Available online at www.sciencedirect.com

SCIENCE DIRECT

ELSEVIER

Genomics 87 (2006) 382 – 391

GENOMICS

www.elsevier.com/locate/ygeno

Comparative genomics of the sperm mitochondria-associated cysteine-rich protein gene

Sabrina K. Hawthorne, Golnaz Goodarzi, Jana Bagarova, Katherine E. Gallant, Rakhee R. Busanelli, Wendy J. Olend, Kenneth C. Kleene *

Department of Biology, University of Massachusetts at Boston, 100 Morrissey Boulevard, Boston, MA 02125, USA

Received 1 July 2005; accepted 20 September 2005
Available online 1 December 2005

Abstract

The sperm mitochondrial cysteine-rich protein (SMCP) is a rapidly evolving cysteine- and proline-rich protein that is localized in the mitochondrial capsule and enhances sperm motility. The sequences of the SMCP protein, gene, and mRNA in a variety of mammals have been compared to understand their evolution and regulation. SMCP can now be reliably identified by its tripartite structure including a short amino-terminal segment; a central segment containing short tandem repeats rich in cysteine, proline, glutamine, and lysine; and a C-terminal segment containing no repeats, few cysteines, and a C-terminal lysine. The *SMCP* gene is located in the epidermal differentiation complex (EDC), a large gene cluster that functions in forming epithelial barriers. Similarities in chromosomal location, molecular function, intron–exon structure, and protein organization argue that *SMCP* originated from an EDC gene and acquired spermatogenic cell-specific transcriptional and translational regulation and a novel cellular function in sperm motility. The *SMCP* 5′ UTR and 3′ UTR contain conserved elements and uORFs that may function in cytoplasmic regulation of gene expression, and the levels of *SMCP* mRNA in human are much lower than in other mammals, a feature of male-biased expression. The evolution of SMCP has been accompanied by changes in the sequence, number, and length of repeat units, including three alleles in dogs. The major proteins associated with the mitochondrial capsule, SMCP and phospholipid hydroperoxide glutathione peroxidase, provide outstanding examples of changes in cellular function driven by selective pressures on sperm motility, an important determinant of male reproductive success.

Keywords: Sperm mitochondria-associated cysteine-rich protein; Sperm mitochondrial capsule; Translational regulation; Epidermal differentiation complex; Tandem repeats; uORF; Unstructured protein; Male-biased expression

The sperm flagellum generates motility, a factor that is under strong evolutionary pressures in the competition to fertilize the egg among the millions of sperm in the ejaculates of individual and multiple males [1]. The flagella of mammalian sperm are modified by three sperm-specific accessory structures, the mitochondrial sheath, the outer dense fibers, and the fibrous sheath, which provide mechanical factors that enhance motility and a scaffold for localization of proteins that produce ATP and regulate motility [2].

The outer membranes of sperm mitochondria are modified by a structure, the mitochondrial capsule, that is crosslinked by disulfide bridges and resistant to SDS [3]. The mitochondrial capsule confers a stiffness on sperm mitochondria that is very different from that of somatic mitochondria: sperm mitochondria do not swell in hypotonic media and retain the characteristic crescent shape of mitochondria in the mitochondrial sheath after being pelleted in the ultracentrifuge [3]. Biochemical and immunocytochemical studies demonstrate that the mitochondrial capsule is associated with several proteins, including two ~20-kDa proteins that were confused for many years, the sperm mitochondria-associated cysteine-rich protein (SMCP), a small, hydrophilic, cysteine- and proline-rich protein, and a sperm mitochondria-specific isoform of the selenoprotein, phospholipid hydroperoxide glutathione peroxidase, PHGPx [3–7]. SMCP replaces the misleading name “mitochondrial capsule selenoprotein.” A careful examination of the conflicting studies of the composition of the mitochondrial capsule suggests that the primary constituent of the capsule is PHGPx [4,6], while SMCP is present in a

* Corresponding author. Fax: +1 617 287 6650.
E-mail address: kenneth.kleene@umb.edu (K.C. Kleene).

0888-7543/$ - see front matter © 2005 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2005.09.010

"granular layer" on the surface of the capsule [3], but rigorous evidence is lacking for this hypothesis.

Targeted deletion of the mouse *Smcp* gene has no discernable effect on the ultrastructure of the mitochondrial sheath, but sperm motility on the inbred background is impaired, resulting in sperm that fail to migrate in the female reproductive tract and penetrate the egg membranes during fertilization [8]. Expression of the *Smcp* mRNA at high levels is restricted to haploid spermatids, and translation is developmentally regulated: the *Smcp* mRNA is stored in translationally repressed free mRNPs in early haploid cells, and it is actively translated in late haploid cells [5,9]. SMCP is a rapidly evolving protein that exhibits little similarity in divergent species [5,10,11], and the features that distinguish it from other cysteine- and proline-rich proteins have never been identified.

The accompanying comparative analysis of the sequences of the *SMCP* gene and protein in various mammals identifies the distinctive features of SMCP and conserved elements of the mRNA that may function in translational regulation. We also establish unexpectedly that the *SMCP* gene is a paralogue of the genes in the epidermal differentiation complex, a large gene cluster that functions in forming epithelial barriers in cornified epithelia, and that alternations in the number of copies of a short repeated unit have played an important part in the evolution of the protein.

Results

SMCP sequences in various mammals

To identify candidate SMCPs, the EST databases were searched with TBLASTN with SMCPs from mouse, human, and bull. The putative *SMCP* ESTs were evaluated by several criteria: (1) the EST was derived from a library containing testis cDNAs, (2) predicted proteins that exhibit the distinctive features identified here, (3) conserved sequences are present in the 5′ UTR and 3′ UTR as identified below. Selected cDNAs were purchased and sequenced completely on both strands. The bull, hamster, dog, and deer mouse *SMCP* mRNAs were RT-PCR amplified using universal primers for conserved sequences in the 5′ UTR and 3′ UTR, and the sequences of the dog and human 5′ UTRs were extended by 5′ RACE.

Structure and chromosomal location of SMCP genes

BLASTN searches of the mouse, rat, and human genomes reveal that the *SMCP* gene is located near the middle of the epidermal differentiation complex (EDC), a cluster of ~50 genes encoding proteins that form epithelial barriers in syntenic locations at chromosome 3F1 in mouse, chromosome 1q21 in human, and chromosome 2q34 in rat [13,14], confirming in situ hybridization studies [10,11]. In each genome, the *SMCP* gene is located between the involucrin gene and a homolog of the human *XP5/LED10* gene [13]. The Discussion assembles several lines of evidence arguing that the *SMCP* gene evolved from an *EDC* gene.

Comparison of the sequences of the *SMCP* cDNAs and genomic DNAs reveal that the *SMCP* gene in mouse, bull, rat, human, and dog contains a single intron, located 20 nt upstream of the *SMCP* ATG translation initiation codon. The intron is 3829 nt in dog, 4134 nt in bull, 4269 nt in mouse, 4147 nt in rat, and 5956 nt in human, and the splice junctions conform to the AG:GT rule [15]. The structure of the mouse, rat, and human *SMCP* genes deduced here agrees with the structures deduced from genomic clones [10–12].

