Evolution of species-specific major seminal fluid proteins in placental mammals by gene death and positive selection

Camille Meslin 1,2,3,4 Michel Laurin 5,7, Isabelle Callebaut 6, Xavier Druart1,2,3,4, Philippe Monget 1,2,3,4
1 UMR85 Physiologie de la Reproduction et des Comportements, INRA, Nouzilly, F-37380, France
2 UMR7247, CNRS, Nouzilly, F-37380, France
3 Université François Rabelais de Tours, F-37041 Tours, France
4 Institut Français du Cheval et de l’Equitation, Nouzilly, F-37380, France
5 CR2P (UMR 7207), Sorbonne Universités, CNRS/MNHN/UPMC, Muséum National d’Histoire Naturelle, Bâtiment de Géologie, Case Postale 48, 57 rue Cuvier, 75005 Paris, France
6 IMPMC, Sorbonne Universités, UPMC Univ Paris 06, UMR CNRS 7590, Muséum National d’Histoire Naturelle, IRD UMR 206, 75005 Paris, France
7 E-mail: laurin@mnhn.fr

Key words: molecular evolution, multivariate phylogenetic pairwise comparisons, reproductive isolation, speciation, comparative biology, pseudogenisation

Abstract

The seminal fluid is a complex substance composed of a variety of secreted proteins and has been shown to play an important role in the fertilisation process in mammals and also in Drosophila. Several genes under positive selection have been documented in some rodents and primates. Our study documents this phenomenon in several other mammalian taxa. We study the evolution of genes that encode for 20 proteins that are quantitatively predominant in the seminal fluid of at least one out of seven domestic animal species. We analyse the amino acid composition of these proteins for positive selection and for the presence of pseudogenes. Genes that disappeared through pseudogenisation include KLK2 in cattle, horse and mice. Traces of positive selection are found in seven genes. The identified amino acids are located in regions exposed to the protein surface, suggesting a role in the interaction of gametes, with possible impact on the process of speciation. Moreover, we found no evidence that the predominance of proteins in seminal fluid and their mode of evolution are correlated, and the uncoupled patterns of change suggest that this result is not due solely to lack of statistical power.

Introduction

Seminal fluid is a complex composite secretion produced in specialised organs of the male reproductive tract of most metazoans as well as in the testis and in the epididymis (Poiani, 2006). It can influence female reproductive physiology and behaviour and is involved in postcopulatory sexual selection and intersexual conflict (Rice, 1996; Clark et al., 1999; Holland and Rice, 1999; Civetta and Clark, 2000; Chapman et al., 2003). In mammals, components of seminal plasma such as proteins have been shown to influence sperm physiology, i.e. through the interaction with the female tract and survival of gametes. The role of seminal fluid has also been extensively explored in Drosophila, in which accessory gland protein (Acp) have been shown
to influence oviposition, sperm storage, as well as female receptivity to remating (Wolfner, 2002; Chapman and Davies, 2004).

Studies of adaptive evolution have revealed several classes of proteins under positive selection, including those involved in gamete recognition or seminal fluid factors such as proteases, lipocalins and chemokines (Swanson et al., 2001, 2004; Galindo et al., 2003; Meslin et al., 2012). This is particularly true in Drosophila and crickets, where genes encoding some seminal proteins are subject to an unusually high rate of adaptive evolution (Swanson et al., 2001; Andrés et al., 2006; Findlay et al., 2009; Walters and Harrison, 2010; Marshall et al., 2011). In mammals, some rodent and primate seminal proteins have also been shown to be under positive selection (Clark and Swanson, 2005; Karn et al., 2008; Ramm et al., 2009; Dean et al., 2011). Variation in evolutionary rates between lineages may be linked to the intensity of postcopulatory sexual selection (Dorus et al., 2004; Wagstaff and Begun, 2005a; Hurle et al., 2007; Ramm et al., 2008), and led to rapid divergence in seminal fluid proteomes (named the Acp-complement in Drosophila) (Begun and Lindfors, 2005; Mueller et al., 2005; Wagstaff and Begun, 2005b, 2007; Begun et al., 2006). In rodents, this evolutionary process is associated with wide interspecific differences in accessory gland protein extracts and relative testis size between species (Ramm et al., 2009) because positive selection at the surface would imply changes in the function of the protein, especially in the binding of partners, rather than changes in its fold. However, rodents and primates are the only mammalian taxa for which the evolution of genes encoding proteins of seminal fluid has been studied.

The protein composition of seminal plasma has been extensively studied in several domestic animal species. However, a limited number of studies have performed systematic cross-species comparative analyses of seminal plasma proteins using high throughput proteomics (Kelly et al., 2006; Moura et al., 2007; Souza et al., 2012). In fact, while the proteome of human seminal plasma has been comprehensively described with an actual list of more than 2000 proteins identified (Pilch and Mann, 2006; Batruch et al., 2011; Milardi et al., 2012), relatively few of the proteins present within the seminal plasma of the major domestic mammalian species have been identified. We have recently performed such a systematic analysis of seminal fluid proteome in our laboratory using a proteomic strategy including liquid chromatography and mass spectrometry (2DLC MS-MS) (Druart et al., 2013).

The quantitatively major proteins were identified in ram, buck, bull, horse, boar, camel and alpaca. As previously observed in rodent and primate seminal fluids, the identity of the prevailing proteins of the seminal fluid is particularly variable in the species sampled by Druart et al. (2013).

The objective of the present work is to study the evolution of these proteins in placental mammals. In particular, the ‘translational robustness hypothesis’, proposed previously (Drummond et al., 2005), suggests that a high level of expression of a protein is associated with a high degree of sequence conservation between species. In the present paper, we have tested a related hypothesis, which we formulate here. Namely, we test if there is a correlation (positive or negative) between protein abundance in the seminal fluid and intensity of positive selection. Moreover, in our previous works, we showed that in some proteins, amino acids under positive selection are more often at the protein surface rather than in the vicinity of catalytic site (Meslin et al., 2012). We then studied if the position of amino acids under positive selection in other proteins is random in the three-dimensional structure of the proteins. For the latter hypothesis, we are particularly interested in discriminating between the protein surface vs hydrophobic core because positive selection at the surface would imply changes in function of the protein, especially in the binding of partners, rather than changes in its fold.

