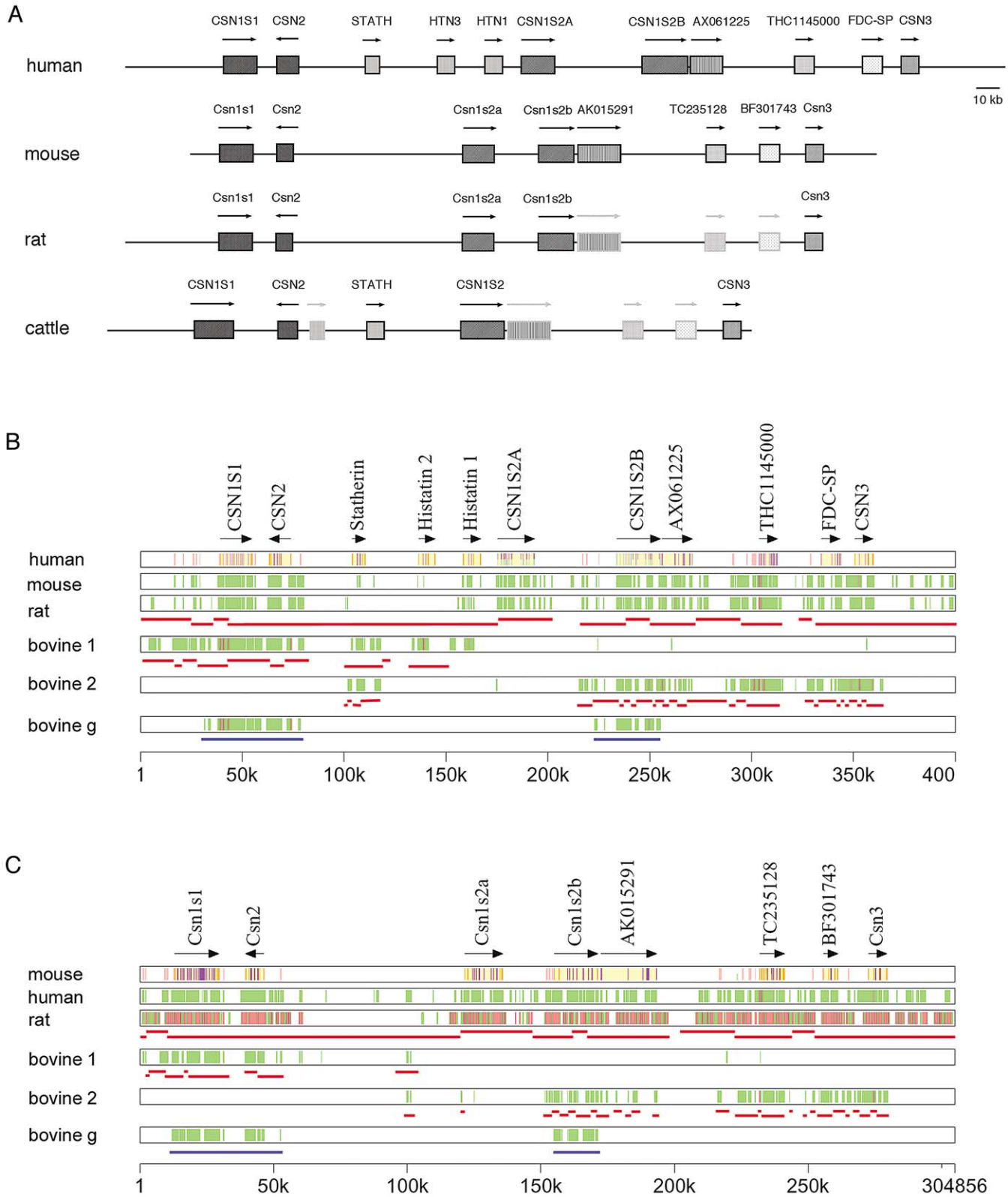


Fig. 1. Casein gene cluster region overview. (A) Overview of organization of the genes identified in the casein gene cluster region in human, mouse, rat, and cow. Orthologous genes are indicated by identical shading, genes whose presence is predicted based on comparative analysis but have not been verified by expression analyses or presence of matching sequences in the EST databases are depicted in light gray. (B) A graphic overview of the multispecies alignment from the MultiPipMaker server. The position and transcriptional orientation of each gene identified in this region are indicated above the image. The colored underlay in the uppermost panel represents various genomic features like introns (light yellow), coding exons (light purple), and 5′ and 3′ UTRs (orange). Sequences that meet specific criteria for conservation in human, mouse, and bovine are colored light red. Lower panels represent pairwise alignments between

Table 1

	NR[b]	NR_NE[b]	NR_IG[b]
% Aligning[a]			
H/M[c]	34.8	28.9	22.9
H/C[c]	57.6	45.3	39.0
M/H[c]	41.5	34.9	27.0
M/C[c]	32.4	21.9	17.9
% Masked[d]			
Human	41	43	46
Mouse	35	38	60
G+C content[e]			
Human	32	31.7	32.5
Mouse	34.1	33.7	34.4

[a] Percentage of nonmasked nucleotides in alignment.

[b] NR, nonrepetitive sequences: repeats identified by RepeatMasker were masked. NR_NE, nonrepetitive nonexonic sequences: repeat and annotated exonic sequences were masked. NR_IG, nonrepetitive intergenic sequences: repeat, exonic, and intronic sequences were masked.

[c] Human or mouse sequences masked as indicated ([b]) were used as reference sequence in PipMaker analysis. H/M, human vs mouse; M/H, mouse vs human; H/C, human vs cow; M/C, mouse vs cow. First species indicates sequence used in calculations.

[d] Percentage of total sequence masked.

[e] G+C content in nonmasked sequence.

genes, indicative of low or no evolutionary constraint (data not shown). One exception, TC235128/THC1145000, has a nonsynonymous vs synonymous substitution ratio of 0.38, suggesting moderate selective evolutionary pressure.

Genes identified in the casein gene cluster regions

In addition to the known genes (*CSN1S1, CSN2, CSN3, STATH, HTN1,* and *HTN3*) there are five predicted genes in the 400 kb containing the human casein gene cluster. Two match EST sequences with predicted proteins of unknown function (AX061225, THC1145000) and one is the recently identified follicular dendritic cell secreted protein (FDC-SP [22]). Two novel α-s2-like genes were identified by comparison to mouse and cow sequences (Table 2 and Figs. 1 and 2) and fall in orthologous positions between CSN2 and CSN3.

Database searches identified three new (noncasein) genes in the 304-kb mouse region that correspond to three (assembled) ESTs (AK015291, TC235128, BF301743). These sequences show similarity to the human ESTs in the orthologous positions (Table 2).

Other than the bovine casein genes no additional bovine transcripts mapped to this region in database searches, but comparative analyses indicate the presence of several conserved regions at positions orthologous with human genes and ESTs, including a bovine *STATH*-like gene.

Nucleotide level analyses of predicted α-s2-like casein genes in the human and rat CSN gene clusters

Two potential human α-s2-like casein genes are found at positions 175466–193353 and 233857–255236 (Fig. 1 and Pip online). Based on the nomenclature of the bovine α-s2-casein gene (*CSN1S2*) and the homology to the rabbit genes (see below) we propose the following gene symbols: *CSN1S2A* for the gene at positions 175–193 kb (α-s2-like casein A) and *CSN1S2B* for the gene at positions 233–255 kb (α-s2-like casein B). *CSN1S2A* has significant homology to both the rabbit α-s2A signal peptide and 3′ UTR. Similarly, *CSN1S2B* has homology to rabbit α-s2B cDNA in exon 2, also encoding the signal peptide. The structure of the putative human *CSN1S2*-like casein genes and their relationship to the bovine and mouse homologues (*CSN1S2* and *Csn1s2a* and *Csn1s2b*, respectively) was examined with PipMaker (http://bio.cse.psu.edu [21]) and *sim4* (http://biom3.univ-lyon1.fr/sim4.html [23]). The combined analyses identified 19 potential exons for *CSN1S2B* and 16 for *CSN1S2A* (light green, light blue, and light pink bars in Fig. 2A and Pip online). One human EST (AW104440) overlaps several putative exons in *CSN1S2B* (indicated by light and dark purple bars in Fig. 2A and Pip online) and initiates from a repetitive element (Mer-1B) at 241685, producing a truncated form of the transcript.

Based on the comparative analysis, primers for RT-PCR were designed using human sequences with the highest homology to bovine and/or mouse exons (Fig. 2). A primer combination for exon 2 and 7 (E2–E7) detected two transcripts, lacking either putative exon 6 or both exons 5 and 6. A 720-bp cDNA encompassing putative exons 2–4, 7–12, and 14–16 of human *CSN1S2A* (see Fig. 2A) was generated using the E2 primer in combination with a dT reverse primer. When this sequence was extended by 10 bp based on homology to include the upstream start codon, its translation revealed a premature stop codon in the first position of exon 4. The resulting 85-nt open reading frame encodes a 27-aa peptide encompassing the signal peptide and the first potential phosphorylation site. It is not known if this transcript is either translated or stable and functional as a peptide. No transcripts that are specific for the *CSN1S2B* gene have been identified in human lactating mammary gland RNA. The same premature stop codon is detected in

human (finished) and mouse (finished), rat (draft contigs), or bovine [draft contigs (bovine 1 and bovine 2) and finished sequence for the bovine *CSN1S1, CSN2,* and intergenic region and *CSN1S2* gene (bovine g)]. Green and red bars represent a measure of the alignment quality, i.e., gap-free alignments of at least 100 bp and 70% identity are colored red, whereas all other alignments are green. The locations of contigs of draft and finished sequences are indicated by red (draft) and blue (finished) horizontal lines underneath the panel representing the pairwise comparison of the particular sequence to the reference sequence. (C) As B for the pairwise comparison between mouse and human, rat, or bovine sequences.