SMCP is expressed at very low levels in epithelial tissues

The realization that the *SMCP* gene is an *EDC* paralogue raises questions whether the *SMCP* promoter is active in epithelial tissues, a question that was not answered by previous Northern blot analyses of nonepithelial somatic tissues [10,11]. RT-PCR using 30 amplification cycles detects a strong ethidium bromide-stained band in testis and weaker bands in uterus, tongue, stomach, skin, spleen, heart, and kidney (data not shown). Fifteen amplification cycles and a Southern blot detect *Smcp* mRNA at high levels in testis and at much lower levels in somatic tissues (Fig. 1). This protocol gives a more accurate indication of the differences in *Smcp* mRNA levels in various tissues than does ethidium bromide staining and demonstrates that *Smcp* is expressed at very low levels in epithelial tissues in which a variety of EDC mRNAs are expressed at high levels [13,14]. RT-PCR of 18S rRNA shows that similar amounts of PCR products are detected in all lanes (Fig. 1), and the absence of ethidium bromide-

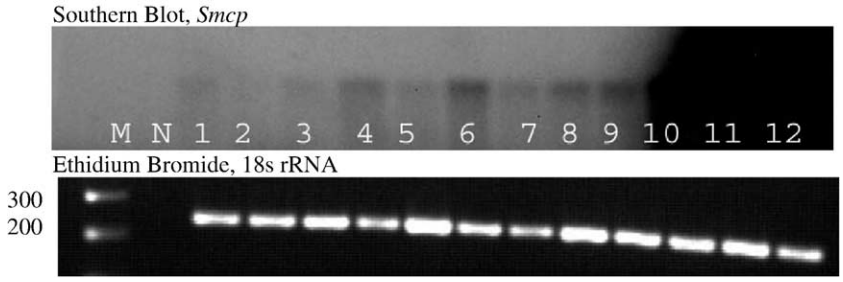

Fig. 1. RT-PCR analysis of the levels of *Smcp* mRNA in various adult mouse tissues. Top: *Smcp* mRNA was RT-PCR amplified for 15 cycles, a point at which ethidium bromide staining detected a PCR product only in testis, Southern blotted, and probed with an *Smcp* probe. Bottom: 18S rRNA was RT-PCR amplified for 15 cycles and the product was detected by ethidium bromide staining. M, 100 bp ladder size marker; N, negative control (no reverse transcriptase); lane 1, uterus; lane 2, stomach; lane 3, esophagus; lane 4, tongue; lane 5, lung; lane 6, spleen; lane 7, skin; lane 8, heart; lane 9, kidney; lane 10, liver; lane 11, brain; lane 12, testis. The sequences of the primers used in the RT-PCR are shown in Table 1.

stained PCR products after 30 cycles of amplification without reverse transcriptase (not shown) indicates that the PCR-products are derived from mRNA. Clearly, the specificity of the *Smcp* promoter differs sharply from those of EDC promoters.

Transcription start site, 5′ UTR, and 3′ UTR

Primer extension analysis was used to compare the transcription start sites of the *SMCP* mRNA in mouse, human, and deer mouse testis RNA, to obtain information that may be critical for identifying elements that regulate transcription and translation in haploid spermatogenic cells. As shown in Fig. 2, the primer extension of deer mouse and mouse testis RNAs produced strong bands at ~180 and 190 nt, respectively, although there is a much weaker band in deer mouse at ~220 bp. In reality, the major start sites in these species are at virtually the same position, because the primers anneal to slightly different positions in each mRNA. The primer extension with human testis RNA produced a faint band at about 140 nt, which corresponds to a 5′ UTR of ~147 nt, significantly shorter than reported previously, ~169 nt [11]. 5′ RACE confirms the position of the major human transcription start site (Fig. 4), but shows that dog uses several transcription start sites, in contrast to the strong preference for single start sites in mouse, rat, deer mouse, and human [10–12].

Unexpectedly, the levels of *SMCP* primer extension products in human are much lower than those of mouse and deer mouse. This does not appear to be due to the quality of the primer because low levels of primer extension products were observed with several primers (data not shown). In addition, Northern blots in which constant amounts of dog, human, and mouse testis RNA were hybridized to the same number of counts of probes of the same length for dog, human, and mouse *SMCP* mRNAs confirm that the *SMCP* mRNA is much more abundant in mouse and dog testis RNA than it is in human testis RNA (Fig. 3). Rehybridization of these blots to protamine 1 (*PRM1*) probes detects similar levels of *PRM1* mRNA in each species. The *PRM1* mRNA was used as a control for the levels of spermatogenic cells in the testes from each source, because it encodes a basic

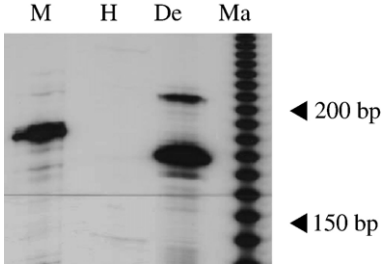

Fig. 2. Primer extension analysis of mouse, deer mouse, and human *SMCP* mRNAs. 105 cpm of each primer (Table 1) was annealed with 10 μg RNA from each species and extended with reverse transcriptase, and the size of the extension products was determined by denaturing PAGE. Lane M, mouse; lane H, human; lane De, deer mouse; Ma, 10 bp DNA ladder.

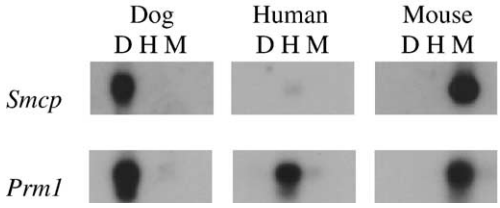

Fig. 3. Northern blot analysis of *SMCP* levels in mouse, deer mouse, and human testis RNA. 2 μg of each total testis RNA was electrophoresed on a 2% formaldehyde agarose gel, transferred to nitrocellulose, and hybridized sequentially to equal cpm of probes for *SMCP* and *PRM1* from each species. M, mouse; H, human; D, deer mouse. The universal primers were used to generate the *SMCP* probe and the primers used to generate the *PRM1* probes are shown in Table 1.

chromosomal structural protein in sperm whose levels are dictated by the size of the genome; hence, its levels are expected to vary little between species. Our results show that *PRM1* mRNA in dog, a species that lacks *PRM2,* is slightly more abundant than in mouse and human, species that express *PRM2.* The *SMCP* and *PRM1* mRNAs are poorly conserved, and each probe hybridizes only to mRNAs of the same species.

