RNA-seq coupled to proteomic analysis reveals high sperm proteome variation between two closely related marine mussel species

Mónica R. Romero[1,2], Andrés Pérez-Figueroa[1], Mónica Carrera[3], Willie J. Swanson[4], David O. F. Skibinski[5], Angel P. Diz[1,2]*

[1]Department of Biochemistry, Genetics and Immunology, Faculty of Biology, University of Vigo, Vigo, Spain

[2]Marine Research Centre, University of Vigo (CIM-UVIGO), Isla de Toralla, Vigo, Spain

[3]Institute of Marine Research, CSIC, Vigo, Spain

[4]Department of Genome Sciences, School of Medicine, University of Washington, Seattle, USA

[5]Institute of Life Science, Swansea University Medical School, Swansea University, Swansea, UK

***Corresponding author**:

Dr. Angel P. Diz

Department of Biochemistry, Genetics and Immunology, Faculty of Biology, University of Vigo, 36310 Vigo, Spain

Tel: +34 986813828

E-mail address: angel.p.diz@uvigo.es

Abstract

Speciation mechanisms in marine organisms have attracted great interest because of the apparent lack of substantial barriers to genetic exchange in marine ecosystems. Marine mussels of the *Mytilus edulis* species complex provide a good model to study mechanisms underlying species formation. They hybridise extensively at many localities and both pre- and postzygotic isolating mechanisms may be operating. Mussels have external fertilisation and sperm cells should show specific adaptations for survival and successful fertilisation. Sperm thus represent key targets in investigations of the molecular mechanisms underlying reproductive isolation. We undertook a deep transcriptome sequencing (RNA-seq) of mature male gonads and a 2DE/MS-based proteome analysis of sperm from *Mytilus edulis* and *M. galloprovincialis* raised in a common environment. We provide evidence of extensive expression differences between the two mussel species, and general agreement between the transcriptomic and proteomic results in the direction of expression differences between species. Differential expression is marked for mitochondrial genes and for those involved in spermatogenesis, sperm motility, sperm-egg interactions, the acrosome reaction, sperm capacitation, ATP reserves and ROS production. Proteins and their corresponding genes might thus be good targets in further genomic analysis of reproductive barriers between these closely related species.

Keywords: Sperm, gonad, external fertilisation, marine invertebrates, reproductive isolation, speciation, proteomics, transcriptomics

Highlights

- *Mytilus* spp. are valuable in reproductive isolation and speciation studies.
- Gametes are key cell targets in investigations of speciation mechanisms.
- *Mytilus* spp. show proteome and transcriptome differences in male gonads and sperm.
- Identified proteins are involved in sperm motility and sperm-egg interactions.
- Joint proteomic and RNA-seq analysis provide candidate proteins for evolution studies.

Significance

Model systems for the study of fertilization include marine invertebrates with external fertilisation, such as abalones, sea urchins and mussels, because of the ease with which large quantities of gametes released into seawater can be collected after induced spawning. Unlike abalones and sea urchins, hybridisation has been reported between mussels of different *Mytilus* spp., which thus makes them very appealing for the study of reproductive isolation at both pre- and post-zygotic levels. There is a lack of empirical proteomic studies on sperm samples comparing different *Mytilus* species, which could help to advance this study. A comparative analysis of sperm proteomes across different taxa may provide important insights into the fundamental molecular processes and mechanisms involved in reproductive isolation. It might also contribute to a better understanding of sperm function and of the adaptive evolution of sperm proteins in different taxa. There is now growing evidence from genomics studies that multiple protein complexes and many individual proteins might have important functions in sperm biology and the fertilisation process. From an applied perspective, the identification of sperm-specific proteins could also contribute to the improved understanding of fertility problems and as targets for fertility control.

1. Introduction

The study of the mechanisms that lead to the formation of new species is of special interest in marine ecosystems due to the lack of obvious barriers to gene flow, and is especially relevant in organisms with a prolonged period of larval dispersion [1]. Many marine species release gametes into seawater, so fertilization occurs externally. Because of this, research on speciation in marine systems has focused on the evolution of gamete recognition systems because of their potential as prezygotic reproductive isolation mechanisms [2-4]. The role of postzygotic mechanisms has been less studied and is controversial [5] despite their potential relevance to maintain the integrity of species [6]. It seems obvious that gametes are key cell targets in investigations of the molecular mechanisms underlying reproductive isolation. Molecular studies on gametes are however quite scarce and largely restricted to a few model organisms. The molecular basis of fertilisation including the sperm-egg recognition system is still a poorly understood, yet basic, biological process [7-8]. In marine invertebrates such studies have focused on sea urchins, starfish, clams, oysters, abalones, sea snails and worms [8-9]. The use of a greater diversity of species has recently been advocated as a good way to shed light on diverse questions that remain open in reproductive biology [10], including the molecular basis of species-specificity gamete interactions during fertilisation.

Sperm are highly differentiated cells with marked genetic, cellular and functional differences from other cell types, reflecting important roles in fertilization, embryonic development, and heredity [11]. The sperm cell has also been put forward as an ideal candidate for proteomic analyses [12], mainly because it is thought to be transcriptionally inert (but see [13]). So far only a few proteomics studies have focussed on sperm cells, mostly in widely studied model organisms (see [4, 14]). The ascidian *Ciona intestinales* [15], the red abalone *Haliotis rufescens* [16], the Pacific oyster *Crassostrea gigas* [17], the king scallop *Pecten maximus* [18] and the marine mussels *Mytilus edulis* [19-20] and *M. galloprovincialis* [21], are the only marine organisms, all of them external fertilisers, currently in the sperm cell proteomic literature. Furthermore to the best of our knowledge, there are no comparative quantitative proteomic studies of sperm of closely related species, with the exception of an analysis of different ungulate and rodent species [14, 22]. A comparative research strategy involving proteomics

should contribute towards elucidating the molecular basis underlying reproductive isolation mechanisms and the evolutionary forces involved, as well as to obtaining a better understanding of basic functional aspects of sperm biology at the molecular level.

Marine mussels from the *Mytilus edulis* complex are represented by three closely related species (*Mytilus edulis, M. galloprovincialis* and *M. trossulus*) that are able to hybridise at some rocky shore areas where their distributions overlap [23]. Hence, mussels represent a good model to address evolutionary hypotheses and study mechanisms underlying the formation of new species. On European coasts, *M. edulis* has a more northerly and *M. galloprovincialis* a more southerly distribution, while *M. trossulus* is mainly restricted to the Baltic Sea area. There are many localities where hybridisation and variable levels of genome introgression occur between the species. Research on *Mytilus* spp. has also attracted attention because of the important mussel aquaculture industry. Marine mussels are external fertilisers with a prolonged planktonic larval stage facilitating dispersal over great distances [24]. In order to preserve their genome integrity, despite extensive hybridisation, different reproductive mechanisms are likely to be operating both at the pre- and postzygotic level, though their relative contribution and underlying molecular mechanisms are not yet well understood. Cross-species fertilisation in *Mytilus* might be prevented to some degree by molecular incompatibilities resulting from the rapid evolution of reproductive proteins. Evidence for positive selection on M7 and M3 sperm lysin protein was provided for sympatric and allopatric populations of *Mytilus* spp. [25-28]. However prezygotic barriers might not be strong enough to prevent introgression due to extensive hybrid zones and wide variation in the genomic introgression rates observed in natural populations [29]. Weaknesses of prezygotic barriers are also suggested by contrasting results from interspecific crosses under laboratory conditions between *Mytilus* spp. [30-36].

The arrival of high-throughput genomics and proteomics techniques is allowing the expansion of classical evolutionary studies over large protein datasets [37]. Despite this advance, less attention is still paid in evolutionary ecology studies to the proteome as compared to the transcriptome or genome, even though the proteome is closer to the molecular phenotype, and thus a more direct target for natural selection [38-40]. The choice of reproductive tissues or gametes as the main focus of research helps to bridge the gap between reproductive phenotypes and underlying molecular mechanisms [37,

41]. A 2-DE based proteomic study using a somatic tissue, the foot, from two sympatric *Mytilus* species (*M. edulis* and *M. galloprovincialis*) and their hybrids showed differences in the protein expression patterns of hybrids when compared with the two parental species, providing evidence compatible with Dobzhansky-Muller incompatibilities (DMI) between both parental genomes in hybrids [42]. Thus postzygotic isolation factors may also have played a role in limiting the degree of introgression among genomes of *Mytilus* spp. New studies using high throughput genomics and proteomics on gametes should provide a significantly better understanding of the molecular mechanisms underlying reproductive isolation and evolution of *Mytilus* spp.

A good strategy when working with less well studied organisms to significantly boost the number and quality of protein identifications obtained through mass spectrometry analysis is to generate a customised protein database, for example through the translation of tissue and species-specific transcriptome datasets available in public databases or obtained from in-house experiments [37]. An additional resource for mussels is a recently published *M. galloprovincialis* genome [43]. However the availability of protein databases derived from transcriptomes provides a useful and complementary tool because of known limitations in the prediction and annotation of genes and posttranscriptional variants [44]. Moreover the combined use of transcriptomic and proteomic data specifically in non-model organisms has been advocated as one of the most useful proteogenomic approaches [45-46], because of its high and proven potential for synergy between the two approaches.

In this study we undertook a deep transcriptome sequencing (RNA-seq) of mature male gonads obtained from *Mytilus edulis* and *M. galloprovincialis* individuals acclimatised for several weeks to common laboratory conditions after collection from their native localities. The results from this study contribute to, 1) providing a tissue *Mytilus*-specific protein database to enhance protein identifications in follow-up proteomic analyses, and 2) providing a preliminary list of candidate gene products with potential involvement in sperm biology, fertilisation and reproductive isolation mechanisms in the two *Mytilus* species. A second complementary analysis based on a 2-DE+MS/MS proteomic approach, with the use of different customised protein databases, including one derived from our transcriptome data, to enhance protein identification, was carried

out directly on sperm samples. This was to assess whether sperm samples from the same two *Mytilus* species and populations, that were acclimatised to common laboratory conditions for several months, presented proteomic differences which would be a consequence of underlying genetic differences between the populations and species. The level of concordance of differential expression results between transcriptome and proteome data is evaluated, while the functional consequence of the observed variation is discussed from an evolutionary perspective in relation to sperm biology, and the potential role of the variation in fertilisation and reproductive isolation.

2. Materials and Methods

Extended versions of Material and Methods for RNA-seq and proteomic analysis are provided in Ref. [47] and File S1 respectively.

2.1. Transcriptome (RNA-seq) analysis of mature male gonad tissues from two *Mytilus* spp.

2.1.1. Sampling and histological analysis

Mussels from *Mytilus edulis* and *Mytilus galloprovincialis* species were collected from rocky shores in Swansea (South Wales, UK) and Ria de Vigo (North-West Spain) respectively during the end of January of 2012, transported to aquarium facilities in the marine station at the University of Vigo (ECIMAT), and kept there in seawater under the same conditions for at least 2 months. This design ensured that all analysed individuals shared the same environmental conditions, and that gene expression differences between species were not therefore the results of differences in the immediate environment [48]. After 2 months, mussels from each species were processed individually. From each mussel, one piece of gonad tissue was immediately snap frozen and preserved in liquid nitrogen for further RNA-seq analysis, while a second piece of the same tissue was used for a histological test to assess the sex and reproductive stage of the mussel. For this purpose gonad tissues were fixed in Davidson´s solution and embedded in paraffin. Paraffin blocks were sectioned at 5µm with a microtome. Tissue sections were deparaffinised, stained with Harris´ hematoxylin and eosin, and examined by light microscopy for a histological study. Finally, 6 individual samples from each

Mytilus species corresponding to reproductively mature male individuals were chosen for RNA extraction (Figure 1).

2.1.2. RNA extraction, mRNA library and Illumina paired-end sequencing

RNA extraction was carried out using a protocol based on the Qiagen RNeasy® Mini kit (Qiagen, Valencia, CA, USA) with tissue homogenization in QIAshredder columns (Qiagen). The quantification of RNA samples was carried out using a NanoDrop 1000 Spectrophotometer (Thermo scientific, DE, USA), and the RNA quality was assessed in an Agilent 2100 bioanalyzer (Agilent Technologies, CA, USA). Total RNA extracts from these selected samples were used to make two pools of 6 individuals each, one pool for each of the two *Mytilus* species. 700 ng of RNA per individual sample was used, so each pool contained 4.2 µg of total RNA. mRNA libraries were generated using the Illumina Truseq Small RNA Preparation kit (Illumina, CA, USA) according to Illumina's TruSeq Small RNA Sample Preparation Guide v2 (low sample protocol). Agarose gel-based selection was carried out to obtain libraries with fragments close to 500 bp in length, and their quality was assessed through Bioanalyzer profiles using a high sensitivity DNA chip. Finally, libraries were quantified, by using quantitative PCR with specific primers complementary to the library adapters and KAPA SYBR FAST Universal qPCR Kit (Kapa Biosystems, MA, USA), and diluted to 12 pM before sequencing. Each library, corresponding to each of the two pools, was analysed in a full line of the flow cell from an Illumina HiScanSQ instrument (Illumina) and using TruSeq SBS v3 chemistry (Illumina) to generate 2 × 100 bases long paired-end reads. After sequencing, data were acquired and analysed by using the Genome Analyzer Sequencing Control Software (SCS 2.6) and Real Time Analyser (RTA 1.6) software from Illumina. A total of 124,102,082 and 111,865,458 raw reads were obtained from the *Mytilus edulis* and *Mytilus galloprovincialis* pooled samples respectively. Raw data were deposited into SRA-NCBI database (BioProject ID: PRJNA451093). The quality control and filtering of nucleotide sequences was carried out as explained in Ref. [47], yielding 187,829,361 confident reads that were used for *de novo* assembly and generation of a consensus transcriptome.

2.1.3. De novo transcriptome assembly and functional annotation

Due to absence of a complete *Mytilus* spp. genome sequence (but see a recently published low-coverage *M. galloprovincialis* genome in [43]), it was necessary to

follow a *de novo* assembly approach in order to build a consensus transcriptome from mature male gonad from both *Mytilus* spp. Thus, reads from both *Mytilus* species were assembled to generate a set of contigs (herein isotigs). The full set of isotigs should represent the majority of transcribed genes in this specific tissue in either one or both *Mytilus* species. This approach allowed the comparison of the expression levels from the different isotigs between samples of the two species. *De novo* transcriptome assembly was carried out by using Velvet followed by Oases software [49-50]. Oases uses the preliminary assembly made by Velvet to complete the assembling of reads into isotigs. Finally, it clusters the isotigs into small groups called loci (synonymous with the term isogroups, also used in the literature), representing the consensus transcriptome of the samples under study. These are not genetic loci, but rather a collection of similar sequences (isotigs), which might include different splice variants, alleles and partial assemblies of longer transcripts. Hence, it might be said that there are different isotigs for each locus (consensus transcript). Nevertheless, many loci contain only one isotig, though some others may contain hundreds of isotigs. The generated consensus transcriptome was annotated against a non-redundant UniProtKB/SwissProt sequence database using the program BlastX [51]. For comparative purposes the annotation was repeated against the published genome of another marine bivalve the Pacific oyster *Crassostrea gigas* [52], against all EST sequences available in NCBI from "Mytilus"[organism], and against two protein databases with sequences retrieved from NCBI either for "Mytilus"[Organism] or "Mollusca"[Organism] using a threshold e-value of $1x10^{-3}$. Functional annotation based on Gene Ontology (GO) terms was performed using the tool Blast2GO [53]. An enrichment analysis of GO terms was carried out for those transcripts that showed significant differences between samples of the two *Mytilus* spp. (see below) using Fisher's exact test with a FDR=5% (see Ref. [47] for further details on method). This might provide some clues about the differences at functional level present in mature male gonad tissue of the two *Mytilus* spp.

2.1.4. Differential expression analyses

In the present study, differential gene expression analysis from mature male gonad tissue (pooled samples) between *Mytilus edulis* and *M. galloprovincialis* was carried using the RNA-seq data at isotig level. In circumstances where one biological replicate is available for each treatment group, methods based on the Negative Binomial (NB)

distribution [54] can be used to make inferences about differential expression between the *Mytilus* species and identify isotigs with higher effect-size. These changes could be supported in complementary studies, for instance by proteomic analysis with an appropriate biological replication (see section 2.2). The pooling approach met the requirements to fulfil one of the main objectives of the current study. This is to generate a tissue-specific *Mytilus* protein database from a high coverage reference transcriptome of both species in order to increase the success of protein identifications in proteomic analysis on sperm cells (see section 2.2). RSEM [55] combined with EBSeq [56] software were used to calculate differential expression (p<0.05, FDR=5%). This pipeline is appropriate in situations where a reference genome is not available, enabling accurate transcript quantification after transcriptomic *de novo* assembly [55], while controlling the false discovery rate (FDR) [57]. Functional annotation and an enrichment analysis for those differentially expressed transcripts was carried out as explained in the above section 2.1.3 and Ref. [47].

2.2. Proteomic analyses of sperm samples from two *Mytilus* spp.

2.2.1. Sampling of mussels and sperm sample collection

Mussels from *Mytilus edulis* and *Mytilus galloprovincialis* species were collected from rocky shores in Swansea (South Wales, UK) and Ria de Vigo (North-West Spain) respectively at different times within the spawning period (end of January and April) in 2012, transported and kept under as far as possible the same laboratory conditions for at least 2 months, in order to minimize the differences between mussel species due to immediate environmental effects (see [48]). After 2 months, mussels were periodically induced to spawn following a thermal shock procedure (see detail in File S1). Sperm samples released into filtered/UV-treated seawater in individual bottles were collected, filtered twice (300 µm and 41 µm sieves), and centrifuged for 10 min at 24400 g, 10ºC. After discarding the supernatant, the pellet containing sperm was resuspended in 150 µl of a 10% glycerol solution, snap frozen in liquid nitrogen, and finally preserved at -80ºC until further analysis. In parallel, a drop of seawater for each sample containing sperm cells was examined under the microscope in order to check that the sperm presented good morphology, high motility and density, otherwise the sample was discarded for any further analysis.

2.2.2. Protein extraction and 2-DE electrophoresis

Proteins were extracted from sperm samples of the two *Mytilus* spp. (10 biological replicates for each *Mytilus* spp. Two of them were run twice) in 0.3-0.5 ml of lysis buffer (7M urea, 2M thiourea, 4% CHAPS, 1% DTT and 1% carrier ampholytes 3-10) aided by sonication on ice (Branson Digital Sonifier 250, CT, USA). After centrifugation for 30 min at 21,000g, at 10^{o}C, the supernatant was stored at -80ºC until electrophoresis. Protein concentration was measured with the Bradford method [58]. Approximately 200 µg of total protein was used for 2-DE. The first dimension electrophoresis was carried out with immobilized pH gradient strips (pH 5-8/17cm, Bio-Rad) in a horizontal electrophoresis apparatus Protean IF System (BioRad) after strip equilibration. The second dimension of gel electrophoresis was carried out in 12.5 % polyacrylamide gels using an EttanDaltsix electrophoresis system (GE Healthcare, Little Chalfont, UK) at 20ºC, 15W/gel, and ~ 6h. Protein spots were visualized using SYPRO-Ruby (Molecular Probes, OR, USA), following the protocol described in [48]. Stained gels were scanned with a Pharox FX Plus molecular imager (BioRad), and 2-DE gel images saved in TIFF file format. The SameSpots vs.4.1 (Nonlinear Dynamics Ltd, Newcastle upon Tyne, UK) software was used for 2-DE gel image and protein spot detection analysis (including background subtraction and normalisation) following the same procedure described in [59]. Normalised protein spot volumes for each 2-DE gel were saved in csv file format for further statistical analyses.

