Cronfa - Swansea University Open Access Repository

This is an author produced version of a paper published in:
Journal of Proteomics

Cronfa URL for this paper:
http://cronfa.swan.ac.uk/Record/cronfa44682

Paper:
Romero, M., Pérez-Figueroa, A., Carrera, M., Swanson, W., Skibinski, D. & Diz, A. (2018). RNA-seq coupled to proteomic analysis reveals high sperm proteome variation between two closely related marine mussel species.
Journal of Proteomics
http://dx.doi.org/10.1016/j.jprot.2018.08.020

Released under the terms of a Creative Commons Attribution Non-Commercial No Derivatives License (CC-BY-NC-ND).

This item is brought to you by Swansea University. Any person downloading material is agreeing to abide by the terms of the repository licence. Copies of full text items may be used or reproduced in any format or medium, without prior permission for personal research or study, educational or non-commercial purposes only. The copyright for any work remains with the original author unless otherwise specified. The full-text must not be sold in any format or medium without the formal permission of the copyright holder.

Permission for multiple reproductions should be obtained from the original author.

Authors are personally responsible for adhering to copyright and publisher restrictions when uploading content to the repository.

http://www.swansea.ac.uk/library/researchsupport/ris-support/

*Manuscript

1

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

4

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

7

8 [1]Department of Biochemistry, Genetics and Immunology, Faculty of Biology,
9 University of Vigo, Vigo, Spain
10 [2]Marine Research Centre, University of Vigo (CIM-UVIGO), Isla de Toralla, Vigo,
11 Spain
12 [3]Institute of Marine Research, CSIC, Vigo, Spain
13 [4]Department of Genome Sciences, School of Medicine, University of Washington,
14 Seattle, USA
15 [5]Institute of Life Science, Swansea University Medical School, Swansea University,
16 Swansea, UK

17

18

19 ***Corresponding author**:

20 Dr. Angel P. Diz

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

23 Tel: +34 986813828

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

25

26 **Abstract**

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

45

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

48 **Highlights**

49 ○ *Mytilus* spp. are valuable in reproductive isolation and speciation studies.

50 ○ Gametes are key cell targets in investigations of speciation mechanisms.

51 ○ *Mytilus* spp. show proteome and transcriptome differences in male gonads and
52 sperm.

53 ○ Identified proteins are involved in sperm motility and sperm-egg interactions.

54 ○ Joint proteomic and RNA-seq analysis provide candidate proteins for evolution
55 studies.

56

57

58

59 **Significance**

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

76

77 **1. Introduction**

78 The study of the mechanisms that lead to the formation of new species is of special
79 interest in marine ecosystems due to the lack of obvious barriers to gene flow, and is
80 especially relevant in organisms with a prolonged period of larval dispersion [1]. Many
81 marine species release gametes into seawater, so fertilization occurs externally. Because
82 of this, research on speciation in marine systems has focused on the evolution of gamete
83 recognition systems because of their potential as prezygotic reproductive isolation
84 mechanisms [2-4]. The role of postzygotic mechanisms has been less studied and is
85 controversial [5] despite their potential relevance to maintain the integrity of species [6].
86 It seems obvious that gametes are key cell targets in investigations of the molecular
87 mechanisms underlying reproductive isolation. Molecular studies on gametes are
88 however quite scarce and largely restricted to a few model organisms. The molecular
89 basis of fertilisation including the sperm-egg recognition system is still a poorly
90 understood, yet basic, biological process [7-8]. In marine invertebrates such studies
91 have focused on sea urchins, starfish, clams, oysters, abalones, sea snails and worms [8-
92 9]. The use of a greater diversity of species has recently been advocated as a good way
93 to shed light on diverse questions that remain open in reproductive biology [10],
94 including the molecular basis of species-specificity gamete interactions during
95 fertilisation.

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

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

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

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

141 41]. A 2-DE based proteomic study using a somatic tissue, the foot, from two sympatric
142 *Mytilus* species (*M. edulis* and *M. galloprovincialis*) and their hybrids showed
143 differences in the protein expression patterns of hybrids when compared with the two
144 parental species, providing evidence compatible with Dobzhansky-Muller
145 incompatibilities (DMI) between both parental genomes in hybrids [42]. Thus
146 postzygotic isolation factors may also have played a role in limiting the degree of
147 introgression among genomes of *Mytilus* spp. New studies using high throughput
148 genomics and proteomics on gametes should provide a significantly better understanding
149 of the molecular mechanisms underlying reproductive isolation and evolution of *Mytilus*
150 spp.

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

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

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

181

182 2. Materials and Methods

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

185 2.1. Transcriptome (RNA-seq) analysis of mature male gonad tissues from two
186 *Mytilus* spp.

187 *2.1.1. Sampling and histological analysis*

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

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

205 *2.1.2. RNA extraction, mRNA library and Illumina paired-end sequencing*

206 RNA extraction was carried out using a protocol based on the Qiagen RNeasy® Mini kit
207 (Qiagen, Valencia, CA, USA) with tissue homogenization in QIAshredder columns
208 (Qiagen). The quantification of RNA samples was carried out using a NanoDrop 1000
209 Spectrophotometer (Thermo scientific, DE, USA), and the RNA quality was assessed in
210 an Agilent 2100 bioanalyzer (Agilent Technologies, CA, USA). Total RNA extracts
211 from these selected samples were used to make two pools of 6 individuals each, one
212 pool for each of the two *Mytilus* species. 700 ng of RNA per individual sample was
213 used, so each pool contained 4.2 µg of total RNA. mRNA libraries were generated using
214 the Illumina Truseq Small RNA Preparation kit (Illumina, CA, USA) according to
215 Illumina's TruSeq Small RNA Sample Preparation Guide v2 (low sample protocol).
216 Agarose gel-based selection was carried out to obtain libraries with fragments close to
217 500 bp in length, and their quality was assessed through Bioanalyzer profiles using a
218 high sensitivity DNA chip. Finally, libraries were quantified, by using quantitative PCR
219 with specific primers complementary to the library adapters and KAPA SYBR FAST
220 Universal qPCR Kit (Kapa Biosystems, MA, USA), and diluted to 12 pM before
221 sequencing. Each library, corresponding to each of the two pools, was analysed in a full
222 line of the flow cell from an Illumina HiScanSQ instrument (Illumina) and using TruSeq
223 SBS v3 chemistry (Illumina) to generate 2 × 100 bases long paired-end reads. After
224 sequencing, data were acquired and analysed by using the Genome Analyzer
225 Sequencing Control Software (SCS 2.6) and Real Time Analyser (RTA 1.6) software
226 from Illumina. A total of 124,102,082 and 111,865,458 raw reads were obtained from
227 the *Mytilus edulis* and *Mytilus galloprovincialis* pooled samples respectively. Raw data
228 were deposited into SRA-NCBI database (BioProject ID: PRJNA451093). The quality
229 control and filtering of nucleotide sequences was carried out as explained in Ref. [47],
230 yielding 187,829,361 confident reads that were used for *de novo* assembly and
231 generation of a consensus transcriptome.

232 *2.1.3. De novo transcriptome assembly and functional annotation*

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

235 follow a *de novo* assembly approach in order to build a consensus transcriptome from
236 mature male gonad from both *Mytilus* spp. Thus, reads from both *Mytilus* species were
237 assembled to generate a set of contigs (herein isotigs). The full set of isotigs should
238 represent the majority of transcribed genes in this specific tissue in either one or both
239 *Mytilus* species. This approach allowed the comparison of the expression levels from
240 the different isotigs between samples of the two species. *De novo* transcriptome
241 assembly was carried out by using Velvet followed by Oases software [49-50]. Oases
242 uses the preliminary assembly made by Velvet to complete the assembling of reads into
243 isotigs. Finally, it clusters the isotigs into small groups called loci (synonymous with the
244 term isogroups, also used in the literature), representing the consensus transcriptome of
245 the samples under study. These are not genetic loci, but rather a collection of similar
246 sequences (isotigs), which might include different splice variants, alleles and partial
247 assemblies of longer transcripts. Hence, it might be said that there are different isotigs
248 for each locus (consensus transcript). Nevertheless, many loci contain only one isotig,
249 though some others may contain hundreds of isotigs. The generated consensus
250 transcriptome was annotated against a non-redundant UniProtKB/SwissProt sequence
251 database using the program BlastX [51]. For comparative purposes the annotation was
252 repeated against the published genome of another marine bivalve the Pacific oyster
253 *Crassostrea gigas* [52], against all EST sequences available in NCBI from
254 "Mytilus"[organism], and against two protein databases with sequences retrieved from
255 NCBI either for "Mytilus"[Organism] or "Mollusca"[Organism] using a threshold e-
256 value of $1x10^{-3}$. Functional annotation based on Gene Ontology (GO) terms was
257 performed using the tool Blast2GO [53]. An enrichment analysis of GO terms was
258 carried out for those transcripts that showed significant differences between samples of
259 the two *Mytilus* spp. (see below) using Fisher's exact test with a FDR=5% (see Ref. [47]
260 for further details on method). This might provide some clues about the differences at
261 functional level present in mature male gonad tissue of the two *Mytilus* spp.
262

263 *2.1.4. Differential expression analyses*

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

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

282 **2.2. Proteomic analyses of sperm samples from two *Mytilus* spp.**

283 *2.2.1. Sampling of mussels and sperm sample collection*

284 Mussels from *Mytilus edulis* and *Mytilus galloprovincialis* species were collected from
285 rocky shores in Swansea (South Wales, UK) and Ria de Vigo (North-West Spain)
286 respectively at different times within the spawning period (end of January and April) in
287 2012, transported and kept under as far as possible the same laboratory conditions for at
288 least 2 months, in order to minimize the differences between mussel species due to
289 immediate environmental effects (see [48]). After 2 months, mussels were periodically
290 induced to spawn following a thermal shock procedure (see detail in File S1). Sperm
291 samples released into filtered/UV-treated seawater in individual bottles were collected,
292 filtered twice (300 µm and 41 µm sieves), and centrifuged for 10 min at 24400 g, 10ºC.
293 After discarding the supernatant, the pellet containing sperm was resuspended in 150 µl
294 of a 10% glycerol solution, snap frozen in liquid nitrogen, and finally preserved at -80ºC
295 until further analysis. In parallel, a drop of seawater for each sample containing sperm
296 cells was examined under the microscope in order to check that the sperm presented
297 good morphology, high motility and density, otherwise the sample was discarded for
298 any further analysis.

299 *2.2.2. Protein extraction and 2-DE electrophoresis*

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

319 *2.2.3. Statistical analyses of 2-DE gels*

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

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

338 *2.2.4. Mass spectrometry analysis and protein identification*

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

362

363 **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

397 low number of protein sequences for *Mytilus* spp. available in protein databases, the
398 Blast analysis showed, as expected, a level of coverage for a protein sequence database
399 (*Mytilus*[organism], NCBInr) of 81.3%. A similar result, a database coverage of 82.7%,
400 was obtained after Blast analyses against all EST sequences available in NCBI for
401 *Mytilus*[organism] that were translated to proteins by using the six-reading frames.
402 Although the redundancy level of these EST sequences is high, the number of sequences
403 is high so it is not surprising to see that a positive match/identification was reached for
404 31,428 (63.2%) of loci from our consensus transcriptome.

405 *3.1.2. Functional annotation of the consensus transcriptome from both Mytilus spp.*
406 From functional analysis using Blast2GO, 12,156 loci were successfully annotated for
407 GO terms (File S3 in Ref. [47]). The annotation was improved after InterProScan
408 analysis, raising the number of successful annotations to 13,283 loci (File S4 in Ref.
409 [47]). This might be interesting because functional information, *e.g.* a peptide signal
410 sequence from the differential expressed sequences between *Mytilus* spp., is still
411 reported despite the inability to get a confident gene/transcript identity during BlastX
412 analysis. The distribution of GO-terms for the full annotated transcriptome at different
413 levels, molecular function (MF), biological process (BP) and cellular component (CC)
414 categories, is displayed in Figure 2a. It is reassuring to see that "reproduction" term is
415 represented in BP category. The dominance of "binding", a general term related to the
416 non-covalent union or interaction of different molecules, in MF is also interesting
417 because when checking MF terms for the more specific tree hierarchy level 3 (Figure 5
418 in Ref. [47]), the highest representation is for protein binding, a term related to
419 interactions among proteins or protein complexes. This category should include sperm
420 proteins involved in sperm-egg interaction. Finally it is interesting to highlight in
421 category CC, in both Figure 2a and Figure 5 in Ref. [47], the high representation for
422 terms related to membrane proteins that potentially include those that might be involved
423 in the sperm-egg recognition mechanisms.