The *cis* elements that specify the *SMCP* transcription in haploid spermatogenic cells are unclear. Single transcription start sites are determined by promoters containing a TATA box located ~25 nt upstream of the start site or an initiator bearing the consensus sequence TCA-G/T-T-T/C [16]. However, neither element is present at the expected positions in the mouse, rat, dog, human, or bull genes, and CLUSTAL alignments reveal that the most conserved element in the 5′ flanking regions of these species, GTCAYARAR (Y, pyrimidine; R, purine), is ~36 nt upstream of the mouse start site (not shown) and does not correspond to the binding site of any transcription factor in the TRANSFAC database. The 5′ flanking regions of the *SMCP* genes are also notable for the paucity of CpG dinucleotides.

In contrast, the *SMCP* 5′ UTR contains several conserved features. The *SMCP* mRNA is one of a relatively small proportion of mammalian mRNA species in which the ATG codon of the *SMCP* upstream reading frames (uORF) is not the ATG closest to the 5′ cap [17]. Specifically, the *SMCP* 5′ UTR of seven species in Fig. 4 contains two uORFs, which regulate mRNA translation by reducing the number of ribosomes that translate downstream reading frames without producing free mRNPs [18]. The uORF closest to the 5′ terminus, uORF1, encodes a hexapeptide (MDSLDC) and is the most conserved sequence of the entire *SMCP* gene, but uORF2 is much more variable in sequence and length, 2 to 8 codons. Since the start codons of both uORFs have purines in the −3 position, they are in a strong context for the initiation of translation and are thus expected to reduce the number of ribosomes that translate the *SMCP* reading frame [17,18]. However, the two *Cetartiodactyla* have only one uORF: the sequence corresponding to uORF1 in pig begins with a GUG codon, and bull lacks uORF2. The 4 bases upstream of the *SMCP* ATG codon are also conserved, GAAG.

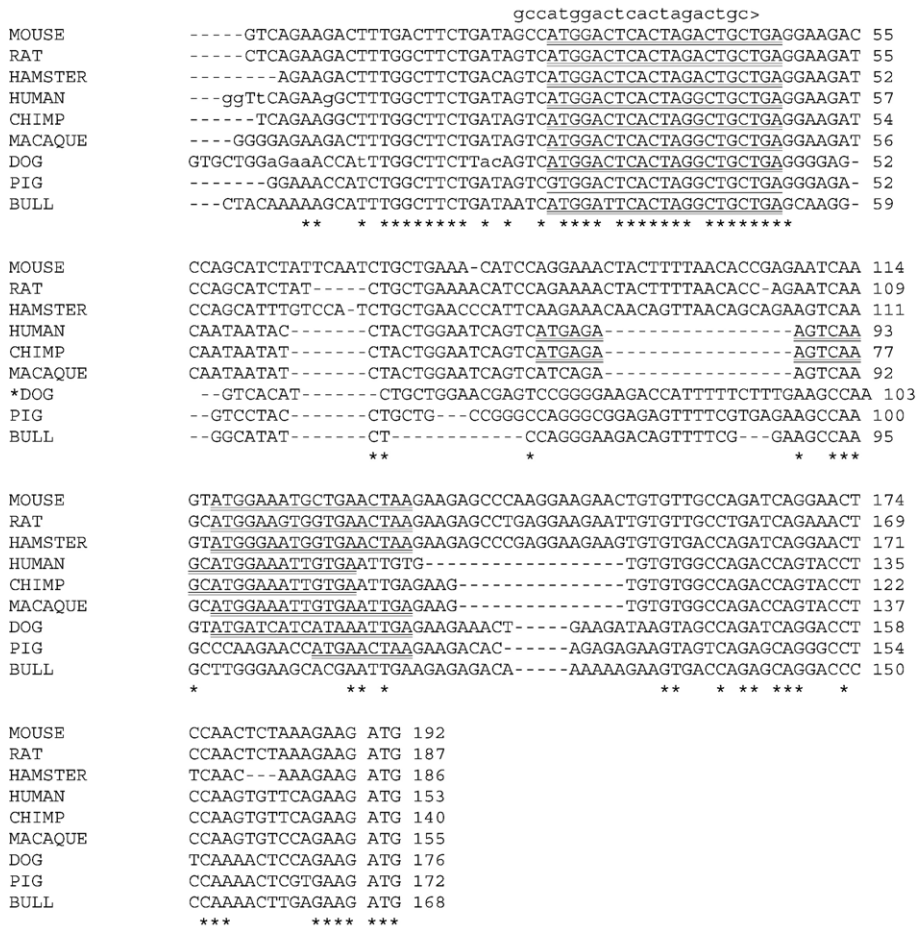

Fig. 4. CLUSTALW alignment of the *SMCP* 5′ UTR in various mammals. The full sequence of the mouse and rat *Smcp* 5′ UTR has been rigorously established by S1 mapping and primer extension ([10,12] and S.K. Hawthorne et al., manuscript in preparation). The 5′ proximal sequences of the dog and human 5′ UTR were determined by 5′ RACE. The 5′ termini of bull, macaque, and chimp were derived from NW 452194, CD767148, and NW 101935, respectively. The sequence of the hamster 5′ UTR was reported by Nam et al. [24]. The 5′ termini of dog and human 5′ RACE clones are indicated by lowercase letters. The uORFs are underlined, and bases that are conserved in all species are marked with asterisks. The sequence of the upstream universal primer is indicated in lowercase letters.

Cis elements close to the 5′ terminus can be important in translational regulation [19,20]. Another highly conserved 8-nucleotide element, located between the transcription start site and uORF1, is a candidate for a sequence in the 5′ UTR that sharply reduces expression of the GFP reporter in early spermatids in transgenic mice (S.K. Hawthorne and K.C. Kleene, manuscript in preparation). In addition, MFOLD predicts a stem loop in ~95% of optimal and suboptimal folds (not shown) near the transcription start site of mouse, rat, and human *SMCP* mRNAs. Stem loops are often binding sites for RNA-binding proteins that regulate cytoplasmic gene expression [19,21]. However, MFOLD does not predict a stem loop in the same positions of the bull and dog 5′ UTRs.

Fig. 5 reveals that the *SMCP* 3′ UTR also contains several conserved sequences, including one that is immediately upstream of the AAUAAA polyadenylation signal, another position that plays a prominent role in translational regulation [22,23]. These sequences are candidates for elements that repress accumulation of the GFP reporter in early spermatids and localize its expression in middle spermatids, but do not sequester the mRNA in free mRNPs (S.K. Hawthorne et al., manuscript in preparation).