2.2.3. Statistical analyses of 2-DE gels

Normalised spot volumes were transformed to a logarithmic scale to fit normality and homoscedasticity assumptions of parametric tests [42]. Spearman's correlation coefficient and coefficient of variation (*CV*) calculations were carried out using the whole protein spot dataset from technical replicates, aiming to assess the experimental reproducibility. Analysis of variance (one-way ANOVA) using the log normalised volume of each protein spot (dependent variable) was carried out to test for significant differences in protein expression patterns in sperms cells of the two *Mytilus* spp., where biological replicates were used to provide the error variance in the analysis. Different corrections to account for the multiple hypothesis testing problem were calculated by using the SGoF+ software v.3.8 [60], thus following the procedure and rationale discussed in Ref. [61]. Heat map analysis was used to group protein spots and individual samples according to their similarity in expression pattern. The heat map and hierarchical clustering analyses were conducted with the R package gplots [62], using

Euclidean distance and the complete linkage method. Chi-square contingency tests were used to compare distributions of ontology terms for the protein spot identification and RNA-seq results, with significance levels determined by bootstrapping using FORTRAN programs written for this purpose and which allow for test of significance of individual rows in contingency tables.

2.2.4. Mass spectrometry analysis and protein identification

The protein spots of interest were visualized on a blue-light DarkReader (Clare Chemical Research, CO, USA), excised and processed following the protocol described in Ref. [48]. Resulting peptides were analyzed in an Orbitrap Elite mass spectrometer coupled to a Proxeon EASY-nLC 1000 UHPLC system (Thermo Fisher, San Jose CA). Peptide separation was performed on RP columns (EASY-Spray column, 50 cm x 75 µm ID, PepMap C18, 2 µm particles, 100 Å pore size, Thermo Scientific) using a 120 min linear gradient from 5 to 25 % of acetonitrile at a flow rate of 300 nL/min. For ionization, the spray voltage used was 1.95 kV, the capillary temperature was 260ºC and the Orbitrap set at 120,000 resolution. A positive mode from 400 to 1,700 amu (1 µscan), 15 data dependent CID MS/MS scans using an isolation window of 2 amu and a normalized collision energy of 35%, with a dynamic exclusion for 80s after the fragmentation event, were used for peptide analysis. Singly charged ions were excluded from MS/MS analysis. MS/MS spectra were searched using PEAKS Studio v.7.0 program (Bioinformatics Solutions Inc., Waterloo, ON, Canada) against three customized protein databases. Databases were made from the tissue and *Mytilus*-specific RNA-seq data provided in this study, EST sequences available in NCBI for four *Mytilus* species retrieved using "Mytilus"[organism] as search term, and protein sequences deposited in NCBInr for "*Mollusca*" [organism] (see further detail in File S1). Positive protein identifications (FDR <1%) were only accepted when at least two matched and one unique peptide sequences were obtained. BlastX analyses against a non-redundant (nr) protein sequence database of all organisms were carried out in order to ascertain the final protein identities of translated EST and RNA-seq sequences using default parameters and a threshold e-value of $1x10^{-6}$.

3. Results

3.1. Transcriptome (RNA-seq) analysis of mature male gonad tissues from *Mytilus edulis* and *M. galloprovincialis*

3.1.1. De novo assembly and Blast analyses of the consensus transcriptome from both Mytilus spp.

RNA-seq analyses of the two pooled samples from mature male gonad tissues, one from *Mytilus edulis* and one from *M. galloprovincialis*, produced more of 200 million 100bp paired-end reads. After filtering steps, more than 187 million reads remained valid to be used for *de novo* assembly, hence the generation of a consensus transcriptome for both *Mytilus* spp. (Table 1). *De novo* assembly produced a total of 97,425 isotigs, grouped in 49,713 loci (see Files S1-S2 in Ref. [47]). Thus a consensus transcriptome for mature male gonads of the two *Mytilus* species was obtained. This provides a reference transcriptome to which individual reads from each pooled sample could be mapped in differential expression analysis. Moreover it provides a tissue and *Mytilus*-specific database that, once translated to six-reading frames, can be used for protein identification in the proteomic studies carried out on sperm samples (see section 3.2.2). The mean (median), maximum and N50 length of isotigs is 706 (434), 13,604 and 1,071 nucleotides, respectively (Table 1). The estimated size calculated for the consensus transcriptome of both *Mytilus* spp. is 35.1 Mb. The redundancy level found for the transcriptome assembly was low (1.5% of loci). Results from Blast analysis against different databases (see Materials and Methods, and Figure 4 in Ref. [47]) are summarised in Table 1. A total of 13,498 sequences (27.2% of total loci) were successfully identified against a non-redundant UniProtKB/SwissProt database. This moderate to low similarity with the database may be due to potential novel genes (or variants) in these two species, whose full genomes had not been sequenced at the time of elaborating this paper. This is supported by the following results. When Blast analysis was carried out against the published and annotated oyster (*C. gigas*) genome [52], another marine bivalve mollusc, the number of positive identifications rose to 18,279 transcripts (36.8%). The relatively modest increase in identifications may be due to the long divergence time between *Mytilus* and *C. gigas* even though they belong to the same phylum and class. This percentage is in line with the identification success (17,529 transcripts, 35.3%) and database coverage (% of sequences from NCBI database giving positive match against our transcriptome) obtained from Blast analysis against protein sequences from Molluscs retrieved from NCBI (Table 1). Despite the

low number of protein sequences for *Mytilus* spp. available in protein databases, the Blast analysis showed, as expected, a level of coverage for a protein sequence database (*Mytilus*[organism], NCBInr) of 81.3%. A similar result, a database coverage of 82.7%, was obtained after Blast analyses against all EST sequences available in NCBI for *Mytilus*[organism] that were translated to proteins by using the six-reading frames. Although the redundancy level of these EST sequences is high, the number of sequences is high so it is not surprising to see that a positive match/identification was reached for 31,428 (63.2%) of loci from our consensus transcriptome.

3.1.2. Functional annotation of the consensus transcriptome from both Mytilus spp.

From functional analysis using Blast2GO, 12,156 loci were successfully annotated for GO terms (File S3 in Ref. [47]). The annotation was improved after InterProScan analysis, raising the number of successful annotations to 13,283 loci (File S4 in Ref. [47]). This might be interesting because functional information, *e.g.* a peptide signal sequence from the differential expressed sequences between *Mytilus* spp., is still reported despite the inability to get a confident gene/transcript identity during BlastX analysis. The distribution of GO-terms for the full annotated transcriptome at different levels, molecular function (MF), biological process (BP) and cellular component (CC) categories, is displayed in Figure 2a. It is reassuring to see that "reproduction" term is represented in BP category. The dominance of "binding", a general term related to the non-covalent union or interaction of different molecules, in MF is also interesting because when checking MF terms for the more specific tree hierarchy level 3 (Figure 5 in Ref. [47]), the highest representation is for protein binding, a term related to interactions among proteins or protein complexes. This category should include sperm proteins involved in sperm-egg interaction. Finally it is interesting to highlight in category CC, in both Figure 2a and Figure 5 in Ref. [47], the high representation for terms related to membrane proteins that potentially include those that might be involved in the sperm-egg recognition mechanisms.

3.1.3. Differential expression analysis between Mytilus edulis and M. galloprovincialis.

A total of 27,233 isotigs (28% of the 97,425 occurring in the transcript assembly) are differentially expressed between pooled samples of the two *Mytilus* spp. at FDR 5%, of which 20,997 (21.6%) are significant at FDR 1%. This corresponds to 14,737 loci (29.6% of 49,713 loci in the transcript assembly) which are significant (in that they

have at least one significant isotig) at FDR 5% of which 11,335 (22.8%) are significant at FDR 1%. Files S5 and S6 in Ref. [47] contain expression and statistical values from this analysis. File S7 in Ref. [47] contains the annotation based on BlastX (see section above 2.1.3) for all transcripts (loci) where a significant differential expression result was found. A total of 4338 (4223 at FDR 1%) differentially expressed loci were successfully annotated after Blast2GO including InterProScan 5.0 [63] analysis. The most relevant result of the GO term enrichment analysis in relation to this study is an overrepresentation of the BP term "reproduction" (Figure 2b). These loci form the main analytical focus in this paper. To pursue this, we chose those functional annotated loci (a total of 309 of the 4338 in total that are differentially expressed) that code for proteins specifically related to fertilisation and sperm biology processes. From these, 61 loci corresponding to 50 different proteins are shortlisted based on the prediction that they have signal peptide or transmembrane domains by using SignalP 4.1 [64] and TMHMM 2.0 [65] servers, available in CGS Technical University of Denmark, respectively, and complemented with results from InterProScan 5.0 analysis described above (Table 2). These types of domains indicate that protein can be either secreted (*e.g.*, present in the sperm acrosomal content) or located in the sperm plasma membrane respectively, hence with high potential to play a role in the sperm-egg recognition system or gamete fusion [16]. We thus wish to specifically focus on these as good candidates for more detailed consideration and perhaps future study. These candidate loci (Table 2) code for proteins that are mainly involved in different steps of spermatogenesis (Cdyl2, Ggnbp2, Nphp1, Rarb, Irs, Iap2, Tmbim6, eif4g2, CtsB, CtsL, CtsL2, Prdm9, Suv39h2), sperm motility (Dnal1, Ropn1, Ift172, Slc26, Slc6a5, Slc9c1), binding of sperm to the egg vitelline coat (Cct2, Cct3, Cct4, Cct5, Cct6a, Cct7, Cct8, Psma2, Ubc8, Pc1, Hya, Spag1, Thbs1, Zan, vitelline coat lysins M3 and M6), acrosome reaction and sperm capacitation (Cdc42, Spa17, CtsB). For each of the above candidate genes (loci), in some cases, isotigs within a locus varied in the nature and extent of differential expression between the two *Mytilus* species, see final two columns in Table 2. The expression differences could have resulted from simple allele differences between the mussels making up the pools, or more complex alternative splicing events producing different protein isoforms in the two species. It also might be the result of differential regulation of expression of the same protein isoform in the mature male gonad of the two different *Mytilus* species. It is important to note that allele differences can have two main different effects at the

molecular phenotype level, either changing the mRNA/protein sequence or acting as expression modifiers. The latter effect can be associated with changes in non-coding usually cis-regulatory regions, though getting direct evidence for this is rather difficult [66].

3.2. Proteomic analysis of sperm cells from *Mytilus edulis* and *M. galloprovincialis*

3.2.1. Two-dimensional electrophoresis (2DE) and differential expression analyses

After applying the quality filter based on comparisons made for each 2DE gel against a pre-defined "gold standard 2D gel", a tool implemented in SameSpots software, two out of ten 2DE gels of sperm samples analysed from the Swansea population (*M. edulis*) were removed from further analysis, while all 2DE gel samples from Vigo population (*M. galloprovincialis*) successfully passed this pre-defined filter (File S2). The analysis of the 2DE gel images produced a final dataset of 727 protein spots (File S3). Results from the reproducibility experiment, where two sperm samples one from each species were analysed twice, permitted the comparison of technical and biological variation. For each of the 727 spots the CV of spot volume was calculated over 10 biological replicates for *M. galloprovincialis* and over 8 biological replicates for *M. edulis*. The technical variation was measured for each species from the sample of two technical replicates for each species. The spot-specific CV values averaged over both spots and species are 41.2 ± 0.29 (SE) and 19.0 ± 0.34 for biological and technical variation respectively. Because of the small number of technical replicates, nonparametric tests were further used to gauge the significance of this difference. Thus of the 727 spots, 638 and 611 had higher CV for biological than technical replication in *M. galloprovincialis* and *M. edulis* respectively. $\chi 2$ tests against a 1:1 expectation were made where the null hypothesis is that higher CV is equally likely for biological and technical replicates. The expected frequencies in each category are thus 363.5:363.5. The $\chi 2$ value is highly significant in each species, even a ratio of 408:319 would be significant at $p<0.001$. Even if spot volume values are not independent for some pairs or groups of spots, this test is highly suggestive of significantly greater CV for biological than technical replicates. In a further test the Spearman correlation was computed over spots between technical replicates within each species. The values are 0.953 and 0.927 for *M. galloprovincialis* and *M. edulis* respectively. The corresponding correlation values between biological replicates vary between 0.767 and 0.895 for *M. galloprovincialis*

and 0.780 and 0.896 for *M. edulis*. Both tests confirm that spot volumes are much more different between biological than technical replicates providing clear evidence of biological signal within each species.

One-way ANOVA ("Species"; fixed factor) for each spot resulted in 17.6% of the protein spots showing significant differences (*a priori* $p<0.05$) in their expression levels between mussel populations from the two *Mytilus* species. After applying several correction methods to control for the type I error using a procedure we have advocated previously [61] (see File S3), most of these spots remained significant, especially when more powerful correction methods were used (*e.g.*, 125 and 123 spots after applying the SGoF+ and SFisher correction respectively). Reassuringly, the *q*-values indicate a low expected false positive rate for the 128 significant spots ($q=0.208$), while fixing a *q*-value at 5% level provides 45 significant spots (Figure 3 and File S3). A heat map including the expression data for the 45 significant spots ($q<0.05$) shows samples for each population in one of two different clusters without any exceptional individuals (Figure 4). The same pattern is observed when the 128 *a priori* significant spots ($p<0.05$) are used (File S4). A Volcano plot (Figure 5) shows important size-effects in either *Mytilus* spp. directions. For example, there are significant differences ($p<0.05$) in expression associated with higher than 1.5 and 2.0 fold differences in 57 and 26 spots respectively comparing *M. galloprovincialis* with *M. edulis*, with higher expression in *M. galloprovincialis*, while 32 and 14 spots follow the same pattern but with opposite fold change direction with higher expression in *M. edulis*.

3.2.2. Protein identification by mass spectrometry (MS)

From a total of 45 candidate protein spots ($q<0.05$; see Figure 3), all except one were successfully identified after the analysis of mass spectrometry data against different customised databases used in this study (Table 3 and File S5). Spots 1101 and 1508 were annotated against protein sequences generated from our RNA-seq dataset, though blast analysis of these RNA sequences against the NCBI protein database did not provide any significant match. It is important to note that in three analysed spots two different proteins were identified with very high confidence, PSMs and scores. These are spot 2164 (Uqcrc2 and Tekt1), spot 705 (Atp5a and Dld) and spot 988 (Acadm and Psmc6). An explanation for this result is that the "protein-pairs" identified for these

spots present similar MW and pI, hence 2-DE analysis was not able to resolve them and they were sampled together when the spots were excised.

There are several spots showing differences in MW and pI (Figure 3) that were identified as the same protein (see Table 3). One possible explanation for this is that these originate by different post-transcriptional or post-translational modifications (PTMs). The correct interpretation of these candidate "multi-spot" proteins is important from a functional viewpoint to prevent misleading conclusions (see Box 2 in [38]). For example, in protein isoforms of Aco2 (spots 1205 and 1241) and Idh3g (spots 1085 and 1087) a concordant pattern of up-regulation in *M. edulis* was observed, whereas protein isoforms for Uqcrc2 (spots 847, 2164a and 2151), Efhc2 (spots 1119, 1134 and 191), es1 (spots 1608, 2039 and 1602), and Glud (spots 589 and 2062) showed a discordant pattern (see Table 3, Figures 3 and 4). Phosphorylation is one of the well-known PTMs that usually implies modification in the pI of phosphorylated protein but little MW change [67]. An advantage of using 2-DE for proteome separation compared to gel-free (shotgun) proteomic approaches is that it provides the possibility of assessing the effects of differential post-translational modifications and different isoform expression between samples [68-70]. The observation of spots resolved in close proximity in the 2-DE gel such as Idh3g (spots 1085 and 1087), Uqcrc (spots 847 and 2164), Glud (spots 589 and 2062), and Tekt2 (spots 814 and 776) is also compatible with differential phosphorylation events in the sperm of the two *Mytilus* spp., and could be verified by further phosphoproteomic analysis [71].

The list of protein identifications from excised spots contained many proteins potentially involved in sperm function. There are proteins involved in cell energy production, hence potentially affecting sperm motility, such as different members of the electron transport chain (ETC) protein complex (Nadufa10, Uqcrc2, Atp5a) or in close relation to ETC (Etfb), while Ppa1, Idh3g, Idh3a, Eno and Ak are other identified enzymes that also contribute to maintain the energetic cellular resources. An interesting observation is that about half of identified proteins are located in mitochondria (Table 3), so playing a role in cellular energy homeostasis either through ETC or different metabolic pathways. Proteins that contribute to flagellum structure could play a role in sperm motility, like Tekt1, Tekt2, Tekt4, and Cnn1. There are also proteins involved in sperm capacitation, for example Aco2, Dld, and Npr1. The identifications include also

different catalytic and regulatory subunits of the proteasome (Psmb2, Psma4, Psmb6, Psmc6, Psmd11, and Psme3). There is a group of identified proteins with a less obvious sperm-specific function role (Acadm, Pfd0110w, Ivd, Efhc2, Glud, Hsd17b10, Prdx5, Sod2, Plc, and an es1 protein).

3.3. Proteomic and transcriptomic differential expression results: in good agreement?

Although gene expression studies based on transcriptomic analysis have relied on mRNA abundance as a good proxy for corresponding protein abundance, results from a number of studies have questioned the validity of this assumption [72]. Substantial posttranscriptional and posttranslational modifications are expected and this can also affect the correlation between protein and transcript levels for many but not all gene products [73]. In this study we have tested the general level of agreement in the direction of the differential expression between proteomics (identified protein spots in Table 3) and transcriptomics data (see Files S5-S6 in Ref. [47]). The data are summarised in File S6 where for both protein and mRNA-seq data E and G are used as abbreviations for *M. edulis* and *M. galloprovincialis*. Worksheet Table S6 of this file lists the protein spots which show differential expression between the two species, and for which of the two species the expression is higher. Then in addition for each spot the number of mRNA isotigs showing differential expression (E>G and G>E) are given in separate columns.

For those protein spots showing higher *M. edulis* protein expression the total number of isotigs over all spots with E>G and G>E are 14 and 26 respectively: with higher *M. galloprovincialis* expression the numbers are 8 and 52. A $\chi 2$ heterogeneity test reveals that the overall preponderance of isotigs with G>E is significant (pooled $\chi 2$ = 31.360 df=1 p=0.000) and that the ratios 14:26 and 8:52 are different (heterogeneity $\chi 2$= 4.507 df=1 p=0.034) (File S6, worksheet Test). Thus spots which show G>E have a tendency towards an excess of isotigs also showing G>E. The data in Table S6 has also been used to directly correlate the fold change values for the proteomics data and for the RNA-seq. The data and plot is given in File S6 worksheet 2Dplot. There is a positive correlation which though weak (Spearman's Rho = 0.126, p=0.210) is nevertheless consistent with the above $\chi 2$ analysis in showing some general correspondence between the two types of data. Expectation of a positive correlation would depend on assumption of

589 generalised up or down regulation for the protein in question. However in general there
590 is not good correspondence between proteomics and transcriptomics data with cellular
591 concentrations of proteins not correlating highly with the abundance of their RNAs [72-
592 73]. This may be related to a number of factors including variation in protein turnover
593 rate, variation in the extent and nature of posttrancriptional and posttranslational
594 modification and measurement error.
595 Given that many isotigs in the overall dataset do not show differential expression, it is
596 of interest to know whether a protein spot with E>G (or G>E) has at least one isotig
597 with differential expression in the same direction. The number of spots showing such
598 agreement can be contrasted with the number of spots for which all isotigs show
599 differential expression but in the opposite direction to that shown by the protein spot.
600 The numbers in these two categories are 28:4 over all spots ($\chi 2$ = 18.000 df=1 p<0.001,
601 for test against 1:1 expectation, see File S6 worksheet Table S6 for further details) and
602 20:4 when counting for protein identities, that is spots for the same protein are counted
603 once only ($\chi 2$= 10.667 df=1 p=0.001). These significant results provide additional
604 evidence for concordance between the two types of expression data. In addition to spots
605 with isotigs showing differential expression, 15 protein spots (32% of the total number
606 of spots) do not have any isotigs showing differential expression (File S6 worksheet
607 Table S6, total spots with "0" in column K). It is important to highlight that four of
608 these protein spots were identified as different proteasome subunits with higher
609 expression in *M. galloprovincialis* sperm (File S6 worksheet Table S6, column D).