Table 2

Human			Mouse			Cow			Function
Gene (gene symbol)	Tissue[a]	Accession No.[b]	Gene (gene symbol)	Tissue[a]	Accession No.[b]	Gene (gene symbol)	Tissue[a]	Accession No.[b]	
α-casein (CSN1S1)	$MG^{l/r}$	NM_001890	α-s1 casein (Cns1s1)	MG^{1}	NM_007784[c]	*α-s1 casein (CSN1S1)*	MG^{1}	M33123 *X59856*	Nutrition
β-casein (CSN2)	MG^{1}	NM_001891[c] *AF027807*	*β-casein (Csn2)*	$MG^{l/r}$	NM_009972 *X13484*	*β-casein (CSN2)*	MG^{1}	X06359 *X14711*	Nutrition
Statherin (STATH)	SG MG	NM_003154	?			*Statherin (STATH)*	SG MG	AY154893	Antimicrobial
Histatin 2 (HTN3)	SG	NM_000200	?			?			Antimicrobial
Histatin 1 (HTN1)	SG	NM_002159	?			?			Antimicrobial
α-s2A casein (CSN1S2A)	MG SG	NM_173085	*α-s2a casein (Csn1s2a)*	MG^{1}	NM_007785	——			Nutrition
α-s2B casein (CSN1S2B)	—	AW10440	*α-s2b casein (Csn1s2b)*	$MG^{1/r}$	NM_009973[c]	*α-s2 casein (CSN1S2)*	$MG^{l/r}SG^{r}$	M16644 *M94327*	Nutrition
AX061225	T	AX061225	AK015291	T	AK015291	?			Transporter
THC1145000	SG MG	AK000520	TC235128	SG MG	AK009298 AA028229	?			Unknown
FDC-SP	SG MG	AF435080	BF301743	SG MG	BF301743	?			Unknown
κ-casein (CSN3)	MG^{1}	NM_005212 *U51899*	*κ-casein (Csn3)*	MG^{1}	NM_007786[c]	*κ-casein (CSN3)*	MG^{1}	X00565	Nutrition

Note. —, not detected; ——, gene not present; ?, presence of gene suggested by comparative analysis, expression not determined yet.

[a] Tissue expression as determined from RT-PCR (r) or for casein genes previously established by Northern blotting and/or from literature (l). MG, mammary gland; SG, salivary gland; T, testis.

[b] GenBank accession number, Refseq is given when available, representative overlapping EST accession numbers are given for EST contigs. Accession No. for genomic sequences are given in italic.

[c] Partial/no 5′-UTR sequence.

CSN1S2B based on the comparison with bovine *CSN1S2* and mouse *Csn1s2b*. For both the *CSN1S2A* and the *CSN1S2B* transcripts amino acid homology extends beyond the premature stop.

The comparative analysis indicated the presence of a rat δ-casein gene at a position orthologous to mouse *Csn1s2b* and human *CSN1S2B*. To determine the sequence of the rat *Csn1s2b* mRNA, RT-PCR was performed with forward and reverse primers corresponding to exons 2 and 12, respectively. Transcripts that correspond to the predicted exons were detected from lactating mammary gland RNA. A 767-bp product (rat δ-casein, AY154894) that extends by homology to encode a 169-aa protein has 55% identity to mouse *Csn1s2b*. The rat *Csn1s2b* transcript differs from the mouse in that exon 3, encompassing the first major site of phosphorylation, is absent although present at the DNA level. Together with the use of an alternative exon 7 (rat) (homologous to bovine exon 10 and putative human CSN1S2B exon 11) this leads to a predicted protein that is 17 aa longer. Based on the genomic alignments these differences seem to result from alternative splicing events (Fig. 2B).

Phylogenetic analyses of α-s2-like casein genes

The comparative analysis indicates the presence of duplicated α-s2-like casein genes in *Homo sapiens*- and Rodentia-containing clades but only one gene in Artiodactyla. MultiPipMaker (http://bio.cse.psu.edu/) comparisons of the α-s2-like casein genes indicate that the rodent δ-casein gene (*Csn1s2b*) and human *CSN1S2B* have more similarity to bovine *CSN1S2* than to the rodent γ-casein gene (*Csn1s2a*) and human *CSN1S2A*. Furthermore, rodent *Csn1s2a* has more similarity to human *CSN1S2A* than to the other α-s2-like genes. This is reflected in the phylogenetic tree that was constructed using the cDNA sequences of the α-s2-like genes of pig, cow, camel, rabbit, mouse, rat, and human, as deduced from RT-PCR for *CSN1S2A* and the predicted cDNA for human *CSN1S2B*. The α s2B-like casein genes of mouse, rat, human, and rabbit cluster with the *CSN1S2* genes of Artiodactyla, whereas the *CSN1S2A*-like genes cluster together (Fig. 2). From this unrooted tree it is not possible to confirm when the duplication occurred; however, it suggests that a duplicated gene existed in the common ancestor of human, rodent, and Artiodactyla. The second α-s2-like gene was lost in the Artiodactyla, while further divergence occurred in both copies in the other species.

Noncasein genes in the casein gene cluster region

The large intervals between the *CSN* genes, 75–100 kb between *CSN2* and *CSN1S2*-like genes and 95–100 kb between *CSN1S2B*-like genes and *CSN3*, and the presence of sequence conservation indicate that these intergenic regions could putatively contain other genes.

The CSN2_CSN1S2 interval: histatin/statherin gene family

Physical mapping and BLAST searches show the presence of the histatin/statherin (*HTN/STATH*) gene family, statherin (*STATH*) and histatins 2 and 1 (*HTN3*, *HTN1*) (in this order), on the human BACs isolated from the casein gene cluster region in concordance with the mapping of a 4-Mb YAC contig on 4q11–q21 by Kärrman et al. [24]. This gene family is located between the *CSN2* and the *CSN1S2*-like genes (Fig. 1 and Pip online). These genes are predominantly expressed in the salivary gland and have evolved from a common ancestral gene [25]. *STATH* and *HTN* expression was confirmed in human submandibular gland (SMG) RNA, by RT-PCR. We also detected expression of statherin but not histatin in human lactating mammary gland RNA. Although it was previously thought that this gene family is primate specific [25], comparative analysis indicates the presence of at least two *HTN/STATH* like genes in cow, and remnants of these genes are present in mouse and rat at the genomic DNA level (Fig. 1 and Pip online). Transcripts were detected by RT-PCR analysis in RNA from cow salivary gland tissue. The detected transcript has an open reading frame with a size similar to that of the human genes and a translated protein with 63% identity to human statherin (BT_STATH, AY154893). No mouse transcripts have been identified yet. The exact number, position, and orientation of *HTN/STATH*-like genes in the bovine genome cannot be assessed until more complete sequences become available.

The CSN1S2_CSN3 interval: other genes or ESTs in the casein gene region

In addition to the *CSN* gene family and the *HTN/STATH* gene family, three sequences (Table 2) map to conserved positions in all species. Furthermore, there is considerable homology between species in the potential promoter regions; however, the function of the predicted proteins is unknown.

AX061225 and AK015291

Both the human and the mouse sequences contain a testis-specific EST (AX061225 and *AK015291*, respectively) that corresponds to positions 262334–270956 in the human sequence (Fig. 3A and Pip online). RT-PCR confirms expression of these genes in human and mouse testis. The human EST lacks putative exon 1 compared to mouse (indicated with an asterisk in Fig. 3A and Pip online). Other differences observed between the human and the mouse ESTs (and RT-PCR product) suggest that differential splicing occurs (Fig. 3A). The protein predicted for *AK015291* is likely to be secreted and has homology in its first 17 aa to the casein α/β block (IPB001588) representing the signal peptide in casein genes.

