Journal of Proteomics 135 (2016) 38–50

Contents lists available at ScienceDirect

Journal of Proteomics

journal homepage: www.elsevier.com/locate/jprot

Cross-species proteomics in analysis of mammalian sperm proteins

Helen L. Bayram [a,1], Amy J. Claydon [b], Philip J. Brownridge [b], Jane L. Hurst [a], Alan Mileham [c], Paula Stockley [a], Robert J. Beynon [b], Dean E. Hammond [d,*]

[a] Mammalian Behaviour and Evolution Group, Institute of Integrative Biology, University of Liverpool, Leahurst Campus, Neston CH64 7TE, UK
[b] Centre for Proteome Research, Institute of Integrative Biology, University of Liverpool, Liverpool L69 7ZB, UK
[c] Genus plc, 1525 River Road, DeForest, WI 53532, USA
[d] Cellular and Molecular Physiology, Institute of Translational Medicine, University of Liverpool, Liverpool L69 3BX, UK

ARTICLE INFO

Article history:
Received 31 August 2015
Received in revised form 28 December 2015
Accepted 29 December 2015
Available online 6 January 2016

Keywords:
Cross-species proteomics
Sperm
Label-free quantification

ABSTRACT

Many proteomics studies are conducted in model organisms for which fully annotated, detailed, high quality proteomes are available. By contrast, many studies in ecology and evolution are conducted in species which lack high quality proteome data, limiting the perceived value of a proteomic approach for protein discovery and quantification. This is particularly true of rapidly evolving proteins in the reproductive system, such as those that have an immune function or are under sexual selection, and can compromise the potential for cross-species proteomics to yield confident identification. In this investigation we analysed the sperm proteome, from a range of ungulates and rodents, and explored the potential of routine proteomic workflows to yield characterisation and quantification in non-model organisms. We report that database searching is robust to cross-species matching for a mammalian core sperm proteome, comprising 623 proteins that were common to most of the 19 species studied here, suggesting that these proteins are likely to be present and identifiable across many mammalian sperm. Further, label-free quantification reveals a consistent pattern of expression level. Functional analysis of this core proteome suggests consistency with previous studies limited to model organisms and has value as a quantitative reference for analysis of species-specific protein characterisation.

Significance: From analysis of the sperm proteome for diverse species (rodents and ungulates) using LC–MS/MS workflows and standard data processing, we show that it is feasible to obtain cross-species matches for a large number of proteins that can be filtered stringently to yield a highly expressed mammalian sperm core proteome, for which label-free quantitative data are also used to inform protein function and abundance.

1. Introduction

A greater understanding of the proteins involved in reproduction can benefit human fertility and animal production. A comparative approach, comparing these proteins across multiple species, can also give insights into evolutionary adaptations e.g. [1,2]. Recent advances in proteomic techniques, particularly quantitative proteomics, have produced indispensable tools in biological research (for reviews of common proteomic techniques and their use in reproductive biology studies see [1,3,4]). However, many proteomic techniques rely on prior knowledge of the expected protein sequences. Databases such as UniProtKB offer a broad resource of protein sequence information that can be used to identify peptides [5]. As 98% of the protein sequences within UniProtKB have been derived from cDNA or genomic sequencing, most of the available protein sequences are reliant on the quantity and quality of DNA or RNA-derived information for that species. As a result, studies of the reproductive proteome to date have been limited to model species supported by extensive, high quality genomic information (e.g. sperm proteomes in *Drosophila* [6–8], house mice [9] and humans [10,11]) or with dedicated genome projects (such as the honeybee, *Apis mellifera* [12]).

Extending proteomic studies to 'non-model' organisms that lack fully annotated genome data is challenging. Protein identification is reliant on cross-species matching [13] and thus, success is based on the taxonomic proximity of high quality proteome data and the rate of divergence of protein sequences. In proteomics based on tandem mass spectrometry, searches of product ion spectra against known protein sequences of different species can score highly, depending on the degree of protein sequence similarity and thus protein homology. The search algorithms will match peptide sequences to homologous proteins from closely related species provided there are limited amino acid substitutions. Work with RNA-seq data implies that this may have only a relatively small effect on data analysis, particularly for highly conserved sequences [14]. The initial aim of this study was to investigate to what extent protein sequence divergence might affect identification following mass spectrometry. Although genomic approaches offer a more

* Corresponding author.
[1] Current address: Department of Ecology and Genetics, Evolutionary Biology Centre, Uppsala University, Norbyvägen 18D, 752 36, Sweden.

http://dx.doi.org/10.1016/j.jprot.2015.12.027
1874-3919/

direct route to the quantification of evolutionary divergence, proteomic methods have the advantage of providing a more accurate depiction of the protein complement. Methods relying on mRNA transcripts are disconnected to the outcome of protein expression, wherein the steady state level of a protein can reflect the balance between translation and degradation or secretion [15]. Proteomic methods could be of particular benefit to sperm proteins that have testis-specific translation e.g. zonadhesin, [16], and therefore wouldn't be identified using RNA level techniques when analysing mature sperm.

Thorough knowledge of the molecular composition of sperm is an essential basis from which to understand the molecular interactions which occur at fertilisation, and thereby may help improve fertility. There have been comprehensive efforts to establish "core sperm proteomes" [8] of *Drosophila* [6–8], mice [9,17], rats [18], rhesus macaque [19], humans [10,20], and non-model insect species (e.g. *Manduca sexta* [21] and *A. mellifera* [12]). A recent comparison of the *Drosophila* and *M. sexta* sperm proteomes suggest many cross species similarities [21]. However, proteins involved in reproduction are predicted to evolve rapidly e.g. [1,22–24], potentially leading to reproductive isolation and speciation [25]. Across diverse sexually reproducing taxa, post-copulatory sexual selection and sexual conflict have been implicated in driving both the rapid evolution, and co-evolution, of male and female reproductive proteins [e.g. 26–29]. Additionally, mammalian sperm contain an extensive range of immunity proteins [24], which are generally known to evolve at a high rate [30]. Comparative proteomic analysis could therefore be compromised for those proteins that are evolving most rapidly, such as those involved in gamete interactions. Nonetheless, there is also some evidence that a majority of proteins present in the reproductive tract [31], and more specifically within sperm [17], are relatively conserved. This combination of proteins that are expected to evolve rapidly and of conserved proteins in sperm offers a useful test for proteomic analysis of non-model organisms. Here we compare which functional groups of sperm proteins are readily identified across different mammalian species.

We analysed sperm proteomes for mammalian species that varied in the availability of genomic, and hence proteomic, data. Our samples were selected from two major mammalian taxonomic groups, rodents and ungulates, for which varying degrees of genomic information is available for a few species (mouse, rat, cattle, pig, and sheep) and also include some 'non-model' relatives (Table 1). We selected this range to test whether phylogenetic similarity allows protein identification in species for which the supporting databases were of limited value. In this study, cauda epididymal samples were chosen for proteomic analysis as they offer a concentrated supply of sperm and associated proteins. After production in mammalian testes, immature sperm travel to the caput epididymis for further maturation and mature sperm are then stored within the cauda epididymis prior to ejaculation. Analysing a sperm rich sample allows comparison of those proteins that may be functionally conserved, such as those involved in metabolism and sperm structure, and proteins that are likely to be species specific, notably those that are essential for sperm–egg interaction.

Table 1
Epididymal sperm samples collected for proteomic analysis.

Species	Common name	N
Rodents		
Apodemus sylvaticus	Wood mouse	3
Microtus agrestis	Field vole	2
Myodes glareolus	Bank vole	2
Rattus norvegicus	Brown Norway rat	2
Rattus norvegicus	Wistar rat	2
Sciurus vulgaris	Red squirrel	1
Ungulates		
Antilope cervicapra	Blackbuck	1
Bos taurus	Domestic cattle	1
Bos taurus indicus	Ankole cattle	2
Cervus alfredi	Prince Alfred's spotted deer	1
Connochaetes gnou	Black wildebeest	1
Equus grevyi	Grevy's zebra	1
Kobus leche	Red lechwe	1
Oryx dammah	Scimitar-horned oryx	1
Oryx gazella	Gemsbok	1
Ovis aries	Domestic sheep	3
Phacochoerus africanus	Common warthog	1
Sus scrofa domesticus	Domestic pig	3
Syncerus caffer	Cape buffalo	1

2. Methods

2.1. Samples

Samples of testicular tissue were collected following castration or death of ungulate species from zoological collections (*Antilope*, *Bos*, *Cervus*, *Connochaetes*, *Equus*, *Kobus*, *Oryx*, *Phacochoerus*, *Sus*, *Syncerus*), domestic ungulates from abattoirs (*Bos*, *Ovis*, *Sus*), and farms (*Ovis*), and from wild and laboratory rodents bred in captivity (*Rattus*), caught from local populations in Cheshire (*Apodemus*, *Myodes*, *Microtus*), or found dead in natural populations (*Sciurus*) (Table 1). Ungulate samples were chilled immediately on collection with the majority processed within 12 h (the longest delay between removal and processing was three days). Scrotal tissue and the tunica vaginalis were removed and the testis washed twice in double distilled water and once in 70% (*v*/*v*) ethanol. The epididymis was cleaved from the testis and a section of approximately 1 cm^3 dissected from the caudal section and placed in a Petri dish. This tissue was scored repeatedly and 400 μl phosphate buffered saline (PBS) was added before collecting the sperm suspension after 5 min of agitation. The Petri dish was then washed with a further 400 μl of PBS that was also collected. The sperm suspension was frozen at −80 °C prior to proteomic analysis. Rodent samples were dissected immediately after death, with the exception of red squirrels that had died of natural causes. In the latter case, although it was not possible to determine how long an animal had been dead, from the quality of the sample it could be assumed that dissection occurred within two days post-mortem. Both epididymides were dissected and the cauda section isolated. The paired cauda epididymides were placed onto a Petri dish with 200 μl of PBS, before being macerated with a scalpel. After 5 min the sperm suspension was stored at −80 °C.