Material and methods

Phylogenetic and syntenic analyses

Twenty-one proteins were chosen for the analysis, based on their predominance (at least about 20 times greater than the average concentration of the seminal plasma proteins) from at least one species among seven domestic mammals. The abundance was based on the identification of proteins by mass spectrometry in the seminal plasma from seven species performed previously (Druart et al., 2013). In this study, seminal plasma proteins were separated by SDS PAGE and imaged after Coomassie Blue staining (Fig. 1 Supplemental data). This staining is commonly used to detect proteins of high abundance given its moderate sensitivity, and also can provide protein quantification as the intensity of staining is positively correlated to protein amount. The main bands observed after SDS PAGE and Coomassie staining were further subjected
to mass spectrometry (MS) to identify their protein content. Each band contains several proteins from which the one exhibiting the maximum number of MS spectra was selected as the major protein of the band (those having at least approximately 20 times the average protein concentration). Therefore, proteins identified according to 1) high intensity staining after SDS PAGE and Coomassie staining and 2) predominant number of MS spectra, were considered quantitatively major components of the seminal plasma. Finally, RNAse10 and MFGE8 have been included in the analysis because they are specific markers of epididymal maturation in ungulates (Castella et al., 2004; Belleannée et al., 2011). Because of this, not all proteins considered by Druart et al. (2013) are considered here.

This study has sampled the genome of nine placental mammal species that have been fully sequenced (Bos taurus Linnaeus, 1758, Canis lupus familiaris Linnaeus, 1758, Equus caballus Linnaeus, 1758, Homo sapiens Linnaeus, 1758, Mus musculus Linnaeus, 1758, Oryctolagus cuniculus Linnaeus, 1758, Pan troglodytes Blumenbach, 1775, Rattus norvegicus Linnaeus, 1758 and Sus scrofa Linnaeus, 1758). We have worked on the version of Ensembl January 2013 (http://jan2013.archive.ensembl.org/index.html), on the following versions of genomes: human (GRCh37), chimpanzee (CHIMP2.1.4), mouse (GRCm38), rat (Rnor_5.0), rabbit (oryCun2), dog (CanFam3.1), pig (Sscrofa10.2), horse (EquCab2), and cattle (UMD3.1). We have chosen these fully sequenced species because it is possible to find pseudogenes and to test the hypothesis of gene loss.

For all identified genes, the corresponding Ensembl protein ID was retrieved from the Ensembl database and submitted to the PhyleasProg web server v2.3 (http://phyleasprog.inra.fr/) (Busset et al., 2011). All reconstructed phylogenetic trees were carefully examined before interpreting selective pressure results, eventually corrected by synteny analysis as previously described (Tian, Pascal, Foucècourt et al., 2009), so that calculations were performed with correct orthologs.

For the comparative analyses on the relationship between protein abundance and intensity of positive selection and for evolutionary rates, we built a reference phylogeny on the nine species on which this paper focuses. The reference phylogeny (S1, 2) generally follows Murphy and Eizirik’s phylogeny for topology and divergence times (Murphy and Eizirik, 2009), with the exceptions of fairly recent divergences, like artiodactyls (Hassanin et al., 2012), hominids (Vignaud et al., 2002), and murines (Rowe et al., 2008).

**Identification of pseudogenes - Inference of positive selection**

A search for pseudogenes was systematically performed by tBlastn in the studied genomes for genes

<table>
<thead>
<tr>
<th>Species</th>
<th>Genes</th>
<th>Accession number (protein)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boar</td>
<td>AQN1</td>
<td>ENSSSCP00000019643</td>
</tr>
<tr>
<td>Boar</td>
<td>AQN3</td>
<td>ENSSSCP00000003222</td>
</tr>
<tr>
<td>Boar</td>
<td>FN1</td>
<td>ENSSSCP00000017132</td>
</tr>
<tr>
<td>Bull</td>
<td>CFH</td>
<td>ENSBTAP00000031480</td>
</tr>
<tr>
<td>Bull</td>
<td>NUCB1</td>
<td>ENSBTAP00000003073</td>
</tr>
<tr>
<td>Bull</td>
<td>NUCB2</td>
<td>ENSBTAP00000023221</td>
</tr>
<tr>
<td>Bull</td>
<td>PLA2G7</td>
<td>ENSBTAP00000025719</td>
</tr>
<tr>
<td>Bull</td>
<td>SPADH1</td>
<td>ENSBTAP00000014297</td>
</tr>
<tr>
<td>Bull</td>
<td>SPADH2</td>
<td>ENSBTAP00000010565</td>
</tr>
<tr>
<td>Buck/Ram (human sequence)</td>
<td>TIMP2</td>
<td>ENSP00000262768</td>
</tr>
<tr>
<td>Bull / Ram / Horse</td>
<td>BSP1</td>
<td>ENSBTAP00000051819</td>
</tr>
<tr>
<td>Ram</td>
<td>PGDS</td>
<td>ENSBTAP00000020065</td>
</tr>
<tr>
<td>Camel (human sequence)</td>
<td>PEBP1</td>
<td>ENSP00000261313</td>
</tr>
<tr>
<td>Camel/Alpaca (human sequence)</td>
<td>QSOX1</td>
<td>ENSP00000356574</td>
</tr>
<tr>
<td>Camel/Alpaca</td>
<td>NGF</td>
<td>ENSBTAP00000009796</td>
</tr>
<tr>
<td>Horse</td>
<td>CRISP3</td>
<td>ENSCAP00000005000</td>
</tr>
<tr>
<td>Horse</td>
<td>KKL1E2</td>
<td>ENSECAP0000009529</td>
</tr>
<tr>
<td>Horse</td>
<td>SAL-1</td>
<td>ENSECAP0000000397</td>
</tr>
<tr>
<td>Boar/ epididymis</td>
<td>RNAse10</td>
<td>ENSSSCP0000002332</td>
</tr>
<tr>
<td>Ram/Boar/Bull/ epididymis</td>
<td>MFGE8</td>
<td>ENSBTAP0000004272</td>
</tr>
</tbody>
</table>

Table 1. List of the 20 proteins studied. The proteins were chosen based on their predominance in the seminal plasma from at least one species among nine species of domestic placental mammals, based on their relative abundance after SDS PAGE and Coomassie staining (Druart et al., 2013). RNAse10 and MFGE8 are specific markers of epididymal maturation in ungulates. For species whose genome is not fully sequenced (camel, alpaca, goat, sheep), the human or the bovine sequences of the proteins were used for analyses.
with no ortholog identified in at least one of the species of interest to test the hypothesis that evolution of seminal fluid in mammals is characterised by a gene loss pattern, as previously described in our laboratory (Tian, Pascal, Fouchécourt, et al., 2009; Meslin et al., 2011). The pseudogene status was inferred in a genome only if we found a stop codon or an indel in the sequence identified by the similarity search in the syntenic locus in comparison with the other species of interest.

