Proteogenomic reconstruction of organ-specific metabolic networks in an environmental sentinel species, the amphipod Gammarus fossarum

Natacha Koenig, Patrice Baa-Puyoulet, Amélie Lafont, Isis Lorenzo-Colina, Vincent Navratil, Maxime Leprêtre, Kevin Sugier, Nicolas Delorme, Laura Garnero, Herve Queau, et al.

Proteogenomic reconstruction of organ-specific metabolic networks in an environmental sentinel species, the amphipod *Gammarus fossarum*

Natacha Koenig[1], Patrice Baa-Puyoulet[2], Amélie Lafont[1], Isis Lorenzo-Colina[1], Vincent Navratil[3], Maxime Leprêtre[1], Kevin Sugier[1], Nicolas Delorme[1], Laura Garnero[1], Hervé Queau[1], Jean-Charles Gaillard[4], Mélodie Kielbasa[4], Sophie Ayciriex[5], Federica Calevro[2], Arnaud Chaumot[1], Hubert Charles[2], Jean Armengaud[4], Olivier Geffard[1], Davide Degli Esposti[1#]

1 INRAE, UR RiverLy, Ecotoxicology Team. Centre de Lyon-Grenoble Auvergne Rhône Alpes, 5 rue de la Doua CS 20244, 69625 Villeurbanne, France

2 INRAE, INSA Lyon, BF2I, UMR203, 69621 Villeurbanne, France

3 PRABI, Rhône-Alpes Bioinformatics Center, Université Lyon 1, Villeurbanne, France, UMS 3601, Institut Français de Bioinformatique, IFB-Core, Évry, France.

4 Université Paris-Saclay, Département Médicaments et Technologies pour la Santé (DMTS), CEA, INRAE, SPI-Li2D, F-30207 Bagnols-sur-Céze, France

5 University of Lyon, CNRS, Institut des Sciences Analytiques, UMR 5280, 5 rue de la Doua, F-69100 Villeurbanne, France

[#]Corresponding Author: Davide Degli Esposti, INRAE, UR RiverLy, Ecotoxicology Team. Centre de Lyon-Grenoble Auvergne Rhône Alpes, 5 rue de la Doua CS 20244, 69625 Villeurbanne, France.

Tel : +33 4 72 20 87 13

e-mail: davide.degli-esposti@inrae.fr

Running title: Organ-specific metabolic pathways in *Gammarus fossarum*

Keywords: amphipods, transcriptomics, proteomics, *Gammarus fossarum*, multi-omics integration, shotgun proteomics, metabolic network.

Abstract

Metabolic pathways are targets of environmental contaminants underlying a large variability of toxic effects throughout biodiversity. However, the systematic reconstruction of metabolic pathways remains limited in environmental sentinel species due to the lack of available genomic data in many taxa of animal diversity. In order to improve the knowledge of the metabolism of sentinel species, in this study we used a multi-omics approach to reconstruct the most comprehensive map of metabolic pathways for a crustacean model in biomonitoring, the amphipod *Gammarus fossarum*.

We revisited the assembly of RNA-seq data by *de novo* approaches drastically reducing RNA contaminants and transcript redundancy. We also acquired extensive mass spectrometry shotgun proteomic data on several organs from *G. fossarum* males and females to identify organ-specific metabolic profiles.

The *G. fossarum* metabolic pathway reconstruction (available through the metabolic database GamfoCyc) was performed by adapting the genomic tool CycADS and we identified 377 pathways representing 7,630 annotated enzymes, 2,610 enzymatic reactions and the expression of 858 enzymes was experimentally validated by proteomics. Our analysis shows organ-specific metabolic profiles, such as an elevated abundance in enzymes involved in ATP biosynthesis and fatty acid beta-oxidation indicative of the high-energy requirement of the gills, or the key anabolic and detoxification role of the hepatopancreatic caeca, as exemplified by the specific expression of the retinoid biosynthetic pathways and glutathione synthesis.

In conclusion, the multi-omics data integration performed in this study provides new resources to investigate metabolic processes in crustacean amphipods and their role in mediating the effects of environmental contaminant exposures in sentinel species.

Abbreviations

BLAST	Basic Local Alignment Search Tool
BUSCO	Benchmarking Universal Single-Copy Orthologues
CycADS	Cyc Annotation Database System
DE	Differentially expressed
EC	Enzyme Commission
FC	Fold Change
FDR	False Discovery Rate
GFF3	General Feature Format 3
G. fossarum	*Gammarus fossarum*
GFBF / GFBM	*Gammarus fossarum* female / male
GO	Gene Ontology
IsoPct	Isoform Percentage
KAAS	KEGG Automatic Annotation Server
KEGG	Kyoto Encyclopedia of Genes and Genomes
KO	KEGG Orthology
MDS	Multi-Dimensional Scaling
MS	Mass Spectrometry
NGS	Next Generation Sequencing
ORF	Open Reading Frame
PGDB	Pathway/Genome Database
ER	Enrichment Ratio
RNA-Seq	RNA Sequencing
RSEM	RNA-Seq by Expectation Maximisation
RT	Retention Time
SBML	Systems Biology Markup Language
SC	Spectral Count

Introduction

Metabolic pathways are potential targets of environmental contaminants, both in humans and in other species (Fritsche et al., 2023; Jordão et al., 2015; Massart et al., 2022). Recent studies showed that some chemicals, such as tributyltin, juvenoid hormones, or bisphenol A, can target and act as endocrine disruptors on lipid metabolism in the model species *Daphnia magna* (Fuertes et al., 2019; Jordão et al., 2016a, 2016b, 2015). These compounds have been shown to affect lipid (e.g. triacylglycerols and cholesterol) distribution, storage, or biosynthesis. Transcriptomic studies have also shown that changes following lipid accumulation include up-regulation of genes involved in fatty acid, glycerophospholipid and glycerolipid metabolism, membrane constituents, and chitin and cuticle biosynthesis pathways (Fuertes et al., 2019).

Although the use of a handful of model species plays a key role in discovering and acquiring knowledge on the molecular mechanisms underlying invertebrate biology and physiology, some obstacles, such as the phylogenetic distance with the species present in the diverse environments to be preserved, hampers their use in assessing the impact of environmental contaminants. Moreover, to improve environmental risk assessment, it is important to consider the biological diversity of a greater number of test species (Ruivo et al., 2022; Santos et al., 2018). In particular, metabolic networks (e.g. hormone synthesis pathways) show strong heterogeneity across phyla (Markov et al., 2009). Thus, strengthening the knowledge of the molecular physiology of sentinel species is an essential step to improve the extrapolation of toxicological effects of environmental contaminants across species. Unfortunately, there is a dramatic lack of genomic data, particularly in aquatic invertebrates used as sentinel species. To overcome this problem, the integration of mRNA sequencing with high-resolution mass spectrometry proteomics, an approach known as proteogenomics, has allowed the establishment of catalogs of thousands of proteins in a few of them, i.e. *Gammarus fossarum* or *Dreissena polymorpha* (Gouveia et al., 2018; Leprêtre et al., 2019; Trapp et al., 2014). Transcriptomic resources for gammarids typically include the pooling of RNAs coming from various individuals (Caputo et al., 2020; Trapp et al., 2014), but recent transcriptomes obtained from different gammarid species were obtained from genotyped single individuals, reducing the risk of chimeric transcript reconstruction (Cogne et al., 2019; Neuparth et al., 2020). In particular, our group contributed to those studies providing individual

transcriptomes from individual male and female gammarids that were assembled separately to maximize mass spectrometry data extraction (Cogne et al., 2019). In this context, it would be of great interest to perform a *de novo* assembly using the RNA-seq data from both male and females organisms in order to improve gene coverage, to have a unique gene set for both sexes and to investigate potential contamination of the organism's RNA from RNAs issued by its microbiome or by potential manipulation bias due to organisms' size (a few centimeters). Similarly, metabolomics and lipidomics approaches in sentinel species such as the crustaceans *Gammarus. fossarum* or *Palaemon serratus* are showing species-specific lipid and metabolite compositions (Fu et al., 2021; Marie et al., 2023). Despite improvements in proteome and transcriptome functional annotations in the last decade, it has been noted that large discrepancies may occur in ontology assignment, especially in certain arthropod orders (McCartney et al., 2023), due to a lack of standardized and systematic bioinformatics methods.

Besides, most molecular studies in aquatic invertebrates used in ecotoxicology are carried out on whole-body sentinel organisms or pools of organisms (Besse et al., 2013; Kunz et al., 2010). This approach may help the identification of toxicity biomarkers (Leprêtre et al., 2022), but, as we have shown, the integration of organ-specific –omics profiles in emergent animal models is very relevant to improve our understanding of the molecular mechanisms (Degli Esposti et al., 2019; Koenig et al., 2021). For instance, the use of coexpression network analyses showed species-specific and orphan proteins involved in gonad maturation and embryonic development (Degli Esposti et al., 2019), or the mechanisms of toxicity of reproductive contaminants (Koenig et al., 2021) in *G. fossarum*. Organ gene expression analysis of newly identified metallothioneins transcripts showed a specific interaction between hepatopancreatic caeca and gills in the detoxification of heavy metals, such as Cadmium (Degli Esposti et al., 2024).

In this context, the present work aimed to provide a complete mapping of the metabolism of *G. fossarum* by adapting the Cyc Annotation Database System (CycADS), particularly suitable for metabolic gene annotation and network reconstruction using genomic data (Vellozo et al., 2011) to the use of transcriptomic data in the freshwater amphipod *G. fossarum.* Moreover, we aimed to integrate new and available proteomic datasets (Leprêtre et al., 2023) to identify organ-specific metabolic profiles

and provide the basis to improve the knowledge of the effects of aquatic contaminants on metabolic pathways of this species and other closely related amphipods.

Materials and methods

Experimental design

To perform a first metabolic reconstruction based on the transcriptomic data available for *G. fossarum* (Cogne et al., 2019), we performed a new *de novo* assembly combining male and female RNA-seq raw data to reduce transcript redundancy and allow the implementation of CycADS. Subsequently, to integrate mRNA and protein expression levels, proteomic data from male and female organs of *G. fossarum* were obtained by label-free high-resolution mass spectrometry. Transcriptomic and proteomic data were integrated by cross-referencing the respective EC numbers to validate the expression of the enzymes identified and the proteomic dataset was further analyzed to investigate organ-specific metabolic pathway expression (Figure 1).

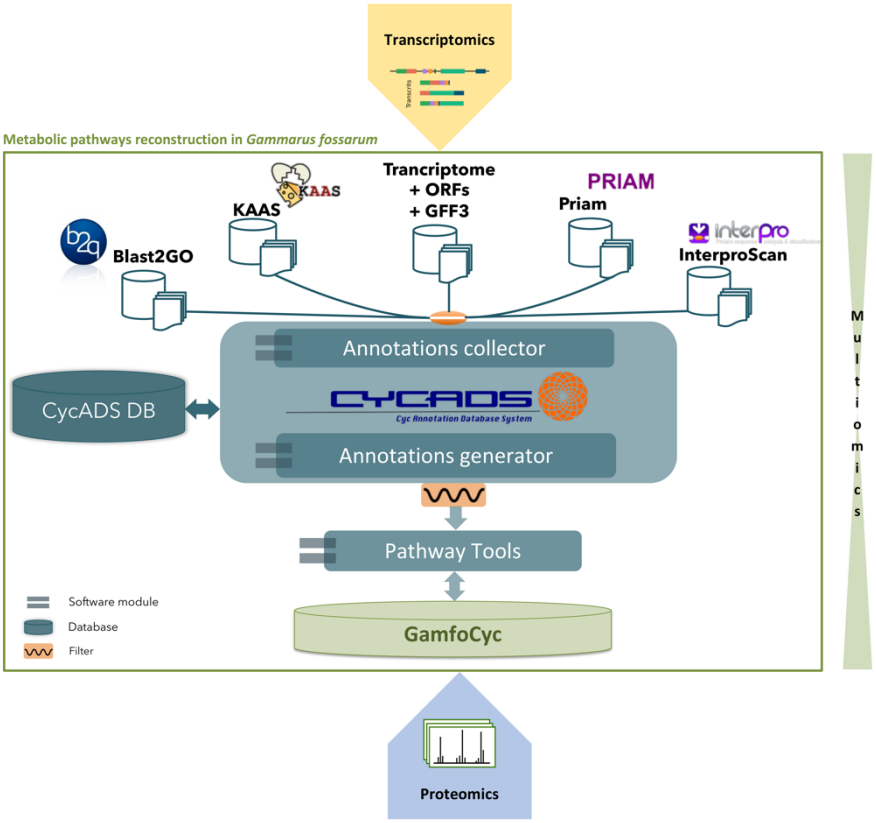

Figure 1. Outline of the methodological approach for the characterization of global metabolism in *Gammarus fossarum* using multi-omics integration involving transcriptomics, proteomics, and the CycADS annotation management system (adapted from (Vellozo et al., 2011). All software were used with their default settings, except for Blast2GO (e-value 10-5 against Swiss-Prot reference database), Interproscan (all sub-methods), Kegg (2 references « for genes » and « for eukaryote »). All annotation were collected by CycADS then extracted without any filter.

Transcriptomic resources

RNA-seq data were retrieved from Cogne et al. (2019). Male and female gammarid data belonged to genotyped individuals of the species *G. fossarum*, subtype B. Illumina TruSeq stranded mRNA Sample Prep kit was used to generate cDNA libraries from poly-A enriched RNAs for each of the samples, which were then sequenced on two lanes of Hiseq3000 (Illumina) using a paired-end read length of 2 × 150bp with the HiSeq Reagent Kits (Illumina, San Diego, California) (Cogne et al. 2019). The raw reads are available in the "NCBI Sequence Reads Archive" under accession number SRR8089722 (“RNAseq GFBM,” 2018) (BioProject PRJNA497972 and BioSample SAMN10259937) for the male and under accession number SRR8089729 (“RNAseq GFBF,” 2018) (BioProject PRJNA497972 and BioSample SAMN10259936) for the female.

De novo transcriptome assembly

Pre-processing of *RNA-Seq data*

A FastQC v0.11.9 analysis of the RNA-seq raw read dataset was performed (Figure S1, Pre-processing step 1), for all forward and reverse reads (male and female samples). Forward reads from male and female samples were merged in a single file, as well as reverse reads, to pool the RNA-seq data in a single paired-end dataset. The mean quality scores, defined as the mean quality value across each base position in the read, were satisfying for every sample (Phred > 28). All sequence lengths were 150 bp (base pair) long.

Trimmomatic v0.36 (Bolger et al., 2014) was used to trim residual adapter sequences and/or low-quality bases (Figure S1, Pre-processing, step 2), with palindrome mode and a PHRED score threshold set to 5 to avoid excessive trimming, according to MacManes' benchmarking (2014). Only paired reads were kept.

