Data in Brief 32 (2020) 106248

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

Transcriptomic data on the transgenerational exposure of the keystone amphipod *Gammarus locusta* to simvastatin

Teresa Neuparth[a,*], André M. Machado[a,b], Rosa Montes[c], Rosario Rodil[c], Susana Barros[a], Nélson Alves[a], Raquel Ruivo[a], Luis Filipe C. Castro[a,b], José B. Quintana[c], Miguel M. Santos[a,b,*]

[a] *CIMAR/CIIMAR—Interdisciplinary Centre of Marine and Environmental Research, University of Porto, Avenida General Norton de Matos, S/N, 4450-208 Matosinhos, Portugal*
[b] *FCUP - Department of Biology, Faculty of Sciences, University of Porto, Porto, Portugal*
[c] *Department of Analytical Chemistry, Nutrition and Food Sciences, IAQBUS - Institute of Research on Chemical and Biological Analysis, Universidade de Santiago de Compostela, R. Constantino Candeira S/N, 15782 Santiago de Compostela, Spain*

ARTICLE INFO

Article history:
Received 5 August 2020
Revised 20 August 2020
Accepted 25 August 2020
Available online 31 August 2020

Keywords:
Gammarus locusta, Transgenerational effects
RNA-Seq
Transcriptome analysis
Differential gene expression
Simvastatin Metabolic pathways

ABSTRACT

The use of transcriptomics data brings new insights and works as a powerful tool to explore the molecular mode of action (MoA) of transgenerational inheritance effects of contaminants of emerging concern. Therefore, in this dataset, we present the transcriptomic data of the transgenerational effects of environmentally relevant simvastatin levels, one of the most prescribed human pharmaceuticals, in the keystone amphipod species *Gammarus locusta*. In summary, *G. locusta* juveniles were maintained under simvastatin exposure up to adulthood (exposed group - F0E) and the offspring of F0E were transferred to control water for the three subsequent generations (transgenerational group - F1T, F2T and F3T).
To gain insights into the biological functions and canonical pathways transgenerationally disrupted by simvastatin, a *G. locusta de novo* transcriptome assembly was produced and the transcriptomic profiles of three individual *G. locusta* fe-

DOI of original article: 10.1016/j.envint.2020.106020
* Corresponding authors.
* Corresponding authors at: CIMAR/CIIMAR—Interdisciplinary Centre of Marine and Environmental Research, University of Porto, Avenida General Norton de Matos, S/N, 4450-208 Matosinhos, Portugal.
E-mail addresses: tneuparth@ciimar.up.pt (T. Neuparth), miguel.santos@fc.up.pt (M.M. Santos).

https://doi.org/10.1016/j.dib.2020.106248

males, per group, over the four generations (F0 to F3) - solvent control groups (F0.C, F1.C, F2.C and F3.C), F0 320 ng/L simvastatin exposed group (F0.320E) and F1 to F3 320 transgenerational group (F1.320T; F2.320T and F3.320T) - were analyzed. Briefly, Illumina HiSeq™ 2500 platform was used to perform RNA sequencing, and due to the unavailability of *G. locusta* genome, the RNA-seq datasets were assembled *de novo* using Trinity and annotated with Trinotate software. After assembly and post-processing steps, 106093 transcripts with N50 of 2371 bp and mean sequence length of 1343.98 bp was produced. BUSCO analyses showed a transcriptome with gene completeness of 97.5 % Arthropoda library profile. The Bowtie2, RSEM and edgeR tools were used for the differential gene expression (DEGs) analyses that allowed the identification of a high quantity of genes differentially expressed in all generations. Finally, to identify the main metabolic pathways affected by the transgenerational effects of SIM across all generations, the DGEs genes were blasted onto KEGG pathways database using the KAAS webserver. The data furnished in this article allows a better molecular understanding of the transgenerational effects produced by simvastatin in the keystone amphipod *G. locusta* and has major implications for hazard and risk assessment of pharmaceuticals and other emerging contaminants. This article is related to the research article entitled "Transgenerational inheritance of chemical-induced signature: a case study with simvastatin [1].