2.2. Proteomic analysis

In total, 30 samples from 19 species were analysed using tandem mass spectrometry. Each sample was defrosted and vortexed for 1 min before protein concentration was determined using a dye binding protein assay. Tryptic digests were performed on 50 μg protein per sample, within a total final volume of 200 μl. Briefly, proteins were denatured using 0.05% (*w*/*v*) RapiGest SF Surfactant (Waters) at 80 °C for 5 min. The samples were then incubated with 3 mM dithiothreitol (60 °C 10 min), followed by 9 mM iodoacetamide (RT 60 min in the dark to reduce and carbamidomethylate the proteins. Trypsin (final concentration, 0.01 μg/μl) was added and the sample was incubated overnight at 37 °C. After 12 h, 1 M hydrochloric acid and additional trypsin (final concentration, 0.015 μg/μl) was added and left to incubate for a further 4 h. Inactivated detergent was precipitated using trifluoroacetic acid (final volume 0.5% (*v*/*v*)) for 45 min at 37 °C. Samples were centrifuged for 90 min at 17,000 ×*g* and 4 °C before being diluted 1:1 with 97:3:0.1 HPLC grade water:MeOH:TFA. From each sample, 1 μl of the diluted digest was resolved over a 90 min linear organic gradient using ultra performance liquid chromatography (Waters nanoAcquity) coupled to an LTQ Orbitrap Velos (ThermoFisher). The mass spectrometry proteomic data have been deposited to the ProteomeXchange Consortium [32] via the PRIDE partner repository with the data set identifier PXD003164.

2.3. Protein identification and quantification

Raw LC–MS/MS peak list files from each experimental sample were searched against a UniProt validated database for all mammalian species using the Andromeda search engine [33] or against species-specific databases within the MaxQuant software suite (version 1.5.8.3) [34]. The minimum required peptide was seven amino acids long and a single missed cleavage was allowed. Cysteine carbamidomethylation (C) was set as a fixed modification and methionine oxidation was allowed as a variable modification. The initial precursor and fragment ion maximum mass deviations were set to 20 ppm and 0.5 Da, respectively. The custom-built UniProt all_mammalian_species ("all mammals") FASTA database contained 66,323 entries across 1878 species. The results of the database search were further processed and statistically evaluated by MaxQuant. Peptide and protein false discovery rates were set to 1%. For protein quantification, intensity-based label-free quantification iBAQ, [35] was used. Multiple iterations of Andromeda–MaxQuant processing against a variety of UniProt databases were carried out in this study, to address the question of cross-species proteomics in our experimental system. These were as follows: a) a "no fractions", individual file-by-file run against the "all mammals" database, to allow tests of LC–MS/MS reproducibility (Supplementary Fig. S1) and obtain search result statistics (Fig. 2); b) a grouped/combined run against the "all mammalian" database, wherein biological replicate files (where available) were combined, within MaxQuant, into distinct fractions/groups during processing; c) a grouped/combined run against only BOVIN (representative of Ungulates) entries in the "all mammalian" database, containing 5985 entries; and d) a grouped/combined run against only MOUSE (representative of rodents) entries in the "all mammals" database, containing 16,657 entries.

To consider the effect of sequence divergence on protein identification dN/dS values of proteins identified in either most or few species were compared. Raw MS data for all species were searched against both the *Bos taurus* and *Mus musculus* UniProt databases using the Andromeda search engine. The resulting protein hits were filtered to only those with at least 3 peptide matches. For proteins present in either at least 16 species or at most 5 species, dN/dS measurements between homologous Mouse and Bovine IDs were calculated from the ensembl *B. taurus* and *M. musculus* marts using the biomaRt package (v. 2.26.1) in Bioconductor [36]. The resultant data were plotted using a kernel density estimation (Fig. 6).

2.4. Protein functional analysis

To automate the process of biological functional term classification (and their enrichment) of protein clusters in our data, we used the R/Bioconductor package clusterProfiler (version 2.4.1) [37]. We used the function 'groupGO', to classify identified proteins based on their projection at a specific level of the gene ontology (GO) hierarchy, specifically focussing on the GO 'Biological Process' (GOBP) class. The function 'enrichGO' was then used to perform enrichment tests for GOBP terms based on a hypergeometric distribution against a background list of all proteins in the corresponding annotation database (for Mouse = *M. musculus*; for Bovine = *B. taurus*). To prevent the high false discovery rates (FDRs) common when using multiple hypothesis testing, we applied a Benjamini–Hochberg p-value threshold of 0.0001 (q-value cut-off = 0.01). To obtain a visual representation of all GO terms – 'Molecular Function' (GOMF), 'Cellular Compartment' (GOCC) and GOBP – linked to our 'core mammalian sperm proteome', in the form of a network, we used g:GOSt, part of the g:Profiler suite of web tools (see http://biit.cs.ut.ee/gprofiler/) [38], and the Enrichment Map application (http://www.baderlab.org/Software/EnrichmentMap) [39] within Cytoscape [40]. The gene group functional profiling tool (g:GOst) options were as follows: significant GO terms only; no electronic GO annotations; minimum and maximum size of functional category set to 3 and 500, respectively; a minimum of 2 queries per GO term; significance threshold determined by the Benjamini–Hochberg FDR testing; p-value cut-off = 1. The resultant output in generic enrichment map format, was loaded into Enrichment Map to organise the data into a network with mutually overlapping protein-sets clustering together, thus easing interpretation (Enrichment Map Tune Parameters: p-value cut-off = 1; q-value cut-off = 1; overlap coefficient of similarity cut-off = 0.5).

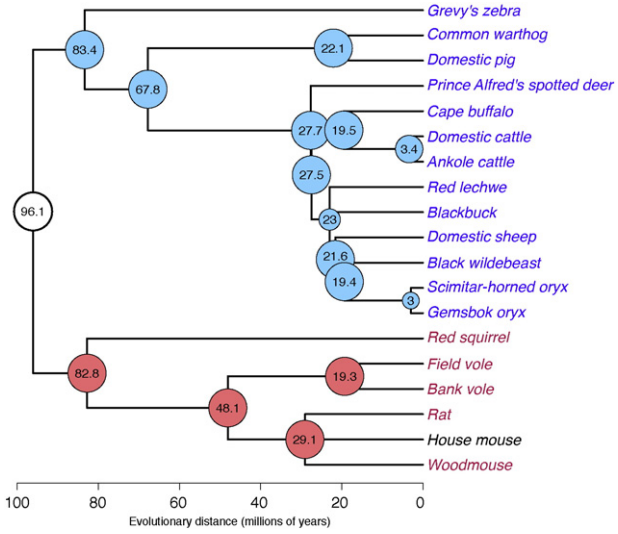


Fig. 1. Phylogeny of species used in this study. A total of 18 species were used in this study (for Latin names, see Table 1). Note that the house mouse is included (black text) because it is a reference database, but was not analysed in this study.

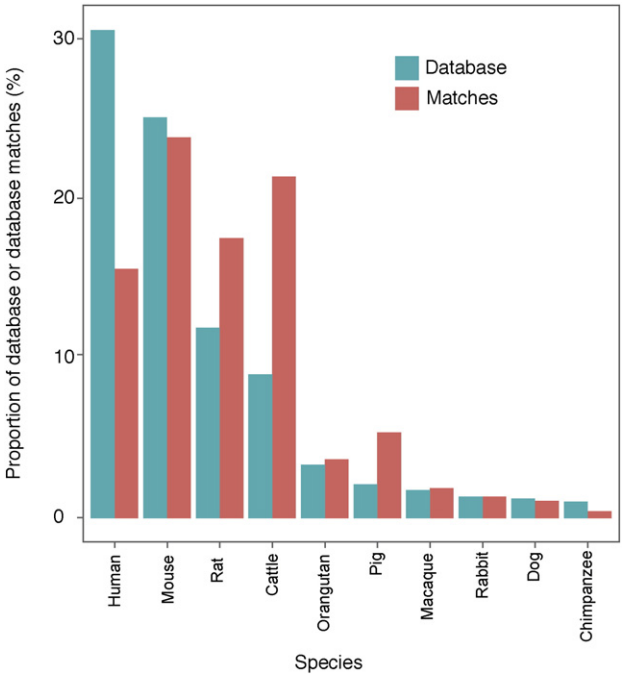


Fig. 2. Species origin of database matches using a mammalian proteome database search. All biological samples were used to direct a database search against a composite database comprising UniProt sequence entries for all mammalian species (green bars). The species distribution of the mammalian database (top ten species) is compared to the distribution of actual database matches (red bars).

3. Results and discussion

3.1. Overall proteomic approach

One of the main purposes of this study was to explore the extent to which it was possible to analyse sperm samples from organisms lacking fully annotated genomes, requiring cross species database matching to assign protein identity and, through label-free proteomics, an approximate measure of protein quantification. Cross-species comparisons could weaken the confidence of the identification and, if fewer peptides were obtained, impact on label-free quantification values. Further, variance in identification would be greater for rapidly evolving proteins, such as those involved in sexual selection and immunity. Whilst understanding these limitations, we aimed to define a common proteome that could act as a reference for other database matches and comparative label-free quantification. To restrict the extent of the cross-species exploration, we concentrated on sperm proteins from two groups of mammals: rodents and ungulates. These groups were selected due to the range of genomic information available for the domestic species, and the diversity of 'non-model' samples available from local sources.

A total of 30 distinct samples were analysed, derived from 19 species (Table 1). These included five rodent species (including 2 inbred rat strains, Wistar and Brown Norway, that were treated as independent biological material) and 13 ungulate species. Rodent and ungulate orders diverged around 100 million years ago (MYA) (Fig. 1). We have included species from each order that offer a range of evolutionary distance to the domestic species; from around 83 MYA for the *Equus* and *Sciurus* genera, to the very closely related *Bos* species which diverged around 3 MYA. Each of the preparations was normalised to a constant protein input into tryptic digestion, and the same proportion of the digested sample was analysed by LC–MS/MS. For all samples, the LC–MS/MS base peak chromatograms attested to the complexity and richness of the digests (results not shown). However, given that the samples were of different biological origin, we did not attempt to align the peptide MS chromatograms. Rather, we analysed each sample independently using the Andromeda search engine in MaxQuant and used the core MaxQuant package for label-free quantification to recover iBAQ values.