610 For the two categories of proteins with expression E>G and G>E, the distribution of
611 number of spots for different ontology terms was determined. This is carried out for two
612 ontology classifications, Cellular Location and Molecular Function, which are derived
613 from the classifications shown in Figure 2. The resulting distributions with further
614 analysis are given in File S6 worksheet Test. The ontology terms having greatest
615 frequency overall are Mitochondrion (43%) and Cytoplasm (20%) for Cellular
616 Location, and Motility (29%), Capacitation (12%) and Acrosome reaction (12%) for
617 Molecular Function. The results of $\chi 2$ contingency tests in which the ontology
618 distributions are compared between E>G and G>E indicate a significant effect overall
619 for both Cellular Location (p=0.002) and Molecular Function (p=0.027). Individual
620 ontology terms which contribute most to the overall effect are Mitochondrion (p=0.000,
621 higher number of spots for E>G), Cytoplasm (p=0.020, higher for G>E), Proteolysis

(p=0.083, higher for G>E), and Tricarboxylic acid cycle (p=0.005, higher for E>G). So while there is a correspondence overall for Cellular Location between the highest frequency terms and those differing in frequency most markedly between species, this is not observed for Molecular Function.

3.4. Customised tissue and species-specific protein databases enhance protein identifications

While identifying peptides from MS data together with the corresponding proteins in model organisms is quite straightforward, the situation becomes more challenging when working with non-model organisms because the availability of genomic and protein sequences in the latter is scarce. However there are different alternatives to overcome this limitation (see [37, 45-46]). For example, the generation of customised protein databases obtained from tissue and species-specific transcriptome datasets (RNA-seq) or from expression sequence tags (ESTs) deposited and available through NCBI. Also *de novo* interpretation of MS/MS spectra can provide complementary results when combined with the use of customised protein databases, specifically in providing information about unknown mutations and PTMs, this latter being also valid for model organisms.

In order to assess whether the use of customised protein sequence databases has improved the quality and quantity of protein identifications in the current study on two *Mytilus* spp., we compared the number of peptide spectrum matches (PSMs), total (TP) and unique peptides (UP) obtained in the identification of 44 protein spots from sperm samples (see section 3.2.2) using 3 different customised databases (see section 2.2.4). Graph displayed in Figure 6, made from data available in File S5, shows that using a protein database made from our consensus tissue and species-specific transcriptome data provide on average across 44 spots better results in terms of a significantly higher number of PSMs (Kruskal-Wallis test; H=25.27, df=2, p<0.0001), TP (H=24.29, df=2, p<0.0001) and UP (H=34.48, df=2, p<0.0001) when compared with the other two protein databases. When these results are inspected in a pair-wise comparison basis, after applying Dunn *post-hoc* test for multiple comparisons, it is worth noting that the customised *Mytilus*-ESTs-based protein database also presented good results for PSMs and TP, but with a significantly lower number of UP, when compared with the RNA-

seq-based protein database (see Figure 6). It is also clear that the results of these two customised *Mytilus* specific protein databases are significantly better than those obtained after using a NCBI[*Mollusca*]-based protein database, except the pairwise comparison between *Mytilus*-ESTs-based and NCBI[*Mollusca*]-based protein databases for UP (see Figure 6). The lower number for UP can be explained by high redundancy found in EST databases. The confirmation that EST sequences from *Mytilus* spp. are generally shorter than protein sequences derived from our RNA-seq project can be easily reached from inspection of matched protein sequences from each database used in the protein spot identifications (see File S5). File S5 also provides useful information about potential PTMs and mutations, ascertained with the PEAKS program through *de novo* interpretation of MS/MS spectra, present in the sequences of proteins to which the different spots were identified.

4. Discussion

4.1 Transcriptomic differences in mature male gonad between two *Mytilus* spp. shed light on proteins with potential involvement in reproductive isolation

Results from transcriptomic experiments using next-generation sequencing technology (RNA-seq) with a focus on different biological questions have been reported for *M. edulis* (e.g. [74], in a study of gene regulation during early development) and *M. galloprovincialis* (e.g. [75], to compare transcript expression profiles in four different tissues). However there has not been any attempt to deep sequence the mature male gonad transcriptome and compare transcriptomic data in these two *Mytilus* species. The current RNA-seq analysis provides evidence of high variation in the mature male gonad transcriptome, with 22.8% of analysed loci differing (at FDR 1%) between *M. galloprovincialis* and *M. edulis* samples. In a high number of instances the differential expression was detected at isotig level within each consensus transcript (locus), with contrasting results among different isotigs within loci, both in terms of effect-size and direction of the expression level between the two *Mytilus* spp. (see Table 2). The RNA transcripts showing different expression in Table 2 are both derived from sperm and have sperm associated GO terms with their protein names. We would thus expect many of these transcripts to be expressed as proteins for specific functioning in this tissue. However in general it cannot be assumed that all isotigs showing differential expression

685 are translated into proteins [76], and it may be that a single transcript is dominant in
686 terms of protein expression [77]. The statistical correspondence in the direction of
687 expression between species for isotigs and protein spots ($\chi 2$ heterogeneity test in File
688 S6, Table S6) give further evidence that some of the isotigs are translated into protein
689 even if it is not possible to pinpoint exactly which isotigs are translated and which are
690 not.

691 Samples from both species shared a common laboratory environment for at least two
692 months. This design often referred to as a common garden experiment (*e.g.* [78]), aims
693 to demonstrate that observed phenotypic differences are mainly attributable to species-
694 specific (genetic) rather than sampling-site environmental differences, and is becoming
695 important for studying adaption in genomic studies [79]). Although acclimation to the
696 same laboratory conditions should help to minimise the effects of local environmental
697 differences between the original sampling sites, some of these environmental effects
698 may be retained permanently even after acclimation for several weeks [80]. When the
699 aim is to compare allopatric population of different species, genetic and local
700 environmental differences may always be confounded, but the long period of
701 acclimation used in the current study (at least 2 months) should have maximised
702 genetically based, as compared with environmentally based, transcriptome differences
703 between the species. Variation between gonadal development stages in transcript
704 abundance have been reported in *M. galloprovincialis* [81]. However in the present
705 study mussels at the same stage of development, according to histological tests, were
706 used in the two species.

707 From the list of genes which show significant expression differences between *M. edulis*
708 and *M. galloprovincialis* at the mRNA level, there are several that produce proteins with
709 functional roles in sperm biology and fertilization (Table 2). Most of these proteins are
710 thus good candidates for evolutionary study due to their potential role in reproductive
711 isolation mechanisms and ultimately in the formation of new species, and are discussed
712 below.

713 *4.1.1 T-complex protein 1 (TCP-1) and ubiquitin-proteasome system (UPS) might be*
714 *involved in intraspecific gamete preference and reproductive isolation in Mytilus spp.*

One of the most important results is the concerted differential expression between the two *Mytilus* spp. for seven out of eight subunits of the T-complex protein 1 (TCP-1). A chaperonin-containing T-complex protein 1 was found in the periacrosomal region of human and mouse sperm heads with an involvement in mediating sperm-ZP interaction [82-83]. Evidence was found to support the view that TCP-1 and the ubiquitin-proteasome system (UPS) might by concerted action be involved in gamete interaction [82-83]. Hence TCP-1 and UPS are good targets for further investigation in relation to involvement in prezygotic reproductive mechanisms that could be operating between *Mytilus* spp. It is possible that differences in the expression level or in the sequence of TCP-1 and UPS related proteins can lead to a preference for intraspecific rather than interspecific fertilisations in *Mytilus* spp. UPS is involved in the process where protein substrates are labelled with different ubiquitins to be later recognised by the 26S proteasome complex machinery for protein substrate degradation playing important roles during sperm capacitation, the acrosome reaction and sperm-egg interactions (reviewed in [84]). Two candidate differentially expressed transcripts found in our study (Table 2) relate to the ubiquitin-proteasome system (UPS). These are the ubiquitin-conjugating enzyme (UBC) E2-24 kDa (Ubc8) and the proteasome subunit alpha type-2 (Psma2). Testis-specific isoforms of the first protein were found in the ascidian *Ciona intestinalis* and rat spermatozoa and a mutant mouse for this enzyme showed alterations in sperm as well as a reduced sperm number and motility [84]. Inactivation of an ubiquitin-conjugating enzyme in *Drosophila* causes male infertility due to abnormal levels of spermatogenesis [85]. It was demonstrated in ascidians, sea urchins and mammals that ubiquitin-conjugating enzymes regulate the penetration of spermatozoa into the vitelline coat (VC) of the egg and degrade the ubiquitinated sperm receptors on the VC (zona pellucida-ZP, in mammals) of eggs during fertilisation, contributing to the avoidance of polyspermy, with some roles also during sperm capacitation and regulation of acrosomal exocytosis (reviewed in [84, 86]. In relation to the second protein (Psma2), sperm proteasomes are released extracellularly as part of the acrosomal content during fertilisation. Together with an intracellular UPS inside the fertilised egg, it seems that animal fertilisation is also dependant of an extracellular UPS driven by the acrosomal exocytosis of different enzymes/proteins, and this mechanism seems to be quite evolutionarily conserved in the animal kingdom with small differences in ascidians compared with sea urchins and mammals. Its functional importance in

748 fertilisation has been empirically confirmed, suggesting that UPS proteins are a good
749 target for controlling fertilisation, and hence reproduction, in different organisms [84].
750 Proteasome subunit alpha was also identified among those proteins with higher
751 expression in *Mytilus edulis* sperm [20].

752 *4.1.2 Other candidate sperm-specific gene products linked to acrosome reaction, sperm-*
753 *egg interaction and rapid evolution*

754 The presence of a beta-n-acetylhexosaminidase (Bre-4) among the candidate proteins is
755 interesting because glycosidic enzymes were observed in the sperm acrosome content
756 and found to be necessary for penetration of the ZP during fertilisation in some
757 mammals, as well as acting as important sperm receptors for the extracellular matrix of
758 the oocyte in ascidians [87-88]. The sperm surface protein SP17 (Spa17) is of interest
759 because it might be involved in spermatogenesis, sperm capacitation, the acrosomal
760 reaction and sperm-egg interactions during fertilisation [89]. Evidence of high Spa17
761 protein expression was obtained in *Mytilus edulis* sperm [20], and in the current study
762 one isoform shows differential expression. Sperm proteins with testis-specific
763 expression have been found to evolve more rapidly on average than proteins expressed
764 in testis alone and in non-reproductive tissues. This is probably due to functional
765 constraints associated with housekeeping tasks of this latter-type of protein (see [90]).
766 The relative contribution of neutral and naturally selected genetic variation has been a
767 long debated and investigated issue during the last 50 years in evolutionary biology
768 [91]. In this context, SP17 was found to evolve rapidly by positive selection in several
769 mammalian species [92]. Similarly zonadhesin protein (Zan) was found to evolve
770 rapidly in primate species [93]. It is a large sperm-specific protein localised in the sperm
771 head within the acrosomal matrix with multiple domains involved in the species-
772 specific recognition of ZP in eggs during fertilisation in mammals (reviewed in [94]).
773 The acrosome content is quite variable between mammals and marine invertebrates. In
774 sea urchins and abalones, bindin and lysin sperm acrosomal proteins are rapidly
775 evolving species-specific proteins that recognise the vitelline coat of the egg
776 (corresponding to ZP in mammals) during fertilisation, while evolution of zonadhesin is
777 also driven by positive selection and involved in the same function in mammals, despite
778 these three proteins being evolutionarily unrelated (reviewed in [2, 94]). The protein
779 structure of zonadhesin is quite conserved despite high aminoacid divergence across
780 different species. A precursor form of zonadhesin protein is produced during

781 spermatogenesis and quickly processed to produce 3 polypeptides of 300, 105 and 45
782 kDa respectively in pig spermatozoa [94]. We provide evidence of four different Zan
783 loci and a total of seven isotigs with differential expression between the *Mytilus* spp., so
784 making this gene a target of interest in further studies of reproductive isolation in
785 *Mytilus* species. Evidence has been actually reported for positive selection acting on the
786 M7 lysin gene in some sympatric and allopatric *Mytilus* populations [25-27, but see 28]
787 and the M3 lysin gene [95]. M3 and M7, together with the less studied M6 lysin, are
788 non-orthologous highly abundant acrosomal proteins responsible for dissolving the egg
789 vitelline envelope during fertilisation [96], so are thought to play an important role in
790 the gamete recognition process. Interestingly in our study we found evidence of
791 differential expression for a total of eight different isotigs of M3 and M6 lysins, but no
792 differential expression of M7 lysin.

793 *4.1.3 Prdm9 and Suv39h2 gene products are promising targets to study postzygotic*
794 *reproductive isolation mechanisms and sex differences in Mytilus spp.*
795 Finally two other candidate gene products displayed in Table 2**,** Prdm9 and Suv39h2,
796 can be highlighted. When two populations that have evolved allopatrically come into
797 secondary contact, gamete compatibility may still occur and hybrid individuals
798 produced as observed for *Mytilus* spp. However hybrids can be sterile or have reduced
799 fitness due to epistatic interactions of alleles from the two diverged genomes. This
800 phenomenon known as Dobzhansky-Muller incompatibility (DMI) can lead to the
801 formation of new species. Only a very few genes responsible for such low hybrid fitness
802 have been discovered so far (see [97]). Prdm9, which shows differential expression for
803 one isotig, is also known as Meisetz, is a histone H3 methyltransferase, and is expressed
804 in mouse testis and ovaries [98]. This gene activates other essential genes for meiosis by
805 means of specific-histone methylation. Sterile hybrid male mice had small testes,
806 spermatogenic arrest and lacked sperm, the same phenotype as observed in null-Prdm9
807 mutant mice [98]. The cause of sterility seems to be DMI generated by epistatic
808 interaction between Prdm9 and other genes located on chromosome X (see [97]). In
809 view of the discoveries in the mouse, we suggest that Prdm9 deserves further attention
810 in evolutionary studies on *Mytilus* spp where reproductive isolation is incomplete. On
811 the other hand, Suv39h2, differentially expressed here in two isotigs, is another histone
812 H3 methyltransferase, and was found to be specifically expressed in mouse adult testes
813 but not ovaries [99] and specifically accumulates with chromatin of the sex

chromosomes silencing their expression during early meiosis. Possibly this protein could be useful for the development of a sex specific marker in *Mytilus*. This is currently lacking in *Mytilus* spp for which there is currently no evidence of sex chromosome dimorphism. For example, Suv39h2 as a target protein in immunofluorescence analysis for detecting differences between males and females.

4.2. Sperm proteome differences between *Mytilus edulis* and *M. galloprovincialis*

In line with the RNA-seq results, proteomic analysis on sperm samples from individuals from *M. edulis* and *M. galloprovincialis* provide evidence of high proteome differences between the species, occurring in 17.6% of protein spots analysed (q=0.208). All mussels were kept under common laboratory conditions for at least 2 months and thus had a long period to acclimate prior to the collection of sperm for proteomics analysis. Following the reasoning given above in the discussion of the transcriptome results, proteome differences between the species can therefore be attributed entirely or in large part to genetic differences between the species. A reassuring result is that from a similar proteomic experiment on sperm samples of individual mussels from a hybrid population at Croyde (UK) with sympatric *M. edulis* and *M. galloprovincialis,* species-specific proteomic patterns were also observed [100], strengthening the evidence that species-specific proteomic differences between mussels raised under similar conditions are genetically based. Although differential expression may be associated with the processes of protein synthesis, post-translational modification, and protein degradation, all may result in variation in protein abundance and have functional implications [101]. From the list of 44 protein spots (q=0.05) with differential expression and identified by MS, there are a number of proteins with key functional roles in sperm biology and fertilization (Table 3) that make them good targets (hereafter candidates) for potential involvement in reproductive isolation mechanisms. A feature of the results shown in Table 3 is that different spots for the same protein may differ in the species in which they show higher expression. Some proteins given in Table 3 which are of particular interest are highlighted and discussed below.

4.2.1 Mitochondrial proteins linked to energy production and antioxidant enzymes are up-regulated in M. edulis

Alterations in ETC-related proteins, and hence in cellular energetic production, have been linked to lack of sperm motility and, hence fertility, in some mammals [102-103],

so any observed differences between the two *Mytilus* species could be the result of their following different adaptive strategies relating to sperm motility. From the list of identified proteins showing differential expression (Table 3), NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 10 (Ndufa10), Cytochrome b-c1 complex subunit 2 (Uqcrc2), and ATP synthase subunit alpha (Atp5a) are nuclear encoded and from the different complexes of the respiratory electron transport chain (ETC) in mitochondria. Remarkably, the list of protein identifications (Table 3) reveals that nearly half of the identified proteins develop their functions and are located in mitochondria. A similar result was observed for highly expressed proteins in the sperm of *Mytilus edulis* [20]. Proteins from these ETC-associated complexes might be implicated in postzygotic isolating mechanisms due to coevolution of nuclear and mitochondrial genomes to ensure appropriate functional interactions between the nuclear and mitochondrial coded protein subunits of these complexes [104-105]. Marine mussels of *Mytilus* spp. as well as other bivalves present an unusual mtDNA inheritance mechanism (termed doubly-uniparental inheritance, DUI) in which distinct mtDNA genomes are passed through the male and female lines of descent and which is coupled to sex determination in these species [106-109] with opportunity for selection to act directly on mtDNA coded sperm proteins. Negative epistatic interactions between nuclear and mitochondrial genomes in hybrids could contribute to the maintenance of species integrity, consistent with observations of DUI disruption in crosses between these two *Mytilus* species [110].

Other identified differentially expressed mitochondrial proteins relate to energy metabolism. These include isocitrate dehydrogenase (Idh3a and Idh3g), aconitate hydratase (Aco2) and dihydrolipoyl dehydrogenase (Dld). Idh3 was identified as having the highest expression levels in a previous proteomics study of *Mytilus edulis* sperm [20]. Low expression levels of Aco2 were reported in human sperm with reduced motility [111], and higher levels during mice sperm capacitation [112]. Deficiency of Dld mature protein was associated with low sperm motility in humans [113], while enzymes of this complex were also related to sperm capacitation and the acrosome reaction in the hamster and humans [114-115]. The higher expression of such proteins in might result in higher ATP production and a fitness advantage under certain ecological and environmental conditions (see section 4.3). However production of ATP through oxidative phosphorylation (OXPHOS) may produce high reactive oxygen species (ROS) in sperm leading to mitochondrial mutations [116] and evolution of a trade-off

between higher OXPHOS and higher activity of antioxidant enzymes to neutralise high ROS production. Related to this is the observation in the present study that the differentially expressed antioxidant enzymes peroxiredoxin-5 (PRDX5) and manganese superoxide dismutase (SOD2) were associated with abnormal sperm and infertility in several mammals [117-118]. SOD activity may have detrimental effects on human sperm motility [119], and PRDX5 might play a role in sperm-egg interaction through the induction of signalling events by means of redox reactions after ZP binding [120].

4.2.2 Up-regulation of rapid energy supply and alternative production pathways in M. galloprovincialis

It is of interest that different species, for different cellular types, could have evolved different strategies and molecular pathways for energy production [121] driven by different ecological or environmental pressures. For example, glutamate dehydrogenase (Glud) converts glutamate to α-ketoglutarate potentially enhancing the activity of the TCA cycle in which α-ketoglutarate is an intermediate. Two spots closely located in the 2-DE map, were identified as Glud. These could be isoforms resulting from different posttranscriptional and posttranslational modifications (*e.g.* phosphorylation) implying functional changes [122] in sperm of both *Mytilus* spp.