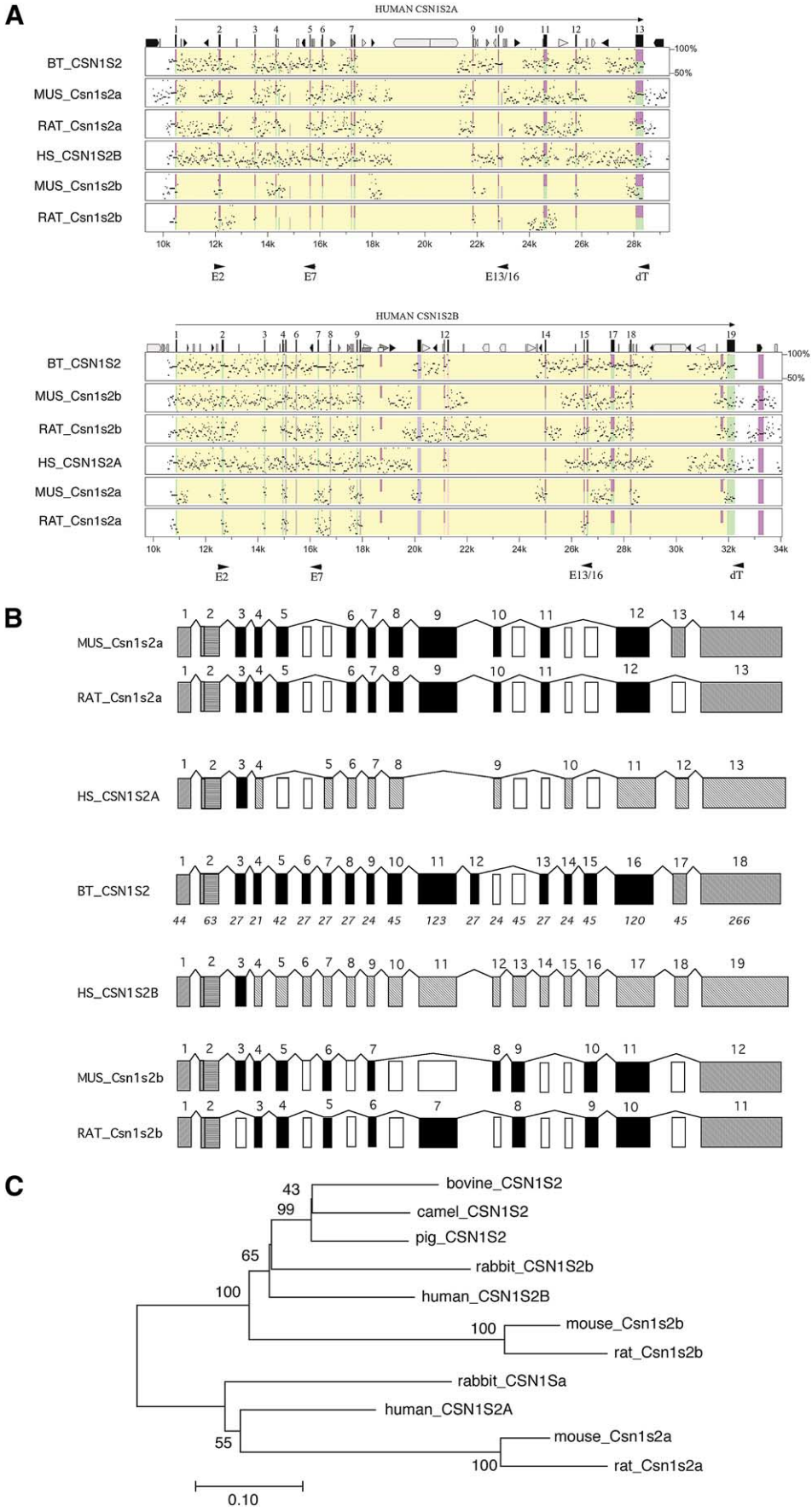
THC1145000 and TC235128

The MultiPipMaker analyses show large regions of sequence conservation between positions 303000 and 313000 that align to the TIGR gene index EST consensus THC1145000 (human) and TC235128 (mouse; Fig. 4B). THC1384761 and THC1384762 are alternative splice clusters for this gene. In comparison with the human sequence, the mouse ESTs contain one extra exon exhibiting very high identity, 87%, with human (indicated with an asterisk in Fig. 4B and MHRC Pip online). Additionally, we identified one rat salivary gland EST (BF417734) in this region, from which two potential exons overlap partially with exons 5 and 7 of TC235128 (light green bars in MHRC Pip online) and two other potential exons appear to be conserved in the human/mouse comparison but not in mouse/cow comparisons. We detected transcripts in both salivary and lactating mammary gland tissue of human and mouse. The predicted proteins are potentially secreted.

FDC-SP

At positions 334222–343332 we mapped the gene encoding follicular dendritic cell-secreted protein (FDC-SP) a protein secreted by follicular dendritic cells isolated from tonsils [22]. The orthologous EST in mouse (BF301743) was isolated from salivary gland tissue. RT-PCR expression analysis identified transcripts for FDC-SP and its mouse ortholog in salivary gland and lactating mammary gland RNA.

Fig. 2. *CSN1S2*-like comparative and phylogenetic analyses. (A) MultiPipMaker, percentage identity plots (pip) for the comparison of the human *CSN1S2*-like gene paralogs *CSN1S2A* (HS_CSN1S2A) and *CSN1S2B* (HS_CSN1S2B) to their orthologous sequences: bovine *CSN1S2* (BT_CSN1S2), mouse γ-casein (MUS_Csn1s2a), rat γ-casein (RAT_Csn1s2a), mouse δ-casein (MUS_Csn1s2b), rat δ-casein (RAT_Csn1s2b) and their respective paralogs. Light yellow indicates intronic regions, light purple indicates exons supported by mRNA or EST evidence. Light blue indicates exons predicted based on comparison to bovine sequences, light pink indicates exons predicted based on comparison to mouse sequences, and light green indicates exons predicted based on homology to mouse and bovine sequences. Location of primers used for RT-PCR analyses is indicated below the pip. (B) Schematic representation of the aligned exon structure of the α-s2-like genes (diagonally hatched blocks indicate 5′ and 3′ UTR, horizontally hatched blocks indicate signal peptide encoding sequences, black blocks represent mature peptide encoding exons, and no-fill blocks represent exons that are not present in mRNA but are present at DNA level. Numbers above blocks indicate exon number; numbers below bovine exons indicate exon size. (C) Phylogenetic tree based on the alignment of cDNA sequences of CSN1S2-like genes of Artiodactyla, cow (bovine_CSN1S2, M16644), camel (camel_CSN1S2, AJ012629), and pig (pig_CSN1S2, X54975); rabbit (rabbit_CSN1S2A, X76907 and rabbit_CSN1S2B, X76909); human (human_CSN1S2A, AY154892 and human_CSN1S2B, from genomic sequence based on comparative analysis); and rodents, mouse (mouse_Csn1s2a, NM_007785 and mouse_Csn1s2b, NM_009973) and rat (rat_Csn1s2a, J00712 and rat_Csn1s2b, AY154894). The tree was constructed with the neighbor-joining method using Jukes and Cantor distance calculations with pair-wise deletion of sequences not in common and bootstrap of 1000. Bootstrap values >40 are indicated at branch points. Branch length is given as number of substitutions per site.


A
HUMAN CSN1S2A
BT_CSN1S2
MUS_Csn1s2a
RAT_Csn1s2a
HS_CSN1S2B
MUS_Csn1s2b
RAT_Csn1s2b
100%
50%
10k 12k 14k 16k 18k 20k 22k 24k 26k 28k
E2 E7 E13/16 dT
HUMAN CSN1S2B
BT_CSN1S2
MUS_Csn1s2b
RAT_Csn1s2b
HS_CSN1S2A
MUS_Csn1s2a
RAT_Csn1s2a
100%
50%
10k 12k 14k 16k 18k 20k 22k 24k 26k 28k 30k 32k 34k
E2 E7 E13/16 dT
B
MUS_Csn1s2a
RAT_Csn1s2a
HS_CSN1S2A
BT_CSN1S2
44 63 27 21 42 27 27 27 24 45 123 27 24 45 27 24 45 120 45 266
HS_CSN1S2B
MUS_Csn1s2b
RAT_Csn1s2b
C
bovine_CSN1S2
camel_CSN1S2
pig_CSN1S2
rabbit_CSN1S2b
human_CSN1S2B
mouse_Csn1s2b
rat_Csn1s2b
rabbit_CSN1Sa
human_CSN1S2A
mouse_Csn1s2a
rat_Csn1s2a
43
99
65
100
100
55
100
0.10