424 *3.1.3. Differential expression analysis between Mytilus edulis and M. galloprovincialis.*
425 A total of 27,233 isotigs (28% of the 97,425 occurring in the transcript assembly) are
426 differentially expressed between pooled samples of the two *Mytilus* spp. at FDR 5%, of
427 which 20,997 (21.6%) are significant at FDR 1%. This corresponds to 14,737 loci
428 (29.6% of 49,713 loci in the transcript assembly) which are significant (in that they

429 have at least one significant isotig) at FDR 5% of which 11,335 (22.8%) are significant
430 at FDR 1%. Files S5 and S6 in Ref. [47] contain expression and statistical values from
431 this analysis. File S7 in Ref. [47] contains the annotation based on BlastX (see section
432 above 2.1.3) for all transcripts (loci) where a significant differential expression result
433 was found. A total of 4338 (4223 at FDR 1%) differentially expressed loci were
434 successfully annotated after Blast2GO including InterProScan 5.0 [63] analysis. The
435 most relevant result of the GO term enrichment analysis in relation to this study is an
436 overrepresentation of the BP term "reproduction" (Figure 2b). These loci form the main
437 analytical focus in this paper. To pursue this, we chose those functional annotated loci (a
438 total of 309 of the 4338 in total that are differentially expressed) that code for proteins
439 specifically related to fertilisation and sperm biology processes. From these, 61 loci
440 corresponding to 50 different proteins are shortlisted based on the prediction that they
441 have signal peptide or transmembrane domains by using SignalP 4.1 [64] and TMHMM
442 2.0 [65] servers, available in CGS Technical University of Denmark, respectively, and
443 complemented with results from InterProScan 5.0 analysis described above (Table 2).
444 These types of domains indicate that protein can be either secreted (*e.g.*, present in the
445 sperm acrosomal content) or located in the sperm plasma membrane respectively, hence
446 with high potential to play a role in the sperm-egg recognition system or gamete fusion
447 [16]. We thus wish to specifically focus on these as good candidates for more detailed
448 consideration and perhaps future study. These candidate loci (Table 2) code for proteins
449 that are mainly involved in different steps of spermatogenesis (Cdyl2, Ggnbp2, Nphp1,
450 Rarb, Irs, Iap2, Tmbim6, eif4g2, CtsB, CtsL, CtsL2, Prdm9, Suv39h2), sperm motility
451 (Dnal1, Ropn1, Ift172, Slc26, Slc6a5, Slc9c1), binding of sperm to the egg vitelline coat
452 (Cct2, Cct3, Cct4, Cct5, Cct6a, Cct7, Cct8, Psma2, Ubc8, Pc1, Hya, Spag1, Thbs1, Zan,
453 vitelline coat lysins M3 and M6), acrosome reaction and sperm capacitation (Cdc42,
454 Spa17, CtsB). For each of the above candidate genes (loci), in some cases, isotigs
455 within a locus varied in the nature and extent of differential expression between the two
456 *Mytilus* species, see final two columns in Table 2. The expression differences could
457 have resulted from simple allele differences between the mussels making up the pools,
458 or more complex alternative splicing events producing different protein isoforms in the
459 two species. It also might be the result of differential regulation of expression of the
460 same protein isoform in the mature male gonad of the two different *Mytilus* species. It is
461 important to note that allele differences can have two main different effects at the

462 molecular phenotype level, either changing the mRNA/protein sequence or acting as
463 expression modifiers. The latter effect can be associated with changes in non-coding
464 usually cis-regulatory regions, though getting direct evidence for this is rather difficult
465 [66].

466 **3.2. Proteomic analysis of sperm cells from *Mytilus edulis* and *M. galloprovincialis***

467 *3.2.1. Two-dimensional electrophoresis (2DE) and differential expression analyses*

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

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

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

515 *3.2.2. Protein identification by mass spectrometry (MS)*

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

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

527 There are several spots showing differences in MW and pI (Figure 3) that were
528 identified as the same protein (see Table 3). One possible explanation for this is that
529 these originate by different post-transcriptional or post-translational modifications
530 (PTMs). The correct interpretation of these candidate "multi-spot" proteins is important
531 from a functional viewpoint to prevent misleading conclusions (see Box 2 in [38]). For
532 example, in protein isoforms of Aco2 (spots 1205 and 1241) and Idh3g (spots 1085 and
533 1087) a concordant pattern of up-regulation in *M. edulis* was observed, whereas protein
534 isoforms for Uqcrc2 (spots 847, 2164a and 2151), Efhc2 (spots 1119, 1134 and 191),
535 es1 (spots 1608, 2039 and 1602), and Glud (spots 589 and 2062) showed a discordant
536 pattern (see Table 3, Figures 3 and 4). Phosphorylation is one of the well-known PTMs
537 that usually implies modification in the pI of phosphorylated protein but little MW
538 change [67]. An advantage of using 2-DE for proteome separation compared to gel-free
539 (shotgun) proteomic approaches is that it provides the possibility of assessing the effects
540 of differential post-translational modifications and different isoform expression between
541 samples [68-70]. The observation of spots resolved in close proximity in the 2-DE gel
542 such as Idh3g (spots 1085 and 1087), Uqcrc (spots 847 and 2164), Glud (spots 589 and
543 2062), and Tekt2 (spots 814 and 776) is also compatible with differential
544 phosphorylation events in the sperm of the two *Mytilus* spp., and could be verified by
545 further phosphoproteomic analysis [71].

546 The list of protein identifications from excised spots contained many proteins
547 potentially involved in sperm function. There are proteins involved in cell energy
548 production, hence potentially affecting sperm motility, such as different members of the
549 electron transport chain (ETC) protein complex (Nadufa10, Uqcrc2, Atp5a) or in close
550 relation to ETC (Etfb), while Ppa1, Idh3g, Idh3a, Eno and Ak are other identified
551 enzymes that also contribute to maintain the energetic cellular resources. An interesting
552 observation is that about half of identified proteins are located in mitochondria (Table
553 3), so playing a role in cellular energy homeostasis either through ETC or different
554 metabolic pathways. Proteins that contribute to flagellum structure could play a role in
555 sperm motility, like Tekt1, Tekt2, Tekt4, and Cnn1. There are also proteins involved in
556 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-

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

665

666 **4. Discussion**

667 **4.1 Transcriptomic differences in mature male gonad between two *Mytilus* spp.**
668 **shed light on proteins with potential involvement in reproductive isolation**

669 Results from transcriptomic experiments using next-generation sequencing technology
670 (RNA-seq) with a focus on different biological questions have been reported for *M.*
671 *edulis* (e.g. [74], in a study of gene regulation during early development) and *M.*
672 *galloprovincialis* (e.g. [75], to compare transcript expression profiles in four different
673 tissues). However there has not been any attempt to deep sequence the mature male
674 gonad transcriptome and compare transcriptomic data in these two *Mytilus* species. The
675 current RNA-seq analysis provides evidence of high variation in the mature male gonad
676 transcriptome, with 22.8% of analysed loci differing (at FDR 1%) between *M.*
677 *galloprovincialis* and *M. edulis* samples. In a high number of instances the differential
678 expression was detected at isotig level within each consensus transcript (locus), with
679 contrasting results among different isotigs within loci, both in terms of effect-size and
680 direction of the expression level between the two *Mytilus* spp. (see Table 2). The RNA
681 transcripts showing different expression in Table 2 are both derived from sperm and
682 have sperm associated GO terms with their protein names. We would thus expect many
683 of these transcripts to be expressed as proteins for specific functioning in this tissue.
684 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 ($\chi2$ 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

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

819 **4.2. Sperm proteome differences between *Mytilus edulis* and *M. galloprovincialis***

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

842 *4.2.1 Mitochondrial proteins linked to energy production and antioxidant enzymes are*
843 *up-regulated in M. edulis*

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

846 so any observed differences between the two *Mytilus* species could be the result of their
847 following different adaptive strategies relating to sperm motility. From the list of
848 identified proteins showing differential expression (Table 3), NADH dehydrogenase
849 [ubiquinone] 1 alpha subcomplex subunit 10 (Ndufa10), Cytochrome b-c1 complex
850 subunit 2 (Uqcrc2), and ATP synthase subunit alpha (Atp5a) are nuclear encoded and
851 from the different complexes of the respiratory electron transport chain (ETC) in
852 mitochondria. Remarkably, the list of protein identifications (Table 3) reveals that nearly
853 half of the identified proteins develop their functions and are located in mitochondria. A
854 similar result was observed for highly expressed proteins in the sperm of *Mytilus edulis*
855 [20]. Proteins from these ETC-associated complexes might be implicated in postzygotic
856 isolating mechanisms due to coevolution of nuclear and mitochondrial genomes to
857 ensure appropriate functional interactions between the nuclear and mitochondrial coded
858 protein subunits of these complexes [104-105]. Marine mussels of *Mytilus* spp. as well
859 as other bivalves present an unusual mtDNA inheritance mechanism (termed doubly-
860 uniparental inheritance, DUI) in which distinct mtDNA genomes are passed through the
861 male and female lines of descent and which is coupled to sex determination in these
862 species [106-109] with opportunity for selection to act directly on mtDNA coded sperm
863 proteins. Negative epistatic interactions between nuclear and mitochondrial genomes in
864 hybrids could contribute to the maintenance of species integrity, consistent with
865 observations of DUI disruption in crosses between these two *Mytilus* species [110].

866 Other identified differentially expressed mitochondrial proteins relate to energy
867 metabolism. These include isocitrate dehydrogenase (Idh3a and Idh3g), aconitate
868 hydratase (Aco2) and dihydrolipoyl dehydrogenase (Dld). Idh3 was identified as having
869 the highest expression levels in a previous proteomics study of *Mytilus edulis* sperm
870 [20]. Low expression levels of Aco2 were reported in human sperm with reduced
871 motility [111], and higher levels during mice sperm capacitation [112]. Deficiency of
872 Dld mature protein was associated with low sperm motility in humans [113], while
873 enzymes of this complex were also related to sperm capacitation and the acrosome
874 reaction in the hamster and humans [114-115]. The higher expression of such proteins in
875 might result in higher ATP production and a fitness advantage under certain ecological
876 and environmental conditions (see section 4.3). However production of ATP through
877 oxidative phosphorylation (OXPHOS) may produce high reactive oxygen species
878 (ROS) in sperm leading to mitochondrial mutations [116] and evolution of a trade-off

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

886 *4.2.2 Up-regulation of rapid energy supply and alternative production pathways in M.*
887 *galloprovincialis*

888 It is of interest that different species, for different cellular types, could have evolved
889 different strategies and molecular pathways for energy production [121] driven by
890 different ecological or environmental pressures. For example, glutamate dehydrogenase
891 (Glud) converts glutamate to α-ketoglutarate potentially enhancing the activity of the
892 TCA cycle in which α-ketoglutarate is an intermediate. Two spots closely located in the
893 2-DE map, were identified as Glud. These could be isoforms resulting from different
894 posttranscriptional and posttranslational modifications (*e.g.* phosphorylation) implying
895 functional changes [122] in sperm of both *Mytilus* spp.

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

904 Phosphagen kinases are involved in intracellular energy transport and temporal
905 buffering of ATP levels, specifically in flagellated cells, and hence probably play a role
906 in sustained sperm motility [125]. The enzyme also influences sperm tail length and
907 flagellar bending [126-127] and sperm-specific isoforms have been reported in various
908 invertebrates [20, 125]. One of these enzymes, arginine kinase (Ak) was differentially
909 expressed here in two spots. Phosphagen molecules also regulate intracellular inorganic
910 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

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

949 4.2.5 *Other identified proteins with sperm-specific functional links*

950 Three different spots with differential expression were identified as the protein EF-hand
951 domain-containing family member C2-like (Efhc2). Sperm proteins with EF-hand
952 domains play a key role in activation of the oocyte during fertilisation in mammals
953 [141], and can also be involved in the acrosome reaction in invertebrates [9] and
954 motility regulation of sperm [142-143]. Three protein spots identified as ES1 also
955 showed differential expression. There is little functional information on this protein
956 though it has been related to differential sperm motility in humans [111]. Other proteins
957 showing differential expression are 3-hydroxyacyl-CoA dehydrogenase type-2
958 (Hsd17b10), potentially involved in the regulation of steroid hormones in reproduction
959 and reported in several molluscs [144], and atrial natriuretic peptide receptor (Npr1)
960 which acts on capacitation, chemotaxis and chemokinesis [145-146] and thus might
961 potentially play a role in species-specific sperm-egg recognition in *Mytilus* spp. driven
962 by chemotaxis signals released from eggs.

963 **4.3. Rapid evolution and sperm function trade-offs may explain species-specific**
964 **proteome differences**

965 4.3.1 *Selective pressures and adaptation in sperm*

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

975 achieve successful fertilisation may be accompanied by selection for eggs that block
976 fertilization to prevent polyspermy. This can lead to rapid co-evolution of proteins in
977 eggs and sperm in the context of sexual conflict. The rapid evolution of sperm proteins
978 has been observed in many animal groups from mammals to different marine
979 invertebrates such as sea urchins, abalones, turban snails, oysters, sea stars and mussels
980 [9, 150-151]. In a comparison of sperm proteins between *M. galloprovincialis* and *M.*
981 *edulis* the highest non-synonymous to synonymous substitutions rates were observed for
982 proteins involved in fertilisation [21]. Sperm limitation should exert strong selection for
983 adaptations increasing the chance of successful fertilisation in marine organisms with
984 external fertilisation [152-153]. These include spawning synchrony, high levels of
985 sperm production, chemotaxis over short distances, and sperm longevity. There is
986 evidence that sperm energetics, for example higher ATP production may enhance sperm
987 performance through an increase in swimming speed [154] and increase the chance of
988 fertilization. But given finite energy resources to allocate to sperm properties and
989 function, trade-offs between sperm traits are expected. For example trade-offs between
990 sperm velocity and longevity occur both within and between species [154-155].
991 However there are numerous complicating factors such as the ability of sperm to
992 maintain flagellar beats with low ATP and high inorganic phosphate levels, or the use of
993 alternative pathways for energy production [121, 156] despite oxidative phosphorylation
994 and glycolysis in the sperm midpiece being the major source of ATP production [123].

995 4.3.2 *Sperm proteins upregulated in M. edulis and M. galloprovincialis*

996 In the present study many proteins connected with sperm function which are
997 upregulated in *M. edulis* or *M. galloprovincialis* (Table 3) have been identified and their
998 properties discussed above (see section 4.2). As contrasting scenarios, selective
999 pressures in the native environments of the two species could be somewhat similar or
1000 quite different. In the former scenario suppose that selection favoured upregulation of
1001 proteins improving motility to enhance fertilisation success. This could be achieved by
1002 upregulating different genes of the same protein in the two species. For example
1003 proteins from different Tektin-2 spots are upregulated in *M. edulis* and *M.*
1004 *galloprovincialis* (Table 3). This differential effect could be achieved by selection or
1005 drift increasing the frequency of different locus specific expression modifiers in the two
1006 species. Alternatively different proteins potentially affecting motility could be
1007 differentially upregulated in the two species. For example, isocitrate dehydrogenase is

1008 upregulated in *M. edulis* and arginine kinase is upregulated in *M. galloprovincialis*
1009 (Table 3). In the latter scenario where selection pressures differ between species,
1010 proteins for quite different traits may obviously be upregulated in the two species.