In order to investigate selective pressure, the PhyLeasProg web server used the CODEML application from the PAML package version 4.4 (Yang, 2007), which allows the ratio \( dN/dS \) to vary across codons and to estimate the probability for each codon to be under positive selection. The alignments were obtained using MUSCLE software (Thompson et al., 1994) and PAL2NAL (Suyama et al., 2006). For the studies of selective pressure, multiple alignments were systematically and carefully examined to avoid false positive results. In particular, amino acids predicted to be under positive selection that were at the boundary of the alignments were not considered. We also eliminated genes for which a signal of positive selection was due to sequence errors in Ensembl according to a comparison with other available sequences from other database such as RefSeq in NCBI.

To evaluate if the intensity of selective pressure varies among sites in the sequences studied, we used Site-Models implemented in PAML (Nielsen and Yang, 1998), which allow the \( \omega \) ratio to vary among sites (Nielsen and Yang, 1998; Yang, 2000). Five models and three comparisons are used in PhyLeasProg: M1a (0 < \( \omega_0 \) < 1 and \( \omega_1 = 1 \)) versus M2a (0 < \( \omega_0 \) < 1, \( \omega_1 = 1 \) and \( \omega_2 > 1 \)) (Wong et al., 2004; Yang et al., 2005), M7 (0 < \( \omega < 1 \)) versus M8 (0 < \( \omega < 1 \) and \( \omega_0 > 1 \)) (Yang, 2000) and M8 versus M8a (0 < \( \omega < 1 \) and \( \omega_0 = 1 \)) (Swanson et al., 2003). LRTs (Likelihood Ratio Tests) were used to compare the log likelihood values (Nielsen and Yang, 1998). Bayes Empirical Bayes (BEB) method (Yang et al., 2005) implemented in PAML was used to estimate posterior probabilities of selection on each codon; probabilities higher than 0.95 were considered significant.

To determine whether some genes in a various species have undergone selective pressure, PhyLeasProg used the branch-site models of PAML (Yang and Nielsen, 2002; Zhang et al., 2005), which estimate different \( dN/dS \) values among branches and among sites. These models allow detection of short episodes of positive selection even if they occur in a small fraction of amino acids. We tested this for all internal and terminal branches. For none of the genes studied did we encounter a sufficient number of paralogs to allow the detection of selective pressures following duplication events. Two models are used to test for positive selection, one model called alternative in which the branch of interest may have a proportion of sites under positive selection, called the foreground branch, and one model called null in which the foreground branch may have different proportions of sites under neutral selection than the background branch. For the alternative model, three classes are defined: \( \omega_0 \): \( dN/dS < 1 \), \( \omega_1 \): \( dN/dS = 1 \) and \( \omega_2 \): \( dN/dS \geq 1 \), while in null model, \( \omega_2 \) is fixed to 1. As for the site model, LRT (Nielsen and Yang, 1998) and BEB (Yang et al., 2005) were used.

Each branch of each phylogenetic tree was tested simultaneously for positive selection. Because we performed multiple-hypothesis tests, we used the \( q \)-value to control the statistical evidence associated with each branch tested. Similar to the \( p \)-value, the \( q \)-value is used to measure the significance in terms of false discovery rate rather than false positive rate. We used the R package QVALUE to compute the \( q \)-values (Storey and Tibshirani, 2003). Positive selection on the foreground branch was considered significant with a threshold of \( q < 10\% \) of false positives. After validation of the branch with the \( q \)-value, we considered only sites with posterior probabilities of Bayes Empirical Bayes analyses superior to 95% or 99% as positively selected. Datasets with less than 10 sequences, the minimum threshold required to obtain significant results, with excessively divergent sequences, or with sequences of genes for which annotations are not reliable were discarded from subsequent analysis.

We tried using an approach described by Pond et al. (2011) to detect sites under positive selection. That approach appears to be statistically sound and more appropriate in the context of our study because it does not force all branches to fit into two rigid classes (‘foreground’ and ‘background’ branches). However, we were unable to easily verify the results (as we did for the results obtained through the method described above), in particular the alignment used in the calculations. We look forward to using a future version of that software (if that becomes available) that will provide users with these alignments.

Modelling of 3D structures

The PSI-BLAST program (Altschul et al., 1997) was used to search for similarities against the non-redundant database. Hydrophobic Cluster Analysis (HCA) (Callebaut et al., 1997) was used to refine the sequence
Rat/Horse KLK2 (Chromosome 10 horse)

69 VRLQGHSLADGDGQ–DVPVRSIPHPYMRSLQMPTFLSPDADNHSNLRLQREPAN 127
  + LG HSL + + G + SI HP YN+ FL+ D LML +L+E
20317007 IGLGLHSL+SNHEEGCMMDANLSIQHPEYNK-----PFLAND-----LMLIKLKESVI

Rat/Mouse KLK2 (Chromosome 7 mouse)

67 NKVRQLQHSLADGDGQ–DVPVRSIPHPYMRSLQMPTFLSPDADNHSNLRLQREPAN 126
  +KVRL QHSL AD D GQ V V SIPH LYN SL+ TFLSPDA +SH+LML QL EPA
4817606 SKVRLDQHLSADEDAQYIVSV*CSIPHSYNMLRKLTF/PSDAGSSHDMLQLSEPA

Human/Cattle TGM4 (Chromosome 22 cattle)

548 YINSLAILDEPVRGFI/AESKEIMASEVFSTFQYEFSIELPNTCRGQL 602
  Y+ L + DD+P+I+GFIIAEI+ES+E+ S+ F SFQY + +E+ + + G L
2075792 YLLGLPMFDDDPIKGFIIAEILESEEMTSQEVFSQYAKLPVEVSH*RQTGLL 2075628

Fig. 1. Identification of marks of pseudogenes. A tBlastn analysis allowed the identification of exons presenting STOP codons within horse and mouse KLK2 pseudogenes and bovine TGM4 pseudogene in corresponding syntenic genomic regions (see Material and Methods). *: STOP codon. The top sequence corresponds to the first species, and the bottom sequence the second species in which the gene is pseudogenised. For each alignment, the upper number corresponds to the amino-acid position and the lower number to the genomic position of nucleotides on the corresponding chromosome.

