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

Camille Meslin, Michel Laurin, Isabelle Callebaut, Xavier Druart, Philippe Monget

Contributions to Zoology, 84 (3) 217-235 (2015)

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 Druart [1, 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.

Contents

Introduction 217
Material and methods 218
Phylogenetic and syntenic analyses 218
Identification of pseudogenes - Inference of positive selection 219
Modelling of 3D structures 220
Comparative analyses 223
Results 225
Identification of pseudogenes 225
Inference of positive selection 225
Position of amino acids under positive selection in the structure of the proteins 226
Relationship between abundance in seminal fluid and positive selection 228
Rate of pseudogenisation 228
Discussion 228
Mode of evolution and protein abundance in seminal fluid 228
Diversification of proteins in seminal fluid 229
Position and function of amino acids under positive selection 231
Conclusion 231
Acknowledgements 232
References 232

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, Fouchecourt *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

Species	Genes	Accession number (protein)
Boar	AQN1	ENSSSCP00000019643
Boar	AQN3	ENSSSCP00000003222
Boar	FN1	ENSSSCP00000017132
Bull	CFH	ENSBTAP00000031480
Bull	NUCB1	ENSBTAP00000003073
Bull	NUCB2	ENSBTAP00000023221
Bull	PLA2G7	ENSBTAP00000025719
Bull	SPADH1	ENSBTAP00000014297
Bull	SPADH2	ENSBTAP00000010565
Buck/Ram (human sequence)	TIMP2	ENSP00000262768
Bull / Ram / Horse	BSP1	ENSBTAP00000051819
Ram	PGDS	ENSBTAP00000020065
Camel (human sequence)	PEBP1	ENSP00000261313
Camel/Alpaca (human sequence)	QSOX1	ENSP00000356574
Camel/Alpaca	NGF	ENSBTAP00000009796
Horse	CRISP3	ENSCAP00000005000
Horse	KLK1E2	ENSECAP00000009529
Horse	SAL-1	ENSECAP00000000397
Boar/ epididymis	RNAse10	ENSSSCP00000002332
Ram/Boar/Bull/ epididymis	MFGE8	ENSBTAP00000004272

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 ω 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 $\omega_S >1$) (Yang, 2000) and M8 *versus* M8a ($0< \omega <1$ and $\omega_S =1$) (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: ω0: dN/dS<1, ω1: dN/dS=1 and ω2: dN/dS≥1, while in null model, ω2 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         VRLGQHSLDADGDTGQ-DVPVRHSIPHPLYNRSLQMPTFLSPDADNSHNLMLRQLREPAN  127
           + LG HSL  + + G   +    SI HP YN+      FL+ D      LML +L+E
20317007   IGLGLHSL*DNHEEGSCMMDANLSIQHPEYNK-----PFLAND------LMLIKLKESVI
```

Rat/Mouse KLK2 (Chromosome 7 mouse)

```
67        NKVRLGQHSLDADGDTGQDVPVRHSIPHPLYNRSLQMPTFLSPDADNSHNLMLRQLREPA  126
          +KVRL QHSL AD D GQ V V  SIPH LYN SL+  TFLSPDA +SH+LML QL EPA
4817606   SKVRLDQHSLSADEDAGQYVSV*CSIPHSLYNMSLRKLTFLSPDAGSSHDLMLLQLSEPA
```

Human/Cattle TGM4 (Chromosome 22 cattle)

```
548       YINSLAILDDEPVIRGFIIAEIVESKEIMASEVFTSFQYPEFSIELPNTGRIGQL  602
          Y+  L + DD+P+I+GFIIAEI+ES+E+  S+ F SFQY +  +E+ +  + G L
2075792   YLLGLPMFDDDPIIKGFIIAEILESEEMTTSQEFVSFQYAKLPVEVSH*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.

