Discovery of eight novel divergent homologs expressed in cattle placenta

Joshua H. Larson, Charu G. Kumar, Robin E. Everts, Cheryl A. Green, Annelie Everts-van der Wind, Mark R. Band, and Harris A. Lewin

Department of Animal Sciences, University of Illinois at Urbana-Champaign, Urbana, Illinois

Submitted 15 December 2005; accepted in final form 23 February 2006

Larson, Joshua H., Charu G. Kumar, Robin E. Everts, Cheryl A. Green, Annelie Everts-van der Wind, Mark R. Band, and Harris A. Lewin. Discovery of eight novel divergent homologs expressed in cattle placenta. Physiol Genomics 25: 405–413, 2006. First published March 22, 2006; doi:10.1152/physiolgenomics.00307.2005.—Ten divergent homologs were identified using a subtractive bioinformatic analysis of 12,614 cattle placenta expressed sequence tags followed by comparative, evolutionary, and gene expression studies. Among the 10 divergent homologs, 8 have not been identified previously. These were named as follows: cattle cerebrum and skeletal muscle-specific transcript 1 (CSSMST1), cattle intestine-specific transcript 1 (CIST1), hepatitis A virus cellular receptor 1 amino-terminal domain-containing protein (HAVCRNDP), prolactin-related proteins 8, 9, and 11 (PRPR8, PRPR9, and PRPR11, respectively) and secreted and transmembrane protein 1A and 1B (SECTM1A and SECTM1B, respectively). In addition, two previously known divergent genes were identified, trophoblast Kunitz domain protein 1 (TKDP1) and a new splice variant of TKDP4. Nucleotide substitution analysis provided evidence for positive selection in members of the PRP gene family, SECTM1A and SECTM1B. Gene expression profiles, motif predictions, and annotations of homologous sequences indicate immunological and reproductive functions of the divergent homologs. The genes identified in this study are thus of evolutionary and physiological importance and may have a role in placental adaptations.

Comparisons between the whole genome sequences of human, mouse, and rat revealed extensive lineage-specific expansions in gene families associated with reproduction, immunity, chemosensation, detoxification, and proteolysis (48, 52). The identification of orthologous relationships for ~80% of genes in the human and mouse genomes indicates that lineage-specific gene family expansion likely represents a major route of mammalian evolution. These lineage-specific paralogs are believed to represent adaptive innovations on the basis of evidence that positive selection has operated on a subset of proteins functioning in reproduction and immunity (48). It is likely that the Cetartiodactyla experienced similar numbers of lineage-specific gene family expansions, given that cetartiodactyls diverged from the common ancestor of the Rodentia and primates ~94 million years ago (MYA), before the rodent-primate divergence ~87 MYA (62). It is important to determine which genes or gene families have been subjected to diversification and what roles these genes played in generating adaptive cetartiodactyl phenotypes.

In the present study, eight new divergent homologs expressed in cattle placenta were identified using a subtractive bioinformatic approach previously developed for the identification of the cattle major histocompatibility complex class I-like gene family A genes (38). The combined approaches of sequence alignment and comparative mapping were used to establish homologous relationships. Biological functions were inferred from the comparison of sequence motif predictions and microarray expression profiles to existing annotations for homologous proteins. Lastly, nucleotide sequence substitution analysis provided evidence for positive selection on two groups of divergent homologs.

MATERIALS AND METHODS

Sequences and bioinformatics. More than 12,600 cattle expressed sequence tags (ESTs) from term placenta (dbEST library no. 3743) were interrogated by basic local alignment search tool (BLAST)N (1) to identify similarity to sequences in the GenBank nonredundant (February 15, 2005 release; cetartiodactyl sequences removed), dbEST (February 25, 2005 release; cetartiodactyl sequences removed), and human genome draft sequence (NCBI build 35.1) databases. Cattle ESTs with significant alignments to these sources were removed from the working set to enrich for highly divergent sequences. An E-value significance threshold of <e^-3 was used for GenBank nonredundant and dbEST databases; E-value <e^-10 was the significance threshold for human genome draft. Repetitive elements were masked using RepeatMasker (http://www.repeatmasker.org/).

The divergent 5′-cattle placenta ESTs were aligned by BLASTN to all other cattle GenBank sequences (excluding cattle placenta library no. 3743). Database sequences that aligned significantly were clustered with the working set of divergent 5′-ESTs using an in-house program that required 95% identity and 40-bp overlap. Consensus sequences were assembled using CAP3 (24). The BLASTN pipeline...
Fig. 1. Microarray expression data for 7 divergent homologs (for details, see RESULTS), using a diverse panel of cattle tissues. Expression is shown as the log2 of the normalized ratio of signal intensities in each cattle tissue compared with a reference standard. The tissues are indicated as follows: CR, cerebrum; HE, heart; KI, kidney; LI, large intestine; LV, liver; MU, skeletal muscle; PL, placenta; RU, rumen; SI, small intestine; SK, skin. Data were excluded when transcript expression level did not exceed background by 3 standard deviations in a particular tissue. Significance of transcript expression level difference between tissue and reference standard was determined by t-test; associated P values are indicated: +P < 0.05, ++P < 0.01, +++P < 0.005, and ++++P < 0.001. Only tissues with significant effects are shown (data not shown for adrenal gland, cerebellum, lung, lymph node, spleen, testis, thalamus, and thymus).