The SMCP protein

Alignment of SMCPs in divergent species is confusing because the protein has few conserved amino acids and varies in length from 103 to 146 amino acids. In addition, a previous sequence of hamster SMCP contained three reading frame shifts [24]. The alignments in closely related species shown in Fig. 6 reveal that the SMCP protein is organized into three segments. Segment 1 at the amino terminus is 8 amino acids long and always contains an aspartic acid and lysine in the third and sixth positions, respectively, and a cysteine or serine in the second position. The middle segment is composed of variable numbers of short tandem repeats. Although the sequences of the repeats vary within and between species, all contain at least one proline or glutamine, ~88% contain two adjacent cysteines, and ~81% contain at least one lysine in position 5, 6, or 7, such as SCCPPKC, QCCPPSP, PCCPPK, PCCPQK, QCCPPKHN, and KCCQSKGN. About 50% of the repeat units are 7 amino acids long, and about 25% are 6 or 8 amino acids long. The central region is rich in 4 amino acids, proline, 31.6%; cysteine, 30.9%; lysine, 12.2%; and glutamine, 9.1%; 3 of which (excluding cysteine) are strongly correlated with

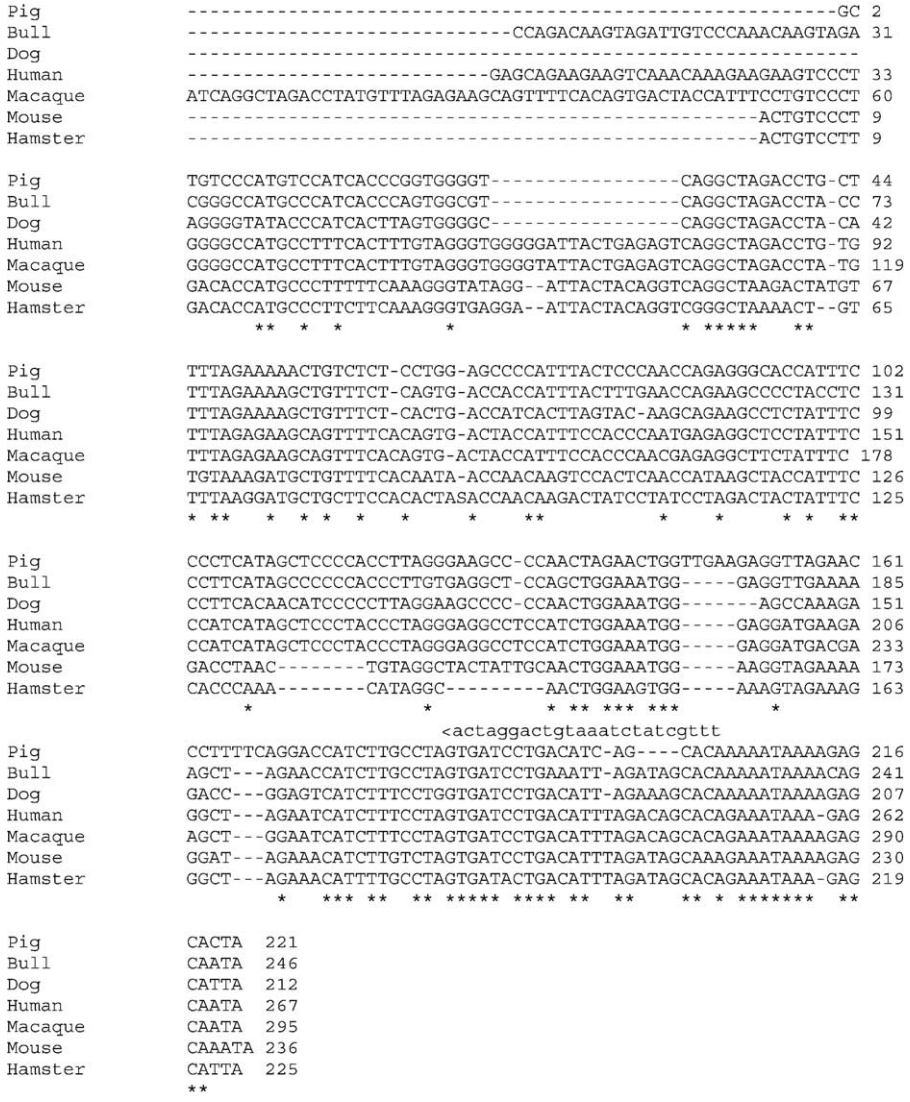

Fig. 5. CLUSTALW alignment of the *SMCP* 3′ UTR in various mammals. The 3′ UTRs have been deduced from cDNAs. The sequences of bull, dog, and macaque 3′ UTRs are derived from genomic sequences, NW 452194, NW 139881, and AB069983, respectively. Nucleotide positions that are conserved in all seven species are marked by asterisks. The sequence of the downstream universal primer is indicated in lowercase letters.

intrinsically unstructured protein segments [25,26]. The junction between the middle segment and the C-terminal segment is marked by two or three consecutive cysteines. The C-terminal segment varies from 42 to 54 amino acids, lacks repeats and adjacent cysteines, and contains zero to two cysteines and a C-terminal lysine, which is conserved despite changes in the position of the stop codon.

Kyte and Doolittle hydrophobicity plots (http://www.expasy.org) show that SMCPs are hydrophilic, that the most hydrophilic region is near the carboxy terminus, and that the only hydrophobic segment in the entire protein is 3 to 4 amino acids long at the junction of the middle and C-terminal segments (data not shown).

Several observations show that the central segment has undergone evolutionary changes in the sequence and expansions/contractions of the tandem repeats. First, the number of repeats varies by more than twofold: 6 in bull and pig, 11 in deer mouse, and at least 14 in house mouse, including 17 amino acids that are a poor match for the repeat unit. Second, variations of an 8-amino-acid repeat unit are repeated five times in bull and six times in all three primates, while the repeat unit is 7 amino acids in pig and dog and 6 or 7 amino acids in rodents. Third, the end of the dog repeat unit contains a sequence, PQKP, which is repeated nine times, while the corresponding sequence in pig is PPKP, and each primate contains two repeats with a motif, KGN, that is absent from the species in other orders. Fourth, RT-PCR amplification of beagle testis RNA produces similar amounts of two bands in agarose gels, which were found to differ by 63 nt (21 amino acids) by cloning and sequence analysis and are best explained by alleles. The boxer *Smcp* genomic sequence in NW_139881 contains a third allele that is 42 nt longer than our shorter allele.

Database searches for SMCP and the absence of SMCP paralogues

Database searches for SMCPs using BLASTP and default conditions detect SMCPs only from closely related species,