1011 A summary of the proteins of Table 3 matched with sperm functional traits is given in
1012 File S7. Column I marks the particular 4.2 sub-sections in which proteins were flagged
1013 as having predominantly higher expression in *M. edulis* (4.2.1) or *M. galloprovincialis*
1014 (4.2.2, 4.2.3, 4.2.4). Column G assigns functional trait terms to the proteins and the
1015 count and % frequency distributions for these terms are given in Figure 7. These
1016 distributions give at least an approximate guide to which sperm traits are upregulated in
1017 the two species. In both species proteins relating to motility are important in this regard.
1018 After this, proteins relating to ATP reserves and perhaps ROS production are important
1019 in *M. edulis* whereas proteins relating to the acrosome reaction, capacitation, and sperm-
1020 egg interaction might be highlighted in *M. galloprovincialis*. On this basis it is possible
1021 to hypothesise that motility is important in both species but particularly *M. edulis*,
1022 whereas in *M. galloprovincialis* proteins relating to sperm maturation and the
1023 fertilization process should be highlighted.

1024 The potential biological consequences of these sperm traits are elaborated in File S7 in
1025 column H. A notable feature is that upregulation of many proteins in Table 3 can be
1026 hypothesised to result in a functional advantage for sperm. In this circumstance red font
1027 is used in columns G and H. For example in *M. edulis*, aconitate hydratase has higher
1028 expression than in *M. galloprovincialis* and this higher expression could be interpreted
1029 as a functional benefit in terms of faster swimming speed or endurance as well as
1030 improved maturation of sperm. By contrast the higher expression of es1 protein in *M.*
1031 *edulis* affecting the sperm trait motility might be hypothesised to reduce motility, a
1032 functional disadvantage, on the basis that lower motility was observed in human sperm
1033 with higher levels of this protein. This is represented by green text font in File S7
1034 columns G and H. Where it is more difficult to arrive at a functional benefit or
1035 disadvantage, black font is used. The counts of the number of spots in which the sperm
1036 trait terms can be flagged with red, green or black font are also given in the final two
1037 columns of Figure 7. There is a clear preponderance of protein spots in which higher
1038 expression can be hypothesised to be a functional benefit in terms of sperm performance
1039 in the species in which this higher expression occurs, the functional benefits being
1040 largely in sperm motility and related traits and the fertilization process.

1041 In both species, the higher expression of proteins associated with various aspects of
1042 sperm function are consistent with positive natural selection towards improved function
1043 and fitness of sperm. Closely related hybridising species such as *M. edulis* and *M.*
1044 *galloprovincialis* might be expected to show few or many differences in expression as a
1045 result of selection pressures arising from ecological forces. The wide range of
1046 differentially expressed proteins observed in the current study is consistent with
1047 evidence from the mouse where a diverse set of 81 different protein genes, including 23
1048 sperm membrane proteins all gave evidence of positive selection [157], and where
1049 proteins involved in sperm-egg interactions in particular show accelerated evolution
1050 [151]. Such a large number of genes involved in sperm function could underline that
1051 there may be a high selection intensity acting on sperm. This may also provide multiple
1052 opportunities for disrupting sperm function. For example it has been reported that in sea
1053 urchins as few as 10 amino acid changes in the protein bindin are needed for complete
1054 gamete incompatibility [158], so limited changes occurring at different loci might have
1055 similar effects.

1056 4.3.3 *Differential expression: implications for hybridization of M. edulis and M.*
1057 *galloprovincialis*

1058 The observation of protein expression differences for many different genes connected
1059 with sperm function has implications for models of hybridization and introgression
1060 between the species. An earlier proteomic study of a hybrid zone between *M. edulis* and
1061 *M. galloprovincialis* using somatic tissue found evidence of high gene expression
1062 variation amongst hybrids consistent with segregation at expression modifier loci as
1063 introgression proceeds [42]. Such segregation of modifiers at many sperm function
1064 related genes differing in protein expression between the species could result lowered
1065 expression or general disruption of expression of these genes depending on dominance
1066 relationships at and epistatic interaction between the modifier loci. This could contribute
1067 to lowered fertility of hybrids or lowered fitness of larvae as has been observed
1068 experimentally between different *Mytilus* spp. [33, 35]. It might also contribute to the
1069 observed disruption of doubly uniparental inheritance (DUI) in crosses between these
1070 two species [159] or other pair of *Mytilus* spp. [160-162].

1071 4.3.4 *Possible influences of environmental variation on sperm function*

1072 *M. edulis* evolved in the North Atlantic whereas *M. galloprovincialis* evolved in the
1073 Mediterranean [101, 162-163]. The most prominent environmental factors that might
1074 have exerted selective influences in the past are first temperature and then salinity
1075 which are both higher in the Mediterranean. These environmental differences persist in
1076 the contrast between Vigo and Swansea today, with seawater temperature about 4°C
1077 higher at Vigo during the spawning season. There is evidence that changes in seawater
1078 temperature may affect sperm function. Thus in *M. galloprovincialis* higher temperature
1079 is associated with lower fertilization rates on average [164] and sperm motility and
1080 linearity of swimming patterns are affected by temperature and its interaction with pH
1081 [165]. This may have fitness consequences as swimming speed has also been associated
1082 with higher fertilisation rates [166]. In some circumstances, for example when
1083 chemoattractants are not present, non-linear swimming patterns may be advantageous to
1084 maximise the chance of fertilisation [147, 167-168]. Other environmental factors may
1085 be important for successful fertilisation for example viscosity which is a function of
1086 temperature and salinity [169]. Factors such as seawater specific gravity and turbulence
1087 may also be important in determining the chance of successful fertilisation [170-171].

1088 4.3.5 *Selective pressures and interpretation of present results*

1089 The historical and current environmental factors affecting *M. edulis* and *M.*
1090 *galloprovincialis* could have generated different selective forces to cause divergence in
1091 sperm phenotype. This could include modification of functional trade-offs between
1092 traits such as swimming speed and endurance [172]. Differential selection modifying
1093 sperm phenotype are expected to cause differences in gene expression which could be
1094 reflected in the observed differences in protein expression as observed in the present
1095 study (Table 3, Figure 7 and File S7). Higher temperature and salinity in the evolution
1096 of *M. galloprovincialis* might relate to another factor, oxygen solubility which is lower
1097 at higher temperature and salinity. Stress from reduced oxygen could impact negatively
1098 on ATP production impacting on energy dependent biological processes such as motility,
1099 swimming speed and endurance in *M. galloprovincialis* from Vigo. In the present study
1100 however it appears that motility related proteins are relatively upregulated in *M. edulis*
1101 whereas proteins involved in sperm maturation and fertilisation are upregulated in *M.*
1102 *galloprovincialis* (Figure 7 and File S7).

1103 4.3.6 *Future studies integrating proteomics and experimental work on sperm*

1104 Clearly relating proteomics data and biochemical interpretations to environmental
1105 factors and to variation between species in sperm functional traits is a complex task for
1106 the future. Measuring intra and interspecific variation in sperm functional traits is in
1107 itself not an easy task [154]. Currently we are not aware of any direct comparative study
1108 of some sperm functional traits, like speed, longevity and movement pattern, between
1109 *M. edulis* and *M. galloprovincialis*. An experimental design in which sperm from *M.*
1110 *edulis* and *M. galloprovincialis* are spawned and their performance in motility and
1111 endurance as well as fertilisation success assessed, at a range of temperature and salinity
1112 conditions would be informative. This could be combined with further proteomics
1113 studies applied to sperm from individual mussels from these experiments. The sperm
1114 phenotype is highly plastic and evidence already exists for genotype-by-environmental
1115 interaction effects on sperm function [172]. An experimental design such as the one
1116 described above should allow detecting main effects and interactions involving species
1117 differences, reflecting genetic adaptation, contemporary environmental variation and
1118 underlying gene expression data. Such approaches could be further extended to the
1119 study of hybrid populations of the two species.

1120 **5. Concluding remarks**

1121 In order to achieve fertilization a sperm must come into contact with an egg and interact
1122 with it appropriately. Proteins mediate the interactions between sperm and egg at each
1123 step of the fertilisation process, and there is growing evidence that multiple protein
1124 complexes might be involved in concert during gamete interaction [82-83]. Species
1125 differences in these proteins are proposed as one of the key factors that lead to species-
1126 specific fertilisation and reproductive isolation. When prezygotic barriers fail, inter-
1127 species hybrids can occur. When this happens, postzygotic barriers play an important
1128 role in preservation of species integrity. We provide evidence of extensive variation in
1129 the mature male gonad transcriptome and sperm proteome in two mussel species, *M.*
1130 *edulis* and *M. galloprovincialis*. From the transcriptome analysis, we provide a
1131 preliminary list of proteins with sperm-specific functions. These functions are related to
1132 sperm-egg interaction, the acrosome reaction, spermatogenesis and motility. From the
1133 proteome analysis, we provide evidence of an overrepresentation of mitochondrial
1134 proteins among those candidate protein spots identified by MS, as well as contrasting
1135 differential expression in isoforms of many proteins. The use of customised species-

1136 specific protein databases significantly enhance both the quantity and quality of protein
1137 identifications, with the use of RNA-seq derived protein databases showing superior
1138 results to other customised databases analysed in this study. Our results provide
1139 evidence of agreement between the transcriptomic and proteomic results in the direction
1140 of expression differences between species. Our results highlight that some candidate
1141 sperm proteins, specifically those relating to sperm motility, ATP reserves, and ROS
1142 production in *M. edulis* and proteins relating to sperm motility, the acrosome reaction,
1143 capacitation and sperm-egg interaction in *M. galloprovincialis* might be good targets in
1144 further genomic analysis of reproductive barriers between closely related species.

1145

1146 **6. Acknowledgements**

1147 The authors are very grateful to Nerea González-Lavín for technical assistance during
1148 sample processing, the staff of CIM-UVIGO for technical support in the marine station
1149 facilities, colleagues from the biochemistry area for sharing equipment and providing
1150 useful advice during histological test interpretation, Susana Catarino and David Arteta
1151 (Progenika Biopharma) for assistance during RNA-seq analysis, Manuel Marcos
1152 (Proteomics service, CACTI, University of Vigo) and José M. Gallardo (IIM-CSIC,
1153 Vigo) for assistance and helpful advice in MS/MS analysis. This work was funded by
1154 the Spanish "*Ministerio de Economía y Competitividad*" (codes BFU2011-22599 and
1155 AGL2014-52062-R)", Fondos Feder and Xunta de Galicia ("*Grupos de Referencia*
1156 *Competitiva*" ED431C 2016-037), and NIH grant HD076862. M.R. Romero was
1157 supported by a predoctoral fellowship from Xunta de Galicia (Campus do Mar).

1158 The authors declare no conflict of interest

1159

1160 **7. References**

1161 [1] Palumbi SR. Genetic divergence, reproductive isolation, and marine
1162 speciation. *Annual Review of Ecology and Systematics* 1994, *25*: 547-572.

1163 [2] Swanson WJ, Vacquier VD. The rapid evolution of reproductive proteins. *Nature*
1164 *Reviews Genetics* 2002, *3*: 137-144.

1165 [3] Lessios HA. Speciation genes in free-spawning marine invertebrates. *Integrative*
1166 *and Comparative Biology* 2011, *51*: 456-465

1167 [4] McDonough CE, Whittington E, Pitnick S, Dorus S. Proteomics of reproductive
1168 systems: towards a molecular understanding of postmating, prezygotic reproductive
1169 barriers. *Journal of Proteomics* 2016, *135*: 26-37.

1170 [5] Corbett-Detig RB, Zhou J, Clark AG, Hartl DL, Ayroles JF. Genetic
1171 incompatibilities are widespread within species. *Nature* 2013, *504*: 135-137.

1172 [6] Orr HA, Presgraves DC. Speciation by postzygotic isolation: forces, genes and
1173 molecules. *BioEssays* 2000, *22*: 1085-1094.

1174 [7] Bianchi E, Doe B, Goulding D, Wright GJ. Juno is the egg Izumo receptor and is
1175 essential for mammalian fertilization. *Nature* 2014, *508*: 483-487.

1176 [8] Wilburn DB, Swanson WJ. From molecules to mating: Rapid evolution and
1177 biochemical studies of reproductive proteins. *Journal of Proteomics* 2016, *135*: 12-25.

1178 [9] Vacquier VD, Swanson WJ. Selection in the rapid evolution of gamete recognition
1179 proteins in marine invertebrates. *Cold Spring Harbor Perspectives in Biology* 2011, *3*:
1180 a002931.

1181 [10] Klinovska K, Sebkova N, Dvorakova-Hortova K. Sperm-egg fusion: a molecular
1182 enigma of mammalian reproduction. *International Journal of Molecular Sciences* 2014,
1183 *15*: 10652-10668.

1184 [11] Oliva R, de Mateo S, Estanyol JM. Sperm cell proteomics. *Proteomics* 2009, *9*:
1185 1004-1017.

1186 [12] Karr TL. Fruit flies and the sperm proteome. *Human Molecular Genetics* 2007, *16*:
1187 R124-R133.

1188 [13] Gur Y, Breitbart H. Protein synthesis in sperm: dialog between mitochondria and
1189 cytoplasm. *Molecular and Cellular Endocrinology* 2008, *282*: 45-55.

1190 [14] Bayram HL, Claydon AJ, Brownridge PJ, Hurst JL, Mileham A, Stockley P,
1191 Beynon RJ, Hammond DE. Cross-species proteomics in analysis of mammalian sperm
1192 proteins. *Journal of Proteomics* 2016, *135*: 38-50.

1193 [15] Nakachi M, Nakajima A, Nomura M, Yonezawa K, Ueno K, Endo T, Inaba K.
1194 Proteomic profiling reveals compartment-specific, novel functions of ascidian sperm
1195 proteins. *Molecular reproduction and development* 2011, *78*: 529-549.

1196 [16] Palmer MR, McDowall MH, Stewart L, Ouaddi A, MacCoss MJ, Swanson WJ.
1197 Mass spectrometry and next-generation sequencing reveal an abundant and rapidly

1198 evolving abalone sperm protein. *Molecular reproduction and development* 2013, *80*:
1199 460-465.

1200 [17] Kingtong S, Kellner K, Bernay B, Goux D, Sourdaine P, Berthelin CH. Proteomic
1201 identification of protein associated to mature spermatozoa in the Pacific oyster
1202 *Crassostrea gigas*. *Journal of Proteomics* 2013, *82:* 81-91.