(Supplemental Table S1; available at the Physiological Genomics website) was used to discard sequences with significant alignments.

Sequencing the 3′-ends of each cDNA clone with a divergent 5′-EST generated full-clone sequences. Bacterial stocks containing cDNA clones with divergent 5′-ESTs were cultured and rearrayed using a Biorobot 8000 (Qiagen, Valencia, CA), and template plasmid DNA was isolated according to the manufacturer’s protocol. Cycle sequencing and electrophoresis were performed using an established protocol (38), substituting an oligo(d)T18V sequencing primer. PHRED (15) was used to remove low-quality sequence (PHRED score <20).

The BLASTN pipeline was implemented to discard clones with 3′-ESTs displaying significant similarity to sequences in the databases, using the significance thresholds described above. The remaining divergent 3′-ESTs were clustered with the divergent 5′-ESTs and all cattle ESTs in dbEST (excluding cattle placenta library no. 3743). The clusters were assembled, and clones with nonoverlapping 5′- and 3′-regions were sequenced by primer walking (38). The BLASTN pipeline was then used to remove full-clone sequences with similarity to database sequences.

The full-length divergent cDNA sequences were interrogated by BLASTX (2) against the GenBank coding sequence database (including eukaryotic sequences; January 9, 2005) to identify distant homologies. Cattle sequences that aligned to database sequences with E-value ≤e−10 were designated as divergent homologs. Full-clone sequences were conceptually translated using SIXFRAME (http://workbench.sdsc.edu/). Usage of all homology-related terms is according to Koonin (36).

Multiple alignments were constructed using CLUSTALX (27). Signal peptides, transmembrane (TM) domains, and conserved domains were predicted using PSORTII (49), TMPred (23), TMHMM 2.0 (http://www.cbs.dtu.dk), and the conserved domain database (44). Glycosylation, glycoporphadylidinositol (GPI) anchors, and additional predictions were carried out using NetNGlyc and NetOGlyc (http://www.cbs.dtu.dk), big-Pi predictor (10), and PredictProtein (36). Homology modeling was performed using Swiss-Model and Swiss-PdbViewer (17). Sequence data have been deposited with the DDBJ/EMBL/GenBank data libraries under the following accession numbers: AY563705–AY563905, AY741372–AY741374, BK005429–BK005441, and CK39994–CK399219.

Radiation hybrid mapping. To confirm predicted homologous relationships, we performed radiation hybrid (RH) mapping and compared the results with positions in the human genome on the basis of comparative mapping information (14). The divergent homologs were mapped using a 5,000-rad cattle-hamster RH panel (69). Primer design, PCR, and mapping were carried out as previously described (Ref. 14 and Supplemental Table S2). Homologous positions in the human genome (NCBI build 35) were identified using the UCSC Genome Browser (31).

Microarray analysis. To elucidate gene expression patterns and to gain insight into biological function, expression profiles were determined for 7 of the 10 transcripts using a 7,653-element cattle cDNA microarray and RNAs collected from placenta and 17 additional cattle tissues (Fig. 1 and M. R. Band, R. E. Everts, Z. L. Liu, D. E. Morin, J. U. Peled, S. L. Rodriguez-Zas, and H. A. Lewin, unpublished data; NCBI GEO: GSE3029). Total RNA was isolated from a Hereford cow term placenta using an established protocol (38). Twenty micrograms of total RNA and titrated RNA spiking controls (5) were used as templates for the synthesis of aminoallyl-labeled cDNA (aa-cDNA) using an indirect labeling protocol (20). First-strand aa-cDNA was purified and fluorescently labeled as previously described (68). The aa-cDNA synthesis and labeling procedures were also carried out on a standard reference RNA sample developed in our laboratory for microarray studies (13). Labeled cDNA samples were purified and hybridized to a microarray consisting of 7,653 cattle spleen- and placenta-derived cDNAs as described previously (13, 68). Slides were scanned using an Axon 4000B scanner, images were analyzed with GenePix 4.0 (Molecular Devices, Union City, CA), and data were normalized with GeneSpring 5.0 (Silicon Genetics, Redwood City, CA) using the approach of Band et al. (5). Microarray data have been deposited with the National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) under the accession no. GSE3029.