3.2. Cross-species matching

We initially searched all samples against a composite mammalian proteome database. Although this contains entries for over 1800 species, it is dominated by relatively few; human, mouse, rat, sheep and cattle account for 87.5% of the sequence entries (Fig. 2). When all sample searches were aggregated, a total of 3689 individual proteins were identified. In many instances, each protein was identified by matching to more than one sequence entry in the mammalian database, although the highest scoring match varied with protein sample and database species. Species within the database that yielded the largest

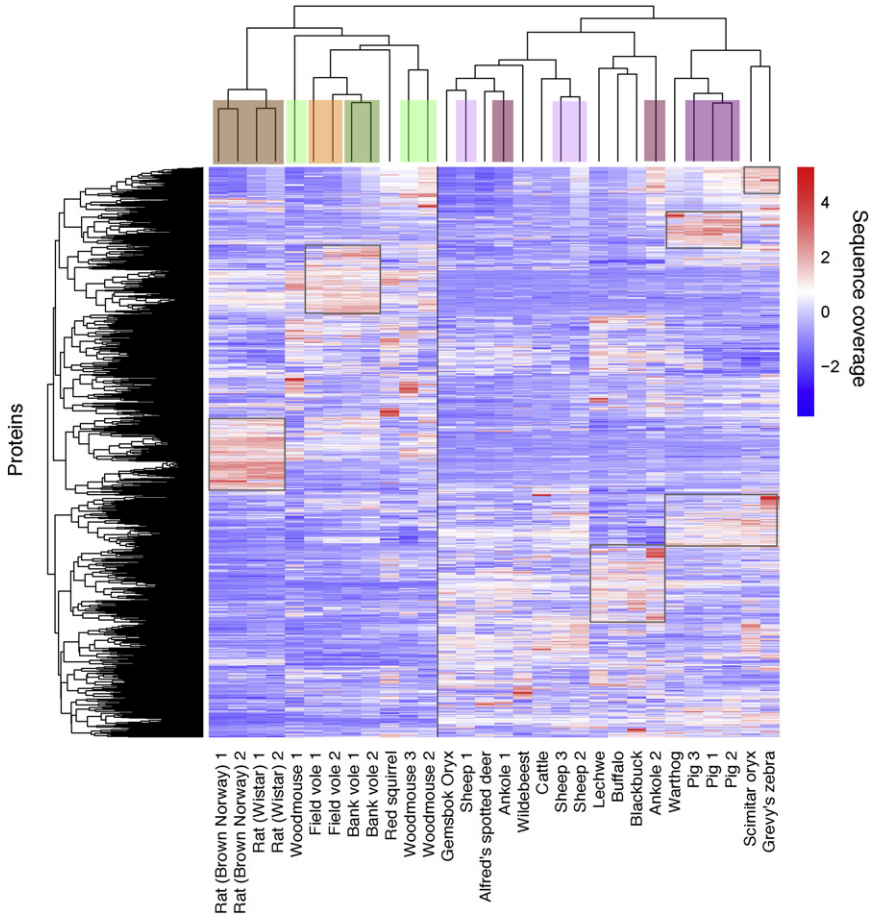


Fig. 3. Hierarchical clustering of test species based on sequence coverage. The entire data set of database matches for each test species was used to direct a hierarchical clustering analysis, with sequence coverage being used as the parameter. Replicate samples are highlighted with a common colour. Most replicates co-cluster, but there were one or two exceptions that might reflect age or sexual maturity. While both had been retrieved after castration, one was from a young male which had only few sperm stored within the epididymis. Interestingly one ankole sample clustered more closely to the lechwe sample, both of which had been collected during castration of particularly young males.

number of matches were the mouse, followed by cattle, rat, human and pig (combined database hits to these species accounted for 82% of the total). This is unsurprising, inasmuch as the two groups of test samples were from ungulates and rodents, and also, that these species are very highly represented in the curated UniProt database. The high proportion of matches to human entries (15% of the total) is likely to reflect the large number of human entries in UniProt; we have not however discarded human entries in the database to preclude duplicate entries for this search.

When the sequence coverage derived by MaxQuant/Andromeda was used to support hierarchical clustering of the entire data set, based on sequence coverage in cross-species database matching, the results were as expected (Fig. 3). First, the highest level of separation emphatically discriminated the rodent and ungulate groups. Secondly, nearly all of the biological replicates from the same species (rat, field vole, bank vole, boar) were clustered to the nearest neighbour. Exceptions were the ankole cattle, where the two samples clustered distinctly within the ungulate group and the sheep and wood mouse, where one of three samples was resolved from the other two. This clustering might reflect the degree of sexual maturation of subjects. The clustering of protein groups also implied groups of proteins that were either specific to rodents and ungulates, either in terms of sequence similarity or abundance; either factor would reduce the measured sequence coverage. It was also evident that species discrimination was attributable in part to clusters of proteins that showed distinctive differences in sequence coverage in a group-specific fashion (outlined with grey boxes in Fig. 3 to highlight some of these protein clusters).

The highest numbers of database hits were obtained from mouse and rat, or bovine and sheep entries in the database (Fig. 2). We therefore searched the proteome data for all ungulates against a simplified database consisting of combined UniProt entries for cattle, sheep, pig and horse, and for all rodents, against a combined mouse and rat UniProt database. We also performed reciprocal searches for each set of samples (rodents or ungulates) against the alternative database; searching rodent samples against the ungulate combined database, and vice versa. As anticipated, there were fewer protein group matches when there was a mismatch between the sample species and the target database (1634 for rodents, 2120 for ungulates). Additionally we compared the distribution of sequence coverage for each combined database search (Fig. 4). The majority of matched proteins exhibited a low percentage of sequence coverage (<20%), although this is similar to the coverage obtained in a recent study of mouse sperm proteins searched solely against a mouse database [41] and is thus not particularly compromised. When matched against the appropriate order-specific database, the distribution favoured identifications with a higher degree of sequence coverage, compared to the reciprocal search. When we restricted the search to a minimum

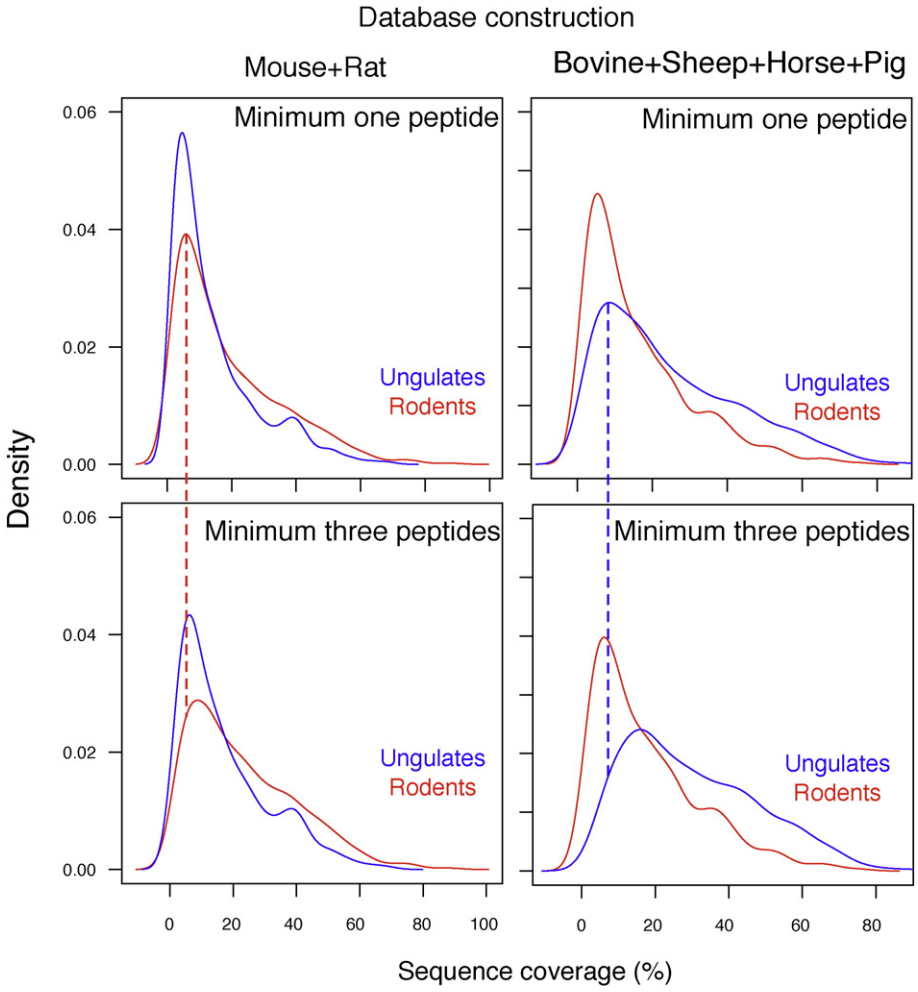


Fig. 4. Distribution of sequence coverage obtained with multi-species database searching. Raw data files from all test species were searched against two databases: rodents (comprising rat and mouse) and ungulates (comprising bovine, sheep, horse and pig). The samples were divided into rodent and ungulate groups, and searches were completed against the cognate or reciprocal databases and the sequence coverage was plotted as a distribution curve using kernel density estimation. Searches were conducted at a FDR of 1%, either unrestricted in peptide number (top panels) or limited to a minimum of three peptides for matches (bottom panels).

of three peptides, the coverage of the cognate matches improved substantially (dotted lines from top to bottom panels) but the reciprocal searches did not show a commensurate gain. This suggests that a gain in protein identification can be achieved by a taxonomically targeted database search and stringent search conditions. Few identifications were only obtained by matches to the human protein database, although these proteins were matched in virtually all samples, suggesting both that they are conserved proteins, and that the equivalent entries were missing for the other species.

Despite the apparently low coverage of proteins in the cross-species database searching, the peptide and protein scores were high and all identifications have been filtered at a FDR of 1%. We examined the relationship between sequence coverage and evolution distance for the entire data set. Further, we compared the mean sequence coverage (%) and the total number of hits at 1% FDR when searched against either the *B. taurus* (cattle) or *M. musculus* (mouse) database. Despite evolutionary divergence of nearly 100 million years, the mean sequence coverage was remarkably stable, ranging from ~20% to ~15% for the bovine database, or from ~15% to ~12% for the mouse database (Fig. 5). The total number of protein hits was equally robust, yielding about 1000 declining to 800 for the bovine database, and 1200 declining to 1000 for the mouse database. Considering the evolutionary separation, the performance was reassuring, whilst recognising that these database hits are a mixture of slowly evolving proteins and more rapidly evolving proteins that matched better to phylogenetic near-neighbours.