Fig. 2. Major/diagnostic proteins present in seminal fluids and phylogenetic results showing positive selection or gene loss. Only one parenthesis before Druart. The list of major proteins in seminal fluids of different mammals comes from our recent work (Druart et al., 2013); see Material and Methods section. The positive selection for the KLK2 gene was previously shown (Marques et al., 2012), as well as the loss of the TGM4 gene in cattle, pig and dog (Tian, Pascal, Fouchécourt, et al., 2009), and of the SAL1 gene in human (Meslin et al., 2011). ND: not determined.
Table 2. Parameter estimates and likelihood scores for branch-site evolutionary models. The Ensembl IDs indicated below the gene names (1st column) refers to the protein IDs used for the numbering of amino acids in the column giving results about the positively selected sites. In the 4th column, $r_0$, $r_1$, and $r_2$ are the proportions of codons subject to purifying, neutral, and positive selection, respectively. $w_0$, $w_1$ and $w_2$ represented $dN/dS$ for each class (purifying, neutral and positive selection, respectively). Legend: *, significant at a 0.05 threshold; **, significant at a 0.01 threshold; ***, significant at a 0.001 threshold.

<table>
<thead>
<tr>
<th>Species</th>
<th>Model</th>
<th>$L$ (Log-likelihood values)</th>
<th>Estimates of parameters</th>
<th>$2\Delta$</th>
<th>Positively selected sites (BEB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFGE8</td>
<td>Null</td>
<td>-10397.24038</td>
<td>$\rho_0 = 0.63$, ($\rho_1 = 0.09$), $\omega_0 = 0.10$, ($\omega_1 = 1$)</td>
<td>12.86</td>
<td>Not allowed</td>
</tr>
<tr>
<td></td>
<td>Alternative</td>
<td>-10390.81054</td>
<td>$\rho_0 = 0.62$, $\rho_1 = 0.10$, ($\rho_2 = 0.28$), $\omega_0 = 0.10$, ($\omega_1 = 1$), $\omega_2 = 3.51$</td>
<td>2 sites&gt;99%: 29K, 52S 20 sites p&gt;95%: 21L, 23A, 6L, 37L, 43Q, 278N, 80N, 92R, 94T, 149L, 152H, 192H, 210T, 214T, 259L, 279V, 281G, 285N, 296S, 312S</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Dog (ENSCAFP00000019013)</td>
<td>Alternative</td>
<td>-10416.95177</td>
<td>$\rho_0 = 0.80$, $\rho_1 = 0.16$, ($\rho_2 = 0.084$), $\omega_0 = 0.11$, ($\omega_1 = 1$), $\omega_2 = 998.97$</td>
<td>1 site p&gt;95%: 433S</td>
</tr>
<tr>
<td>RNAse10</td>
<td>Null</td>
<td>-2199.80156</td>
<td>$\rho_0 = 0.54$, ($\rho_1 = 0.37$), $\omega_0 = 0.11$, ($\omega_1 = 1$)</td>
<td>16.38</td>
<td>Not allowed</td>
</tr>
<tr>
<td></td>
<td>Alternative</td>
<td>-2191.60921</td>
<td>$\rho_0 = 0.57$, $\rho_1 = 0.37$, ($\rho_2 = 0.06$), $\omega_0 = 0.12$, ($\omega_1 = 1$), $\omega_2 = 15.59$</td>
<td>2 sites p&gt;95%: 45Q, 126P</td>
<td></td>
</tr>
<tr>
<td>RNAse10</td>
<td>Null</td>
<td>-2200.02319</td>
<td>$\rho_0 = 0.56$, ($\rho_1 = 0.41$), $\omega_0 = 0.11$, ($\omega_1 = 1$)</td>
<td>16.83</td>
<td>Not allowed</td>
</tr>
<tr>
<td></td>
<td>Alternative</td>
<td>-2191.60921</td>
<td>$\rho_0 = 0.57$, $\rho_1 = 0.37$, ($\rho_2 = 0.06$), $\omega_0 = 0.12$, ($\omega_1 = 1$), $\omega_2 = 15.59$</td>
<td>1 site p&gt;95%: 129K</td>
<td></td>
</tr>
<tr>
<td>RNAse10</td>
<td>Null</td>
<td>-2199.97492</td>
<td>$\rho_0 = 0.56$, ($\rho_1 = 0.40$), $\omega_0 = 0.11$, ($\omega_1 = 1$)</td>
<td>9.70</td>
<td>Not allowed</td>
</tr>
<tr>
<td></td>
<td>Alternative</td>
<td>-2195.12434</td>
<td>$\rho_0 = 0.57$, $\rho_1 = 0.38$, ($\rho_2 = 0.04$), $\omega_0 = 0.12$, ($\omega_1 = 1$), $\omega_2 = 5.76$</td>
<td>1 site p&gt;95%: 127G</td>
<td></td>
</tr>
<tr>
<td>CRISP3</td>
<td>Null</td>
<td>-4781.19482</td>
<td>$\rho_0 = 0.50$, ($\rho_1 = 0.41$), $\omega_0 = 0.12$, ($\omega_1 = 1$)</td>
<td>7.62</td>
<td>Not allowed</td>
</tr>
<tr>
<td></td>
<td>Alternative</td>
<td>-4777.38631</td>
<td>$\rho_0 = 0.53$, $\rho_1 = 0.44$, ($\rho_2 = 0.03$), $\omega_0 = 0.12$, ($\omega_1 = 1$), $\omega_2 = 237.23$</td>
<td>1 site p&gt;95%: 168M</td>
<td></td>
</tr>
</tbody>
</table>
### Table 2