Substitution analysis. Ratios of nonsynonymous to synonymous substitutions (ds/ds, or ω) were determined for the divergent homologs using the phylogenetic analysis by maximum likelihood (PAML) software package (74) to identify evidence of positive Darwinian selection as a contributing factor in their sequence divergence. Sequences used for the substitution analyses included the following: cattle CCSMST1 (BK005429), chimpanzee LOC453815 (XM_510728), mouse BCO03965 (XM_128528), cattle HAVCRNDP (BK005433), cattle predicted HAVCR1 (BF043390), cattle SECTM1A (BK005436), cattle SECTM1B (BK005437), cattle PRP1 (NM_174159), cattle PRP2 (M27239), cattle PRP3 (NM_174160),
cattle PRP4 (M33269), cattle PRP5 (NM_212448), cattle PRP6 (X59504), cattle PRP7 (AB187564), cattle PRP8 (BK005439), cattle PRP9 (BK005440), and cattle PRP11 (BK005438). Signal peptide and TM regions were omitted. The YN00 program in PAML was used to estimate global $d_\text{S}/d_\text{S}$ ($\omega_0$) for each group of aligned sequences using the method of Yang and Nielsen (75). Values of $\omega_0 > 1.0$ indicate that positive selection acted on the sequences analyzed (50). For the PRP genes, the CODEML program in PAML was used to identify variation in selection intensity. The data were modeled using maximum likelihood methods (76), and the results were compared to obtain a test statistic. Three comparisons were performed. Model M1, a neutrality model that constrained $\omega$ to be either 0 or 1, was compared with both M2, a selection model that added an additional $\omega$ ratio class estimated from the data, and M3, a selection model that used an unconstrained discrete distribution to model classes of $\omega$ ratios. Our analysis used three discrete classes for M3. We also compared M7, a continuous distribution neutrality model that estimates $\omega$ using a beta function limited to the interval from 0 to 1, with M8, a continuous distribution selection model that adds an additional class of sites with $\omega$ estimated from the data and not constrained to the interval between 0 and 1. A test statistic of twice the negative value of the difference in log likelihood values generated under each model was obtained from the difference in the number of model parameters (M2 vs. M1, df = 2; M3 vs. M1, df = 4; M8 vs. M7, df = 2). Posterior probabilities for PRP sites under positive selection were generated under M8.

RESULTS

A starting set of 12,614 term cattle placenta 5'-ESTs was reduced to 2,379 divergent 5'-ESTs after BLASTN subtraction and further reduced to 1,595 divergent sequences after BLASTN subtraction was implemented for the assembled consensus sequences (Supplemental Table S1). Sequencing of 3'-ESTs was successful for 1,180 clones with 5'-ESTs in the working set, and BLASTN subtraction of the 3'-ESTs reduced the number of putative divergent clones to 776. BLASTN subtraction of full-clone sequences reduced the working set to 368 high-probability divergent transcripts with no significant nucleotide similarity to sequences derived from organisms outside the Cetartiodactyla. Distant peptide homologies were identified for 10 transcripts using BLASTX (Supplemental Table S3). Conceptual translations of the 10 divergent cattle homologs were characterized with respect to predicted signal peptides, conserved domains, and TM regions (Table 1). Orthologous and/or paralogous relationships are proposed below for each cattle sequence on the basis of their alignment and comparative mapping data.

CCSMST1. A translation of the cattle transcript BK005429 shares 64 and 59% sequence identity with predicted proteins XP_510728 of chimpanzee and XP_128528 of mouse, respectively (Supplemental Fig. S1). The genes associated with these protein sequences were thus classified as putative orthologs, and orthology was subsequently confirmed by RH mapping (Supplemental Table S4). The cattle gene was named cattle cerebrum and skeletal muscle-specific transcript 1 (CCSMST1) on the basis of its transcript expression profile (Fig. 1). Alignment of the CCSMST1 transcript to the current draft of the cattle genome (build 2.0) showed identity to genomic contig NW_930368 assigned to BTA25, supporting the RH mapping data. The first 52 bases of the transcript make up the first exon, which contains the 5'-untranslated region (UTR) and the majority of the signal peptide. A 78-bp intron separates the first exon from a second exon containing the remainder of the coding sequence and 3'-UTR. Substitution analysis revealed $\omega_0 = 0.556$ for the CCSMST1 orthologs, suggesting that nonsynonymous substitutions may be selected against in this gene.