To reduce RNA contamination possibly coming from organisms’ microbiome and sample manipulation, we identified nucleic sequences using the taxonomic sequence classifier Kraken2 v2.1.2 (Wood et al., 2019; Wood and Salzberg, 2014) (Figure S1, Pre-processing, step 3). Taxon information was obtained from the NCBI taxonomy database (Federhen, 2012). Taxonomic classification was run using a complete set of reference libraries, comprising RefSeq complete sets of proteins from Archaea,

Bacteria, Fungi, Plant, and Protozoa taxa as well as plasmid and viruses and GRCh38 human proteins (O'Leary et al., 2016).

To further exclude from the raw reads any rRNA sequence residues, SortMeRNA v4.3.4 (Kopylova et al., 2012) was used for filtering rRNA sequences using SILVA database (Quast et al., 2013) with representative small and large subunit rRNA sequences from Archaea, Bacteria, and Eukarya (Figure S1, Pre-processing, step 4).

Last quality control was performed to ensure that data met assembly requirements and to retrieve pre-processing statistics with the MultiQC v1.12 tool (Ewels et al., 2016) (Figure S1, Pre-processing, step 5).

Assembly pipeline

De novo assembly of *G. fossarum* transcriptome was performed using the Trinity v2.14.0 pipeline (Grabherr et al., 2011; Haas et al., 2013) (Figure S1, Normalization and Assembly, step 1). We combined the standard procedure using the three independent modules (Inchworm, Chrysalis, and Butterfly) in addition to the *in silico* normalization procedure to reduce assembly time and errors.

A general assembly metrics report, including total trinity transcripts, median contig length, average contig, total assembled bases, etc. was obtained with a Trinity embedded script (https://github.com/trinityrnaseq/trinityrnaseq/blob/330b6fe9c65af0e203c5620708cce0fd6f57ceab/util/TrinityStats.pl). TransDecoder v5.5.0 (Haas, 2018) was used to identify candidate coding regions within transcript sequences (Figure S1, Normalization and Assembly, step 2). We estimated the completeness and redundancy of processed data with the Benchmarking Universal Single-Copy Orthologs (BUSCO) v5.3 (Simão et al., 2015), based on universal single-copy orthologs in the OrthoDB database (Figure S1, Normalization and Assembly, step 3) (Kriventseva et al., 2019). The *arthropoda_odb10* database, made up of 1,013 arthropod single-copy orthologs was used, as it is the closest taxonomically to gammarids. BUSCO was run in transcriptome mode with the default options.

To reduce the redundancy of alternative transcripts, we first chose to filter out possible transcript artifacts and transcripts with low expression levels as recommended by (Haas et al., 2013). Then, we evaluated the percentage of expression level (abundance) for a given transcript compared to all transcripts within an isoform cluster (Haas et al., 2013), the so-called isoform percentage (IsoPct), to

retain the highest abundant isoform by applying the "*Highest Iso Only*" option (Figure S1, Normalization and Assembly, step 4). To compute transcript isoform abundances as normalized expression values (Fragments Per Kilobase of transcript per Million FPKM), the RSEM (Li and Dewey, 2011) and Bowtie2 v2.4.5 (Langmead and Salzberg, 2012) tools were used.

To assess the quality of the final assembly, we performed a new BUSCO analysis (Figure S1, Normalization and Assembly, step 5).

Functional annotation of reference transcriptome by CycADS and metabolic network reconstruction

CycADS v1.36 is a system that allows the standardization of genome annotation, executed by command line and customizable with a configuration file (Vellozo et al., 2011). In this study, in order to adapt a transcriptomic assembly for CycADS, we created a structural annotation file (GFF3) with TransDecoder v5.5.0 (Haas, 2018). Then, CycADS was used according to the methodology described by Vellozo et al. (2011) (Figure 1).

In particular, we performed a functional annotation on the assembled transcriptome with pipelines on local machine, namely Blast2GO v2.5 (Conesa et al., 2005; Conesa and Götz, 2008), UniProtKB/Swiss-Prot protein database (The UniProt Consortium, 2021), Priam v2_JAN_18 (Claudel-Renard et al., 2003), InterproScan v5.31-70.0 (Jones et al., 2014), and the KAAS-KEGG v2.1 online pipeline (Moriya et al., 2007). Functional information was collected (KEGG Orthology (KO), Enzyme Commission (EC) number, and Gene Ontology (GO)) from the annotation methods outputs using the CycADS annotation collector module (Figure 1). All annotations were then extracted and written in an enriched "Pathological file" (PF) which was used in the "Pathway Tools" compartment (Karp et al., 2010) to perform the metabolic reconstruction and generate the corresponding BioCyc Pathway/Genome Database (PGDB), named GamfoCyc (Figure 1). Software settings are given in Figure 1.

For the following analyses, to focus on the metabolic network, we chose to work only on the subset of enzymes involved in the metabolic pathways (see results and discussion part).

The MetExplore v2.30.8 platform (Cottret et al., 2018) has been developed to explore metabolic pathways, manipulate the metabolic network graph, and map omics data. Here, we used the online platform to analyze the completeness of each identified metabolic pathway. The network was therefore

exported as SBML and BioPAX files containing gene products, genetic enzyme complexes, reactions, pathways, and metabolites and was then imported into MetExplore.

The GamfoCyc database was added to the Arthropodacyc metabolic database collection (available at http://arthropodacyc.cycadsys.org/) and the MetExplore database (available at https://metexplore.toulouse.inrae.fr/GAMFO_GFB).

Proteomic resources

Proteomics data were acquired from three males and three females of *G. fossarum* species. Six organs or anatomical regions (cephalon gills, caeca, intestine, male gonads, female gonads, and the rest of the body) were sampled and collected for each organism and frozen in liquid nitrogen before protein extraction. Gills and caeca proteomes were previously described in Leprêtre et al., 2023.

Protein extraction

Each organ was directly dissolved in 40 µL of LDS sample buffer (Invitrogen), except for the cephalon and the rest of the body, which were previously ground in LDS by adding a 4-mm steel ball and then using a tissue homogenizer. Samples were subjected to 1 min of sonication (transonic 780H sonicator) and were heated for 5 min at 95 °C. Organ shreds were completely dissolved in LDS sample buffer, and then 35 µL of each sample was subjected to SDS-PAGE on a 10-well 4-12% gradient NuPAGE (Invitrogen) for 10 min at 150 V with MES buffer. Gels were stained with Coomassie Blue Safe dye (Invitrogen) and destained overnight with water. The total protein in each well was extracted as a single polyacrylamide strip and processed for further decoloring and iodoacetamide treatment. Proteins were proteolyzed with sequencing grade trypsin (Roche) using 0.01% Protease-MAX surfactant (Promega).

Mass spectrometry

The resulting peptide mixtures were analyzed in data-dependent acquisition mode with an Orbitrap Exploris high-resolution mass spectrometer (MS) (ThermoFisher) coupled to an UltiMate 3000 LC system (Dionex-LC Packings), operated as described previously (Ramos-Nascimento et al., 2023).

Protein identification and quantification by spectral counting

Peak lists were generated with Mascot Daemon software (version 2.3.2; Matrix Science) using the data import filter extract_msn.exe (Thermo). Data import filter options were set to 400 (minimum mass),

5,000 (maximum mass), 0 (clustering tolerance), 0 (intermediate scans), and 1,000 (threshold), as described previously (Christie-Oleza et al., 2012). MS/MS spectra were assigned to peptide sequences with the Mascot Daemon 2.3.2 search engine (Matrix Science) against the custom database derived from the assembled transcriptome. Protein detection was validated with at least 1 specific peptide and 2 peptides in total.

Two tables were obtained for each individual from mass spectrometry. The first table (Table MS1 available at https://doi.org/10.57745/HMQVCS) corresponds to the peptides (32,925 unique male peptide sequences, 39,018 unique female peptide sequences) detected by mass spectrometry and their associated characteristics (associated protein identifier, sequence, length, mass/charge (*m/z*) ratio, modifications, retention time, Mascot score, etc.). The second table (Table MS2, available at https://doi.org/10.57745/HMQVCS) contains the protein abundance measurements for each sample expressed as "spectral count" (SC). Indeed, spectral counts are an estimate of the abundance of a protein in the samples through the number of MS/MS spectra that are associated (Liu et al., 2004). The more abundant a protein is, the more likely its peptides will be frequently fragmented in the mass spectrometer. For the male samples, we have a matrix of 5,073 identified proteins (Table MS2 GFBM). For the female samples, we have a matrix of 5,678 identified proteins (Table MS2 GFBF).

Proteomic data integration

We retrieved EC numbers from transcriptomic annotation (GamfoCyc) (n=4,033) and EC numbers of proteins identified in shotgun proteomics (all organs combined) (n=4,150; available at https://doi.org/10.57745/HMQVCS). These two lists of ECs were mapped onto the metabolic maps of several pathways in the KEGG database (Kanehisa and Goto, 2000) through to the KEGG Mapper collection of tools (Kanehisa et al., 2022; Kanehisa and Sato, 2020). This approach enables us to observe the ability of shotgun proteomics to confirm the presence of putative proteins found in transcriptomics data.

To retain proteins whose abundance may be low and specific to an organ and exclude those whose detection is incidental, only proteins identified by at least one spectral count and present in at least 3 samples were retained (Gregori et al., 2013). From the filtered table, differential analysis of protein abundance in organs was performed using the R package EdgeR (Robinson et al., 2010) version 3.32.1.

Although this function was originally designed for RNA-Seq count data, it is also applicable to spectral count data in proteomics (Gregori et al., 2013). The selection of differentially expressed (DE) proteins was based on a False Discovery Rate (FDR) threshold of 0.05 and an absolute change in expression of 2 (|Fold Change| > 2) (Table S2).

Pathway analysis

We collected the list of DE proteins found in abundance in gills (n=653), caeca (n=635), male gonads (n=343), and female gonads (n=187), and used them as an input list for the pathway enrichment analysis. We performed an overrepresentation analysis (ORA) (Khatri et al., 2012) from the WebGestalt server (Liao et al., 2019), using the KEGG pathways databases (Kanehisa et al., 2022). The tool was used with the following options: FDR (Benjamini-Hocheberg adjusting method), at least three genes from the input list in the enriched category, and the whole potential coding transcriptome (n=63,639 ORFs) as the reference background. We obtained an enrichment ratio (ER), which is the number of observed proteins divided by the number of expected proteins from each KEGG category in the n-protein list (Liao et al., 2019). We thus obtained a percentage of proteins overrepresented in the pathway of interest. We chose to use the "weighted set cover" method to reduce the redundancy of the gene sets in the enrichment result to identify the most representative significant gene sets for visualization (Liao et al., 2019).

Results and discussion

A reference transcriptome for the sentinel species Gammarus fossarum (subtype B)

In this work, we pooled female and male individual RNA-seq datasets from *G. fossarum* (subtype B) individuals to obtain a new reference transcriptome for this species.

The number of merged male and female reads is 181,487,648 (around 90.7 M forward and 90.4 M reverse) (Table 1). We identified 23,361,906 reads (12.9%) (Table 1) as belonging to different taxa using Kraken 2 (Wood et al., 2019; Wood and Salzberg, 2014) and we identified bacterial RNAs as the main contaminant source (5.34%), followed by green plants RNAs (2.05%) and fungi (1.1%). *Homo sapiens* RNAs represented the single species main contaminant source (2.05%) (Figure S2A). In parallel, the search for rRNAs with SortMeRNA (Kopylova et al., 2012) highlighted 674,405 reads (0.37%) coming from ribosomal databases (Table 1), of which 63,230 matched a hit in the SILVA archaeal database

(Quast et al., 2013), 110,652 in the bacterial database, and 474,224 in the eukaryotic database (Figure S2B). The total number of read after the contamination removal with Kraken2 and SortMeRNA amounted at 157,400,267 (86.73%) and were used for the *de novo* assembly of *G. fossarum* transcriptome and its metabolism reconstruction.

Table 1. Summary of the pre-processing results.

Step	Raw Data	Trimmomatic	Kraken2	SortMeRNA	Total
Input reads (n)	181,487,296	181,436,578	158,074,674	157,400,267	157,400,267
Dropped reads (n)	-	50,718	23,361,906	674,405	24,087,029
Remaining reads (%)	100%	99.97 %	87.10 %	86.73 %	86.73 %

The results of our pre-processing step show that RNA contamination from food sources, organisms' microbiomes, and even experimental manipulation must be considered to limit artifacts of transcriptome assembly of novel species. To our knowledge, this has not been done for most environmental species (Caputo et al., 2020; Cogne et al., 2019; Llorente et al., 2020). In parallel, these results show the interest of considering RNA contaminants to reconstruct the holotranscriptome (host transcriptome and microbiota transcriptome) of *G. fossarum*, similarly to recent hologenome approaches performed in *Daphnia magna* or the metaproteomic studies recently tested in gammarids (Chaturvedi et al., 2023, Gouveia et al 2020).

Since the initial assembly strategy of the two original transcriptomes (Cogne et al., 2019) may involve the potential presence of redundant transcripts and isoforms, we took into account the expression level of the transcripts and to keep only the most highly expressed transcript (i.e., the highest isoform) in the refined final assembly (Haas et al., 2013). After excluding lowly expressed isoforms with the RSEM tool, the refined reference transcriptome contained 71.10% (302,024) of the original transcripts (Table 2). Compared with the primary assembly the number of transcripts under 200 bases decreased to close to zero in the filtered assembly ("n_under_200" from 49 to 2, Table 2). The percentage of contigs with an ORF is higher in the new transcriptome assembly ("mean_orf_percent" from about 48% to about 54%, Table 2).

Table 2. Metrics for the *Gammarus fossarum* transcriptome assemblies.

ASSEMBLY STATISTICS [a]	PRIMARY GFB ASSEMBLY	HIGHEST ISOFORM ONLY GFB ASSEMBLY
n_seqs	424,800	302,024
smallest	161	186
largest	31,223	24,766
n_bases	340,819,579	178,666,341
mean_len	802.28409	591.56214
n_under_200	49	2
n_over_1k	88,262	38,028
n_over_10k	392	74
n_with_orf	78,824	42,997
mean_orf_percent	47.86537	54.13628
n90	297	259
n70	670	424
n50	1,440	801
n30	2 640	1,648
n10	5,151	3,678
gc	0.42831	0.4215
bases_n	0	0

[a] "n_seqs": number of contigs in the assembly; "smallest" (resp. largest): size of the smallest (resp. largest) contig; "n_bases": number of bases included in the assembly; "mean_len": mean length of the contigs; "n_under_X" (resp. over): number of contigs shorter (resp. greater) than X bases long; "n_with_orf": number of contigs that had an open reading frame; "mean_orf": for contigs with an ORF, mean percentage of the contig covered by the ORF; "NX": largest contig size at which at least X% of bases are contained in contigs at least this length; "gc": percentage of bases that are G or C, "bases_n": number of bases that are N.