To further investigate the effect of protein evolution on protein identification within mammalian sperm we compared dN/dS ratios for proteins that were identified within most species, to those identified in only a few species (Fig. 6). Comparing the ratio of synonymous and non-synonymous amino acid substitutions of homologous protein sequences is a universal method of assessing sequence divergence. Higher values, close to or over 1, are indicative of positive Darwinian selection. Values nearer to 0 suggest stabilising selection. For the proteins identified within at least 16 out of 19 species here, from hits against both the Bovine and Mouse databases, the majority of proteins give dN/dS values below 0.2. This suggests that these sequences are highly conserved. In contrast, the majority of proteins found in 5 or fewer species had dN/dS values greater than 0.2. Although higher, the dN/dS values for these proteins did not indicate positive Darwinian selection. The extensive sequence conservation may in part be due to the cross-species matching used here missing very rapidly evolving proteins. However, these results show that a majority of mammalian sperm proteins are conserved and so there is value in describing a core mammalian sperm proteome.

3.3. Quantification of proteins using intensity based label free approaches

Because we used multiple species, for which neither highly annotated genome data nor curated proteome entries were available, label-free quantification was only feasible for cross-species matched proteins. Thus, we searched the rodent samples against a mouse database and searched the ungulate samples against a bovine database in order to obtain comparable protein lists. This avoids the complexity of multiple matches to entries derived from different species. The mouse and bovine databases were selected on the basis of the maximum number of entries. From these data, we were able to calculate a label-free abundance value based on the iBAQ statistic calculated by MaxQuant, based on the total precursor ion intensities of each matched peptide, divided by the total number of peptides that were theoretically possible (calculated by in silico protein digestion of tryptic peptides, without missed cleavage, between 7 and 30 amino acids long). Due to the nature of the species of origin, single samples were often all that was available. However, further confidence in the use of iBAQ values for cross species comparisons came from those samples for which we were able to obtain biological replicates (Supplementary Fig. 1). We confirmed that replicate samples were more similar than the samples from different species (results not shown) and

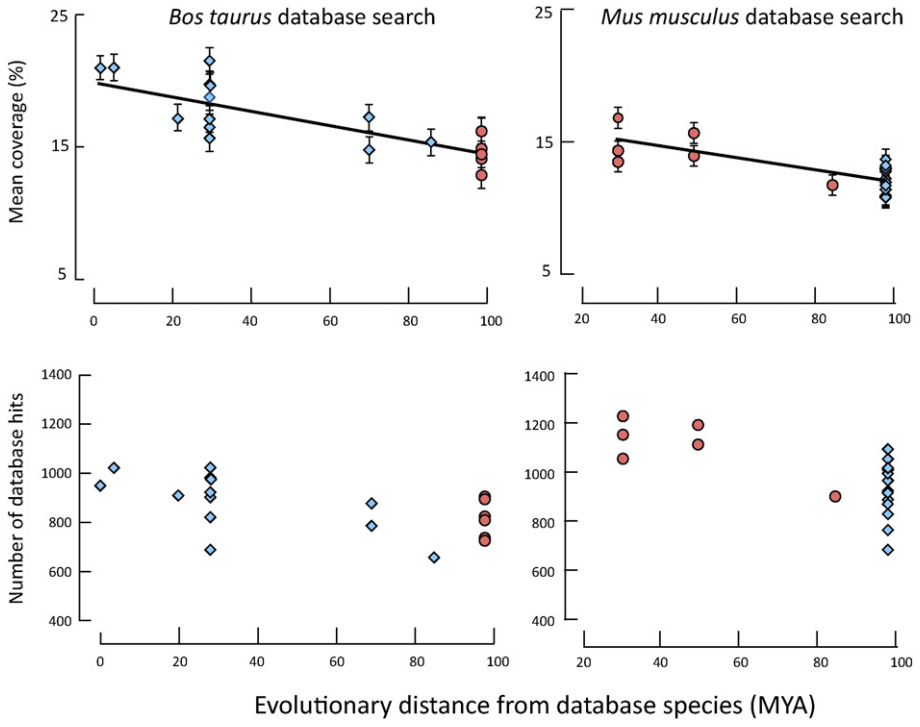


Fig. 5. Performance of database search in relation to evolutionary separation between species. For each study species (13 ungulates, 6 rodents) the MS/MS data files were searched using the Andromeda search engine against either a bovine or house mouse UniProt database. For biological replicates, the individual analyses were averaged. The sequence coverage (upper panels, mean ± SD, n variable) was expressed as a percentage of the entire protein sequence, and the total number of hits (lower panels) were filtered on a minimum of two peptides at a FDR of 1%.

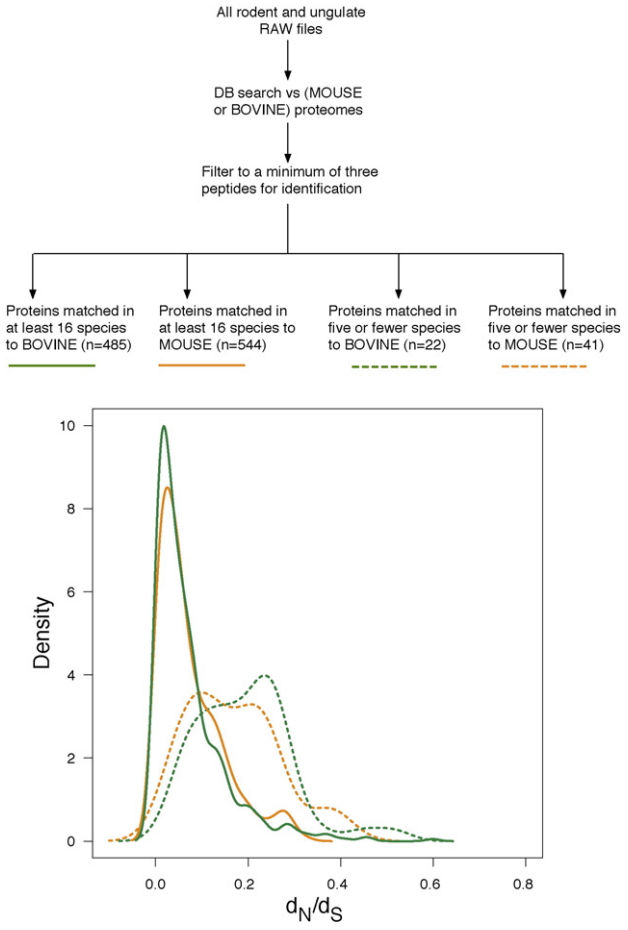


Fig. 6. Distribution of dN/dS values plotted as a distribution curve using kernel density estimation. The dN/dS values of proteins identified with a minimum of three peptides in five or fewer species (dashed lines) or at least 16 species (solid lines), either to a mouse proteome database (orange lines) or a bovine proteome database (green lines). For the proteins in each set, dN/dS values were obtained from biomaRt (as explained in the Materials and methods section) and plotted using a kernel density estimation.

we were confident in the inclusion of proteome analyses from single biological samples that lacked replicates.

We searched each ungulate sample against the bovine database and searched each rodent sample against the mouse database. For all species, the distribution of abundance in each sample spanned an iBAQ $\log_{10}$ intensity of about six orders of magnitude (Fig. 7). All samples were comparable in their dynamic range of protein expression, and the most notable distinction between samples was the number of proteins that were quantified, ranging from ca. 1400 to as low as 700 for a few species (Grevy's zebra, common warthog, blackbuck and scimitar horned oryx for ungulates, red squirrel for rodents). This was conceivably explicable by the phylogenetic distance from the model organisms in the protein databases. However, all samples revealed large numbers of proteins, even with the requirement for cross-species matching.

3.4. Protein identity and function

We attempted to build a 'mammalian core sperm proteome' — the set of proteins that had appeared in almost all biological samples, irrespective of whether cross-species matching was required. These were defined as proteins that were identified by the Andromeda search engine when rodent samples were searched against the mouse database, and ungulate samples against the bovine database. We defined an inclusive proteome (Supplementary workbook W1) that included proteins present in as few of one of each of the rodent and ungulate species. For the more restricted 'mammalian core sperm proteome', we stipulated that each protein had to be represented in at least 5 rodent species and at least 11 ungulate species. Given the variable history of the samples, and the lack of an annotated proteome, these core protein sets are likely to contain proteins that are well conserved, such that searches against model organisms would yield identification. This assumption has been confirmed by considering the dN/dS values for proteins that are identified within at least 16 species here, which are remarkably low (Fig.6). In total, 1224 proteins were retained in the inclusive proteome that reduced to 623 proteins in the core sperm proteome by our filtering criteria. (Fig. 8a, b). For these two sets, approximate quantification was obtained by label-free iBAQ scores from MaxQuant. The correlation between the rodent and ungulate 1224 protein set was good, but was considerably improved when the data set was restricted to the 623 protein core proteome (the Spearman's rank coefficient increased from 0.69 to 0.8).

The quality of the correlation between the iBAQ protein expression levels comparing the rodent and ungulate samples suggested that it might be feasible to define a core mammalian sperm proteome across these species. Whilst these proteins are likely to define common metabolic and structural processes, as well as those related to spermatic development and sperm function, their consistency in abundance across the range of rodents and ungulates was a promising approach to establishing a common quantitative data set that could be used as a reference for expression of other, more variable proteins. We also define the 623 protein proteome in terms of the relative expression when rodents and ungulates were compared (Fig 7c). Virtually all of the proteins (95%) were within a ± 1 log range with relatively few lying outside those boundaries. In addition, some proteins were only observed in either rodents or ungulates (Supplementary workbook W1).