ATP production through the glycolytic pathway in the sperm is compartmentalised in the principal piece of the flagellum, and this ATP source may be important in the sperm motility process known as hyperactivation [123]. The glycolytic enzyme enolase (Eno) was also differentially expressed. Disruption of expression of this enzyme sperm causes sperm structural defects and male infertility in the mouse [124]. In general glycolytic ATP is produced faster but less efficiently than ATP from aerobic pathways. Thus a trade-off between speed and amount of ATP production in sperm cells might also be of functional significance in sperm.

Phosphagen kinases are involved in intracellular energy transport and temporal buffering of ATP levels, specifically in flagellated cells, and hence probably play a role in sustained sperm motility [125]. The enzyme also influences sperm tail length and flagellar bending [126-127] and sperm-specific isoforms have been reported in various invertebrates [20, 125]. One of these enzymes, arginine kinase (Ak) was differentially expressed here in two spots. Phosphagen molecules also regulate intracellular inorganic phosphate levels [128] and play an important role in sperm motility, capacitation, the

911 acrosome reaction and sperm-egg fusion [129]. Inorganic pyrophosphate (PPi) is
912 degraded by pyrophosphatase 1 (Ppa1) for which one differentially expressed spot was
913 identified. PPi enhances sperm proteasome activity, of key importance for the sperm-
914 egg interaction during fertilization [129]. Interestingly several differentially expressed
915 spots related to the proteasome complex have been identified in the present study (see
916 section 4.2.3).

917 4.2.3 *Up-regulation of sperm proteasome activity in M. galloprovincialis: contrasting*
918 *transcriptomic and proteomic results*
919 Six protein spots were identified as different structural (alpha), catalytic (beta) and
920 regulatory subunits of the proteasome complex (Psma4, Psmb2, Psmb6, Psme3, Psmc6
921 and Psmd11). The important role of the ubiquitin-proteasome system (UPS) during
922 fertilization, including sperm capacitation, acrosome reaction and sperm-ZP binding,
923 has been considered in section 4.1. It is notable that all these had higher expression in
924 *M. galloprovincialis* (Table 3 and File S6) suggesting that this species could have
925 evolved specific regulatory mechanisms that increase the abundance of these proteins in
926 sperm cells. Interestingly several proteasome subunit alpha components were also
927 identified in *M. edulis* eggs and linked to the molecular mechanism underlying doubly-
928 uniparental inheritance (DUI, see section 4.2.1) of mtDNA in *Mytilus* spp. [59, 130].
929 Sperm mitochondria are labelled through ubiquitination during spermatogenesis [131]
930 and thus marked for elimination by the proteasome complex in the fertilised oocyte.
931 Three of the differentially expressed transcripts (Table 2) are two prohibitins (Phb and
932 Phb2) and sequestosome-1 (Sqstm1). Prohibitins play a role in mtDNA inheritance
933 [132], and are targets for ubiquitination in sperm mitochondria [133] while Sqstm1 has
934 been linked to sperm mitophagy in mammals [134]. Thus there may be a link between
935 the observed species-specific expression differences of these proteins in this study and
936 disruption in DUI reported in inter-specific crosses [110], and of relevance to
937 Dobzhansky-Muller incompatibilities (DMI) in hybrids between these species.

938 4.2.4 *Higher expression of tektins suggests high motility sperm in M. galloprovincialis*
939 Another interesting functional group of proteins showing differential expression are
940 tektins. Six spots were identified as three different tektin proteins (Tekt1, Tekt2 and
941 Tekt4) (Table 3). Of these, five had higher expression in *M. galloprovincialis*. Tektins
942 are cytoskeletal proteins of the sperm flagellum and involved in sperm motility and

flagellar bending. Differences in expression between normal and low motility sperm in humans were reported for Tekt1 and Tekt2, and Tekt4 was found to be essential for proper coordinated beating of the flagellum and for fertility [135-139]. Tektin expression occurs in the sperm acrosomal region perhaps indicating some specific role during fertilisation (see [135]) and has been implicated in flagellar bending and motility patterns [135, 140].

4.2.5 *Other identified proteins with sperm-specific functional links*

Three different spots with differential expression were identified as the protein EF-hand domain-containing family member C2-like (Efhc2). Sperm proteins with EF-hand domains play a key role in activation of the oocyte during fertilisation in mammals [141], and can also be involved in the acrosome reaction in invertebrates [9] and motility regulation of sperm [142-143]. Three protein spots identified as ES1 also showed differential expression. There is little functional information on this protein though it has been related to differential sperm motility in humans [111]. Other proteins showing differential expression are 3-hydroxyacyl-CoA dehydrogenase type-2 (Hsd17b10), potentially involved in the regulation of steroid hormones in reproduction and reported in several molluscs [144], and atrial natriuretic peptide receptor (Npr1) which acts on capacitation, chemotaxis and chemokinesis [145-146] and thus might potentially play a role in species-specific sperm-egg recognition in *Mytilus* spp. driven by chemotaxis signals released from eggs.

4.3. Rapid evolution and sperm function trade-offs may explain species-specific proteome differences

4.3.1 *Selective pressures and adaptation in sperm*

In external fertilisers such as mussels, sperm are expected to be under a variety of selective pressures relating to the different biological strategies for fertilisation and the ecological and environmental challenges they experience. Mussel settlements are patchy along rocky shores, and population density may vary considerably on a geographic or seasonal basis. Even though there may be synchronous spawning of eggs and sperm, the impact of varying gamete density and the role of sperm limitation is unclear [147-148]. If sperm density is too low then the probability of successful fertilisation may be low: on the other hand if sperm density is too high polyspermy may occur also resulting in incomplete fertilisation [149]. With sexual conflict, competition between sperm to

achieve successful fertilisation may be accompanied by selection for eggs that block fertilization to prevent polyspermy. This can lead to rapid co-evolution of proteins in eggs and sperm in the context of sexual conflict. The rapid evolution of sperm proteins has been observed in many animal groups from mammals to different marine invertebrates such as sea urchins, abalones, turban snails, oysters, sea stars and mussels [9, 150-151]. In a comparison of sperm proteins between *M. galloprovincialis* and *M. edulis* the highest non-synonymous to synonymous substitutions rates were observed for proteins involved in fertilisation [21]. Sperm limitation should exert strong selection for adaptations increasing the chance of successful fertilisation in marine organisms with external fertilisation [152-153]. These include spawning synchrony, high levels of sperm production, chemotaxis over short distances, and sperm longevity. There is evidence that sperm energetics, for example higher ATP production may enhance sperm performance through an increase in swimming speed [154] and increase the chance of fertilization. But given finite energy resources to allocate to sperm properties and function, trade-offs between sperm traits are expected. For example trade-offs between sperm velocity and longevity occur both within and between species [154-155]. However there are numerous complicating factors such as the ability of sperm to maintain flagellar beats with low ATP and high inorganic phosphate levels, or the use of alternative pathways for energy production [121, 156] despite oxidative phosphorylation and glycolysis in the sperm midpiece being the major source of ATP production [123].

4.3.2 *Sperm proteins upregulated in M. edulis and M. galloprovincialis*

In the present study many proteins connected with sperm function which are upregulated in *M. edulis* or *M. galloprovincialis* (Table 3) have been identified and their properties discussed above (see section 4.2). As contrasting scenarios, selective pressures in the native environments of the two species could be somewhat similar or quite different. In the former scenario suppose that selection favoured upregulation of proteins improving motility to enhance fertilisation success. This could be achieved by upregulating different genes of the same protein in the two species. For example proteins from different Tektin-2 spots are upregulated in *M. edulis* and *M. galloprovincialis* (Table 3). This differential effect could be achieved by selection or drift increasing the frequency of different locus specific expression modifiers in the two species. Alternatively different proteins potentially affecting motility could be differentially upregulated in the two species. For example, isocitrate dehydrogenase is

upregulated in *M. edulis* and arginine kinase is upregulated in *M. galloprovincialis* (Table 3). In the latter scenario where selection pressures differ between species, proteins for quite different traits may obviously be upregulated in the two species.

A summary of the proteins of Table 3 matched with sperm functional traits is given in File S7. Column I marks the particular 4.2 sub-sections in which proteins were flagged as having predominantly higher expression in *M. edulis* (4.2.1) or *M. galloprovincialis* (4.2.2, 4.2.3, 4.2.4). Column G assigns functional trait terms to the proteins and the count and % frequency distributions for these terms are given in Figure 7. These distributions give at least an approximate guide to which sperm traits are upregulated in the two species. In both species proteins relating to motility are important in this regard. After this, proteins relating to ATP reserves and perhaps ROS production are important in *M. edulis* whereas proteins relating to the acrosome reaction, capacitation, and sperm-egg interaction might be highlighted in *M. galloprovincialis*. On this basis it is possible to hypothesise that motility is important in both species but particularly *M. edulis*, whereas in *M. galloprovincialis* proteins relating to sperm maturation and the fertilization process should be highlighted.

The potential biological consequences of these sperm traits are elaborated in File S7 in column H. A notable feature is that upregulation of many proteins in Table 3 can be hypothesised to result in a functional advantage for sperm. In this circumstance red font is used in columns G and H. For example in *M. edulis*, aconitate hydratase has higher expression than in *M. galloprovincialis* and this higher expression could be interpreted as a functional benefit in terms of faster swimming speed or endurance as well as improved maturation of sperm. By contrast the higher expression of es1 protein in *M. edulis* affecting the sperm trait motility might be hypothesised to reduce motility, a functional disadvantage, on the basis that lower motility was observed in human sperm with higher levels of this protein. This is represented by green text font in File S7 columns G and H. Where it is more difficult to arrive at a functional benefit or disadvantage, black font is used. The counts of the number of spots in which the sperm trait terms can be flagged with red, green or black font are also given in the final two columns of Figure 7. There is a clear preponderance of protein spots in which higher expression can be hypothesised to be a functional benefit in terms of sperm performance in the species in which this higher expression occurs, the functional benefits being largely in sperm motility and related traits and the fertilization process.

In both species, the higher expression of proteins associated with various aspects of sperm function are consistent with positive natural selection towards improved function and fitness of sperm. Closely related hybridising species such as *M. edulis* and *M. galloprovincialis* might be expected to show few or many differences in expression as a result of selection pressures arising from ecological forces. The wide range of differentially expressed proteins observed in the current study is consistent with evidence from the mouse where a diverse set of 81 different protein genes, including 23 sperm membrane proteins all gave evidence of positive selection [157], and where proteins involved in sperm-egg interactions in particular show accelerated evolution [151]. Such a large number of genes involved in sperm function could underline that there may be a high selection intensity acting on sperm. This may also provide multiple opportunities for disrupting sperm function. For example it has been reported that in sea urchins as few as 10 amino acid changes in the protein bindin are needed for complete gamete incompatibility [158], so limited changes occurring at different loci might have similar effects.

4.3.3 *Differential expression: implications for hybridization of M. edulis and M. galloprovincialis*

The observation of protein expression differences for many different genes connected with sperm function has implications for models of hybridization and introgression between the species. An earlier proteomic study of a hybrid zone between *M. edulis* and *M. galloprovincialis* using somatic tissue found evidence of high gene expression variation amongst hybrids consistent with segregation at expression modifier loci as introgression proceeds [42]. Such segregation of modifiers at many sperm function related genes differing in protein expression between the species could result lowered expression or general disruption of expression of these genes depending on dominance relationships at and epistatic interaction between the modifier loci. This could contribute to lowered fertility of hybrids or lowered fitness of larvae as has been observed experimentally between different *Mytilus* spp. [33, 35]. It might also contribute to the observed disruption of doubly uniparental inheritance (DUI) in crosses between these two species [159] or other pair of *Mytilus* spp. [160-162].

4.3.4 *Possible influences of environmental variation on sperm function*

M. edulis evolved in the North Atlantic whereas *M. galloprovincialis* evolved in the Mediterranean [101, 162-163]. The most prominent environmental factors that might have exerted selective influences in the past are first temperature and then salinity which are both higher in the Mediterranean. These environmental differences persist in the contrast between Vigo and Swansea today, with seawater temperature about 4^{o}C higher at Vigo during the spawning season. There is evidence that changes in seawater temperature may affect sperm function. Thus in *M. galloprovincialis* higher temperature is associated with lower fertilization rates on average [164] and sperm motility and linearity of swimming patterns are affected by temperature and its interaction with pH [165]. This may have fitness consequences as swimming speed has also been associated with higher fertilisation rates [166]. In some circumstances, for example when chemoattractants are not present, non-linear swimming patterns may be advantageous to maximise the chance of fertilisation [147, 167-168]. Other environmental factors may be important for successful fertilisation for example viscosity which is a function of temperature and salinity [169]. Factors such as seawater specific gravity and turbulence may also be important in determining the chance of successful fertilisation [170-171].

4.3.5 *Selective pressures and interpretation of present results*

The historical and current environmental factors affecting *M. edulis* and *M. galloprovincialis* could have generated different selective forces to cause divergence in sperm phenotype. This could include modification of functional trade-offs between traits such as swimming speed and endurance [172]. Differential selection modifying sperm phenotype are expected to cause differences in gene expression which could be reflected in the observed differences in protein expression as observed in the present study (Table 3, Figure 7 and File S7). Higher temperature and salinity in the evolution of *M. galloprovincialis* might relate to another factor, oxygen solubility which is lower at higher temperature and salinity. Stress from reduced oxygen could impact negatively on ATP production impacting on energy dependent biological processes such as motility, swimming speed and endurance in *M. galloprovincialis* from Vigo. In the present study however it appears that motility related proteins are relatively upregulated in *M. edulis* whereas proteins involved in sperm maturation and fertilisation are upregulated in *M. galloprovincialis* (Figure 7 and File S7).

4.3.6 *Future studies integrating proteomics and experimental work on sperm*

Clearly relating proteomics data and biochemical interpretations to environmental factors and to variation between species in sperm functional traits is a complex task for the future. Measuring intra and interspecific variation in sperm functional traits is in itself not an easy task [154]. Currently we are not aware of any direct comparative study of some sperm functional traits, like speed, longevity and movement pattern, between *M. edulis* and *M. galloprovincialis*. An experimental design in which sperm from *M. edulis* and *M. galloprovincialis* are spawned and their performance in motility and endurance as well as fertilisation success assessed, at a range of temperature and salinity conditions would be informative. This could be combined with further proteomics studies applied to sperm from individual mussels from these experiments. The sperm phenotype is highly plastic and evidence already exists for genotype-by-environmental interaction effects on sperm function [172]. An experimental design such as the one described above should allow detecting main effects and interactions involving species differences, reflecting genetic adaptation, contemporary environmental variation and underlying gene expression data. Such approaches could be further extended to the study of hybrid populations of the two species.

5. Concluding remarks

In order to achieve fertilization a sperm must come into contact with an egg and interact with it appropriately. Proteins mediate the interactions between sperm and egg at each step of the fertilisation process, and there is growing evidence that multiple protein complexes might be involved in concert during gamete interaction [82-83]. Species differences in these proteins are proposed as one of the key factors that lead to species-specific fertilisation and reproductive isolation. When prezygotic barriers fail, inter-species hybrids can occur. When this happens, postzygotic barriers play an important role in preservation of species integrity. We provide evidence of extensive variation in the mature male gonad transcriptome and sperm proteome in two mussel species, *M. edulis* and *M. galloprovincialis*. From the transcriptome analysis, we provide a preliminary list of proteins with sperm-specific functions. These functions are related to sperm-egg interaction, the acrosome reaction, spermatogenesis and motility. From the proteome analysis, we provide evidence of an overrepresentation of mitochondrial proteins among those candidate protein spots identified by MS, as well as contrasting differential expression in isoforms of many proteins. The use of customised species-

specific protein databases significantly enhance both the quantity and quality of protein identifications, with the use of RNA-seq derived protein databases showing superior results to other customised databases analysed in this study. Our results provide evidence of agreement between the transcriptomic and proteomic results in the direction of expression differences between species. Our results highlight that some candidate sperm proteins, specifically those relating to sperm motility, ATP reserves, and ROS production in *M. edulis* and proteins relating to sperm motility, the acrosome reaction, capacitation and sperm-egg interaction in *M. galloprovincialis* might be good targets in further genomic analysis of reproductive barriers between closely related species.

6. Acknowledgements

The authors are very grateful to Nerea González-Lavín for technical assistance during sample processing, the staff of CIM-UVIGO for technical support in the marine station facilities, colleagues from the biochemistry area for sharing equipment and providing useful advice during histological test interpretation, Susana Catarino and David Arteta (Progenika Biopharma) for assistance during RNA-seq analysis, Manuel Marcos (Proteomics service, CACTI, University of Vigo) and José M. Gallardo (IIM-CSIC, Vigo) for assistance and helpful advice in MS/MS analysis. This work was funded by the Spanish "*Ministerio de Economía y Competitividad*" (codes BFU2011-22599 and AGL2014-52062-R)", Fondos Feder and Xunta de Galicia ("*Grupos de Referencia Competitiva*" ED431C 2016-037), and NIH grant HD076862. M.R. Romero was supported by a predoctoral fellowship from Xunta de Galicia (Campus do Mar).

The authors declare no conflict of interest

7. References

[1] Palumbi SR. Genetic divergence, reproductive isolation, and marine speciation. *Annual Review of Ecology and Systematics* 1994, *25*: 547-572.

[2] Swanson WJ, Vacquier VD. The rapid evolution of reproductive proteins. *Nature Reviews Genetics* 2002, *3*: 137-144.

[3] Lessios HA. Speciation genes in free-spawning marine invertebrates. *Integrative and Comparative Biology* 2011, *51*: 456-465

[4] McDonough CE, Whittington E, Pitnick S, Dorus S. Proteomics of reproductive systems: towards a molecular understanding of postmating, prezygotic reproductive barriers. *Journal of Proteomics* 2016, *135*: 26-37.

[5] Corbett-Detig RB, Zhou J, Clark AG, Hartl DL, Ayroles JF. Genetic incompatibilities are widespread within species. *Nature* 2013, *504*: 135-137.

[6] Orr HA, Presgraves DC. Speciation by postzygotic isolation: forces, genes and molecules. *BioEssays* 2000, *22*: 1085-1094.

[7] Bianchi E, Doe B, Goulding D, Wright GJ. Juno is the egg Izumo receptor and is essential for mammalian fertilization. *Nature* 2014, *508*: 483-487.

[8] Wilburn DB, Swanson WJ. From molecules to mating: Rapid evolution and biochemical studies of reproductive proteins. *Journal of Proteomics* 2016, *135*: 12-25.

[9] Vacquier VD, Swanson WJ. Selection in the rapid evolution of gamete recognition proteins in marine invertebrates. *Cold Spring Harbor Perspectives in Biology* 2011, *3*: a002931.

[10] Klinovska K, Sebkova N, Dvorakova-Hortova K. Sperm-egg fusion: a molecular enigma of mammalian reproduction. *International Journal of Molecular Sciences* 2014, *15*: 10652-10668.

[11] Oliva R, de Mateo S, Estanyol JM. Sperm cell proteomics. *Proteomics* 2009, *9*: 1004-1017.

[12] Karr TL. Fruit flies and the sperm proteome. *Human Molecular Genetics* 2007, *16*: R124-R133.

[13] Gur Y, Breitbart H. Protein synthesis in sperm: dialog between mitochondria and cytoplasm. *Molecular and Cellular Endocrinology* 2008, *282*: 45-55.

[14] Bayram HL, Claydon AJ, Brownridge PJ, Hurst JL, Mileham A, Stockley P, Beynon RJ, Hammond DE. Cross-species proteomics in analysis of mammalian sperm proteins. *Journal of Proteomics* 2016, *135*: 38-50.

[15] Nakachi M, Nakajima A, Nomura M, Yonezawa K, Ueno K, Endo T, Inaba K. Proteomic profiling reveals compartment-specific, novel functions of ascidian sperm proteins. *Molecular reproduction and development* 2011, *78*: 529-549.