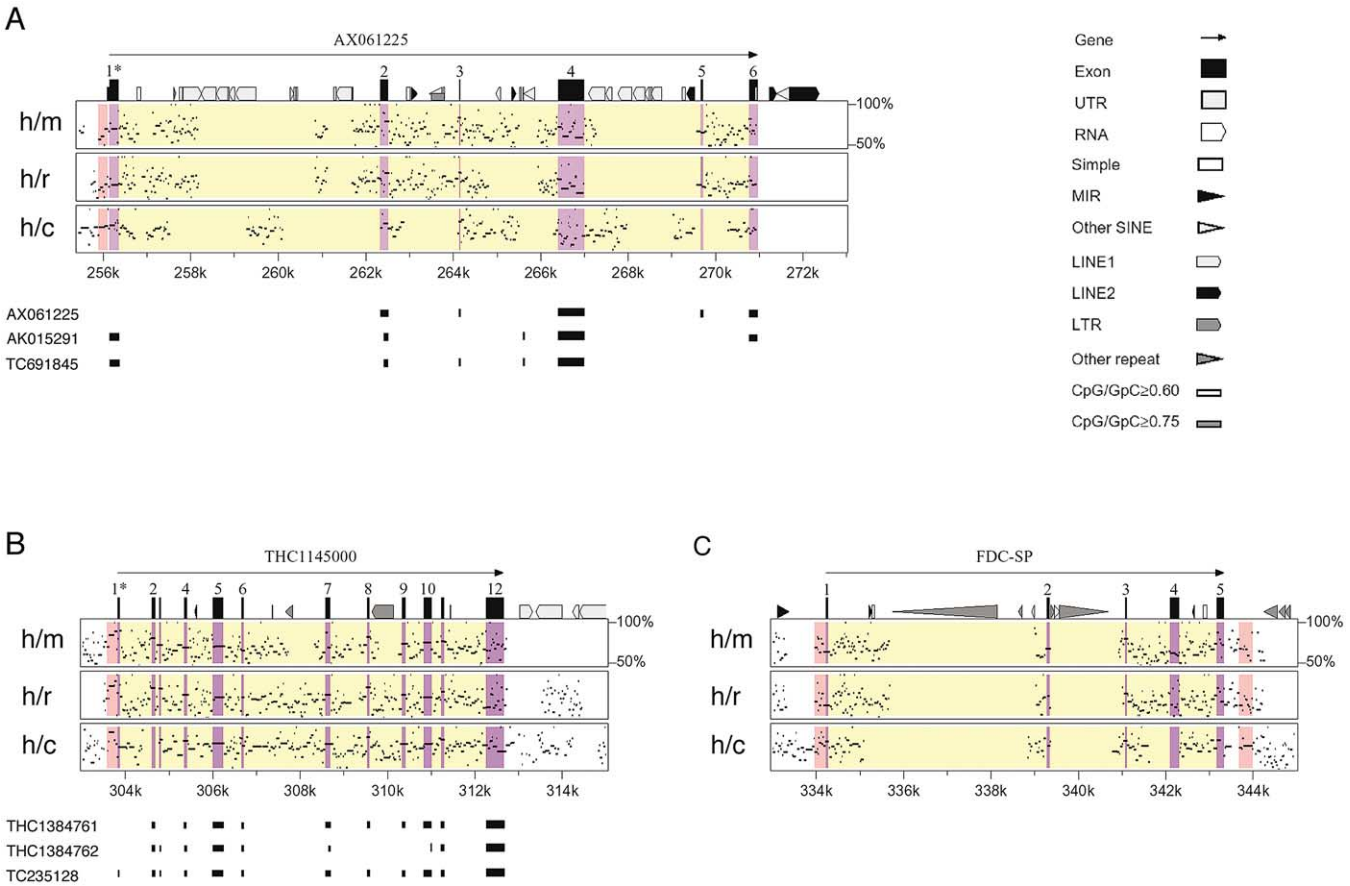
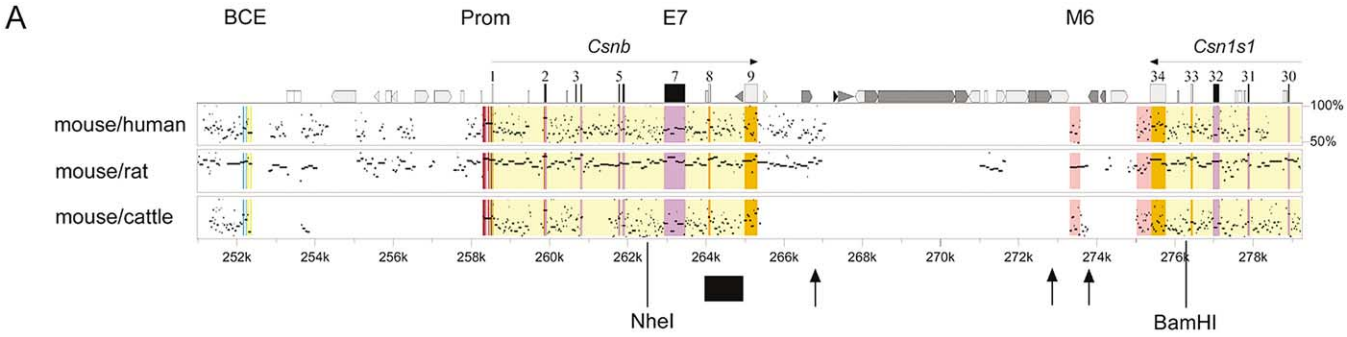


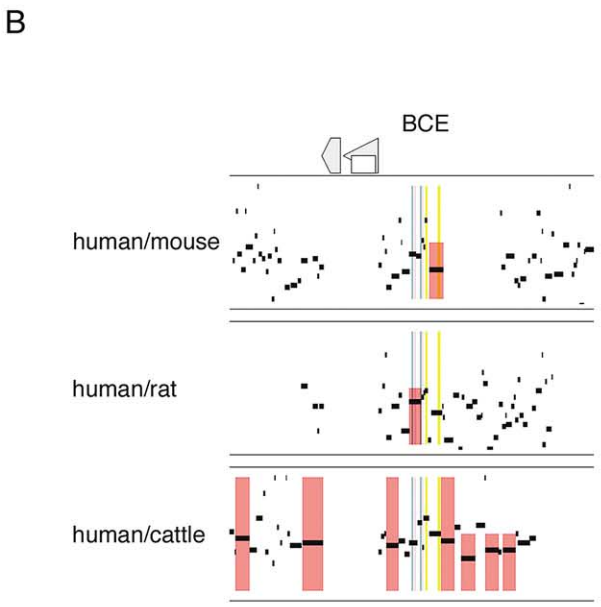
Fig. 3. Genes identified in the *CSN1S2_CSN3* interval. Portions of the pip containing the genes show the pattern of conservation between the human/mouse (h/m), human/rat (h/r), and human/cow (h/c) sequences. Alternative splice forms in human and/or mouse are indicated beneath the pip. (A) AX061225 (human, patent), AK015291 (mouse, testis), TC691845 (mouse TIGR gene index, spermatid). (B) THC1145000 (human TIGR gene index), splits into alternative splice forms THC1384761 and THC 1384762; TC235128 (mouse TIGR gene index). (C) FDC-SP. Light yellow indicates intronic regions; light purple indicates exons supported by cDNA or EST evidence; light red indicates regions outside of transcriptional units that are conserved in human, mouse, and bovine. Asterisk indicates exon sequence present in mouse EST or TIGR gene index sequence but not in human EST.

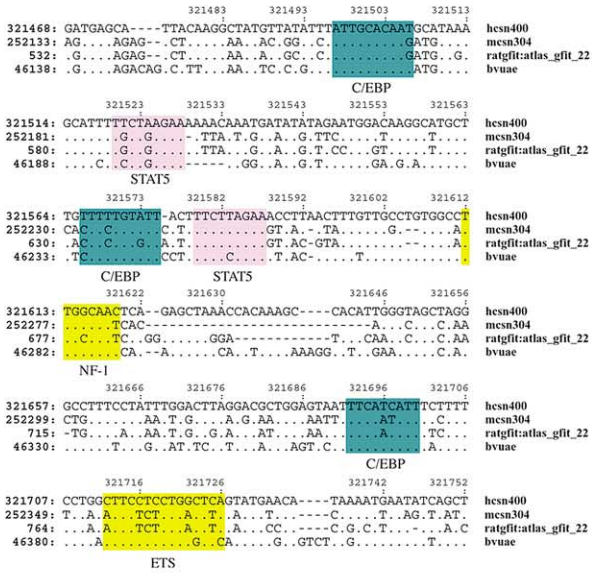
Identification of noncoding evolutionarily conserved regions (N-ECRs)

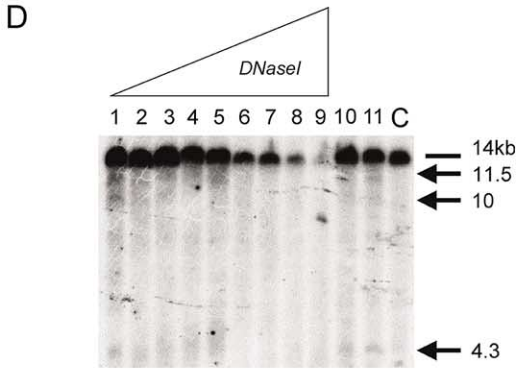
One of the goals of this study is to identify N-ECRs that serve as indicators of potential regulatory regions [3,16]. Comparisons between human and cow show the most similarity, whereas the mouse/cow comparison appears to be the most stringent (similarity M/C < H/M < H/C, see Table 1). We identified 28 regions outside of transcriptional units (genes) that were conserved in all three comparisons (M/C, H/M, H/C) (Table 3). All regions present in comparisons with mouse were also present in human/cow comparisons, but not all human/mouse or human/cow regions were present in mouse/cow comparisons. Of the 28 regions conserved in all three species, 11 are located immediately 5′ or 3′ of genes and often overlap the 5′- or 3′-terminal exon, 5 overlap ESTs or genes not present/identified in mouse (AA865052 unspliced EST, *HTN* and *STATH*, not further

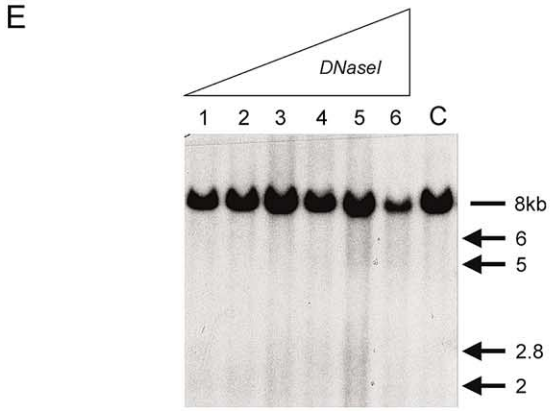
Fig. 4. β-Casein enhancer (BCE) and conserved region (M6) analysis. (A) Pip comparing the mouse *Csn2_Csn1s1* gene and 5′ flanking region to human, rat, and cow *CSN2* sequences. The BCE and proximal promoter are marked by colored bars representing the different transcription factor binding sites identified, C/EBP, dark cyan; STAT5, light pink; NF1, ETS, or YY1, light yellow; GR, dark red; and TATA, dark blue. See also B and C. (B) Close-up of BCE region in pip with human sequence as reference. The different transcription factor binding sites are indicated by colored bars (see A). Full-length light red bars indicate sequence with >70% homology over >100 bp (100_70), half-length light red bars indicate sequences with >58% homology over >100 bp (100_58). (C) Nucleotide-level alignment view of BCE region generated by MultiPipMaker. Transcription factor binding sites previously identified are indicated, see also A. (D and E) DNase I hypersensitive site analysis of M6 region in the mouse and cow *Csn2_Csn1s1* intergenic region. Nuclei from lactating mammary gland tissue of mouse (D) and cow (E) were analyzed with increasing amounts of DNase I (mouse, lanes 1–9, lanes 10 and 11 same DNase I-treated DNA as lane 1; cow, lanes 1–6). Lane C, control non-DNase-I-treated DNA. Genomic DNA was digested with *Nhe*I/*Bam*HI (mouse) or *Nhe*I (cow). Southern blot analysis was performed with radionucleotide-labeled probes as indicated in A (mouse) and E (cow). Sizes and locations of DNase I-sensitive fragments are indicated with arrows.

