Proteogenomics of *Gammarus fossarum* to Document the Reproductive System of Amphipods*§

Judith Trapp‡§, Olivier Geffard‡¶, Gilles Imbert§, Jean-Charles Gaillard§, Anne-Hélène Davin§, Arnaud Chaumot‡, and Jean Armengaud§¶

Because of their ecological importance, amphipod crustacea are employed worldwide as test species in environmental risk assessment. Although proteomics allows new insights into the molecular mechanisms related to the stress response, such investigations are rare for these organisms because of the lack of comprehensive protein sequence databases. Here, we propose a proteogenomic approach for identifying specific proteins of the freshwater amphipod *Gammarus fossarum*, a keystone species in European freshwater ecosystems. After deep RNA sequencing, we created a comprehensive ORF database. We identified and annotated the most relevant proteins detected through a shotgun tandem mass spectrometry analysis carried out on the proteomes from three major tissues involved in the organism’s reproductive function: the male and female reproductive systems, and the cephalon, where different neuroendocrine glands are present. The 1,873 mass-spectrometry-certified proteins represent the largest crustacean proteomic resource to date, with 218 proteins being lineage specific. Comparative proteomics between the male and female reproductive systems indicated key proteins with strong sexual dimorphism. Protein expression profiles during spermatogenesis at seven different stages highlighted the major gammarid proteins involved in the different facets of reproduction. *Molecular & Cellular Proteomics* 13: 10.1074/mcp.M114.038851, 3612–3625, 2014.

Next-generation proteomics, relying on ultra-rapid and sub-parts-per-million mass spectrometry analyzers, is able to offer in-depth insights into the molecular players sustaining the physiology of complex organisms. Identifying and quantitating thousands of proteins has become, over the past decade, a routine task for most proteomic platforms with the development of high-throughput shotgun proteomics. The interpretation of large-scale MS/MS data is only possible if a high-quality database of nucleic acid sequences is available. Homology-driven proteomics using cross-species matching is a first alternative if the genome of interest is unknown, and *de novo* sequencing (i.e. interpretation of MS/MS data to establish the exact sequence of each peptide from scratch) is another possibility. However, major drawbacks of these two approaches lead to a scarcity of results as soon as a non-model organism, distantly related to a sequenced organism, is analyzed. Indeed, only highly conserved and ubiquitous proteins will be identified and carefully annotated with such approaches. It is noteworthy that lineage-specific genes are more likely to be linked to the organism’s unique biology, as demonstrated by the characterization of the genome of *Daphnia pulex*, the water flea, a freshwater microcrustacean whose orphan genes have been shown to be among the most ecoresponsive (1). Consequently, gene products from non-model organisms responding to environmental challenges are overlooked.

Among amphipods, the genus *Gammarus* represents the greatest number of epigean freshwater species distributed throughout the Northern Hemisphere. In Europe, two closely related species, *Gammarus pulex* and *Gammarus fossarum*, are keystone species for freshwater ecosystems. They are commonly used as sentinel species in freshwater risk assessment, for several reasons (2). First, they are widespread and found throughout a large habitat range, where they often occur at high densities. Second, they occupy a large trophic repertoire: herbivores, predators, and detritivores playing a major role in leaf-litter breakdown processes (3). They also constitute a food reserve for macroinvertebrates and fish. Finally, gammarids can be easily maintained in the laboratory or used for *in situ* bioassays (2), in which one can assess the impact of pollutants by measuring molecular markers related to diverse modes of action, such as neurotoxicity (4), as well as by using life-history-trait reproductive features (5). Alterations of sexual phenotype (intersexuality) have been reported *in situ* (6), as well as alterations by xenobiotics of various physiological parameters related to reproductive success (i.e.
gametogenesis, embryogenesis, fecundity, or molt (5). However, the molecular mechanisms involved in these reproductive impairments are unknown. A major reason for this is that hormones and proteins involved in the regulation of reproductive function in *G. fossarum*, and in crustacea in general, have until recently been largely misunderstood (7). Although *Gammarus* is a relevant ecotoxicological animal model, comprehensive genomic resources are lacking for this genus, and also for crustacea within the tree of life in general. To date, only the genome of the zooplanktonic branchiopod *D. pulex*, a laboratory-cultured organism with clonal reproduction, has been released (1). This organism is a popular ecotoxicological model because of its ease of rearing, but the large phylogenetic distances between crustacea, and the ecology of other classes of crustacea, lead some to question the insightfulness of this model (8). Furthermore, in the case of assessing reprotoxicity, different life-history strategies (e.g. parthenogenic versus sexual reproduction) can have different consequences concerning vulnerability to pollutants (for example, different physiological manifestations).

A novel approach that intimately combines genomics and proteomics, namely, proteogenomics, emerged (reviewed in Refs. 9–11) after the pioneering work of Yates et al. (12), which used combined expressed sequence tag (EST) databases translated into six reading frames for the interpretation of MS/MS spectra. Proteogenomics sensu stricto consists of better annotating genomes by means of high-throughput proteomic data, and this approach has been exemplified for several bacteria (13, 14) and eukaryotes (15, 16). Numerous protein-encoding genes can be identified using mass spectrometry data when missed by automatic annotation software programs, or their structures can be corrected through experimental validation of their N-terminal translational starts. Recently, proteogenomics-style approaches in which genome-sequencing data are used directly to interpret large proteomic datasets, and vice versa, have flourished, although they are not directly aimed at genome annotation or reannotation (17–19). On this basis, a sensu lato definition of proteogenomics has been discussed to quality projects based on multi-omics data, with closely linked nucleic acid and protein information, including, typically, the use of six-frame translation (20). Typically, the use of six-frame translation (20). In this vein, Ning and Nesvizhskii (21), Nagarai et al. (22), Wang et al. (23), and Woo et al. (24) noted the interest of sample-specific protein databases established from RNA-Seq data, which can enable a straightforward analysis of MS/MS data. Although comprehensive coverage of transcriptomes via RNA-Seq has been achieved for several marine species, studies on amphipods remain scarce and have been restricted to *Parhyale hawaiensis* (25, 26) and *Melita plumulosa* (27).

Here, we have documented gammarid reproductive function through the identification and characterization of novel proteins. For this, we used a proteogenomic approach aimed at quickly pinpointing the proteins from *G. fossarum*, a model of interest in terms of ecotoxicology. After deep RNA-Seq of total RNAs of specific organs from individuals sampled in a wild population of gammarids, we created a six-reading-frame translation protein database for interpreting a large shotgun tandem mass spectrometry dataset obtained in *G. fossarum* specimens. We detail and discuss the proteomes of key physiological organs: the female reproductive system, the male reproductive system, and the cephalon, where several neuroendocrine glands are located. To highlight protein candidates involved in reproductive physiology, we first performed comparative proteomics between the male and female reproductive systems to pinpoint key proteins with strong sexual dimorphism. Subsequently, protein expression profiles during spermatogenesis at seven different stages were analyzed to uncover new proteins that could be directly related to spermatogenesis in *G. fossarum*.

**EXPERIMENTAL PROCEDURES**

**Sampling and Maintenance of Animals**—The amphipods were sampled at La Tour du Pin (latitude, 45°56′44″; longitude, 5°45′11″, upstream of the Bourbre River (mid-eastern France). This site has good water quality according to data records from the Réseau National de Bassin (French Watershed Biomonitoring Network) and accommodates high densities of gammarids. Sexually mature organisms were collected by kick sampling using a net. They were sieved (>2 mm), stored in plastic bottles containing ambient freshwater, and quickly transported to the laboratory. During an acclimatization period of at least 15 days, the organisms were kept in 30-l tanks continuously supplied with aerated, uncontaminated groundwater adjusted to the sampling site conductivity (i.e. 600 μS cm⁻¹). A 16/8 h light/dark photoperiod was maintained, and the temperature was kept at 12 °C ± 1 °C. The gammarids were fed *ad libitum* with alder leaves (*Alnus glutinosa*) and supplemented with freeze-dried worms (*Tubifex tubifex*) twice a week.