1203 [18] Boonmee A, Berthelin CH, Kingtong S, Pauletto M, Bernay B, Adeline B, Suquet
1204 M, Sourdaine P, Kellner, K. Differential protein expression during sperm maturation and
1205 capacitation in an hermaphroditic bivalve, *Pecten maximus* (Linnaeus, 1758). *Journal of*
1206 *Molluscan Studies* 2016, *82*: 575-584.

1207 [19] Bartel M, Hartmann S, Lehmann K, Postel K, Quesada H, Philipp EER, Heilmann
1208 K, Micheel B, Stuckas H. Identification of sperm proteins as candidate biomarkers for
1209 the analysis of reproductive isolation in *Mytilus*: a case study for the enkurin locus.
1210 *Marine Biology* 2012, *159*: 2195

1211 [20] Diz AP, Dudley E, Skibinski DOF. Identification and characterization of highly
1212 expressed proteins in sperm cells of the marine mussel *Mytilus edulis*. *Proteomics* 2012,
1213 *12*: 1949-1956.

1214 [21] Zhang Y, Mu H, Lau SC, Zhang Z, Qiu JW. Sperm proteome of *Mytilus*
1215 *galloprovincialis*: Insights into the evolution of fertilization proteins in marine
1216 mussels. *Proteomics* 2015, *15*: 4175-4179.

1217 [22] Vicens A, Borziak K, Karr TL, Roldan ERS, Dorus S. Comparative sperm
1218 proteomics in mouse species with divergent mating systems. *Molecular Biology and*
1219 *Evolution* 2017, *34*, 1403-1416.

1220 [23] Koehn RK. The genetics and taxonomy of species in the genus
1221 *Mytilus*. *Aquaculture* 1991, *94*: 125-145.

1222 [24] Seed R. Ecology, in: Bayne, B.L. (Ed.). Marine mussels: their ecology and
1223 physiology. International Biological Programme 1976, 10: 13-65.

1224 [25] Riginos C, McDonald JH. Positive selection on an acrosomal sperm protein, M7
1225 lysin, in three species of the mussel genus *Mytilus*. *Molecular Biology and Evolution*
1226 2003, *20*: 200-207.

1227 [26] Riginos C, Wang D, Abrams AJ. Geographic variation and positive selection on M7
1228 lysin, an acrosomal sperm protein in mussels (*Mytilus* spp.). *Molecular Biology and*
1229 *Evolution* 2006, *23*: 1952-1965.

1230 [27] Springer SA, Crespi BJ. Adaptive gamete-recognition divergence in a hybridizing
1231 *Mytilus* population. *Evolution* 2007, *61*: 772-783.

1232 [28] McCartney MA, Lima TG. Evolutionary consequences of introgression at M7
1233 Lysin, a gamete recognition locus, following secondary contact between blue mussel
1234 species. *Integrative and Comparative Biology* 2011, *51*: 474-484.

1235 [29] Brannock PM, Hilbish TJ. Hybridization results in high levels of sterility and
1236 restricted introgression between invasive and endemic marine blue mussels. *Marine*
1237 *Ecology Progress Series* 2010, *406*: 161-171.

1238 [30] Beaumont AR, Abdul-Matin AKM, Seed R. Early development, survival and
1239 growth in pure and hybrid larvae of *Mytilus edulis* and *M. galloprovincialis*. *Journal of*
1240 *Molluscan Studies* 1993, *59*: 120-123.

1241 [31] Beaumont AR, Turner G, Wood AR, Skibinski DO. Hybridisations between *Mytilus*
1242 *edulis* and *Mytilus galloprovincialis* and performance of pure species and hybrid veliger
1243 larvae at different temperatures. *Journal of Experimental Marine Biology and Ecology*
1244 2004, *302*: 177-188.

1245 [32] Bierne N, David P, Boudry P, Bonhomme F. Assortative fertilization and selection
1246 at larval stage in the mussels *Mytilus edulis and M. galloprovincialis*. *Evolution*
1247 2002, *56*: 292-298.

1248 [33] Bierne N, Bonhomme F, Boudry P, Szulkin M, David P. Fitness landscapes support
1249 the dominance theory of post-zygotic isolation in the mussels *Mytilus edulis* and *M.*
1250 *galloprovincialis*. *Proceedings of the Royal Society of London B: Biological Sciences*
1251 2006, *273*: 1253-1260.

1252 [34] Rawson PD, Slaughter C, Yund PO. Patterns of gamete incompatibility between the
1253 blue mussels *Mytilus edulis* and *M. trossulus*. *Marine Biology* 2003, *143*: 317-325.

1254 [35] Miranda MBB, Innes DJ, Thompson RJ. Incomplete reproductive isolation in the
1255 blue mussel (*Mytilus edulis* and *M. trossulus*) hybrid zone in the Northwest Atlantic:
1256 role of gamete interactions and larval viability. *The Biological Bulletin* 2010, *218*: 266–
1257 281.

1258 [36] Klibansky LK, McCartney MA. Conspecific sperm precedence is a reproductive
1259 barrier between free-spawning marine mussels in the northwest Atlantic *Mytilus* hybrid
1260 zone. *PloS One* 2014, *9*: e108433.

1261 [37] Findlay GD, Swanson WJ. Proteomics enhances evolutionary and functional
1262 analysis of reproductive proteins. *Bioessays* 2010, *32*: 26-36.

1263 [38] Diz AP, Martínez-Fernández M, Rolán-Alvarez E. Proteomics in evolutionary
1264 ecology: linking the genotype with the phenotype. *Molecular Ecology* 2012, *21*: 1060-
1265 1080.

1266 [39] Baer B, Millar AH. Proteomics in evolutionary ecology. *Journal of Proteomics*
1267 2016, *135*: 4-11.

1268 [40] Diz AP, Calvete JJ. Ecological proteomics: is the field ripe for integrating
1269 proteomics into evolutionary ecology research?. *Journal of Proteomics* 2016, *135*: 1-3.

1270 [41] Palumbi SR. Speciation and the evolution of gamete recognition genes: pattern and
1271 process. *Heredity* 2009, *102*: 66-76.

1272 [42] Diz AP, Skibinski DOF. Evolution of 2-DE protein patterns in a mussel hybrid
1273 zone. *Proteomics* 2007, *7*: 2111-2120.

1274 [43] Murgarella, M, Puiu D, Novoa B, Figueras A, Posada D, Canchaya C. A first
1275 insight into the genome of the filter-feeder mussel *Mytilus galloprovincialis*. *PloS one*
1276 2016, *11*: e0151561.

1277 [44] Findlay GD, MacCoss MJ, Swanson WJ. Proteomic discovery of previously
1278 unannotated, rapidly evolving seminal fluid genes in *Drosophila*. *Genome Research*
1279 2009, *19*: 886-96.

1280 [45] Armengaud J, Trapp J, Pible O, Geffard O, Chaumot A, Hartmann EM. Non-model
1281 organisms, a species endangered by proteogenomics. *Journal of Proteomics* 2014, *105*:
1282 5-18.

1283 [46] Nesvizhskii AI. Proteogenomics: concepts, applications and computational
1284 strategies. *Nature methods* 2014, *11*: 1114-1125.

1285 [47] Diz AP, Romero MR, Pérez-Figueroa A, Swanson WJ, Skibinski DOF. RNA-seq
1286 data from mature male gonads of marine mussels *Mytilus edulis* and *M.*
1287 *galloprovincialis*. *Data in Brief* 2018 (submitted).

1288 [48] Atasaral-Şahin Ş, Romero MR, Cueto R, González-Lavín N, Marcos M, Diz AP.
1289 Subtle tissue and sex-dependent proteome variation in mussel (*Mytilus*
1290 *galloprovincialis*) populations of the Galician coast (NW Spain) raised in a common
1291 environment. *Proteomics* 2015, *15*: 3993-4006.

1292 [49] Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de
1293 Bruijn graphs. *Genome Research* 2008, *18*: 821-829.

1294 [50] Schulz MH, Zerbino DR, Vingron M, Birney E. Oases: robust de novo RNA-seq
1295 assembly across the dynamic range of expression levels. *Bioinformatics* 2012, *28*: 1086-
1296 1092.

1297 [51] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment
1298 search tool. *Journal of Molecular Biology* 1990, *215*: 403-410.

1299 [52] Zhang G, Fang X, Guo X, Li L, Luo R, Xu F, Yang P, Zhang L, Wang X, Qi H,
1300 Xiong Z *et al*. The oyster genome reveals stress adaptation and complexity of shell
1301 formation. *Nature* 2012, *490*: 49-54.

1302 [53] Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. Blast2GO: a
1303 universal tool for annotation, visualization and analysis in functional genomics research.
1304 *Bioinformatics* 2005, *21*: 3674-3676.

1305 [54] Zhang ZH, Jhaveri DJ, Marshall VM, Bauer DC, Edson J, et al. A comparative
1306 study of techniques for differential expression analysis on RNA-Seq data. Plos One
1307 2014, 9: e103207

1308 [55] Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data
1309 with or without a reference genome. *BMC Bioinformatics* 2011, *12*: 1.

1310 [56] Leng N, Dawson JA, Thomson JA, Ruotti V, Rissman AI, Smits BMG, Haag JD,
1311 Gould MN, Stewart RM, Kendziorski C. EBSeq: an empirical Bayes hierarchical model
1312 for inference in RNA-seq experiments. *Bioinformatics* 2013, *29*: 1035-1043.

1313 [57] Schurch NJ, Schofield P, Gierliński M, Cole C, Sherstnev A, Singh V, Wrobel N,
1314 Gharbi K, Simpson GG, Owen-Hughes T, Blaxter M, Barton GJ. How many biological
1315 replicates are needed in an RNA-seq experiment and which differential expression tool
1316 should you use? *RNA* 2016, *22*:1-13

1317 [58] Ramagli LS, Rodriguez LV. Quantitation of microgram amounts of protein in two-
1318 dimensional polyacrylamide gel electrophoresis sample buffer. *Electrophoresis* 1985, 6:
1319 559–563.

1320 [59] Diz AP, Dudley E, Cogswell A, MacDonald BW, Kenchington EL, Zouros E,
1321 Skibinski DOF. Proteomic analysis of eggs from *Mytilus edulis* females differing in
1322 mitochondrial DNA transmission mode. *Molecular & Cellular Proteomics* 2013, *12*:
1323 3068-3080.

1324 [60] Carvajal-Rodriguez A, de Uña-Alvarez J. Assessing significance in high-
1325 throughput experiments by sequential goodness of fit and q-value estimation. *PLoS One*
1326 2011, *6*: e24700.

1327 [61] Diz AP, Carvajal-Rodríguez A, Skibinski DOF. Multiple hypothesis testing in
1328 proteomics: a strategy for experimental work. *Molecular & Cellular Proteomics*
1329 2011, *10*: M110-004374.

1330 [62] Warnes GR, Bolker B, Bonebakker L, Gentleman R, Liaw WHA, Lumley T,
1331 Maechler M, Magnusson A, Moeller S, Schwartz M, Venables B. gplots: Various R
1332 Programming Tools for Plotting Data 2014. R package version 2.14.2. https://CRAN.R-
1333 project.org/package=gplots

1334 [63] Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R.
1335 InterProScan: protein domains identifier. *Nucleic Acids Research* 2005, 33: W116-
1336 W120.

1337 [64] Petersen TN, Brunak S, Heijne von G, Nielsen H. SignalP 4.0: Discriminating
1338 signal peptides from transmembrane regions. *Nature Methods* 2011, 8: 785-786.

1339 [65] Krogh A, Larsson B, Von Heijne G, Sonnhammer EL. Predicting transmembrane
1340 protein topology with a hidden Markov model: application to complete genomes.
1341 *Journal of Molecular Biology* 2001, 305: 567-580.

1342 [66] Buckland PR. Allele-specific gene expression differences in humans. *Human*
1343 *Molecular Genetics* 2004, *13*, R255-R260.

1344 [67] Yamagata A, Kristensen DB, Takeda Y, Miyamoto Y, Okada K, Inamatsu M,
1345 Yoshizato K. Mapping of phosphorylated proteins on two-dimensional polyacrylamide
1346 gels using protein phosphatase. *Proteomics* 2002, *2*: 1267-1276.

1347 [68] Morandell S, Stasyk T, Grosstessner-Hain K, Roitinger E, Mechtler K, Bonn GK,
1348 Huber LA. Phosphoproteomics strategies for the functional analysis of signal
1349 transduction. *Proteomics* 2006, *6*: 4047-4056.

1350 [69] Rabilloud T, Lelong C. Two-dimensional gel electrophoresis in proteomics: a
1351 tutorial. *Journal of Proteomics* 2011, *74*: 1829-1841.

1352 [70] Diz AP, Álvarez-Rodríguez M, Romero MR, Rolán-Alvarez E, Galindo J. Limited
1353 proteomic response in the marine snail *Melarhaphe neritoides* after long-term emersion.
1354 *Current Zoology* 2017, *63*, 487-493.

1355 [71] Thingholm TE, Jensen ON, Larsen MR. Analytical strategies for
1356 phosphoproteomics. *Proteomics* 2009, *9*: 1451-1468.

1357 [72] Feder ME, Walser JC. The biological limitations of transcriptomics in elucidating
1358 stress and stress responses. *Journal of Evolutionary Biology* 2005, *18*: 901-910.

1359 [73] Vogel C, Marcotte EM. Insights into the regulation of protein abundance from
1360 proteomic and transcriptomic analyses. *Nature Reviews Genetics* 2012, *13*: 227-232.

1361 [74] Bassim S, Tanguy A, Genard B, Moraga D, Tremblay R. Identification of *Mytilus*
1362 *edulis* genetic regulators during early development. *Gene* 2014, *551*: 65-78.

1363 [75] Moreira R, Pereiro P, Canchaya C, Posada D, Figueras A, Novoa B. RNA-Seq in
1364 *Mytilus galloprovincialis*: comparative transcriptomics and expression profiles among
1365 different tissues. *BMC genomics* 2015, *16*: 728.