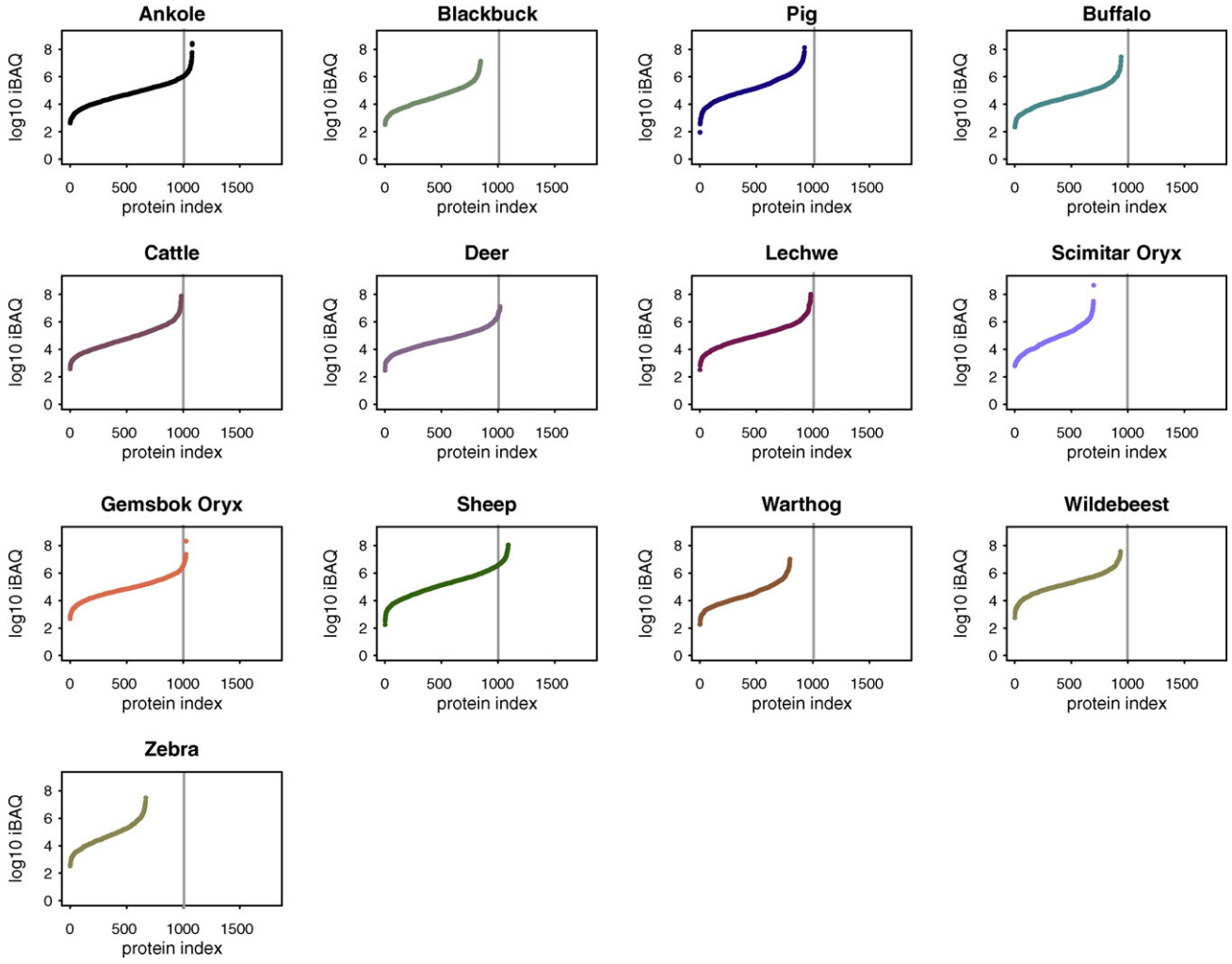
3.5. Functional analysis of mammalian core sperm proteome

Having established a 'core mammalian sperm proteome', we conducted functional analysis on the 623 proteins that constituted this core. These proteins had been identified as proteins present in the majority of rodent or ungulate sperm proteomes, and the quantitative data were extracted from either the UniProt mouse protein database or the UniProt bovine database, depending on which target database yielded the more confident match. Thus, the core protein list was a hybrid of protein matches from either database, although the proteins had matched to entries in each (see Supplementary workbook W1). First, we analysed the separate protein lists that had matched most strongly to either the mouse or the bovine database. Analysis of gene ontology terms (level 3) revealed good representation of proteins associated with sexual reproduction (Fig. 9). Furthermore, the results from these two database searches were highly comparable.

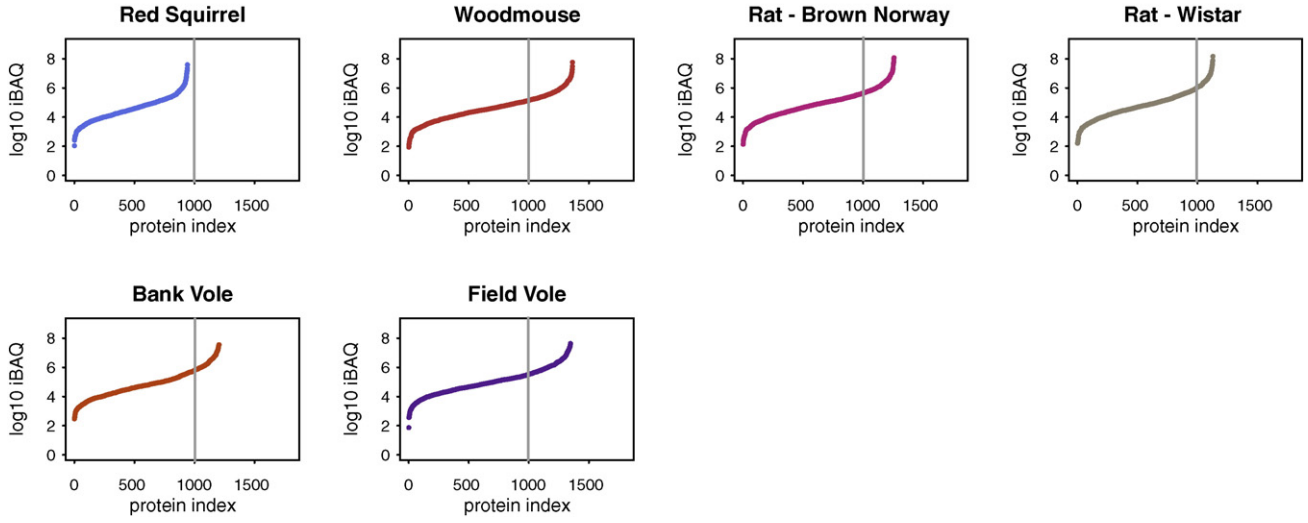
To analyse the core mammalian sperm proteome (623 proteins), we created a 'working' protein list by replacing the term '_BOVIN' with '_MOUSE' for any entries that had yielded the strongest match by reference to the bovine database. This was necessary as it was not possible to directly perform functional analysis on proteins that contained identifiers from two different species (mouse or bovine). This core list of the 623 'mouse' proteins was then used for further functional analysis, making use of the mouse annotation resources.

Fig. 7. Quantitative profiling of proteomes from different species. All ungulate samples were searched against the bovine database, and all rodent samples were searched against the mouse database, both at an FDR of 1%. Where biological replicates were obtained, these were averaged prior to plotting. Each sample was searched independently, and the protein abundances obtained by label free quantification were ranked and plotted in ascending order. In this plot, each point represents a single protein. All analyses are scaled to the same axis limits for number of proteins and for abundance, to aid comparison between samples. As a further guide, a vertical line at 1000 proteins identified has been superimposed in grey on each plot.

Ungulates, searched aganist bovine protein database



Rodents, searched aganist mouse protein database



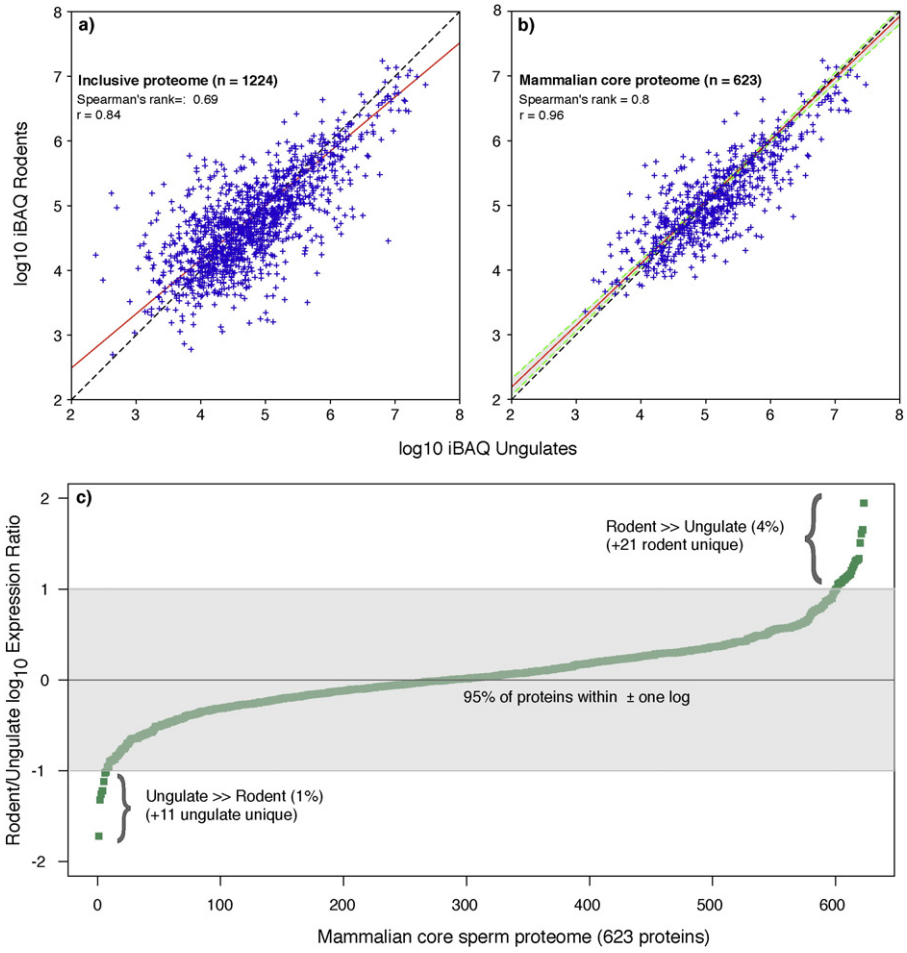


Fig. 8. Quantitative profiling of the 'core' sperm proteome. The rodent and ungulate samples were grouped and searched against the mouse or bovine database. From the search results, common groups of proteins were derived that were present in both rodent and ungulate groups, and from these, the iBAQ quantification values were correlated. Two analyses were conducted; the 'inclusive core' proteome (top panel) refers to proteins that were present in one or more of both rodents and ungulates whereas the 'frequent core' proteome (bottom panel) refers to proteins that were present in 5 or more out of the 6 rodent species/strains and 11 or more of the ungulate species.

Inevitably, this has to be seen as a compromise, and the resultant functional analysis is an overview of the functional competencies of the core proteome rather than a detailed and final analysis. Until fully annotated proteomes are available for each species, the necessity for such a cross-species process is unavoidable. However, comprehensive GO analysis was feasible.