**Preparation of Biological Samples**—Gammarids in amplexus were selected to cover a wide distribution range of body sizes (from 1 to 2 cm) and dissected under stereomicroscope magnification, as described by Lacaze et al. (28). The third segment of the urosoma was cut with microscissors, the cephalon was removed from the body with fine forceps, and the attached caeca were cut. Dorsal and ventral cuticles were then excised. The male gonads, including the seminal vesicle, were gently removed with fine forceps, as were oocytes. For RNA library preparation, each tissue was collected as one tissue-type sample using 11, 9, 57, and 173 organisms for cephalons, caeca, oocytes, and testes, respectively. For the proteomes, only adult organisms sampled using 2-mm and 2.5-mm sieves and determined to be *G. fossarum* on the basis of phenotypic criteria were selected. Each tissue was collected as one tissue-type sample using 44, 26, and 64 organisms for cephalon, oocyte, and testis tissues, respectively. Proteomes were investigated in duplicate for cephalons and in triplicate for reproductive tissues. All pooled samples were immediately frozen in liquid nitrogen, in the presence of either RNAlater RNA Stabilization Reagent (Qiagen, Courtaboeuf, France) or cold protein lysis buffer (50 mM Tris, pH 7.8, 6 mM urea, 2 mM thiourea, 0.1% (v/v) Triton X-100, 4% CHAPS, 65 mM DTT, and 2% Protease Inhibitor Complete Mini, EDTA-free (Roche)), and stored at −80 °C until required.
Construction of a Paired-end RNA-Seq Library and Illumina Sequencing—Tissues were disrupted with a TissueRuptor (Qiagen), and total RNA was isolated using the RNeasy Mini Kit (Qiagen) according to the manufacturer’s instructions. Total RNAs for each tissue sample were quantified using a NanoDrop 2000 spectrophotometer (Thermo Scientific; Wilmington, DE), and their qualities were verified using an Agilent Bioanalyzer 2100 RNA Nanochip (Agilent Technologies, Santa Clara, CA). Similar amounts of total RNAs from each tissue sample were pooled in order to create an equimolar sample for library construction representing the four tissues equally. The cDNA library for Illumina mRNA-Seq (Illumina, Hayward, CA) was constructed with 5 μg of total RNA. The manufacturer’s protocol was carried out with some modifications. In brief, poly-A-containing mRNA molecules were purified using poly-T oligos that were attached to magnetic beads. The purified mRNA was fragmented by the addition of 5× fragmentation buffer (Illumina) and heated to 94 °C in a thermocycler for 5 min. This fragmentation time is the default setting yielding nucleic acid fragments of ~250 bp. First-strand cDNA was synthesized using random primers to eliminate the general bias toward the 3’ end of the transcripts. Second-strand cDNA synthesis was performed by adding GEX second-strand buffer (Illumina), dNTPs, RNaseH, and DNA polymerase I and then incubating for 2.5 h at 16 °C. Second-strand cDNA was further subjected to end repair, A-tailing, and adapter ligation, as recommended. Purified cDNA templates were enriched by 15 cycles of PCR for 10 s at 98 °C, 30 s at 65 °C, and 30 s at 72 °C using PE1.0 and PE2.0 primers and Fastart TaqDNA polymerase (Roche). The samples were cleaned using QIAquick PCR purification columns and eluted in 30 μl of elution buffer (Qiagen). Purified cDNA libraries were quantified using a 2100 Bioanalyzer DNA 100 Chip (Agilent). Sequencing was performed on the Illumina/Solexa Genome Analyzer according to the manufacturer’s instructions.

Contig Assembly and Creation of a Customized Protein Sequence Database—EST sequence analysis and assembly were performed by the Skuldtech, Montpellier, France Company. ESTs were assembled into clusters using the Velvet and MIRA software programs. In order to obtain highly reliable consensus sequences, overlapping identity percentage and minimum overlapping length parameters were set to 90% and 30 bp, respectively. Contigs resulting from the assembly of multiple sequences are referred to as unique sequences, and sequences that do not fit into any assemblies are called singletons. Singletons are mainly chimeras or artifacts from cDNA synthesis, sequencing, or assembly, resulting in an increase of protein sequence errors, inappropriate frame shifts, and premature ORF terminations. Thus, only the unique sequences were translated into six reading frames, resulting in a customized protein sequence database, named GFOSS, containing 1,311,444 sequences totaling 289,084,257 amino acids.

Protein Extraction and Trypsin Digestion—For protein extraction, tissues were disrupted with 2 ml of cold protein lysis buffer with the TissueRuptor instrument (Qiagen). Homogenates were sonicated for 3 min using a Hielscher Compact Lab Homogenizer (amplitude of 1,311,444 sequences totaling 289,084,257 amino acids.

Protein Extraction and Trypsin Digestion—For protein extraction, tissues were disrupted with 2 ml of cold protein lysis buffer with the TissueRuptor instrument (Qiagen). Homogenates were sonicated for 3 min using a Hielscher Compact Lab Homogenizer (amplitude of 30% and cycle 0.3). Samples were centrifuged for 15 min at 12,000 × g and 4 °C. Clear supernatants were collected and delipidated by the addition of 3 volumes of ethanol/diethyl ether solution (1:1, v/v) per 1 volume of sample, as previously reported by Simon et al. (29). The mixtures were vortexed and incubated on ice for 10 min. Samples were then centrifuged at 16,000 × g for 15 min at 4 °C. The resulting supernatants were removed. The pellets were washed twice with 200 μl of cold acetone and then resuspended in 300 μl of trypsin buffer (50 mM ammonium bicarbonate, 1 mM urea, and 0.01% ProteaseMAX surfactant (Promega, Charbonnieres les Bains, France)). During the extraction procedure, samples were kept on ice. The protein concentration of the samples was determined in microplates using the Coomassie Standard Protein Assay Kit (Interchim, Montluçon, France) with bovine serum albumin as the standard. The protein samples were adjusted to 2 μg μl⁻¹ via dilution with trypsin buffer. Samples were treated with DTT (1 mM final concentration) for 15 min at 56 °C under constant agitation to reduce cysteines. These were alkylated with 3,75 mi iodoacetamide for 15 min at room temperature in the dark. Proteolysis was performed overnight with 2% sequencing-grade trypsin (Roche) dissolved in 0.01% trifluoroacetic acid.

Isoelectric Focusing of Peptides—After digestion, peptides were resolved by their isoelectric point, according to the manufacturer’s instructions, using an Agilent 3100 OFFGEL Fractionator (G3100 AA, Agilent) with a High Resolution Kit (i.e., pH 3–10) (5188–6424, Agilent). A total of 250 μg of digested peptides was loaded into the 24-well fractionation device and focused at a 50-μA current and 200 μW of power until 50 kVh was reached. After focusing, samples were washed online with Micro SpinColumns (#74–4601, Harvard Biosciences, Les Ulis, France) according to the manufacturer’s instructions. Peptides were eluted with 50 μl of 50% aqueous acetoniitrile solution containing 0.1% trifluoroacetic acid.

Nano-LC-MS/MS Analysis—Nano-LC-MS/MS experiments were performed on an LTQ-Orbitrap XL hybrid mass spectrometer (Thermo-Fisher) coupled to an UltiMate 3000 LC system (Dionex-LC Packings, Courtaboeuf, France). Peptide samples (2 μl) were loaded and desalted online on a reverse-phase precolumn (C18 PepMap 100 column, LC Packings) and then resolved on a nanoscale C18 PepMap™ 100 capillary column (LC Packings) at a flow rate of 0.3 μl/min with a gradient of CH3CN, 0.1% formic acid prior to injection into the ion trap mass spectrometer. Peptides were separated using a 90-min gradient from 5% to 60% solvent B (0.1% HCOOH, 80% CH3CN). Solvent A was 0.1% HCOOH, 100% H2O. Full-scan mass spectra were measured from m/z 300 to 1800 with the LTQ-Orbitrap XL mass spectrometer in data-dependent mode using the TOP3 strategy. In brief, a scan cycle was initiated with a full scan of high mass accuracy in the Orbitrap followed by MS/MS scans in the linear ion trap on the three most abundant ions.

Proteomic Profiles of Male Reproductive Tissues during Spermatogenesis—Expression profiles were performed on G. fossarum males at seven different spermatogenesis stages, which have been described previously (28): before copulation (male in amplexus) and after copulation (Days 0, 1, 2, 3, 4, and 7). Precopulates were individually selected and isolated in 500-ml polyethylene beakers under the same experimental conditions as mentioned above. All beakers were checked daily to determine the date of reproduction, attested to by the female ecdisis (5). For each stage, five biological replicates were performed. Male reproductive tissues were sampled individually, immediately frozen in liquid nitrogen, and stored at −80 °C until required. Proteins were directly dissolved in 40 μl of LDS sample buffer (Invitrogen). Samples were sonicated for 1 min in a transonic 780H sonicator and boiled for 5 min at 95 °C. Proteins were resolved via SDS-PAGE with a short migration of 10 min at 150 V on 4%-12% gradient, 10-well NuPAGE gels (Invitrogen) run with MES buffer (Invitrogen). Gels were stained with Coomassie Blue Safe stain (Invitrogen). After overnight destaining with water, the whole protein content from each well was extracted as a sole polycrylamide band and processed for further destaining and iodoacetamide treatment. The samples were then proteolyzed with sequencing-grade trypsin (Roche) using 0.01% ProteaseMAX surfactant (Promega). The resulting peptide mixtures were diluted 1:20 in 0.1% trifluoroacetic acid and analyzed via nano-LC-MS/MS.