1366 [76] Ezkurdia I, del Pozo A, Frankish A, Rodriguez JM, Harrow J, Ashman K, Valencia
1367 A, Tress ML. Comparative proteomics reveals a significant bias toward alternative
1368 protein isoforms with conserved structure and function. *Molecular Biology and*
1369 *Evolution* 2012, *29*: 2265-2283.

1370 [77] Gonzàlez-Porta M, Frankish A, Rung J, Harrow J, Brazma A. Transcriptome
1371 analysis of human tissues and cell lines reveals one dominant transcript per gene.
1372 *Genome Biology* 2013, *14*: R70.

1373 [78] Pelini SL, Diamond SE, MacLean H, Ellison AM, Gotelli NJ, Sanders NJ, Dunn
1374 RR. Common garden experiments reveal uncommon responses across temperatures,
1375 locations, and species of ants. *Ecology and Evolution* 2012, *2*: 3009-3015.

1376 [79] de Villemereuil P, Gaggiotti OE, Mouterde M, Till-Bottraud I. Common garden
1377 experiments in the genomic era: new perspectives and opportunities. *Heredity* 2016,
1378 *116*: 249-254.

1379 [80] Sanford E, Kelly MW. Local adaptation in marine invertebrates. *Annual Review of*
1380 *Marine Science* 2011, *3:* 509-535.

1381 [81] Banni M, Negri A, Mignone F, Boussetta H, Viarengo A, Dondero F. Gene
1382 expression rhythms in the mussel *Mytilus galloprovincialis* (Lam.) across an annual
1383 cycle. *PLoS One* 2011, *6*: e18904.

1384 [82] Dun MD, Smith ND, Baker MA, Lin M, Aitken RJ, Nixon B. The chaperonin
1385 containing TCP1 complex (CCT/TRiC) is involved in mediating sperm-oocyte
1386 interaction. *Journal of Biological Chemistry* 2011, *286*: 36875-36887.

1387 [83] Redgrove KA, Anderson AL, Dun MD, McLaughlin EA, O'Bryan MK, Aitken RJ,
1388 Nixon B. Involvement of multimeric protein complexes in mediating the capacitation-
1389 dependent binding of human spermatozoa to homologous zonae pellucidae.
1390 *Developmental Biology* 2011, *356*: 460-474.

1391 [84] Sutovsky P. Sperm proteasome and fertilization. *Reproduction* 2011, *142*: 1-14.

1392 [85] Hershko A, Ciechanover A. The ubiquitin system. *Annual Review of Biochemistry*
1393 1998, *67*: 425-479.

1394 [86] Zimmerman S, Sutovsky P. The sperm proteasome during sperm capacitation and
1395 fertilization. *Journal of Reproductive Immunology* 2009, *83*: 19-25.

1396 [87] Miranda PV, González-Echeverría F, Blaquier JA, Mahuran DJ, Tezón JG.
1397 Evidence for the participation of β-hexosaminidase in human sperm–zona pellucida
1398 interaction in vitro. *Molecular human reproduction* 2000, *6*: 699-706.

1399 [88] Koyanagi R, Honegger TG. Molecular cloning and sequence analysis of an
1400 ascidian egg β-N-acetylhexosaminidase with a potential role in fertilization.
1401 *Development, Growth & Differentiation* 2003, *45*: 209-218.

1402 [89] Chiriva-Internati M, Gagliano N, Donetti E, Costa F, Grizzi F, Franceschini B,
1403 Albani E, Levi-Setti PE, Gioia M, Jenkins M, Cobos E, Kast WM. Sperm protein 17 is
1404 expressed in the sperm fibrous sheath. *Journal of Translational Medicine* 2009, *7*: 61.

1405 [90] Schumacher J, Rosenkranz D, Herlyn H. Mating systems and protein–protein
1406 interactions determine evolutionary rates of primate sperm proteins. *Proceedings of the*
1407 *Royal Society of London B: Biological Sciences* 2014, *281*: 20132607.

1408 [91] Enard D, Depaulis F, Crollius HR. Human and non-human primate genomes share
1409 hotspots of positive selection. *PLoS Genetics* 2010, *6*: e1000840.

1410 [92] Swanson WJ, Nielsen R, Yang Q. Pervasive adaptive evolution in mammalian
1411 fertilization proteins. *Molecular Biology and Evolution* 2003, *20*: 18-20.

1412 [93] Gasper J, Swanson WJ. Molecular population genetics of the gene encoding the
1413 human fertilization protein zonadhesin reveals rapid adaptive evolution. *The American*
1414 *Journal of Human Genetics* 2006, *79*: 820-830.

1415 [94] Tardif S, Cormier N. Role of zonadhesin during sperm–egg interaction: a species-
1416 specific acrosomal molecule with multiple functions. *Molecular Human Reproduction*
1417 2011, *17*: 661-668.

1418 [95] Lima TG, McCartney MA. Adaptive evolution of M3 lysin—a candidate gamete
1419 recognition protein in the *Mytilus edulis* species complex. *Molecular Biology and*
1420 *Evolution* 2013, *30*: 2688-2698.

1421 [96] Takagi T, Nakamura A, Deguchi R, Kyozuka KI. Isolation, characterization, and
1422 primary structure of three major proteins obtained from *Mytilus edulis* sperm. *Journal of*
1423 *Biochemistry* 1994, *116*: 598-605.

1424 [97] Flachs P, Bhattacharyya T, Mihola O, Piálek J, Forejt J, Trachtulec Z. Prdm9
1425 incompatibility controls oligospermia and delayed fertility but no selfish transmission in
1426 mouse intersubspecific hybrids. *PloS One* 2014, *9*: e95806.

1427 [98] Mihola O, Trachtulec Z, Vlcek C, Schimenti JC, Forejt J. A mouse speciation gene
1428 encodes a meiotic histone H3 methyltransferase. *Science* 2009, *323*: 373-375.

1429 [99] O'Carroll D, Scherthan H, Peters AH, Opravil S, Haynes AR, Laible G, Rea S,
1430 Schmid M, Lebersorger A, Jerratsch M, Sattler L, Mattei MG, Denny P, Brown SDM,
1431 Schweizer D, Januwein T. Isolation and characterization of Suv39h2, a second histone
1432 H3 methyltransferase gene that displays testis-specific expression. *Molecular and*
1433 *Cellular Biology* 2000, *20*: 9423-9433.

1434 [100] Romero MR. Evolutionary and functional proteomic study of gametes from
1435 *Mytilus edulis* and *M. galloprovincialis* integrating different -omics technologies
1436 (Unpublished doctoral dissertation 2017). University of Vigo, Spain.

1437 [101] Tomanek L. Environmental proteomics of the mussel *Mytilus*: implications for
1438 tolerance to stress and change in limits of biogeographic ranges in response to climate
1439 change. *Integrative and Comparative Biology* 2012, *52*: 648-664.

1440 [102] Ruiz-Pesini E, Diez C, Lapeña AC, Pérez-Martos A, Montoya J, Alvarez E,
1441 Arenas J, López-Pérez MJ. Correlation of sperm motility with mitochondrial enzymatic
1442 activities. *Clinical Chemistry* 1998, *44*: 1616-1620.

1443 [103] Zhao C, Huo R, Wang FQ, Lin M, Zhou ZM, Sha JH. Identification of several
1444 proteins involved in regulation of sperm motility by proteomic analysis. *Fertility and*
1445 *Sterility* 2007, *87*: 436-438.

1446 [104] Dowling DK, Friberg U, Lindell J. Evolutionary implications of non-neutral
1447 mitochondrial genetic variation. *Trends in Ecology & Evolution* 2008, *23*: 546-54.

1448 [105] Skibinski DOF, Ghiselli F, Diz AP, Milani L, Mullins JG. Structure-related
1449 differences between cytochrome oxidase I proteins in a stable heteroplasmic
1450 mitochondrial system. *Genome Biology and Evolution* 2017, *9*: 3265-3281.

1451 [106] Skibinski DOF, Gallagher C, Beynon CM. Sex-limited mitochondrial-DNA
1452 transmission in the marine mussel *Mytilus edulis*. *Genetics* 1994, *138*: 801-809.

1453 [107] Skibinski DOF, Gallagher C, Beynon CM. Mitochondrial-DNA inheritance.
1454 *Nature* 1994, *368*: 817-818.

1455 [108] Zouros E, Ball AO, Saavedra C, Freeman KR. An unusual type of mitochondrial-
1456 DNA inheritance in the blue mussel *Mytilus*. *Proceedings of the National Academy of*
1457 *Sciences of the USA* 1994, *91*: 7463-7467.

1458 [109] Zouros E, Ball AO, Saavedra C, Freeman KR. Mitochondrial-DNA inheritance.
1459 *Nature*, 1994, *368*: 818.

1460 [110] Wood AR, Turner G, Skibinski DOF, Beaumont AR. Disruption of doubly
1461 uniparental inheritance of mitochondrial DNA in hybrid mussels (*Mytilus edulis*×*M.*
1462 *galloprovincialis*). *Heredity* 2003, *91*: 354-360.

1463 [111] Amaral A, Paiva C, Attardo Parrinello C, Estanyol JM, Ballescà JL, Ramalho-
1464 Santos J, Oliva R. Identification of proteins involved in human sperm motility using
1465 high-throughput differential proteomics. *Journal of Proteome Research* 2014, *13*: 5670-
1466 5684.

1467 [112] Zhao C, Guo XJ, Shi ZH, Wang FQ, Huang XY, Huo R, Zhu H, Wang XR, Liu JY,
1468 Zhou ZM, Sha JH. Role of translation by mitochondrial-type ribosomes during sperm
1469 capacitation: an analysis based on a proteomic approach. *Proteomics* 2009, *9*: 1385-
1470 1399.

1471 [113] Martínez-Heredia J, de Mateo S, Vidal-Taboada JM, Ballescà JL, Oliva R.
1472 Identification of proteomic differences in asthenozoospermic sperm samples. *Human*
1473 *Reproduction* 2008, *23*: 783-791.

1474 [114] Kumar V, Kota V, Shivaji S. Hamster sperm capacitation: role of pyruvate
1475 dehydrogenase A and dihydrolipoamide dehydrogenase 1. *Biology of Reproduction*
1476 2008, *79*: 190-199.

1477 [115] Siva AB, Panneerdoss S, Sailasree P, Singh DK, Kameshwari DB, Shivaji S.
1478 Inhibiting sperm pyruvate dehydrogenase complex and its E3 subunit, dihydrolipoamide
1479 dehydrogenase affects fertilization in syrian hamsters. *PloS One* 2014, *9*: e97916.

1480 [116] Ghiselli F, Breton S, Milani L. Mitochondrial activity in gametes and uniparental
1481 inheritance: a comment on 'What can we infer about the origin of sex in early
1482 eukaryotes?'. *Philosophical Transactions of the Royal Society B* 2018, *373*: 20170147.

1483 [117] Ji G, Gu A, Wang Y, Huang C, Hu F, Zhou, Y *et al*. Genetic variants in antioxidant
1484 genes are associated with sperm DNA damage and risk of male infertility in a Chinese
1485 population. *Free Radical Biology and Medicine* 2012, *52*: 775-780.

1486 [118] Saadi HAS, van Riemsdijk E, Dance AL, Rajamanickam GD, Kastelic JP,
1487 Thundathil JC. Proteins associated with critical sperm functions and sperm head shape
1488 are differentially expressed in morphologically abnormal bovine sperm induced by
1489 scrotal insulation. *Journal of Proteomics* 2013, *82*: 64-80.

1490 [119] Kobayashi T, Miyazaki T, Natori M, Nozawa S. Protective role of superoxide
1491 dismutase in human sperm motility: superoxide dismutase activity and lipid peroxide in
1492 human seminal plasma and spermatozoa. *Human Reproduction* 1991, *6*: 987-991.

1493 [120] van Gestel RA, Brewis IA, Ashton PR, Brouwers JF, Gadella BM. Multiple
1494 proteins present in purified porcine sperm apical plasma membranes interact with the
1495 zona pellucida of the oocyte. *Molecular Human Reproduction* 2007, *13*: 445-454.

1496 [121] Müller M, Mentel M, van Hellemond JJ, Henze K, Woehle C, Gould SB, Yu RY,
1497 van der Giezen M, Tielens AG, Martin WF. Biochemistry and evolution of anaerobic
1498 energy metabolism in eukaryotes. *Microbiology and Molecular Biology Reviews* 2012,
1499 *76*: 444-495.

1500 [122] Bunik V, Artiukhov A, Aleshin V, Mkrtchyan G. Multiple forms of glutamate
1501 dehydrogenase in animals: structural determinants and physiological implications.
1502 *Biology* 2016, *5*: 53.

1503 [123] Visconti PE. Sperm bioenergetics in a nutshell. *Biology of reproduction* 2012, *87*:
1504 72-1.

1505 [124] Nakamura N, Dai Q, Williams J, Goulding EH, Willis WD, Brown PR, Eddy EM.
1506 Disruption of a spermatogenic cell-specific mouse enolase 4 (eno4) gene causes sperm
1507 structural defects and male infertility. *Biology of reproduction* 2013, *88*: 90.

1508 [125] Strong SJ, Ellington WR. Horseshoe crab sperm contain a unique isoform of
1509 arginine kinase that is present in midpiece and flagellum. *Journal of Experimental*
1510 *Zoology Part A: Ecological Genetics and Physiology* 1993, *267:* 563-571.

1511 [126] Tombes RM, Shapiro BM. Energy transport and cell polarity: Relationship of
1512 phosphagen kinase activity to sperm function. *Journal of Experimental Zoology* 1989,
1513 *251*: 82-90.

1514 [127] Tombes RM, Brokaw CJ, Shapiro BM. Creatine kinase-dependent energy
1515 transport in sea urchin spermatozoa. Flagellar wave attenuation and theoretical analysis
1516 of high energy phosphate diffusion. *Biophysical Journal* 1987, *52*: 75-86.

1517 [128] Ellington WR. Evolution and physiological roles of phosphagen systems. *Annual*
1518 *Review of Physiology* 2001, *63*: 289-325.