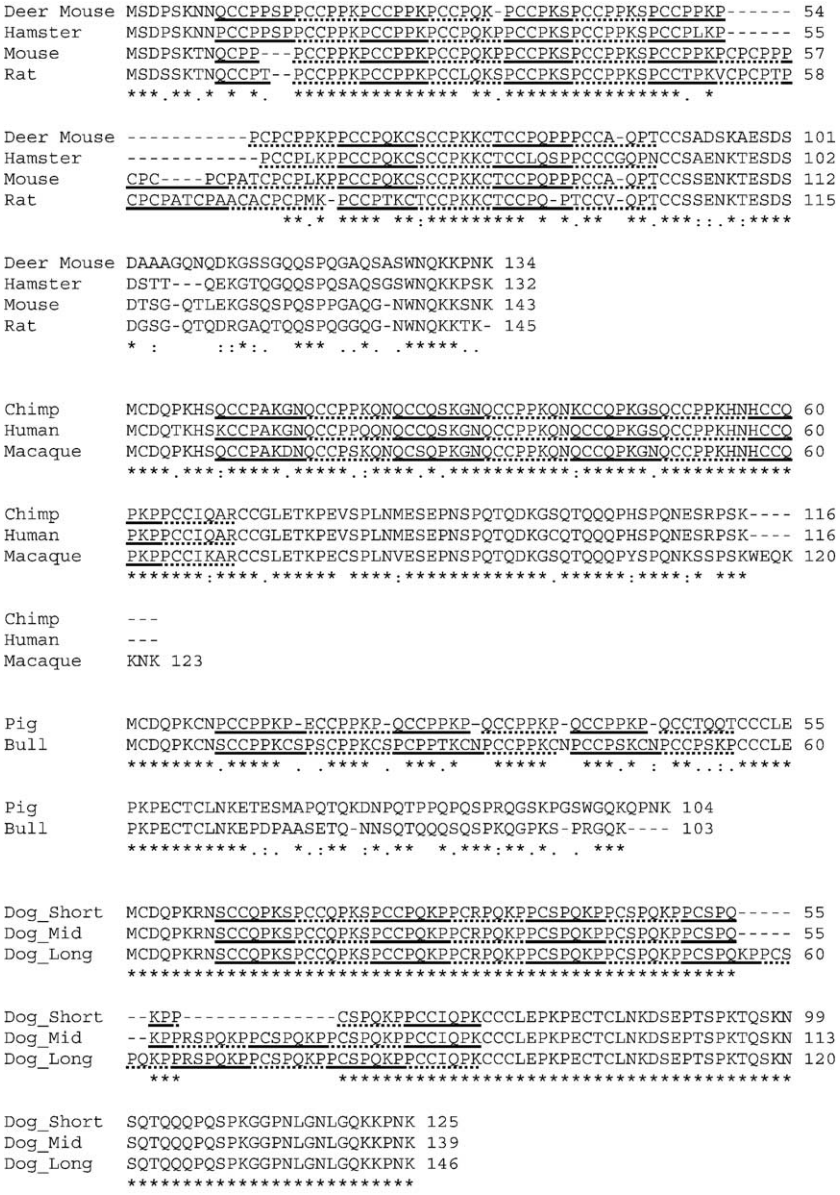

Fig. 6. Multiple sequence alignment of predicted SMCP sequences in various mammals. CLUSTALW multiple alignments were performed on closely related species to eliminate the confusing gaps produced by alignments of distantly related species. The alignments were adjusted manually to illustrate the short repeats in the middle segment, and alternating short repeats are underlined with lines and dashes. The sequences of chimpanzee and macaque SMCP are derived from virtual translation of NW_102379, AB069983, and CD767148, respectively.

because the low-complexity filter eliminates the central, low-complexity segment, a characteristic feature of SMCP. Database searches with SMCP without the low-complexity filter detect closely related SMCPs and a variety of unrelated proline- and/or cysteine-rich proteins such as dipteran small proline-rich proteins, plant cell-wall proteins, and mammalian keratin-associated proteins (not shown). In addition, the variations in the sequence, length, and number of repeat units create many gaps in alignments of SMCPs in distantly related species that do not correspond to repeat units. The ratio of SMCPs to unrelated proteins can be improved dramatically with PHI-BLAST, using either of two motifs, [QP]CCP or [PQ]Cx [3,4,5] K.

Having clearly defined the *SMCP* gene and protein, it is possible to determine whether *SMCP* has generated copies. Searches of the mouse, rat, human, bull, and dog genomes with BLASTN (word size 7) and FASTA3.3 (ktuple 4) with the appropriate *SMCP* cDNA detects only the *SMCP* gene with ~100% identity in the genomes of the same species and no paralogues with lower sequence identity. Similarly, unfiltered TBLASTN searches using protein query SMCP detects unrelated cysteine- and proline-rich proteins, but no sequences encoding proteins with the distinctive features identified here. These findings support the contention that genes that are expressed specifically in meiotic and haploid spermatogenic cells are deficient in generating retroposons [27].

Discussion

SMCP is a rapidly evolving protein, which can now be unequivocally recognized by its small size; a low-complexity

central region containing short repeats rich in cysteine, proline, glutamine, and lysine; a carboxy-terminal region with little cysteine and no repeats; and a small number of conserved amino acids. In addition, the *SMCP* gene can be recognized by the expression of its mRNA at high levels in testis; conserved 5′ UTR and 3′ UTR sequences; a single, large intron 20 nt upstream of the *SMCP* start codon; and its location in the heart of the EDC gene complex. As a uniformly hydrophilic protein, it is likely that SMCP belongs to the very large class of protein regions in higher eukaryotes, known as intrinsically unstructured, which lack hydrophobic globular domains [26]. In addition, ~52% of the amino acids in the SMCP repeat unit, proline, glutamine, and lysine, are correlated with intrinsically unstructured protein regions [25,26].

The structure of SMCP affords insights into the early confusion of SMCP with PHGPx. Presumably, early workers did not anticipate that bull and rat SMCP would differ greatly in size, 103 vs 146 amino acids, and that SMCP, a hydrophilic lysine-rich protein, would be extremely susceptible to degradation by trypsin, an enzyme that was used in purifying the rat mitochondrial capsules [4,6] but was not used in purifying the bull capsules [3].

Similarities in function and sequence argue that the *SMCP* gene is an EDC paralogue. The EDC contains about 50 genes encoding involucrin, loricrin, fillagrin, 10–20 small proline-rich proteins, ~18 late envelope proteins, and ~10 S100 calcium-binding proteins. which collectively function as structural constituents of cornified epithelia in skin and internal epithelial linings [13,14,28,29]. Tissue-specific expression of EDC proteins creates differences in the permeability, strength, and flexibility of various epithelial barriers. SMCP and many EDC proteins are made up of three segments, including a central segment consisting of repeated units, and are constituents of SDS-resistant complexes [3,13,14,28]. Unstructured regions promote protein–protein associations and flexibility [26], properties that are likely shared by SMCP and many EDC proteins. In addition, *SMCP* and many EDC genes have an intron in almost the same position in the 5′ UTR, 20 nt vs 19–23 nt upstream of the ATG codon, and lack introns in the coding region and 3′ UTR [14,28–31]. The similarities in *SMCP* and EDC genes and proteins argue that the progenitor of the *SMCP* gene is not located in another part of the genome.

SMCP has also evolved patterns of expression and functions that differ from those of EDC genes. The distinctive SMCP repeat unit is not repeated in other EDC proteins, the *SMCP* gene has a TATA-independent promoter while many EDC genes have TATA-dependent promoters, and the *SMCP* 5′ UTR and 3′ UTR contain conserved elements that may repress translation in early haploid cells that are absent from EDC mRNAs. The defects in sperm motility and male fertility produced by the *SMCP* gene knockout [8] are very different from the functions of EDC proteins in forming epithelial barriers. In addition, EDC proteins and SMCP are crosslinked in insoluble, keratinous complexes by different mechanisms: EDC proteins are primarily crosslinked by N^{ε}-(γ-glutamyl)lysine isodipeptide bonds formed by transglutaminases, and secondarily by disulfide bridges [32], while bull sperm SMCP is completely solubilized by β-mercaptoethanol and SDS [3], implying that SMCP is crosslinked only by disulfide bridges despite the presence of many glutamine and lysine residues. Thus, SMCP has retained functions similar to those of EDC proteins at the molecular level while acquiring radically new functions at the cellular level. This makes excellent biological sense considering that the selective pressures on sperm, sperm competition, and coevolution with female traits that influence fertilization differ sharply from the selective pressures on somatic cells [1].