Specifications Table

Subject	Environmental Science; Pollution
Specific subject area	Transcriptomics, Ecotoxicology, Hazard and risk assessment
Type of data	Tables Figures
How data were acquired	Illumina Hiseq 2500
Data format	Raw sequencing data Analyzed data
Parameters for data collection	Mature female amphipods (*Gammarus locusta*)
Description of data collection	*Gammarus locusta* females derived from a permanent stock culture settled at the Interdisciplinary Centre of Marine and Environmental Research (CIIMAR), periodically renewed with animals collected in the south margin of the Sado estuary (38°27Ń, 08°43Ẃ), Portugal.
Data source location	Institution: Interdisciplinary Centre of Marine and Environmental Research (CIIMAR) City/Town/Region: Matosinhos / Porto Country: Portugal
Data accessibility	All versions of the transcriptome can be consulted in the following link: https://figshare.com/s/1110e0d14fcc6a275acb. Raw data of RNA Seq analysis are available on Sequence Read Archive (SRA) database and connected with BioProject PRJNA600472 (https://www.ncbi.nlm.nih.gov/sra/?term=SRP241176); other data are presented with the article.
Related research article	T. Neuparth, A.M. Machado, R. Montes, R. Rodil, S. Barros, N. Alves, R. Ruivo, L.F.C. Castro, J.B. Quintana, M.M. Santos, Transgenerational inheritance of chemical-induced signature: a case study with simvastatin, Environment International. DOI: 10.1016/j.envint.2020.106020

Value of the Data

- This data article reports a comprehensive transcriptome dataset of *Gammarus locusta* females transgenerational exposed to one of the most prescribed human pharmaceuticals – Simvastatin.
- The RNA seq raw datasets will be valuable for ecotoxicologists to support the identification of transcripts for *Gammarus locusta* related species with unavailable transcriptome; and can be used to design probes and primers for gene expression.
- The gene expression dataset here presented can be used as a reference to explore all the metabolic pathways affects by simvastatin.
- The entire data set can be useful to improve the hazard and risk assessment of contaminant-induced transgenerational effects and thus support integrating transgenerational inheritance into risk assessment frameworks.

1. Data Description

This dataset contain data on the Transcriptomic analysis of *Gammarus locusta* females transgenerational exposed to simvastatin across generations (F0 to F3). RNA-seq was performed independently in three individual *G. locusta* females, per group, over four generations. Total RNA was extracted using the Illustra RNAspin Mini RNA Isolation Kit, and sequenced by Illumina Hiseq 2500. The raw datasets were cleaned, de novo assembled, annotated and the differential gene expression and KEGG pathways analyzed. The data consists of 16 excel tables and figures (Annex 1–16) that are provided as a supplementary file to this publication. Additionally, the quantification of simvastatin levels in the water is presented as Table A in the Section 2.2. This report presents data from the research article entitled "Transgenerational inheritance of chemical-induced signature: a case study with simvastatin, Environment International."

2. Experimental design, materials and methods

2.1. Amphipods collection and culturing

G. locusta, used in this article, were derived from a permanent stock culture settled at the Interdisciplinary Centre of Marine and Environmental Research (CIIMAR) since 2009, periodically renewed with animals collected in the south margin of the Sado estuary (38°27Ń, 08°43Ẃ), Portugal, a site with a wealthy zoobenthic community without direct exposure to contamination sources [2]. The *G. locusta* culturing is established at 18–20 °C in aquaria with natural filtered seawater (33–35‰). A sediment layer of about 1 cm and small stones, to simulate the natural habitat of this species, were placed in each aquarium. The amphipods are feed with the macroalgae *Ulva sp*. The sediment, stones and food are periodically collected from a beach in the north of Portugal devoid of direct contamination sources (Aguda beach, Portugal – 41°02′55.2"N 8°39′16.6"W). The culturing is developed in a semi-static system with water renewed twice a week to clean the system and distribute the animals by size. During this process, the water is sieved through a battery of screens of decreasing mesh size (1500, 1000, 450 and 250 μm) to separate the animals in four size classes (offspring, juveniles, sub-adults and adults).