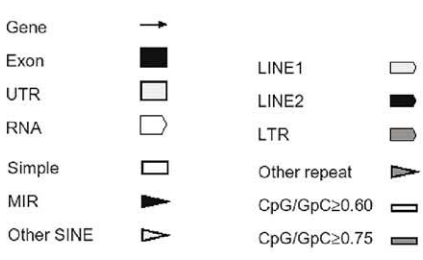

A
BCE
Prom
E7
M6
Csnb
Csn1s1
mouse/human
mouse/rat
mouse/cattle
100%
50%
252k
254k
256k
258k
260k
262k
264k
266k
268k
270k
272k
274k
276k
278k
NheI
BamHI


B
BCE
human/mouse
human/rat
human/cattle


C
C/EBP
STAT5
C/EBP
STAT5
NF-1
C/EBP
ETS
hesn400
mesn304
ratgfit:atlas_gfit_22
bvuae


D
DNaseI
1 2 3 4 5 6 7 8 9 10 11 C
14kb
11.5
10
4.3


E
DNaseI
1 2 3 4 5 6 C
8kb
6
5
2.8
2


Gene
Exon
UTR
RNA
Simple
MIR
Other SINE
LINE1
LINE2
LTR
Other repeat
CpG/GpC≥0.60
CpG/GpC≥0.75

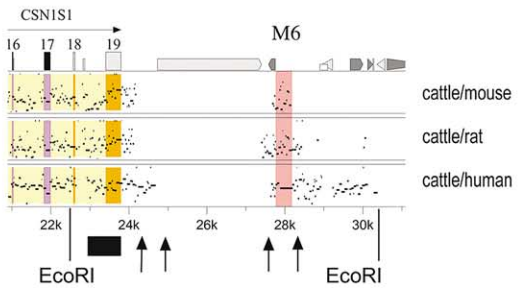

CSN1S1
M6
cattle/mouse
cattle/rat
cattle/human
22k
24k
26k
28k
30k
EcoRI
EcoRI

Table 3

ECR[a]		Mouse/cow Size (% ID)	Human/mouse Size (% ID)	Human/cow Size (% ID)
M1A		**118 (58)**[b]	**115 (68)**[b]	**235 (72)**[b]
M1ext		**450 (64)**	**450 (70)**	**487 (74)**
M2		**194 (62)**	**193 (69)**	**210 (73)**
M3A		**483 (65)**	**502 (65)**	**571 (72)**
M3B		**353 (59)**	**171 (64)**	**123 (67)**|[c] **143 (73)**[b]
M3C		**47 (72)**[b]	**47 (72)**[b]	**25 (68)**[b]
M4	CSN1-5′[d]	529 (69)	489 (69)	500 (79)
		169 (75)	176 (74)	162 (86)
M5	CSN1-3′[e]	323 (67)	373 (70)	355 (78)
M6	**MII a/b3′**	**259 (61)**	**256 (63)**	**315 (71)**
M7	CSN2-3′	NA	NA	NA
M8	CSN2-5′	212 (74)	233 (76)	306 (75)
		123 (74)[b]	157 (76)[b]	123 (80)[b]
M9	**BCE**	**444 (66)**	**490 (66)**	**500 (73)**
M10	STATH+5	NA	NA	NA
M11	STATH	NA	NA	NA
M12	HTN3	NA	NA	NA
M13	HTN3	NA	NA	NA
M14		**297 (66)**	**371 (65)**	**306 (80)**
M15		**182 (69)| 281 (64)**	**187 (67)| 281 (67)**	**582 (74)**
M16	Mouse/rat/human Csn1s2b	—	183 (64)[a]	—
M17	CSN1S2B/Csn1s2b	245 (71)	255 (70)	254 (75)
M18	AX/AK-5′	245 (62)	201 (62)	216 (77)
M19		**755 (66)**	**763 (67)**	**859 (76)**
M20A		**195 (68)**	**114 (69)**	**258 (79)**
M20ext		**520 (66)**	**418 (68)**	**464 (78)**
M21		**696 (68)**	**678 (69)**	**764 (77)**
M22		**152 (70)**	**157 (73)**	**153 (77)**
M23	THC-5′	213 (80)	224 (83)	222 (87)
		145 (83)[b]	116 (86)[b]	108 (94)[b]
M24	EST AA	NA	NA	NA
M25		**293 (61)**	**301 (68)**	**348 (70)**
M26	FDC-SP-5′	231 (66)	238 (71)	252 (78)
M27	PDC-SP-3′	248 (68)	257 (65)	300 (77)
M28	CSN3-5′	107 (72)	101 (84)	102 (75)
M29	CSN3-3′	204 (67)	208 (71)	211 (77)

Note. NA, not analyzed; bold indicates ECR located in intergenic regions, not immediately flanking genes.

[a] Indicated as light red bars in pip-plot (see supplemental data).

[b] Ungapped alignment.

[c] |, alignment gap >25 bp.

[d] -5′, sequence immediately 5′ of gene.

[e] -3′, sequence immediately 3′ of gene.

analyzed), and 12 are in noncoding distal regions (bold in Table 3).

BCE (M9)

One of the conserved regions coincides with a β-casein enhancer (BCE), previously identified in the distal promoter region of the human and cow β-casein gene (*CSN2*) [11,12]. In human, this BCE is located 4.5 kb upstream of the gene and in cow, 1.6 kb. The homologous region in mouse is located 6.5 kb upstream of *Csn2* (Fig. 4A) and roughly at −3.5 kb in rat. This region was, therefore, used as a paradigm for the identification of potential regulatory regions. Previous analyses have used the criteria of >100 bp and >70% homology of ungapped sequence alignment to identify potential regulatory regions from comparative analyses (100_70) [3]. In the casein gene region, only the proximal promoters show conservation that is ungapped and around 75% (see Table 3). When we applied the 100_70 or a 100_58 (>100 bp and >58%) rule several N-ECRs were identified (only three in human/mouse with 100_70). Some of these overlapped between H/M and H/C comparisons. In the BCE region several 100_70 and 100_58 ECRs were identified. However, upon closer inspection they barely overlapped between human and mouse and flanked the previously described BCE sequence in cow (Fig. 4B). Potential *trans*-acting factor binding sites identified in the human and cow BCE show conservation of all these sites, although different levels of identity and numbers of gaps were present (Fig. 4C).

We applied the tool “strong-hits” (from the PipTools

Table 4

	Hum. CSN 400 kb	1.4 MB 4q13[a]	Hum genome draft[b]	*Mus* CSN 204 kb	*Mus* genome draft[b]	Bov. CSN[c] 396 kb
SINE	5.77	12.14	13.64	3.71	8.22	8.07
LINE	25.78	24.87	20.99	22.37	19.20	29.10
LTR	4.71	8.03	8.55	4.04	9.87	2.57
DNA elements	2.91	2.96	3.03	0.28	0.88	0.91
Unclassified	0.38		0.15	0.95	0.38	0
Total int.[d]	39.56	48.23	46.36	31.35	38.55	40.65
Rest[e]	1.84	1.41	1.25	4.19	2.63	1.56
Total	41.40	49.64	47.61	35.53	42.18	42.22
GC	34.35	37.00	41	36.47	42	37.42

[a] NT_006281.
[b] From [26,30].
[c] Contigs including ≃20 kb overlap.
[d] Interspersed repeats.
[e] Small RNA, satellites, simple repeats, low complexity.

package; Elnitski et al., 2002; available at http://bio.cse.psu.edu/) to calculate the overall conservation of the N-ECRs using different parameters per comparison, which depended on the BLASTZ output for a particular region (Table 3). The concise files were edited by hand to generate regions of comparable size (human/cow tended to generate very large ECRs with the more relaxed parameters). The results in Table 3 show that the conservation in these ECRs is similar between human/mouse and mouse/cow, >60%, but almost always <70%, and not different from the average similarity for these species. Human/cow ECRs are all >70%, reflecting an overall sequence similarity in this region of 73%. The BCE was identified as a 444- to 500-bp region with 66–73% conservation. Preliminary analysis of the histone acetylation status using ChIP assays on the BCE in lactogenic hormone (prolactin, glucocorticoid, and insulin)-stimulated mouse mammary epithelial cells indicates a hormone-dependant acetylation of H3 and H4 in the BCE, similar to the proximal promoter (M. Kallesen, personal communication.). This suggests functionality of BCE in the mouse.

M6.

In earlier studies we identified a DNase I hypersensitive side (HS) in the CSN1_CSN2 intergenic region of mouse and cow (Figs. 4D and 4E). In the human/cow comparison a 314-bp, 71% conserved (ungapped) ECR coincides with this HS. In mouse/human and mouse/cow an ~260-bp (at 61–63% identity) ECR is identified using the gapped approach (M6). The potential functionality of this ECR is under investigation.

At present we cannot definitively rule out the possibility that the N-ECRs represent unknown genes or are part of nearby transcripts. However, they do not overlap with any exon predictions, ESTs, or ORFs. Thus, they represent good candidates for regions involved in the regulation of adjacent or more distal genes. It is clear that the introduction of a third species increases the stringency of the ECR search.

Repeat analysis

The repeat content in the genomic sequences was analyzed with RepeatMasker (http://ftp.genome.washington.edu/RM/RepeatMasker.html; version 07/07/2001) (Table 4). The repeat distribution in the human sequence deviates slightly from the human genome average ([26] Table 4). For instance, the human *CSN* gene region shows a diminished total interspersed repeat content, a higher percentage of LINE repeats, and a significantly lower SINE content compared to the human genome draft [26]. Human chromosome 4 (HSA4) is one of several chromosomes that have an above-average LINE content (others include HSA13, 15, and 18 and the X chromosome [27]). The average SINE content in the genome is comparable to that of the 4q13 region, although both are twofold higher than the *CSN* region. The LTR content is also lower in the human *CSN* region compared to the 4q13 region and the genome average.