Spermatogenic cells appear to be deficient in generating processed retroposons because Southern blots demonstrate that a large number of mRNAs that are expressed at high levels in meiotic and haploid spermatogenic cells are single copy [27]. This observation is unexpected because retrogenes are created in the germ line; hence it is expected that genes that are expressed at high levels in spermatogenic cells would be prolific generators of processed pseudogenes. However, many members of gene families encoding somatic cell-specific isoforms have spawned many processed pseudogenes, while their paralogues encoding sperm-specific isoforms have generated no processed pseudogenes [27]. The present genomic analysis supports this conclusion for bull, dog, mouse, rat, and human *SMCP* using the more sensitive detection methods of BLASTN, FASTA, and TBLASTN.

The primer extension and Northern blot analyses indicate that the levels of *SMCP* mRNA are drastically lower in testicular RNA from human than in dog, mouse, and deer mouse. This represents a species-specific difference in the levels of *SMCP* mRNA, rather than a difference in the proportion of spermatogenic cells in the testes of each species, because the levels of protamine 1 mRNA are similar in our samples of human, dog, and mouse testis RNA. Although few studies have compared mRNA levels in spermatogenic cells in various mammals, the levels of lactate dehydrogenase C and transition protein 2 mRNAs differ dramatically between rats and mice and between rodents and humans, respectively [33,34]. In addition, microarray analyses of various *Drosophila* species demonstrate that the levels of various mRNA species in male germ cells are much more variable than are mRNA levels in somatic tissues and female germ cells [35]. The variability in *SMCP* mRNA levels, minimal conservation of the *SMCP* promoter, and lack of strict conservation in the uORFs suggest that the levels of the SMCP protein may be regulated by different combinations of transcriptional and posttranscriptional mechanisms in different mammals, possibly related to the strong selective pressures and genetic conflicts affecting gene expression in spermatogenic cells [35,36].

Studies in transgenic mice reveal that the *SMCP* 5′ UTR and 3′ UTR repress expression of the GFP reporter by different mechanisms (S.K. Hawthorne et al., in preparation): the *SMCP* 5′ UTR represses translation in free mRNPs, possibly accompanied by small polysomes produced by uORFs, while the 3′ UTR prevents accumulation of GFP without blocking the initiation of translation, a form of translational repression that is mediated by microRNAs [37]. This study identifies conserved sequences in the *SMCP* 3′ UTR and 5′ UTR that may function

in both mechanisms of translational repression as well as uORFs that may reduce the proportion of ribosomes that translate the *SMCP* reading frame [18]. Although many mRNAs that are expressed in mammalian spermatogenic cells contain uORFs [38], the *SMCP* mRNA is the only mRNA in which the uORFs have been shown to be conserved. Transgenic studies also demonstrate that the mouse *Prm1* mRNA is regulated by three mechanisms of translational control [23,39]. The significance of multiple mechanisms of translational control is unclear. Possible explanations include minimizing deleterious effects of premature expression of SMCP in early haploid cells because each mechanism alone is inadequate to suppress translation completely, fine-tuning developmental changes in protein levels, and regulatory conflicts generated by evolutionary pressures on male reproductive success [36].

In summary, the *SMCP* gene encodes a protein that functions in sperm motility and fertilization, processes that are under the intense selection and genetic conflicts associated with sexual selection in males [1,38]. SMCP has two properties that are often associated with rapid evolutionary change, expression in sperm and a segment containing short tandem repeats, which in other genes evolve by expansions and contractions of the number of repeats produced by unequal crossovers and gene conversion as well as nucleotide substitutions [40,41]. In fact, the EDC superfamily and involucrin provide striking examples of evolutionary changes by repeat expansion/contraction [13,14,30]. Owing to confusion of SMCP with PHGPx and the difficulties in detecting SMCP with standard database search tools, SMCP has been overlooked in evolutionary studies of unstructured and sperm-specific proteins. The elucidation of the *SMCP* sequences here will facilitate future studies of the evolution and regulation of the *SMCP* gene.

Materials and methods

RNA purification and RT-PCR of mRNA

Adult inbred male deer mice (*Peromyscus maniculatus bairdii*) were purchased from the Peromyscus Genetic Stock Center (University of South Carolina, Columbia, SC, USA). The deer mice and *Mus musculus* (strain CD-1) were sacrificed by CO_2 asphyxiation, the testes were removed and stripped of the tunica albuginea, and RNAs were extracted from the seminiferous tubules using the Trizol reagent (Invitrogen, Bethesda, MD, USA). RNAs from a variety of mouse somatic tissues were purified similarly. RNAs were digested with RQ1 DNase I (Promega Biotec, Madison, WI, USA) to eliminate DNA. Total human and beagle testis RNA was purchased from Ambion (Austin, TX, USA) and Biochain Institute, Inc. (Breakwater, CA, USA), respectively. Total testis RNA from golden hamster (*Mesocricetus auratus*) and bull was generously donated, respectively, by Drs. H. Wang and D.L. Kilkpatrick (University of Massachusetts Medical Center, Worcester, MA, USA) and Dr. John Glomset (University of Washington, Seattle, WA, USA).

cDNA was synthesized with Superscript II reverse transcriptase (Invitrogen) using 10 pmol of reverse primer and 5 μg of total testis RNA according to the recommendations of the supplier and amplified with *Taq* polymerase using the following conditions: 1.5 mM $MgCl_2$, 95°C for 2 min, 10–30 cycles (95°C for 1 min, 72°C for 2 min), final extension at 72°C for 5 min. The sequences and annealing temperatures of all of the primers used in this study are contained in Table 1.