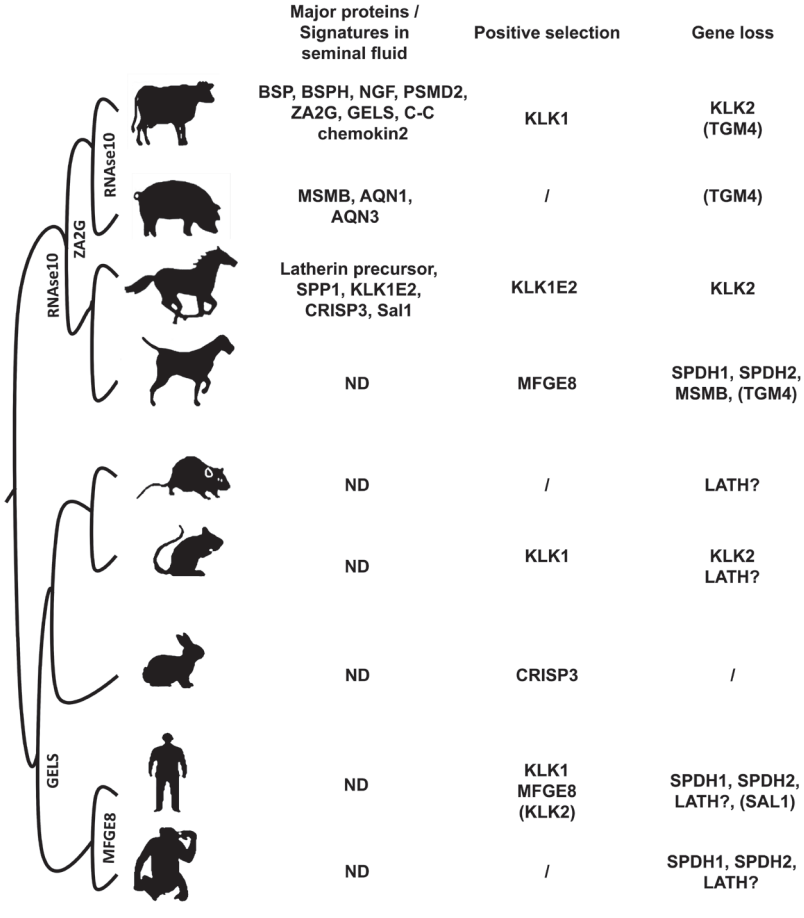

Fig. 2. Major/diagnostic proteins present in seminal fluids and phylogenetic results showing positive selection or gene loss. Only one parenthesis before Druart. The list of major proteins in seminal fluids of different mammals comes from our recent work (Druart *et al.*, 2013); see Material and Methods 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, ρ0, ρ1, and ρ2 are the proportions of codons subject to purifying, neutral, and positive selection, respectively. ω0, ω1 and ω2 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.

Species	Model	*L* (Log-likelihood values)	Estimates of parameters	2Δ*l*	Positively selected sites (BEB)
MFGE8 Stem of Hominidae (ENSP00000268150)	Null	-10397.24038	ρ0 = 0.63, (ρ1= 0.09), ω0 = 0.10, (ω1 = 1)	12.86 **	Not allowed
	Alternative	-10390.81054	ρ0 = 0.62, ρ1 = 0.10, (ρ2 = 0.28), ω0 = 0.10, (ω 1 = 1), ω2 = 3.51		2 sites>99%: 29K, 52S 20 sites p>95%: 21L, 23A, 6I, 37L, 43Q, 278N, 80N, 92R, 94T, 149L, 152H, 192H, 210T, 214T, 259L, 279V, 281G, 285N, 296S, 312S
MFGE8 Dog (ENSCAFP00000019013)	Null	-10425.96658	ρ0 = 0.76, (ρ1= 0.16), ω0 = 0.11, (ω1 = 1)	18.03 ***	Not allowed
	Alternative	-10416.95177	ρ0 = 0.80, ρ1 = 0.16, (ρ2 = 0.084), ω0 = 0.11, (ω1 = 1), ω2 = 998.97		1 site p>95%: 433S
RNAse10 Stem of Artiodactyla (ENSSSCP00000002332)	Null	-2199.80156	ρ0 = 0.54, (ρ1=0.37), ω0 = 0.11, (ω1 = 1)	16.38 ***	Not allowed
	Alternative	-2191.60921	ρ0 = 0.57, ρ1 = 0.37, (ρ2 = 0.06), ω0 = 0.12, (ω1 = 1), ω2 = 15.59		2 sites p>95%: 45Q, 126P
RNAse10 Stem of Euungulata (ENSECAP00000001027)	Null	-2200.02319	ρ0 = 0.56, (ρ1=0.41), ω0 = 0.11, (ω1 = 1)	16.83 ***	Not allowed
	Alternative	-2191.60921	ρ0 = 0.57, ρ1 = 0.37, (ρ2 = 0.06), ω0 = 0.12, (ω1 = 1), ω2 = 15.59		1 site p>95%: 129K
RNAse10 Stem of Fereuungulata Dog/Horse/Artiodactyles ancestor (ENSCAFP00000008085)	Null	-2199.97492	ρ0 = 0.56, (ρ1= 0.40), ω0 = 0.11, (ω1 = 1)	9.70 **	Not allowed
	Alternative	-2195.12434	ρ0 = 0.57, ρ1 = 0.38, (ρ2 = 0.04), ω0 = 0.12, (ω1 = 1), ω2 = 5.76		1 site p>95%: 127G
CRISP3 Rabbit (ENSOCUP00000001631)	Null	-4781.19482	ρ0 = 0.50, (ρ1=0.41), ω0 = 0.12, (ω1 = 1)	7.62 **	Not allowed
	Alternative	-4777.38631	ρ0 = 0.53, ρ1 = 0.44, (ρ2 = 0.03), ω0 = 0.12, (ω1 = 1), ω2=237.23		1 site p>95%: 168M

cont. Table 2

Species	Model	*L* (Log-likelihood values)	Estimates of parameters	2Δ*l*	Positively selected sites (BEB)
KLK1 Human (ENSP00000301420)	Null	-4650.41596	ρ0 = 0.31, (ρ1=0.18), ω0 = 0.18, (ω1 = 1)	9.51 **	Not allowed
	Alternative	-4645.66156	ρ0 = 0.60, ρ1 = 0.35, (ρ2 = 0.05), ω0 = 0.18, (ω1 = 1), ω2=155.94		1 site p>95%: 244A
KLK1 Mouse (ENSMUSP00000082577)	Null	-4650.77256	ρ0 = 0.55, (ρ1=0.32), ω0 = 0.17, (ω1 = 1)	5.74 *	Not allowed
	Alternative	-4647.90052	ρ0 = 0.60, ρ1 = 0.33, (ρ2 = 0.06), ω0 = 0.18, (ω1 = 1), ω2 =8.79		1 site p>95%: 172Y
KLK1E2 Horse (ENSECAP00000009529)	Null	-4647.52224	ρ0 = 0.50, (ρ1=0.29), ω0 = 0.16, (ω1 = 1)	19.22 ***	Not allowed
	Alternative	-4637.91026	ρ0 = 0.54, ρ1 = 0.31, (ρ2 = 0.14), ω0 = 0.17, (ω1 = 1), ω2 =14.52		2 sites p>99%: 159L, 190T 1 site p>95%: 240N
KLK1 Cattle (ENSBTAP00000024677)	Null	-4649.50604	ρ0 = 0.50, (ρ1= 0.29), ω0 = 0.17, (ω1 = 1)	6.66 **	Not allowed
	Alternative	-4646.17684	ρ0 = 0.58, ρ1 = 0.34, (ρ2 = 0.08), ω0 = 0.17, (ω1 = 1), ω2 = 11.30		2 site p>95%: 19F, 56A

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 binarise 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 (BSP1, 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-

lires) 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 (Batruch *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.