To evaluate the impact of the pre-processing strategy as well as the decrease in redundancy of isoforms in our filtered assembly, we performed a BUSCO analysis. The BUSCO results show that more than 92% of the 1,013 *Arthropoda* orthologous genes (sum of complete and fragmented) are found in the primary transcriptome, and 90% in the refined transcriptome (Figure 2). This indicates that our refined *G. fossarum* transcriptome assembly did not lose much of the expected biological information from its phylogenetic membership. About 38% (381 out of 1,013) of the complete genes were found in a single copy in the primary transcriptome, against about 63% (638 out of 1,013) in the refined transcriptome (Figure 2). Thus, our pipeline applied to the RNA-seq data led to an increased presence of single-copy genes (+ 25%) and a decreased number of duplicated orthologs (- 28%), with only a slight increase (of about 2%) in the number of missing genes. These results suggest that using a biology-driven approach that takes into consideration RNA origins and level of expression may increase the species-specificity of *de novo* assemblies. The remaining duplicated genes in the refined transcriptome could

come either from different haplotypes or isoforms still present in the assembly or from true duplications in the amphipod family, a hypothesis that should be tested using new genomic resources. Additionally, it must be noted that the BUSCO arthropod database derives from OrthoDB, which contains only *Daphnia pulex* as a crustacean species (Kriventseva et al., 2019), indicating an under-representation of crustaceans in the ortholog database. The completeness of the transcriptomes is positively correlated to the proportion of complete BUSCO genes, but this evaluation may be biased by the number of species close to the species of interest (Amil-Ruiz et al., 2021; Seppey et al., 2019). In summary, this new assembly provides a common annotation of *G. fossarum* male and female organisms with improved transcript coverage and reduced redundancy and external RNA contamination.

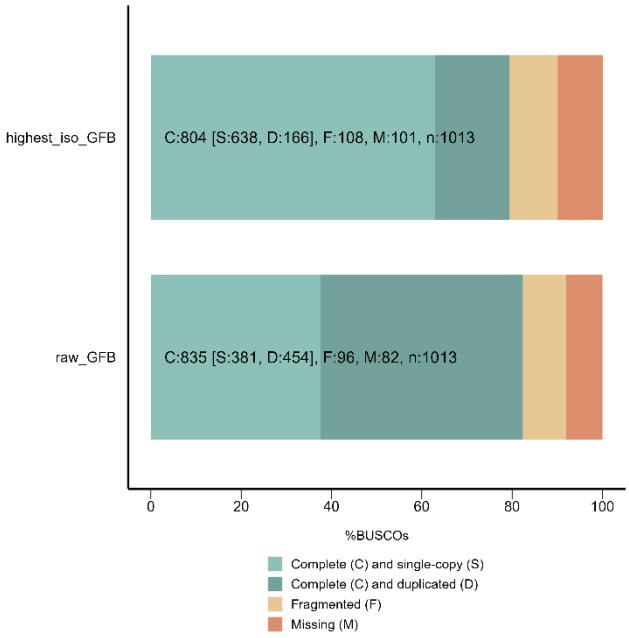

Figure 2. Completeness analysis of the pre-filtered (raw_GFB) and filtered reference (highest_iso_GFB) transcriptomes of *G. fossarum* via BUSCO of arthropods.

Development of the metabolism database (GamfoCyc) dedicated to Gammarus fossarum.

The refined transcriptome was then annotated by the CycADS pipeline (Vellozo et al., 2011) to create a database of the metabolic pathways of *G. fossarum*. This database, named GamfoCyc, was obtained from the analysis of 63,639 polypeptides/ORFs sequences (available at https://doi.org/10.57745/HMQVCS) from 42,997 contigs (Table 2). Four methods were used to conduct the functional annotation of the *G. fossarum* transcriptome, collecting 7,630 enzymes classified as

complete, i.e. EC numbers with all the 4 levels of enzyme classification defined representing 1,764 distinct EC numbers (Figure 3A). Of these 7,630 enzymes, 4,033 are involved in the metabolic pathways identified by BioCyc and thus constitute the *G. fossarum* metabolic network. All functional analyses were therefore performed on this subset of 4,033 enzymes. KAAS and BLAST2GO were the major sources of annotation with 1,332 and 1,166 annotations respectively, compared to Interpro (447 annotations) and Priam (412 annotations) (Figure 3A). There were 194 EC numbers common to all methods (Figure 3A). We can also note that 840 EC numbers are not cross-referenced, which is due on the one hand to the different annotation strategies that do not target the same features (i.e., local alignment vs sequence profile-based searches or functional domain searches) and on the other hand to the incomplete EC numbers (i.e., not all classification levels are defined) which are not taken into account by PathwayTools (Figure 3A).

The collection of functional annotations allowed the reconstruction of metabolic pathways, and the mapping of 7,630 enzymes onto these pathways (Table S1). Our filtering method does not identify annotation artifacts (e.g., transcripts may be incomplete). In fact, from a biological point of view, if two isoforms are annotated as conducting the same enzymatic reaction, the Pathway Tools tool keeps both annotations. In this case, the presence of true isozymes is also possible.

In order to assess the network quality by measuring the completeness of the metabolic pathways in GamfoCyc, the MetExplore platform (Cottret et al., 2018) was used to assess the percentage of reactions annotated with enzymes in pathways. A comparison with the metabolic network of two model organisms, *Drosophila melanogaster* and *Daphnia pulex,* was also performed. The results showed that the GamfoCyc database contains 374 annotated metabolic pathways composed of 2,610 reactions (Table S1). Among the annotated metabolic pathways, 295 pathways (79 %) contained more than 75% of reactions annotated with enzymes. Only 4 pathways (1%) contained less than 25% of annotated reactions (Figure 3B).

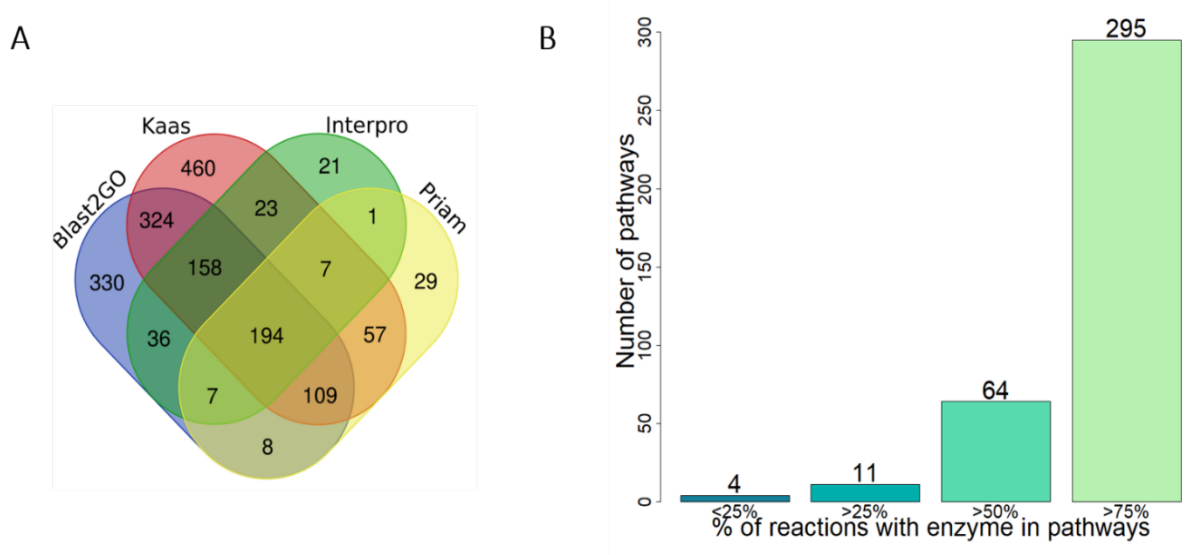

Figure 3. Functional annotation of the reference transcriptome (A) Venn diagram of EC annotation in *Gammarus fossarum* according to the different methods (Blast2GO, KAAS-KEGG, InterproScan, Priam) of the CycADS pipeline (B) Completeness analysis of metabolic pathways annotation via MetExplore.

These latter are involved in eumelanin biosynthesis, iron-sulfur cluster biosynthesis, protein nitrosylation/denitrosylation, and glycosphingolipids biosynthesis. Interestingly, glycosphingolipids are rarely identified in the lipidome studies on gammarids (Arambourou et al., 2018; Fu et al., 2021) and our metabolic reconstruction sheds light on these previous findings suggesting that this pathway is either absent or not detectable at gene expression level in *G. fossarum*. In the case of eumelanin biosynthesis, scarce knowledge of this metabolic pathway in *Arthropoda* may explain this result, since arthropod cuticular melanin has been shown to be different from mammalian epidermal melanins (Barek et al., 2018).

We therefore compared our results with data from these species to illustrate the relevance of our strategy for reconstructing the metabolic pathways of a non-model organism from the transcriptome. As a comparison, in the *Drosophila melanogaster* database (genome version release 6 plus ISO1MT, (Hoskins et al., 2015), 4,965 enzymes are annotated and in the *Daphnia pulex* database (genome version jgi_v11 geneset, (Colbourne et al., 2011) 3,672 enzymes are annotated (Baa-Puyoulet et al., 2016; Consuegra et al., 2020; McQuilton et al., 2012, Nordberg et al., 2014). As both species have annotated genomes, the large number of enzymes found in *G. fossarum* (n=7,630) from the transcriptomes could be explained by the presence of multiple annotated transcripts for the same enzyme, despite the reduced

redundancy of our assembly. The reconstructed global metabolism of *D. melanogaster* contains 2,201 reactions for 353 pathways (Figure S3). Regarding the completeness of the annotation of the globality of the pathways, 277 pathways (78.4%) contain more than 75% of reactions with annotated enzymes, and 5 pathways (1.4%) contain less than 25% of reactions with annotated enzymes (Figure S3). Thus, the results obtained for *G. fossarum* transcriptome are qualitatively similar to those of a model organism for which the reconstruction of the global metabolic pathways was performed using a sequenced and annotated genome. This work shows that the exploitation of transcriptomic data is therefore feasible and relevant in a non-model species by adapting the CycADS tool to this type of data and opens new perspectives for other non-model species for which data are already available.

Integration of proteomic data for metabolic annotation

After the reconstruction of *G. fossarum* metabolic pathways from the transcriptomic data, we validated the presence of the enzymes and studied the organ-specific metabolic profiles by integrating the shotgun mass spectrometry proteomic data from six distinct organs or anatomical regions (cephalon, gills, caeca, intestine, gonads, and the rest of the body) sampled individually from three male and three female *G. fossarum B* individuals. In female organs, we obtained 35,515 peptides of different sequences and 3,643 proteins and a cumulative 466,942 spectral counts for all 18 samples analyzed (Figure S4A). In male organs, we obtained 29,174 peptides of different sequences and 3,133 proteins (validated with at least one specific peptide and 2 peptides in total) and cumulatively 353,123 spectral counts for all 18 samples analyzed (Figure S4A). By cross-referencing male and female tables of spectral counts, the overall proteomic profile of *G. fossarum* was composed of 4,150 proteins in total. This refined version of the *G. fossarum* reference transcriptome allowed for increasing the proteomic catalog of this species by more than two-fold compared to the first reported proteogenomic study (n=1,873) (Trapp et al., 2014).

In total, 858 enzymes annotated in the metabolic pathways were experimentally validated by mass spectrometry measurements. This represents 21.3% of the enzymes mapped on the transcriptome (n=4,033) and 20.7% of the proteins identified by shotgun proteomics (n=4,150) (Figure S4).

To find and visualize where enzymes are involved in metabolic reactions, we chose to map the enzymes annotated by the transcriptome and validated in proteomics onto the metabolic pathways. As

an example, the mapping of the enzymes validated by proteomics on the beta-oxidation pathway (more than 75% of the reactions annotated) is shown in Figure 4. In contrast, the glycosphingolipid biosynthesis pathway, (less than 25% reactions annotated) showed only two enzymes identified by shotgun proteomics (Figure S5).

The integration of proteomic data that we proposed in this work provides the first multi-omics and experimental validation of metabolic pathways in an environmental sentinel species.

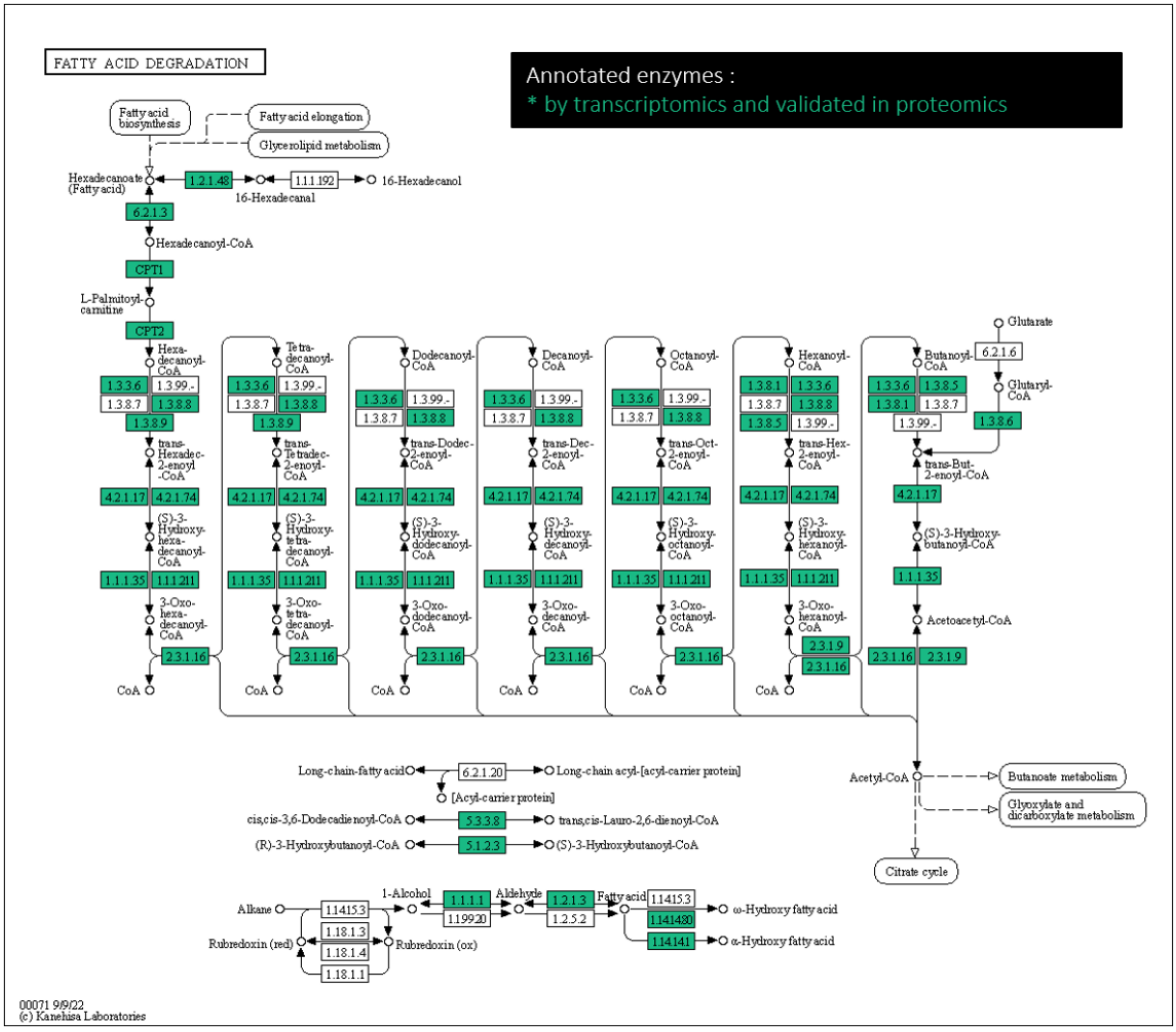

Figure 4: Example of mapped enzymes of *Gammarus fossarum* involved in the beta-oxidation pathway. Pathway module (functional unit of gene sets in metabolic pathway) is highlighted in green when the correspondent EC was annotated by transcriptomics and validated in proteomics.