[16] Palmer MR, McDowall MH, Stewart L, Ouaddi A, MacCoss MJ, Swanson WJ. Mass spectrometry and next-generation sequencing reveal an abundant and rapidly

evolving abalone sperm protein. *Molecular reproduction and development* 2013, *80*: 460-465.

[17] Kingtong S, Kellner K, Bernay B, Goux D, Sourdaine P, Berthelin CH. Proteomic identification of protein associated to mature spermatozoa in the Pacific oyster *Crassostrea gigas*. *Journal of Proteomics* 2013, *82:* 81-91.

[18] Boonmee A, Berthelin CH, Kingtong S, Pauletto M, Bernay B, Adeline B, Suquet M, Sourdaine P, Kellner, K. Differential protein expression during sperm maturation and capacitation in an hermaphroditic bivalve, *Pecten maximus* (Linnaeus, 1758). *Journal of Molluscan Studies* 2016, *82*: 575-584.

[19] Bartel M, Hartmann S, Lehmann K, Postel K, Quesada H, Philipp EER, Heilmann K, Micheel B, Stuckas H. Identification of sperm proteins as candidate biomarkers for the analysis of reproductive isolation in *Mytilus*: a case study for the enkurin locus. *Marine Biology* 2012, *159*: 2195

[20] Diz AP, Dudley E, Skibinski DOF. Identification and characterization of highly expressed proteins in sperm cells of the marine mussel *Mytilus edulis*. *Proteomics* 2012, *12*: 1949-1956.

[21] Zhang Y, Mu H, Lau SC, Zhang Z, Qiu JW. Sperm proteome of *Mytilus galloprovincialis*: Insights into the evolution of fertilization proteins in marine mussels. *Proteomics* 2015, *15*: 4175-4179.

[22] Vicens A, Borziak K, Karr TL, Roldan ERS, Dorus S. Comparative sperm proteomics in mouse species with divergent mating systems. *Molecular Biology and Evolution* 2017, *34*, 1403-1416.

[23] Koehn RK. The genetics and taxonomy of species in the genus *Mytilus*. *Aquaculture* 1991, *94*: 125-145.

[24] Seed R. Ecology, in: Bayne, B.L. (Ed.). Marine mussels: their ecology and physiology. International Biological Programme 1976, 10: 13-65.

[25] Riginos C, McDonald JH. Positive selection on an acrosomal sperm protein, M7 lysin, in three species of the mussel genus *Mytilus*. *Molecular Biology and Evolution* 2003, *20*: 200-207.

[26] Riginos C, Wang D, Abrams AJ. Geographic variation and positive selection on M7 lysin, an acrosomal sperm protein in mussels (*Mytilus* spp.). *Molecular Biology and Evolution* 2006, *23*: 1952-1965.

[27] Springer SA, Crespi BJ. Adaptive gamete-recognition divergence in a hybridizing *Mytilus* population. *Evolution* 2007, *61*: 772-783.

[28] McCartney MA, Lima TG. Evolutionary consequences of introgression at M7 Lysin, a gamete recognition locus, following secondary contact between blue mussel species. *Integrative and Comparative Biology* 2011, *51*: 474-484.

[29] Brannock PM, Hilbish TJ. Hybridization results in high levels of sterility and restricted introgression between invasive and endemic marine blue mussels. *Marine Ecology Progress Series* 2010, *406*: 161-171.

[30] Beaumont AR, Abdul-Matin AKM, Seed R. Early development, survival and growth in pure and hybrid larvae of *Mytilus edulis* and *M. galloprovincialis*. *Journal of Molluscan Studies* 1993, *59*: 120-123.

[31] Beaumont AR, Turner G, Wood AR, Skibinski DO. Hybridisations between *Mytilus edulis* and *Mytilus galloprovincialis* and performance of pure species and hybrid veliger larvae at different temperatures. *Journal of Experimental Marine Biology and Ecology* 2004, *302*: 177-188.

[32] Bierne N, David P, Boudry P, Bonhomme F. Assortative fertilization and selection at larval stage in the mussels *Mytilus edulis and M. galloprovincialis*. *Evolution* 2002, *56*: 292-298.

[33] Bierne N, Bonhomme F, Boudry P, Szulkin M, David P. Fitness landscapes support the dominance theory of post-zygotic isolation in the mussels *Mytilus edulis* and *M. galloprovincialis*. *Proceedings of the Royal Society of London B: Biological Sciences* 2006, *273*: 1253-1260.

[34] Rawson PD, Slaughter C, Yund PO. Patterns of gamete incompatibility between the blue mussels *Mytilus edulis* and *M. trossulus*. *Marine Biology* 2003, *143*: 317-325.

[35] Miranda MBB, Innes DJ, Thompson RJ. Incomplete reproductive isolation in the blue mussel (*Mytilus edulis* and *M. trossulus*) hybrid zone in the Northwest Atlantic: role of gamete interactions and larval viability. *The Biological Bulletin* 2010, *218*: 266–281.

[36] Klibansky LK, McCartney MA. Conspecific sperm precedence is a reproductive barrier between free-spawning marine mussels in the northwest Atlantic *Mytilus* hybrid zone. *PloS One* 2014, *9*: e108433.

[37] Findlay GD, Swanson WJ. Proteomics enhances evolutionary and functional analysis of reproductive proteins. *Bioessays* 2010, *32*: 26-36.

[38] Diz AP, Martínez-Fernández M, Rolán-Alvarez E. Proteomics in evolutionary ecology: linking the genotype with the phenotype. *Molecular Ecology* 2012, *21*: 1060-1080.

[39] Baer B, Millar AH. Proteomics in evolutionary ecology. *Journal of Proteomics* 2016, *135*: 4-11.

[40] Diz AP, Calvete JJ. Ecological proteomics: is the field ripe for integrating proteomics into evolutionary ecology research?. *Journal of Proteomics* 2016, *135*: 1-3.

[41] Palumbi SR. Speciation and the evolution of gamete recognition genes: pattern and process. *Heredity* 2009, *102*: 66-76.

[42] Diz AP, Skibinski DOF. Evolution of 2-DE protein patterns in a mussel hybrid zone. *Proteomics* 2007, *7*: 2111-2120.

[43] Murgarella, M, Puiu D, Novoa B, Figueras A, Posada D, Canchaya C. A first insight into the genome of the filter-feeder mussel *Mytilus galloprovincialis*. *PloS one* 2016, *11*: e0151561.

[44] Findlay GD, MacCoss MJ, Swanson WJ. Proteomic discovery of previously unannotated, rapidly evolving seminal fluid genes in *Drosophila*. *Genome Research* 2009, *19*: 886-96.

[45] Armengaud J, Trapp J, Pible O, Geffard O, Chaumot A, Hartmann EM. Non-model organisms, a species endangered by proteogenomics. *Journal of Proteomics* 2014, *105*: 5-18.

[46] Nesvizhskii AI. Proteogenomics: concepts, applications and computational strategies. *Nature methods* 2014, *11*: 1114-1125.

[47] Diz AP, Romero MR, Pérez-Figueroa A, Swanson WJ, Skibinski DOF. RNA-seq data from mature male gonads of marine mussels *Mytilus edulis* and *M. galloprovincialis*. *Data in Brief* 2018 (submitted).

[48] Atasaral-Şahin Ş, Romero MR, Cueto R, González-Lavín N, Marcos M, Diz AP. Subtle tissue and sex-dependent proteome variation in mussel (*Mytilus galloprovincialis*) populations of the Galician coast (NW Spain) raised in a common environment. *Proteomics* 2015, *15*: 3993-4006.

[49] Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* 2008, *18*: 821-829.

[50] Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* 2012, *28*: 1086-1092.

[51] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *Journal of Molecular Biology* 1990, *215*: 403-410.

[52] Zhang G, Fang X, Guo X, Li L, Luo R, Xu F, Yang P, Zhang L, Wang X, Qi H, Xiong Z *et al*. The oyster genome reveals stress adaptation and complexity of shell formation. *Nature* 2012, *490*: 49-54.

[53] Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 2005, *21*: 3674-3676.

[54] Zhang ZH, Jhaveri DJ, Marshall VM, Bauer DC, Edson J, et al. A comparative study of techniques for differential expression analysis on RNA-Seq data. Plos One 2014, 9: e103207

[55] Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 2011, *12*: 1.

[56] Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BMG, Haag JD, Gould MN, Stewart RM, Kendziorski C. EBSeq: an empirical Bayes hierarchical model for inference in RNA-seq experiments. *Bioinformatics* 2013, *29*: 1035-1043.

[57] Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V, Wrobel N, Gharbi K, Simpson GG, Owen-Hughes T, Blaxter M, Barton GJ. How many biological replicates are needed in an RNA-seq experiment and which differential expression tool should you use? *RNA* 2016, *22*:1-13

[58] Ramagli LS, Rodriguez LV. Quantitation of microgram amounts of protein in two-dimensional polyacrylamide gel electrophoresis sample buffer. *Electrophoresis* 1985, 6: 559–563.

[59] Diz AP, Dudley E, Cogswell A, MacDonald BW, Kenchington EL, Zouros E, Skibinski DOF. Proteomic analysis of eggs from *Mytilus edulis* females differing in mitochondrial DNA transmission mode. *Molecular & Cellular Proteomics* 2013, *12*: 3068-3080.

[60] Carvajal-Rodriguez A, de Uña-Alvarez J. Assessing significance in high-throughput experiments by sequential goodness of fit and q-value estimation. *PLoS One* 2011, *6*: e24700.

[61] Diz AP, Carvajal-Rodríguez A, Skibinski DOF. Multiple hypothesis testing in proteomics: a strategy for experimental work. *Molecular & Cellular Proteomics* 2011, *10*: M110-004374.

[62] Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw WHA, Lumley T, Maechler M, Magnusson A, Moeller S, Schwartz M, Venables B. gplots: Various R Programming Tools for Plotting Data 2014. R package version 2.14.2. https://CRAN.R-project.org/package=gplots

[63] Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R. InterProScan: protein domains identifier. *Nucleic Acids Research* 2005, 33: W116-W120.

[64] Petersen TN, Brunak S, Heijne von G, Nielsen H. SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nature Methods* 2011, 8: 785-786.

[65] Krogh A, Larsson B, Von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology* 2001, 305: 567-580.

[66] Buckland PR. Allele-specific gene expression differences in humans. *Human Molecular Genetics* 2004, *13*, R255-R260.

[67] Yamagata A, Kristensen DB, Takeda Y, Miyamoto Y, Okada K, Inamatsu M, Yoshizato K. Mapping of phosphorylated proteins on two-dimensional polyacrylamide gels using protein phosphatase. *Proteomics* 2002, *2*: 1267-1276.

[68] Morandell S, Stasyk T, Grosstessner-Hain K, Roitinger E, Mechtler K, Bonn GK, Huber LA. Phosphoproteomics strategies for the functional analysis of signal transduction. *Proteomics* 2006, *6*: 4047-4056.

[69] Rabilloud T, Lelong C. Two-dimensional gel electrophoresis in proteomics: a tutorial. *Journal of Proteomics* 2011, *74*: 1829-1841.

[70] Diz AP, Álvarez-Rodríguez M, Romero MR, Rolán-Alvarez E, Galindo J. Limited proteomic response in the marine snail *Melarhaphe neritoides* after long-term emersion. *Current Zoology* 2017, *63*, 487-493.

[71] Thingholm TE, Jensen ON, Larsen MR. Analytical strategies for phosphoproteomics. *Proteomics* 2009, *9*: 1451-1468.

[72] Feder ME, Walser JC. The biological limitations of transcriptomics in elucidating stress and stress responses. *Journal of Evolutionary Biology* 2005, *18*: 901-910.

[73] Vogel C, Marcotte EM. Insights into the regulation of protein abundance from proteomic and transcriptomic analyses. *Nature Reviews Genetics* 2012, *13*: 227-232.

[74] Bassim S, Tanguy A, Genard B, Moraga D, Tremblay R. Identification of *Mytilus edulis* genetic regulators during early development. *Gene* 2014, *551*: 65-78.

[75] Moreira R, Pereiro P, Canchaya C, Posada D, Figueras A, Novoa B. RNA-Seq in *Mytilus galloprovincialis*: comparative transcriptomics and expression profiles among different tissues. *BMC genomics* 2015, *16*: 728.

[76] Ezkurdia I, del Pozo A, Frankish A, Rodriguez JM, Harrow J, Ashman K, Valencia A, Tress ML. Comparative proteomics reveals a significant bias toward alternative protein isoforms with conserved structure and function. *Molecular Biology and Evolution* 2012, *29*: 2265-2283.

[77] Gonzàlez-Porta M, Frankish A, Rung J, Harrow J, Brazma A. Transcriptome analysis of human tissues and cell lines reveals one dominant transcript per gene. *Genome Biology* 2013, *14*: R70.

[78] Pelini SL, Diamond SE, MacLean H, Ellison AM, Gotelli NJ, Sanders NJ, Dunn RR. Common garden experiments reveal uncommon responses across temperatures, locations, and species of ants. *Ecology and Evolution* 2012, *2*: 3009-3015.

[79] de Villemereuil P, Gaggiotti OE, Mouterde M, Till-Bottraud I. Common garden experiments in the genomic era: new perspectives and opportunities. *Heredity* 2016, *116*: 249-254.

[80] Sanford E, Kelly MW. Local adaptation in marine invertebrates. *Annual Review of Marine Science* 2011, *3:* 509-535.

[81] Banni M, Negri A, Mignone F, Boussetta H, Viarengo A, Dondero F. Gene expression rhythms in the mussel *Mytilus galloprovincialis* (Lam.) across an annual cycle. *PLoS One* 2011, *6*: e18904.

[82] Dun MD, Smith ND, Baker MA, Lin M, Aitken RJ, Nixon B. The chaperonin containing TCP1 complex (CCT/TRiC) is involved in mediating sperm-oocyte interaction. *Journal of Biological Chemistry* 2011, *286*: 36875-36887.

[83] Redgrove KA, Anderson AL, Dun MD, McLaughlin EA, O'Bryan MK, Aitken RJ, Nixon B. Involvement of multimeric protein complexes in mediating the capacitation-dependent binding of human spermatozoa to homologous zonae pellucidae. *Developmental Biology* 2011, *356*: 460-474.

[84] Sutovsky P. Sperm proteasome and fertilization. *Reproduction* 2011, *142*: 1-14.

[85] Hershko A, Ciechanover A. The ubiquitin system. *Annual Review of Biochemistry* 1998, *67*: 425-479.

[86] Zimmerman S, Sutovsky P. The sperm proteasome during sperm capacitation and fertilization. *Journal of Reproductive Immunology* 2009, *83*: 19-25.

[87] Miranda PV, González-Echeverría F, Blaquier JA, Mahuran DJ, Tezón JG. Evidence for the participation of β-hexosaminidase in human sperm–zona pellucida interaction in vitro. *Molecular human reproduction* 2000, *6*: 699-706.

[88] Koyanagi R, Honegger TG. Molecular cloning and sequence analysis of an ascidian egg β-N-acetylhexosaminidase with a potential role in fertilization. *Development, Growth & Differentiation* 2003, *45*: 209-218.

[89] Chiriva-Internati M, Gagliano N, Donetti E, Costa F, Grizzi F, Franceschini B, Albani E, Levi-Setti PE, Gioia M, Jenkins M, Cobos E, Kast WM. Sperm protein 17 is expressed in the sperm fibrous sheath. *Journal of Translational Medicine* 2009, *7*: 61.

[90] Schumacher J, Rosenkranz D, Herlyn H. Mating systems and protein–protein interactions determine evolutionary rates of primate sperm proteins. *Proceedings of the Royal Society of London B: Biological Sciences* 2014, *281*: 20132607.

[91] Enard D, Depaulis F, Crollius HR. Human and non-human primate genomes share hotspots of positive selection. *PLoS Genetics* 2010, *6*: e1000840.

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

[93] Gasper J, Swanson WJ. Molecular population genetics of the gene encoding the human fertilization protein zonadhesin reveals rapid adaptive evolution. *The American Journal of Human Genetics* 2006, *79*: 820-830.

[94] Tardif S, Cormier N. Role of zonadhesin during sperm–egg interaction: a species-specific acrosomal molecule with multiple functions. *Molecular Human Reproduction* 2011, *17*: 661-668.

[95] Lima TG, McCartney MA. Adaptive evolution of M3 lysin—a candidate gamete recognition protein in the *Mytilus edulis* species complex. *Molecular Biology and Evolution* 2013, *30*: 2688-2698.

[96] Takagi T, Nakamura A, Deguchi R, Kyozuka KI. Isolation, characterization, and primary structure of three major proteins obtained from *Mytilus edulis* sperm. *Journal of Biochemistry* 1994, *116*: 598-605.

[97] Flachs P, Bhattacharyya T, Mihola O, Piálek J, Forejt J, Trachtulec Z. Prdm9 incompatibility controls oligospermia and delayed fertility but no selfish transmission in mouse intersubspecific hybrids. *PloS One* 2014, *9*: e95806.

[98] Mihola O, Trachtulec Z, Vlcek C, Schimenti JC, Forejt J. A mouse speciation gene encodes a meiotic histone H3 methyltransferase. *Science* 2009, *323*: 373-375.

[99] O'Carroll D, Scherthan H, Peters AH, Opravil S, Haynes AR, Laible G, Rea S, Schmid M, Lebersorger A, Jerratsch M, Sattler L, Mattei MG, Denny P, Brown SDM, Schweizer D, Januwein T. Isolation and characterization of Suv39h2, a second histone H3 methyltransferase gene that displays testis-specific expression. *Molecular and Cellular Biology* 2000, *20*: 9423-9433.

[100] Romero MR. Evolutionary and functional proteomic study of gametes from *Mytilus edulis* and *M. galloprovincialis* integrating different -omics technologies (Unpublished doctoral dissertation 2017). University of Vigo, Spain.

[101] Tomanek L. Environmental proteomics of the mussel *Mytilus*: implications for tolerance to stress and change in limits of biogeographic ranges in response to climate change. *Integrative and Comparative Biology* 2012, *52*: 648-664.

[102] Ruiz-Pesini E, Diez C, Lapeña AC, Pérez-Martos A, Montoya J, Alvarez E, Arenas J, López-Pérez MJ. Correlation of sperm motility with mitochondrial enzymatic activities. *Clinical Chemistry* 1998, *44*: 1616-1620.

[103] Zhao C, Huo R, Wang FQ, Lin M, Zhou ZM, Sha JH. Identification of several proteins involved in regulation of sperm motility by proteomic analysis. *Fertility and Sterility* 2007, *87*: 436-438.

[104] Dowling DK, Friberg U, Lindell J. Evolutionary implications of non-neutral mitochondrial genetic variation. *Trends in Ecology & Evolution* 2008, *23*: 546-54.

[105] Skibinski DOF, Ghiselli F, Diz AP, Milani L, Mullins JG. Structure-related differences between cytochrome oxidase I proteins in a stable heteroplasmic mitochondrial system. *Genome Biology and Evolution* 2017, *9*: 3265-3281.

[106] Skibinski DOF, Gallagher C, Beynon CM. Sex-limited mitochondrial-DNA transmission in the marine mussel *Mytilus edulis*. *Genetics* 1994, *138*: 801-809.

[107] Skibinski DOF, Gallagher C, Beynon CM. Mitochondrial-DNA inheritance. *Nature* 1994, *368*: 817-818.

[108] Zouros E, Ball AO, Saavedra C, Freeman KR. An unusual type of mitochondrial-DNA inheritance in the blue mussel *Mytilus*. *Proceedings of the National Academy of Sciences of the USA* 1994, *91*: 7463-7467.

[109] Zouros E, Ball AO, Saavedra C, Freeman KR. Mitochondrial-DNA inheritance. *Nature*, 1994, *368*: 818.

[110] Wood AR, Turner G, Skibinski DOF, Beaumont AR. Disruption of doubly uniparental inheritance of mitochondrial DNA in hybrid mussels (*Mytilus edulis*×*M. galloprovincialis*). *Heredity* 2003, *91*: 354-360.