Proteins	Species	Template (pdb code, sequence identity)	Position on the 3D structure models		Other information, remarks
			Exposed	Buried	
MFGE8	human	1F7E (41 %)	EGF 26I, 29K,37L, 43Q,52S		See text (43Q)
		1SDD (29 %)	C1 domain : N78, N80		C1/C2 interface
			C1 domain : 92R, 94T, 149L, 152H, 192H, 210T, 214T		See text (92R, 94T, 149L, 152H, 214T)
			C2 domain : 259L, 279V, 281G, 285N, 296S, 312S		See text (259L, 279V, 281G, 285N, 312S)
	dog		C2 domain : 433S		
CRISP3	rabbit	1RC9 (53 %)		168M	
KLK1	human	1SPJ (100 %) 1SGF (62 %) in complex with NGF 2GVZ (56 %)		244A	See text
	mouse	1SGF (100 %) 2GVZ (50 %)	172Y		
KLK1	horse	2GVZ (100%) 1SPJ (56 %) 1SGF (52 %) in complex with NGF	159L, 200T, 240N		See text
	cattle	1SPJ (71 %) 1SGF (61 %) in complex with NGF 2GVZ (58 %)	56A		

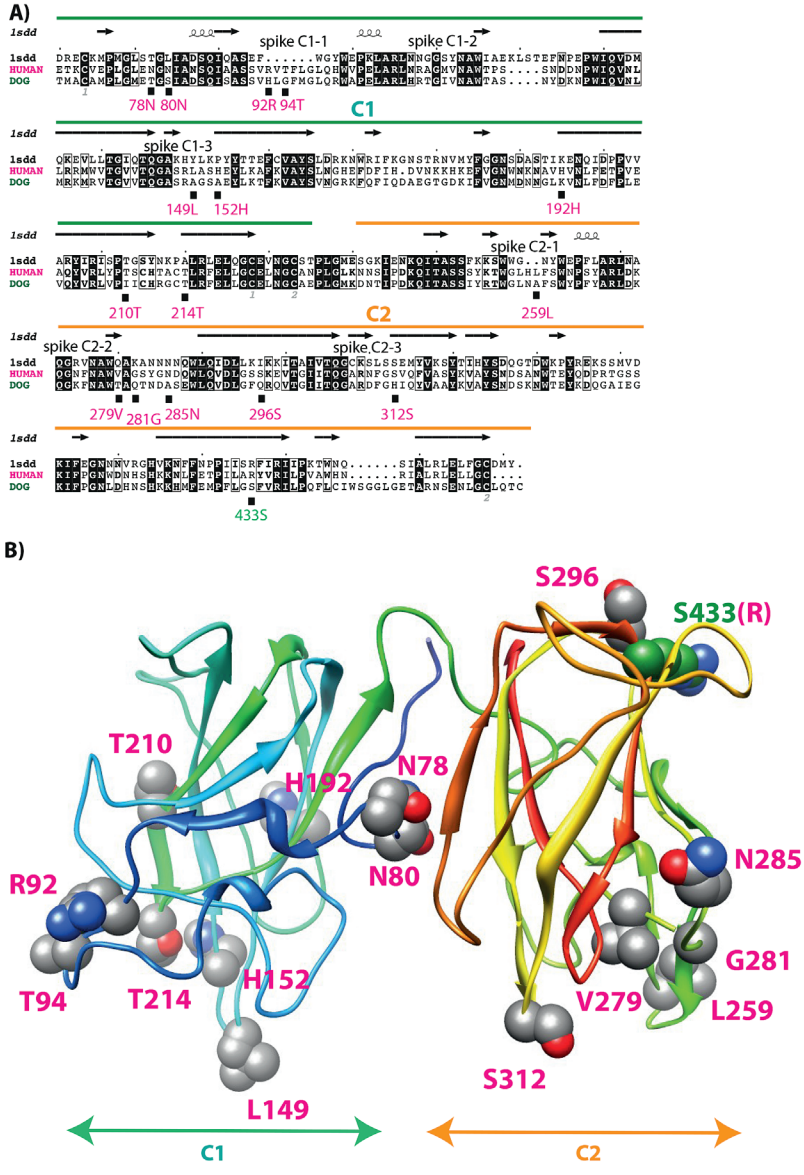

Fig. 3. Positively selected positions within the MFGE8 F5/8 type C domains. A) Human and dog MFGE8 sequences are aligned with that of bovine factor Va, whose 3D structure (pdb 1sdd) serves as a template for modelling. The secondary structures, as observed from the 3D structure of bovine factor Va, are reported on top, together with the position of the two domains (C1 and C2). The positions of positively selected amino acids are reported at the bottom. B) Ribbon representation of a 3D structure model of human MFGE8, constructed using bovine factor Va as a template and after the alignment shown in panel A. Secondary structures are rainbow coloured and the positively selected amino acids are shown with atomic details.