Organ-specific metabolism through proteomic profiling

We retained 2,190 proteins across the proteomic dataset to investigate the organs-specific basal metabolism in gammarids, after filtering out lowly abundant proteins and spectral count normalization.

A multivariate analysis of the data shows organ-specific protein profiles is shown in Figure 5. Proteome profiles showed an expected sexual dimorphism in the gonads, and a general weak variability in biological replicates (Figure 5). In fact, inter-sex comparisons in each single organ showed few significantly differentially expressed proteins. For example, male and female gills differed by around 1% (n=21) of their proteome (Figure S6). Similar results were found for the caeca, approximately, with 1.6% (n=35) of proteins found differentially expressed between male and females (Figure S7). These results suggest a low inter-sex variability between *G. fossarum* organs at the proteome level, except for the gonads (Figure 5). Intestines showed higher individual variability compared to other organs, probably due to a lower protein content (mirrored by a lower spectral count compared to other organs) in intestine protein extracts, and therefore a more difficult reproducibility for these organs.

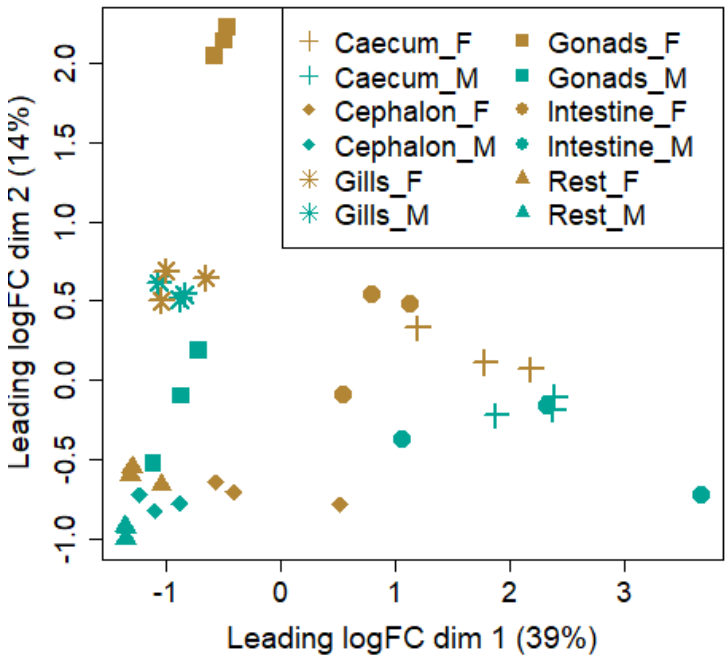

Figure 5 : Multi-dimensional scaling (MDS) plot of expression profiles of male (*_M) and female (*_F) samples in *Gammarus fossarum* proteome.

In order to investigate organ-specific metabolic profiles, we first compared the proteomes and the metabolic pathways in the gills and the caeca compared to the other organs (Figure 6,Table S2). In the gills we found 653 proteins more abundant than the other organs (FDR<0.05, FC>2), of which 237 are

annotated in GamfoCyc (Figure 6A, Table S2). Proteins involved in energy metabolism were predominant in the gills, in particular the pathways of the TCA cycle with an ER of 30.17% (FDR<2.2e-16) (Figure 6B, Table S5). Proteins involved in oxidative phosphorylation and the degradation of fatty acids were also enriched, with an enrichment of around 20% (FDR <2.2e-16) and 16% (FDR=1.58e-10), respectively (Table S5). Similar results have been shown previously in gammarids (Leprêtre et al., 2023) and may be explained by the extremely energy demanding processes the gills are responsible for, such as oxygen uptake, acid–base balance, and osmotic and ionic regulation (Henry and Wheatly, 1992; Péqueux, 1995). As an example, osmoregulation is suggested to represent 11% of the total energy budget in *Gammarus pulex* (Felten et al., 2008). On the other hand, oxidative phosphorylation is the most effective energy-release process in animals (Dimroth et al., 2000; Jin et al., 2019). This pathway also appears to be modulated by abiotic conditions such as salinity in aquatic organisms (Bal et al., 2021), and pollutants in *Gammarus* spp. (Kunz et al., 2010) or plays a role in the adaptation of *Gammarus lacustris* to environmental conditions in the Tibetan region (Jin et al., 2019). Altogether, these results point out a key role of cellular respiration based on ATP biosynthesis and fatty acid consumption as a major energy resource in the gills of *G. fossarum*.

In the caeca, 635 proteins were more abundant (FDR<0.05, FC>2) compared to other organs (Figure 6C, Table S2). Among these DE proteins, 228 were concomitantly annotated in GamfoCyc. Indeed, gammarid caeca have been previously identified as important organs for amino acids, lipids, carbohydrates, and vitamins (Leprêtre et al., 2023), but not specific pathways were described. The glycan degradation pathway was overrepresented in the caecum with an ER of 26.90% (FDR<2.2^{e}-16) (Figure 6D, Table S6). This pathway has been reported to be involved in growth factor signaling and morphogenesis in arthropods (Scanlan et al., 2015) and the expression of genes involved in this pathway was altered by exposure to certain flame retardants (Scanlan et al., 2015). Proteins involved in the retinol metabolism pathway were also enriched in the caecum in the present study to around 23% (FDR=5.93e-08) (Table S6). Retinoids play crucial roles in many physiological processes, including embryonic development, morphogenesis, and cellular differentiation in vertebrates (Ghyselinck and Duester, 2019), and were also shown to be responsible for the development of crustaceans and insects (Liñán-Cabello et al., 2002; Nakamura et al., 2007). The retinoid system has been explored scarcely in crustaceans

(Liñán-Cabello et al., 2002). However, some retinoids (oxidated forms of retinoate, retinoate isomers, and retinaldehyde isomers) were quantified in *G. fossarum* and were shown to fluctuate during the reproductive cycle in females and their levels were affected by methoprene, a juvenile hormone analog (Gauthier et al., 2023). Another over-represented KEGG pathway in the gammarid caeca was the glutathione metabolism with an ER of about 12% (FDR<2.2^{e}-16) (Figure 6D, Table S6). Glutathione is known to be involved in detoxification, sequestering labile metals such as cadmium and preventing cytotoxicity (Khan et al., 2012). Recent studies by Gestin et al. (2023, 2022, 2021), have shown that gammarid caeca bioaccumulate and eliminate cadmium. These metabolic pathways may play a role in protecting against heavy metal exposure and oxidative stress following environmental variations.

Then we focused on the metabolic pathways present in the gonads. In total, 343 proteins were more abundant in the testes (FDR<0.05 and FC>2), while 187 DE proteins were overexpressed in the ovaries (FDR<0.05 and FC>2) (Figure S8A). Among these 530 DE proteins (male and female combined), 149 are annotated in GamfoCyc (Figure S8B).

In male gonads, glycolysis was the most enriched metabolic pathway with around 29% of enrichment ratio (ER) (FDR<2.2e-16) (Figure S8C and Table S3). Glycolysis is a pathway by which ATP molecules are formed from glucose to supply energy to the organism and has been highlighted as a key pathway in spermatogenesis in various vertebrate species. For instance, glycolysis has been shown to play an important role in early and late spermatogenesis in *Drosophila hydei* (Geer et al., 1972). Previous observations made using co-expression networks in gammarid male and female gonads showed that glycolysis was highly enriched in testes (Degli Esposti et al., 2019). Notably, through integrated metabolomics and transcriptomics analyses of *Macrobrachium nipponense* testes, it was found that glycolysis/gluconeogenesis and the tricarboxylic acid (TCA) cycle may play an essential role in promoting the process of male sexual differentiation and development by supplying ATP (Jin et al., 2020).

In the female gonads, metabolic pathways involved in glutathione and purine metabolism were found overrepresented with respectively around 11% and 10% of enrichment and an FDR of less than 0.05 (Figure S8D and Table S4). While predominant, amino acid metabolisms such as those involving alanine, aspartate, and glutamate were not significantly enriched. The glutathione system is involved in

detoxification and oxidative stress response. Interestingly, in the crustacean *Metapenaeus ensis*, glutathione peroxidase was found to be specifically expressed in early ovaries, suggesting that the glutathione peroxidase might play a pivotal role in preventing oocytes from oxidative damage and thus in crustacean reproduction (Wu and Chu, 2010; Xia et al., 2013). Many aspects of oocyte maturation, such as mitosis and meiosis include DNA and RNA synthesis and rearrangements, and thus require an intense use purine nucleotides. Purines also act as enzyme cofactors, participate in cellular signaling, act as phosphate group donors to generate cellular energy (Carter et al., 2008).

Conclusions

Our analysis demonstrates that transcriptomic data can be exploited by specific annotation systems such as CycADS to improve the annotation of metabolic pathways in species lacking genomic resources and the integration with proteomics data can contribute to improve the phenotypic characterization at the organ level and avoid potential biases (i.e., false negatives or false positives) coming from single omics approaches (Ge et al., 2003; Reeves et al., 2009). The reported results highlight the value of applying omics approaches on organs of small crustaceans to assess their metabolic specificities. Moreover, in the case of sentinel species used to assess the impact of contaminants on the ecosystems, this work put the basis to investigate the ability of environmental contaminants to disrupt metabolism depending on the target organ.

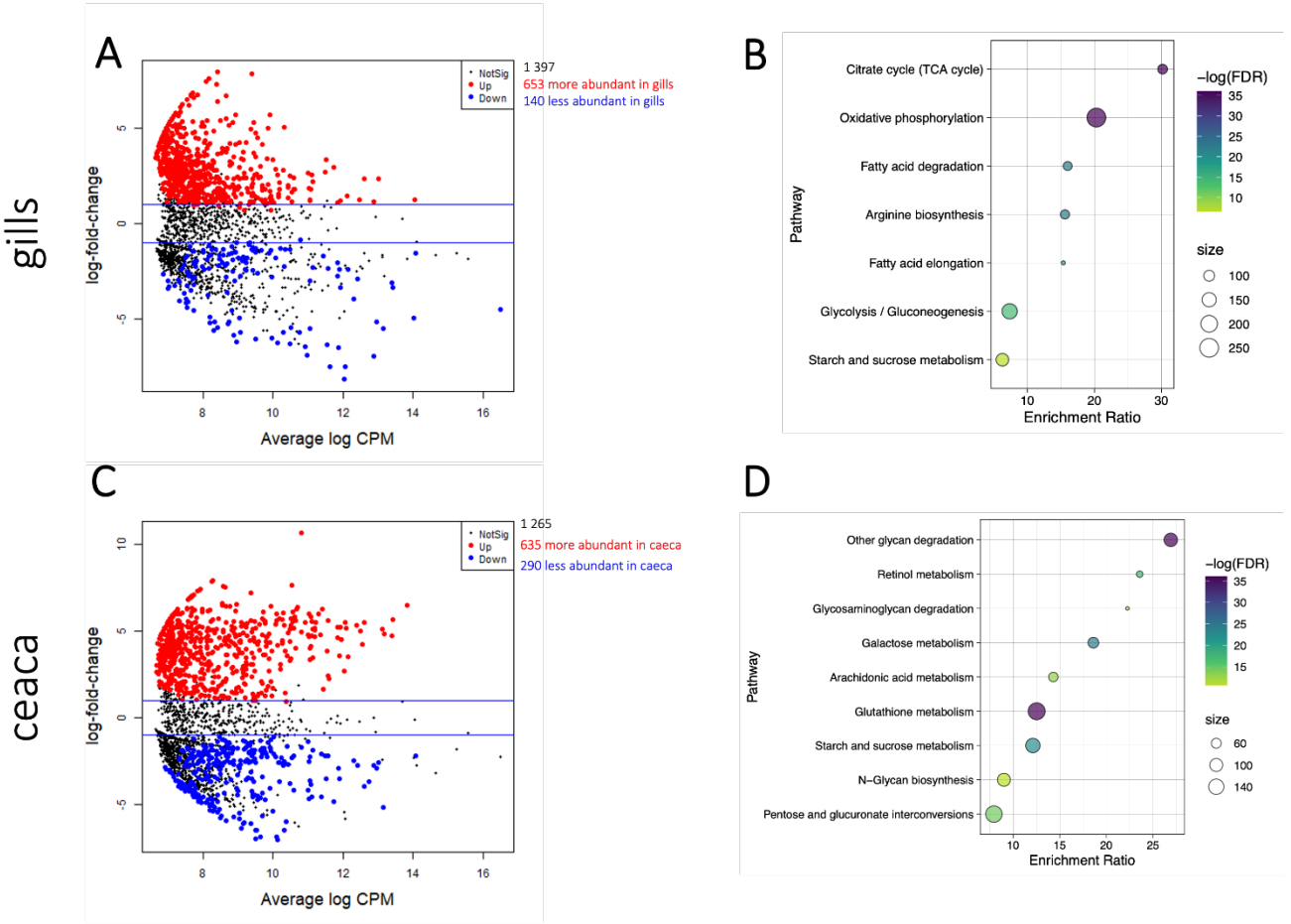

Figure 6. Differential protein abundances in gills and caeca of *G. fossarum* compared with the other organs (FDR < 0.05, LFC>2). (A) The MA plot, in red the most abundant proteins in gills, in blue the least abundant proteins in gills, (B) Enrichment plot for KEGG pathway categories in gills, the size of the dots is proportional to the number of genes present in the pathways, all the pathways presented have an FDR > 0.05, the more significant the enrichment, the darker the dot. (C) The MA plot in caeca. (D) Enrichment plot for KEGG pathway categories in the caeca.

Data access

The sequences of the original reads are available in the "NCBI Sequence Reads Archive" under accession number SRR8089722 (“RNAseq GFBM,” 2018) (BioProject PRJNA497972 and BioSample SAMN10259937) for the male and under accession number SRR8089729 (“RNAseq GFBF,” 2018) (BioProject PRJNA497972 and BioSample SAMN10259936) for the female.