2.2. Analytical quantification of Simvastatin (SIM) in water by liquid chromatography–tandem mass spectrometry

The quantification of Simvastatin (SIM) levels in the water was determined in each treatment of the exposed and transgenerational groups twice during each generation, at zero hours after one of the water changes (time 0) and at 72 h, immediately before the next water change (Table A). The water was sampled at mid-column of each aquarium (100 ml per replicate). Thus, two samples of seawater per treatment (bulk samples from each treatment replicate) were collected and stored at -20°C until quantification by solid-phase extraction (SPE) and Liquid Chromatography–Tandem Mass Spectrometry (LC–MS/MS) following the protocol described in Barros et al. [3]. Mevastatin (MEV) was used as Internal Standard. In addition to SIM and MEV, the quantification of their corresponding hydroxyacid forms (SIMHA and MEVHA) was also performed in order to find out the total amount of SIM in the samples. The final amount of SIM in water was calculated as sum of both lactone and hydroxyacid forms.

Table A
Measured concentrations of SIM in water samples collected in duplicate from each treatment in the four generations before and after one of the water renewal, at 0 and 72 h. Data are expressed as mean ± standard error.

Time	Generation	Solvent control	SIM 32E 32 ng/L[a]	SIM 64E 64 ng/L[a]	SIM 320E 320 ng/L[a]	SIM 64T	SIM 320T
T 0 h	F0	n.d.[b]	35.2 ± 2.4	78.9 ± 0.7	345.2 ± 22.8	–	-
T 72 h (before water renew)		n.d.[b]	25.4 ± 0.7	58.7 ± 2.9	309.1 ± 15.5	–	-
T 0 h	F1	n.d.[b]	-	73.8 ± 1.8	254.1 ± 52.6	n.d.[b]	n.d.[b]
T 72 h (before water renew)		n.d.[b]	-	37.0 ± 7.1	74.2 ± 13.3	n.d.[b]	n.d.[b]
T 0 h	F2	n.d.[b]	-	72.5 ± 8.8	sl [c]	n.d.[b]	n.d.[b]
T 72 h (before water renew)		n.d.[b]	-	27.2 ± 2.1	sl [c]	n.d.[b]	n.d.[b]
T 0 h	F3	n.d.[b]	-	84.2 ± 0.8	247.2 ± 5.5	n.d.[b]	n.d.[b]
T 72 h (before water renew)		n.d.[b]	-	17.3 ± 0.1	45.0 ± 11.5	n.d.[b]	n.d.[b]

[a] nominal concentrations;
[b] below detection limit (n.d.);
[c] not determined due to technical problem – samples were lost (sl); (-) not measured

Briefly, 100 mL of sample were spiked with the IS and passed through a 60 mg Oasis HLB (Waters) cartridge, after washing and drying the cartridge, the analytes were recovered in 5 mL of MeOH. These methanolic extracts were concentrated to 100 µL and injected in the LC-MS/MS (Varian) system. The chromatographic separation was carried out in a Luna C18 (Phenomenex) column using a gradient program with acidified (formic acid 0.1%) water and methanol as mobile phases. Two multiple-reaction monitoring (MRM) transitions were used for each compound as quantifier and qualifier respectively. These transitions were (precursor > fragment ion, m/z values): 441 > 325 and 419 > 199 for SIM, 435 > 319 and 435 > 341 for SIMHA, 413 > 311 and 391 > 185 for MEV and finally, 407 > 305 and 407 > 327 for MEVHA. Quantification was performed by matrix-matched calibration, using real seawater spiked with SIM in a calibration range between 5 and 1000 ng/L. The determination coefficient (R2) for the sum of SIM and SIMHA was 0.9964. The trueness and precision of the method, evaluated as recovery (R) and relative standard deviation (RSD), were 91% and 8% (at 50 ng/L). The method limits of quantification (MQLs) for SIM and SIMHA were 1.5 and 1.2 ng/L, respectively. SIM was not detected in solvent control group. SIM concentrations at T0h were close to the nominal ones. At T72h, total SIM concentration decays, but was still present in all treatments, except solvent control and transgenerational groups (Table A).