[111] Amaral A, Paiva C, Attardo Parrinello C, Estanyol JM, Ballescà JL, Ramalho-Santos J, Oliva R. Identification of proteins involved in human sperm motility using high-throughput differential proteomics. *Journal of Proteome Research* 2014, *13*: 5670-5684.

[112] Zhao C, Guo XJ, Shi ZH, Wang FQ, Huang XY, Huo R, Zhu H, Wang XR, Liu JY, Zhou ZM, Sha JH. Role of translation by mitochondrial-type ribosomes during sperm capacitation: an analysis based on a proteomic approach. *Proteomics* 2009, *9*: 1385-1399.

[113] Martínez-Heredia J, de Mateo S, Vidal-Taboada JM, Ballescà JL, Oliva R. Identification of proteomic differences in asthenozoospermic sperm samples. *Human Reproduction* 2008, *23*: 783-791.

[114] Kumar V, Kota V, Shivaji S. Hamster sperm capacitation: role of pyruvate dehydrogenase A and dihydrolipoamide dehydrogenase 1. *Biology of Reproduction* 2008, *79*: 190-199.

[115] Siva AB, Panneerdoss S, Sailasree P, Singh DK, Kameshwari DB, Shivaji S. Inhibiting sperm pyruvate dehydrogenase complex and its E3 subunit, dihydrolipoamide dehydrogenase affects fertilization in syrian hamsters. *PloS One* 2014, *9*: e97916.

[116] Ghiselli F, Breton S, Milani L. Mitochondrial activity in gametes and uniparental inheritance: a comment on 'What can we infer about the origin of sex in early eukaryotes?'. *Philosophical Transactions of the Royal Society B* 2018, *373*: 20170147.

[117] Ji G, Gu A, Wang Y, Huang C, Hu F, Zhou, Y *et al*. Genetic variants in antioxidant genes are associated with sperm DNA damage and risk of male infertility in a Chinese population. *Free Radical Biology and Medicine* 2012, *52*: 775-780.

[118] Saadi HAS, van Riemsdijk E, Dance AL, Rajamanickam GD, Kastelic JP, Thundathil JC. Proteins associated with critical sperm functions and sperm head shape are differentially expressed in morphologically abnormal bovine sperm induced by scrotal insulation. *Journal of Proteomics* 2013, *82*: 64-80.

[119] Kobayashi T, Miyazaki T, Natori M, Nozawa S. Protective role of superoxide dismutase in human sperm motility: superoxide dismutase activity and lipid peroxide in human seminal plasma and spermatozoa. *Human Reproduction* 1991, *6*: 987-991.

[120] van Gestel RA, Brewis IA, Ashton PR, Brouwers JF, Gadella BM. Multiple proteins present in purified porcine sperm apical plasma membranes interact with the zona pellucida of the oocyte. *Molecular Human Reproduction* 2007, *13*: 445-454.

[121] Müller M, Mentel M, van Hellemond JJ, Henze K, Woehle C, Gould SB, Yu RY, van der Giezen M, Tielens AG, Martin WF. Biochemistry and evolution of anaerobic energy metabolism in eukaryotes. *Microbiology and Molecular Biology Reviews* 2012, *76*: 444-495.

[122] Bunik V, Artiukhov A, Aleshin V, Mkrtchyan G. Multiple forms of glutamate dehydrogenase in animals: structural determinants and physiological implications. *Biology* 2016, *5*: 53.

[123] Visconti PE. Sperm bioenergetics in a nutshell. *Biology of reproduction* 2012, *87*: 72-1.

[124] Nakamura N, Dai Q, Williams J, Goulding EH, Willis WD, Brown PR, Eddy EM. Disruption of a spermatogenic cell-specific mouse enolase 4 (eno4) gene causes sperm structural defects and male infertility. *Biology of reproduction* 2013, *88*: 90.

[125] Strong SJ, Ellington WR. Horseshoe crab sperm contain a unique isoform of arginine kinase that is present in midpiece and flagellum. *Journal of Experimental Zoology Part A: Ecological Genetics and Physiology* 1993, *267:* 563-571.

[126] Tombes RM, Shapiro BM. Energy transport and cell polarity: Relationship of phosphagen kinase activity to sperm function. *Journal of Experimental Zoology* 1989, *251*: 82-90.

[127] Tombes RM, Brokaw CJ, Shapiro BM. Creatine kinase-dependent energy transport in sea urchin spermatozoa. Flagellar wave attenuation and theoretical analysis of high energy phosphate diffusion. *Biophysical Journal* 1987, *52*: 75-86.

[128] Ellington WR. Evolution and physiological roles of phosphagen systems. *Annual Review of Physiology* 2001, *63*: 289-325.

[129] Yi YJ, Sutovsky M, Kennedy C, Sutovsky P. Identification of the inorganic pyrophosphate metabolizing, ATP substituting pathway in mammalian spermatozoa. *PloS One* 2012, *7*: e34524.

[130] Diz AP, Dudley E, MacDonald BW, Piña B, Kenchington EL, Zouros E, Skibinski DOF. Genetic variation underlying protein expression in eggs of the marine mussel *Mytilus edulis*. *Molecular & Cellular Proteomics* 2009, *8*: 132-144.

[131] Sutovsky P, Moreno RD, Ramalho-Santos J, Dominko T, Simerly C, Schatten G. Ubiquitinated sperm mitochondria, selective proteolysis, and the regulation of mitochondrial inheritance in mammalian embryos. *Biology of Reproduction* 2000, *63*: 582-590.

[132] Berger KH, Yaffe MP. Prohibitin family members interact genetically with mitochondrial inheritance components in *Saccharomyces cerevisiae*. *Molecular and Cellular Biology* 1998, *18*: 4043-4052.

[133] Sutovsky P, Van Leyen K, McCauley T, Day BN, Sutovsky M. Degradation of paternal mitochondria after fertilization: implications for heteroplasmy, assisted reproductive technologies and mtDNA inheritance. *Reproductive Biomedicine Online* 2004, *8*: 24-33.

[134] Song WH, Yi YJ, Sutovsky M, Meyers S, Sutovsky P. Autophagy and ubiquitin–proteasome system contribute to sperm mitophagy after mammalian fertilization. *Proceedings of the National Academy of Sciences of the USA* 2016, *113*: E5261-E5270.

[135] Roy A, Lin YN, Agno JE, DeMayo FJ, Matzuk MM. Absence of tektin 4 causes asthenozoospermia and subfertility in male mice. *The FASEB Journal* 2007, *21*: 1013-1025.

[136] Siva AB, Kameshwari DB, Singh V, Pavani K, Sundaram CS, Rangaraj N, Deenadayal M, Shivaji S. Proteomics-based study on asthenozoospermia: differential expression of proteasome alpha complex. *Molecular Human Reproduction* 2010, *16*: 452-462.

[137] Thepparat T, Katawatin S, Vongpralub T, Duangjinda M, Thammasirirak S, Utha A. Separation of bovine spermatozoa proteins using 2D-PAGE revealed the relationship between tektin-4 expression patterns and spermatozoa motility. *Theriogenology* 2012, *77*: 1816-1821.

[138] Ashrafzadeh A, Karsani SA, Nathan S. Mammalian sperm fertility related proteins. *International Journal of Medical Sciences* 2013, *10*: 1649.

[139] Bhilawadikar R, Zaveri K, Mukadam L, Naik S, Kamble K, Modi D, Hinduja I. Levels of Tektin 2 and CatSper 2 in normozoospermic and oligoasthenozoospermic men

and its association with motility, fertilization rate, embryo quality and pregnancy rate. *Journal of Assisted Reproduction and Genetics* 2013, *30*: 513-523.

[140] Mariappa D, Aladakatti RH, Dasari SK, Sreekumar A, Wolkowicz M, van der Hoorn F, Seshagiri PB. Inhibition of tyrosine phosphorylation of sperm flagellar proteins, outer dense fiber protein-2 and tektin-2, is associated with impaired motility during capacitation of hamster spermatozoa. *Molecular Reproduction and Development* 2010, *77*: 182-193.

[141] Nomikos M, Sanders JR, Parthimos D, Buntwal L, Calver BL, Stamatiadis P, Smith A, Clue M, Sideratou Z, Swann K, Lai FA. Essential role of the EF-hand domain in targeting sperm phospholipase Cζ to membrane phosphatidylinositol 4, 5-bisphosphate (PIP2). *Journal of Biological Chemistry* 2015, *290*: 29519-29530.

[142] Wang D, Hu J, Bobulescu IA, Quill TA, McLeroy P, Moe OW, Garbers DL. A sperm-specific Na^+/H^+ exchanger (sNHE) is critical for expression and in vivo bicarbonate regulation of the soluble adenylyl cyclase (sAC). *Proceedings of the National Academy of Sciences* 2007, *104*: 9325-9330.

[143] Di Sole F, Vadnagara K, Moe OW, Babich V. Calcineurin homologous protein: a multifunctional Ca^{2+}-binding protein family. *American Journal of Physiology-Renal Physiology* 2012, *303*: F165-F179.

[144] Fernandes D, Loi B, Porte C. Biosynthesis and metabolism of steroids in molluscs. *The Journal of Steroid Biochemistry and Molecular Biology* 2011, *127*: 189-195.

[145] Rotem R, Zamir N, Keynan N, Barkan D, Breitbart H, Naor Z. Atrial natriuretic peptide induces acrosomal exocytosis of human spermatozoa. *American Journal of Physiology-Endocrinology and Metabolism* 1998, *274*: E218-E223.

[146] Zhang M, Hong H, Zhou B, Jin S, Wang C, Fu M, Wang S, Xia, G. The expression of atrial natriuretic peptide in the oviduct and its functions in pig spermatozoa. *Journal of Endocrinology* 2006, *189*: 493-507.

[147] Stewart DT, Jha M, Breton S, Hoeh WR, Blier PU. No effect of sperm interactions or egg homogenate on sperm velocity in the blue mussel, *Mytilus edulis* (Bivalvia: Mytilidae). *Canadian Journal of Zoology* 2012, *90*: 1291-1296.

[148] Beekman M, Nieuwenhuis B, Ortiz-Barrientos D, Evans JP. Sexual selection in hermaphrodites, sperm and broadcast spawners, plants and fungi. *Philosophical Transactions of the Royal Society B* 2016, *371*: 20150541.

[149] Bode M., Marshall DJ. The quick and the dead? Sperm competition and sexual conflict in sea. *Evolution* 2007, *61*: 2693-2700.

[150] Sunday JM, Hart MW. Sea star populations diverge by positive selection at a sperm-egg compatibility locus. *Ecology and Evolution* 2013, *3*: 640-654.

[151] Vicens A, Lüke L, Roldan ERS. Proteins involved in motility and sperm-egg interaction evolve more rapidly in mouse spermatozoa. *Plos One* 2014, *9*:e91302.

[152] Yund PO. How severe is sperm limitation in natural populations of marine free-spawners?. *Trends in Ecology & Evolution* 2000, *15*: 10-13.

[153] Evans JP, Sherman CD. Sexual selection and the evolution of egg-sperm interactions in broadcast-spawning invertebrates. *The Biological Bulletin* 2013, *224*: 166-183.

[154] Fitzpatrick JL, Lüpold S. Sexual selection and the evolution of sperm quality. *Molecular Human Reproduction* 2014, *20*: 1180-1189.

[155] Levitan DR. Sperm velocity and longevity trade off each other and influence fertilization in the sea urchin *Lytechinus variegatus*. *Proceedings of the Royal Society of London B: Biological Sciences* 2000, *267*: 531-534.

[156] Kamp G, Büsselmann G, Lauterwein J. Spermatozoa: models for studying regulatory aspects of energy metabolism. *Experientia* 1996, *52*: 487-494.

[157] Dorus S, Wasbrough ER, Busby J, Wilkin EC, Karr TL. Sperm proteomics reveals intensified selection on mouse sperm membrane and acrosome genes. *Molecular Biology and Evolution* 2010, *27*: 1235-1246.

[158] Zigler KS, McCartney MA, Levitan DR, Lessios HA. Sea urchin bindin divergence predicts gamete compatibility. *Evolution* 2005, *59*: 2399-2404.

[159] Kenchington EL, Hamilton L, Cogswell A, Zouros E. Paternal mtDNA and maleness are co-inherited but not causally linked in Mytilid mussels. *PLoS One* 2009, 4: e6976.

[160] Brannock PM, Roberts MA, Hilbish TJ. Ubiquitous heteroplasmy in Mytilus spp. resulting from disruption in doubly uniparental inheritance regulation. *Marine Ecology Progress Series* 2013, *480*: 131-143.

[161] Śmietanka B, Burzyński A. Disruption of doubly uniparental inheritance of mitochondrial DNA associated with hybridization area of European *Mytilus edulis* and *Mytilus trossulus* in Norway. *Marine Biology* 2017, *164*: 209.

[162] Skibinski DOF, Beardmore JA, Cross TF. Aspects of the population-genetics of *Mytilus* (Mytilidae, Mollusca) in the British-Isles. *Biological Journal of the Linnean Society* 1983, *19*: 137-183.

[163] Saarman NP, Kober KM, Simison WB, Pogson GH. Sequence-based analysis of thermal adaptation and protein energy landscapes in an invasive blue mussel (*Mytilus galloprovincialis*). *Genome Biology and Evolution* 2017, *9*: 2739-2751.

[164] Eads AR, Evans JP, Kennington WJ. Plasticity of fertilization rates under varying temperature in the broadcast spawning mussel, *Mytilus galloprovincialis*. *Ecology and Evolution* 2016, *6*: 6578-6585.

[165] Eads AR, Kennington WJ, Evans JP. Interactive effects of ocean warming and acidification on sperm motility and fertilization in the mussel *Mytilus galloprovincialis*. *Marine Ecology Progress Series* 2016, *562*: 101-111.

[166] Oliver M, Evans JP. Chemically moderated gamete preferences predict offspring fitness in a broadcast spawning invertebrate. *Proceedings of the Royal Society of London B: Biological Sciences* 2014, *281*: 20140148.

[167] Fitzpatrick JL, Simmons LW, Evans JP. Complex patterns of multivariate selection on the ejaculate of a broadcast spawning marine invertebrate. *Evolution* 2012, *66*: 2451-2460.

[168] Liu G, Innes D, Thompson RJ. Quantitative analysis of sperm plane circular movement in the blue mussels *Mytilus edulis*, *M. trossulus* and their hybrids. *Journal of Experimental Zoology Part A: Ecological Genetics and Physiology* 2011, *315*: 280-290.

[169] El-Dessouky HT, Ettouney HM. Fundamentals of salt water desalination. Elsevier, 2002, Pages 525-563.

[170] Ospina-Álvarez A, Palomera I, Parada C. Changes in egg buoyancy during development and its effects on the vertical distribution of anchovy eggs. *Fisheries Research* 2012, *117*: 86-95.

[171] Sundby S, Kristiansen T. The principles of buoyancy in marine fish eggs and their vertical distributions across the world oceans. *Plos One* 2015, *10*: e0138821

[172] Reinhardt K, Dobler R, Abbott J. An ecology of sperm: sperm diversification by natural selection. *Annual Review of Ecology, Evolution, and Systematics* 2015, *46*: 435-459.

Figure Legends

Figure 1: Histological tests of mature male gonads of the six *Mytilus edulis* (a-f) and six *M. galloprovincialis* (g-l) mussels selected to make each pool for RNA-seq analysis. There are two different zoom views (see 500 and 50 µm scale respectively, above and below) shown for each histology test and individual mussel. Ac: male gonadal follicles with spermatozoa (sp), where heads (hd) and flagella (fl) can be seen and differentiated. Adipogranular (ag) and vesicular connective tissue (cv) cells can be found between the spermatic acini.

Figure 2: a) Distribution of Level 2 GO terms of loci annotated in three ontological categories: biological process (BP), molecular function (MF) and cellular component (CC). Note that only those GO terms with annotations in at least 100 and 10 loci, for BP and MF respectively are shown. **b**) Enrichment analysis results for GO terms in differentially expressed loci between mature male gonads of the two *Mytilus* spp. according to Fisher's exact test (FDR<0.05). DE: differentially expressed, ND: not differentially expressed set of loci defined after RSEM analysis. Length of bars represents the percentage of loci annotated for each term in the DE (blue bars) and ND (red bars) sets. A blue longer than red bar indicates that that GO term is overrepresented in the differentially expressed loci. GO terms are grouped by their ontological category (BP, MF, CC), and within category, GO terms are displayed sorted by increasing *p*-values.

Figure 3: 2DE gels showing sperm proteome from a representative *Mytilus galloprovincialis* and *M. edulis* mussel respectively. 45 spots that showed significant differences between the two *Mytilus* populations and species ($q \leq 0.05$) and were identified (all except one) by MS (see Table 3) are numbered and encircled.

Figure 4: Hierarchical clustering and heat map made using log normalised expression data for the 45 protein spots of sperm samples that showed significant differences in expression level ($q \leq 0.05$) between the two *Mytilus* species and populations (SW:

Swansea, VG: Vigo) and were identified (all except one) by MS (see Figure 3). Each column and row contains information for an individual mussel and protein spot respectively. The numbers on the right are the protein spot numbers to each of which isattached an abbreviation that corresponds to gene name that code for the identified protein (see Table 3). Note that for two identified protein spots (1101 and 1508) there are no gene name abbreviations available. Cells are coloured according to z-scores, showing up-regulation (red) or down-regulation (green) of protein spot volumes in the individual mussels compared with average expression values calculated from all mussel samples.

Figure 5: Volcano plot made with the 727 sperm protein spots analysed by 2DE. Log_2 of the ratio of average expression values between Swansea and Vigo populations (FC) plotted against log_{10} of *p*-values derived from the one-way ANOVA analysis. Note that positive and negative Log_2 (FC) values mean higher expression on average in samples from Vigo (*M. galloprovincialis*) and Swansea (*M. edulis*), respectively. Grey (FC>1.5) and black (up to 1.5 FC) represent non-significant protein spots (p>0.05), while colour represents protein spots significant after one-way ANOVA ($p \leq 0.05$); blue, <1.5 FC; red, between 1.5 and 2.0 FC; green, >2.0 FC.

Figure 6: Comparative results of protein spot identifications by MS using different customised protein databases (see Materials and Methods). Bars represent the total number of peptide spectrum matches (PSMs), total peptides (TP) and unique peptides (UP), expressed as percentage, obtained against each of the three protein databases made from: 1) RNA-seq data from the current study (RNA), 2) EST sequences available in NCBI from Mytilus[organism] (EST), and 3) protein sequences available in NCBI for Mollusca[organism] (NCBI). *: $p<0.001$, ns: not significant, for Kruskal-Wallis and *post-hoc* pairwise tests (after Dunn correction to account for multiple comparisons) between the different protein databases either for the total number of PSMs, TP or UP.

Figure 7: Summary of counts and percentages of sperm trait and functional terms for proteins having higher expression in *M. edulis* and *M. galloprovincialis*. The data is

derived from Table 3 and from File S7, worksheet Table S7 where it is further elaborated (see captions of Tables S6-S7). Columns 2-5 give the counts and % values of sperm trait terms assigned to proteins having higher expression in *M. edulis* and *M. galloprovincialis*. Red and green fill indicate higher and lower % values in each row. Columns 6 and 7 indicate the number of occurrences of terms according to a tentative hypothesis on perceived benefit of higher expression to the species at the head of the columns (in red font) or perceived disadvantage (green font). Black font indicates that a conclusion in relation to benefit or disadvantage could not easily be made.

Table 1: Summary results from RNA-seq data and annotation through Blast analysis against different databases: 1) all protein sequences available in SwissProt (UniProtKB/SwissProt), 2) the Pacific oyster *Crassostrea gigas* genome (Oyster_Genome), 3) all EST sequences available in NCBI from "Mytilus", 4) protein sequences retrieved from NCBI for "Mytilus" (NCBI_MytProt), and 5) protein sequences retrieved from NCBI for "Mollusca" (NCBI_MolluscaProt). See further details in materials and methods.