Analysis of transcription start sites

The transcription start site of the *SMCP* gene in deer mouse, *M. musculus,* and human testes was determined by primer extension. Oligonucleotides were labeled with [γ-^{32}P]ATP and T4 polynucleotide kinase and purified by spin-chromatography on Sephadex G10. Each ^{32}P-labeled oligonucleotide (1×10^5

Table 1
Oligonucleotide primers for RT-PCR, 5′ RACE, and primer extension

Sequence[a]	RNA target, procedure[b]	Product size[c] (anneal. temp.[d])
TTTGCTATCTAAATGTCAGGATCA	Universal *SMCP,* DS, RT-PCR	~650–750 nt (R 42°C; P 53°C)
GCCATGGACTCACTAGACTGC	Universal *SMCP,* US, PCR	–
TCTTCTTTAGAGTTGGAGTTCCTGA	Mouse *Smcp,* DS, primer ext.	(R 49°C)
CACTTGGAGGTACTGGTCGG	Human *SMCP,* DS, primer ext.	(R 51°C)
AGTTGAAGTTCCTGATCTGGTC	Deer mouse *Smcp,* DS, primer ext.	(R 48°C)
CAGCACTTGGGCTGAATG	Dog *Smcp,* DS1, 5′ RACE	(R 42°C)
GGTTTCTGCGGGCAACATGG	Dog *Smcp,* DS2, 5′ RACE	(P 60°C)
CTGGCAGCAGTGATTGTGTT	Human *SMCP,* DS1, RT, 5′ RACE	(R 42°C)
CTGCTGTGGTGGGCAGCATT	Human *SMCP,* DS2, PCR, 5′ RACE	(P 62°C)
GAGCCGGAGCAGATATTACC	Human *PRM1,* DS, RT-PCR	243 nt (R 42°C; P 55°C)
CCTTAGCAGGCTCCTGATTTT	Human *PRM1,* US, 5′ RACE	–
AGATGTGGCGAGATGCTCTT	Mouse *Prm1,* DS, RT-PCR	231 nt (R 42°C; P 55°C)
ATGGCCAGATACCGATGCT	Mouse *Prm1,* US, PCR	–
AGGGCATCAAACAGATACCG	Dog *Prm1,* DS, RT-PCR	238 nt (R 42; P 55°C)
GGGTGGCATGTTCAGGAG	Dog *Prm1,* US, PCR	–
TGGTGATTGAGAGCCCTTCT	Mouse *Smcp,* DS, RT-PCR, Fig. 1	392 nt (R 42°C; P 57°C)
CAACTCTAAAGAAGATGAGTGATCCA	Mouse *Smcp,* US, PCR, Fig. 1	–
AGTCGGCATCGTTTATGGTC	Mouse 18S rRNA, DS, RT-PCR	246 nt (R 42°C; P 57°C)
CCGCAGCTAGGAATAATGGA	Mouse 18S rRNA, US, PCR	–

[a] Sequence of primer, 5′ to 3′.
[b] The RNA species to which the primer anneals and the procedure in which the primer was used. Upstream and downstream primers are abbreviated US and DS, respectively.
[c] The predicted size of the PCR product. The predicted sizes always agreed with the observed size in agarose gels.
[d] The annealing temperature that was used in PCR (P) or reverse transcriptase reactions (R).

cpm) was annealed to total testis RNA at the temperature indicated in Table 1, followed by extension with Thermoscript reverse transcriptase for 30 min at the annealing temperature, followed by extension at 65°C for 30 min using buffers supplied with the enzyme (Invitrogen). The size of the extension product was determined by electrophoresis on an 8 M urea polyacrylamide gel and autoradiography [12] and calibration with the 10 bp ladder (New England Biolabs, Beverly, MA, USA) that had been labeled with T4 DNA polynucleotide kinase and [γ-^{32}P]ATP.

The sequences of the dog and human *SMCP* mRNAs were extended by 5′ RACE using a kit from Invitrogen. Briefly, testis RNA was copied with Superscript II reverse transcriptase and the primers shown in Table 1. The cDNAs were tailed with terminal transferase and dCTP and PCR-amplified with a second reverse primer and the abridged anchor primer provided in the kit (Invitrogen 50359), PCR products were ligated into pGEM-T, and individual clones were sequenced.

DNA sequencing

PCR products were purified by agarose gel electrophoresis, extracted from the gel with a GeneClean kit (Bio101, Carlsbad, CA, USA), and sequenced either directly or after ligation into pGEM-T (Promega Biotec) and purification of the plasmids (Qiagen). The plasmid containing pig *Smcp* cDNA (AW480579) was purchased from BacPac Resources (Oakland, CA, USA). Sequences were determined in the UMass Boston Molecular Genomics Laboratory on both strands of DNAs derived from two to four independent PCRs or clones. The sequences of bull, deer mouse, hamster, and pig and two alleles of dog *SMCP* are deposited under Accession Nos. AY796023, AY788097, AY788096, AY788095, AY788098, and DQ103332, respectively. Sequences were aligned with CLUSTALW 1.82 (European Bioinformatics Institute).

Acknowledgments

This research was supported by NSF Grants MCB-9874491 and BCE-0348497 to K.C.K. We thank Drs. H. Wang and D.L. Kilkpatrick (University of Massachusetts Medical Center) and Dr. John Glomset (University of Washington, Seattle, WA, USA) for the generous donation of bull and hamster testis RNA.