MASCOT Database Mining—Peak lists were generated with Mascot Daemon software (version 2.3.2, Matrix Science, London, United Kingdom) using the extract msn.exe data import filter (Thermo). Data import filter options were set to 400 (minimum mass), 5000 (maximum mass), 0 (grouping tolerance), 0 (intermediate scans), and 1000 (threshold), as described previously by Christie-Oleza et al. (30). MS/MS spectra were assigned to peptide sequences with the Mascot Daemon 2.3.2
search engine (Matrix Science) against the NCBInr protein sequence database or the GFOSS database. Searches for tryptic peptides were performed with the following parameters: full-trypsin specificity, mass tolerance of 5 ppm on the parent ion and 0.5 Da on the MS/MS, static modifications of carboxyamidomethylated Cys ($\Delta H_1100157.0215$), and dynamic modifications of oxidized Met ($\Delta H_1100115.9949$). The maximum number of missed cleavages was set at two. All peptide matches with a Mascot peptide score less than a $p$ value of 0.05 were filtered by IRMa 1.30.4 software. A protein was considered validated when at least two different peptides were detected in the same experiment. The false-positive rate for protein identification was estimated, through a search with a reverse decoy database, as less than 0.1% using the same parameters in both the NCBInr protein sequence database and the GFOSS database.

Statistical Data Treatment for Comparative Proteomics—Spectral counts (the number of spectra recorded per protein) were extracted from the spectra-to-peptide dataset, and the normalized spectral abundance factor for each protein was calculated as described previously (31). For comparative proteomics between male and female reproductive tissues, the data were analyzed with the Tfold module of the PatternLab software program. No specific normalization of spectral counts was applied, as strictly equal quantities of protein samples were analyzed. Fold change and $p$ value cutoffs were set at 2 and 0.05, respectively. The false discovery rate (Benjamini–Hochberg $q$-value) was 10%. The F-Stringency factor was systematically optimized. For analysis of the proteome dynamics along the different spermatogenesis stages, the data were clustered with the TrendQuest module of PatternLab. We retained in the final list of proteins for temporal profiling the proteins validated with at least two distinct peptides when merging all the data. Again, no specific normalization of spectral counts was applied. Proteins observed in at least three of the replicate samples were considered for clustering; the minimum signal required was 5. Cluster health was set at 0.9, and the minimum number of items per cluster was three.

Sequence Analysis—A BLASTp search was performed with the NCBI website facilities using the nonredundant protein sequences. The BlastX algorithm (version 2.2.15, GenBank release number 166) was used with a threshold E-value set at $1E^{-10}$.

Data Repository—This whole transcriptome project has been deposited in the EMBL Nucleotide Sequence Database under the accession number PRJEB5098 (www.ebi.ac.uk/ena/data/view/PRJEB5098). The mass spectrometry proteomics data have been deposited in the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org/) via the PRIDE partner repository (32) with the dataset identifier PXD000576.

RESULTS

General Proteogenomic Strategy for Discovering Key Proteins Involved in Reproduction from a Nonmodel Complex Organism—To represent as accurately as possible the real ecology of gammarids, we sampled heterogeneously sized organisms from a field population located in eastern central France. Because our focus was on proteins involved in reproduction, only sexually mature organisms were chosen, by selecting those in precopulatory mate guarding (Fig. 1, panel A). We first established a database of nucleic-acid-derived protein coding sequences that could be a generic platform for ecotoxicology-related proteomic studies on gammarids, as shown in Fig. 2. For this, we extracted total RNAs from four key tissues: the male and female reproductive tissues (Fig. 1, panels C and D); the cephalon, where several neuroendocrine glands are located (Fig. 1, panel B); and the hepatopancreatic caeca (Fig. 1, panel B), involved in xenobiotic detoxification and energy acquisition (digestive enzyme secretion, food absorption, and nutrient storage). Equal amounts of RNA from each tissue were pooled and sequenced in a single Illumina sequencing run (Fig. 2). The resulting nucleic acid information was used to create a protein database named GFOSS that could be used for proteogenomic identification of proteins. For this, each mRNA contig was systematically translated into the six possible reading frames, considering all possible putative ORFs. As in previous proteogenomic studies, tandem

Fig. 1. Images of G. fossarum in amplexus and sampled tissues. A, sexually mature organisms in precopulatory mate guarding. B, the cephalon and the four attached hepatopancreatic caeca. C, a female oocyte. D, transversal view of a male gammarid with attached reproductive tissue. M, male reproductive tissue (orange); F, female reproductive tissue (black); C, cephalon; H, hepatopancreatic caeca; ovrd., oviduct.
mass spectrometry data could be assigned with such a database (23). The protein catalog established from experimental evidence allows the selection of the appropriate reading direction and frame of each protein-coding mRNA contig. As indicated in Fig. 2, we recorded a large dataset on the proteomes from three tissues by resolving each sample via OFFGEL electrophoresis and performing 24 nano-LC-MS/MS runs per sample. In order to identify proteins related to reproductive function, we first performed automatic functional prediction based on cross-species homology. To discover new proteins involved in this function, we then compared the male and female reproductive tissues with a label-free comparative approach. We also focused our attention on the proteome dynamics of the male reproductive tissue to indicate new proteins related to spermatogenesis in *G. fossarum*.

**Transcriptome Characterization via RNA-Seq**—A normalized cDNA library was constructed from the pool of mRNAs extracted from the male and female reproductive tissues, the hepatopancreatic caeca, and the cephalon of sexually mature gammarids. The library generated 323,731,722 paired-end reads of ~100 bp. After low-quality read filtering, de novo assembly of the Illumina sequences was carried out with a two-stage strategy. First, 1,406,956 sequences were preassembled with an N50 of 334 nt using the Velvet program. A dataset of 218,574 contigs with N50 and N25 of 799 nt and 1,221 nt, respectively, was then created with the Mira software program. The average guanine-cytosine (GC) content was 47.9%. The total length of mRNA sequences was 145,008,280 nt. A total of 386,238 reads referred to as singletons were not incorporated into the final assembly. Among the assembled contigs, 127,332 (58%) produced significant hits against the nonredundant NCBI database when we used an E-value threshold of 1E^-10 and the BLASTp search tool (supplemental Table S1). We extracted the taxonomic information associated with the best protein homolog for each significant hit. The greatest number of contigs (8,927, or 4%) showed the most similarity to protein sequences from the crustacean *D. pulex*, and another significant number (4,053, or 2%) were similar to proteins from the beetle *Tribolium castaneum*. It is noteworthy that few similarities were found to proteins annotated from the amphipod *P. hawaiensis* using this metric (221 contigs), although the maternal and embryonic transcriptomes are available for this organism (26).

In addition, 33 BLASTp results were annotated as cytochrome c oxidase subunit 1 (CO1) from diverse species, mainly crustacea (supplemental Table S1). Because this gene has been proposed as a genetic marker for species assignment in Gammarus (33), we performed a thorough search for CO1 sequences from GFOSS via BLASTn using 376-bp sequences available for organisms from a regional study covering the area of origin of our population (accession numbers X97163–X97185 (34)). Phylogenetic tree reconstruction (based on the classification proposed by Meyran *et al.* (34)) revealed the presence of different *G. fossarum* lineages concomitant with *G. pulex* lineages. The wild population sampled for RNA-Seq was thus an assembly of different lineages of the *G. pulex/fossarum* species complex. Our RNA-Seq-derived database thus constitutes a resource that is potentially of interest for studies in gammarids other than the *G. fossarum* species.