CIST1. A translation of the cattle transcript BK005441 shares 49 and 30% sequence identity with the predicted human CIST1 ortholog and chimpanzee protein XP_530016, respectively (Supplemental Fig. S2). The cattle ortholog was named cattle intestine-specific transcript 1 (CIST1) on the basis of its transcript expression profile (Fig. 1). Query of the cattle draft genomic sequence with the CIST1 transcript sequence revealed identity to genomic contig NW_931809 assigned to BTA7, supporting the RH mapping data (Supplemental Table S4). The genomic structure consists of a 175-bp exon consisting of the 5'-UTR and signal peptide, a 2,603-bp intron, a 390-bp exon, a 1,375-bp intron, a 152-bp exon, a 118-bp intron, and a 114-bp exon containing the remainder of the coding sequence and 3'-UTR, respectively. Cattle CIST1 contains an extracellular serine/threonine/proline (S/T/P)-rich mucin-like domain; these three residue types constitute 56% of this region. Nearly all of the serine and threonine residues in this region represent the serine/threonine/proline (S/T/P)-rich mucin-like domain; these three residue types constitute 56% of this region. Nearly all of the serine and threonine residues in this region represent O-linked glycosylation sites with significant probability. CIST1 and its orthologs were not evaluated by substitution analysis methods because of extensive nonconserved regions in the coding sequences.

HAVCRNDP. A translation of the cattle transcript BK005433 shares 48, 47, and 58% sequence identity with

### Table 1. Bioinformatic characterization of the divergent homolog protein sequences

<table>
<thead>
<tr>
<th>Protein Accession No.</th>
<th>Cattle Nomenclature</th>
<th>Length, AA</th>
<th>Signal Peptide</th>
<th>NCBI Conserved Domains</th>
<th>TM Domain</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAA05601</td>
<td>cattle cerebrum and skeletal muscle-specific transcript 1 (CCSMST1)</td>
<td>135</td>
<td>1–15</td>
<td>none</td>
<td>78–99</td>
</tr>
<tr>
<td>DAA05613</td>
<td>cattle intestine-specific transcript 1 (CIST1)</td>
<td>235</td>
<td>1–31</td>
<td>none</td>
<td>183–205</td>
</tr>
<tr>
<td>DAA05605</td>
<td>hepatitis A virus cellular receptor 1 NH2-terminal domain-containing protein (HAVCRNDP)</td>
<td>131</td>
<td>1–20</td>
<td>22–108 (immunoglobulin; smart00409)</td>
<td>none</td>
</tr>
<tr>
<td>DAA05611</td>
<td>prolactin-related protein 8 (PRP8)</td>
<td>238</td>
<td>1–36</td>
<td>36–237 (somatotropin; pfam00103)</td>
<td>none</td>
</tr>
<tr>
<td>DAA05612</td>
<td>prolactin-related protein 9 (PRP9)</td>
<td>239</td>
<td>1–36</td>
<td>36–237 (somatotropin; pfam00103)</td>
<td>none</td>
</tr>
<tr>
<td>DAA05610</td>
<td>prolactin-related protein 11 (PRP11)</td>
<td>238</td>
<td>1–36</td>
<td>36–237 (somatotropin; pfam00103)</td>
<td>none</td>
</tr>
<tr>
<td>DAA05608</td>
<td>secreted and transmembrane protein 1A (SECTM1A)</td>
<td>185</td>
<td>1–27</td>
<td>none</td>
<td>none</td>
</tr>
<tr>
<td>DAA05609</td>
<td>secreted and transmembrane protein 1B (SECTM1B)</td>
<td>202</td>
<td>1–28</td>
<td>none</td>
<td>none</td>
</tr>
<tr>
<td>AAFP61247</td>
<td>trophoblast Kunitz domain protein 1 (TKDP1)</td>
<td>351</td>
<td>1–20</td>
<td>293–344 (Kunitz; smart00131)</td>
<td>none</td>
</tr>
<tr>
<td>DAA05607</td>
<td>trophoblast Kunitz domain protein 4 (TKDP4)</td>
<td>155</td>
<td>1–20</td>
<td>86–138 (Kunitz; smart00131)</td>
<td>none</td>
</tr>
</tbody>
</table>

NCBI, National Center for Biotechnology Information.
human hepatitis A virus cellular receptor 1 (HAVCR1), mouse
T cell immunoglobulin and mucin domain-containing mole-
cule 1 (TIM-1), and a conceptual translation of cattle EST
BF043390, (the putative cattle HAVCR1 ortholog), respec-
tively (Supplemental Fig. S3). Query of dbEST with the
BK005433 transcript revealed numerous ESTs aligning with
100% identity and no evidence of alternate splicing. Query
of the cattle draft genomic sequence with the BK005433
transcript showed identity to the unassigned cattle genomic
contig NW_973071. However, RH mapping demonstrated
HAVCRNDP to be located on BTA7, in proximity to IL12B
(Supplemental Table S4). The genomic structure consists of a
210-bp exon containing 5'UTR, a 5,146-bp intron, a 68-bp
exon containing 5'UTR and the majority of the signal peptide,
a 1,954-bp intron, a 330-bp exon, a 2,779-bp intron, and a
239-bp exon consisting of the remainder of the coding se-
quence and 3'UTR, respectively. The BK005433 transcript
does not align with the cattle genomic sequence containing
cattle HAVCR1, indicating that these loci represent distant
paralogs. The cattle gene was thus given the name hepatitis A
virus cellular receptor 1 amino-terminal domain-containing
protein (HAVCRNDP). The HAVCRNDP molecule terminates
at approximately the start of the mucin-like region found in the
HAVCR1 proteins (16, 25). HAVCRNDP was significantly
overexpressed relative to the reference standard in liver, small
intestine, and skin tissues (Fig. 1). Substitution analysis of the
cattle HAVCRNDP and HAVCR1 paralogs was performed on
the conserved immunoglobulin (Ig) domain region only. The
HAVCRNDP paralogs had ωs = 0.770, indicating that positive
selection does not act on this gene or that nonsynonymous
substitutions are weakly selected against.