<table>
<thead>
<tr>
<th>Species</th>
<th>Model</th>
<th>(L) (Log-likelihood values)</th>
<th>Estimates of parameters</th>
<th>(2\Delta)</th>
<th>Positively selected sites (BEB)</th>
</tr>
</thead>
<tbody>
<tr>
<td>KLK1 Human</td>
<td>Null</td>
<td>-4650.41596</td>
<td>(\rho_0 = 0.31, (\rho_1 = 0.18), \omega_0 = 0.18, (\omega_1 = 1))</td>
<td>9.51</td>
<td><strong>Not allowed</strong></td>
</tr>
<tr>
<td>(ENSP00000301420)</td>
<td>Alternative</td>
<td>-4645.66156</td>
<td>(\rho_0 = 0.60, \rho_1 = 0.35, (\rho_2 = 0.05), \omega_0 = 0.18, (\omega_1 = 1), \omega_2 = 155.94)</td>
<td>1 site p&gt;95%: 244A</td>
<td></td>
</tr>
<tr>
<td>KLK1 Mouse</td>
<td>Null</td>
<td>-4650.77256</td>
<td>(\rho_0 = 0.55, (\rho_1 = 0.32), \omega_0 = 0.17, (\omega_1 = 1))</td>
<td>5.74</td>
<td><strong>Not allowed</strong></td>
</tr>
<tr>
<td>(ENSMUSP00000082577)</td>
<td>Alternative</td>
<td>-4647.90052</td>
<td>(\rho_0 = 0.60, \rho_1 = 0.33, (\rho_2 = 0.06), \omega_0 = 0.18, (\omega_1 = 1), \omega_2 = 8.79)</td>
<td>1 site p&gt;95%: 172Y</td>
<td></td>
</tr>
<tr>
<td>KLK1E2 Horse</td>
<td>Null</td>
<td>-4647.52224</td>
<td>(\rho_0 = 0.50, (\rho_1 = 0.29), \omega_0 = 0.16, (\omega_1 = 1))</td>
<td>19.22</td>
<td>*<strong>Not allowed</strong></td>
</tr>
<tr>
<td>(ENSECAP00000009529)</td>
<td>Alternative</td>
<td>-4637.91026</td>
<td>(\rho_0 = 0.54, \rho_1 = 0.31, (\rho_2 = 0.14), \omega_0 = 0.17, (\omega_1 = 1), \omega_2 = 14.52)</td>
<td>2 sites p&gt;99%: 159L, 190T, 1 site p&gt;95%: 240N</td>
<td></td>
</tr>
<tr>
<td>KLK1 Cattle</td>
<td>Null</td>
<td>-4649.50604</td>
<td>(\rho_0 = 0.50, (\rho_1 = 0.29), \omega_0 = 0.17, (\omega_1 = 1))</td>
<td>6.66</td>
<td><strong>Not allowed</strong></td>
</tr>
<tr>
<td>(ENSBTAP00000024677)</td>
<td>Alternative</td>
<td>-4646.17684</td>
<td>(\rho_0 = 0.58, \rho_1 = 0.34, (\rho_2 = 0.08), \omega_0 = 0.17, (\omega_1 = 1), \omega_2 = 11.30)</td>
<td>2 sites p&gt;95%: 19F, 56A</td>
<td></td>
</tr>
</tbody>
</table>

alignments prior to modelling. Phyre (Bennett-Lovsey et al., 2008) was used for fold recognition and Chimera (Pettersen et al., 2004) for manipulation of three-dimensional structures. MODELLER 9v10 (Martí-Renom et al., 2000) was used for homology modelling.

**Comparative analyses**

To test the hypothesis that protein abundance in the seminal fluid (Table S3) is correlated with the amount of positive selection on the genes encoding them (Table S4), we performed a modified version of the pairwise comparison test (Maddison, 2000). It was not possible to use the more popular phylogenetic independent contrasts (FIC) (Felsenstein, 1985) because preliminary tests onto our timetree performed using the PDAP-PDTREE module (Midford et al., 2008) for Mesquite (Maddison and Maddison, 2015) yielded very highly significant artefacts, which indicated that these data do not follow a Brownian motion evolutionary model on this tree. This is not surprising because some of our data are categorical, and others are meristic, and the Brownian model (and FIC) was designed for continuous data. This results partly from the fact that the available abundance data are not truly quantitative because they were obtained by observing electrophoresis gels. Thus, they were scored as a discrete character (1 for absence or low abundance; 2 for moderate abundance; 3 for high abundance). Thus, the pairwise comparison test, designed for discrete data, is more appropriate. To maximise power, we had to binaryize the data because Mesquite’s pair selector contrasting pairs varying in both characters simultaneously.
(the only informative pairs) only works with binary data. This was not problematic because very few taxa displayed variation, so the loss of information was slight. Thus, for abundance, out of 20 characters for which we have data (S table 1), only three (BSPI, NUCB1, NUCB2) display more than two states, and for intensity of positive selection, out of the same 20 characters, only three (KLK1, RNAse10, and MFGE8) display more than two states. We also checked that the test performed more poorly when working on the original data (not binarised) with the pair selector that draws the highest number of pairs without considering character state distribution. Note that neither way of selecting pairs depends on our hypotheses about how characters may be correlated to each other. However, with only 9 terminal taxa, no single set of pairs of taxa could possibly yield significant results because at most 4 pairs could be drawn (if character distribution were optimal for this test), and the lowest probability that the test could yield is $2^{-4}=0.0625$, just above the statistical significance threshold before corrections for multiple tests are applied. It would be even more difficult to get positive results after such corrections; this would require even more data.

In our case, given that the null hypothesis to be tested is the same for all pairs of characters (that for each gene, there is no correlation between protein abundance in the seminal fluid and the amount of positive selection in the gene), a solution to increase power is to compile the number of pairs suggesting positive and negative correlation (separately). A simple binomial test can then be performed (we used the GraphPad software available at http://www.graphpad.com/quickcalcs/binomial1/) to test the hypothesis that at least that many positive or negative results (whichever is greater) can be obtained by chance alone if there is no correlation. In our case, the alternative hypothesis is that the relationship between positive selection and abundance is positive because both should relate to the functional importance of the gene and its protein, so a one-tailed test is appropriate. The number of tests is simply the number of pairs showing positive or negative covariation, and the probability of each result is 0.5. We suggest calling this test (which is new, as far as we know) a ‘multivariate phylogenetic pairwise comparison test’. This is a simple multivariate extension of Maddison’s pairwise comparison test (Maddison, 2000), but we add the ‘phylogenetic’ to the name because non-phylogenetic pairwise comparisons also exist in the literature (Stoline, 1981; Rees et al., 2005). Note that this test should be used only if the same null hypothesis can apply to all pairs of characters to be tested. It should not be used if there were a priori reasons to believe that only some characters were correlated to each other, or if the direction of the correlation were expected to vary between pairs of characters.

We also used Pagel’s maximum likelihood-based test to assess the same correlations, in the hope that it would have more power (Pagel, 1994). However, that test only works on binary data. To assess statistical significance, we performed 100 simulations for each analysis.

We also tested the null hypothesis that sites under positive selection are randomly distributed on the protein surface. The alternative is that more (or fewer) are exposed to the solvent than predicted by chance. For each protein, we calculated the probability of observing, among the amino acids under positive selection of known position (Table S5; for some, the position is currently unknown and these were not considered in the calculations), at least as many amino acids exposed to the solvent. For this, we determined the proportion of amino acids (under selection or not) exposed to the solvent (Table S6); this gives the probability that each amino acid under positive selection is exposed to the solvent, under the null hypothesis. This was done using two tests: first, a binomial distribution, in the program GraphPad http://www.graphpad.com/quickcalcs/binomial1/; second, Fisher’s exact test, in Statistica 6 (by StatSoft France). The global test of the null hypothesis is the product of probabilities of the individual tests (each of which bears on a single protein in a single species).