**Establishment of Protein Catalogs in G. fossarum**—Our proteogenomic experiment led to a global dataset of 1,148,367 recorded MS/MS spectra. These indicated a total of 7,644 peptide sequences and certified 1,873 protein sequences. A total of 218 (12%) proteins were lineage-specific proteins (i.e. orphans) and are described here for the first time. This relatively high number of mass-spectrometry-certified orphans indicates
the phylogenetic distance of *G. fossarum* among the already-sequenced animal models and suggests that crustacea in general are poorly characterized in terms of molecular biology. Considering the taxonomic information obtained through the species identity of the most closely related homologous sequence, a relatively low number of proteins showed significant sequence similarities to protein sequences from *D. pulex* (159, 8%) and from the beetle *T. castaneum* (70, 4%). A summary plot reflecting the global proteome is presented in Fig. 3. As shown in Fig. 3A, where the distribution of the proteins in terms of label-free quantitation is shown for groups of proteins ranked from the most abundant to the least abundant, the ratio of orphans was remarkably constant. Fig. 3B indicates the occurrence of the MS/MS-detected proteins in the different tissues. The specific study of these proteins appears worthy because of the relatively minimal overlap between tissues. The relatively low number of RNA contigs certified by mass spectrometry in the study may be explained by differences in (i) the dynamic range between proteome and transcriptome; (ii) the nature of the samples analyzed, as the hepatopancreatic caeca were not included in the proteomic study; and (iii) the redundancy of RNA sequences that might be due to site-specific variants resulting from the heterogeneity of the animal population, diversity in terms of splicing variants, sequencing errors with the Illumina high-throughput approach, and some assembly errors resulting from the fact that no genome backbone was available for the assembly. The assignment rate of the MS/MS spectra was low, especially with the OFFGEL approach (5%) relative to the gel-based shotgun approach when analyzing the same tissue (*i.e.* testis tissues), but this figure agrees with a previous interpretation of data acquired for two other complex animals (35). Indeed, the high dynamic range of the proteome in these complex organisms, as well as the diversity of the proteome, leads to noisy MS/MS spectra for low-abundance proteins. Post-translational modifications could be an additional factor of complexity.

The Cephalon Proteome—When we used the NCBI database, a relatively low number of MS/MS spectra (3,103) could be assigned among the 231,098 MS/MS spectra recorded. The 878 unique peptide sequences (supplemental Table S2) proved the presence of 135 proteins (supplemental Table S3). Using our customized, RNA-Seq-derived database, 10,647 MS/MS spectra could be assigned, indicating the existence of 2,576 distinct peptide sequences (supplemental Table S4). In this case, a total of 546 mRNA-translated products (supplemental Table S5) were identified. Using the BLASTp tool, we observed that 505 (92%) exhibited significant sequence similarities with an already functionally annotated protein from another sequenced organism (E-value less than 1E−10). Of note, relative to the results obtained with the NCBI database, our customized database led to a large increase in the number of identified peptides (∼3-fold) and, consequently, proteins (∼4-fold).

Ovary and Testis Proteomes—Using the GFOSS database, among the 700,078 MS/MS spectra recorded, we were able to assign a total of 36,516 (supplemental Table S6). These validated a large list of peptides (5,435 unique peptide sequences), which allowed the identification of 1,219 mRNA-translated products (supplemental Table S7). Among these, 1,069 proteins (88%) showed significant similarities to previously annotated proteins from other organisms. The other 150 proteins (12%) can be referred to as orphans. The
level of orphans is higher (1.5×) for these reproductive tissues than has been observed for the cephalon proteome, highlighting the lack of knowledge regarding the key molecular players for both reproductive tissues. Furthermore, among the five proteins contributing 10% of the ovary and testis proteomes in terms of quantity, as estimated based on the normalized spectral abundance factor, one (I.D. 68651) does not have any functional annotation. Fig. 4 shows the complementary strategy followed for establishing the protein profiles of male reproductive systems. These were analyzed at seven different spermatogenesis stages, on five biological replicates. Using the GFOSS database, among the 217,191 MS/MS spectra recorded, we were able to assign a total of 50,793 (supplemental Table S8). These validated a large list of peptides (2,699 unique peptide sequences) and allowed label-free quantitation of 914 translated contigs (supplemental Table S9). Regarding the previous analysis of the male reproductive tissue proteome, 437 additional products were newly identified here. Among the 914 proteins, 853 items (93%) showed significant similarities to previously annotated proteins from other organisms. The other 61 proteins (7%) can be referred to as orphans. In this global dataset, only three proteins contributed 10% of the proteome in terms of quantity: two histones (I.D. 13250 histone H2A and I.D. 65834 histone H3.2-like) and one protein (I.D. 100349) for which the function could not be predicted.

Identification of G. fossarum Proteins Associated with Reproductive Function

Automatic Functional Annotation—The functional annotation of the 1,873 identified proteins was performed via cross-species prediction based on sequence similarity searches. The function of a total of 1,216 proteins (65%) could be annotated with relative confidence (threshold E-value set at 1E−10), as marked biological differences between phylogenetically distant organisms may lead to the acquisition by sequence-related proteins of very different functions that can be differentiated only through functional studies, and not by means of sequence comparison alone. We identified 66 mRNA-translated products that may contribute to the G. fossarum reproductive process, on the basis of this functional annotation complemented by our biological expertise in crustacean physiology. These were grouped into families according to the biological processes in which they might be involved (supplemental Table S10).

A total of six contigs were markedly related to enzymes involved in the hormonal regulation of steroids and isoprenoid methyl farnesoate (MF) synthesis. Four were identified as having a high probability of being involved in MF metabolism. Two isoforms of farnesolic acid O-methyltransferase (FAMeT) were identified in our survey; one was exclusively detected in the male reproductive system, and the other was also detected in the cephalon. Based on the phylogenetic analysis by Hui et al. (36), both contigs are integrated among the crustacean FAMeT family (supplemental Data File S1), sustaining their involvement in MF biosynthesis. Additionally, two proteins that might be involved in sexual hormone catabolism were sequenced from testis. Although steroids are substrates for cytochromes P450 (CYP) (37), two CYP isoforms were identified in the reproductive tissues. One member of the CYP12A2 CYP family was identified in the female reproductive tissue, and one member of the CYP4C39 family was identified in male reproductive tissue. In the cuticle degradation process, closely linked to the molt cycle, which is intri-
cately synchronized with female reproduction in amphipods (5), seven contigs annotated as chitinase, an enzyme involved in the hydrolysis of chitin glycosidic bonds, were found in our G. fossarum protein catalog. Two isoforms were identified in the cephalon, and one isoform was observed in the female reproductive tissue. The four other contigs were identified in testis at different spermatogenesis stages. Among the seven contigs, five were related to the protein family analyzed previously by Huang et al. (38) (supplemental Data File S1), of which four belonged to protein groups unrelated to the molting process (chitinase 4 and 1) and one (I.D. 12415) belonged to the chitinase 2 group.

Remarkably, a large number of oogenesis-related proteins were detected via tandem mass spectrometry. Among these mRNA-translated products, 45 were lipoproteins (i.e., major yolk proteins or potential components of major yolk proteins), and five were receptors involved in the incorporation of the yolk proteins. Based on the prediction of function, the existence of 20 mRNA-translated products annotated as vitellogenin (VTG) was identified, 7 with close similarity to an existing sequence from G. fossarum (29). Some contigs indicated the same protein, leading to a total of 14 distinct proteins. Apart from these vitellogenins, a protein showing the most similarity to a decapod apolipocrustacein (the major egg yolk protein in decapods according to Avarre et al. (39)), was detected. The 23 other contigs were referred to by the generic term “clotting” in their predicted functions. Taking into account a fold change of at least 2.0 and a confidence threshold of 0.95 (p value < 0.05), 129 contigs were significantly more frequently detected in the male tissue than the female tissue (Figs. 5C and 5D).