1519 [129] Yi YJ, Sutovsky M, Kennedy C, Sutovsky P. Identification of the inorganic
1520 pyrophosphate metabolizing, ATP substituting pathway in mammalian spermatozoa.
1521 *PloS One* 2012, *7*: e34524.

1522 [130] Diz AP, Dudley E, MacDonald BW, Piña B, Kenchington EL, Zouros E, Skibinski
1523 DOF. Genetic variation underlying protein expression in eggs of the marine mussel
1524 *Mytilus edulis*. *Molecular & Cellular Proteomics* 2009, *8*: 132-144.

1525 [131] Sutovsky P, Moreno RD, Ramalho-Santos J, Dominko T, Simerly C, Schatten G.
1526 Ubiquitinated sperm mitochondria, selective proteolysis, and the regulation of
1527 mitochondrial inheritance in mammalian embryos. *Biology of Reproduction* 2000, *63*:
1528 582-590.

1529 [132] Berger KH, Yaffe MP. Prohibitin family members interact genetically with
1530 mitochondrial inheritance components in *Saccharomyces cerevisiae*. *Molecular and*
1531 *Cellular Biology* 1998, *18*: 4043-4052.

1532 [133] Sutovsky P, Van Leyen K, McCauley T, Day BN, Sutovsky M. Degradation of
1533 paternal mitochondria after fertilization: implications for heteroplasmy, assisted
1534 reproductive technologies and mtDNA inheritance. *Reproductive Biomedicine Online*
1535 2004, *8*: 24-33.

1536 [134] Song WH, Yi YJ, Sutovsky M, Meyers S, Sutovsky P. Autophagy and ubiquitin–
1537 proteasome system contribute to sperm mitophagy after mammalian fertilization.
1538 *Proceedings of the National Academy of Sciences of the USA* 2016, *113*: E5261-E5270.

1539 [135] Roy A, Lin YN, Agno JE, DeMayo FJ, Matzuk MM. Absence of tektin 4 causes
1540 asthenozoospermia and subfertility in male mice. *The FASEB Journal* 2007, *21*: 1013-
1541 1025.

1542 [136] Siva AB, Kameshwari DB, Singh V, Pavani K, Sundaram CS, Rangaraj N,
1543 Deenadayal M, Shivaji S. Proteomics-based study on asthenozoospermia: differential
1544 expression of proteasome alpha complex. *Molecular Human Reproduction* 2010, *16*:
1545 452-462.

1546 [137] Thepparat T, Katawatin S, Vongpralub T, Duangjinda M, Thammasirirak S, Utha
1547 A. Separation of bovine spermatozoa proteins using 2D-PAGE revealed the relationship
1548 between tektin-4 expression patterns and spermatozoa motility. *Theriogenology* 2012,
1549 *77*: 1816-1821.

1550 [138] Ashrafzadeh A, Karsani SA, Nathan S. Mammalian sperm fertility related
1551 proteins. *International Journal of Medical Sciences* 2013, *10*: 1649.

1552 [139] Bhilawadikar R, Zaveri K, Mukadam L, Naik S, Kamble K, Modi D, Hinduja I.
1553 Levels of Tektin 2 and CatSper 2 in normozoospermic and oligoasthenozoospermic men

1554 and its association with motility, fertilization rate, embryo quality and pregnancy rate.
1555 *Journal of Assisted Reproduction and Genetics* 2013, *30*: 513-523.

1556 [140] Mariappa D, Aladakatti RH, Dasari SK, Sreekumar A, Wolkowicz M, van der
1557 Hoorn F, Seshagiri PB. Inhibition of tyrosine phosphorylation of sperm flagellar
1558 proteins, outer dense fiber protein-2 and tektin-2, is associated with impaired motility
1559 during capacitation of hamster spermatozoa. *Molecular Reproduction and Development*
1560 2010, *77*: 182-193.

1561 [141] Nomikos M, Sanders JR, Parthimos D, Buntwal L, Calver BL, Stamatiadis P,
1562 Smith A, Clue M, Sideratou Z, Swann K, Lai FA. Essential role of the EF-hand domain
1563 in targeting sperm phospholipase Cζ to membrane phosphatidylinositol 4, 5-
1564 bisphosphate (PIP2). *Journal of Biological Chemistry* 2015, *290*: 29519-29530.

1565 [142] Wang D, Hu J, Bobulescu IA, Quill TA, McLeroy P, Moe OW, Garbers DL. A
1566 sperm-specific Na^+/H^+ exchanger (sNHE) is critical for expression and in vivo
1567 bicarbonate regulation of the soluble adenylyl cyclase (sAC). *Proceedings of the*
1568 *National Academy of Sciences* 2007, *104*: 9325-9330.

1569 [143] Di Sole F, Vadnagara K, Moe OW, Babich V. Calcineurin homologous protein: a
1570 multifunctional Ca^{2+}-binding protein family. *American Journal of Physiology-Renal*
1571 *Physiology* 2012, *303*: F165-F179.

1572 [144] Fernandes D, Loi B, Porte C. Biosynthesis and metabolism of steroids in
1573 molluscs. *The Journal of Steroid Biochemistry and Molecular Biology* 2011, *127*: 189-
1574 195.

1575 [145] Rotem R, Zamir N, Keynan N, Barkan D, Breitbart H, Naor Z. Atrial natriuretic
1576 peptide induces acrosomal exocytosis of human spermatozoa. *American Journal of*
1577 *Physiology-Endocrinology and Metabolism* 1998, *274*: E218-E223.

1578 [146] Zhang M, Hong H, Zhou B, Jin S, Wang C, Fu M, Wang S, Xia, G. The
1579 expression of atrial natriuretic peptide in the oviduct and its functions in pig
1580 spermatozoa. *Journal of Endocrinology* 2006, *189*: 493-507.

1581 [147] Stewart DT, Jha M, Breton S, Hoeh WR, Blier PU. No effect of sperm
1582 interactions or egg homogenate on sperm velocity in the blue mussel, *Mytilus edulis*
1583 (Bivalvia: Mytilidae). *Canadian Journal of Zoology* 2012, *90*: 1291-1296.

1584 [148] Beekman M, Nieuwenhuis B, Ortiz-Barrientos D, Evans JP. Sexual selection in
1585 hermaphrodites, sperm and broadcast spawners, plants and fungi. *Philosophical*
1586 *Transactions of the Royal Society B* 2016, *371*: 20150541.

1587 [149] Bode M., Marshall DJ. The quick and the dead? Sperm competition and sexual
1588 conflict in sea. *Evolution* 2007, *61*: 2693-2700.

1589 [150] Sunday JM, Hart MW. Sea star populations diverge by positive selection at a
1590 sperm-egg compatibility locus. *Ecology and Evolution* 2013, *3*: 640-654.

1591 [151] Vicens A, Lüke L, Roldan ERS. Proteins involved in motility and sperm-egg
1592 interaction evolve more rapidly in mouse spermatozoa. *Plos One* 2014, *9*:e91302.

1593 [152] Yund PO. How severe is sperm limitation in natural populations of marine free-
1594 spawners?. *Trends in Ecology & Evolution* 2000, *15*: 10-13.

1595 [153] Evans JP, Sherman CD. Sexual selection and the evolution of egg-sperm
1596 interactions in broadcast-spawning invertebrates. *The Biological Bulletin* 2013, *224*:
1597 166-183.

1598 [154] Fitzpatrick JL, Lüpold S. Sexual selection and the evolution of sperm quality.
1599 *Molecular Human Reproduction* 2014, *20*: 1180-1189.

1600 [155] Levitan DR. Sperm velocity and longevity trade off each other and influence
1601 fertilization in the sea urchin *Lytechinus variegatus*. *Proceedings of the Royal Society of*
1602 *London B: Biological Sciences* 2000, *267*: 531-534.

1603 [156] Kamp G, Büsselmann G, Lauterwein J. Spermatozoa: models for studying
1604 regulatory aspects of energy metabolism. *Experientia* 1996, *52*: 487-494.

1605 [157] Dorus S, Wasbrough ER, Busby J, Wilkin EC, Karr TL. Sperm proteomics reveals
1606 intensified selection on mouse sperm membrane and acrosome genes. *Molecular*
1607 *Biology and Evolution* 2010, *27*: 1235-1246.

1608 [158] Zigler KS, McCartney MA, Levitan DR, Lessios HA. Sea urchin bindin
1609 divergence predicts gamete compatibility. *Evolution* 2005, *59*: 2399-2404.

1610 [159] Kenchington EL, Hamilton L, Cogswell A, Zouros E. Paternal mtDNA and
1611 maleness are co-inherited but not causally linked in Mytilid mussels. *PLoS One* 2009, 4:
1612 e6976.

1613 [160] Brannock PM, Roberts MA, Hilbish TJ. Ubiquitous heteroplasmy in Mytilus spp.
1614 resulting from disruption in doubly uniparental inheritance regulation. *Marine Ecology*
1615 *Progress Series* 2013, *480*: 131-143.

1616 [161] Śmietanka B, Burzyński A. Disruption of doubly uniparental inheritance of
1617 mitochondrial DNA associated with hybridization area of European *Mytilus edulis* and
1618 *Mytilus trossulus* in Norway. *Marine Biology* 2017, *164*: 209.

1619 [162] Skibinski DOF, Beardmore JA, Cross TF. Aspects of the population-genetics of
1620 *Mytilus* (Mytilidae, Mollusca) in the British-Isles. *Biological Journal of the Linnean*
1621 *Society* 1983, *19*: 137-183.

1622 [163] Saarman NP, Kober KM, Simison WB, Pogson GH. Sequence-based analysis of
1623 thermal adaptation and protein energy landscapes in an invasive blue mussel (*Mytilus*
1624 *galloprovincialis*). *Genome Biology and Evolution* 2017, *9*: 2739-2751.

1625 [164] Eads AR, Evans JP, Kennington WJ. Plasticity of fertilization rates under varying
1626 temperature in the broadcast spawning mussel, *Mytilus galloprovincialis*. *Ecology and*
1627 *Evolution* 2016, *6*: 6578-6585.

1628 [165] Eads AR, Kennington WJ, Evans JP. Interactive effects of ocean warming and
1629 acidification on sperm motility and fertilization in the mussel *Mytilus galloprovincialis*.
1630 *Marine Ecology Progress Series* 2016, *562*: 101-111.

1631 [166] Oliver M, Evans JP. Chemically moderated gamete preferences predict offspring
1632 fitness in a broadcast spawning invertebrate. *Proceedings of the Royal Society of*
1633 *London B: Biological Sciences* 2014, *281*: 20140148.

1634 [167] Fitzpatrick JL, Simmons LW, Evans JP. Complex patterns of multivariate
1635 selection on the ejaculate of a broadcast spawning marine invertebrate. *Evolution* 2012,
1636 *66*: 2451-2460.

1637 [168] Liu G, Innes D, Thompson RJ. Quantitative analysis of sperm plane circular
1638 movement in the blue mussels *Mytilus edulis*, *M. trossulus* and their hybrids. *Journal of*
1639 *Experimental Zoology Part A: Ecological Genetics and Physiology* 2011, *315*: 280-290.

1640 [169] El-Dessouky HT, Ettouney HM. Fundamentals of salt water desalination.
1641 Elsevier, 2002, Pages 525-563.

1642 [170] Ospina-Álvarez A, Palomera I, Parada C. Changes in egg buoyancy during
1643 development and its effects on the vertical distribution of anchovy eggs. *Fisheries*
1644 *Research* 2012, *117*: 86-95.

1645 [171] Sundby S, Kristiansen T. The principles of buoyancy in marine fish eggs and their
1646 vertical distributions across the world oceans. *Plos One* 2015, *10*: e0138821

1647 [172] Reinhardt K, Dobler R, Abbott J. An ecology of sperm: sperm diversification by
1648 natural selection. *Annual Review of Ecology, Evolution, and Systematics* 2015, *46*: 435-
1649 459.

1650

1651 **Figure Legends**

1652

1653 **Figure 1:** Histological tests of mature male gonads of the six *Mytilus edulis* (a-f) and
1654 six *M. galloprovincialis* (g-l) mussels selected to make each pool for RNA-seq analysis.
1655 There are two different zoom views (see 500 and 50 μm scale respectively, above and
1656 below) shown for each histology test and individual mussel. Ac: male gonadal follicles
1657 with spermatozoa (sp), where heads (hd) and flagella (fl) can be seen and differentiated.
1658 Adipogranular (ag) and vesicular connective tissue (cv) cells can be found between the
1659 spermatic acini.

1660

1661 **Figure 2: a**) Distribution of Level 2 GO terms of loci annotated in three ontological
1662 categories: biological process (BP), molecular function (MF) and cellular component
1663 (CC). Note that only those GO terms with annotations in at least 100 and 10 loci, for BP
1664 and MF respectively are shown. **b**) Enrichment analysis results for GO terms in
1665 differentially expressed loci between mature male gonads of the two *Mytilus* spp.
1666 according to Fisher's exact test (FDR<0.05). DE: differentially expressed, ND: not
1667 differentially expressed set of loci defined after RSEM analysis. Length of bars
1668 represents the percentage of loci annotated for each term in the DE (blue bars) and ND
1669 (red bars) sets. A blue longer than red bar indicates that that GO term is overrepresented
1670 in the differentially expressed loci. GO terms are grouped by their ontological category
1671 (BP, MF, CC), and within category, GO terms are displayed sorted by increasing *p*-
1672 values.

1673

1674 **Figure 3:** 2DE gels showing sperm proteome from a representative *Mytilus*
1675 *galloprovincialis* and *M. edulis* mussel respectively. 45 spots that showed significant
1676 differences between the two *Mytilus* populations and species ($q \leq 0.05$) and were
1677 identified (all except one) by MS (see Table 3) are numbered and encircled.