We have established the minimal rate of pseudogenisation in our dataset by dividing the number of events, inferred through maximum parsimony while considering pseudogenisation to be irreversible (Table S7), on the tree by Faith’s (Faith, 1992) phylogenetic diversity index (the sum of branch lengths, where the lengths represent evolutionary time), which was calculated using the StratigraphicTools for Mesquite (Josse et al., 2006). Note that given that our sample includes only 9 species out of the more than 5000 species of mammals, we probably under-estimated the number of transitions in our dataset because a more intensive taxonomic sampling would have probably shown that some of the events that we consider to be synapomorphic of a large clade, may well be convergent. For instance, pseudogenisation of ENS/AQN/SPADH is here considered a synapomorphy of Euarchontoglires (because it is a pseudogene in all sampled euarchontog-
laries) and this occurred convergently in the dog. However, if other euarchontoglires retained this gene, this would indicate that more than one pseudogenisation of this gene occurred in euarchontoglires.

Results

Identification of pseudogenes

Our present search for pseudogenes showed that the KLK2 gene, which is under positive selection in some primates, has also been lost in cattle, horse and mouse (examples of traces of pseudogenes in Fig. 1). Genes encoding BSPH1 and -2 have also been lost in human, chimpanzee and dog. Figure 2 recapitulates all the events of gene loss across species (right).

Inference of positive selection

We studied the evolution of the proteins identified by proteomic/orbitrap analysis in seminal fluid from domestic animals as previously described in human (Ba-truch et al., 2011; Milardi et al., 2012). We chose the proteins that are also shown to be highly expressed in the seminal fluid of at least one of the sampled domestic animal species (bull, ram, billy goat, boar, stallion and rabbit); see Methods section and Druart et al. (2013).

Table 3. Position of the amino acids under positive selection in the 3D structure models. Positions on the 3D structures were assessed by computing the accessible solvent area (ASA) and through visual inspection.

<table>
<thead>
<tr>
<th>Proteins</th>
<th>Species</th>
<th>Template (pdb code, sequence identity)</th>
<th>Position on the 3D structure models</th>
<th>Other information, remarks</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFGE8</td>
<td>human</td>
<td>1F7E (41 %)</td>
<td>EGF 26I, 29K, 37L, 43Q, 52S</td>
<td>See text (43Q)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1SDD (29 %)</td>
<td>C1 domain : N78, N80</td>
<td>C1/C2 interface</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>C1 domain : 92R, 94T, 149L, 152H, 192H, 210T, 214T</td>
<td>See text (92R, 94T, 149L, 152H, 214T)</td>
</tr>
<tr>
<td>dog</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CRISP3</td>
<td>rabbit</td>
<td>1RC9 (53 %)</td>
<td>168M</td>
<td></td>
</tr>
<tr>
<td>KLK1</td>
<td>human</td>
<td>1SPJ (100 %)</td>
<td></td>
<td>244A</td>
</tr>
<tr>
<td></td>
<td></td>
<td>1SGF (62 %) in complex with NGF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mouse</td>
<td></td>
<td>1SGF (100 %)</td>
<td>172Y</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2GVZ (50 %)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>KLK1</td>
<td>horse</td>
<td>2GVZ (100%)</td>
<td>159L, 200T, 240N</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1SPJ (56 %)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1SGF (52 %) in complex with NGF</td>
<td></td>
<td></td>
</tr>
<tr>
<td>cattle</td>
<td></td>
<td>1SPJ (71 %)</td>
<td>56A</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>1SGF (61 %) in complex with NGF</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>2GVZ (58 %)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
None of the studied genes exhibited positive selection on site, but two showed significant positive selection on branch-site (Table 2; Fig. 2): KLK1 in human, cattle, mouse and horse (KLK1E2 in horse is one of the three co-orthologs of human KLK1), and CRISP3 in rabbit. Both genes highly expressed in (and markers of) epididymis are also under positive selection: RNase10 in the stem of Fereungulata (which includes carnivorans, perissodactyls and artiodactyls) and in the stem of the sampled artiodactyls, and MFGE8 in a stem-hominid, and in dog (Fig. 2). It was not possible to determine without ambiguity if genes encoding proteins of the SPADH/AQN family and of the BSP family evolved under positive selection, due to the particularly high divergence of protein sequences, impairing accurate and unambiguous alignment.

**Position of amino acids under positive selection in the structure of the proteins**

The 3D structures of the proteins (or protein domains) were modelled when their sequences could be significantly related to, and aligned with those of proteins with known 3D structures. Sequence identities range be-
between 29 and 94 %, giving rise to models with accuracy at least comparable to low-resolution experimental structures. This qualitative analysis showed that most of the positively selected positions (40 out of 42 for which 3D structure information is available) are located in regions exposed to solvent, suggesting site-specific selective pressures reflecting the functional context, rather than a structural constraint (Table 3).

For some genes (KLK1 and 2, MFG8), our tests clearly reject the null hypothesis that the amino acids under positive selection are randomly located in the proteins; far more appear to be exposed to the solvent than expected under the null hypothesis (Table S8). For RNAse10, the null hypothesis could not be rejected, although this probably reflects lack of power linked to the small amount of data for this gene. For CRISP3, the probability could not be computed at all. Globally, our results clearly suggest that more amino acids under positive selection are exposed to the solvent than expected by chance alone, although a global probability could not be computed because one of the individual tests yielded a probability not distinguishable from zero (hence, this probability could not be multiplied with the others). Binomial and Fisher’s exact tests gave very similar results.

For illustration purpose, we present here the inferred 3D structures for MFG8 and KLK, as they have multiple positions under positive selection that might be of functional importance for the binding properties of MFG8 and the protease activity of KLK.