PRP8, PRP9, and PRP11. The PRP molecules described
here were designated PRP8, PRP9, and PRP11 to conform with
the revised PRP nomenclature (Refs. 9, 58; Fig. 2). Cattle
PRP8, encoded by transcript BK005439, is most closely related
to PRP1 (98% identity), whereas PRP9, encoded by transcript
BK005440, is most closely related to PRP5 (97% identity), and
PRP11, encoded by transcript BK005438, is most closely
related to PRP3 (85% identity). Each of the newly described
PRP transcripts have been uniquely localized to genomic
sequence with 100% identity (PRP8 to AAFC01015348, PRP9
to AAFC01043996, PRP11 to AAFC01039561), indicating that
these sequences represent paralogs. However, definitive proof
of paralogy or allelism for all the PRPs awaits complete
sequencing of this region in the cattle genome. PRP genes have
not been identified in the human genome; however, the cattle
PRP genes on BTA23 (Supplemental Table S4) share a con-
served chromosome position with the human prolactin (PRL)

Fig. 2. Protein alignment of the cattle prolactin-related protein (PRP) family including PRP1 (NP_776584), PRP2 (AAA30728), PRP3 (NP_776585), PRP4
(AAA30740), PRP5 (NP_997613), PRP6 (CAA42092), PRP7 (BAD74059), PRP8 (DA05611), PRP9 (DA05612), and PRP11 (DA05610). Alignment gaps are
represented by periods, and aligned PRP sites sharing identity with cattle PRP1 are represented by dashes. The signal peptide region (sp) is represented by
a solid underscore appearing above the alignment. Universally conserved cysteine and tryptophan residues are annotated with Cs and Ws, respectively, beneath
the alignment. Predicted N-linked glycosylation motifs are represented by underlined text. Positively selected sites are indicated by asterisks above the
alignment, and posterior probabilities associated with each positively selected site (see Table 2) are represented within the alignment by normal text (probability
>0.90), bold text (probability >0.95), and italic and bold text (probability >0.99).
The PRP8 locus on HSA6p22.3 (9). PRP8 was significantly overexpressed relative to the reference standard in heart, skeletal muscle, and placenta tissues, and PRP11 was significantly overexpressed relative to the reference standard in placenta tissue (Fig. 1). There was no sequence element on the array corresponding to PRP9, so its expression pattern could not be evaluated.

Substitution analysis showed the PRP paralogs to have $\omega_i = 1.457$, indicating that these genes have been subjected to positive selection (Ref. 50 and Supplemental Fig. S4). We found significant heterogeneity in selection intensity within the PRP molecules and identified 21 molecular sites appearing to be under a high degree of positive selection (Table 2, Fig. 2). Twenty of the twenty-one selected sites were mapped onto the tertiary structure of human PRL [Molecular Modeling Database (MMDB): 1N9D; Fig. 3]. Fifteen selected residues are outwardly directed, indicating that the positive selection observed at these residue positions is likely exerted by an interaction with another molecule. There are no outwardly directed positively selected sites between PRP1 residues 117 and 180, a region with little evidence for positive selection and corresponding to helices 2 and 3 of the predicted PRP molecule on the basis of homology modeling (11, 30). Two relative “hotspots” of positively selected residues flank this region. The first hotspot, from residues 108 to 116, corresponds to the extended loop preceding helix 2. The second hotspot, from residues 181 to 198, is contained within the extended loop distal to helix 3. Thirty-two prolactin receptor-binding residues identified in the ovine PL molecule (11) were mapped onto the PRL molecule, and only four PL receptor-binding residues were found to overlap with the PRP sites under positive selection.