Original mass spectrometry data are available via the PRIDE repository with the dataset identifiers PXD040344. Peptide sequence data and spectral count data are available on https://doi.org/10.57745/HMQVCS.

All R scripts combining proteomic data integration, differential analysis, and metabolic pathway analysis are available on Github (https://github.com/NatachaKoenig/MultiomicsAnnotationGFB).

All functional annotations and metabolic network data are available on entrepot.recherche.data.gouv.fr (https://doi.org/10.57745/HMQVCS).

Competing interests

The authors declare no competing interests.

Acknowledgements

The authors benefitted from the French GDR “Aquatic Ecotoxicology” framework which aims at fostering stimulating scientific discussions and collaborations for integrative approaches. This work has been supported by the APPROve project (ANR-18- CE34-0013-01) and by the Fédération de Recherche BioEEnVis (GamfoCyc project).

Credit Author statement

NK: formal analysis, investigation, visualization, writing - original draft, writing - review & editing.

DDE: conceptualization, funding acquisition supervision, writing - original draft, writing - review & editing

OG: funding acquisition, review & editing

HQ, LG, ND, JCG, MK: methodology.

AL, PBP, ML and KS: formal analysis.

VN, AC, OG, ML, KS, HC, FC, SA, JA: writing - review & editing.

References

Amil-Ruiz, F., Maria Herruzo-Ruiz, A., Fuentes-Almagro, C., Baena-Angulo, C., Manuel Jimenez-Pastor, J., Blasco, J., Alhama, J., Michan, C., 2021. Constructing a de novo transcriptome and a reference proteome for the bivalve *Scrobicularia plana*: Comparative analysis of different assembly strategies and proteomic analysis. Genomics 113, 1543–1553. https://doi.org/10.1016/j.ygeno.2021.03.025

Arambourou, H., Fuertes, I., Vulliet, E., Daniele, G., Noury, P., Delorme, N., Abbaci, K., Barata, C., 2018. Fenoxycarb exposure disrupted the reproductive success of the amphipod *Gammarus fossarum* with limited effects on the lipid profile. PLOS ONE 13, e0196461. https://doi.org/10.1371/journal.pone.0196461

Baa-Puyoulet, P., Parisot, N., Febvay, G., Huerta-Cepas, J., Vellozo, A.F., Gabaldón, T., Calevro, F., Charles, H., Colella, S., 2016. ArthropodaCyc: a CycADS powered collection of BioCyc databases to analyse and compare metabolism of arthropods. Database (Oxford) 2016. https://doi.org/10.1093/database/baw081

Bal, A., Panda, F., Pati, S.G., Das, K., Agrawal, P.K., Paital, B., 2021. Modulation of physiological oxidative stress and antioxidant status by abiotic factors especially salinity in aquatic organisms. Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology 241, 108971. https://doi.org/10.1016/j.cbpc.2020.108971

Barek, H., Sugumaran, M., Ito, S., Wakamatsu, K., 2018. Insect cuticular melanins are distinctly different from those of mammalian epidermal melanins. Pigment Cell Melanoma Res 31, 384–392. https://doi.org/10.1111/pcmr.12672

Besse, J.-P., Coquery, M., Lopes, C., Chaumot, A., Budzinski, H., Labadie, P., Geffard, O., 2013. Caged *Gammarus Fossarum* (Crustacea) as a robust tool for the characterization of bioavailable contamination levels in continental waters: Towards the determination of threshold values. Water Research 47, 650–660. https://doi.org/10.1016/j.watres.2012.10.024

Bolger, A.M., Lohse, M., Usadel, B., 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170

Caputo, D.R., Robson, S.C., Werner, I., Ford, A.T., 2020. Complete transcriptome assembly and annotation of a critically important amphipod species in freshwater ecotoxicological risk assessment: *Gammarus Fossarum*. Environment International 137, 105319. https://doi.org/10.1016/j.envint.2019.105319

Carter, N.S., Yates, P., Arendt, C.S., Boitz, J.M., Ullman, B., 2008. Purine and pyrimidine metabolism in *Leishmania*. Adv. Exp. Med. Biol. 625, 141–154. https://doi.org/10.1007/978-0-387-77570-8_12

Chaturvedi, A., Li, X., Dhandapani, V., Marshall, H., Kissane, S., Cuenca-Cambronero, M., Asole, G., Calvet, F., Ruiz-Romero, M., Marangio, P., Guigó, R., Rago, D., Mirbahai, L., Eastwood, N., Colbourne, J.K., Zhou, J., Mallon, E., Orsini, L., 2023. The hologenome of *Daphnia magna* reveals possible DNA methylation and microbiome-mediated evolution of the host genome. Nucleic Acids Research 51, 9785–9803. https://doi.org/10.1093/nar/gkad685

Christie-Oleza, J.A., Miotello, G., Armengaud, J., 2012. High-throughput proteogenomics of *Ruegeria pomeroyi*: seeding a better genomic annotation for the whole marine *Roseobacter* clade. BMC Genomics 13, 73. https://doi.org/10.1186/1471-2164-13-73

Claudel-Renard, C., Chevalet, C., Faraut, T., Kahn, D., 2003. Enzyme-specific profiles for genome annotation: PRIAM. Nucleic Acids Research 31, 6633–6639.

Cogne, Y., Degli-Esposti, D., Pible, O., Gouveia, D., François, A., Bouchez, O., Eché, C., Ford, A., Geffard, O., Armengaud, J., Chaumot, A., Almunia, C., 2019. De novo transcriptomes of 14 gammarid individuals for proteogenomic analysis of seven taxonomic groups. Sci. Data 6, 1–7. https://doi.org/10.1038/s41597-019-0192-5

Colbourne, J.K., Pfrender, 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., Cáceres, C.E., Carmel, L., Casola, C., Choi, J.-H., Detter, J.C., Dong, Q., Dusheyko, S., Eads, B.D., Fröhlich, T., Geiler-Samerotte, K.A., Gerlach, D., Hatcher, P., Jogdeo, S., Krijgsveld, J., Kriventseva, E.V., Kültz, D., Laforsch, C., Lindquist, E., Lopez, J.,

Manak, J.R., Muller, J., Pangilinan, J., Patwardhan, R.P., Pitluck, S., Pritham, E.J., Rechtsteiner, 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., Lucas, S.M., Robertson, H.M., Bork, P., Koonin, E.V., Zdobnov, E.M., Grigoriev, I.V., Lynch, M., Boore, J.L., 2011. The Ecoresponsive Genome of *Daphnia pulex*. Science 331, 555–561. https://doi.org/10.1126/science.1197761

Conesa, A., Götz, S., 2008. Blast2GO: A comprehensive suite for functional analysis in plant genomics. International Journal of Plant Genomics 2008, 1–12. https://doi.org/10.1155/2008/619832

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

Cottret, L., Frainay, C., Chazalviel, M., Cabanettes, F., Gloaguen, Y., Camenen, E., Merlet, B., Heux, S., Portais, J.-C., Poupin, N., Vinson, F., Jourdan, F., 2018. MetExplore: collaborative edition and exploration of metabolic networks. Nucleic Acids Res 46, W495–W502. https://doi.org/10.1093/nar/gky301

Degli Esposti, D., Almunia, C., Guery, M.-A., Koenig, N., Armengaud, J., Chaumot, A., Geffard, O., 2019. Co-expression network analysis identifies gonad- and embryo-associated protein modules in the sentinel species *Gammarus Fossarum*. Sci. Rep. 9, 7862. https://doi.org/10.1038/s41598-019-44203-5

Degli Esposti, D., Lalouette, A., Gaget, K., Lepeule, L., Chaabi, Z., Leprêtre, M., Espeyte, A., Delorme, N., Quéau, H., Garnero, L., Calevro, F., Chaumot, A., Geffard, O., 2024. Identification and organ-specific patterns of expression of two metallothioneins in the sentinel species *Gammarus fossarum*. Comp. Biochem. Physiol. B Biochem. Mol. Biol. 269, 110907. https://doi.org/10.1016/j.cbpb.2023.110907

Dimroth, P., Kaim, G., Matthey, U., 2000. Crucial role of the membrane potential for ATP synthesis by F(1)F(o) ATP synthases. J. Exp. Biol. 203, 51–59. https://doi.org/10.1242/jeb.203.1.51

Ewels, P., Magnusson, M., Lundin, S., Käller, M., 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics 32, 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Federhen, S., 2012. The NCBI Taxonomy database. Nucleic Acids Res. 40, D136–D143. https://doi.org/10.1093/nar/gkr1178

Felten, V., Charmantier, G., Charmantier-Daures, M., Aujoulat, F., Garric, J., Geffard, O., 2008. Physiological and behavioural responses of *Gammarus pulex* exposed to acid stress. Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology 147, 189–197. https://doi.org/10.1016/j.cbpc.2007.09.006

Fritsche, K., Ziková-Kloas, A., Marx-Stoelting, P., Braeuning, A., 2023. Metabolism-disrupting chemicals affecting the liver: screening, testing, and molecular pathway identification. Int. J. Mol. Sci. 24, 2686. https://doi.org/10.3390/ijms24032686

Fu, T., Knittelfelder, O., Geffard, O., Clement, Y., Testet, E., Elie, N., Touboul, D., Abbaci, K., Shevchenko, A., Lemoine, J., Chaumot, A., Salvador, A., Degli-Esposti, D., Ayciriex, S., 2021. Shotgun lipidomics and mass spectrometry imaging unveil diversity and dynamics in *Gammarus fossarum* lipid composition. iScience 24, 102115. https://doi.org/10.1016/j.isci.2021.102115

Fuertes, I., Jordão, R., Piña, B., Barata, C., 2019. Time-dependent transcriptomic responses of *Daphnia magna* exposed to metabolic disruptors that enhanced storage lipid accumulation. Environmental Pollution 249, 99–108. https://doi.org/10.1016/j.envpol.2019.02.102

Gauthier, M., Daniele, G., Giroud, B., Lafay, F., Vulliet, E., Jumarie, C., Garric, J., Boily, M., Geffard, O., 2023. The retinoid metabolism of *Gammarus Fossarum* is disrupted by exogenous all-trans retinoic acid, citral, and methoprene but not by the technical formulation of glyphosate. Ecotoxicology and Environmental Safety 252, 114602. https://doi.org/10.1016/j.ecoenv.2023.114602

Ge, H., Walhout, A.J.M., Vidal, M., 2003. Integrating "omic" information: a bridge between genomics and systems biology. Trends Genet. 19, 551–560. https://doi.org/10.1016/j.tig.2003.08.009

Geer, B.W., Martensen, D.V., Downing, B.C., Muzyka, G.S., 1972. Metabolic changes during spermatogenesis and thoracic tissue maturation in *Drosophila hydei*. Developmental Biology 28, 390–406. https://doi.org/10.1016/0012-1606(72)90022-X

Gestin, O., Lacoue-Labarthe, T., Coquery, M., Delorme, N., Garnero, L., Dherret, L., Ciccia, T., Geffard, O., Lopes, C., 2021. One and multi-compartments toxico-kinetic modeling to understand metals' organotropism and fate in *Gammarus fossarum*. Environ. Int. 156, 106625.

Gestin, O., Lacoue-Labarthe, T., Delorme, N., Garnero, L., Geffard, O., Lopes, C., 2023. Influence of the exposure concentration of dissolved cadmium on its organotropism, toxicokinetic and fate in *Gammarus Fossarum*. Environment International 171, 107673. https://doi.org/10.1016/j.envint.2022.107673

Gestin, O., Lopes, C., Delorme, N., Garnero, L., Geffard, O., Lacoue-Labarthe, T., 2022. Organ-specific accumulation of cadmium and zinc in *Gammarus Fossarum* exposed to environmentally relevant metal concentrations. Environmental Pollution 308, 119625. https://doi.org/10.1016/j.envpol.2022.119625

Ghyselinck, N.B., Duester, G., 2019. Retinoic acid signaling pathways. Development 146, dev167502. https://doi.org/10.1242/dev.167502

Gouveia, D., Bonneton, F., Almunia, C., Armengaud, J., Quéau, H., Degli-Esposti, D., Geffard, O., Chaumot, A., 2018. Identification, expression, and endocrine-disruption of three ecdysone-responsive genes in the sentinel species *Gammarus Fossarum*. Sci. Rep. 8, 3793. https://doi.org/10.1038/s41598-018-22235-7

Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., Thompson, D.A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B.W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., Regev, A., 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29, 644–652. https://doi.org/10.1038/nbt.1883

Gregori, J., Sánchez, A., Villanueva, J., 2013. msmsTests: LC-MS/MS Differential Expression Tests. R package version 1.14. 0.

Haas, BJ., 2018. TransDecoder v 5.5.0. URL https://github.com/TransDecoder/TransDecoder

Haas, B.J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P.D., Bowden, J., Couger, M.B., Eccles, D., Li, B., Lieber, M., MacManes, M.D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C.N., Henschel, R., LeDuc, R.D., Friedman, N., Regev, A., 2013. De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. Nat. Protoc. 8, 10.1038/nprot.2013.084. https://doi.org/10.1038/nprot.2013.084

Henry, R.P., Wheatly, M.G., 1992. Interaction of respiration, ion regulation, and acid-base balance in the everyday life of aquatic crustaceans. American Zoologist 32, 407–416. https://doi.org/10.1093/icb/32.3.407

Hoskins, R.A., Carlson, J.W., Wan, K.H., Park, S., Mendez, I., Galle, S.E., Booth, B.W., Pfeiffer, B.D., George, R.A., Svirskas, R., Krzywinski, M., Schein, J., Accardo, M.C., Damia, E., Messina, G., Méndez-Lago, M., de Pablos, B., Demakova, O.V., Andreyeva, E.N., Boldyreva, L.V., Marra, M., Carvalho, A.B., Dimitri, P., Villasante, A., Zhimulev, I.F., Rubin, G.M., Karpen, G.H., Celniker, S.E., 2015. The Release 6 reference sequence of the Drosophila melanogaster genome. Genome Res 25, 445–458. https://doi.org/10.1101/gr.185579.114

Jin, S., Bian, C., Jiang, S., Sun, S., Xu, L., Xiong, Y., Qiao, H., Zhang, W., You, X., Li, J., Gong, Y., Ma, B., Shi, Q., Fu, H., 2019. Identification of candidate genes for the plateau adaptation of a tibetan amphipod, *Gammarus lacustris*, through integration of genome and transcriptome sequencing. Front. Genet. 10. https://doi.org/10.3389/fgene.2019.00053

Jin, S., Hu, Y., Fu, H., Sun, S., Jiang, S., Xiong, Y., Qiao, H., Zhang, W., Gong, Y., Wu, Y., 2020. Analysis of testis metabolome and transcriptome from the oriental river prawn (*Macrobrachium nipponense*) in response to different temperatures and illumination times. Comparative Biochemistry and Physiology Part D: Genomics and Proteomics 34, 100662. https://doi.org/10.1016/j.cbd.2020.100662

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics 30, 1236–1240.