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 Fereuungulata (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-

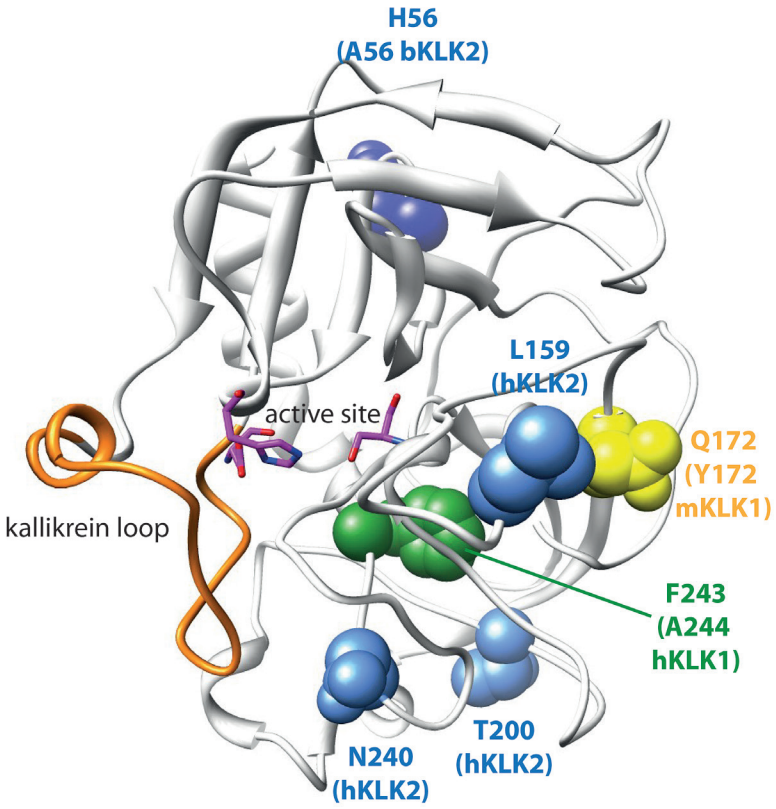

Fig. 4. Positively selected positions within KLK1 and KLK2. The positions under positive selection from different species are reported on the experimental 3D structure of horse KLK1E2, used as reference (pdb 1gvz). Amino acids under positive selection in horse KLK1E2 (hKLK1E2) are in light blue, in bovine KLK2 (bKLK2) in dark blue, in human KLK1 (hKLK1) in green, and in mouse KLK1 (mKLK1) in yellow.

tween 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, MFGE8), 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 MFGE8 and KLK, as they have multiple positions under positive selection that might be of functional importance for the binding properties of MFGE8 and the protease activity of KLK.

According to domain databases, MFGE8, 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 2pqs, 3bn6) or related proteins (bovine factor Va - pdb 1sdd, human factor VIII – pdb 3cdz, 2r7e, human neuropilin – pdb 2qqj, 2qqk, 2qqm, 2 qqo, 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 1ssd (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, 152H, 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