The low SINE content in the *CSN* region correlates with a GC content that is lower than the genome average, 34.5% vs 41%, respectively. The GC content in the larger 4q13 region (37%) is close to the chromosome 4 average of 38% [28]. A low G+C has been correlated to low SINE/high LINE content and low gene content [29]. However, this latter relationship is not observed in the *CSN* region, because as we report above, there are at least 11 genes in a 320-kb (400 kb, including 40 kb flanking on each side) region, or roughly 1 gene/29 kb (1/36) while the HSA4 average is 0.23 genes/29 kb (or 1 gene every 125 kb) [26].

The repeat and GC content in the mouse *Csn* gene cluster is very similar to that in the human region and the deviations from the genome average mimic those observed for the human *CSN* region (Table 4 [30]). A lower overall repeat content in the mouse correlates with a more compact size of this region compared to human (260 vs 320 kb, respectively). In addition the region of the mouse casein gene cluster has fewer genes [8]. The shorter length combined with fewer genes results in an average gene density (1/33 kb) that is comparable to the orthologous human sequence. The GC

content of 36.4% is much lower than the 42% genome average reported [30].

The bovine casein gene cluster region has a gene density that is similar to the human and mouse regions, 1/32 kb. The SINE-to-LINE ratio is similar in the cow sequence even though the SINE and LINE content is somewhat higher than in human and mouse. This could reflect a difference in the content of the repeats in ruminants. The GC content is 37.47%.

Discussion

Comparative sequence analysis is a very powerful tool to identify genes, either totally unknown or by homology to known genes in a different species, as demonstrated here for the α-s2-like genes and histatin/statherin genes. Gene prediction tools are not very efficient in predicting genes with small exons. Therefore comparative analysis and multiple sequence alignments can be used to identify such genes in different species. The assembly of genome sequencing efforts can also be facilitated by multispecies comparisons [31], although repetitive elements and paralogs can complicate these analyses.

Most genes identified in the casein gene cluster region are expressed in the mammary and/or salivary gland and possibly other secretory epithelial tissues. We observed low levels of statherin expression in human and bovine lactating mammary glands, in addition to its expression in SMG. Furthermore, we detected α-s2-like casein expression in bovine and human salivary and lactating mammary glands. In fact, the α-s2-like casein genes and the *HTN/STATH* gene family share similarity in gene structure and function. For instance, the *HTN/STATH* proteins have a signal peptide with similarity to the casein signal peptides that is identifiable via a BLOCKs analysis. At the gene level the bovine and human α-s2-like genes exhibit homology with exon 2 of the *HTN1/3* and *STATH* genes, which encode the signal peptide, as identified in MultiPipMaker alignments, and both gene families have similarly sized small exons. The *HTN/STATH* genes encode proteins that have antimicrobial properties and play a role in calcium homeostasis in saliva. Several naturally occurring peptides derived from caseins have antimicrobial and opioid properties [4] and caseins are the major regulators of calcium homeostasis in milk. In addition, the protein kinases that phosphorylate *STATH* and *HTN* share similarities to the casein kinases in the mammary gland that phosphorylate caseins *in vivo* [32]. These observations coupled with the physical linkage and conservation suggest a common ancestry for the two gene families. However, their high rate of divergence precludes detection of more direct evidence for such a relationship. The predicted proteins, encoded by the neighboring genes FDC-SP and AX061225/AK015291, contain some homology to casein signal peptides and are potentially secreted. FDC-SP expression was identified in follicular dendritic cells in tonsil, trachea, lymph node, and prostate and at low levels in thyroid, stomach, and colon. Mammary and salivary gland tissues were not analyzed in the study by Marshall et al. [22] but we have identified transcripts for both the human and the mouse gene in these tissues. FDC-SP binds B cells and might play a role in the immune response in epithelial tissues.

Another gene family that encodes secretory proteins expressed in the salivary and lachrymal gland is located in the distal region 3′ of the κ-casein gene (about 50–100 kb). In humans these are proline-rich protein (*PB/Prol3*), proline-rich-like protein (*PBI/Prol5*), and basic proline-rich protein (*BPLP/Prol1*). In mouse and rat they are designated *msg1/smr1* (and *msg3/smr3*) and *msg2/smr2*. This whole chromosomal region lacks any detectable similarity in comparison to nonmammalian vertebrate sequences. This could be due to the high rate of divergence and/or because these genes have evolved functions that are required only in mammals. The genes in this region have roles in host defense, nutrition, and calcium homeostasis. These findings pose interesting questions regarding the evolution of this gene region and the role of the genes in it in secretory epithelial tissues.

It has been suggested that the genome can be divided into domains in accordance with the regulated expression of groups of adjacent genes. There are many examples of evolutionarily and functionally related gene clusters in mammals [33–38]. In *Saccharomyces cerevisiae* pairs or triplets of adjacent genes display similar expression patterns that also correlate to function and often share distal regulatory elements [39]. The globin gene clusters and HOX genes are examples of evolutionarily and functionally related gene clusters with shared control elements in higher eukaryotes. In human the clustering of highly expressed genes has been observed, which was shown to be the consequence to a large extent of the clustering of “housekeeping” genes [40,41]. However, clustering of genes with specific functions (tissue-specific) was not ruled out. Expression profiling in *Drosophila* identified groups of adjacent and coregulated genes, but no obvious functional relation was found or correlation to chromosomal banding pattern or structure [42]. Spellman and Rubin [42] suggest that the similarity in regulation is conferred at the level of chromatin structure based on two observations: first, that the gene regions are large, 200 kb, encompassing on average 15 genes; second, that one or two genes within a group frequently show differential expression. Here we describe the clustering of groups of genes that share evolutionary and/or functional relations as well as spatial expression patterns. *CSN* genes are highly expressed in lactating mammary gland epithelium, while *STATH* and *HTN1*and *HTN3* are highly expressed in salivary gland. However, low levels of each were detected in the reciprocal tissues. This could be explained by a model in which the general chromatin structure for this genomic domain, containing mammalian-specific genes, is accessible in epithelial cells while the local expression levels are regulated by (e.g.) hormonal cues in specific tissues.

The casein gene cluster region is characterized by a low GC content (34%) and the low SINE and higher LINE content that has been correlated with GC-poor genome regions. This corresponds to the observation that HSA4 is, in general, GC poor and that 4q13 has been reported to be among the GC-poorest regions in the genome (<37% GC) [43]. However, with a gene density of 1/36 kb the gene content in this region does not correlate with what is usually observed in GC-poor regions (1/50–150 kb). Although it is not as high as for GC-rich regions (1/5–15 kb) [29], the casein gene cluster is almost 3.5 times more gene dense than the HSA4 average (1/36 vs 1/125) [26]. Compared to other regions with low GC content, HSA4q22 [44], HSA13 [45], and X [17,46], only the 13q22 region has a GC and repeat distribution similar to that of the *CSN* region. This region is composed of a cluster of four genes in approximately 600 kb (1/150) flanked by a gene desert of 500 kb. The genes in this region are very large with vast introns, which has been described as characteristic of AT-rich regions [29]. The casein genes seem to represent an intermediate type, consisting of many small exons (5–19 exons, 21–500 bp) with intron sizes averaging around 1 kb and the sizes of transcriptional units (genes) varying between 7 and 20 kb. The other genes identified in this region certainly mimic many of these features. Genes in GC-poor regions are thought to predominantly represent developmentally and/or tissue-specifically expressed genes [29]. The gene region studied here seems to contain genes that are expressed mainly in secretory tissues and may be regulated in a developmental fashion.

We find conservation of regions immediately flanking identified genes, which often overlap with the 5′ and 3′ terminal exons. Such conservation could be a result of both a functional constraint on the immediate flanking regions, which is obvious for the 5′ flanking sequences as these orchestrate the transcription initiation, and their close proximity to the actual gene-encoding sequences, such that their conservation results from evolutionary constraints on the protein-coding sequences. Another observation made from these analyses is the disparity in location in the different species of certain conserved putative regulatory regions relative to the start site of transcription. For example, the BCE was first identified in the 5′ flanking region of the bovine *CSN2* gene, 1.6 kb from the transcription start [12]. The homologous regions in human, mouse, and rat are 2–5 kb farther upstream. Transient transfection experiments have suggested a functional importance for the human BCE (−4.5 kb) [11] and preliminary data suggest that the mouse BCE (6.5 kb upstream) also exhibits changes in its chromatin organization following exposure to lactogenic hormones (Michelle Kallesen, personal communication). These differences in spacing between upstream enhancers and more proximal regulatory elements may account for differences in the developmental timing of β-casein gene expression in the different mammals, although they do not appear to be critical for maintaining tissue specificity.

These studies demonstrate the power of large-scale alignment approaches for identifying potential regulatory elements. They also show how traditional linear alignments of sequences just a few kilobases upstream of a gene will limit the identification of conserved elements whose position relative to a gene is affected by various insertion/deletion events (e.g., insertion of repeats). Similarly we have detected conserved regions about 10 and 15 kb upstream of the human *CSN1S2B* gene that are only 2.5 and 5.5 kb upstream of the bovine *CSN1S2* gene and within 2.5 kb upstream of mouse *Csn1s2b*. In this case, the difference in localization could be one of the reasons for the lack of expression of the human *CSN1S2B* gene. However, this illustrates again that alignments of limited regions flanking a gene may miss conserved elements potentially important for understanding gene regulation.