References

[1] T. Birkhead, Promiscuity: An Evolutionary History of Sperm Competition, Harvard Univ. Press, Cambridge, MA, 2000.
[2] E.M. Eddy, K. Toshimori, D.A. O'Brien, Fibrous sheath of mammalian spermatozoa, Microsc. Res. Tech. 61 (2003) 103–115.
[3] V. Pallini, B. Baccetti, A.G. Burrini, A peculiar cysteine-rich polypeptide related to some unusual properties of mammalian sperm mitochondria, in: D.W. Fawcett, J.M. Bedford (Eds.), The Spermatozoon, Urban & Schwartzenberg, Baltimore/Munich, 1979, pp. 141–151.
[4] H.I. Calvin, G.W. Cooper, E. Wallace, Evidence that selenium in rat sperm is associated with a cysteine-rich structural protein of the mitochondrial capsules, J. Reprod. Fertil. 81 (1981) 1–11.
[5] L. Cataldo, K. Baig, R. Oko, M.-A. Mastrangelo, K.C. Kleene, Developmental expression, intracellular localization and selenium content of the cysteine-rich protein associated with the mitochondrial capsules of mouse sperm, Mol. Reprod. Dev. 45 (1996) 320–331.
[6] F. Ursini, S. Heim, M. Kiess, M. Maiorino, A. Rover, J. Wissing, L. Flohe, Dual function of the selenoprotein PHGPx during sperm maturation, Science 285 (1999) 1393–1396.
[7] K. Nayernia, M. Diaconu, G. Aumuller, G. Wennemuth, I. Schwandt, K. Kleene, W. Engel, Phospholipid hydroperoxide glutathione peroxidase: expression pattern during testicular development in mouse and evolutionary conservation in spermatozoa, Mol. Reprod. Dev. 67 (2004) 458–464.
[8] K. Nayernia, I.M. Adham, E. Burkhardt-Goote, J. Neesen, M. Rieche, S. Wolf, U. Sancken, K. Kleene, W. Engel, Asthenozoospermia in mice with a targeted deletion of the sperm mitochondria-associated cysteine-rich protein (*Smcp*) gene, Mol. Cell. Biol. 22 (2002) 3046–3052.
[9] K.C. Kleene, Poly(A) shortening accompanies the activation of translation of five mRNAs during spermiogenesis in the mouse, Development 106 (1989) 367–373.
[10] I.M. Adham, D. Tessmann, K.A. Soliman, D. Murphy, H. Kremling, C. Szpirer, W. Engel, Cloning, expression, and chromosomal localization of the rat mitochondrial capsule selenoprotein gene (MCS): the reading frame does not contain potential UGA selenocysteine codons, DNA Cell Biol. 15 (1996) 159–166.
[11] H.M. Aho, D. Schwemmer, D. Tessmann, G. Murphy, W. Mattei, W. Engel, I.M. Adham, Isolation, expression and chromosomal localization of the human mitochondrial capsule selenoprotein gene (MSCP), Genomics 32 (1996) 84–190.
[12] I.M. Karimpour, M. Cutler, D. Shih, K.C. Kleene, Sequence of the gene encoding the mitochondrial capsule selenoprotein of mouse sperm: identification of three in-phase TGA selenocysteine codons, DNA Cell Biol. 11 (1992) 693–699.
[13] D. Marshall, M.J. Hardman, K.M. Nield, C. Byrne, Differentially expressed late constituents of the epidermal cornified envelope, Proc. Natl. Acad. Sci. USA 98 (2001) 13031–13036.
[14] S. Patel, T. Kartasova, J.A. Segre, Mouse *Sprr* locus: tandem array of coordinately regulated genes, Mamm. Genome 14 (2003) 140–146.
[15] R.A. Padgett, P.J. Grabowski, M.M. Konarska, S. Seiler, P.A. Sharp, Splicing of messenger RNA precursors, Annu. Rev. Biochem. 55 (1986) 1119–1150.
[16] K. Lo, S.T. Smale, Generality of a functional initiator consensus sequence, Gene 1992 (1996) 13–22.
[17] M. Kozak, The scanning model for translation: an update, J. Cell Biol. 108 (1989) 229–241.
[18] D.R. Morris, A.P. Geballe, Upstream reading frames as regulators of mRNA translation, Mol. Cell. Biol. 20 (2000) 8633–8642.
[19] E.A. Leibold, H.N. Munro, Cytoplasmic protein binds in vitro to a highly conserved sequence in the 5′ untranslated region of ferritin heavy- and light-subunit mRNAs, Proc. Natl. Acad. Sci. USA 85 (1988) 2171–2175.
[20] O. Meyuhas, E. Hornstein, Translational control of TOP mRNAs, in: N. Sonenberg, J.W.B. Hershey, M.B. Mathews (Eds.), Translational Control of Gene Expression, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 2000, pp. 671–694.
[21] A.J. Robertson, J.T. Howard, Z. Dominski, B.J. Schnackenberg, J.L. Sumerel, J.J. McCarthy, J.A. Coffman, W.F. Marzluff, The sea urchin stem–loop-binding protein: a maternally expressed protein that probably functions in expression of multiple classes of histone mRNA, Nucleic Acids Res. 32 (2004) 811–818.
[22] R. Mendez, J.D. Richter, Translational control by CPEB: a means to the end, Nat. Rev. Mol. Cell Biol. 2 (2001) 521–529.
[23] J. Zhong, A.H.F.M. Peters, K. Kafer, R.E. Braun, A highly conserved sequence essential for translational repression of the protamine 1 messenger RNA in murine spermatids, Biol. Reprod. 64 (2001) 1784–1789.
[24] S.-Y. Nam, S. Maeda, M. Fujisawa, M. Kurohmaru, Y. Hayashi, Cloning and expression of mitochondrial capsule selenoprotein gene in the golden hamster, J. Vet. Med. Sci. 60 (1998) 1113–1118.
[25] S. Lise, D.T. Jones, Sequence patterns associated with disordered regions in proteins, Proteins 58 (2005) 144–150.
[26] H.J. Dyson, P.E. Wright, Intrinsically unstructured proteins and their functions, Nat. Rev. Mol. Cell. Biol. 6 (2005) 197–208.
[27] X. Zhong, K.C. Kleene, cDNA copies of the testis-specific lactate dehydrogenase (LDH-C) mRNA are present in spermatogenic cells in mice, but processed pseudogenes are not derived from mRNAs that are expressed in haploid and late meiotic spermatogenic cells, Mamm. Genome 10 (1999) 6–12.
[28] S. Gibbs, R. Funeman, J. Wiegant, A.G. van Kessel, P. van De Putte, C. Backendorf, Molecular characterization and evolution of the SPRR

family of keratinocyte differentiation markers encoding small proline-rich proteins, Genomics 16 (1993) 630–637.
[29] H.J. Song, G. Poy, N. Darwiche, H.J. Song, G. Poy, N. Darwiche, U. Lichti, T. Kuroki, P.M. Steinert, T. Kartasova, Mouse *Sprrr2* genes: a clustered family of genes showing differential expression in epithelial tissues, Genomics 55 (1999) 28–34.
[30] B. Delhomme, P. Djian, Expansion of mouse involucrin by intra-allelic repeat addition, Gene 252 (2000) 195–207.
[31] K. Yoneda, D. Hohl, O.W. McBride, M. Wanl, K.U. Cehrs, W.W. Idler, P.M. Steinert, The human loricrin gene, J. Biol. Chem. 267 (1992) 18060–18066.
[32] M. Simon, The epidermal cornified envelope and its precursors, in: I. Leigh, E. Land, F. Watt (Eds.), The Keratinocyte Handbook, Cambridge Univ. Press, Cambridge, UK, 1994, pp. 275–292.
[33] K.I. Salehi-Ashtiani, E. Goldberg, Differences in regulation of testis specific lactate dehydrogenase in rat and mouse occur at multiple levels, Mol. Reprod. Dev. 35 (1993) 1–7.
[34] G. Schlueter, H. Kremling, W. Engel, The gene for human transition protein 2: nucleotide sequence, assignment to the protamine gene cluster, and evidence for its low expression, Genomics 143 (1992) 377–383.
[35] C.D. Meiklejohn, J. Parsch, D.L. Hartl, Rapid evolution of male-biased expression in *Drosophila*, Proc. Natl. Acad. USA 100 (2003) 9894–9899.
[36] K.C. Kleene, Sexual selection, genetic conflict, selfish genes and the atypical patterns of gene expression in spermatogenic cells, Dev. Biol. 277 (2005) 16–26.
[37] P.H. Olsen, V. Ambros, The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of translation, Dev. Biol. 216 (1999) 671–680.
[38] K.C. Kleene, A possible meiotic function of the peculiar patterns of gene expression in mammalian spermatogenic cells, Mech. Dev. 106 (2001) 3–23.
[39] M.A. Fajardo, H.S. Haugen, C.H. Clegg, R.E. Braun, Separate elements in the 3′ untranslated region of the mouse protamine 1 mRNA regulate translational repression and activation during murine spermatogenesis, Dev. Biol. 191 (1997) 42–52.
[40] W.K. Swanson, V.D. Vacquier, The rapid evolution of reproductive proteins, Nat. Rev. Genet. 3 (2002) 137–144.
[41] P. Tompa, Intrinsically unstructured proteins evolve by repeat expansion, Bioessays 25 (2003) 847–855.