2.3. In silico clean-up of raw datasets and de novo transcriptome assembly

The 563 M of PE sequenced reads were analysed. The FastQC (v.0.11.8) software (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) was used to provide a complete quality report of each dataset and to select the best parameters (LEADING:5 TRAILING:5 SLIDINGWINDOW:4:5 MINLEN:36) to use in Trimmomatic program (v.0.38) [4]. Notwithstanding, the Trimmomatic tool removed adaptors and low-quality reads, while establishing the minimum quality threshold across all datasets. The Rcorrector (v.1.0.3) [5] and Centrifuge (v.1.0.3-beta) [6] were then applied. The centrifuge filter was applied in two steps. Firstly, each read dataset was taxonomically classified against the nucleotide database of NCBI (nt-NCBI) using the previously configured centrifuge index (/ftp.ccb.jhu.edu/pub/infphilo/centrifuge/data/nt_2018_3_3.tar.gz) and the following parameters (–exclude-taxids 6656 –min-hitlen 50). Second, the datasets and all reads were filtered and classified as belonging to other phylum than Arthropoda were removed with a house shell script.

To perform the reference transcriptome assembly several methods were considered, however no genomic and transcriptomic datasets or any references are available online for *G. locusta* species. Thus, *de* novo assembly method, Trinity (v.2.8.4) [7], was chosen following the protocol applied by Haas and co-workers [8]. To run Trinity, all samples were concatenated in two files (*_1.fastq / *_2.fastq) and inputted into the assembler with Non-standard parameters, SS_lib_type; RF; max_memory 245G.

At the end of the clean-up process, the FastQC tool showed that about 15 % of the total number reads were removed, remaining 559 M PE reads to use for further analyses (Annex 1 and Annex 2). All the clean read datasets were submitted to Sequence Read Archive (SRA) database of NCBI, and can be consulted under the Bio project number PRJNA600472 (https://www.ncbi.nlm.nih.gov/sra/?term=SRP241176).

2.4. Decontamination, open reading frame (ORF) prediction, and assessment of transcriptome assembly

The transcriptome assembly decontamination and assessments were done using three distinct approaches. The blast-n (v.2.9.0) searches were conducted in two databases, nt-NCBI (downloaded directly from NCBI at 30/03/2019) and UniVec (downloaded directly from NCBI in 02/04/2019) databases, using the parameters: nt-NCBI - evalue 1e-5; -max_target_seqs 1; -perc_identity 90; -max_hsps 1; and minimum alignment length of 100 bp; and UniVec -reward 1; -penalty -5; -gapopen 3; -gapextend 3; -dust yes; -soft_masking true; -evalue 700; -searchsp 1750000000000. The conserved parameters applied in nt-NCBI database allowed the detection of biological cross contaminations and removal of all contigs with match hit in any species out of Arthropoda phylum. On the other hand, the UniVec database was used to remove sequences of nucleic acids with vector origin (all sequences with match hit were removed). At the end of the decontamination/validation step, the first version of transcriptome assembly was generated (1st version - Raw transcriptome assembly (Annex 2).

Next, the open reading frames from the raw assembly were predicted using the TranDecoder software (v.5.3.0) [8], following the author guidelines (https://transdecoder.github.io/). Briefly, blast-p (v.2.9.0) searches were applied against Swiss-Prot database (downloaded directly from UniProtKB database in 12/04/2019) [9], and protein profile searches, with hmmer2 package (v.2.4i) [10], against PFAM database (downloaded directly from https://pfam.xfam.org/ in 12/04/2019) [11]. Further, all searches were integrated in one last ORF prediction, being applied a final cut-off of 100 amino acids, and all ORF's – complete, '5 and 3 prime partial or internal considered as valid. The ORF predictions allowed to produce the second version of transcriptome (2nd version - Protein coding assembly – All transcripts with ORF (Annex 2)), mainly focused on protein-coding transcripts and using a similar approach to Dylus et al. [12]. Importantly, only 15.61% of the initial transcripts matched coding ORFs with 100 or more amino acids. As ex-

pected, this percentage is relatively low due to the intraspecific variation of the 24 individuals used to generate the transcriptome assembly.