Comparative Proteomics of Male and Female Reproductive Tissues—In order to quickly pinpoint new candidates involved in the reproductive process, a cross-comparison of protein abundances in male and female reproductive tissues was performed with a label-free shotgun strategy. Among the 1,219 mRNA-translated products detected in the reproductive tissues, 401 (33%) were shared by both males and females, as shown in Fig. 5A. Subsequently, for a detailed comparison of the most abundant proteins between both tissues, spectral-count-based quantitative proteomics was performed. A multiple-hypothesis-testing Student’s t test was applied on our MS/MS dataset by means of the Tfold module of the PatternLab software program (40). Supplemental Table S7 lists the calculated protein variations between male and female reproductive tissues and their p value confidence. Fig. 5B presents the distribution of the proteins according to their fold change and p value. Taking into account a fold change of at least 2.0 and a confidence threshold of 0.95 (p value < 0.05), 129 contigs were significantly more frequently detected in the male tissue than the female tissue, and 75 were significantly more frequently detected in the female than the male tissue (Figs. 5C and 5D).
Table I: The most differentially detected proteins in female and male reproductive tissues

| Contig I.D. | Probable function               | Fold change | p value   | Number of MS/MS spectra | Number of peptide sequences |
|------------|--------------------------------|-------------|-----------|--------------------------|-----------------------------|
| Female reproductive tissue | | | | | |
| 25127      | Clottable protein              | 197         | 0.4E-2   | 589                      | 8                           |
| 277        | Vitellogenin                   | 192         | 0.8E-2   | 574                      | 37                          |
| 19261      | No predicted function          | 175         | 3.8E-2   | 873                      | 20                          |
| 276        | Clottable protein              | 91          | 0.3E-2   | 271                      | 23                          |
| 194758     | Apolipocrustacein              | 91          | 0.7E-2   | 271                      | 29                          |
| Male reproductive tissue | | | | | |
| 2029       | Projectin                      | 34          | 3.7E-2   | 100                      | 28                          |
| 6444       | l-connectin                    | 20          | 1.1E-2   | 58                       | 11                          |
| 3157       | No predicted function          | 19          | 0.1E-2   | 55                       | 9                           |
| 22255      | Myosin                         | 19          | 2.1E-2   | 132                      | 5                           |
| 608        | No predicted function          | 18          | 2.7E-2   | 51                       | 8                           |

Among the contigs more frequently detected in the female reproductive tissue, 40% (31) were orphans. Thirty-three percent (25) belonged to the large lipid transfer protein (LLTP) superfamily, including 14 vitellogenins (11 proteins), one apolipocrustacein protein, and 10 clottable proteins (7 proteins). These were all classified as major yolk proteins in the previous section, but these mRNA-translated products proved to be female specific. Among other female-specific products, the previously mentioned cytochromes P450 and three heat-shock proteins (70 kDa and 21 kDa) were identified. We also observed proteins with housekeeping functions, such as energy production (two ATP synthases), amino acid metabolism (β-1-pyruvyl-5-carboxylate dehydrogenase), and cytoskeletal structure maintenance (tubulin).

Table I presents the characteristics of the five contigs with the strongest sexual dimorphism for each sex. In females, the contig 19261-encoded protein (frame 6) that exhibited the strongest sexual dimorphism (a fold change between female and male tissues of 175) was an orphan, and the others were part of the LLTP family.

In male reproductive tissue, most of the contigs (41%) could not be functionally annotated because of the low sequence similarity to already-annotated proteins. Remarkably, a set of 40 contigs (31%) were involved in various aspects of cellular shaping by cytoskeletal structure maintenance. Additional proteins were found to take part in energy metabolism (seven contigs), immunity (three contigs), amino acid metabolism (three contigs), and hydrolysis of glycosidic linkage (three contigs). One isoform of juvenile hormone epoxide hydrolase proved to be specific to the male reproductive tissue. The membrane protein flotillin-1 (I.D. 5059), which was included in the five contigs contributing to 1/10 of the reproductive proteome, was specific to this tissue. Of note, seven contigs were calcium dependent or involved in calcium homeostasis. Relative to the proteome of the female organ, fewer proteins (10) (8%) were orphans. Finally, among the contigs that were significantly more frequently detected in the testis (Table I), three were involved in cytoskeletal structure maintenance, and the two others had no predicted function.

Identification of Spermatogenesis-related Proteins by Means of Protein Dynamics—In order to indicate new proteins that could be directly related to spermatogenesis in G. fossarum, we analyzed the proteome dynamics of the male reproductive tissue. For this, we correlated the temporal patterns of proteins detected in testis sampled at seven different spermatogenesis stages: Day -1 (i.e. pre-copulatory mate guarding), Day 0, Day 1, Day 2, Day 3, Day 4, and Day 7 after copulation. Among the 914 proteins identified in this analysis, a total of 17 proteins presented interesting clustered temporal variations. Of these, nine proteins were devoid of functional annotation (53%), and three were more frequently detected in male reproductive tissue in the comparative analysis: one protein with no predicted function (I.D. 608), a calcium transporter (I.D. 4227), and a hemocytin-like protein (I.D. 461). Five distinct trends were delineated (Fig. 6). The characteristics of the corresponding proteins from each cluster are listed in supplemental Table S9. These different trends demonstrate the complexity of the dynamic variation patterns of the proteome during spermatogenesis. As shown in Fig. 6, trend A cluster, comprising three proteins, was defined by a general steady state until Day 3, then decreased expression at Day 4, followed by maximal expression at Day 7. Trend B cluster, also comprising three proteins, showed an expression maximum at Day 0 followed by a continuous decrease until Day 3 and then an increase until Day 7. Trend C cluster highlighted three proteins whose expression decreased until Day 1 and then underwent a continuous increase until Day 3. The three proteins from the trend D cluster exhibited increasing expression until Day 1 or Day 2 and then a decrease from Day 2 to Day 4. Trend E was the dominant pattern, with five proteins, including one male-specific protein with no predicted function (I.D. 608), with expression continuously decreasing until Day 2 and then increasing from Day 3 to Day 4. Both trend D and trend E are somehow correlated with spermatogenesis in G. fosarum.
**DISCUSSION**

Although crustacea are one of the largest and most species-diverse groups of animals, inhabiting all aquatic and terrestrial environments, little comprehensive molecular infor-

*Gammarus fossarum*, as the protein increase at Day 0 or over the two first days corresponds to the beginning of spermatogenesis. The maxima of the three other trends (at Day 3 or Day 4) correlate with the change in spermatozoon production rate.

**Fig. 6.** Clustered temporal abundance variations of proteins during spermatogenesis, before and after copulation. The different trends were obtained from clustering the spectral count using the TrendQuest module of the PatternLab program and are displayed in A–E. Cluster health was set at 0.9, and the minimum number of items per cluster was three.
Proteogenomics of Gammarus fossarum

Proteogenomics is available for this major arthropod group. Today, the only available crustacean genome is that of Daphnia magna, a Branchiopod Cladocera (1). According to pancrustacean phylogeny, branchiopoda are more closely related to hexapoda (insects) than to malacostraca (including decapods and amphipods) (41). Consequently, the large wealth of information obtained for the amphipod Gammarus, at both transcriptomic and proteomic levels, is unique and will allow further comparisons when other malacostraca are studied. Moreover, such molecular information is a key resource for understanding amphipod biology and enables the development of new biomarkers for in situ toxicological bioassays. Indeed, although daphnids, with their short generation time, are laboratory workhorses, their ecology (lentic ecosystem) and their small size hinder their application in the field in running water. In contrast, the ecology of G. fossarum makes it ideally suited for river biomonitoring. Furthermore, its size enables the direct targeting of specific organs, allowing the study of specific physiological functions, such as reproduction. We identified 218,574 contigs in the Gammarus transcriptome. We observed that ~40% of the contigs could be Gammarus-specific genes, as they presented poor sequence similarity to previously known genes. Thus, as shown with the cephalon proteome, a shotgun proteomic approach based on a dedicated RNA-Seq-derived database greatly improved MS/MS data interpretation relative to the use of current protein sequence databases from public resources. Here, the experimental difficulty of our approach is in the dissection of relatively large quantities of organs from small animals (~1 mm in size) to extract sufficient mRNA and protein material. However, the interpretation of data is relatively straightforward, as the proteins identified and quantified by means of the shotgun nano-LC-MS/MS approach give a relatively comprehensive view of the G. fossarum molecular players. We recently proposed a quantitative LC-MS/MS assay for the quantitation of vitellogenin as a biomarker of endocrine disruption (29). The extensive protein catalog described here will allow further development of toxicological biomarkers, with the aim of rapid evaluation of the impact of pollutants in freshwaters.

In amphipods, molting and oogenic cycles are closely linked. Although the new exoskeleton is still flexible after ecdisis, mature oocytes are spawned through the oviducts into the marsupium and are immediately fertilized by the male. The molting process is controlled by steroid molting hormones, called ecdysonoids. Ecdysonoid biosynthetic reactions are catalyzed by P450 mono-oxygenases and associated reductases (43). In this proteogenomic survey, a cytochrome P450 mono-oxygenase isoform (CYP450, I.D. 100255) was identified in the male reproductive tissue, and another one (I.D. 122081) was shown to be tissue-specific for the female organ. On the basis of their organ localization, the two identified proteins might be involved in ecdysteroid metabolism. In addition to ecdysteroid action, the hormonal system of arthropods (insects and crustacea) relies on sesquiterpene molecules such as farnesoid acid, MF, and the juvenile hormone (JH). In crustacea, the end product of the sesquiterpenoid pathways seems to be MF, formed after the methylation of farnesoid acid through the action of FAMEt (36). Here two FAMEt isoforms were identified, one exclusively detected in the male reproductive system and the other also detected in the cephalon. The functions of MF in crustacea are, to date, elusive, but this hormone can be considered as a determinant of male sex, inducing the production of male offspring in daphnids (44). It also stimulates gonadal maturation; MF has been proven to stimulate testicular development in the crab Carcinus maenas (45). Moreover, in the shrimp Litopenaeus vannamei, RNA interference of putative FAMEt suggests that this enzyme is essential for the molt process and regulates molt-related genes (46). Although the mechanism of biosynthesis of MF is well established, little is known about its catabolism. In insects, JH inactivation is processed by JH carboxysterase and epoxide hydrolase. Here, homologs of both enzymes were identified in male reproductive tissue, but the presence of JH in G. fossarum remains to be established experimentally. For pest control in agriculture, the JH/MF hormonal machinery is targeted directly by molecular analogues of this hormone. Consequently, these act as endocrine-disrupting chemicals for non-target organisms such as G. fossarum. The identification of proteins involved in MF metabolism will allow the further development of endocrine disruption biomarkers in crustacea.