According to domain databases, MFG8, also known as lactadherin, contains two EGF-like domains, followed by a tandem of discoidin/F5/8 type C domains (C1 and C2). The positively selected positions of the EGF-like domains are exposed to solvent, without clustering in a particular region of the surface exposed to the solvent. The Q43 position, in the vicinity of the integrin-binding RGD motif, is included in a large loop, within the second EGF-like domain. The two Discoidin/F5/8C domains bind to anionic phospholipids of cellular membranes (Raymond et al., 2009). We know several 3D structures of F5/8 type C domains of lactadherins (bovine – pdb 2pqq, 3bn6) or related proteins (bovine factor Va - pdb 1sdd, human factor VIII – pdb 3ezd, 2rte, human neuropilin – pdb 2qqj, 2qqk, 2qqm, 2qqo, 2orx,...). We selected a template in which the two
domains are present in tandem (rather than the isolated F5/8 type C domains of lactadherin), in order to get information on domain interface. The chosen template, bovine factor Va - pdb 1issd (Adams et al., 2004), has the best score according to the Phyre fold recognition program (E-value 1.6 $10^{-16}$), and shares 29 % amino acid sequence identity with human MFGE8. The alignment was manually refined (Fig. 3A) and the positions of the positively selected sites were reported on the obtained 3D model of the human MFGE8 C1 and C2 domains (Fig. 3B). Interestingly, several sites under positive selection (92R, 94T, 149L, 214T, 259L, 279V, 281G, 285N, 312S) are located within or in the vicinity of the three β-hairpin loops (referred to as ‘spikes’) of the two F5/8 type C domains, which are aligned in an edge-to-edge configuration. These spikes form pockets that are thought to allow interaction with phospholipids and membrane interaction (Shao et al., 2008).

Domain databases indicate that KLK has the typical fold of chymotrypsin-like proteases, consisting of two beta-barrels, which form a cleft in which the catalytic triad is formed. We considered the experimental 3D structure of horse KLK1E2 (pdb (1gvz; Carvalho et al., 2002)) for mapping the positions of amino acids under positive selection in KLK1, KLK1E2 and KLK2 from different species (Fig. 4). Our results suggest that five out of the six amino acids of KLK are located on the protein surface, one in the first beta-barrel domain (KLK2 H56, the amino acid equivalent to A56 in bovine KLK2) and four in the second one (KLK2 L159, T200 and N240, as well as Q172, the amino acid equivalent to Y172 in mouse KLK1). None of these are involved in the active site, or in the kallikrein loop (in orange in Fig. 3), which has a direct role in the control and selectivity of the enzyme activity. None of these residues (or equivalent ones) are likely to be involved in ligand binding, when considering the structure of mouse KLK1 in complex with NGF (Bax et al., 1997; data not shown). The remaining sixth amino acid under positive selection (KLK2 F243, the amino acid equivalent to A244 in KLK1 one) contributes to the hydrophobic core of the second beta-barrel, and is located within a strand forming a wall of the enzymatic pocket.

### Relationship between abundance in seminal fluid and positive selection

Only three genes (CRISP3, KLK1, and MFGE8) exhibit variation in both abundance of their product protein in the seminal fluid and in the presence of positive selection. Thus, when contrasting pairs of characters for which both vary within pairs (which is only possible on binarised data, in the Mesquite implementation), only one positive and two negative pairs were found, which is consistent with a random association (p = 0.5). When using the ‘most pairs’ selector in Mesquite (which draws the highest number of pairs, irrespective of character state), only one positive and two neutral pairs were found, which is also not significant (p = 1). Pagel’s test yielded lower probabilities (Pagel, 1994), with the lowest being for KLK1, but this is not significant (p = 0.14; S 1, sheets ‘Pagel 1994 test’ and ‘Abundance, selection correl.’). The lack of correlation is also confirmed by a visual inspection of the evolutionary changes implied by the data, as this can best be done from mirror trees. For instance, for KLK 1 (Fig. 5), positive selection is found (in increasing number of amino acids) in the mouse, cow, and horse, but of these three taxa, only the horse has increased abundance of the gene product in its semen (in all other taxa in our sample, it is absent or in low abundance). In MFGE8 (Fig. 6), the dog (one amino acid) and humans (22 amino acids) exhibit positive selection, but only the cow has an abundance of the gene product in its semen.

### Rate of pseudogenisation

The sampled phylogenetic biodiversity is 613 Ma. Our data imply minimally 6 pseudogenisation events, which gives a global rate (for the 20 considered genes) of about 0.0098 events/lineage/Ma, or a rate of about 0.00048 pseudogenisations/lineage/gene/Ma.

### Discussion

**Mode of evolution and protein abundance in seminal fluid**

Data from several previous studies that have identified the most abundant proteins in the seminal fluid of domestic animals allow testing hypotheses about the evolution of relevant genes. Our recent study that showed a particularly high proteome diversity of seminal fluid between species suggested this diversity was potentially associated with attributes of male reproductive physiology (Druart et al., 2013). Our negative results concerning the possible correlation between the abundance of proteins in the seminal fluid and the
presence of positive selection in the gene encoding it in the same species (obtained by multivariate phylogenetic pairwise comparisons) should be viewed with caution because of the low power of our test, itself resulting from the low number of genes, taxa, and the limited variability of the relevant characters in our dataset. Nevertheless, our results do not lend any support to the hypothesis that both characters are positively correlated. The apparent absence of correlation between the predominance of a protein in seminal fluid in one species and its evolution under positive selection, which is confirmed by visual inspection of the data (Figs 5-6), is compatible with the 'translational robustness hypothesis' proposed before (Drummond et al., 2005). According to this hypothesis, genes with high expression evolve slowly, which avoids protein misfolding.

**Diversification of proteins in seminal fluid**

The present work suggests that the high diversity of proteins present in seminal fluid of mammals is associated with a species-specific evolutionary pattern of the corresponding genes by fairly frequent pseudogenisation, high expression diversity, and positive selection. Pseudogenisation has been previously demonstrated for TGM4 and semenogelin genes in some ape species (Jensen-Seaman and Li, 2003). We also have previously shown that TGM4 has also been lost in cattle, horse, dog, and likely several other mammalian species (Tian, Pascal, Fouchécourt, et al., 2009) and that the ortholog of porcine Sal1 and Major allergen Equine C1 Precursor has also been lost in human as well as in the Neanderthal genome (Meslin et al., 2011). It is difficult to determine if the pseudogenisa-

Fig. 5. Lack of correlation between abundance (left) and number of amino acids under positive selection in KLK1. Parsimony optimizations performed in Mesquite (Maddison and Maddison, 2015).
tion rate of 0.00048 events/lineage/gene/Ma is especially high because such rates have seldom been reported in the literature. We reported before, from a different set of 69 genes and a different taxonomic sample, rates ranging from 0 (in teleosts) to 0.016 (in eutherians) (Meslin et al., 2012). The latter value, to be meaningfully compared with our rates, has to be converted into a rate per gene, which gives about 0.00023 events/lineage/gene/Ma for eutherians. Given that our sample is also composed of eutherians, the 20 genes studied here appear to have undergone more pseudogenisation than most of the 69 genes studied previously (Meslin et al., 2012).