Number of reads (raw / filtered)	235,967,540 / 187,829,361
Number of Isotigs	97,425
Number of Loci	49,713
Maximum sequence length (bp)	13,604
Mean / Median sequence length (bp)	706 / 434
N50 length (bp)	1,071
Number of Loci identified following:	
BlastX (UniProtKB/SwissProt)	13,498 (27.1% of total loci)*
tBlastX (Oyster_Genome)	18,279 (36.8%)
tBlastX (NCBI_MytESTs)	31,428 (63.2%); *database coverage [56,253 of total 67,990 MytEST sequences (82.7%)]*
BlastX (NCBI_MytProt)	2,234 (4.5%); *database coverage [5,153 of total 6338 MytProt sequences (81.3%)]*
BlastX (NCBI_MolluscaProt)	17,529 (35.3%); *database coverage [70,317 of total 190,951 MolluscaProt sequences (36.8%)]*

(*) 13,283 loci were functionally annotated using Blast2GO, including InterProScan.

Table 2: Transcripts (loci) showing significant differences (FDR 1% at isotig level) in expression of mature
edulis (mussels from Swansea, E) and *M. galloprovincialis* (mussels from Vigo, G), with GO or protein na
term string "SPERM*" OR "FERT*" and a prediction that they have a signal peptide (SP) or a transmembra
this later information coming from SignalP 4.1, TMHMM 2.0 and InterProScan 5.0 analysis. Transcripts
Blast2GO against UniProt-SwissProt database [all organisms], but protein names below are derived by che
protein database. The numbers of significant isotigs from each locus (FDR 1%) with higher expression
galloprovincialis (E<G) and vice-versa (G>E) are also displayed.

Transcript #	Gene name	Protein name (nrNCBI [Mollusca])	Function
Locus_2854	Iap2	Apoptosis 2 inhibitor [*C. gigas*]	Spermatogenesis, acrosome rea
Locus_3972	Tmbim6	Bax inhibitor-1 protein [*M. galloprovincialis*]	Spermatogenesis, acrosome rea
Locus_9050	Bre-4	Beta-1,4-N-acetylgalactosaminyltransferase bre-4 [*C. gigas*]	Sperm-egg interaction
Locus_1384	CtsB	Cathepsin B [*C. ariakensis*]	Spermatogenesis, acrosome rea
Locus_175 Locus_2547	CtsL	Cathepsin L [*C. gigas*]	Spermatogenesis, acrosome rea
Locus_587	CtsL2	Cathepsin L2 cysteine protease [*P. fucata*]	Spermatogenesis, acrosome rea
Locus_6135	Cdc42	Cell division cycle 42 [*Mytilus* sp. ZED-2008]	Sperm capacitation, acrosome r
Locus_24960	Cht3	Chitinase-3 [*H. cumingii*]	Sperm-egg interaction
Locus_6902	Cdyl2	Chromodomain Y-like protein 2 [*C. gigas*]	Spermatogenesis
Locus_1290	Cng	Cyclic nucleotide-gated channel rod photoreceptor sub.	Spermatogenesis

		alpha [*C. gigas*]	
Locus_1433	Dnal1	Dynein light chain 1, axonemal, partial [*C. gigas*]	Sperm motility
Locus_2552	Eif4g2	Eukaryotic translation initiation factor 4 gamma 2 [*C. gigas*]	Spermatogenesis
Locus_5126	Ggnbp2	Gametogenetin-binding protein 2 [*C. gigas*]	Spermatogenesis
Locus_134	Hsp90	Heat shock protein 90 [*M. galloprovincialis*]	Spermatogenesis
Locus_22899	Prdm9	Histone-lysine N-methyltransferase PRDM9 [*C. gigas*]	Spermatogenesis
Locus_18746	Suv39h2	Histone-lysine N-methyltransferase SUV39H2 [*C. gigas*]	Spermatogenesis
Locus_6027	Hya	Hyaluronidase [*C. gigas*]	Sperm-egg interaction
Locus_1259 Locus_12988	Irs	Insulin-related peptide receptor [*P. fucata*]	Spermatogenesis
Locus_5663	Ift172	Intraflagellar transport protein 172 homolog, predicted [*A. californica*]	Sperm motility
Locus_2244	Imp2	Mitochondrial inner membrane protease subunit 2 [*C. gigas*]	Spermatogenesis
Locus_10336	Nphp1	Nephrocystin-1 [*C. gigas*]	Spermatogenesis
Locus_9945	Pmca	Plasma membrane calcium ATPase [*P. fucata*]	Sperm motility
Locus_1143	Phb	Prohibitin [*O. tankahkeei*]	Spermatogenesis
Locus_1157	Phb2	Prohibitin-2-like, predicted [*A. californica*]	Spermatogenesis
Locus_19017	Pc1	Prohormone convertase 1 [*H. diversicolor sup.*]	Sperm-egg interaction, sperm capacitation, sperm motility
Locus_2686	Psma2	Proteasome subunit alpha type-2 [*C. gigas*]	Sperm capacitation, acrosome r

Locus_29609	Rarb	Retinoic acid receptor beta [*C. gigas*]	Spermatogenesis
Locus_29136	Ropn1	Ropporin-1-like protein [*C. gigas*]	Spermatogenesis, sperm motilit
Locus_815	Sqstm1	Sequestosome-1 [*C. gigas*]	Spermatogenesis
Locus_9081	Slc6a5	Sodium- and chloride-dependent glycine transporter 2 [*C. gigas*]	Sperm motility
Locus_3269	Slc9c1	Sodium/hydrogen exchanger 10 [*C. gigas*]	Spermatogenesis, sperm motilit
Locus_29004	Spatc1	Speriolin [*C. gigas*]	Spermatogenesis
Locus_13213	Spa17	Sperm surface protein Sp17 [*C. gigas*]	Spermatogenesis, sperm-egg in sperm capacitation, acrosome r
Locus_12286	Spag1	Sperm-associated antigen 1 [*C. gigas*]	Sperm-egg interaction
Locus_1176 **Locus_10277**	Srsf4	Splicing factor, arginine/serine-rich 4 [*C. gigas*]	Spermatogenesis
Locus_18976	Samd7	Sterile alpha motif domain-containing protein 7 [*C. gigas*]	Spermatogenesis
Locus_1959	Slc26	Sulfate transporter-like, predicted [*A. californica*]	Sperm motility
Locus_4801	Cct2	T-complex protein 1 (TCP-1) subunit beta [*C. gigas*]	Sperm-egg interaction
Locus_586	Cct4	T-complex protein 1 (TCP-1) subunit delta [*C. gigas*]	Sperm-egg interaction
Locus_1374	Cct5	T-complex protein 1 (TCP-1) subunit epsilon [*C. gigas*]	Sperm-egg interaction
Locus_24738	Cct7	T-complex protein 1 (TCP-1) subunit eta [*C. gigas*]	Sperm-egg interaction
Locus_22131 **Locus_25048** **Locus_36832**	Cct3	T-complex protein 1 (TCP-1) subunit gamma [*C. gigas*]	Sperm-egg interaction
Locus_20775	Cct8	T-complex protein 1 (TCP-1) subunit theta [*C. gigas*]	Sperm-egg interaction

Locus_188	Cct6a	T-complex protein 1 (TCP-1) subunit zeta [*C. gigas*]	Sperm-egg interaction
Locus_8047 **Locus_29534**	Thbs1	Thrombospondin-1 [*C. gigas*]	Sperm-egg interaction
Locus_17402	Ubc8	Ubiquitin-conjugating enzyme E2-24 kDa [*C. gigas*]	Spermatogenesis
Locus_39229 **Locus_25485**	M3	vitelline coat lysin M3 [*M. edulis*]	Sperm-egg interaction
Locus_24 **Locus_30388**	M6	vitelline coat lysin M6 [*M. edulis*]	Sperm-egg interaction
Locus_3846	Zfr	Zinc finger RNA-binding protein [*C. gigas*]	Spermatogenesis
Locus_1040 **Locus_1240** **Locus_1570** **Locus_2570**	Zan	Zonadhesin [*C. gigas*]	Sperm-egg interaction

Table 3: Identification by MS/MS of 44 out of 45 protein spots (see Fig. 3) from sperm that showed significant differe
species and populations of mussels (*M. galloprovincialis* from Vigo *vs M. edulis* from Swansea). Gene, the name of t
from UniProt) that code for the protein sequence described in "Protein id" column. FC, fold change, defined as the
expression in either *M. galloprovincialis* (G) or *M. edulis* (E) mussel species. The databases from which an identificati
given in the Database column: EST, expression sequence tags from *Mytilus* spp. available in Genbank, RNA, seque
NCBI, protein sequences from *Mollusca* available in NCBI (see Materials and Methods).

Spot	Gene	Protein id	FC	Database	Cellular location	M
1205	Aco2	Aconitate hydratase	1.7 E	RNA, NCBI	Mitochondrion	T
1241			2.0 E	RNA, NCBI		c
1272	Ak	*Arginine kinase*	2.0 G	RNA, EST, NCBI	Cytoplasm	P
1744			2.1 G	RNA, EST, NCBI		
705a	Atp5a	*ATP synthase subunit alpha*	1.8 E	RNA, EST, NCBI	Mitochondrion	R
430	Npr1	*Atrial natriuretic peptide receptor 1*	1.7 G	RNA, NCBI	Membrane	H
1074	Cnn1	*Calponin protein*	4.2 G	RNA, EST, NCBI	Cytoskeleton	A
2151	Uqcrc2	Cytochrome b-c1 complex subunit 2	1.9 G	RNA, EST	Mitochondrion	R
2164a			1.6 G	RNA, EST		
847			5.6 E	RNA, EST		
705b	Dld	*Dihydrolipoyl dehydrogenase*	1.8 E	RNA, EST, NCBI	Mitochondrion	C
1119	Efhc2	EF-hand domain-containing family	4.7 G	RNA, EST, NCBI	Ubiquitous	C
1134		member C2	2.4 E	RNA, EST, NCBI		
191			2.3 G	RNA, EST, NCBI		

1536	Etfb	*Electron transfer flavoprotein subunit beta*	1.4 G	RNA, EST, NCBI	Mitochondrion	R m
801	Eno	*Enolase*	1.6 G	RNA, NCBI	Cytoplasm	C
1608	es1	es1 protein	2.1 G	RNA, EST	Mitochondrion	U
2039			3.5 E	RNA, EST, NCBI		
1602			2.4 G	RNA, EST		
589	Glud	*Glutamate dehydrogenase*	2.0 G	RNA, EST, NCBI	Mitochondrion	C
2062			2.5 E	RNA, EST, NCBI		
1265	Ppa1	*Inorganic pyrophosphatase*	2.6 G	RNA, EST, NCBI	Cytoplasm	H c
1094	Idh3a	*Isocitrate dehydrogenase [NAD] subunit alpha*	1.2 E	RNA, EST, NCBI	Mitochondrion	T
1085	Idh3g	*Isocitrate dehydrogenase [NAD] subunit gamma*	3.1 E	RNA, EST, NCBI	Mitochondrion	T
1087			1.5 E	RNA, EST		
1012	Ivd	*Isovaleryl-CoA dehydrogenase*	2.6 G	RNA, EST, NCBI	Mitochondrion	A
988a	Acadm	*Medium-chain specific acyl-CoA dehydrogenase*	1.7 G	RNA, EST, NCBI	Mitochondrion	E
2038	Sod2	*Mitochondrial manganese superoxide dismutase*	3.4 E	RNA, EST, NCBI	Mitochondrion	A
949	Ndufa10	*NADH dehydrogenase [ubiquinone] 1 alpha subcomplex subunit 10*	4.5 G	RNA, EST	Mitochondrion	R m
2108	Plc	*Perlucin*	1.8 E	RNA	Extracellular region	S
1919	Prdx5	*Peroxiredoxin-5*	2.2 E	EST, NCBI	Mitochondrion	A in
1322	Psme3	*Proteasome activator complex subunit 3*	1.7 G	RNA, EST	Cytoplasm, nucleus	P a
1503	Psma4	*Proteasome subunit alpha type-4*	1.9 G	RNA, EST, NCBI	Cytoplasm, nucleus	P a

1795	Psmb2	*Proteasome subunit beta type-2*	2.0 G	RNA, EST, NCBI	Cytoplasm, nucleus	P a
1778	Psmb6	*Proteasome subunit beta type-6*	1.6 G	EST, NCBI	Cytoplasm, nucleus	P a
590	Pfd0110w	*Reticulocyte-binding protein PFD0110w isoform X3*	1.7 G	RNA, EST	Membrane	C
2164b	Tekt1	*Tektin-1*	1.6 G	EST, NCBI	Cytoskeleton	M
1258	Tekt2	*Tektin-2*	1.7 G	RNA, EST, NCBI	Cytoskeleton	M
2084			2.5 E	RNA, EST, NCBI		
776			2.3 G	RNA, EST, NCBI		
814			2.7 G	RNA, EST, NCBI		
753	Tekt4	*Tektin-4*	1.9 G	RNA, EST, NCBI	Cytoskeleton	M
1508	--	*Uncharacterized protein LOC105318227*	3.5 G	RNA, EST	--	--
1101	--	*Uncharacterized protein ZK1073.1 isoform X2*	3.7 G	RNA, EST	--	--
988b	Psmc6	*26S protease regulatory subunit 10B*	1.7 G	EST, NCBI	Cytoplasm, nucleus	P a
901	Psmd11	*26S proteasome non-ATPase regulatory subunit 11*	1.9 G	RNA, EST, NCBI	Cytoplasm, nucleus	P a
1606	Hsd17b10	*3-hydroxyacyl-CoA dehydrogenase type-2*	1.5 E	RNA, EST, NCBI	Mitochondrion	E
97[#]			1.7 G			

[#]: due to technical problems this protein spot was not identified by MS. Note that three spots (705, 988 and 2164) were identified as