Regarding morphogenesis, the major constituent of the arthropod exoskeleton is chitin, a linear polymer of N-acetyl-D-glucosamine. During the premolt stage of crustacea and insects, the organisms secrete enzymes for chitin degradation in order to recycle the old exoskeleton while synthesizing a new one. In crustacea, measurement of the activity of chitinases, as a determinant of male sex, inducing the production of male offspring in daphnids (44). It also stimulates gonadal maturation; MF has been proven to stimulate testicular development in the crab Carcinus maenas (45). Moreover, in the shrimp Litopenaeus vannamei, RNA interference of putative FAMEt suggests that this enzyme is essential for the molt process and regulates molt-related genes (46). Although the mechanism of biosynthesis of MF is well established, little is known about its catabolism. In insects, JH inactivation is processed by JH carboxysterase and epoxide hydrolase. Here, homologs of both enzymes were identified in male reproductive tissue, but the presence of JH in G. fossarum remains to be established experimentally. For pest control in agriculture, the JH/MF hormonal machinery is targeted directly by molecular analogues of this hormone. Consequently, these act as endocrine-disrupting chemicals for non-target organisms such as G. fossarum. The identification of proteins involved in MF metabolism will allow the further development of endocrine disruption biomarkers in crustacea.

Among the most abundant proteins involved in the female reproductive process identified in our study, most were members of the LLTP superfamily. These proteins are structural scaffolds for the assembly of lipoproteins, as they bind a multitude of lipid species. Apolipoproteins, clottable proteins, and VTG are members of the LLTP superfamily. Generally, VTG is considered the major yolk protein in most oviparous organisms (39). In the present study, we identified a total of 45...
Several proteins that were functionally annotated in our study, mainly on the basis of sequence similarities to decapod protein homologs, for which functional annotation is confusing because of neofunctionalization of orthologous VTG gene to clotting function (39). Among all these candidates potentially involved in the female reproductive process, only 25, pinpointing the existence of 18 distinct proteins, were found to be preferentially expressed in female reproductive tissue via comparative proteomics. Evidence of multiple copies of vitellogenin genes had already been obtained through targeted PCR amplification: two copies in D. magna (48) and four copies in the mosquito Culex tarsalis (49). Our study, the first ever performed for an arthropod at such a large scale at the protein level, revealed unsuspected diversity of LLTP members.

Most of the female-specific proteins were orphans (40%), including the most abundant one (I.D. 19261), with no putative conserved domains detected. This suggests that the major yolk protein might not be structurally related to LLTPs in gammarids. This hypothesis is reinforced by molecular data obtained from other species. The major yolk protein is a transferrin-like, iron-binding protein in sea urchins (50), and it is a lipase-related protein in higher diptera (51). The existence of many gammarid-specific proteins, such as the large variety of LLTPs revealed by our analysis, might suggest that major yolk proteins are diverse and possibly genus specific, with this diversity being explained by the variety of carried lipid species. It is noteworthy that the female organisms used in our experiment were in the last phase of their reproductive cycle, but the abundances of the different proteins detected here should change during the course of the cycle. Given the essential role of yolk proteins in ensuring the development of a new generation, these results, together with our knowledge of the G. fossarum ovarian cycle and embryonic development (5), generate multiple research hypotheses, and further experiments will be required in order to ascertain the proteins’ functions.

Regarding proteins involved in male reproductive function, we also employed an additional strategy for uncovering molecular players involved in the spermatogenesis process. The first protein catalog from testis tissue was carried out on a specific spermatogenesis stage, but proteome dynamics of this tissue was studied at seven different stages, leading to the identification of 437 additional proteins. Proteome dynamics during spermatogenesis revealed five distinct temporal trends, suggesting that spermatogenesis is a highly complex temporal event. The majority of the proteins delineated in these trends were devoid of functional prediction. Indeed, the spermatozoon cell is the most diverse cell type in the animal kingdom, and few investigations have been carried out on genes that regulate spermatogenesis in crustacea. Thus, these proteins are interesting candidates for potential involvement, directly or indirectly, in male reproductive function. Several proteins that were functionally annotated in our study are crucial for the integrity of reproductive function. For example, flotillins, members of a ubiquitous family of membrane proteins, are involved in numerous cellular processes and are also essential for spermatozoon functionality. In spermatozoa, the acrosome is a cap-like structure that enables fusion with the egg, and proteins from the flotillin family have been shown to be involved in its biogenesis and functionality (52). In our study, flotillin-1 (I.D. 5059) proved to be abundant, as it was included in the five most detected proteins from the reproductive proteome, and was classified as male specific. Among acrosomal enzymes, glycohydrolases are believed to digest the extracellular glycoalyx surrounding the egg (53). The three glycohydrolases classified as male specific in this study are potentially acrosomal enzymes involved in sperm–egg fusion.

In order to ensure sperm viability, a “programmed degradation” of spermatozoa before they become obsolete is carried out by the immune system. In the male shrimp L. vannamei, progressive melanization of spermatozoa has been documented for elimination through an innate immune response (54). Melanization is initiated by the activation of prophenoloxidase, a protein that was found to be male specific. Two other male-specific proteins involved in innate immunity were identified: a hemocyanin-like protein and a transglutaminase (55). During our spermatogenesis kinetics experiment we also detected several isoforms of epididymal sperm-binding 1, which has been proposed as an indicator of sperm defectiveness (56).

In our proteogenomic survey, we identified a strong prevalence in the male reproductive tissue of calcium-dependent proteins and proteins involved in calcium homeostasis. Indeed, regulation of internal Ca^{2+} concentration is believed to be crucial for sperm maturation, capacitation, and sperm–egg interaction and is consequently subject to strict spatiotemporal control. One sarcoplasmic calcium-binding protein and three isoforms of sarcoplasmic/endoplasmic reticulum calcium ATPase were detected, with one isoform (I.D. 4227) that showed dynamic changes during spermatogenesis. These proteins, previously identified in the acrosome from vertebrates, are believed to sequester Ca^{2+} within this organelle, supporting the hypothesis that the acrosome acts as an intracellular Ca^{2+} storage compartment (57). In addition, three proteins identified as calcium dependent were classified as male specific: transglutaminase and two isoforms of copine-8. The functions of copine-8 are elusive, but in humans the gene is predominantly expressed in prostate and testis, suggesting an important role in the development and regulation of male sexual characteristics (58).

In conclusion, we have generated a proteogenomic dataset for the amphipod G. fossarum using Illumina-Solexa pyrosequencing and shotgun proteomics that paves the way for several studies related to this ecotoxicological sentinel organism. Here, we focused our analysis on the reproductive process in order to propose new biomarkers related to its
Proteogenomics of *Gammarus fossarum*

disturbance. To our knowledge, this represents the first extra-large proteogenomic resource for a crustacean. At this stage, we were able to certify via tandem mass spectrometry the presence of 1,873 proteins in *G. fossarum*. Further analyses are needed to go deeper into this catalog. Although the results of functional annotation by sequence similarities were quite limited because of the absence of other datasets for this lineage, we were able to highlight interesting candidates involved in various processes related to reproduction, such as hormonal metabolism, morphogenesis, sperm viability, and calcium homeostasis. We also found a high diversity of vitellogenin-like proteins in the female organ. However, confidence in these functional annotations is low, as for any new model organism. For this reason, we performed a comparative analysis of the abundance of proteins in male and female reproductive tissues and quickly pinpointed specific candidates involved in the reproductive process. Interestingly, we uncovered a substantial proportion of orphan proteins, with a significantly higher ratio in reproductive tissue than in the cephalon. Reproductive proteins are known to evolve more rapidly because of divergent natural selection leading to molecular reproductive isolation in closely related species (42). Finally, several orphans were found in abundance in male and female reproductive tissues that were classified as sex specific and thus are potentially involved, directly or indirectly, in reproductive function. In order to better understand their functions, and the respective molecular mechanisms, new investigations should now be performed with the protein sequence knowledge detailed in the present study as a starting point.