**SECTM1A and SECTM1B.** Cattle secreted and transmembrane protein 1A (SECTM1A), encoded by transcript BK005436, shares 32 and 27% sequence identity with the human and mouse SECTM1 proteins, respectively (Supplemental Fig. S5). Cattle SECTM1B, encoded by transcript BK005437, shares 30 and 26% sequence identity with the human and mouse molecules, respectively. SECTM1A and SECTM1B are 72% identical. The loci were named **SECTM1A** and **SECTM1B** to differentiate the two cattle paralogs, because only a single **SECTM1** gene has been described in other species. SECTM1A contains a hydrophobic region from residues 167 to 184 preceded by a GPI anchor site at residue 163. Only partial sequences of **SECTM1A** and **SECTM1B** are present within the cattle genome draft sequence contigs (GenBank: NW_929537 and NW_929532, respectively), so determination of the genomic organization of the SECTM1 paralogs awaits completion of the cattle genome sequence. RH mapping demonstrated both **SECTM1A** genes to be located on BTA19, in proximity to tubulin-specific chaperone D (TBCD) (Supplemental Table S4). There were no sequence elements on the array corresponding to **SECTM1A** or **SECTM1B**, so expression patterns could not be evaluated. Substitution analysis revealed that the **SECTM1** paralogs had $\omega_i = 1.137$, suggesting that these genes have been subjected to positive selection (50).

**TKDP1 and TKDP4.** Cattle trophoblast Kunitz domain protein 1 (TKDP1) and TKDP4 were recently characterized (37, 42). The TKDP4 molecule identified in our study, encoded by transcript BK005435, is a splice variant lacking residues 32–59 of the previously described TKDP4 protein (42). TKDP genes have not been identified in the human genome; however, the TKDP genes on BTA13 (Supplemental Table S4) share a conserved chromosomal position with SPINLW1 on HSA20q13.12, which encodes a protein with a COOH-terminal Kunitz domain (53) and may share a common ancestry with the cattle TKDP genes. TKDP4 was significantly overexpressed relative to the reference standard in cerebrum, skeletal muscle, and placenta tissues, but TKDP1 did not show any preference in tissue expression pattern (Fig. 1). The TKDP paralogs were represented in normal text (probability $> 0.90$), bold text (probability $> 0.95$), and italic and bold text (probability $> 0.99$).

**Table 2. Likelihood ratio tests of $\omega$ variation and identification of molecular sites under positive selection in the cattle prolactin-related proteins**

<table>
<thead>
<tr>
<th>Sequences</th>
<th>$-2(M2$ vs. $M1)$</th>
<th>$-2(M3$ vs. $M1)$</th>
<th>$-2(M7$ vs. $M8)$</th>
<th>Parameter Estimates for $M8$</th>
<th>Positively Selected Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cattle PRPs</td>
<td>37.7</td>
<td>37.7</td>
<td>37.6</td>
<td>$p_1 = 0.362$, $p_2 = 0.638$, $\omega = 3.199$, $\beta = 0.074$, $0.014$</td>
<td>47, 50, 63, 69, 73, 108, 112, 113, 114, 116, 139, 162, 181, 185, 186, 192, 195, 198, 214, 217, 235</td>
</tr>
</tbody>
</table>

*M1, M2, M3, M7, and M8 refer to maximum likelihood models of $\omega$ ratios, and $-2(M2$ vs. $M1)$, $-2(M3$ vs. $M1)$, and $-2(M8$ vs. $M7)$ indicate the negative of 2 times the log likelihood difference between the selection and neutral models compared. The resulting test statistics are highly significant ($P < 0.005$). For $M8$, $p_1$ is the proportion of positively selected sites, $p_2$ is the proportion of sites not under positive selection, $\omega$ is the $d_2/d_0$ ratio for the selected sites, and $\beta$ describes the beta distribution function. Positively selected prolactin-related protein (PRP) sites are presented according to their numbered positions in the PRP1 preprotein sequence. Posterior probabilities for positively selected sites are represented in normal text (probability $> 0.90$), bold text (probability $> 0.95$), and italic and bold text (probability $> 0.99$).
were excluded from substitution analysis due to heterogeneous domain organization.

DISCUSSION

The identification of genes that are highly divergent among closely related species is an important step in understanding the phenotypic changes that accompany adaptive evolution. The advent of high-throughput DNA sequencing and associated genome databases permit a more global search for those genes that diversify through copy number and primary sequence. In the absence of complete genome sequence, EST collections offer a means to search for divergent homologs as well as transcripts and genes that are unique to a given species (38). In the present study, cattle placenta was chosen as a source of ESTs, because the placenta is among the most phenotypically divergent mammalian tissues and has been shown to express novel paralogs such as the PAG and TKDP genes (42, 71). Our subtractive BLAST approach yielded eight previously unknown divergent homologs, the previously identified TKDP1 gene, and a novel splice variant of TKDP4. On the basis of cattle-human comparative mapping data (14), each divergent cattle gene shows conserved chromosomal position with respect to its predicted human ortholog (Supplemental Table S4). Gene expression analyses using microarrays proved to be highly effective in determining their tissue distributions and for primary functional annotations. For example, expression patterns were used to name CIST1 and CCSMST1, two genes with unannotated orthologs in other species. Our results support these computationally predicted orthologous gene models, thus demonstrating that our approach is also useful for annotating chimpanzee, human, mouse, and other genomes.