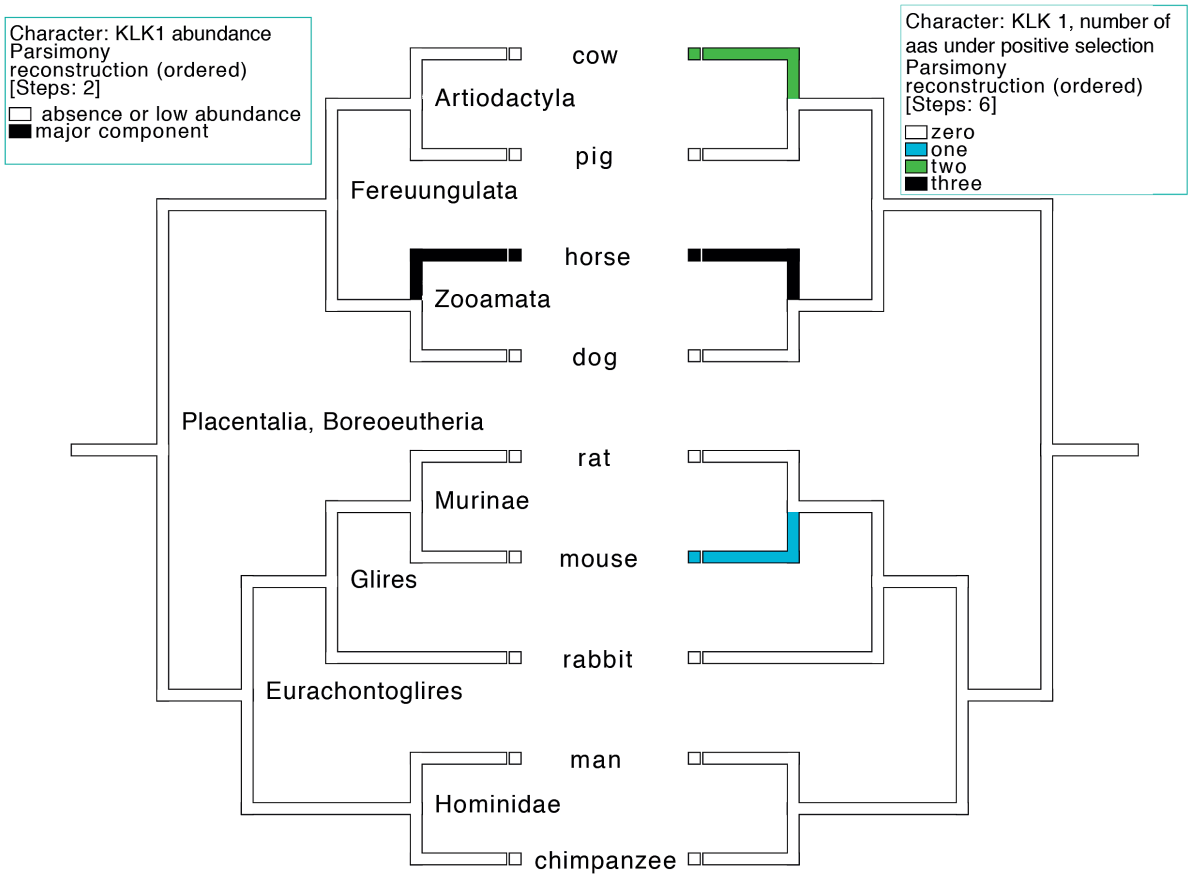

Fig. 5. Lack of correlation between abundance (left) and number of amino acids under positive selection in KLK1. Parsimony optimizations performed in Mesquite (Maddison and Maddison, 2015).

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-

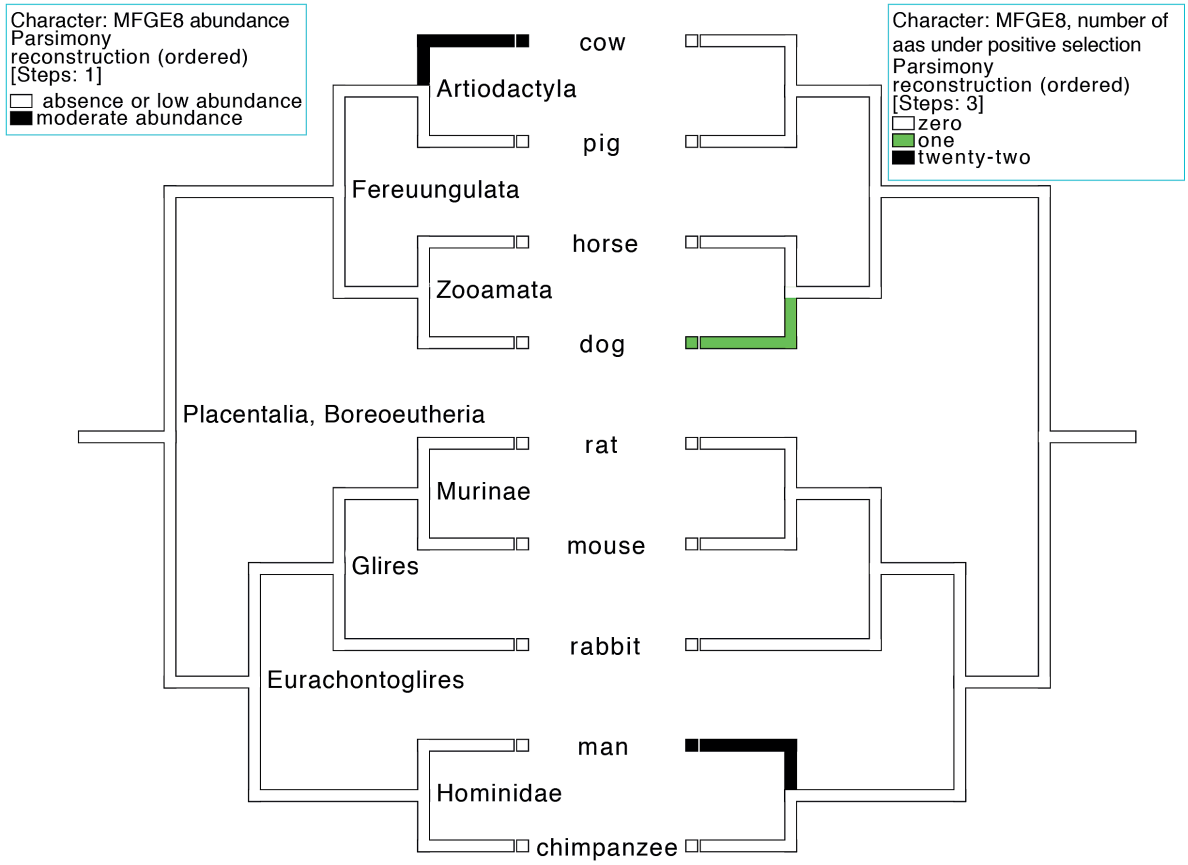

Fig. 6. Lack of correlation between abundance (left) and number of amino acids under positive selection in MFGE8. Parsimony optimizations performed in Mesquite (Maddison and Maddison, 2015).

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 (Fornůsková *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, polyginous, 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

Adams TE, Hockin MF, Mann KG, Everse SJ. 2004. The crystal structure of activated protein C-inactivated bovine factor Va: Implications for cofactor function. *Proceedings of the National Academy of Sciences of the United States of America* 101: 8918-8923.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic acids research* 25: 3389-3402.

Andrés JA, Maroja LS, Bogdanowicz SM, Swanson WJ, Harrison RG. 2006. Molecular evolution of seminal proteins in field crickets. *Molecular Biology and Evolution* 23: 1574-1584.

Batruch I, Lecker I, Kagedan D, Smith CR, Mullen BJ, Grober E, Lo KC, Diamandis EP, Jarvi KA. 2011. Proteomic analysis of seminal plasma from normal volunteers and post-vasectomy patients identifies over 2000 proteins and candidate biomarkers of the urogenital system. *Journal of proteome research* 10: 941-953.