Furthermore, there is no reason to assume that all regulatory elements will manifest themselves in sequence comparisons as ungapped highly conserved regions of a certain length (thresholds used in [3,47]). Regulatory elements usually contain clusters of *trans*-acting factor binding sites that individually do not contain more than 10 bp and that have varying affinities for their respective factors. Various protein–DNA and protein–protein interactions may take place and subsequent modifications to the proteins and DNA conformation result from these interactions. While spacing and clustering of these sites may be important, a certain amount of variability might not critically affect function. Moreover these binding sites are not necessarily highly conserved compared to their defined consensus binding sites, and the reduced affinity of the binding site might actually be an essential part of its function within the context of the regulatory element [48]. Insertions and/or deletions between the binding sites, although not abrogating function, may obscure the identification of the regulatory elements when using arbitrarily set criteria like >70% over 100 bp (as shown in the case of the BCE). Thus, different approaches are needed to identify such potential regulatory elements. A better approximation can be made from a gapped homology analysis, in combination with the cross-comparison of several relatively closely related species (H/M, H/C, M/C) as shown here and by Paulsen et al. [16]. This multispecies approach may be combined with a scoring of the alignments based on length and identity as described by Flint et al. [49]. The identification of regulatory elements could be aided by combining this type of homology search with a search for clustered binding sites of factors known to be involved in the regulation of the region under investigation; rVista [50] and TraFaC [51] are based on similar approaches. However, once again low-affinity binding sites might be hard to identify in such an approach, even combined with identification of conserved regions. A more objective pattern search in combination with analysis of conserved regions could enable the identification of new binding sites/regulatory elements. One should also keep in mind that some factors involved in chromatin remodeling

have fairly nonspecific binding sites, and we do not have a clear definition for such elements as matrix-attachment sites, boundary elements, and origins of replication that may contribute to the regulation of genes and genomic domains. Nevertheless, this is an era of exciting possibilities to elucidate the complexity of the interface of sequence composition, gene regulation, and evolution by the comparative analysis of DNA sequences from different species.

Materials and methods

BAC isolation and characterization

To isolate BAC clones containing the casein gene cluster or parts of it the following BAC libraries were screened: RPCI-11 human male BAC library, RPCI-23 mouse female C57BL/6J BAC library, and RPCI-42 bovine male BAC library. The following probes were used to isolate and characterize BACs: for human, human β-casein exon 7 probe [6], human κ-casein exon 4 [6], bovine α-s2-casein cDNA [52], human α-casein cDNA [6], histatin, and statherin [53]; for mouse, mouse α-casein cDNA, β, δ, γ, and κ [5]; for bovine, α-s1, β-casein cDNA, α-s2, κ [52].

Libraries were screened by colony hybridization and isolated BAC clones were mapped using pulsed-field gel electrophoresis and Southern blot analysis and compared to previously established maps for the casein gene cluster [5–7].

BACs with minimal overlap were selected and sequenced by the Baylor College of Medicine Human Genome Center. Two mouse BACs were sequenced to completion (RP23-440C1, AC074046 and RP23-423C8, AC022298), two other BACs covering a larger region are still in draft sequence (RP23-457P20, AC087037 and RP23-354F6, AC69075). Human BAC RP11-529K3 was sequenced to completion (AC063956), human BACs RP11-16908 (AC074273) and RP11-922D2 (AC060228) are still in progress (3.6× and 3.2× coverage). Bovine BACs RP42-73E7 (AC134934) and RP42-254I13 (AC134173) are in draft sequence (~4× coverage).

Computational analysis

For comparative sequence analysis the following sequences were used: a 400-kb region encompassing the human casein gene cluster assembled using Sequencher (version 4.1, Gene Codes Corp.) from the finished sequence of five overlapping BACs (AC106884, AC105347, AC104811, AC063956, and AC108941, a total of 692 kb, also present in NT_030640) (this sequence was compared to previously established maps and sequences); 304 kb of mouse sequence assembled from two finished BACs (AC074046 and AC022298) containing the casein genes and extended with one contig of BAC RP23-354F6 (AC069075); contigs of draft sequence (AC134173 and AC134934, ~4× coverage) and preexisting data covering the bovine locus (CSN1_CSN2 region M33123, X06359, and AY15495; bovine α-s2, M94327); and sequence subtracted from rat megacontig NM_042906 (1345–1600 kb) that includes contigs of BAC clones CH230-222E16 (AC105536, 5× coverage) and CH230-73M13 (AC097835, 5× coverage). Repeat elements and G+C content were characterized using RepeatMasker, version 07/07/2001 (A.F.A. Smit and P. Green, RepeatMasker at http://ftp.genome.washington.edu/RM/RepeatMasker.html). The sequences were analyzed using BLAST to the nonredundant, EST, TIGR, and HTGS databases (http://www.ncbi.nlm.nih.gov/BLAST/, http://www.tigr.org/tdb/tgi/) and other computational genome analysis tools to identify the presence and positions of known genes and ESTs and the potential functions of their gene products (http://psort.nibb.ac.jp, http://www.blocks.fhcrc.org). The majority of gene prediction tools failed in most cases to predict even the known casein genes; those predictions are not discussed here. FGENES (http://genomic.sanger.ac.uk/gf/gfb.html) predicted the positions of some of the genes and exons. *sim4* (http://biom3.univ-lyon1.fr/sim4.html (23)) was used to align the cDNAs and ESTs to the genomic sequences and generate exon files for PipMaker and MultiPipMaker.

PipMaker (21) and MultiPipMaker (both available at http://bio.cse.psu.edu), World Wide Web servers that align two and three or more genomic sequences, respectively, providing both graphical and nucleotide-level views of the results, were used for comparative analysis. ClustalW [54] at the multiple-alignment Web server of the UK HGMP Resource Center (http://www.hgmp.mrc.ac.uk/Registered/Webapp/magi) was used to generate alignments; alignments were edited according to the conservation of exon boundaries, and further phylogenetic and molecular evolutionary analysis were performed using MEGA version 2.2 [55].

Expression analysis

Total RNA was isolated from mouse and bovine lactating mammary gland and salivary gland tissue using Trizol (Invitrogen, Inc.) according to the manufacturer's instructions. RNA from lactating breast tissue was recovered from human milk fat, based on the findings that significant amounts of cytoplasm from lactating cells are included with some milk fat globules at secretion [56]. RNA was isolated using Trizol as follows: 10 to 20 ml of fresh milk (no more than a few hours stored on ice) was centrifuged at 4000 rpm for 30 min at 4°C. The fat plug was recovered from the top of the tube and weighed. Trizol reagent was added as recommended in the manufacturer's protocol and the mixture was shaken vigorously until completely mixed. The manufacturer's protocol was further followed for isolation of RNA. Human salivary gland RNA was a kind gift from Dr. D. Dickinson (Medical College of Georgia, Augusta, GA), mouse and human testis RNA were kindly provided by Dr. A. Cooney (Baylor College of Medicine, Houston, TX).

Expression analysis of genes located in the casein gene cluster was performed using RT-PCR. Primers were designed based on the alignments of cDNAs or ESTs to the genomic sequences and based on homologies in the alignments of the sequences of the three species. To verify the expression of genes for which no cDNA or EST was available or to confirm expression in unusual tissues, PCR products were cloned into the pGEM-T Easy vector (Promega) and sequenced.

DNase I analysis

DNase I-hypersensitive site analysis on lactating mammary gland tissue of mouse and cow was performed as described in Li and Rosen [57]. Genomic DNA was isolated from the DNase I-treated nuclei and digested with *Nhe*I and *Bam*HI (mouse) or *Nhe*I (cow). Southern blot analysis was performed with radionucleotide-labeled probes.

Acknowledgments

We thank Steve Scherer (Baylor College of Medicine, Human Genome Center) for screening of the human BAC library; Wei-Wen Cai (Baylor College of Medicine, Molecular and Human Genetics) for the opportunity to screen the mouse BAC library; Mike Metzker, Kim Worley, and John Bouck (Baylor College of Medicine, Human Genome Center) for assistance during sequencing of the isolated BACs; David Needleman (UT–Houston, Health Science Center) for sequencing in early stages of this project; Suzan Schanler for technical assistance; and David Hewett-Emmett (UT–Houston) for his phylogenetic expertise. Human salivary gland RNA and HTN/STATH cDNAs were kindly provided by Douglas Dickinson (Medical College of Georgia, School of Dentistry); Anthony Capuco (USDA) provided bovine salivary gland and mammary gland tissue. Austin Cooney (Baylor College of Medicine, Molecular and Cellular Biology) provided human and mouse testis RNA. We thank Beverly and Delilah Walsh, Jacqueline and Rosalyn Elliott for human milk samples. This work was supported by USDA-CNRC-CRIS 6250-51000-039-00D.