Jordão, R., Campos, B., Piña, B., Tauler, R., Soares, A.M.V.M., Barata, C., 2016a. Mechanisms of Action of Compounds That Enhance Storage Lipid Accumulation in Daphnia magna. Environ. Sci. Technol. 50, 13565–13573. https://doi.org/10.1021/acs.est.6b04768

Jordão, R., Casas, J., Fabrias, G., Campos, B., Piña, B., Lemos, M.F.L., Soares, A.M.V.M., Tauler, R., Barata, C., 2015. Obesogens beyond vertebrates: lipid perturbation by tributyltin in the crustacean *Daphnia magna*. Environ. Health Perspect. 123, 813–819. https://doi.org/10.1289/ehp.1409163

Jordão, R., Garreta, E., Campos, B., Lemos, M.F.L., Soares, A.M.V.M., Tauler, R., Barata, C., 2016b. Compounds altering fat storage in *Daphnia magna*. Sci. Total Environ. 545–546, 127–136. https://doi.org/10.1016/j.scitotenv.2015.12.097

Kanehisa, M., Goto, S., 2000. KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res 28, 27–30. https://doi.org/10.1093/nar/28.1.27

Kanehisa, M., Sato, Y., 2020. KEGG Mapper for inferring cellular functions from protein sequences. Protein Science 29, 28–35. https://doi.org/10.1002/pro.3711

Kanehisa, M., Sato, Y., Kawashima, M., 2022. KEGG mapping tools for uncovering hidden features in biological data. Protein Science 31, 47–53. https://doi.org/10.1002/pro.4172

Karp, P.D., Paley, S.M., Krummenacker, M., Latendresse, M., Dale, J.M., Lee, T.J., Kaipa, P., Gilham, F., Spaulding, A., Popescu, L., 2010. Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology. Briefings in Bioinformatics 11, 40–79.

Khan, F.R., Bury, N.R., Hogstrand, C., 2012. Copper and zinc detoxification in *Gammarus pulex* (L.). Journal of Experimental Biology 215, 822–832. https://doi.org/10.1242/jeb.062505

Khatri, P., Sirota, M., Butte, A.J., 2012. Ten years of pathway analysis: current approaches and outstanding challenges. PLOS Computational Biology 8, e1002375. https://doi.org/10.1371/journal.pcbi.1002375

Koenig, N., Almunia, C., Bonnal-Conduzorgues, A., Armengaud, J., Chaumot, A., Geffard, O., Degli Esposti, D., 2021. Co-expression network analysis identifies novel molecular pathways associated with cadmium and pyriproxyfen testicular toxicity in *Gammarus Fossarum*. Aquatic Toxicology 235, 105816. https://doi.org/10.1016/j.aquatox.2021.105816

Kopylova, E., Noé, L., Touzet, H., 2012. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics 28, 3211–3217. https://doi.org/10.1093/bioinformatics/bts611

Kriventseva, E.V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F.A., Zdobnov, E.M., 2019. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Research 47, D807–D811. https://doi.org/10.1093/nar/gky1053

Kunz, P., Kienle, C., Gerhardt, A., 2010. *Gammarus* spp. in aquatic ecotoxicology and water quality assessment: toward integrated multilevel tests. Reviews of environmental contamination and toxicology 205, 1–76. https://doi.org/10.1007/978-1-4419-5623-1_1

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods 9, 357–359. https://doi.org/10.1038/nmeth.1923

Leprêtre, M., Almunia, C., Armengaud, J., Salvador, A., Geffard, A., Palos-Ladeiro, M., 2019. The immune system of the freshwater zebra mussel, *Dreissena polymorpha*, decrypted by proteogenomics of

hemocytes and plasma compartments. J. Proteomics 202, 103366. https://doi.org/10.1016/j.jprot.2019.04.016

Leprêtre, M., Degli Esposti, D., Sugier, K., Espeyte, A., Gaillard, J.-C., Delorme, N., Duflot, A., Bonnard, I., Coulaud, R., Boulangé-Lecomte, C., Xuereb, B., Palos Ladeiro, M., Geffard, A., Geffard, O., Armengaud, J., Chaumot, A., 2023. Organ-oriented proteogenomics functional atlas of three aquatic invertebrate sentinel species. Sci. Data 10, 643. https://doi.org/10.1038/s41597-023-02545-w

Leprêtre, M., Geffard, A., Palos Ladeiro, M., Dedourge-Geffard, O., David, E., Delahaut, L., Bonnard, I., Barjhoux, I., Nicolaï, M., Noury, P., Espeyte, A., Chaumot, A., Degli-Esposti, D., Geffard, O., Lopes, C., 2022. Determination of biomarkers threshold values and illustration of their use for the diagnostic in large-scale freshwater biomonitoring surveys. Environmental Sciences Europe 34, 115. https://doi.org/10.1186/s12302-022-00692-2

Li, B., Dewey, C.N., 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12, 323. https://doi.org/10.1186/1471-2105-12-323

Liao, Y., Wang, J., Jaehnig, E.J., Shi, Z., Zhang, B., 2019. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. Nucleic Acids Research 47, W199–W205. https://doi.org/10.1093/nar/gkz401

Liñán-Cabello, M. a., Paniagua-Michel, J., Hopkins, P. m., 2002. Bioactive roles of carotenoids and retinoids in crustaceans. Aquaculture Nutrition 8, 299–309. https://doi.org/10.1046/j.1365-2095.2002.00221.x

Liu, H., Sadygov, R.G., Yates, J.R., 2004. A model for random sampling and estimation of relative protein abundance in shotgun proteomics. Anal. Chem. 76, 4193–4201. https://doi.org/10.1021/ac0498563

Llorente, L., Herrero, Ó., Aquilino, M., Planelló, R., 2020. *Prodiamesa olivacea*: *de novo* biomarker genes in a potential sentinel organism for ecotoxicity studies in natural scenarios. Aquatic Toxicology 227, 105593. https://doi.org/10.1016/j.aquatox.2020.105593

MacManes, M., 2014. On the optimal trimming of high-throughput mRNA sequence data. Frontiers in Genetics 5.

Marie, B., Coulaud, R., Boulangé-Lecomte, C., Foucault, P., Lance, É., Duflot, A., Xuereb, B., 2023. Dataset on metabolome dimorphism in different organs of mature *Palaemon serratus* prawn. Data in Brief 48, 109038. https://doi.org/10.1016/j.dib.2023.109038

Markov, G.V., Tavares, R., Dauphin-Villemant, C., Demeneix, B.A., Baker, M.E., Laudet, V., 2009. Independent elaboration of steroid hormone signaling pathways in metazoans. Proceedings of the National Academy of Sciences 106, 11913–11918. https://doi.org/10.1073/pnas.0812138106

Massart, J., Begriche, K., Corlu, A., Fromenty, B., 2022. Xenobiotic-induced aggravation of metabolic-associated fatty liver disease. Int. J. Mol. Sci. 23, 1062. https://doi.org/10.3390/ijms23031062

McCartney, N., Kondakath, G., Tai, A., Trimmer, B.A., 2023. Functional annotation of insecta transcriptomes: A cautionary tale from Lepidoptera. Insect Biochemistry and Molecular Biology 104038. https://doi.org/10.1016/j.ibmb.2023.104038

McQuilton, P., St. Pierre, S.E., Thurmond, J., the FlyBase Consortium, 2012. FlyBase 101 – the basics of navigating FlyBase. Nucleic Acids Research 40, D706–D714. https://doi.org/10.1093/nar/gkr1030

Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A.C., Kanehisa, M., 2007. KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic acids research 35, W182–W185.

Nakamura, A., Stiebler, R., Fantappié, M.R., Fialho, E., Masuda, H., Oliveira, M.F., 2007. Effects of retinoids and juvenoids on moult and on phenoloxidase activity in the blood-sucking insect Rhodnius prolixus. Acta Tropica 103, 222–230. https://doi.org/10.1016/j.actatropica.2007.06.009

Neuparth, T., Machado, A.M., Montes, R., Rodil, R., Barros, S., Alves, N., Ruivo, R., Castro, L.F.C., Quintana, J.B., Santos, M.M., 2020. Transcriptomic data on the transgenerational exposure of the keystone amphipod *Gammarus locusta* to simvastatin. Data in Brief 32, 106248. https://doi.org/10.1016/j.dib.2020.106248

Nordberg, H., Cantor, M., Dusheyko, S., Hua, S., Poliakov, A., Shabalov, I., Smirnova, T., Grigoriev, I.V., Dubchak, I., 2014. The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. Nucleic Acids Research 42, D26–D31. https://doi.org/10.1093/nar/gkt1069

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bao, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., Farrell, C.M., Goldfarb, T., Gupta, T., Haft, D., Hatcher, E., Hlavina, W., Joardar, V.S., Kodali, V.K., Li, W., Maglott, D., Masterson, P., McGarvey, K.M., Murphy, M.R., O'Neill, K., Pujar, S., Rangwala, S.H., Rausch, D., Riddick, L.D., Schoch, C., Shkeda, A., Storz, S.S., Sun, H., Thibaud-Nissen, F., Tolstoy, I., Tully, R.E., Vatsan, A.R., Wallin, C., Webb, D., Wu, W., Landrum, M.J., Kimchi, A., Tatusova, T., DiCuccio, M., Kitts, P., Murphy, T.D., Pruitt, K.D., 2016. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 44, D733-745. https://doi.org/10.1093/nar/gkv1189

Péqueux, A., 1995. Osmotic regulation in crustaceans. Journal of Crustacean Biology 15, 1–60. https://doi.org/10.2307/1549010

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., Glöckner, F.O., 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research 41, D590–D596. https://doi.org/10.1093/nar/gks1219

Ramos-Nascimento, A., Grenga, L., Haange, S.-B., Himmelmann, A., Arndt, F.S., Ly, Y.-T., Miotello, G., Pible, O., Jehmlich, N., Engelmann, B., von Bergen, M., Mulder, E., Frings-Meuthen, P., Hellweg, C.E., Jordan, J., Rolle-Kampczyk, U., Armengaud, J., Moeller, R., 2023. Human gut microbiome and metabolite dynamics under simulated microgravity. Gut Microbes 15, 2259033. https://doi.org/10.1080/19490976.2023.2259033

Reeves, G.A., Talavera, D., Thornton, J.M., 2009. Genome and proteome annotation: organization, interpretation and integration. J. R. Soc. Interface 6, 129–147. https://doi.org/10.1098/rsif.2008.0341

Rna-seq transcriptome *Gammarus Fossarum* B female [WWW Document], 2018. . NCBI Sequence Read Archive. URL https://www.ncbi.nlm.nih.gov/sra/SRR8089729 (accessed 12.16.21).

Rna-seq transcriptome *Gammarus Fossarum* B male [WWW Document], 2018. . NCBI Sequence Read Archive. URL https://www.ncbi.nlm.nih.gov/sra/SRR8089722 (accessed 12.16.21).

Robinson, M.D., McCarthy, D.J., Smyth, G.K., 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26, 139–140.

Ruivo, R., Sousa, J., Neuparth, T., Geffard, O., Chaumot, A., Castro, L.F.C., Degli-Esposti, D., Santos, M.M., 2022. From extrapolation to precision chemical hazard assessment: the ecdysone receptor case study. Toxics 10, 6. https://doi.org/10.3390/toxics10010006

Santos, M.M., Ruivo, R., Capitão, A., Fonseca, E., Castro, L.F.C., 2018. Identifying the gaps: Resources and perspectives on the use of nuclear receptor based-assays to improve hazard assessment of emerging contaminants. Journal of hazardous materials 358, 508–511.

Scanlan, L.D., Loguinov, A.V., Teng, Q., Antczak, P., Dailey, K.P., Nowinski, D.T., Kornbluh, J., Lin, X.X., Lachenauer, E., Arai, A., Douglas, N.K., Falciani, F., Stapleton, H.M., Vulpe, C.D., 2015. Gene transcription, metabolite and lipid profiling in eco-indicator *Daphnia magna* indicate diverse mechanisms of toxicity by legacy and emerging flame-retardants. Environ. Sci. Technol. 49, 7400–7410. https://doi.org/10.1021/acs.est.5b00977

Seppey, M., Manni, M., Zdobnov, E.M., 2019. BUSCO: assessing genome assembly and annotation completeness, in: Kollmar, M. (Ed.), Gene Prediction: Methods and Protocols, Methods in Molecular Biology. Springer, New York, NY, pp. 227–245. https://doi.org/10.1007/978-1-4939-9173-0_14

Simão, F.A., Waterhouse, R.M., Ioannidis, P., Kriventseva, E.V., Zdobnov, E.M., 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31, 3210–3212. https://doi.org/10.1093/bioinformatics/btv351

The UniProt Consortium, 2021. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research 49, D480–D489. https://doi.org/10.1093/nar/gkaa1100

Trapp, J., Geffard, O., Imbert, G., Gaillard, J.-C., Davin, A.-H., Chaumot, A., Armengaud, J., 2014. Proteogenomics of *Gammarus fossarum* to document the reproductive system of amphipods. Molecular & Cellular Proteomics 13, 3612–3625. https://doi.org/10.1074/mcp.M114.038851

Vellozo, A.F., Véron, A.S., Baa-Puyoulet, P., Huerta-Cepas, J., Cottret, L., Febvay, G., Calevro, F., Rahbé, Y., Douglas, A.E., Gabaldón, T., Sagot, M.-F., Charles, H., Colella, S., 2011. CycADS: an annotation database system to ease the development and update of BioCyc databases. Database (Oxford) 2011, bar008. https://doi.org/10.1093/database/bar008

Wood, D.E., Lu, J., Langmead, B., 2019. Improved metagenomic analysis with Kraken 2. Genome Biology 20, 257. https://doi.org/10.1186/s13059-019-1891-0

Wood, D.E., Salzberg, S.L., 2014. Kraken: ultrafast metagenomic sequence classification using exact alignments. Genome Biology 15, R46. https://doi.org/10.1186/gb-2014-15-3-r46

Wu, L.T., Chu, K.H., 2010. Characterization of an ovary-specific glutathione peroxidase from the shrimp Metapenaeus ensis and its role in crustacean reproduction. Comparative Biochemistry and Physiology Part B: Biochemistry and Molecular Biology 155, 26–33. https://doi.org/10.1016/j.cbpb.2009.09.005

Xia, X.-F., Zheng, J.-J., Shao, G.-M., Wang, J.-L., Liu, X.-S., Wang, Y.-F., 2013. Cloning and functional analysis of glutathione peroxidase gene in red swamp crayfish *Procambarus clarkii*. Fish Shellfish Immunol. 34, 1587–1595. https://doi.org/10.1016/j.fsi.2013.03.375

Supplementary Information

Table S1. Summary table of GamfoCyc database statistics (extracted from the "Special SmartTables feature")

All compounds of *G. fossarum*	All enzymes of *G. fossarum*	All pathways of *G. fossarum*	All reactions of *G. fossarum*	All transporters of *G. fossarum*
1,903	7,630**	377	2,610*	310

* including 2,328 enzymatic reactions
** of which 4,033 are associated with metabolic pathways and 3,597 are isolated reactions

Table S2. Differential expression analysis result tables for male gonads, female gonads, caeca, and gills in metabolic pathways in *Gammarus fossarum*.