A few examples illustrate how this diversity appeared. The gene encoding KLK2, a kallikrein expressed in the prostate in humans, was previously shown to be lost in several primates (Gorilla gorilla Savage 1847, Papio anubis Lesson 1827, (Marques et al., 2012)), and under positive selection in others, as were two other genes encoding the proteases ACPP and TGM4 (Clark and Swanson, 2005). In the present study, we found that KLK2 has been lost in cattle, horse, and mouse (i.e. we found traces of pseudogenes), probably independently because close relatives of these taxa retain this gene. The situation of the TGM4 gene is different. It is present in birds, squamates, platypus, several primates and at least three rodents, but is absent in all sampled laurasiatherians. Thus, it may have been lost before the appearance of Laurasiatheria. Interestingly KLK1, a paralog of KLK2, a major protein found in equine seminal fluid (named KLK1E2), is under positive selection in cattle, horse, mouse as well as in human. KLK1 seems to be mainly expressed in the kidney, the pancreas and the salivary glands in the mouse (http://www.ncbi.nlm.nih.gov/UniGene/clust.cgi?ORG=Hs&CID=123107&MAXE
ST=94), so some investigations are needed to confirm its presence in the seminal fluid of other species (except horse). Nevertheless, this suggests that at least in horse, KLK1 may replace KLK2 for an important biological function in the seminal fluid.

**Position and function of amino acids under positive selection**

Our data suggest that amino acids under positive selection are more often exposed to the solvent than expected by chance. However, this result should be viewed with caution because it is based on a relatively low number of amino acids, and this result reflects the data of only some of the sampled proteins; for others, we could not reject the null hypothesis. Some studies have led to similar conclusions, on an ad hoc basis, but without addressing this issue at a large scale. For instance, amino acids under positive selection were identified in Toll-like receptor (Fornûsková et al., 2013), which are likely involved in species-specific recognition of lipopolysaccharide of gram-negative bacteria. In the particular case of our medium-scale study, more data on the position of amino acids and of associated 3D structures will need to be gathered to reach firm conclusions on this point. Our new results about the position of amino acids suggest that positive selection affected preferentially amino acids involved in interactions with partners rather than other functions of the proteins, as most such amino acids are located at the surface rather than in the vicinity of an eventual enzymatic pocket.

Ultimately, we have confirmed here that MFGE8 was under positive selection in the dog and in the human/chimpanzee clade. This protein binds to the zona pellucida of unfertilised (but not fertilised) oocytes, because recombinant protein or specific antibody raised against MFGE8 competitively inhibit sperm-egg interaction. For this protein, positioning the positively selected amino acids on a 3D model was particularly informative. In particular, ten sites under positive selection (92R, 94T, 149L, 152H, 214T, 259L, 279V, 281G, 285N, 312S) are located within or in the vicinity of the three β-hairpin loops also called ‘spikes’, which allow interaction with phospholipids and between membranes (Rodrigues et al., 2013). Interestingly, among the whole family of F5/8 type C domains, there is a particularly high variability in the domain interfaces. One can then hypothesise that the position of the amino acids under positive selection on a same platform, displayed by the alignment of the spikes, provides support to the hypothesis that both F5/8 type C domains participate in a species-dependent function of MFGE8. More generally and as observed in our previous work on the evolution of genes encoding Odorant Binding Proteins and proteins involved in gamete fertilisation, amino acids under positive selection are located almost always at the surface of the proteins rather than in the vicinity of the enzymatic pocket or other functional domain (Meslin et al., 2011, 2012). This suggests that this evolution is driven by species-dependent interaction with partners, as described for example for the positively selected sites on the surface glycoprotein (G) of infectious hematopoietic necrosis virus (LaPatra et al., 2008).

**Conclusion**

In conclusion we suggest that the high diversity of proteomes of mammalian seminal fluids is associated with a particular evolutionary process that includes positive selection and gene loss in various species. Some genes such as those encoding for Kallikreins (identified by others and in the present study) are under positive selection in at least one species (*Homo sapiens*) and have been lost in other taxa (*Gorilla gorilla, Papio anubis* for KLK2, *Bos taurus* for TGM4); some genes encode the prevailing proteins of the seminal fluid of some species but have been lost in other taxa (*Homo sapiens* for SAL1) or are under positive selection in some species (CRISP3). Among all the genes studied however, KLK1 seems to be under positive selection in the greatest number of species, and it has not been lost in the sampled species. Overall, this diversity in expression and this rapid evolution may contribute to the diversity of mating systems and may explain part of the loss in interspecific fecundity. The link between mating system (polyandrous, polygynous, or nomogamous) and the evolution of seminal proteins and their genes in placental mammals, and the potential impact of domestication by human on the evolution of these molecules are unclear. The different mating systems might exert slightly different selective pressures on seminal proteins and thus, contribute to their diversity. It is also possible that domesticated breeds have been reproductively isolated from their wild conspecifics and have thus displayed their own evolutionary dynamics; this mechanism may also have contributed to diversity of seminal proteins in the sampled taxa. These problems could be investigated by systematically studying the evolution of the genes coding for semen proteins in several ancestral and domesticated
animals from *Bos, Sus* and other species. However, testing this hypothesis would require far more data and is beyond the scope of our study.

In any case, our results show conclusively that in proteins of the seminal fluid, amino acids under positive selection appear to be located mostly at the surface of the protein and may suggest a role in gamete interaction. However, we found no evidence of a link between the intensity of positive selection and protein abundance in the seminal fluid.

**Acknowledgements**

We are grateful to Marie-Claire Orgebin-Crist for the helpful discussion and for their critical reading of an earlier version of the draft. We thank Gilles Didier and Manuela Royer-Carenzi for advice on probability calculations, anonymous reviewers for comments that improved the draft, and the Contributions to Zoology editorial staff for its efficient handling of the paper.

**References**


Online supplementary material

S1. Mesquite Nexus file including the reference tree and the data that were analysed by multiple phylogenetic pairwise comparisons (protein abundance data and presence of positive selection), as well as presence of genes.

S2. Reference tree in pdf format.

S3. Abundance of each protein in the seminal fluid.


S5. Position of sites under positive selection.

S6. Proportion of sites exposed in each gene or protein.

S7. Presence of genes.

S8. Test of the hypothesis according to which aminoacids under positive selection are randomly distributed. Legends Supplemental data.