CCSMST1 appears to be expressed on the cell surface on the basis of its signal peptide and TM region (Table 1, Supplemental Fig. S1). Although there is insufficient evidence for positive selection over the entire molecule, the ω is a stringent statistic and not likely to detect positive selection in subregions (12). Detection of heterogeneous selection intensity in the CCSMST1 genes is not possible at this time because of a lack of sufficient orthologs for a meaningful analysis (3).

The CIST1 signal peptide and TM domain indicate that the protein is expressed on the cell surface (Table 1, Supplemental Fig. S2). The high expression of CIST1 in large intestine, small intestine, rumen, and kidney tissues suggests that the protein has a gastrointestinal and genitourinary function (Fig. 1). Numerous O-glycosylation sites in the S/T/P-rich mucin-like domain indicate that CIST1 is highly glycosylated. Low peptide sequence conservation in the extracellular domain among CIST1 orthologs suggests that CIST1 function is mediated by its O-linked carbohydrate moieties.

HAVCR1 of humans and African green monkeys is an integral membrane protein bound by the hepatitis A virus (HAV) (16, 29). Cattle HAVCRNDP is the first HAVCR1 homolog described consisting of only a single secreted NH2-terminal HAVCR1-like Ig domain (Supplemental Fig. S3). Because of the absence of the mucin-like, TM, and intracellular HAVCR1 domains, the function of HAVCRNDP is likely to be specialized and dependent on the Ig domain. The Ig domain of African green monkey HAVCR1 is necessary and sufficient for HAV binding and neutralization, and the HAVCR1 mucin-like domain is necessary for viral genome uncoating (59, 60, 65). A secreted, soluble, HAVCR1-type Ig domain molecule could potentially bind a HAV-like virus, prevent viral uncoating, and limit viral infectivity. Thus HAVCRNDP may function as an antiviral molecule in cattle. Certain Picornaviruses (of which HAV is a member) infect cattle, including the apthoviruses that cause foot and mouth disease and are currently regarded as among the most economically important veterinary pathogens worldwide (35). The high expression of HAVCRNDP in liver and small intestine (Fig. 1) may neutralize such pathogens at sites of infectivity and clear them from the animal. Substitution analysis showed a ω <1, but this does not eliminate the possibility that molecular subregions could be subject to positive selection. Given current data, we hypothesize that HAVCRNDP arose by duplication and divergence of the HAVCR1 locus during ruminant evolution as an adaptive response to an enteric pathogen(s).

Cattle HAVCR1 contains seven previously described genes expressed in cattle placenta throughout gestation (26, 33, 57, 63, 64, 67, 73). PRP proteins are believed to function in implantation, placental development, and fetomaternal interaction throughout gestation, yet no known receptor has been identified (19, 32–34, 46, 72). PRP1 interacts with α2-macroglobulin, which may facilitate presentation of PRP molecules to their molecular targets or allow PRP molecules to regulate the activity of other α2-macroglobulin-bound molecules (28, 34). Heterogeneity in glycosylation is the major structural difference among the PRPs and may ultimately contribute to the functional characteristics of each paralog (34, 77). High expression of PRP8 and PRP11 in term placenta is consistent with the expression profiles observed for other PRPs (Fig. 1). Surprisingly, PRP8 was highly expressed in heart and skeletal muscle.

Substitution analysis demonstrated that the PRPs have been subject to positive selection (Supplemental Fig. S4). Twenty-one high-probability (>90%) selected sites were identified, revealing two hotspots of positive selection flanking PRP helices 2 and 3, helices that may be under purifying selection (Figs. 2 and 3, Table 2). This structural association suggests that purifying selection resulted from a constraining influence such as that imposed by a receptor or ligand on a docking surface necessary for molecular interaction. The α2-macroglobulin molecule may exert this influence, as it performs multiple intrinsic and carrier molecule functions (28, 34) that likely constrain peptide sequence mutability. The positions of the positive selection hotspots next to this putative docking site suggest that the hotspot residues may introduce significant diversity at a functional site(s) in the complex. There appears to be no association between PRP sites under positive selection and receptor-binding sites identified in ovine PL.