References

[1] K. Sumiyama, et al., Genomic structure and functional control of the Dlx3-7 bigene cluster, Proc. Natl. Acad. Sci. USA 99 (2002) 780–785.
[2] U. DeSilva, et al., Generation and comparative analysis of approximately 3.3 Mb of mouse genomic sequence orthologous to the region of human chromosome 7q11.23 implicated in Williams syndrome, Genome Res. 12 (2002) 3–15.
[3] G.G. Loots, et al., Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons, Science 288 (2000) 136–140.
[4] N.P. Shah, Effects of milk-derived bioactives: An overview, Br. J. Nutr. 84 (Suppl 1) (2000) S3–10.
[5] M. Rijnkels, D. Wheeler, H.A. de Boer, F.R. Pieper, Structure and expression of the mouse casein gene locus, Mamm. Genome 8 (1997) 9–15.
[6] M. Rijnkels, E. Meershoek, H.A. de Boer, F.R. Pieper, Physical map and localization of the human casein gene locus, Mamm. Genome 8 (1997) 285–286.
[7] M. Rijnkels, P.M. Kooiman, H.A. de Boer, F.R. Pieper, Organization of the bovine casein gene locus, Mamm. Genome 8 (1997) 148–152.
[8] J.-C. Mercier, J.-L. Vilotte, Structure and function of milk protein genes, J. Dairy Sci. 76 (1993) 3079–3098.
[9] J.M. Rosen, C. Zahnow, A. Kazansky, B. Raught, Composite response elements mediate hormonal and developmental regulation of milk protein gene expression, Biochem. Soc. Symp. 63 (1998) 101–113.
[10] J.M. Rosen, S.L. Wyszomierski, D. Hadsell, Regulation of milk protein gene expression, Annu. Rev. Nutr. 19 (1999) 407–436.
[11] P. Winklehner-Jennewein, et al., A distal enhancer region in the human β-casein gene mediates the response to prolactin and glucocorticoid hormones, Gene 217 (1998) 127–139.
[12] C. Schmidhauser, et al., A novel transcriptional enhancer is involved in the prolactin- and extracellular matrix-dependent regulation of β-casein gene expression, Mol. Biol. Cell. 3 (1992) 699–709.
[13] S. Pierre, et al., A distal region enhances the prolactin induced promoter activity of the rabbit α s1-casein gene, Mol. Cell. Endocrinol. 87 (1992) 147–156.
[14] A. Gerencser, et al., Comparative analysis on the structural features of the 5′ flanking region of κ-casein genes from six different species, Genet. Sel. Evol. 34 (2002) 117–128.
[15] M.A. Groenen, R.J. Dijkhof, J.J. van der Poel, R. van Diggelen, E. Verstege, Multiple octamer binding sites in the promoter region of the bovine α s2-casein gene, Nucleic Acids Res. 20 (1992) 4311–4318.
[16] M. Paulsen, et al., Comparative sequence analysis of the imprinted Dlk1–Gtl2 locus in three mammalian species reveals highly conserved genomic elements and refines comparison with the Igf2–H19 region, Genome Res. 11 (2001) 2085–2094.
[17] C. Chureau, et al., Comparative sequence analysis of the X-inactivation center region in mouse, human, and bovine, Genome Res. 12 (2002) 894–908.
[18] W. Miller, So many genomes, so little time, Nat. Biotechnol. 18 (2000) 148–149.
[19] W.J. Murphy, et al., Resolution of the early placental mammal radiation using Bayesian phylogenetics, Science 294 (2001) 2348–2351.
[20] M. Nei, P. Xu, G. Glazko, Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms, Proc. Natl. Acad. Sci. USA 98 (2001) 2497–2502.
[21] S. Schwartz, et al., PipMaker—A web server for aligning two genomic DNA sequences, Genome Res. 10 (2000) 577–586.
[22] A.J. Marshall, et al., FDC-SP, a novel secreted protein expressed by follicular dendritic cells, J. Immunol. 169 (2002) 2381–2389.
[23] L. Florea, et al., Web-based visualization tools for bacterial genome alignments, Nucleic Acids Res. 28 (2000) 3486–3496.
[24] C. Karrman, B. Backman, M. Dixon, G. Holmgren, K. Forsman, Mapping of the locus for autosomal dominant amelogenesis imperfecta (*AIH2*) to a 4-Mb YAC contig on chromosome 4q11–q21, Genomics 39 (1997) 164–170.
[25] L.M. Sabatini, T. Ota, E.A. Azen, Nucleotide sequence analysis of the human salivary protein genes HIS1 and HIS2, and evolution of the STATH/HIS gene family, Mol. Biol. Evol. 10 (1993) 497–511.
[26] E.S. Lander, et al., Initial sequencing and analysis of the human genome, Nature 409 (2001) 860–921.
[27] A. Pavlicek, et al., Similar integration but different stability of Alus and LINEs in the human genome, Gene 276 (2001) 39–45.
[28] A. Pavlicek, J. Paces, O. Clay, G. Bernardi, A compact view of isochores in the draft human genome sequence, FEBS Lett. 511 (2002) 165–169.

[29] G. Bernardi, Isochores and the evolutionary genomics of vertebrates, Gene 241 (2000) 3–17.
[30] R.H. Waterston, et al., Initial sequencing and comparative analysis of the mouse genome, Nature 420 (2002) 520–562.
[31] R. Chen, J.B. Bouck, G.M. Weinstock, R.A. Gibbs, Comparing vertebrate whole-genome shotgun reads to the human genome, Genome Res. 11 (2001) 1807–1816.
[32] M.S. Lamkin, P. Lindhe, Purification of kinase activity from primate parotid glands, J. Dent. Res. 80 (2001) 1890–1894.
[33] F. Grosveld, et al., The regulation of expression of human β-globin genes, Prog. Clin. Biol. Res. 251 (1987) 133–144.
[34] B.K. Jones, B.R. Monks, S.A. Liebhaber, N.E. Cooke, The human growth hormone gene is regulated by a multicomponent locus control region, Mol. Cell. Biol. 15 (1995) 7010–7021.
[35] L. Bélanger, S. Roy, A. Allard, New albumin gene 3′ adjacent to the α1-fetoprotein locus, J. Biol. Chem. 269 (1994) 5481–5484.
[36] N.S. Shachter, Y. Zhu, A. Walsh, J.L. Breslow, J.D. Smith, Localization of a liver-specific enhancer in the apolipoprotein E/C-I/C-II gene locus, J. Lipid Res. 34 (1993) 1699–1707.
[37] A. Walsh, et al., Intestinal expression of the human apoA-I gene in transgenic mice is controlled by a DNA region 3′ to the gene in the promoter of the adjacent convergently transcribed apoC-III gene, J. Lipid Res. 34 (1993) 617–623.
[38] V. Mahdavi, A.P. Chambers, G.B. Nadal, Cardiac α- and β-myosin heavy chain genes are organized in tandem, Proc. Natl. Acad. Sci. USA 81 (1984) 2626–2630.
[39] B.A. Cohen, R.D. Mitra, J.D. Hughes, G.M. Church, A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression, Nat. Genet. 26 (2000) 183–186.
[40] M.J. Lercher, A.O. Urrutia, L.D. Hurst, Clustering of housekeeping genes provides a unified model of gene order in the human genome, Nat. Genet. 31 (2002) 180–183.
[41] H. Caron, et al., The human transcriptome map: Clustering of highly expressed genes in chromosomal domains, Science 291 (2001) 1289–1292.
[42] P.T. Spellman, G.M. Rubin, Evidence for large domains of similarly expressed genes in the *Drosophila* genome, J. Biol. 1 (2002) 5.
[43] G. Bernardi, Misunderstandings about isochores, Gene 276 (Part 1) (2001) 3–13.
[44] J.W. Touchman, et al., Human and mouse α-synuclein genes: Comparative genomic sequence analysis and identification of a novel gene regulatory element, Genome Res. 11 (2001) 78–86.
[45] L.J. Kurihara, et al., Candidate genes required for embryonic development: A comparative analysis of distal mouse chromosome 14 and human chromosome 13q22, Genomics 79 (2002) 154–161.
[46] A.M. Mallon, et al., Comparative genome sequence analysis of the Bpa/Str region in mouse and man, Genome Res. 10 (2000) 758–775.
[47] I. Dubchak, et al., Active conservation of noncoding sequences revealed by three-way species comparisons, Genome Res. 10 (2000) 1304–1306.
[48] J. Jiang, M. Levine, Binding affinities and cooperative interactions with bHLH activators delimit threshold responses to the dorsal gradient morphogen, Cell 72 (1993) 741–752.
[49] J. Flint, et al., Comparative genome analysis delimits a chromosomal domain and identifies key regulatory elements in the α globin cluster, Hum. Mol. Genet. 10 (2001) 371–382.
[50] G.G. Loots, I. Ovcharenko, L. Pachter, I. Dubchak, E.M. Rubin, rVista for comparative sequence-based discovery of functional transcription factor binding sites, Genome Res. 12 (2002) 832–839.
[51] A.G. Jegga, et al., Detection and visualization of compositionally similar cis-regulatory element clusters in orthologous and coordinately controlled genes, Genome Res. 12 (2002) 1408–1417.
[52] M. Rijnkels, P.M. Kooiman, P.J.A. Krimpenfort, H.A. de Boer, F.R. Pieper, Expression analysis of the individual bovine β-, α s2- and κ-casein genes in transgenic mice, Biochem. J. 311 (1995) 929–937.
[53] D.P. Dickinson, A.L. Ridall, M.J. Levine, Human submandibular gland statherin and basic histidine-rich peptide are encoded by highly abundant mRNA's derived from a common ancestral sequence, Biochem. Biophys. Res. Commun. 149 (1987) 784–790.
[54] D.G. Higgins, J.D. Thompson, T.J. Gibson, Using CLUSTAL for multiple sequence alignments, Methods Enzymol. 266 (1996) 383–402.
[55] S. Kumar, K. Tamura, I.B. Jakobsen, M. Nei, MEGA2: Molecular evolutionary genetics analysis software, Bioinformatics 17 (2001) 1244–1245.
[56] G.E. Huston, S. Patton, Factors related to the formation of cytoplasmic crescents on milk fat globules, J. Dairy Sci. 73 (1990) 2061–2066.
[57] S. Li, J.M. Rosen, Glucocorticoid regulation of rat whey acidic protein gene expression involves hormone-induced alterations of chromatin structure in the distal promoter region, Mol. Endocrinol. 8 (1994) 1328–1335.