1678

1679 **Figure 4:** Hierarchical clustering and heat map made using log normalised expression
1680 data for the 45 protein spots of sperm samples that showed significant differences in
1681 expression level ($q \leq 0.05$) between the two *Mytilus* species and populations (SW:

1682 Swansea, VG: Vigo) and were identified (all except one) by MS (see Figure 3). Each
1683 column and row contains information for an individual mussel and protein spot
1684 respectively. The numbers on the right are the protein spot numbers to each of which
1685 isattached an abbreviation that corresponds to gene name that code for the identified
1686 protein (see Table 3). Note that for two identified protein spots (1101 and 1508) there
1687 are no gene name abbreviations available. Cells are coloured according to z-scores,
1688 showing up-regulation (red) or down-regulation (green) of protein spot volumes in the
1689 individual mussels compared with average expression values calculated from all mussel
1690 samples.

1691

1692 **Figure 5:** Volcano plot made with the 727 sperm protein spots analysed by 2DE. Log_2
1693 of the ratio of average expression values between Swansea and Vigo populations (FC)
1694 plotted against log_{10} of *p*-values derived from the one-way ANOVA analysis. Note that
1695 positive and negative Log_2 (FC) values mean higher expression on average in samples
1696 from Vigo (*M. galloprovincialis*) and Swansea (*M. edulis*), respectively. Grey (FC>1.5)
1697 and black (up to 1.5 FC) represent non-significant protein spots ($p>0.05$), while colour
1698 represents protein spots significant after one-way ANOVA ($p \leq 0.05$); blue, <1.5 FC; red,
1699 between 1.5 and 2.0 FC; green, >2.0 FC.

1700

1701 **Figure 6:** Comparative results of protein spot identifications by MS using different
1702 customised protein databases (see Materials and Methods). Bars represent the total
1703 number of peptide spectrum matches (PSMs), total peptides (TP) and unique peptides
1704 (UP), expressed as percentage, obtained against each of the three protein databases
1705 made from: 1) RNA-seq data from the current study (RNA), 2) EST sequences available
1706 in NCBI from Mytilus[organism] (EST), and 3) protein sequences available in NCBI for
1707 Mollusca[organism] (NCBI). *: $p<0.001$, ns: not significant, for Kruskal-Wallis and
1708 *post-hoc* pairwise tests (after Dunn correction to account for multiple comparisons)
1709 between the different protein databases either for the total number of PSMs, TP or UP.

1710

1711 **Figure 7**: Summary of counts and percentages of sperm trait and functional terms for
1712 proteins having higher expression in *M. edulis* and *M. galloprovincialis*. The data is

1713 derived from Table 3 and from File S7, worksheet Table S7 where it is further
1714 elaborated (see captions of Tables S6-S7). Columns 2-5 give the counts and % values of
1715 sperm trait terms assigned to proteins having higher expression in *M. edulis* and *M.*
1716 *galloprovincialis*. Red and green fill indicate higher and lower % values in each row.
1717 Columns 6 and 7 indicate the number of occurrences of terms according to a tentative
1718 hypothesis on perceived benefit of higher expression to the species at the head of the
1719 columns (in red font) or perceived disadvantage (green font). Black font indicates that a
1720 conclusion in relation to benefit or disadvantage could not easily be made.

1721

1722 **Table 1:** Summary results from RNA-seq data and annotation through Blast analysis
1723 against different databases: 1) all protein sequences available in SwissProt
1724 (UniProtKB/SwissProt), 2) the Pacific oyster *Crassostrea gigas* genome
1725 (Oyster_Genome), 3) all EST sequences available in NCBI from "Mytilus", 4) protein
1726 sequences retrieved from NCBI for "Mytilus" (NCBI_MytProt), and 5) protein
1727 sequences retrieved from NCBI for "Mollusca" (NCBI_MolluscaProt). See further
1728 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%)]*

1729 (*) 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 chec 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 re
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 n
430	Npr1	*Atrial natriuretic peptide receptor 1*	1.7 G	RNA, NCBI	Membrane	H c
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		n
847			5.6 E	RNA, EST		
705b	Dld	*Dihydrolipoyl dehydrogenase*	1.8 E	RNA, EST, NCBI	Mitochondrion	C n
1119	Efhc2	EF-hand domain-containing family	4.7 G	RNA, EST, NCBI	Ubiquitous	C
1134		member C2	2.4 E	RNA, EST, NCBI		n
191			2.3 G	RNA, EST, NCBI		

1536	Etfb	*Electron transfer flavoprotein subunit beta*	1.4 G	RNA, EST, NCBI	Mitochondrion	R n
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 **2062**	Glud	*Glutamate dehydrogenase*	2.0 G 2.5 E	RNA, EST, NCBI RNA, EST, NCBI	Mitochondrion	C
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 **1087**	Idh3g	*Isocitrate dehydrogenase [NAD] subunit gamma*	3.1 E 1.5 E	RNA, EST, NCBI RNA, EST	Mitochondrion	T
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 n
2108	Plc	*Perlucin*	1.8 E	RNA	Extracellular region	S
1919	Prdx5	*Peroxiredoxin-5*	2.2 E	EST, NCBI	Mitochondrion	A i
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 a

Figure 1

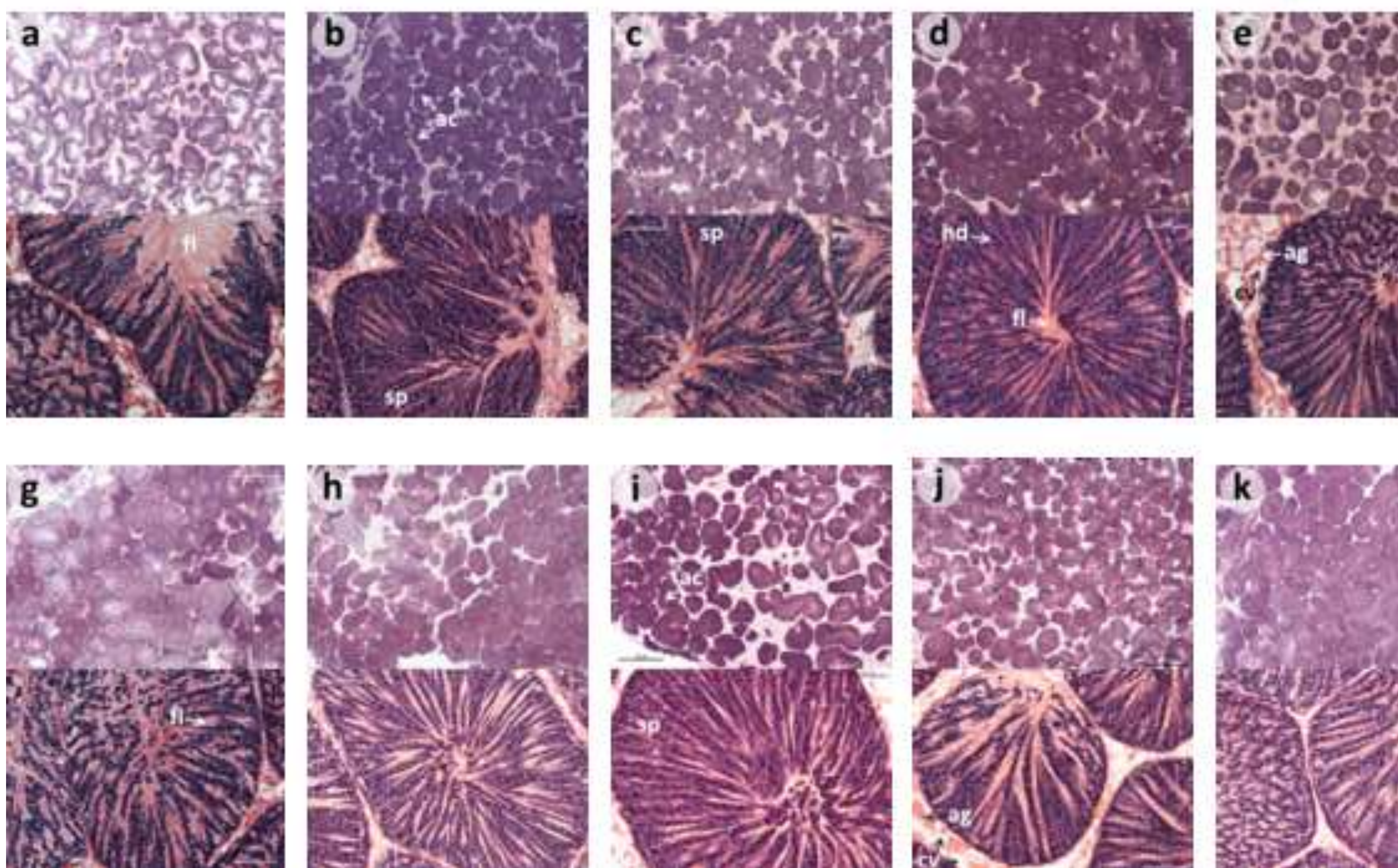


Figure 2

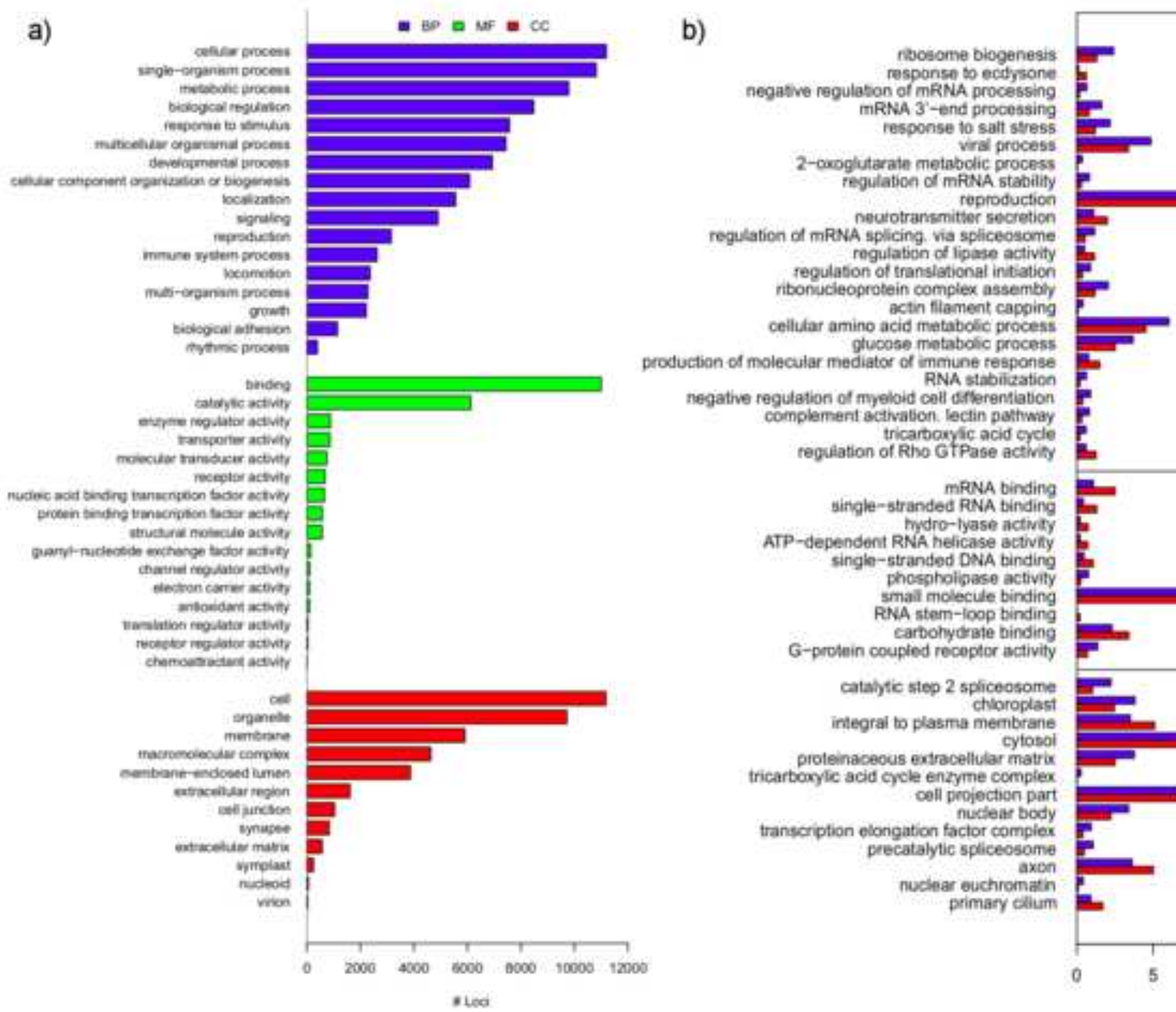
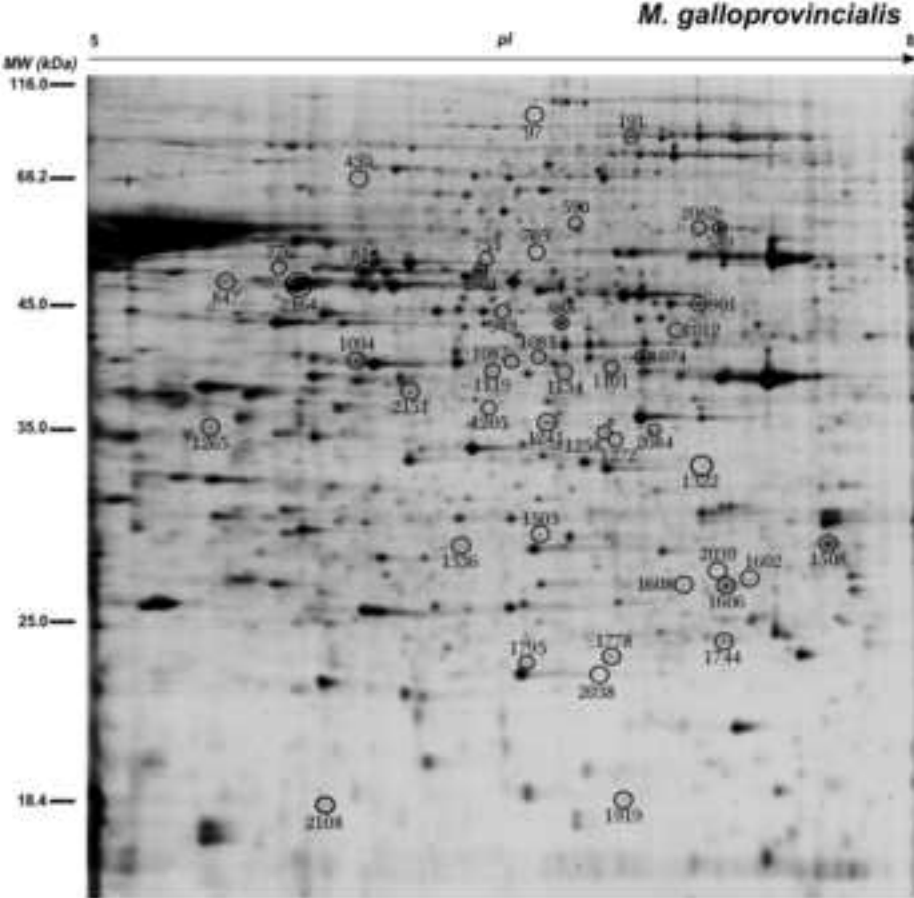