To further reduce the level of redundancy in the dataset, a third version of transcriptome assembly was generated (3rd version - Protein coding assembly – Unigenes with ORF (Annex 2)). To that aim, "unigenes" (groups of transcripts clustered by shared sequence content) were collected from the Protein coding assembly considering the clustering previously done using the Trinity assembler. Contrarily to the Trinity pipeline, which classifies the longest nucleotide sequence per cluster as "Unigene", for that propose "Unigenes" were classified as the sequence per cluster with the longest predicted ORF. In the end of this process, the three versions of *G. locusta* transcriptome assembly were inspected with the BUSCO analyses, against three libraries profiles (Eukaryota, Metazoa and Arthropoda) and general quality metric scripts of Trinity and Transrate softwares. The final version of transcriptome assembly yield 106,093 transcripts with an N50 sequence length of 2371bp, mean sequence length 1343.98 bp and a longest sequence of 43269 bp. Reinforcing the high quality of the transcriptome, the Buscos searches of complete, fragmented and missing single-copy orthologs showed a high level of gene completeness. All transcriptomes showed between 99.1–99.7%, 97.0–97.5%, and 97.5–98.0% of total Buscos found, in Eukaryota, Metazoa and Arthropoda lineage-specific profile libraries, respectively (Annex 2). These values are comparable, or in some cases better, than other *Gammarus sp*. transcriptomes, already published (e.g. [13–17]). All versions of the transcriptome were submitted to the online Figshare repository and can be consulted in the following link: https://figshare.com/s/1110e0d14fcc6a275acb.

2.5. Transcriptome annotation

The transcriptome annotation was performed using the Trinotate tool [18]. To perform the blast searches, distinct blast tools and databases were used: 3 blast tools, blast-x, blast-n and blast-p (v.2.9.0), and four databases, nt-NCBI (downloaded directly from NCBI in 30/03/2019 and built locally with makeblastdb application), non-redundant database of proteins of NCBI (nr-NCBI, downloaded directly from NCBI, in fasta format in 12/04/2019, and built with DIAMOND software (v.0.9.24) [19], Uniref90 and Swiss-Prot databases of Uniprot [20] (downloaded directly from Uniprot at 12/04/2019 and built with makedb application of DIAMOND software). While blast-n searches were conducted in nt-NCBI database with blast-n (with both megablast and dc-megablast algorithms) and an e-value cut-off of 1e-5, the blast-p/x searches were performed against nr-NCBI, Swiss-Prot and Uniref90 databases with blast tools of DIAMOND software, using the specific settings (-p 24 -k 1 -b 4 -e 1e-5 –more-sensitive). To integrate the information of all functional annotations in a single report, the Trinotate-provided SQLite Database template (downloaded and built in 12/4/2019) was initially used, which generated the .xls report of the first results (blast-x/p of Uniref90 and Swiss-Prot outputs as well as PFAM, GO terms and eggnog results) with an e-value cut-off of 1e-5. After that, with in-house shell scripts the remaining databases results (Blast-x/p of nr-NCBI database and Blast-n against nt-NCBI database (megablast and dc-megablast algorithms)) were added to the final .xls report. Notwithstanding, and using this report as input, a subset report to the third version of transcriptome assembly was produced.

After completion of the functional annotation, about 65022 of 106093 transcripts (3rd version of transcriptome assembly) were annotated in at least one of the analysed databases. The blast-x analysis with the nr-NCBI database generated the higher number of hits (61967), while the blast-n (megablast) yielded the lowest number of annotated transcripts (16642) (Annex 2). Given that few genomic resources are available for the Crustacea subphylum, annotated genomes such as *Penaeus vannamei* (Acession Number of NCBI RefSeq database - GCF_003789085.1) [21] and *Hyalella azteca* (Acession Number of NCBI RefSeq database - GCF_000764305.1) [22] were crucial to perform the correct identification and validation of *G. locusta* transcriptomic sequences. Notwithstanding, in nr/nt-NCBI databases, the majority of blast-x/p/n hits were against these

species. The full annotation report and the subset report can be consulted in the Figshare repository and using the following link: https://figshare.com/s/1110e0d14fcc6a275acb.