First, we identified GO terms for 'Biological Processes' that were significantly enriched — these emphasised core metabolic processes and nucleotide and nucleoside metabolism (Fig. 10). We also examined the distribution of core GO terms at two hierarchical levels (Fig. 10; the analysis at all levels from 3 to 8 is provided in Supplementary Fig. S2). At level 3, there was a clear emphasis on proteins associated with reproduction, and at level 7, processes such as capacitation, zona pellucida binding and spermatid differentiation and development were represented by multiple proteins. In agreement with previous studies, the mammalian core sperm proteome reflects a considerable capacity for core metabolism to support sperm development and fertilisation [e.g. 10,19,42]. Additional network analyses are provided in Supplementary Figs. S3 and S4. The functional classification was clarified further by an analysis of the proteome using KEGG metabolic pathway terms, using the DAVID resource [43]. The most abundant groupings were characterised by reference to disease processes, but the term with the lowest p value was that of the proteasome (see below). However, strongly represented were multiple pathways of carbohydrate, lipid and amino acid metabolism, consistent with a high level of metabolic potential in the sperm cell (Fig. 11).

Further insight is gained by examination of specific groups of proteins. A recent paper [41] examined the mouse sperm proteome in caput, corpus and cauda epididymis and identified multiple components of the 20S proteasome and the 19S regulator complex that together constitute the 26S proteasome. The sperm proteasome has also been identified within the Rhesus macaque (*Macaca mulatta*) [19]. The 26S complex may play a role in capacitation and the acrosome reaction [44], and these sperm derived proteasomes degrade the zona pellucida prior to fertilisation [45]. We examined our mammalian data set for the proteins of the 20S and 19S complexes; these were relatively abundant proteins. In addition to all 14 20S subunits (PSA1–7, PSB1–7), 19 proteins from the 19S complex were also identified (PRS4, PRS6A, PSDE, PSMD1, PRS7, PSMD2, PSMD3, PSME1, PSMD6, PRS10, PRS6B, PSD11, PRS8, PSMD7, PSD12, PSD13, PSME2, PSMD4, PSMD8). These proteins were compared quantitatively, using label-free quantification (Fig. 12a). Most of the core 20S proteasome subunits clustered in a tight region at the upper end of the protein abundance distribution, although two proteins PSA7 and PSB7 seemed to yield lower abundance values, with greater discrepancy between the rodent and ungulate values. This may reflect a particular issue in label free quantification of those specific subunits, or a genuine deficit in those subunits. Although sperm have immune potential, there was no evidence for the 20S immunoproteasome subunits. Regulators

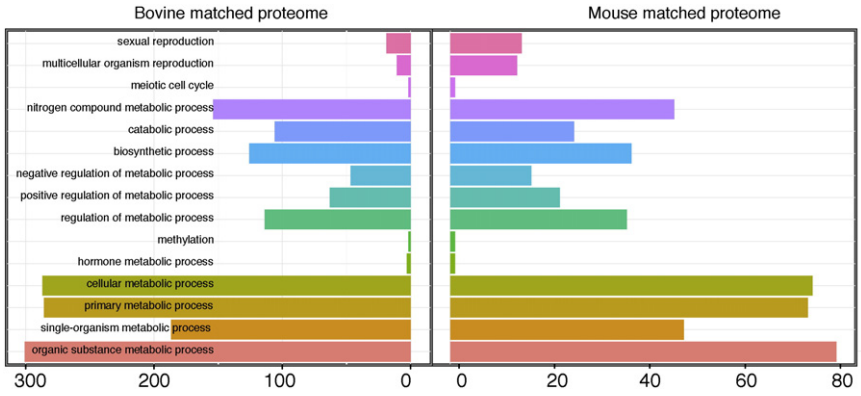


Fig. 9. Gene Ontology Biological Processes (GOBP) analysis of core proteomes. The frequencies of the top 15 level 3 GOBP terms were calculated for proteins that had matched most strongly to either the bovine (left) or mouse (right) database, as described in the main text. The values plotted are the frequency with which the level 3 term was assigned to an annotated protein in the list in each case.

of the proteasome were also detected well, and these proteins also clustered at about one order of magnitude lower abundance than the 20S core particle. Interestingly, both proteasome regulators PSME1 and PSME2 were detected. These proteins are involved in the regulation of immunoproteasome activity, and alter the peptide cleavage pattern of the 20S proteolytic complex. The presence of these proteins, combined with the failure to detect the immunoproteasome 20S subunits might imply a different role of the complex in sperm, conceivably generating specific peptides to assist sperm function.

We also examined a subset of proteins relevant to sperm maturation, structure and development (Fig. 12b). The outer dense fibre proteins (ODF) and A-kinase anchor proteins (AKAP) make up much of the sperm fibrous sheath [46,47]. Two outer dense fibre proteins (ODFP1 and 2) were amongst the highest abundance proteins in both ungulate and rodent sperm proteomes, a third member of the family (ODFP3A) was quantified at about 1% of the abundance of the other two proteins. However, the overall expression level was very comparable. The A-kinase anchor proteins (AKAP3 and AKAP4) were also highly abundant in both proteomes. Mammalian fertilisation relies upon free swimming sperm binding to the extracellular egg-coat — the zona pellucida [48]. Proteins within the zona pellucida have an elevated rate of evolution [26,27] and a role in taxon specificity of gamete recognition [49]; therefore sperm proteins that bind to the zona pellucida are likely to coevolve to maintain fertilisation ability. In accordance with this, 22% of genes encoding for sperm membrane proteins exhibit accelerated evolution [17].

Zonadhesin, sperm surface protein 17, and zona pellucida binding proteins assist with sperm binding to and penetration of the zona pellucida [50–53]. Sperm equatorial segment protein 1 and Izumo sperm–egg fusion protein allow the sperm to fuse to the oolemma, with the latter protein being required for this process to occur and so is essential to fertilisation [54,55]. These proteins displayed remarkable consistency of expression between the two groups of species, suggesting that even fertilisation-specific proteins retain a similar level of relative expression at the protein level (Fig.12b).

Spermadhesins are a family of ungulate-specific proteins that modulate the initial binding of sperm with the zona pellucida [53,56,57]. The ancestral spermadhesin gene region has been deleted from rodent genomes [58]. As predicted, although the spermadhesins are identifiable in the unfiltered data set (Supplementary workbook W1) they were not present in the 623 protein core proteome because of the absence of expression in the rodent proteome. We found little evidence for similar proteins being present within rodent sperm, however there was a single peptide match within the field vole and wood mouse samples — these are likely to be false positives and were present at very low intensities. This protein was identified within most ungulates tested, including the samples that contained very little sperm. It is therefore likely to be found within diverse ungulate species and not specific to the domestic species.

4. Conclusions

In this investigation, we explored whether standard proteomic workflows can be used to perform comparative proteomics of sperm. We chose sperm because of the presence of proteins that are under intense selection pressure and so could be anticipated to elicit poor matches in proteomic database searches against model organisms. Although we were able to use a mammalian protein database, this could be reduced to sub-set databases of ungulates and rodents, or even *Bos* and *Mus*, with only a small impact on the number of database matches. We were able to define between 700 and 1400 proteins within each sperm sample; however the sequence coverage of many of the matched proteins was low. Poor sequence coverage here may be due to low protein abundance, or the failure to achieve satisfactory cross-species matching in database searching. The requirement for precursor ion mass matching and product ion alignments has the outcome that only conserved regions of protein sequences can elicit matches. Nonetheless, we were able to construct an acceptable model of a core sperm proteome that could be used for quantitative comparisons, at least for the mapping of abundance of other proteins in relation to the core proteome. The core proteome proteins also have dN/dS statistics that are consistent with low rates of evolution, suggesting minimal sexual selection upon these proteins, whereas other, less frequently observed proteins are more highly evolving. It is also clear that extensive cross-species proteomics requires advanced tools that include peptide sequencing *de novo*. Sequencing techniques can recover the sequence of peptides without the necessity of a database match [59]. Indeed, *de novo* sequencing increases the accuracy of database searching when studying species for which protein sequence information is incomplete [60–62]. However, it is also feasible to embark on the application of RNAseq data to inform proteome construction [63], a relatively inexpensive route to construction of a workable proteome for database searches. If such data were available, there would be considerable interest in the use of our data set to explore the gains that are attainable with such approaches.

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.jprot.2015.12.027.

Transparency document

The Transparency document associated with this article can be found in the online version.