Bax B, Blundell TL, Murray-Rust J, McDonald NQ. 1997. Structure of mouse 7S NGF: a complex of nerve growth factor with four binding proteins. *Structure* 5: 1275-1285.

Begun DJ, Lindfors HA. 2005. Rapid evolution of genomic Acp complement in the melanogaster subgroup of *Drosophila*. *Molecular Biology and Evolution* 22: 2010-2021.

Begun DJ, Lindfors HA, Thompson ME, Holloway AK. 2006. Recently evolved genes identified from *Drosophila yakuba* and *D. erecta* accessory gland expressed sequence tags. *Genetics* 172: 1675-1681.

Belleannée C, Labas V, Teixeira-Gomes A-P, Gatti JL, Dacheux J-L, Dacheux F. 2011. Identification of luminal and secreted proteins in bull epididymis. *Journal of proteomics* 74: 59-78.

Bennett-Lovsey RM, Herbert AD, Sternberg MJ, Kelley LA. 2008. Exploring the extremes of sequence/structure space with ensemble fold recognition in the program Phyre. *Proteins: Structure, Function, and Bioinformatics* 70: 611-625.

Burtnick LD, Urosev D, Irobi E, Narayan K, Robinson RC. 2004. Structure of the N-terminal half of gelsolin bound to actin: roles in severing, apoptosis and FAF. *The EMBO journal* 23: 2713-2722.

Busset J, Cabau C, Meslin C, Pascal G. 2011. PhyleasProg: a user-oriented web server for wide evolutionary analyses. *Nucleic acids research* 39: W479-W485.

Callebaut I, Labesse G, Durand P, Poupon A, Canard L, Chomilier J, Henrissat B, Mornon J. 1997. Deciphering protein sequence information through hydrophobic cluster analysis (HCA): current status and perspectives. *Cellular and Molecular Life Sciences CMLS* 53: 621-645.

Carvalho AL, Sanz L, Barettino D, Romero A, Calvete JJ, Romão MJ. 2002. Crystal structure of a prostate kallikrein isolated from stallion seminal plasma: a homologue of human PSA. *Journal of molecular biology* 322: 325-337.

Castella S, Fouchécourt S, Teixeira-Gomes AP, Vinh J, Belghazi M, Dacheux F, Dacheux J-L. 2004. Identification of a member of a new RNase a family specifically secreted by epididymal caput epithelium. *Biology of reproduction* 70: 319-328.

Chapman T, Arnqvist G, Bangham J, Rowe L. 2003. Sexual conflict. *Trends in Ecology & Evolution* 18: 41-47.

Chapman T, Davies SJ. 2004. Functions and analysis of the seminal fluid proteins of male *Drosophila melanogaster* fruit flies. *Peptides* 25: 1477-1490.

Cho S, Beintema JJ, Zhang J. 2005. The ribonuclease A superfamily of mammals and birds: identifying new members and tracing evolutionary histories. *Genomics* 85: 208-220.

Civetta A, Clark AG. 2000. Correlated effects of sperm competition and postmating female mortality. *Proceedings of the National Academy of Sciences* 97: 13162-13165.

Clark AG, Begun DJ, Prout T. 1999. Female× male interactions in *Drosophila* sperm competition. *Science* 283: 217-220.

Clark NL, Swanson WJ. 2005. Pervasive adaptive evolution in primate seminal proteins. *PLoS Genetics* 1: e35.

Dean MD, Findlay GD, Hoopmann MR, Wu CC, MacCoss MJ, Swanson WJ, Nachman MW. 2011. Identification of ejaculated proteins in the house mouse (*Mus domesticus*) via isotopic labeling. *BMC genomics* 12: 306.

Dorus S, Evans PD, Wyckoff GJ, Choi SS, Lahn BT. 2004. Rate of molecular evolution of the seminal protein gene SEMG2 correlates with levels of female promiscuity. *Nature genetics* 36: 1326-1329.

Druart X, Rickard J, Mactier S, Kohnke P, Kershaw-Young C, Bathgate R, Gibb Z, Crossett B, Tsikis G, Labas V. 2013. Proteomic characterization and cross species comparison of mammalian seminal plasma. *Journal of proteomics* 91: 13-22.

Drummond DA, Bloom JD, Adami C, Wilke CO, Arnold FH. 2005. Why highly expressed proteins evolve slowly. *Proceedings of the National Academy of Sciences of the United States of America* 102: 14338-14343.

Faith DP. 1992. Conservation evaluation and phylogenetic diversity. *Biological Conservation* 61: 1-10.

Felsenstein J. 1985. Phylogenies and the comparative method. *American Naturalist*: 1-15.

Findlay GD, MacCoss MJ, Swanson WJ. 2009. Proteomic discovery of previously unannotated, rapidly evolving seminal fluid genes in *Drosophila*. *Genome research* 19: 886-896.

Fornůsková A, Vinkler M, Pagès M, Galan M, Jousselin E, Cerqueira F, Morand S, Charbonnel N, Bryja J, Cosson J-F. 2013. Contrasted evolutionary histories of two Toll-like receptors (Tlr4 and Tlr7) in wild rodents (MURINAE). *BMC evolutionary biology* 13: 194.

Galindo BE, Vacquier VD, Swanson WJ. 2003. Positive selection in the egg receptor for abalone sperm lysin. *Proceedings of the National Academy of Sciences* 100: 4639-4643.