Acknowledgments—We thank K. Abbaci, G. Le Goff, G. Miotello, G. Steinmetz, H. Quéau, and the PRIDE Team for their technical assistance, as well as P. C. Carvalho (COPPE, Rio de Janeiro, Brazil) for adapting the export option of the PatternLab program.

* J. Trapp was supported by a fellowship from the Institut National de Recherche en Sciences et Technologies pour l’Environnement et l’Agriculture (France) and from the Program Transversal de Toxicologie of the Commissariat de l’Energie Atomique et aux Energies Alternatives (France). We thank the Institut National de Recherche en Sciences et Technologies pour l’Environnement et l’Agriculture (France) and the Commissariat de l’Energie Atomique et aux Energies Alternatives (France) for financial support.

This article contains supplemental material.

1. To whom correspondence should be addressed: Jean Armen-gaud, Laboratoire de Biochimie des Systèmes Perturbés, CEA Marc-coule, DSV, IEBEB, SBTM, LBSP, F-30207 BAGNOLS-SUR-CEZE, France, Tel.: 00-33-4-66-79-68-02, Fax: 00-33-4-66-79-19-05, E-mail: jean.armengaud@cea.fr; Olivier Geffard, Laboratoire d’écotoxicologie, Irstea Villeurbanne, Unité de Recherche MALY, CS70077, F-69626 VILLEURBANNE, France, Tel.: 00-33-4-72-20-87-58, E-mail: olivier.geffard@irstea.fr.

REFERENCES

1. Colbourne, J. K., Pfenner, M. E., Gilbert, D., Thomas, W. K., Tucker, A., Oakley, T. H., Tokishita, S., Aerts, A., Arnold, G. J., Basu, M. K., Bauer, D. J., Caceres, C. E., Carmel, L., Casola, C., Choi, J. H., Deter, J. C., Dong, Q., Dusheyko, S., Eads, B. D., Frohlich, T., Geiler-Samerotte, K. A., Gerlach, D., Hatcher, P., Jogdeo, S., Kriegsvid, J., Kriventseva, E. V., Kutz, D., Laforsch, C., Lindquist, E., Lopez, J., Manak, J. R., Muller, J., Pangilinan, J., Patwardhan, R. P., Pittuck, S., Pritham, E. J., Rechsteiner, A., Rho, M., Rogozin, I. B., Sakarya, O., Salamov, A., Schaack, S., Shapiro, H., Shiga, Y., Skalitzky, C., Smith, Z., Souvorov, A., Sung, W., Tang, Z., Tsuchiya, D., Tu, H., Vos, H., Wang, M., Wolf, Y. I., Yamagata, H., Yamada, T., Ye, Y., Shaw, J. R., Andrews, J., Crease, T. J., Tang, H., Lu, S., Robertson, H. M., Bork, P., Koonin, E. V., Zdobnov, E. M., Grigoriev, I. V., Lynch, M., and Boore, J. L. (2011) The ecoregressive genome of Daphnia pulex. Science 331, 555–561.

2. Kunz, P. Y., Kienle, C., and Gerhardt, A. (2010) Gammarus spp. in aquatic ecotoxicology and water quality assessment: toward integrated multi-level tests. Rev. Environ. Contam. Toxicol. 205, 1–76.

3. Dangles, O., and Guerold, F. (2001) Linking shredders and leaf litter processes: insights from an acidic stream study. Int. Rev. Hydrobiol. 86, 395–406.

4. Xuereb, B., Chaumont, A., Mons, R., Garric, J., and Geffard, O. (2009) Acetylcholinesterase activity in Gammarus fossarum (Crustacea Amphipoda) intrinsic variability, reference levels, and a reliable tool for field surveys. Aquat. Toxicol. 93, 225–233.

5. Geffard, O., Xuereb, B., Chaumont, A., Geffard, A., Biagianti, S., Noel, C., Abbaci, K., Garric, J., Chaumontier, G., and Charmantier-Daures, M. (2010) Ovarian cycle and embryonic development in Gammarus fossarum: application for reproductive toxicity assessment. Environ. Toxicol. Chem. 29, 2249–2259.

6. Jungmann, D., Ladewig, V., Ludwichowski, K. U., Petzsch, P., and Nagel, R. (2004) Intersexuality in Gammarus fossarum—A common inducible phenomenon? Arch. Hydrobiol. 159, 511–529.

7. Lessio, C. (2007) Crustacean endocrine toxicology: a review. Ecotoxicology 16, 61–81.

8. Kulkarni, D., Gergs, A., Hommen, U., Ratte, H. T., and Preuss, T. G. (2013) A plea for the use of copepods in freshwater ecotoxicology. Environ. Sci. Pollut. Res. Int. 20, 75–85.

9. Ansong, C., Purvine, S. O., Adkins, J. N., Lipton, M. S., and Smith, R. D. (2008) Proteogenomics: needs and roles to be filled by proteomics in genome annotation. Brief Funct. Genomic Proteomic 7, 50–62.

10. Renuse, S., Chaerkady, R., and Pandey, A. (2011) Proteomics, Proteomics 11, 620–630.

11. Armengaud, J., Hartmann, E. M., and Bland, C. (2013) Proteogenomics for environmental microbiology. Proteomics 13, 2731–2742.

12. Yates, J. R., 3rd, Eng, J. K., and McCormack, A. L. (1995) Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. Anal. Chem. 67, 3202–3210.

13. Gallien, S., Perroudou, C., Deshayes, C., Reyart, J. M., Van Dorselaer, A., Poch, O., Schaeffer, C., and Lecompte, O. (2009) Ortho-proteogenomics: multiple proteomes investigation through orthology and a new MS-based protocol. Genome Res. 19, 128–135.

14. Venter, E., Smith, R. D., and Payne, S. H. (2011) Proteogenomic analysis of bacteria and archaea: a 46 organism case study. PLoS One 6, e22657.

15. Castellana, N. E., Payne, S. H., Shen, Z., Stanke, M., Bafna, V., and Briggs, S. P. (2008) Discovery and revision of Arabidopsis genes by proteogenomics. Proc. Natl. Acad. Sci. U.S.A. 105, 21034–21038.

16. Castellana, N. E., Shen, Z., He, Y., Walley, J. W., Cassidy, C. J., Briggs, S. P., and Bafna, V. (2014) An automated proteogenomic method uses mass spectrometry to reveal novel genes in Zea mays. Mol. Cell. Proteomics 13, 157–167.

17. Rubiano-Labrador, C., Bland, C., Miottello, G., Guerin, P., Pible, O., Baena, S., and Armengaud, J. (2014) Proteogenomic insights into salt tolerance by a halotolerant alpha-proteobacterium isolated from an Andean saline spring. J. Proteomics 97, 36–47.

18. Wilmes, P., Andersson, A. F., Lefsrud, M. G., Wexler, M., Shah, M., Zhang, B., Hettich, R. L., Bond, P. L., VerBerkmoes, N. C., and Banfield, J. F. (2008) Community proteogenomics highlights microbial strain-variant protein expression within activated sludge performing enhanced biological phosphorus removal. ISME J. 2, 853–864.

19. Christie-Oleza, J. A., Miottello, G., and Armengaud, J. (2013) Proteogenomic definition of biomarkers for the large Roseobacter clade and application for a quick screening of new environmental isolates. J. Proteome Res. 12, 5331–5339.