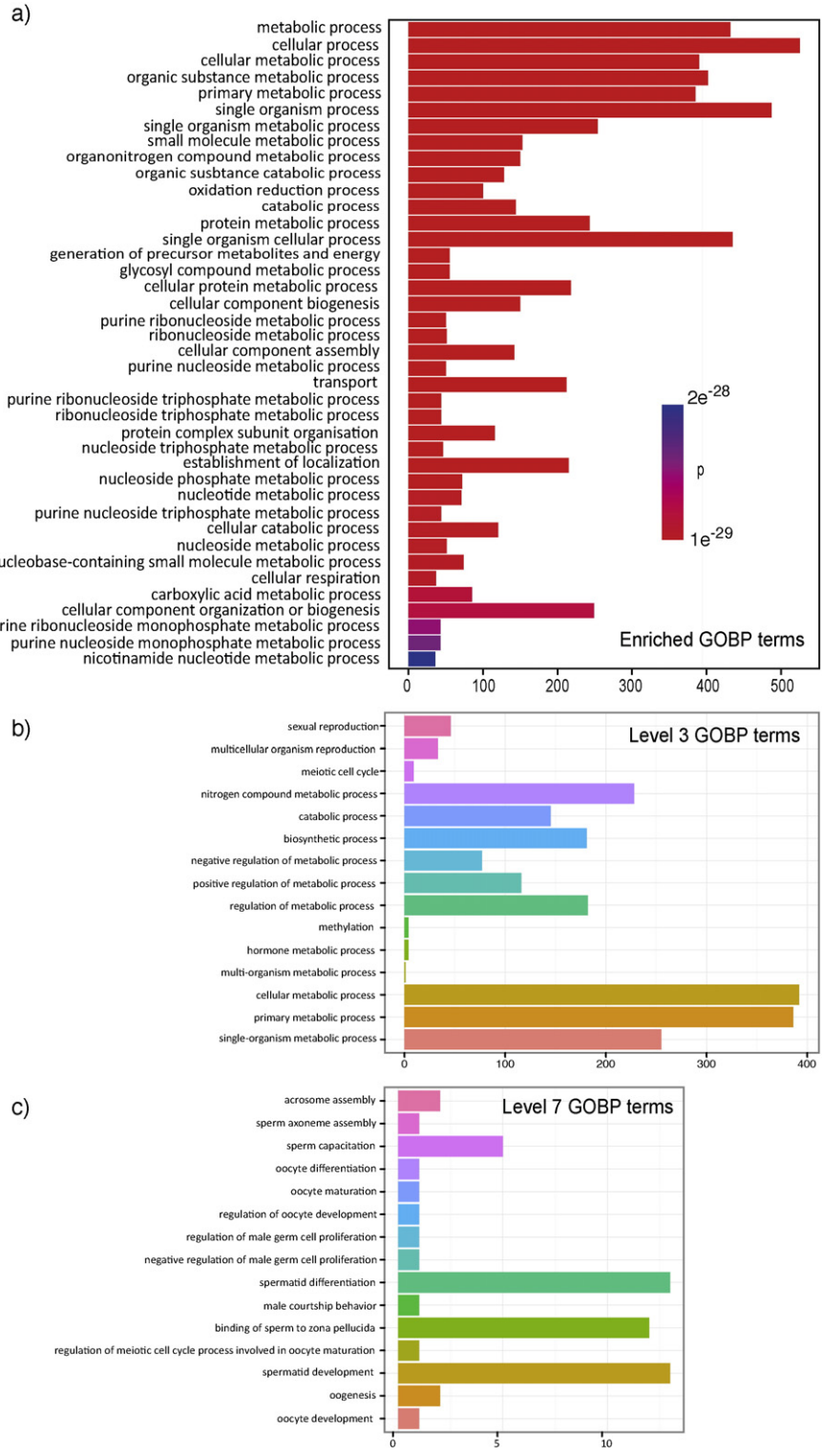


Fig. 10. Gene Ontology Biological Processes (GOBP) analysis of the mammalian core sperm proteome. The top 40 enriched GOBP terms, based on a hypergeometric distribution (panel a), within our 'inclusive core' sperm proteome (see main text for definition) together with the frequencies of the top 15 level 3 (panel b) or level 7 (panel c) GOBP terms assigned to the proteins in this list. The values plotted are the frequency with which the level 3 (or level 7) term was assigned to an annotated protein in the list in each case.

Acknowledgments

Samples were collected with the assistance of members of the Veterinary Pathology group, University of Liverpool; Ben Jones, Julian Chantrey, and Tim Dale. Thanks also to Steve Unwin, Olly O'Mally, Amy Campbell, E M & T Jackson, and Emily Logie for the additional assistance with sourcing samples. This study was funded by a BBSRC CASE award with Genus plc awarded to HB. The work was also supported by a research grants from the Natural Environment Research Council, UK (grant number NE/I013008/1).

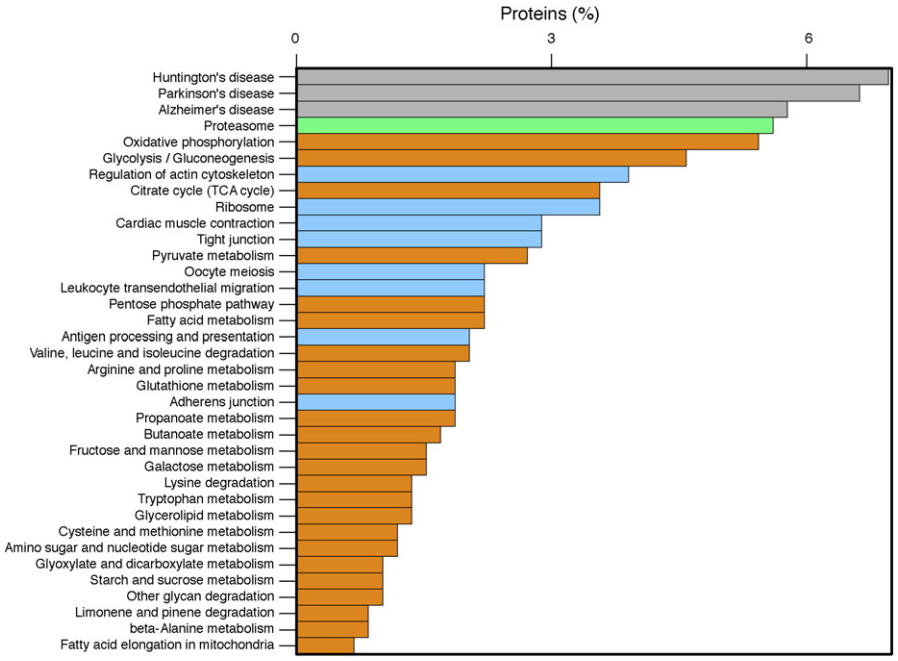


Fig. 11. Predominant KEGG metabolic pathway terms in the mammalian core sperm proteome. The mammalian core sperm proteome was analysed in terms of KEGG metabolic pathways using DAVID, which was able to access data for 590 of our 623 proteins. The most highly enriched KEGG groups were recovered, explaining 354 of the 623 proteins and the categories were ranked according to the number of proteins in each category — note that these do not rank with p-value (see Supplementary workbook W2). The top three terms are associated with disease processes (grey), metabolic functions are orange and the proteasome term is green.

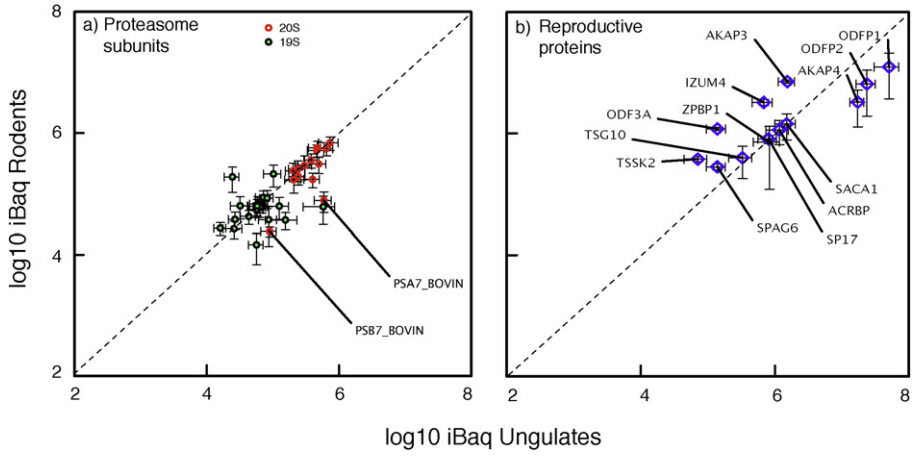


Fig. 12. Quantitative profiling of specific protein groups. The quantitative profile of the rodent and ungulate members of the core sperm proteome were highlighted and plotted as a scattergraph comparing the mean values (±SEM) for the proteins in either the rodent or ungulate members of the data set. Each data point is labelled with the UniProt ID of the protein. Panel a) 20S and 19S regulatory proteins of the proteasome, panel b) proteins directly associated with sperm function. The abundance data are plotted on the same scale as Fig. 7 for ease of comparison.

References

[1] D.B. Wilburn, W.J. Swanson, From molecules to mating: rapid evolution and biochemical studies of reproductive proteins, J. Proteome 135 (2016) 12–25.

[2] S.A. Ramm, L. McDonald, J.L. Hurst, R.J. Beynon, P. Stockley, Comparative proteomics reveals evidence for evolutionary diversification of rodent seminal fluid and its functional significance in sperm competition, Mol. Biol. Evol. 26 (1) (2009) 189–198.

[3] G.J. Arnold, T. Frohlich, Dynamic proteome signatures in gametes, embryos and their maternal environment, Reprod. Fertil. Dev. 23 (1) (2011) 81–93.

[4] P.C. Wright, J. Noirel, S.-Y. Ow, A. Fazeli, A review of current proteomics technologies with a survey on their widespread use in reproductive biology investigations, Theriogenology 77 (4) (2012) 738–765, e52.

[5] UniProt Consortium, Activities at the Universal Protein Resource (UniProt), Nucleic Acids Res. 42 (2014), D191–198 (Database issue).

[6] E.R. Wasbrough, S. Dorus, S. Hester, J. Howard-Murkin, et al., The *Drosophila melanogaster* sperm proteome-II (DmSP-II), J. Proteome 73 (11) (2010) 2171–2185.

[7] S. Dorus, S.A. Busby, U. Gerike, J. Shabanowitz, et al., Genomic and functional evolution of the *Drosophila melanogaster* sperm proteome, Nat. Genet. 38 (12) (2006) 1440–1445.

[8] T.L. Karr, Fruit flies and the sperm proteome, Hum. Mol. Genet. 16 (R2) (2007) R124–R133.

[9] M.A. Baker, L. Hetherington, G.M. Reeves, R.J. Aitken, The mouse sperm proteome characterized via IPG strip prefractionation and LC–MS/MS identification, Proteomics 8 (8) (2008) 1720–1730.

[10] A. Amaral, J. Castillo, J. Ramalho-Santos, R. Oliva, The combined human sperm proteome: cellular pathways and implications for basic and clinical science, Hum. Reprod. Update 20 (1) (2014) 40–62.

[11] M.A. Baker, R.J. Aitken, Proteomic insights into spermatozoa: critiques, comments and concerns, Expert. Rev. Proteomics 6 (6) (2009) 691–705.

[12] V. Poland, H. Eubel, M. King, C. Solheim, et al., Stored sperm differs from ejaculated sperm by proteome alterations associated with energy metabolism in the honeybee *Apis mellifera*, Mol. Ecol. 20 (12) (2011) 2643–2654.