Hassanin A, Delsuc F, Ropiquet A, Hammer C, van Vuuren BJ, Matthee C, Ruiz-Garcia M, Catzeflis F, Areskoug V, Nguyen TT. 2012. Pattern and timing of diversification of Cetartiodactyla (Mammalia, Laurasiatheria), as revealed by a comprehensive analysis of mitochondrial genomes. *Comptes rendus biologies* 335: 32-50.

Holland B, Rice WR. 1999. Experimental removal of sexual selection reverses intersexual antagonistic coevolution and removes a reproductive load. *Proceedings of the National Academy of Sciences* 96: 5083-5088.

Hurle B, Swanson W, Green ED. 2007. Comparative sequence analyses reveal rapid and divergent evolutionary changes of the WFDC locus in the primate lineage. *Genome research* 17: 276-286.

Jensen-Seaman MI, Li W-H. 2003. Evolution of the hominoid semenogelin genes, the major proteins of ejaculated semen. *Journal of molecular evolution* 57: 261-270.

Josse S, Moreau T, Laurin M. 2006. Stratigraphic tools for Mesquite. *See* http: //mesquiteproject. *org/packages/stratigraphicTools*.

Karn RC, Clark NL, Nguyen ED, Swanson WJ. 2008. Adaptive evolution in rodent seminal vesicle secretion proteins. *Molecular Biology and Evolution* 25: 2301-2310.

Kelly VC, Kuy S, Palmer DJ, Xu Z, Davis SR, Cooper GJ. 2006. Characterization of bovine seminal plasma by proteomics. *Proteomics* 6: 5826-5833.

LaPatra SE, Evilia C, Winston V. 2008. Positively selected sites on the surface glycoprotein (G) of infectious hematopoietic necrosis virus. *Journal of general virology* 89: 703-708.

Maddison WP. 2000. Testing character correlation using pairwise comparisons on a phylogeny. *Journal of Theoretical Biology* 202: 195-204.

Maddison WP, Maddison DR. 2015. Mesquite: a modular system for evolutionary analysis. Version 3.02.

Marques PI, Bernardino R, Fernandes T, Green ED, Hurle B, Quesada V, Seixas S. 2012. Birth-and-death of KLK3 and KLK2 in primates: evolution driven by reproductive biology. *Genome biology and evolution* 4: 1331-1338.

Marshall JL, Huestis DL, Garcia C, Hiromasa Y, Wheeler S, Noh S, Tomich JM, Howard DJ. 2011. Comparative proteomics uncovers the signature of natural selection acting on the ejaculate proteomes of two cricket species isolated by postmating, prezygotic phenotypes. *Molecular Biology and Evolution* 28: 423-435.

Martí-Renom MA, Stuart AC, Fiser A, Sánchez R, Melo F, Šali A. 2000. Comparative protein structure modeling of genes and genomes. *Annual review of biophysics and biomolecular structure* 29: 291-325.

Meslin C, Brimau F, Meillour PN, Callebaut I, Pascal G, Monget P. 2011. The evolutionary history of the SAL1 gene family in eutherian mammals. *BMC evolutionary biology* 11: 148.

Meslin C, Mugnier S, Callebaut I, Laurin M, Pascal G, Poupon A, Goudet G, Monget P. 2012. Evolution of genes involved in gamete interaction: evidence for positive selection, duplications and losses in vertebrates. *PloS one* 7: e44548.

Midford P, Garland Jr T, Maddison W. 2008. PDAP package of Mesquite, version 1.16. *See* http: //mesquiteproject. *org/pdap_mesquite*.

Milardi D, Grande G, Vincenzoni F, Messana I, Pontecorvi A, De Marinis L, Castagnola M, Marana R. 2012. Proteomic approach in the identification of fertility pattern in seminal plasma of fertile men. *Fertility and sterility* 97: 67-73. e61.

Moura AA, Chapman DA, Koc H, Killian GJ. 2007. A comprehensive proteomic analysis of the accessory sex gland fluid from mature Holstein bulls. *Animal reproduction science* 98: 169-188.

Mueller JL, Ram KR, McGraw LA, Qazi MB, Siggia ED, Clark AG, Aquadro CF, Wolfner MF. 2005. Cross-species comparison of *Drosophila* male accessory gland protein genes. *Genetics* 171: 131-143.

Murphy WJ, Eizirik E. 2009. Placental mammals (Eutheria). In: Hedges SB, S. K, editors. The timetree of life. New York: Oxford Univeristy Press. p. 474-474.

Nag S, Ma Q, Wang H, Chumnarnsilpa S, Lee WL, Larsson M, Kannan B, Hernandez-Valladares M, Burtnick LD, Robinson RC. 2009. Ca2+ binding by domain 2 plays a critical role in the activation and stabilization of gelsolin. *Proceedings of the National Academy of Sciences* 106: 13713-13718.

Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148: 929-936.

Pagel M. 1994. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society of London. Series B: Biological Sciences* 255: 37-45.

Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. 2004. UCSF Chimera—a visualization system for exploratory research and analysis. *Journal of computational chemistry* 25: 1605-1612.

Pilch B, Mann M. 2006. Large-scale and high-confidence proteomic analysis of human seminal plasma. *Genome biology* 7: R40.

Poiani A. 2006. Complexity of seminal fluid: a review. *Behavioral Ecology and Sociobiology* 60: 289-310.