2.6. Differential gene expression analyses

The differential gene expression (DEGs) analyses were performed with the 2nd transcriptome assembly version and the clean reads. The first part of the DGE analyses were performed with the Trinity the pipeline scripts (abundance_estimates_to_matrix.pl script, under the defaults (Haas et al., 2013)). The Bowtie2 (v.2.3.5) [23] software was applied to map the reads and the RSEM (v.1.3.0) [24] tool to estimate the transcript abundance. At the end of this process two matrices of counts, in TPM (transcript per million), were generated, one at transcript level and another at gene level. Importantly, the gene level counts were obtained from the sum of transcript expected counts, in TPM's, of each transcript belonging to the same gene (gene and transcript clustering was done previously by Trinity pipeline). Posteriorly, the matrix of counts (gene level) was imported to the Degust (v.4.1.1) [25] platform (http://degust.erc.monash.edu/) and all genes with less than one count per million mapped reads in at least two samples were filtered out. To calculate the DGE, the edgeR (v.3.26.8) package (Robinson et al., 2010) of R (v.3.6.1) was used along the normalization scale with the trimmed mean of M-values (TMM) method [26]. The multidimensional scaling (MDS) plot was applied to analyse the variance between samples and generations. Importantly, only filtered and normalized data were used in MDS and DEGs analyses. To determine DEG, the exposed or transgenerational samples of each generation were compared against the respective control (F0.C vrs F0.320E; F1.C vrs F1.320T; F2.C vrs F2.320T; F3.C vrs F3.320T), and all genes with False Discovery Rate – corrected (FDR) p-value < 0.05 and log2|fold change| ≥ 2 were considered differentially expressed. In addition, to have a broad overview of the DGEs genes across all generations, three additional and conservative thresholds were applied, (FDR) p-value < 0.01 and log2|fold change| ≥ 2, (FDR) p-value < 0.001 and log2| | ≥ 4, (FDR) p-value < 0.0001 and log2|fold change| ≥ 4. In the end, all the heatmaps were done using the Heatmapper tool [27] with the clustering method (Average Linkage) and the Distance measurement method (Pearson) applied to the rows of the dendrogram.

The use of relaxed parameters (p-value < 0.05 and fold change ≥ 2) in DEGs analyses allowed the identification of a high quantity of genes differentially expressed in all generations, F0 = 1482, F1 = 2753, F2 = 90, F3 = 2017 (Annex 3 to Annex 9). All generations exhibited three to nine times more down-regulated genes than up- regulated genes. The ratio of down/up-regulated genes is even clearer in DGEs conservative analyses (p-values < 0.01, 0.001, 0.0001, and fold changes ≥ 2, 4), at least in F0, F1 and F3 generations (Annex 3 and Annex 9). In contrast, the F2 generation shows a distinct pattern when compared to F0, F1 and F3, presenting a lower number of differentially expressed genes when considering p-value < 0.05 and fold change ≥ 2 and no genes in conservative analyses ((p-values < 0.01, 0.001, 0.0001, and fold changes ≥ 2, 4). Furthermore, the same pattern is revealed by multidimensional scaling analysis where one of the control replicates (F2.C.R2) clusters with transgenerational and exposed samples while one of the transgenerational samples (F2.320T.R3) clusters with the control samples. The obtained results for the F2 generation are unexpected, although no experimental or analytical error have been detected. Notwithstanding, the remaining samples clustered as expected, with a clear segregation between control and transgenerational or exposed samples (Annex 10). This is evident between F0 and F3, which show the transgenerational effects of simvastatin on *G. locusta*. Importantly, and despite the different pattern of the F2 generation, 51 differentially expressed unigenes were found across all generations, 728 in at least 3 generations (F0, F1 and F3) and 1778 in two or more generations (Annex 11). To analyze the gene content in detail, the functional annotation was used, which showed 1088 (F0), 2012 (F1), 74 (F2), 1465 (F3) genes with one hit in at least one database. In addition, further comparisons between generations F0 and F3 (F0-F3) were performed. Importantly, from a total of 806 common and differentially expressed genes, 765 and 38 genes kept the down and up-regulation patterns from F0 to F3, while 1 and 2 genes changed their expression pattern from up to down or *vice versa* (Annex 8 and Annex 11).