Table S2 is provided as a separate Excel file with four tabs: differentially expressed (DE) proteins in male gonads, DE proteins in female gonads, DE proteins in gills, and DE proteins in caeca.

Table S3. KEGG pathway enrichment analysis of the most abundant proteins in male gonads.

geneSet	description	size	overlap	enrichmentRatio	pValue	FDR
ko00010	Glycolysis / Gluconeogenesis	172	27	29.13	<2.2e-16	<2.2e-16
ko03050	Proteasome	170	15	16.37	3.75E-14	4.90E-12
ko04510	Focal adhesion	49	6	22.72	2.70E-07	2.05E-05
ko00040	Pentose and glucuronate interconversions	164	9	10.18	3.14E-07	2.05E-05
ko04151	PI3K-Akt signaling pathway	90	7	14.43	6.32E-07	3.30E-05
ko04015	Rap1 signaling pathway	123	7	10.56	5.16E-06	2.03E-04
ko00500	Starch and sucrose metabolism	124	7	10.48	5.44E-06	2.03E-04
ko04910	Insulin signaling pathway	29	4	25.60	1.77E-05	5.77E-04
ko05410	Hypertrophic cardiomyopathy	35	4	21.21	3.80E-05	1.10E-03
ko00071	Fatty acid degradation	79	5	11.74	7.17E-05	1.87E-03

Table S4. KEGG pathway enrichment analysis of the most abundant proteins in female gonads.

geneSet	description	size	overlap	enrichmentRatio	pValue	FDR
ko00480	Glutathione metabolism	176	6	11.59	1.49E-05	3.90E-03
ko00230	Purine metabolism	491	8	5.54	1.14E-04	1.03E-02
ko04145	Phagosome	160	5	10.63	1.18E-04	1.03E-02
ko00250	Alanine, aspartate and glutamate metabolism	61	3	16.72	7.94E-04	5.18E-02
ko00190	Oxidative phosphorylation	254	4	5.36	6.94E-03	3.62E-01
ko00604	Glycosphingolipid biosynthesis - ganglio series	4	1	85.01	1.17E-02	4.92E-01
ko04013	MAPK signaling pathway - fly	166	3	6.15	1.32E-02	4.92E-01
ko00030	Pentose phosphate pathway	66	2	10.30	1.63E-02	5.32E-01
ko00020	Citrate cycle (TCA cycle)	87	2	7.82	2.73E-02	7.93E-01
ko04142	Lysosome	231	3	4.42	3.12E-02	8.13E-01

Table S5. KEGG pathway enrichment analysis of the most abundant proteins in gills versus all other organs.

geneSet	description	size	overlap	enrichmentRatio	pValue	FDR
ko00190	Oxidative phosphorylation	254	53	20.29	<2.2e-16	<2.2e-16
ko00020	Citrate cycle (TCA cycle)	87	27	30.17	<2.2e-16	<2.2e-16
ko00071	Fatty acid degradation	79	13	16.00	1.81E-12	1.58E-10
ko00220	Arginine biosynthesis	81	13	15.60	2.53E-12	1.65E-10
ko00062	Fatty acid elongation	57	9	15.35	7.06E-09	3.69E-07
ko04022	cGMP-PKG signaling pathway	79	10	12.31	9.42E-09	4.10E-07
ko00010	Glycolysis / Gluconeogenesis	172	13	7.35	3.35E-08	1.10E-06
ko04151	PI3K-Akt signaling pathway	90	10	10.80	3.38E-08	1.10E-06
ko04210	Apoptosis	55	6	10.61	2.19E-05	6.34E-04
ko00500	Starch and sucrose metabolism	124	8	6.27	4.64E-05	1.21E-03

Table S6. KEGG pathway enrichment analysis of the most abundant proteins in caeca versus all other organs.

geneSet	description	size	overlap	enrichmentRatio	pValue	FDR
ko00480	Glutathione metabolism	176	22	12.52	<2.2e-16	<2.2e-16
ko00511	Other glycan degradation	108	29	26.90	<2.2e-16	<2.2e-16
ko00052	Galactose metabolism	70	13	18.60	2.44E-13	2.12E-11
ko00500	Starch and sucrose metabolism	124	15	12.12	2.42E-12	1.58E-10
ko03010	Ribosome	727	32	4.41	4.86E-12	2.53E-10
ko00830	Retinol metabolism	34	8	23.57	1.36E-09	5.93E-08
ko00040	Pentose and glucuronate interconversions	164	13	7.94	1.34E-08	4.99E-07
ko00590	Arachidonic acid metabolism	56	8	14.31	8.79E-08	2.87E-06
ko00531	Glycosaminoglycan degradation	27	6	22.26	2.39E-07	6.94E-06
ko00510	N-Glycan biosynthesis	100	9	9.02	7.90E-07	2.06E-05

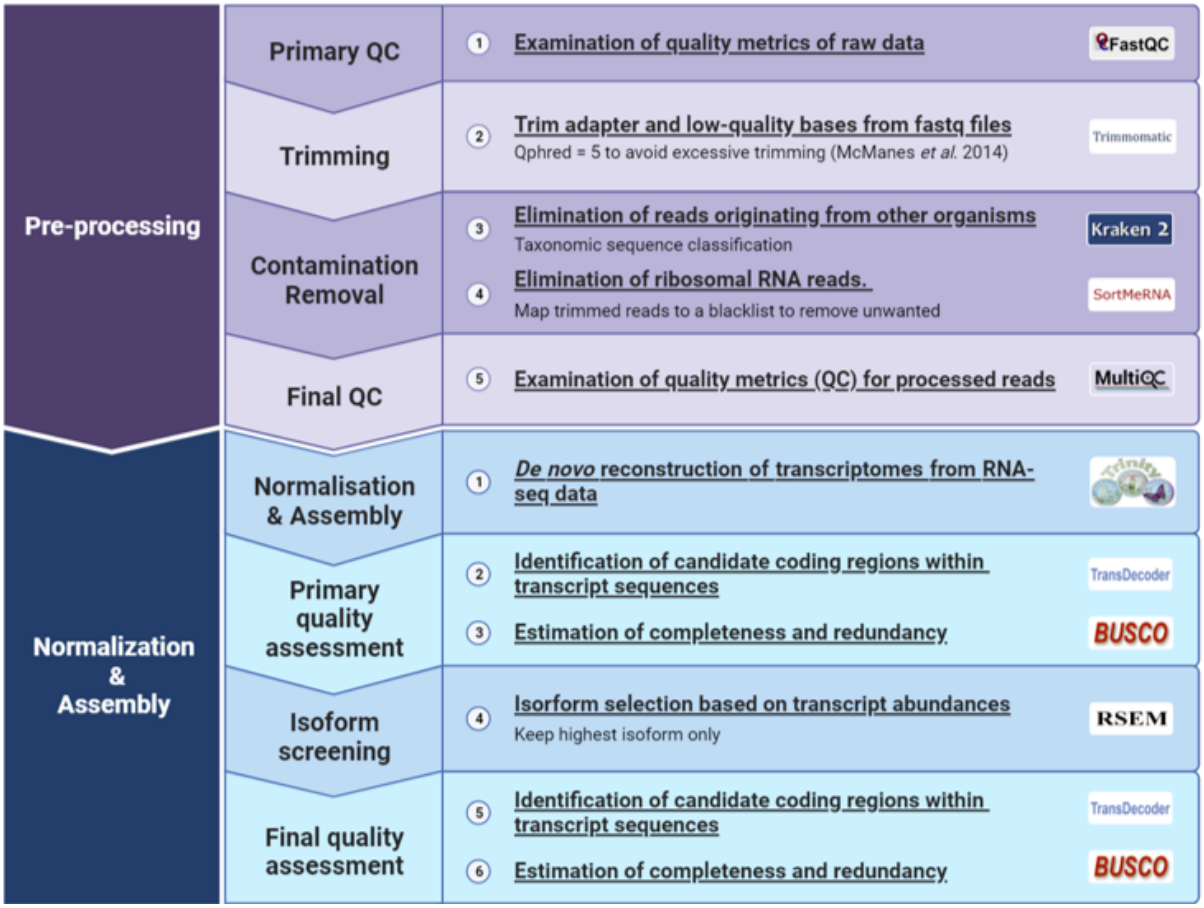

Figure S1. Reads pre-processing and de novo transcriptome assembly pipeline overview.

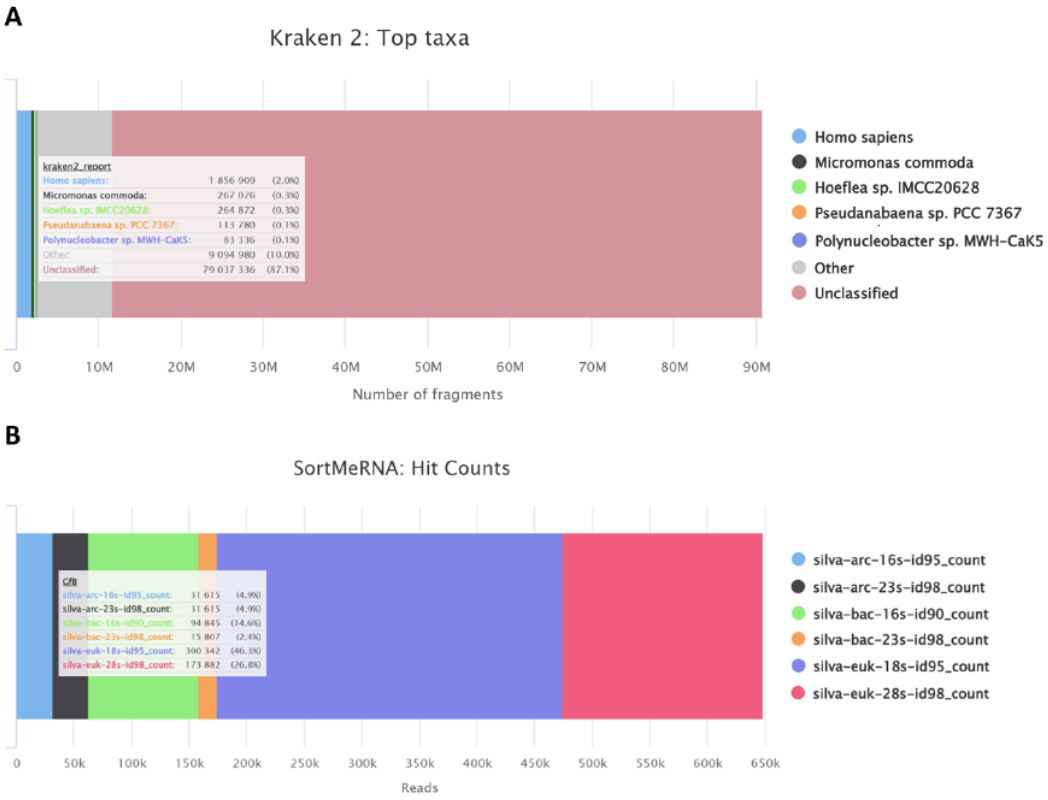

Figure S2. Barplots of the number of fragments/sequences classified for each taxon assigned by (A) Kraken2 and (B) SortMeRNA.

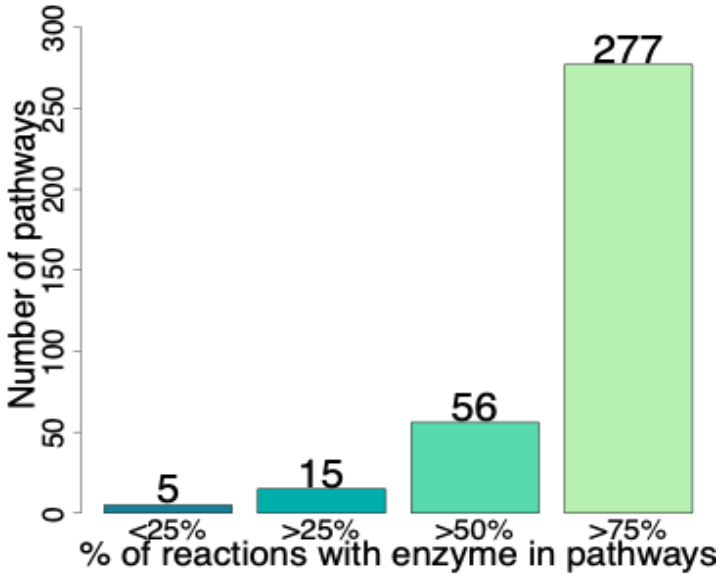

Figure S3. Completion of the global pathways of *Drosophila melanogaster* by MetExplore.

A

Database	Female			Male		
	Peptides	Proteins	SC	Peptides	Proteins	SC
GfB	35,515	3,643	466,942	29,174	3,133	353,123

B

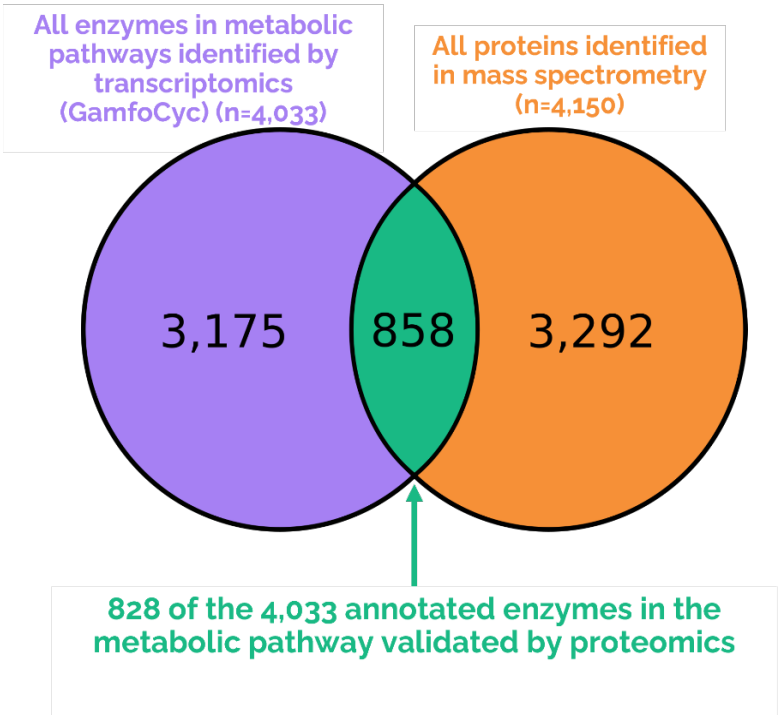

Figure S4. Integration of proteomic data for metabolic annotation (A) Mass spectrometry results interpreted with GfB transcriptome-derived database, SC: spectral count (B) Venn diagram of Gammarus fossarum enzymes annotated by CycADS and validated using proteomic.