Figure 1

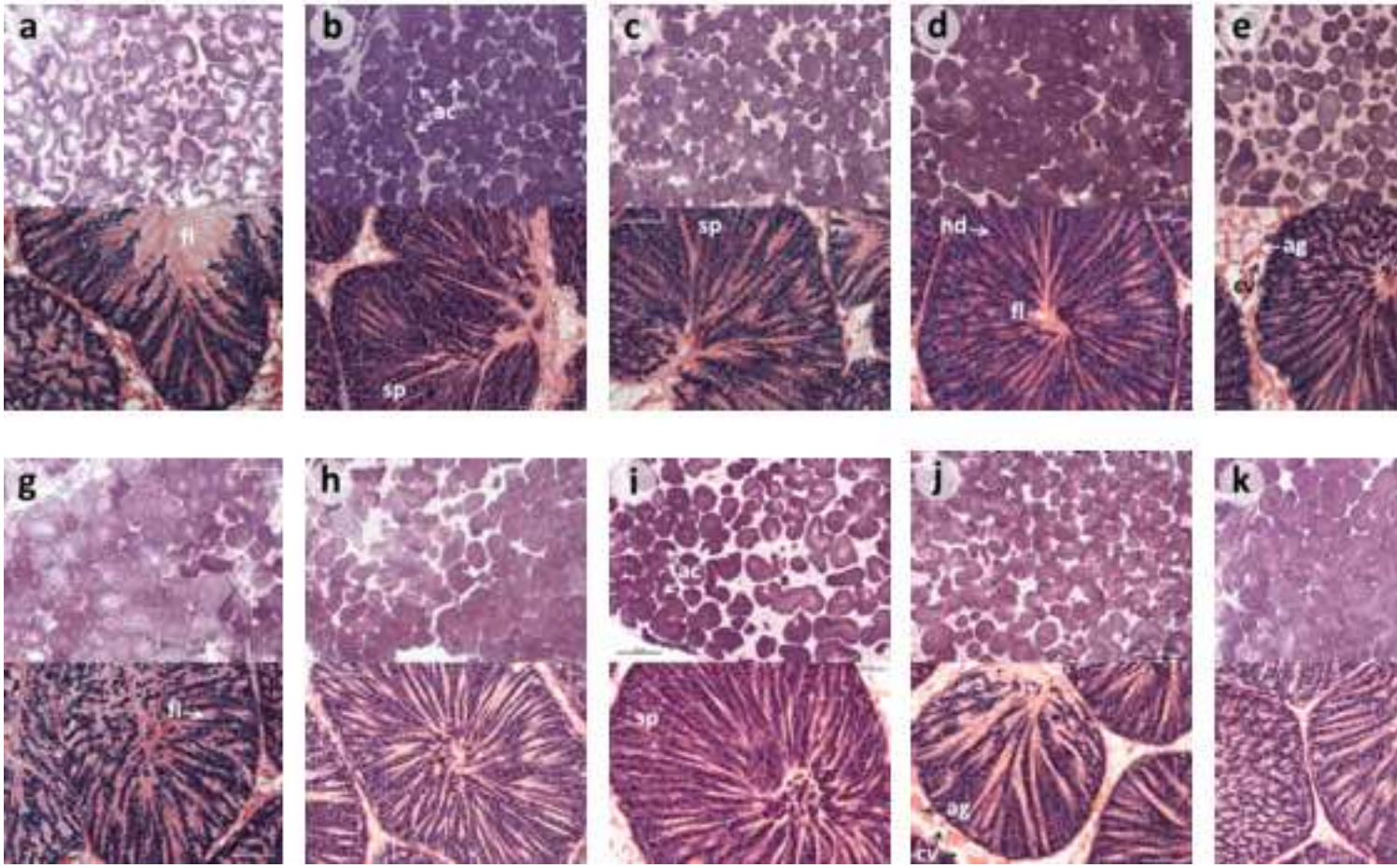

Figure 2

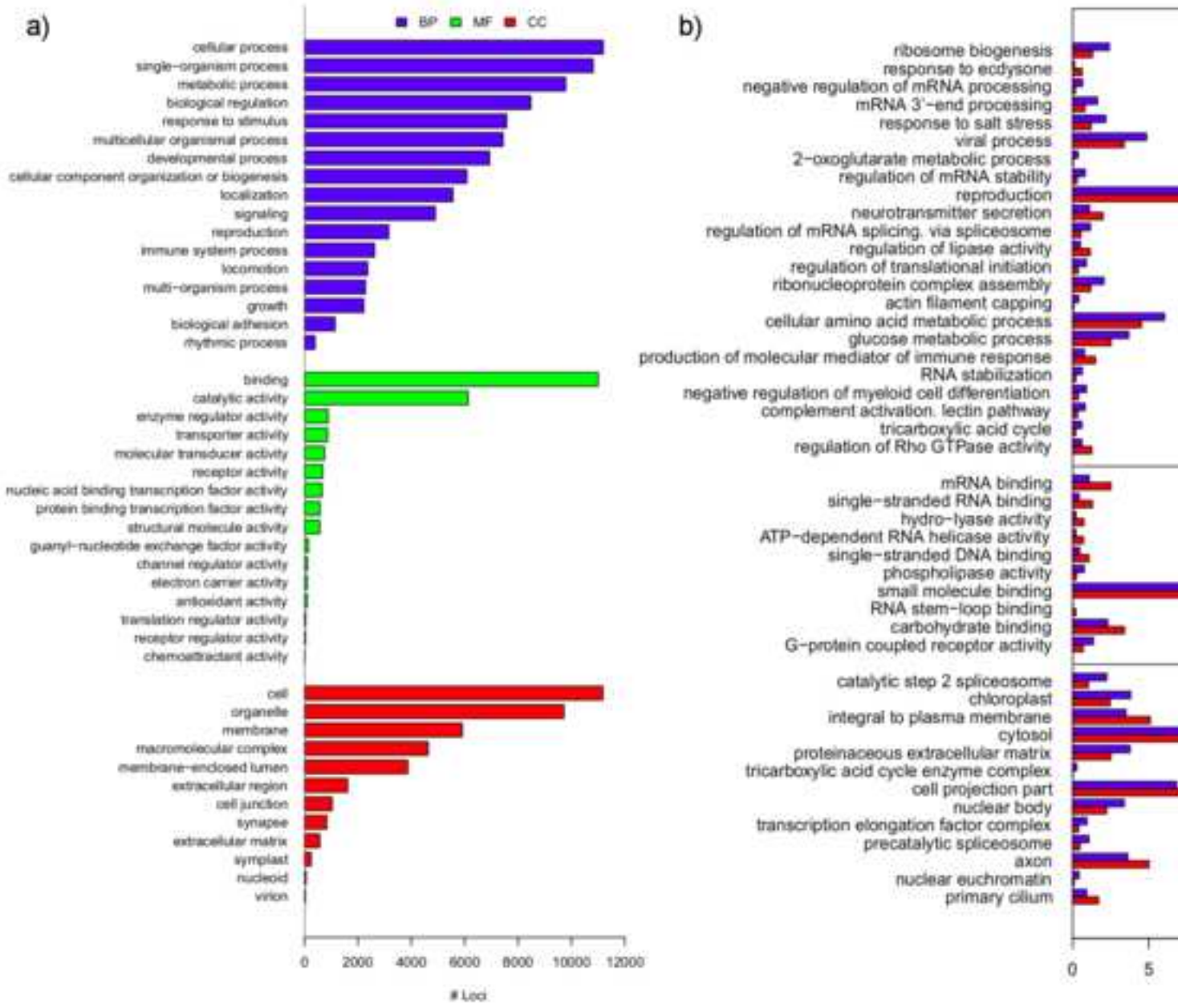
a)
BP
MF
CC
cellular process
single−organism process
metabolic process
biological regulation
response to stimulus
multicellular organismal process
developmental process
cellular component organization or biogenesis
localization
signaling
reproduction
immune system process
locomotion
multi−organism process
growth
biological adhesion
rhythmic process
binding
catalytic activity
enzyme regulator activity
transporter activity
molecular transducer activity
receptor activity
nucleic acid binding transcription factor activity
protein binding transcription factor activity
structural molecule activity
guanyl−nucleotide exchange factor activity
channel regulator activity
electron carrier activity
antioxidant activity
translation regulator activity
receptor regulator activity
chemoattractant activity
cell
organelle
membrane
macromolecular complex
membrane−enclosed lumen
extracellular region
cell junction
synapse
extracellular matrix
symplast
nucleoid
virion
0
2000
4000
6000
8000
10000
12000
Loci
b)
ribosome biogenesis
response to ecdysone
negative regulation of mRNA processing
mRNA 3'−end processing
response to salt stress
viral process
2−oxoglutarate metabolic process
regulation of mRNA stability
reproduction
neurotransmitter secretion
regulation of mRNA splicing, via spliceosome
regulation of lipase activity
regulation of translational initiation
ribonucleoprotein complex assembly
actin filament capping
cellular amino acid metabolic process
glucose metabolic process
production of molecular mediator of immune response
RNA stabilization
negative regulation of myeloid cell differentiation
complement activation, lectin pathway
tricarboxylic acid cycle
regulation of Rho GTPase activity
mRNA binding
single−stranded RNA binding
hydro−lyase activity
ATP−dependent RNA helicase activity
single−stranded DNA binding
phospholipase activity
small molecule binding
RNA stem−loop binding
carbohydrate binding
G−protein coupled receptor activity
catalytic step 2 spliceosome
chloroplast
integral to plasma membrane
cytosol
proteinaceous extracellular matrix
tricarboxylic acid cycle enzyme complex
cell projection part
nuclear body
transcription elongation factor complex
precatalytic spliceosome
axon
nuclear euchromatin
primary cilium
0
5

Figure 3

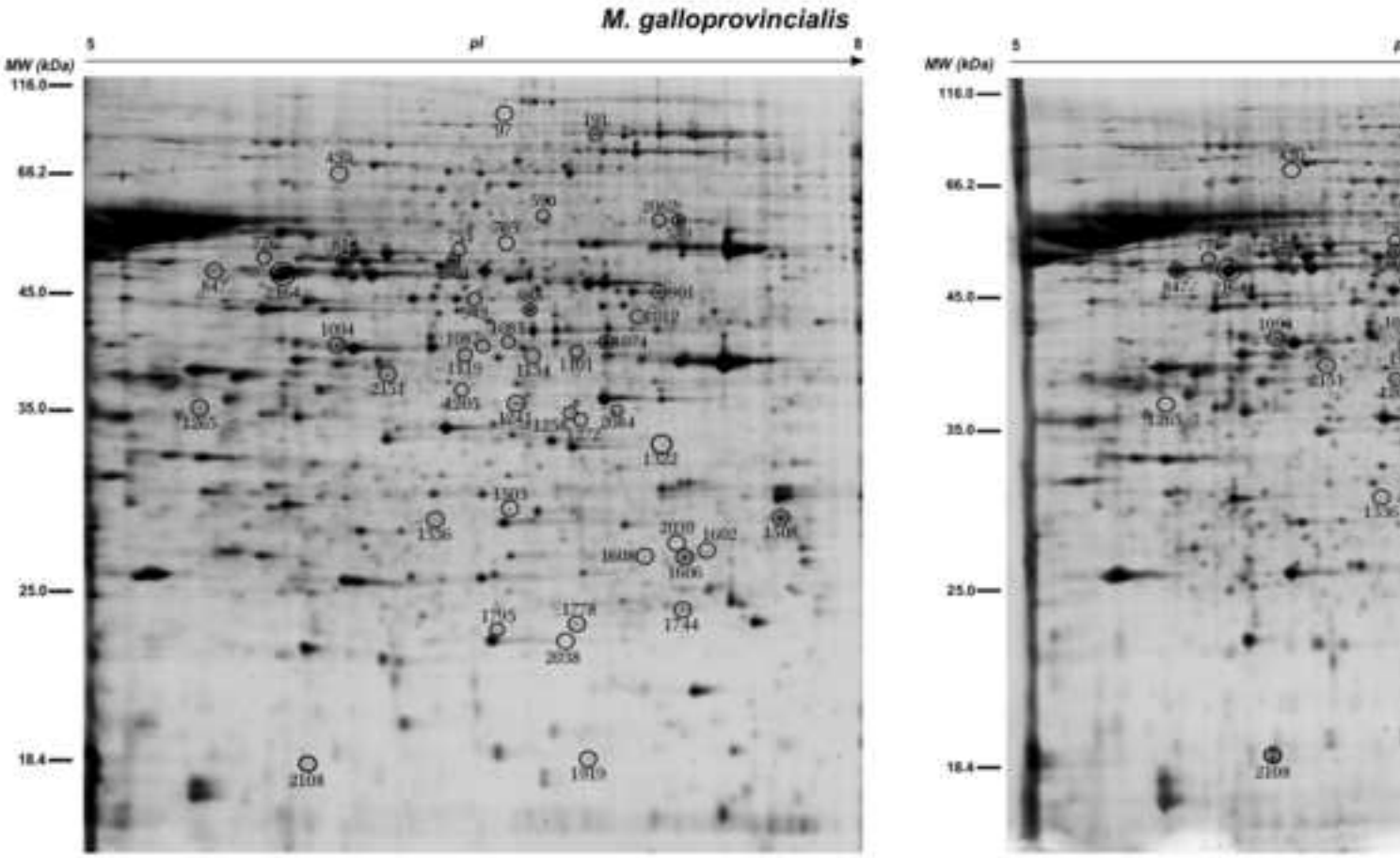

Figure 4

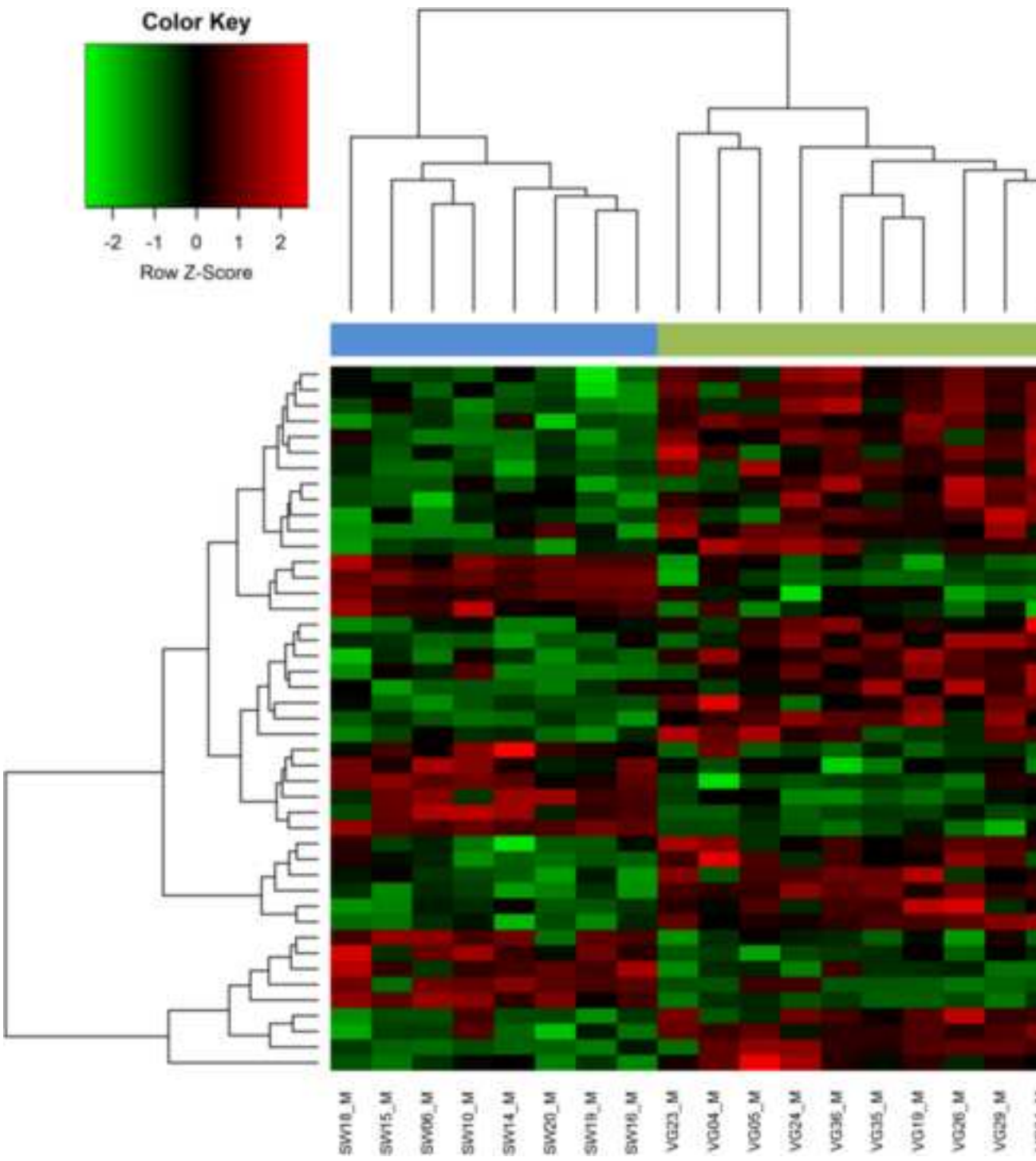
Color Key
-2
-1
0
1
2
Row Z-Score
SWI18_M
SWI15_M
SWI06_M
SWI10_M
SWI14_M
SWI20_M
SWI19_M
SWI16_M
VG03_M
VG04_M
VG05_M
VG24_M
VG36_M
VG35_M
VG19_M
VG26_M
VG29_M

Figure 5

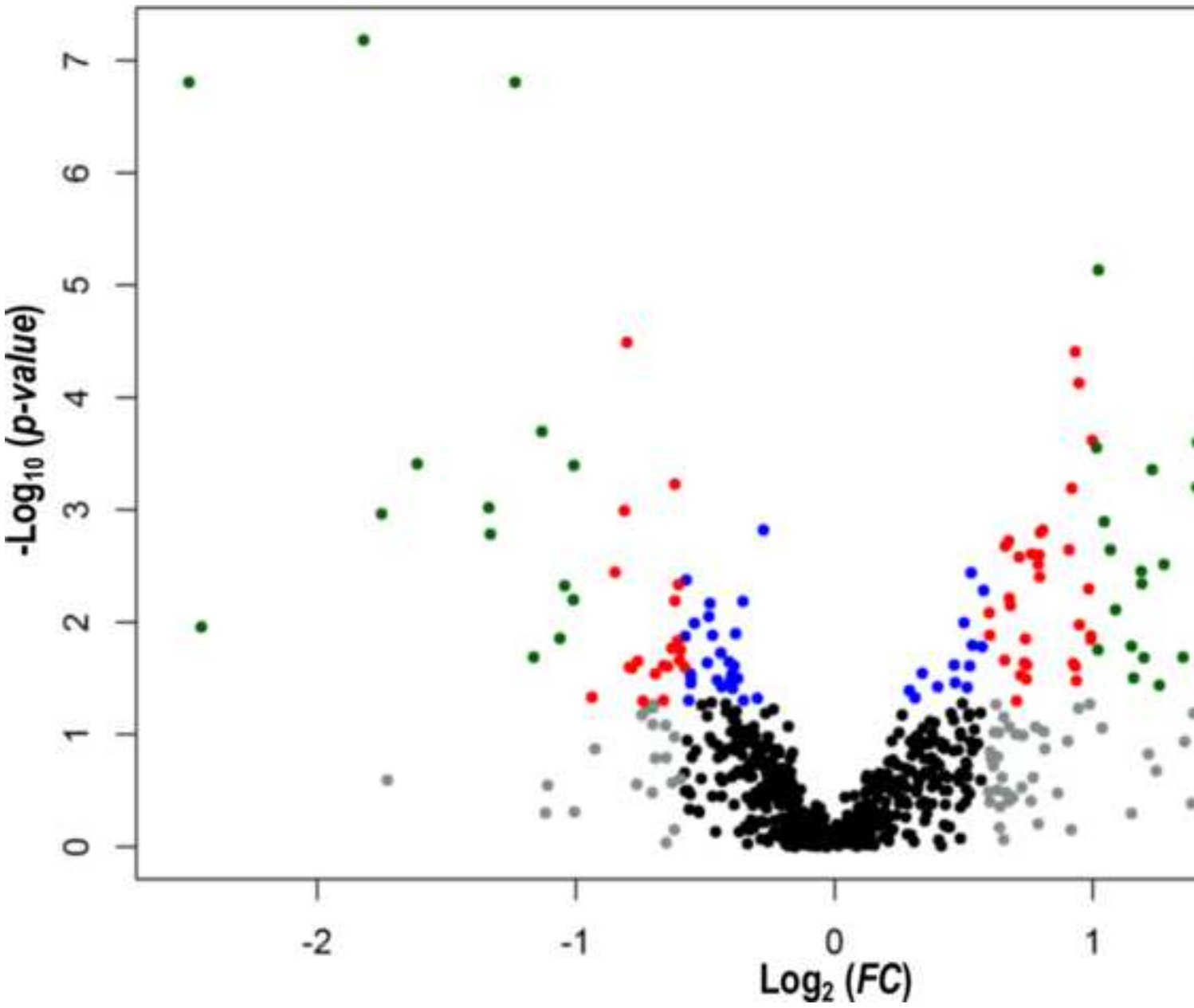

Figure 6

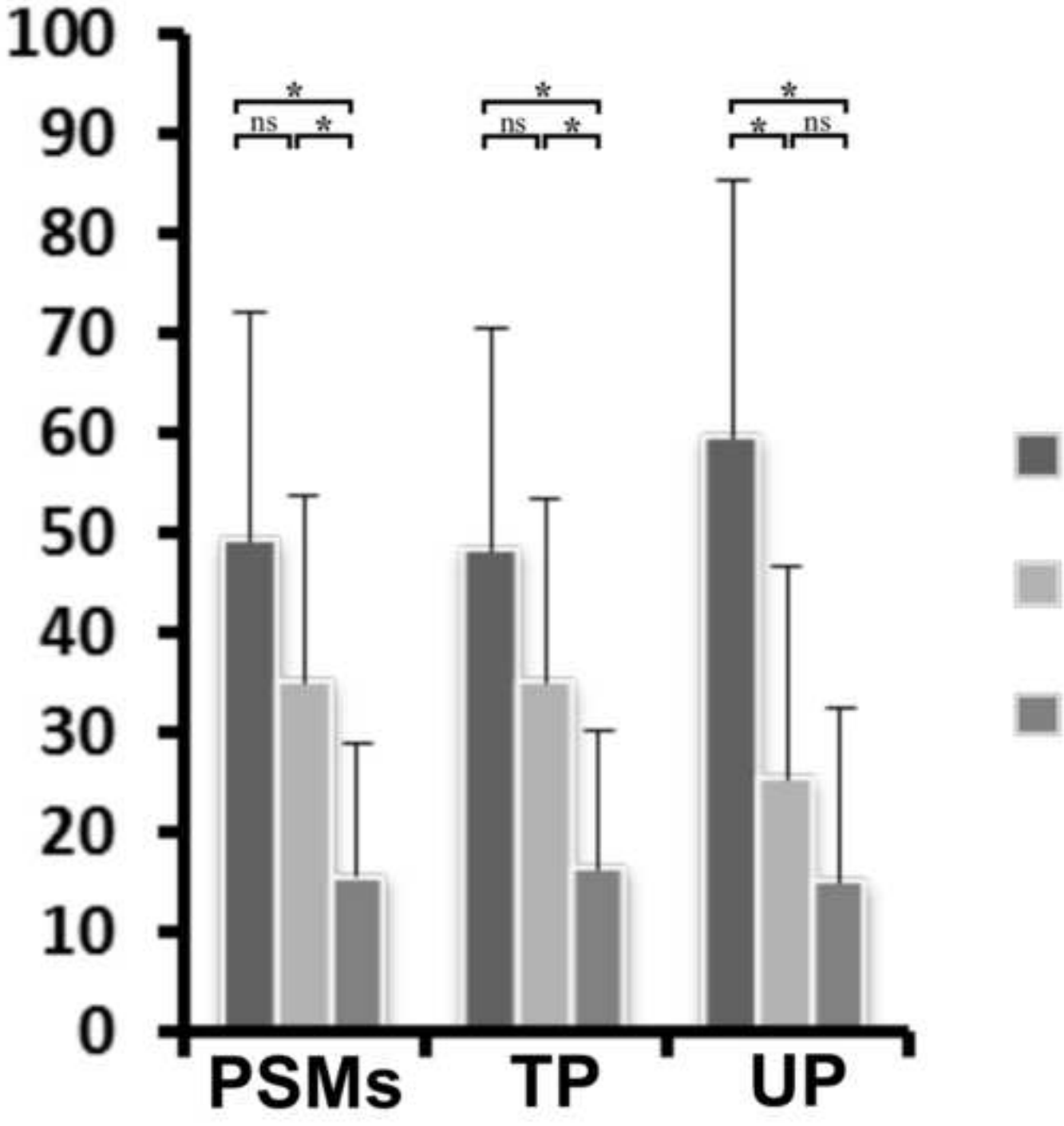

Figure 7

Sperm trait terms	Higher protein expression in *M. edulis*		Higher protein expression in *M. galloprovincialis*		P benefi *M. ed*
	Count	%	Count	%	Cou
Acrosome reaction	2	7	8	13	2
Alternative energetic pathways	0	0	4	7	0
ATP reserves	6	20	5	8	6
Capacitation	3	10	8	13	3
Chemotaxis	0	0	1	2	0
Motility	12	40	19	31	11,
Oxidative stress control	2	7	0	0	2
ROS production	3	10	0	0	3
Sperm-egg interaction	1	3	8	13	1
Swimming pattern	1	3	8	13	1