Cattle SECTM1A and SECTM1B represent the first discovery of paralogous SECTM1 loci in any species. Tight linkage and high sequence similarity compared with human and mouse SECTM1 suggest that SECTM1A and SECTM1B originated by duplication of an ancestral SECTM1 locus in the cetartiodactyl lineage. In humans and mice, SECTM1 is a predicted TM molecule with an NH2-terminal Ig-like extracellular domain that interacts with CD7 (40, 61). The signal peptide and NH2-glycosylation sites suggest that cattle SECTM1A and SECTM1B are extracellular glycoproteins (Supplemental Fig. S5), although the GPI anchor may allow SECTM1A to function as a cell surface receptor or soluble ligand depending on...
lipase activity and accessibility (18). Given the apparent role of CD7 in the activation of γδ T cells (4, 7), the expression of SECTM1 molecules by term placenta may induce proliferation of γδ T cells to the elevated levels observed in the ruminant neonate (21, 41). The SECTM1 molecules may also modulate maternal immunity (22, 39).

Substitution analysis indicated that the cattle SECTM1 paralogs have been subjected to positive selection. The existence of two cattle SECTM1 proteins may represent the need for both secreted and membrane-associated CD7 ligands in cattle. Alternately, one of the cattle SECTM1 molecules may function as the CD7 ligand, while the other acquired novel function. The latter possibility is more consistent with the sequence divergence of the cattle molecules and the evidence for positive selection.

The TKDP family in cattle is a group of five molecules containing COOH-terminal Kunitz-type domains and novel NH2-terminal domains (37, 42). In the present study, we identified an alternately spliced TKDP4 transcript, suggesting that other TKDP transcripts may be alternatively spliced, as proposed for TKDP1 (42). The TKDP expression data support previous analyses showing that TKDP1 is no longer expressed in term placenta or in any other tissue and that TKDP4 is highly expressed in near-term placenta (37, 42). The expression of TKDP4 in cerebrum and skeletal muscle is of unknown functional significance. The TKDPs are hypothesized to regulate cellular differentiation by inhibiting growth factor proteinases or to protect the conceptus by inhibition of maternal T cell granymes (37, 42). Cattle TKDP4 inhibits trypsin and plasmin in vitro, but other TKDPs showed no protease inhibition, suggesting either a high degree of target proteinase specificity or an alternate function (43). An additional hypothesis, formulated by MacLean et al. (43), is that TKDP molecules regulate the activity of voltage-gated ion channels similar to their Kunitz domain-containing homologs in snake venom. The expression of TKDP4 in cerebrum and skeletal muscle, two tissues reliant on the conduction of action potentials, adds credence to this hypothesis and suggests that the study of extrapolateral expression of TKDP genes could provide clues to additional functions.

One consistent theme in our results is that the divergent homologs are predicted to be expressed extracellularly without exception, despite clear differences in their ancestry. The divergent paralogs show a greater degree of overt positive selection compared with the divergent orthologs, likely because single-copy orthologs are under a greater degree of functional constraint against nonsynonymous sequence substitution than are duplicated paralogous genes. The latter may perform redundant functions and create a more permissive environment for nonsynonymous substitution (51). More sophisticated analyses may be required to detect positive selection within small orthologous sequence groupings and probably require structural and functional data as a guide. Substitution analysis of the CIST1 and TKDP genes could not be performed because of the extensive diversity of domain sequences and organization, but their discovery as divergent homologs suggests that gross divergence in domain structure contributes unique species- or lineage-specific molecules and may represent an adaptive modality in mammals.

Although the functions discussed for the divergent genes are based on bioinformatic predictions and derived functional annotations of homologs, the two functional attributes prevalent in our discoveries are immunity and reproduction. These two biological functions represent major foci of genomic diversity identified in other mammalian species (48, 52). Functional characterization of the divergent proteins identified in this study is required to evaluate practical applications of our findings and is expected to provide insight into the molecular events that guided ruminant evolution.

NOTE ADDED IN PROOF

While this paper was in review, the discovery of two novel prolactin-related protein paralogs, designated PRP8 and PRP9, was reported (Ushizawa K, Takahashi T, Hosoe M, Kaneyama K, and Hashizume K. Cloning and expression of two new prolactin-related proteins, prolactin-related protein-VIII and -IX, in bovine placenta. Reprod Biol Endocrinol 3: 68, 2005). To avoid confusion in nomenclature, the PRP8 gene reported in this study (GenBank accession no. BK005439) will be renamed PRP12. In addition, the PRP9 gene reported in this study (GenBank accession no. BK005440) will be renamed PRP13.

ACKNOWLEDGMENTS

We thank John Morse, Chris Wright, and Laura Guest (W. M. Keck Center for Comparative and Functional Genomics, University of Illinois at Urbana-Champaign) for clone rearraying and sequencing services.

GRANTS

This work was made possible by a grant from the United States Department of Agriculture, National Research Initiative, Project No. AG2002-35205-11625. J. Larson was supported in part by a Cargill Research Excellence Fellowship in Animal Genomics.

REFERENCES


Physiol Genomics • VOL 25 • www.physiolgenomics.org