20. Armengaud, J., Trapp, J., Pible, O., Geffard, O., Chaumont, A., and Hartmann, E. M. (2014) Non-model organisms, a species endangered by...
proteogenomics. J. Proteomics 105, 5–18
21. Ning, K., and Nesvizhskii, A. I. (2010) The utility of mass spectrometry-based proteomic data for validation of novel alternative splice forms reconstructed from RNA-Seq data: a preliminary assessment. BMC Bioinformatics 11 Suppl 11, S14
22. Nagaraj, N., Wilsinowski, J. R., Geiger, T., Cox, J., Kircher, M., Kelso, J., Paabo, S., and Mann, M. (2011) Deep proteome and transcriptome mapping of a human cancer cell line. Mol. Syst. Biol. 7, 548
23. Wang, X., Sieblos, R. J. C., Wang, D., Halvey, P. J., Tabb, D. L., Liebler, D. C., and Zhang, B. (2012) Protein identification using customized protein sequence databases derived from RNA-seq data. J. Proteome Res. 11, 1009–1017
24. Woo, S., Cha, S. W., Merrihew, G., He, Y., Castellana, N., Guest, C., MacCoss, M., and Sfarna, V. (2014) Proteogenomic database construction driven from large scale RNA-seq data. J. Proteome Res. 13, 21–28
25. Blythe, M. J., Malia, S., Everall, R., Shih, Y. H., Lemay, V., Moreton, J., Wilson, R., and Aboobaker, A. A. (2012) High through-put screening of the Parthyle hawaiience mRNAs and microRNAs to aid comparative developmental studies. PLoS One 7, e33784
26. Zeng, V., Villanueva, K. E., Ewen-Campen, B. S., Alves, F., Browne, W. E., and Ektavour, C. G. (2011) De novo assembly and characterization of a maternal and developmental transcriptome for the emerging model crustacean Parthyle hawaiience. BMC Genomics 12, 581
27. Hook, S. E., Osborn, H. L., Spadaro, D. A., and Simpson, S. L. (2014) Assessing mechanisms of toxicant response in the amphipod Melita plumulosa through transcriptomic profiling. Aquat. Toxicol. 146, 247–257
28. Lacaze, E., Devaux, A., Jubeaux, G., Mons, R., Gardette, M., Bony, S., Garric, J., and Geffard, O. (2011) DNA damage in Gammarus fossarum sperm as a biomarker of genotoxic pressure: intrinsic variability and reference level. Sci. Total Environ. 409, 3230–3236
29. Simon, R., Jubeaux, G., Chaumont, A., Lemoine, J., Geffard, O., and Salavador, A. (2010) Mass spectrometry assay as an alternative to the enzyme-linked immunosorbent assay test for biomarker quantitation in ecdysteroidology: application to vitellogenin in Crustacea (Gammarus fossarum). J. Chromatogr. A 1217, 5109–5115
30. Christie-Oleza, J. A., Miotello, G., and Armengaud, J. (2012) High-throughput proteogenomics of Ruergea pomeroyi: seeding a better genomic annotation for the whole marine Roseostracae clade. BMC Genomics 13, 73
31. Christie-Oleza, J. A., Pina-Villalonga, J. M., Bosch, R., Nogales, B., and Armengaud, J. (2012) Comparative proteogenomics of twelve Roseostracae exoproteomes reveals different adaptive strategies among these marine bacteria. Mol. Cell. Proteomics 11, M111.013110
32. Vizzcaino, J. A., Cote, R. G., Csdoros, A., Dianes, J. A., Fabregat, A., Foster, J. M., Gris, J., Alpi, E., Birim, M., Contell, J., O’Kelly, G., Schoenenger, A., Ovelleiro, D., Perez-Riverol, Y., Reisinger, F., Rios, D., Wang, R., and Hermjakob, H. (2010) The PRoteomics IDENTifications (PRIDE) database associated tools: status in 2013. Nucleic Acids Res. 41, D1053–D1069
33. Weiss, M., Macher, J. N., Seefeldt, M. A., and Taberlet, P. (1997) Taxonomic status and evolutionary relationships of some species of the genus Gammarus (Crustacea, Amphipoda) deduced from mitochondrial DNA sequences. Mol. Phylogenet. Evol. 17, 251–267
34. Jones, M. O., and Boyle, S. M. (1998) Cytochromes P450 in crustaceans. Comp. Biochem. Physiol. C Pharmacol. Toxicol. Endocrinol. 121, 157–172
35. Hwang, Q. S., Yan, J. H., Tang, J. Y., Yao, Y. M., Xie, X. L., Wang, Y., Wei, X. O., Yan, Q. H., and Chen, Q. X. (2010) Cloning and tissue expressions of seven chitinase family genes in Litopenaeus vannamei. Fish Shellfish Immunol. 29, 75–81
36. Avarre, J. C., Lubzens, E., and Babin, P. J. (2007) Apolipoprotecinase, formerly vitellogenin, is the major egg yolk precursor protein in decapod crustaceans and is homologous to insect apolipopophin II/1 and vertebrate apolipoprotein B. BMC Evol. Biol. 7, 3
37. Carvalho, P. C., Fischer, J. S., Chen, E. I., Yates, J. R., 3rd, and Barbosa, V., C. (2008) PatternLab for proteomics: a tool for differential shotgun proteomics. BMC Bioinformatics 9, 316
38. Armengaud, J. (2012) Comparative proteomics uncovers the signature of natural selection acting on the ejaculatory proteins of two cricket species isolated by postmating, prezygotic phenotypes. Mol. Biol. Evol. 28, 423–435
39. Mykles, D. L. (2011) Ecdysosterone metabolism in crustaceans. J. Steroid Biochem. Mol. Biol. 127, 196–203
40. Lampt, W., Lampt, K. P., and Larsson, P. (2012) Induction of male production in clones of Daphnia pulex by the juvenile hormone methyl farnesolate under short photoperiod. Comp. Biochem. Physiol. C Pharmacol. Toxicol. 156, 130–133
41. Nagaraju, G. P., and Borst, D. W. (2008) Methyl farnesolate couples environmental changes to testicular development in the crab Carcinus maenas. J. Exp. Biol. 211, 2773–2778
42. Hui, J. H., Toпе, S. S., and Chan, S. M. (2008) Characterization of the putative farnesolic acid O-methyltransferase (LVFAMet) cDNA from white shrimp, Litopenaeus vannamei: evidence for its role in molting. Peptides 29, 229–236
43. Noguez, S., Quilhermino, L., and Guimarães, L. (2011) Biochemical and locomotor responses of Carcinus maenas exposed to the serotonin reuptake inhibitor fluoxetine. Chemosphere 85, 967–976
44. Tokishita, S., Kato, Y., Kobayashi, T., Nakamura, S., Ohta, T., and Yamagata, H. (2008) Organization and repression by juvenile hormone of a vitellogenin gene cluster in the crustacean, Daphnia magna. Biochem. Biophys. Res. Commun. 365, 362–370
45. Chen, S., Armstead, J. S., Provost-Javier, K. N., Sakamoto, J. M., and Rasgon, J. L. (2010) Duplication, concerted evolution and purifying selection drive the evolution of mosquito vitellogenin genes. BMC Evol. Biol. 10, 142
46. Brooks, J. M., and Wessel, G. M. (2002) The major yolk protein in sea urchins is a transferrin-like, iron binding protein. Dev. Biol. 245, 1–12
47. Sappington, T. W. (2002) The major yolk proteins of higher Diptera are homologs of a class of minor yolk proteins in lepidoptera. J. Mol. Evol. 55, 470–475
48. Wu, Y., Chen, X., Wang, S., Jiang, M., Zheng, B., Zhou, Q., Bi, Y., Zhou, Z., Huang, X., and Sha, J. (2012) Flotillin-2 is an acrosome-related protein involved in mouse spermiogenesis. J. Biomed. Res. 26, 278–287
49. Tulsiani, D. R., Abou-Hall, A., Loeser, C. R., and Pereira, B. M. (1998) The biological and functional significance of the sperm acrosome and acrosomal enzymes in mammalian fertilization. Exp. Cell Res. 240, 151–164
50. Parson, S., Raviv, S., Shechter, A., and Sagi, A. (2006) Males also have their time of the month! Cyclic disposal of old spermatophores, timed by the molt cycle, in a marine shrimp. J. Exp. Biol. 209, 4974–4983
51. Vazquez, L., Alpuache, J., Maldonado, G., Agundis, C., Pereyra-Morales, A., and Zenteno, E. (2009) Review: immunity mechanisms in crustaceans. Innate Immun. 15, 179–186
52. D’Amours, O., Frenette, E., Bourdeau, L. J., Allard, N., Leclerc, P., Blondin, P., and Sullivan, R. (2012) Epidymosymes transfer epidymal sperm binding protein 1 (ELSPBP1) to dead spermatozoa during epididymal transit in bovine. Biol. Reprod. 87, 94
53. Jimenez-Gonzalez, C., Michelangeli, F., Harper, C. V., Barratt, C. L., and Publicover, S. J. (2006) Calcium signalling in human spermatozoa: a specialized “toolbox” of channels, transporters and stores. Hum. Reprod. Update 12, 253–267
54. Malta, R., Grigorov, D. N., Bera, T. K., Pastan, I. H., and Lee, B. (2003) Cloning, molecular characterization, and expression analysis of copine 8. Biochem. Biophys. Res. Commun. 303, 842–847