The high rate of annotated DGEs allowed the identification of other important families of differentially expressed genes in the four generations (e.g., genes related with cuticle metabolism – Annex 12 or with epigenetic regulation – Annex 13) using an e-value cut-off of 1e-45.

2.7. KEGG pathways analyses

To identify the main metabolic pathways affected by the transgenerational effects of SIM across all generations, the DGEs genes were blasted onto KEGG pathways database using the KAAS webserver [28]. The blast was performed with the single-directional best hit (SBH) method, the proteins corresponding to the differential expressed genes (collected from the third version of the transcriptome), and 710,890 reference sequences of 40 manually selected species (Annex 14).

All KEGG Orthology (KO) results and KEGG pathways, per generation, were manually scrutinized and analyzed using the Trinotate functional annotation (Annex 15). The subsequent analyses focused on the understanding of the transgenerational effects in pathways and genes commonly expressed in F0–F3 generations. Taking into account the large number of genes overlapping both generations, the search was restricted to annotated genes in some pathways of interest (Carbohydrate metabolism – Glycolysis/gluconeogenesis, Pentose phosphate pathway, Pentose and glucoronate interconversions, Fructose and mannose metabolism, Galactose metabolism and Amino sugar and nucleotide sugar metabolism; Metabolism of terpenoids and polyketides – Mevalonate pathway and sesquiterpenoid biosynthesis/degradation, Ecdysteroid biosynthesis; Lipids metabolism – Biosynthesis of unsaturated fatty acids metabolism) using an e-values cut-off of 1e-45 (Annex 16). All the selected gene sequences were re-validated via protein and nucleotide alignment against several other species, and re-blasted in the nt/nr-NCBI database. In the end of this analysis, we collected KEGG assignments to 756 of 3785 proteins (DGEs in F0, F1, F2 and F3). In F0-F3 proteins we found about of 75 KEGG ID's (The KEGG IDs can be consulted in Annex 3 to Annex 8 and Annex 12, 13 and 16).

Ethics statement

All experiments have been approved by the CIIMAR ethical committee and by CIIMAR Managing Animal Welfare Body (ORBEA) according to the European Union Directive 2010/63/EU on the protection of animals used for scientific purposes.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This article was developed under the Transobesogen project - Trans-phyletic obesogenic responses: from epigenetic modules to transgenerational environmental impacts (reference PTDC/CTA-AMB/31544/2017 - NORTE-01-0145-FEDER-031544), cofunded by Portugal 2020, the European Union through the ERDF and by the Portuguese Foundation for Science and Technology – FCT. This article was also supported by FCT through national funds (UIDB/04423/2020; UIDP/04423/2020), by the Spanish Agencia Estatal de Investigación (CTM2017-84763-C3-2-R) and by the Galician Council of Culture, Education and Universities (ED431C2017/36), cofounded by ERDF. A PhD grant awarded to Susana Barros acknowledges the doctoral grant attributed by FCT with reference PD/BD/143090/2018.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.dib.2020.106248.

References

[1] T. Neuparth, A.M. Machado, R. Montes, R. Rodil, S. Barros, N. Alves, R. Ruivo, L.F.C. Castro, J.B. Quintana, M.M. Santos, Transgenerational inheritance of chemical-induced signature: a case study with simvastatin, Environ. Int.. DOI 10.1016/j.envint.2020.106020.

[2] T. Neuparth, F.O. Costa, M.H. Costa, Effects of temperature and salinity on life history of the marine amphipod *Gammarus locusta*. Implications for ecotoxicological testing, Ecotoxicology 11 (2002) 61–73 https://doi.org/10.1023/a:1013797130740.