Pond SLK, Murrell B, Fourment M, Frost SD, Delport W, Scheffler K. 2011. A random effects branch-site model for detecting episodic diversifying selection. *Molecular Biology and Evolution*: msr125.

Ramm SA, McDonald L, Hurst JL, Beynon RJ, Stockley P. 2009. Comparative proteomics reveals evidence for evolutionary diversification of rodent seminal fluid and its functional significance in sperm competition. *Molecular Biology and Evolution* 26: 189-198.

Ramm SA, Oliver PL, Ponting CP, Stockley P, Emes RD. 2008. Sexual selection and the adaptive evolution of mammalian ejaculate proteins. *Molecular Biology and Evolution* 25: 207-219.

Raymond A, Ensslin MA, Shur BD. 2009. SED1/MFG-E8: A Bi-Motif protein that orchestrates diverse cellular interactions. *Journal of cellular biochemistry* 106: 957-966.

Rees GN, Baldwin DS, Watson GO, Perryman S, Nielsen DL. 2005. Ordination and significance testing of microbial community composition derived from terminal restriction fragment length polymorphisms: application of multivariate statistics. *Antonie van Leeuwenhoek* 86: 339-347.

Rice WR. 1996. Sexually antagonistic male adaptation triggered by experimental arrest of female evolution. *Nature* 381: 232-234.

Rodrigues M, Souza C, Martins J, Rego J, Oliveira J, Domont G, Nogueira F, Moura A. 2013. Seminal plasma proteins and their relationship with sperm motility in Santa Ines rams. *Small Ruminant Research* 109: 94-100.

Rowe KC, Reno ML, Richmond DM, Adkins RM, Steppan SJ. 2008. Pliocene colonization and adaptive radiations in Australia and New Guinea (Sahul): Multilocus systematics of the old endemic rodents (Muroidea: Murinae). *Molecular phylogenetics and evolution* 47: 84-101.

Shao C, Novakovic VA, Head JF, Seaton BA, Gilbert GE. 2008. Crystal structure of lactadherin C2 domain at 1.7 Å resolution with mutational and computational analyses of its membrane-binding motif. *Journal of Biological Chemistry* 283: 7230-7241.

Souza CEA, Rego J, Lobo CH, Oliveira JTA, Nogueira F, Domont GB, Fioramonte M, Gozzo FC, Moreno FB, Monteiro-Moreira A. 2012. Proteomic analysis of the reproductive tract fluids from tropically-adapted Santa Ines rams. *J Proteomics* 75: 4436-4456.

Stoline MR. 1981. The status of multiple comparisons: simultaneous estimation of all pairwise comparisons in one-way ANOVA designs. *The American Statistician* 35: 134-141.

Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* 100: 9440-9445.

Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic acids research* 34: W609-W612.

Swanson WJ, Clark AG, Waldrip-Dail HM, Wolfner MF, Aquadro CF. 2001. Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. *Proceedings of the National Academy of Sciences* 98: 7375-7379.

Swanson WJ, Nielsen R, Yang Q. 2003. Pervasive adaptive evolution in mammalian fertilization proteins. *Molecular Biology and Evolution* 20: 18-20.

Swanson WJ, Wong A, Wolfner MF, Aquadro CF. 2004. Evolutionary expressed sequence tag analysis of *Drosophila* female reproductive tracts identifies genes subjected to positive selection. *Genetics* 168: 1457-1465.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research* 22: 4673-4680.

Tian X, Pascal G, Fouchécourt S, Pontarotti P, Monget P. 2009. Gene birth, death, and divergence: the different scenarios of reproduction-related gene evolution. *Biology of reproduction* 80: 616-621.

Vignaud P, Duringer P, Mackaye HT, Likius A, Blondel C, Boisserie J-R, De Bonis L, Eisenmann V, Etienne M-E, Geraads D. 2002. Geology and palaeontology of the Upper Miocene Toros-Menalla hominid locality, Chad. *Nature* 418: 152-155.

Wagstaff BJ, Begun DJ. 2007. Adaptive evolution of recently duplicated accessory gland protein genes in desert *Drosophila*. *Genetics* 177: 1023-1030.

Wagstaff BJ, Begun DJ. 2005a. Comparative genomics of accessory gland protein genes in *Drosophila melanogaster* and *D. pseudoobscura*. *Molecular Biology and Evolution* 22: 818-832.

Wagstaff BJ, Begun DJ. 2005b. Molecular population genetics of accessory gland protein genes and testis-expressed genes in *Drosophila mojavensis* and *D. arizonae*. *Genetics* 171: 1083-1101.

Walters JR, Harrison RG. 2010. Combined EST and proteomic analysis identifies rapidly evolving seminal fluid proteins in *Heliconius* butterflies. *Molecular Biology and Evolution* 27: 2000-2013.

Wolfner M. 2002. The gifts that keep on giving: physiological functions and evolutionary dynamics of male seminal proteins in *Drosophila*. *Heredity* 88: 85-93.

Wong WS, Yang Z, Goldman N, Nielsen R. 2004. Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. *Genetics* 168: 1041-1051.

Yang Z. 2000. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. *Journal of molecular evolution* 51: 423-432.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Molecular Biology and Evolution* 24: 1586-1591.

Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Molecular Biology and Evolution* 19: 908-917.

Yang Z, Wong WS, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. *Molecular Biology and Evolution* 22: 1107-1118.

Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Molecular Biology and Evolution* 22: 2472-2479.

Received: 14 October 2014
Revised and accepted: 4 March 2015
Published online: 4 September 2015
Editor: J.W. Arntzen

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.

S4. Positive selection.

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.