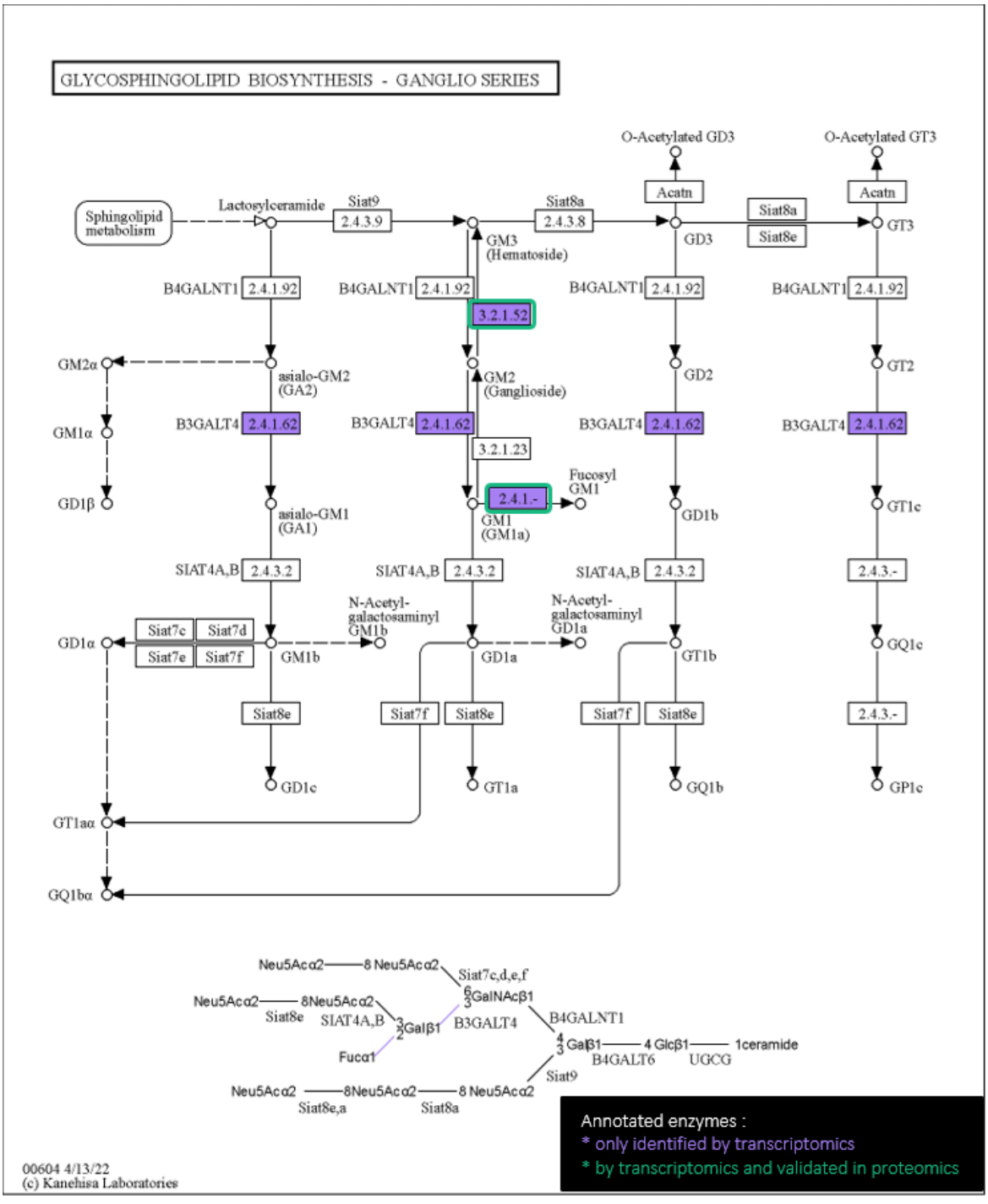

Figure S5. Example of mapped enzymes of *Gammarus fossarum* involved in the biosynthesis of glycosphingolipid pathway. Pathway module (functional unit of gene sets in metabolic pathway) is highlighted in green when the correspondent EC was annotated by transcriptomics and validated in proteomics and highlight in purple when the correspondent EC was annotated only by transcriptomics.

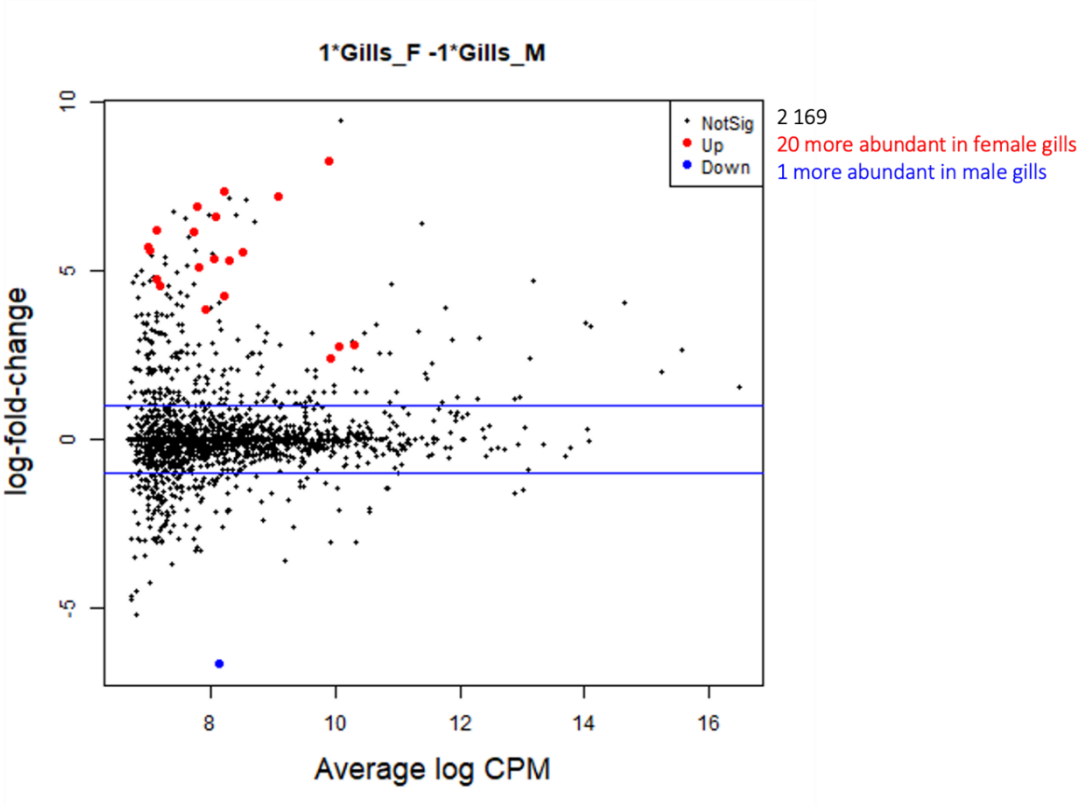

Figure S6. MA plot of differential analysis of female versus male gills

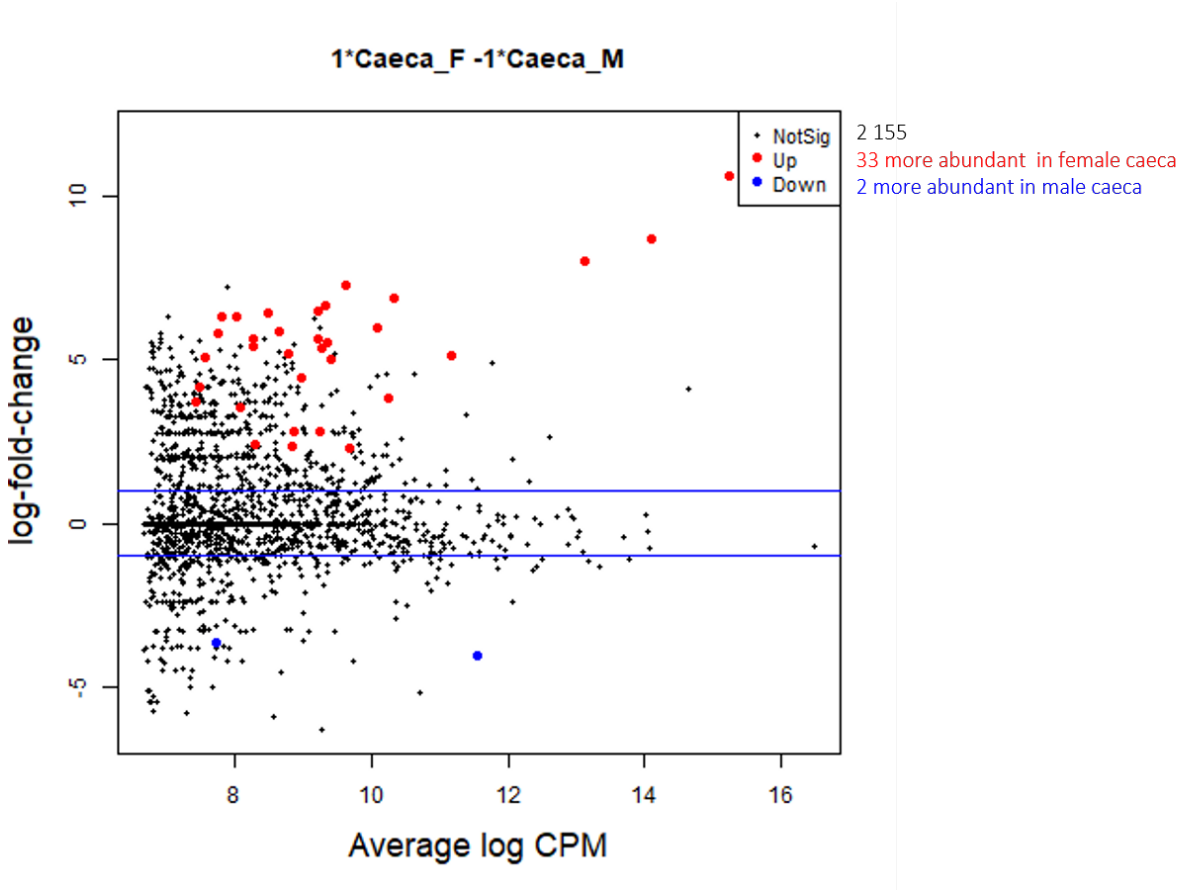

Figure S7. MA plot of differential analysis of male versus female caeca

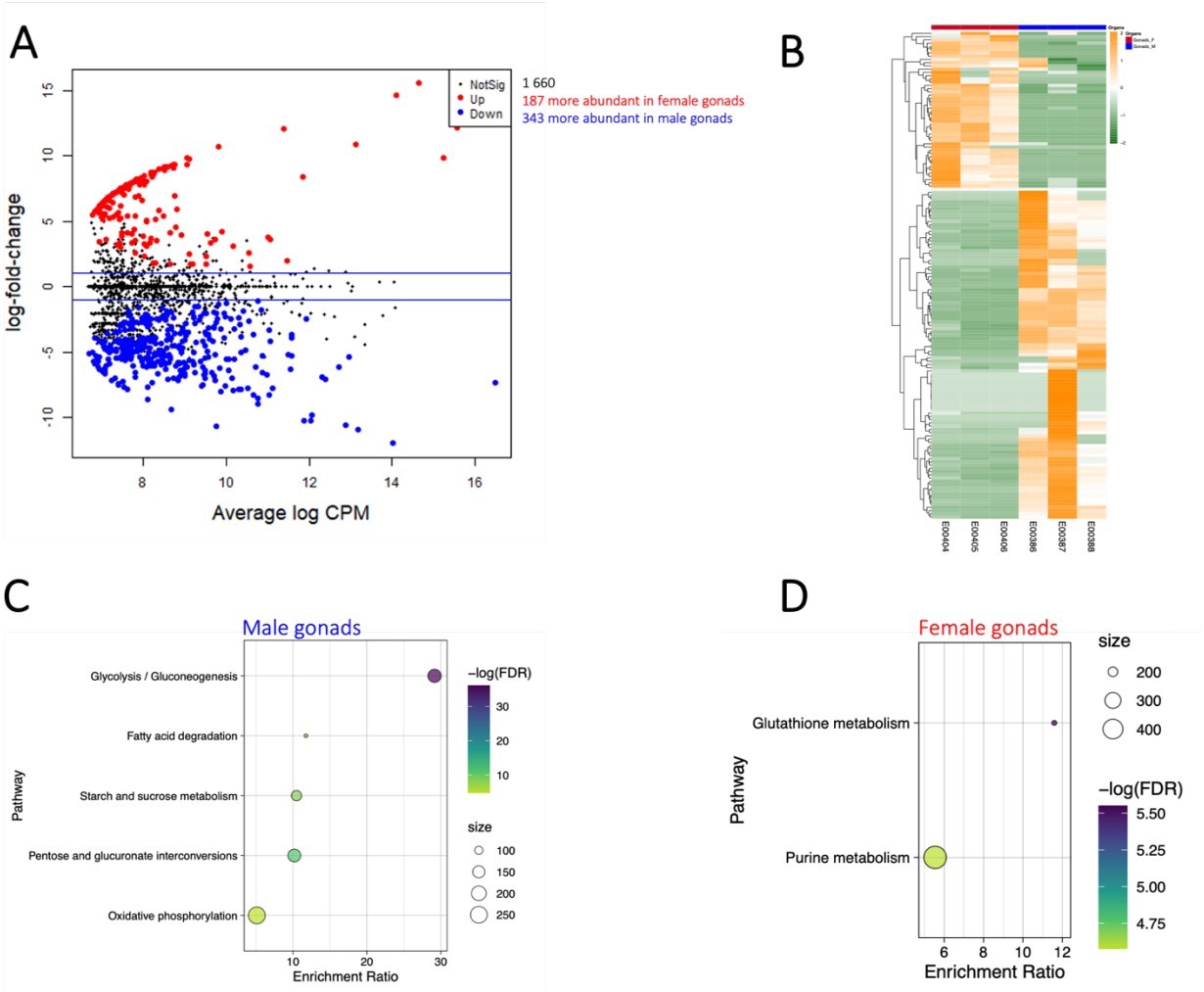

Figure S8. Differential protein abundances between gammarid gonads (FDR < 0.05, LFC > 2). (A) The MA plot, in red the most abundant proteins in female gonads, in blue the most abundant proteins in male gonads, (B) the heatmap of significantly abundant proteins validated by proteomics in the female gonads (red) and male gonads (blue). KEGG pathways enrichment plot for male gonads (C) and female gonads (D), the size of the dots is proportional to the number of genes present in the pathways, all the pathways presented have an FDR > 0.05, the more significant the enrichment, the darker the dot.