[13] J.C. Wright, R.J. Beynon, S.J. Hubbard, Cross Species Proteomics, in: S.J. Hubbard, A.R. Jones (Eds.), Proteome Bioinformatics, Methods in Molecular Biology™, Humana

Press 2010, pp. 123–135 ([online] Available from: http://dx.doi.org/10.1007/978-1-60761-444-9_9).
[14] E.A. Hornett, C.W. Wheat, Quantitative RNA-Seq analysis in non-model species: assessing transcriptome assemblies as a scaffold and the utility of evolutionary divergent genomic reference species, BMC Genomics 13 (2012) 361.
[15] S.P. Gygi, Y. Rochon, B.R. Franza, R. Aebersold, Correlation between protein and mRNA abundance in yeast, Mol. Cell. Biol. 19 (3) (1999) 1720–1730.
[16] H. Herlyn, H. Zischler, The molecular evolution of sperm zonadhesin, Int. J. Dev. Biol. 52 (5–6) (2008) 781–790.
[17] S. Dorus, E.R. Wasbrough, J. Busby, E.C. Wilkin, T.L. Karr, Sperm proteomics reveals intensified selection on mouse sperm membrane and acrosome genes, Mol. Biol. Evol. 27 (6) (2010) 1235–1246.
[18] M.A. Baker, L. Hetherington, G. Reeves, J. Müller, R.J. Aitken, The rat sperm proteome characterized via IPG strip prefractionation and LC–MS/MS identification, Proteomics 8 (11) (2008) 2312–2321.
[19] S. Skerget, M. Rosenow, A. Polpitiya, K. Petritis, et al., The Rhesus macaque (*Macaca mulatta*) sperm proteome, Mol. Cell. Proteomics 12 (11) (2013) 3052–3067.
[20] R. Oliva, S.D. Mateo, J. Castillo, R. Azpiazu, et al., Methodological advances in sperm proteomics, Hum. Fertil. 13 (4) (2010) 263–267.
[21] E. Whittington, Q. Zhao, K. Borziak, J.R. Walters, S. Dorus, Characterisation of the *Manduca sexta* sperm proteome: genetic novelty underlying sperm composition in Lepidoptera, Insect Biochem. Mol. Biol. 62 (2015) 183–193.
[22] N.L. Clark, J.E. Aagaard, W.J. Swanson, Evolution of reproductive proteins from animals and plants, Reproduction 131 (1) (2006) 11–22.
[23] W.J. Swanson, V.D. Vacquier, The rapid evolution of reproductive proteins, Nat. Rev. Genet. 3 (2) (2002) 137–144.
[24] S. Dorus, S. Skerget, T.L. Karr, Proteomic discovery of diverse immunity molecules in mammalian spermatozoa, Syst. Biol. Reprod. Med. 58 (4) (2012) 218–228.
[25] S. Gavrilets, Rapid evolution of reproductive barriers driven by sexual conflict, Nature 403 (6772) (2000) 886–889.
[26] W.J. Swanson, Z. Yang, M.F. Wolfner, C.F. Aquadro, Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals, Proc. Natl. Acad. Sci. 98 (5) (2001) 2509–2514.
[27] L.M. Turner, H.E. Hoekstra, Adaptive evolution of fertilization proteins within a genus: variation in ZP2 and ZP3 in deer mice (*Peromyscus*), Mol. Biol. Evol. 23 (9) (2006) 1656–1669.
[28] S. Dorus, P.D. Evans, G.J. Wyckoff, S.S. Choi, B.T. Lahn, Rate of molecular evolution of the seminal protein gene SEMG2 correlates with levels of female promiscuity, Nat. Genet. 36 (12) (2004) 1326–1329.
[29] S.A. Ramm, P.L. Oliver, C.P. Ponting, P. Stockley, R.D. Emes, Sexual selection and the adaptive evolution of mammalian ejaculate proteins, Mol. Biol. Evol. 25 (1) (2008) 207–219.
[30] A.G. Clark, M.B. Eisen, D.R. Smith, C.M. Bergman, et al., Evolution of genes and genomes on the *Drosophila* phylogeny, Nature 450 (7167) (2007) 203–218.
[31] M.D. Dean, N.L. Clark, G.D. Findlay, R.C. Karn, et al., Proteomics and comparative genomic investigations reveal heterogeneity in evolutionary rate of male reproductive proteins in mice (*Mus domesticus*), Mol. Biol. Evol. 26 (8) (2009) 1733–1743.
[32] J.A. Vizcaíno, R.G. Côté, A. Csordas, J.A. Dianes, et al., The Proteomics Identifications (PRIDE) database and associated tools: status in 2013, Nucleic Acids Res. 41 (D1) (2013) D1063–D1069.
[33] J. Cox, N. Neuhauser, A. Michalski, R.A. Scheltema, et al., Andromeda: a peptide search engine integrated into the MaxQuant environment, J. Proteome Res. 10 (4) (2011) 1794–1805.
[34] J. Cox, M. Mann, MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification, Nat. Biotechnol. 26 (12) (2008) 1367–1372.
[35] B. Schwanhäusser, D. Busse, N. Li, G. Dittmar, et al., Global quantification of mammalian gene expression control, Nature 473 (7347) (2011) 337–342.
[36] D. Smedley, S. Haider, S. Durinck, L. Pandini, et al., The BioMart community portal: an innovative alternative to large, centralized data repositories, Nucleic Acids Res. 43 (W1) (2015) W589–W598.
[37] G. Yu, L.-G. Wang, Y. Han, Q.-Y. He, clusterProfiler: an R package for comparing biological themes among gene clusters, Omics J. Integr. Biol. 16 (5) (2012) 284–287.
[38] J. Reimand, T. Arak, J. Vilo, g:Profiler—a web server for functional interpretation of gene lists (2011 update), Nucleic Acids Res. (2011), gkr378.
[39] D. Merico, R. Isserlin, O. Stueker, A. Emili, G.D. Bader, Enrichment map: a network-based method for gene-set enrichment visualization and interpretation, PLoS One 5 (11) (2010), e13984.
[40] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, et al., Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Res. 13 (11) (2003) 2498–2504.
[41] S. Skerget, M.A. Rosenow, K. Petritis, T.L. Karr, Sperm proteome maturation in the mouse epididymis, PLoS One 10 (11) (2015), e0140650.
[42] A. Amaral, J. Castillo, J.M. Estanyol, J.L. Ballescà, et al., Human sperm tail proteome suggests new endogenous metabolic pathways, Mol. Cell. Proteomics 12 (2) (2013) 330–342.
[43] D.W. Huang, B.T. Sherman, R.A. Lempicki, Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources, Nat. Protoc. 4 (1) (2009) 44–57.
[44] S. Zimmerman, P. Sutovsky, The sperm proteasome during sperm capacitation and fertilization, J. Reprod. Immunol. 83 (1–2) (2009) 19–25.
[45] S.W. Zimmerman, G. Manandhar, Y.-J. Yi, S.K. Gupta, et al., Sperm proteasomes degrade sperm receptor on the egg zona pellucida during mammalian fertilization, PLoS One 6 (2) (2011), e17256.
[46] C. Petersen, L. Füzesi, S. Hoyer-Fender, Outer dense fibre proteins from human sperm tail: molecular cloning and expression analyses of two cDNA transcripts encoding proteins of approximately 70 kDa, Mol. Hum. Reprod. 5 (7) (1999) 627–635.
[47] P.R. Brown, K. Miki, D.B. Harper, E.M. Eddy, A-kinase anchoring protein 4 binding proteins in the fibrous sheath of the sperm flagellum, Biol. Reprod. 68 (6) (2003) 2241–2248.
[48] P.M. Wassarman, Mammalian fertilization: molecular aspects of gamete adhesion, exocytosis, and fusion, Cell 96 (2) (1999) 175–183.
[49] M.A. Avella, B. Baibakov, J. Dean, A single domain of the ZP2 zona pellucida protein mediates gamete recognition in mice and humans, J. Cell Biol. 205 (6) (2014) 801–809.
[50] H.R. Burkin, D.J. Miller, Zona pellucida protein binding ability of porcine sperm during epididymal maturation and the acrosome reaction, Dev. Biol. 222 (1) (2000) 99–109.
[51] Z. Gao, D.L. Garbers, Species diversity in the structure of zonadhesin, a sperm-specific membrane protein containing multiple cell adhesion molecule-like domains, J. Biol. Chem. 273 (6) (1998) 3415–3421.
[52] D.M. Hardy, D.L. Garbers, A sperm membrane protein that binds in a species-specific manner to the egg extracellular matrix is homologous to von Willebrand factor, J. Biol. Chem. 270 (44) (1995) 26025–26028.
[53] E. Topfer-Petersen, J.J. Calvete, Sperm-associated protein candidates for primary zona pellucida-binding molecules: structure–function correlations of boar spermadhesins, J. Reprod. Fertil. 50 (Supplement) (1996) 55–61.
[54] Y. Fujihara, M. Murakami, N. Inoue, Y. Satouh, et al., Sperm equatorial segment protein 1, SPESP1, is required for fully fertile sperm in mouse, J. Cell Sci. 123 (9) (2010) 1531–1536.
[55] N. Inoue, M. Ikawa, A. Isotani, M. Okabe, The immunoglobulin superfamily protein Izumo is required for sperm to fuse with eggs, Nature 434 (7030) (2005) 234–238.
[56] I. Caballero, J.M. Vazquez, M.A. Gil, J.J. Calvete, et al., Does seminal plasma PSP-I/PSP-II spermadhesin modulate the ability of boar spermatozoa to penetrate homologous oocytes in vitro? J. Androl. 25 (6) (2004) 1004–1012.
[57] G. Tedeschi, E. Oungre, M. Mortarino, A. Negri, et al., Purification and primary structure of a new bovine spermadhesin, Eur. J. Biochem. 267 (20) (2000) 6175–6179.
[58] B. Haase, C. Schlötterer, M.E. Hundrieser, H. Kuiper, et al., Evolution of the spermadhesin gene family, Gene 352 (2005) 20–29.
[59] V. Dancík, T.A. Addona, K.R. Clauser, J.E. Vath, P.A. Pevzner, De novo peptide sequencing via tandem mass spectrometry, J. Comput. Biol. 6 (3–4) (1999) 327–342.
[60] B. Ma, K. Zhang, C. Hendrie, C. Liang, et al., PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry, Rapid Commun. Mass Spectrom. 17 (20) (2003) 2337–2342.
[61] J. Zhang, L. Xin, B. Shan, W. Chen, et al., PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification, Mol. Cell. Proteomics 11 (4) (2012), M111.010587.
[62] B. Ma, G. Lajoie, De Novo Interpretation of Tandem Mass Spectra, Current Protocols in Bioinformatics/Editorial Board, Andreas D. Baxevanis. [et Al.] Chapter 13 p. Unit 13.10, 2009.
[63] V.C. Evans, G. Barker, K.J. Heesom, J. Fan, et al., De novo derivation of proteomes from transcriptomes for transcript and protein identification, Nat. Methods 9 (12) (2012) 1207–1211.