Figure 3



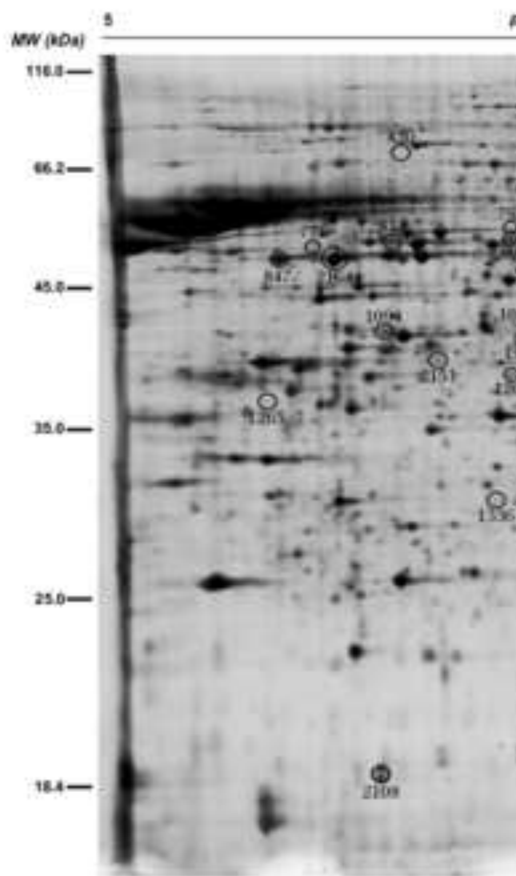


Figure 4

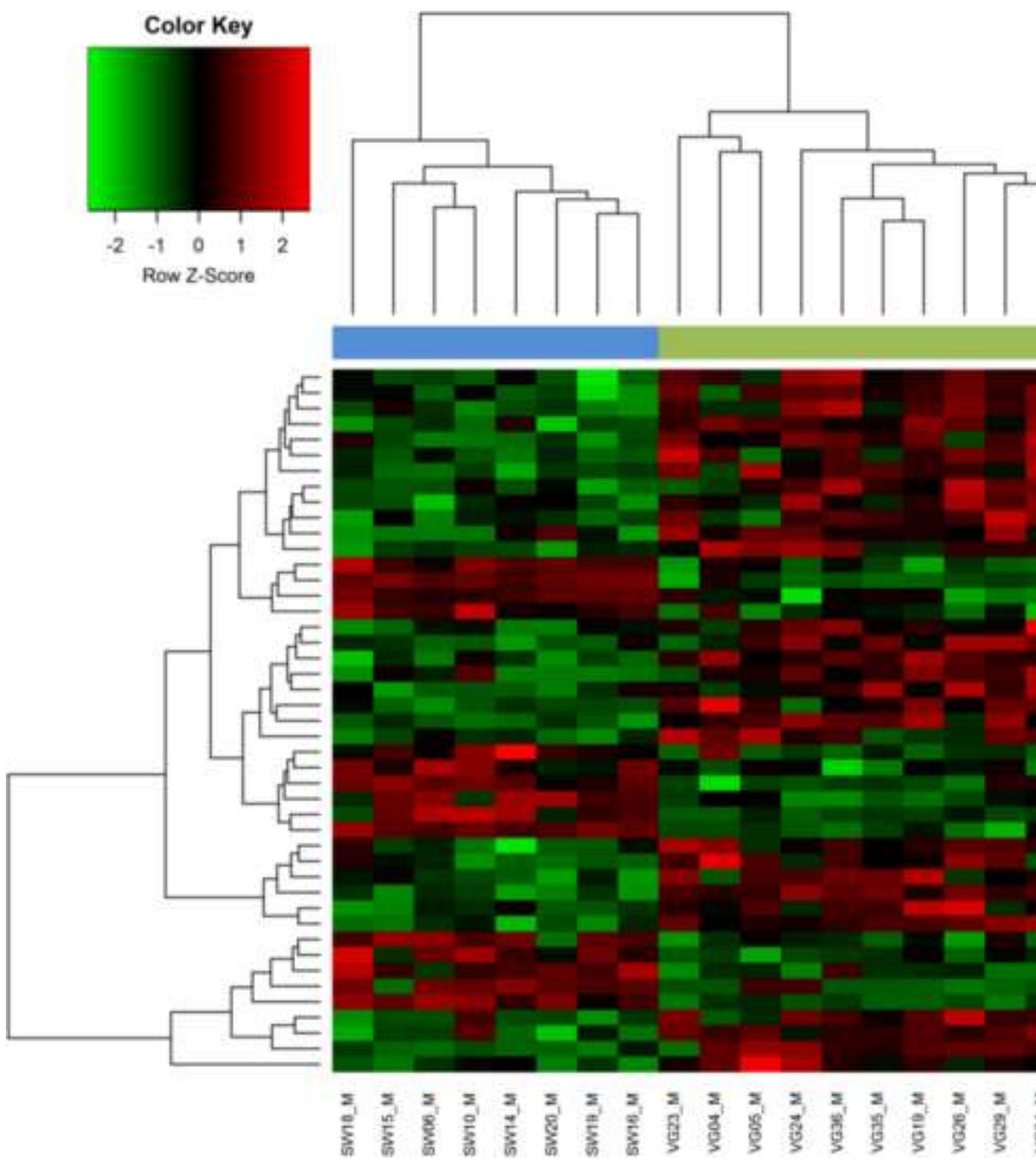


Figure 5

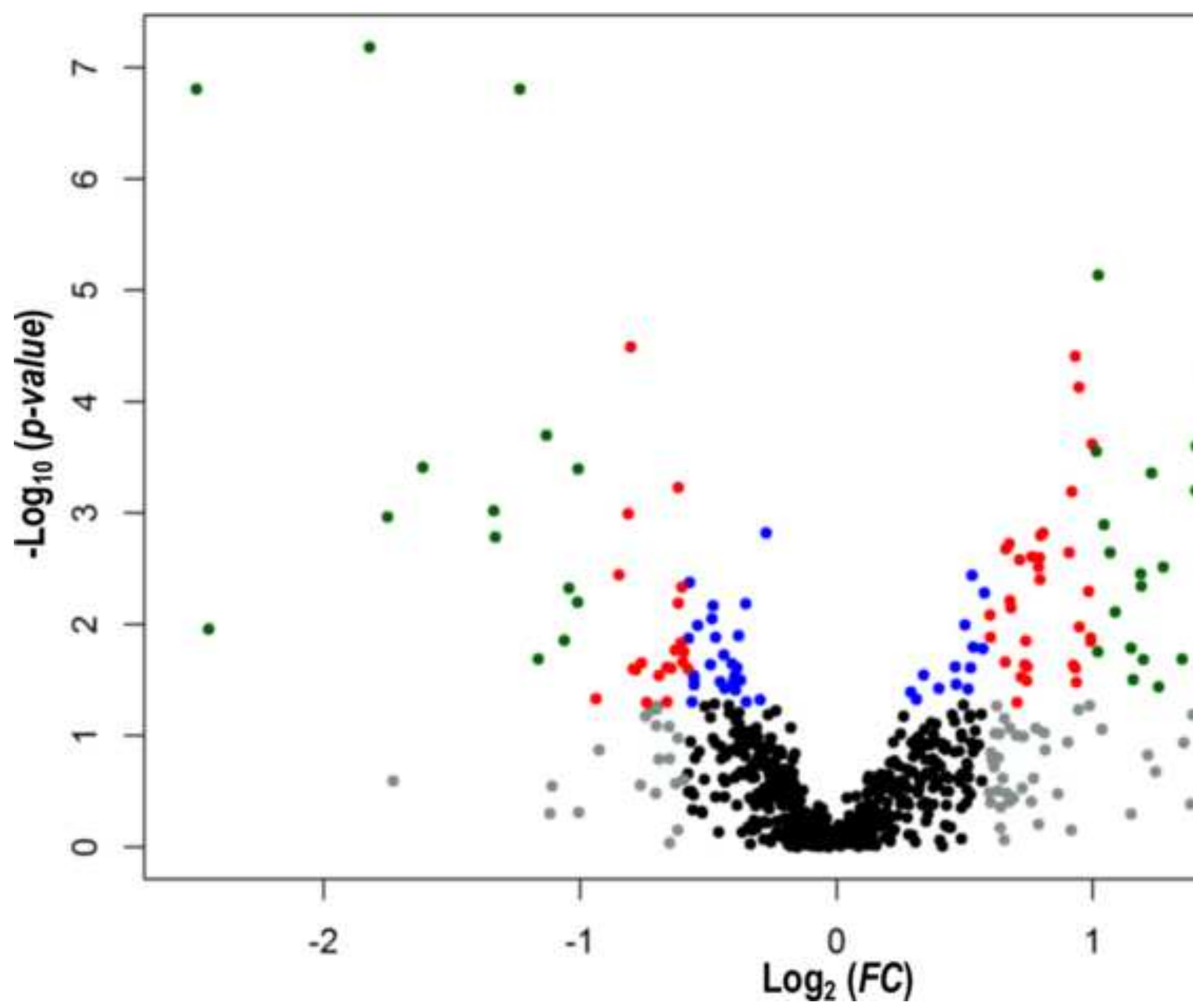


Figure 6

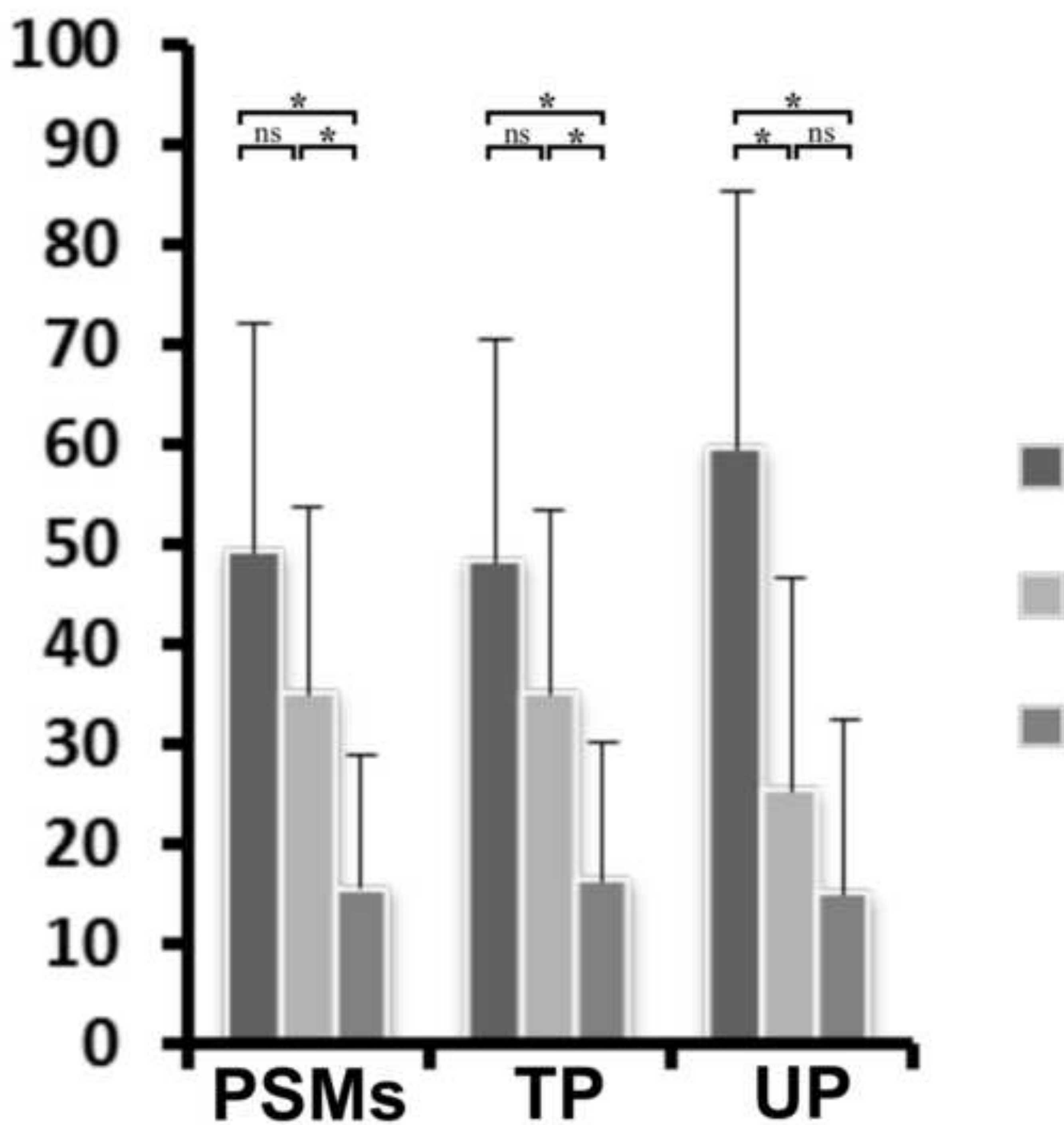


Figure 7

Sperm trait terms	Higher protein expression in *M. edulis* Count	%	*M. galloprovincialis* Count	%	benefi *M. ed* 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