[3] S. Barros, R. Montes, J.B. Quintana, R. Rodil, A. André, A. Capitão, J. Soares, M.M. Santos, T. Neuparth, Chronic environmentally relevant levels of simvastatin disrupt embryonic development, biochemical and molecular responses in zebrafish (*Danio rerio*), Aquat. Toxicol. 201 (2018) 47–57 https://doi.org/10.1016/j.aquatox.2018.05.014.

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

[5] L. Song, L. Florea, Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads, Gigascience 4 (2015) 48 https://doi.org/10.1186/s13742-015-0089-y.

[6] D. Kim, L. Song, F.P. Breitwieser, S.L. Salzberg, Centrifuge: rapid and sensitive classification of metagenomic sequences, Genome Res. 26 (2016) 1721–1729 https://doi.org/10.1101/gr.210641.116.

[7] M. Grabherr, B. Haas, M. Yassour, J.Z. Levin, D.A. Thompson, I. Amit, et al., Full-length transcriptome assembly from RNA-seq data without a reference genome, Nat. Biotechnol. 29 (2011) 644–652 https://doi.org/10.1038/nbt.1883.

[8] B. Haas, A. Papanicolaou, M. Yassour, M. Grabherr, P.D. Blood, J. Bowden, et al., De novo transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis, Nat. Protoc. 8 (2013) 1494–151 https://doi.org/10.1038/nprot.2013.084 .

[9] The UniProt Consortium, UniProt: the universal protein knowledgebase, Nucleic Acids Res. 45 (2017) D158–D169 https://doi.org/10.1093/nar/gkw1099.

[10] R.D. Finn, J. Clements, S.R. Eddy, HMMER web server: interactive sequence similarity searching, Nucleic Acids Res. 39 (2011) W29–W37 https://doi.org/10.1093/nar/gkr367.

[11] M. Punta, P.C. Coggill, R.Y. Eberhardt, J. Mistry, J. Tate, C. Boursnell, et al., The Pfam protein families database, Nucleic Acids Res. 40 (2012) D290–D301 https://doi.org/10.1093/nar/gkr1065.

[12] D.V. Dylus, A. Czarkwiani, L.M. Blowes, M.R. Elphick, P. Oliveri, Developmental transcriptomics of the brittle star *Amphiura filiformis* reveals gene regulatory network rewiring in echinoderm larval skeleton evolution, Genome Biol. 19 (2018) 26 https://doi.org/10.1186/s13059-018-1402-8.

[13] D.R. Caputo, S.C. Robson, I. Werner, AT. Ford, Complete transcriptome assembly and annotation of a critically important amphipod species in freshwater ecotoxicological risk assessment: *Gammarus fossarum*, Environ. Int. 137 (2020) 105319 https://doi.org/10.1016/j.envint.2019.105319.

[14] D.B. Carlini, D.W. Fong, The transcriptomes of cave and surface populations of *Gammarus minus* (Crustacea: Amphipoda) provide evidence for positive selection on cave downregulated transcripts, PLoS One 12 (2017) e0186173 https://doi.org/10.1371/journal.pone.0186173.

[15] Y. Cogne, D. Degli-Esposti, O. Pible, D. Gouveia, A. François, O. Bouchez, et al., De novo transcriptomes of 14 gammarid individuals for proteogenomic analysis of seven taxonomic groups, Sci. Data 6 (2019) 184 https://doi.org/10.1038/s41597-019-0192-5.

[16] S. Jin, C Bian, S. Jiang, S. Sun, L. Xu, Y. Xiong, et al., Identification of candidate genes for the plateau adaptation of a tibetan amphipod, *Gammarus lacustris*, through integration of genome and transcriptome sequencing, Front. Genet. 10 (2019) 53 https://doi.org/10.3389/fgene.2019.00053.

[17] M. Truebano, O. Tills, J.I. Spicer, Embryonic transcriptome of the brackishwater amphipod Gammarus chevreuxi, Mar. Genom. 28 (2016) 5–6 https://doi.org/10.1016/j.margen.2016.02.002.

[18] DM Bryant, K Johnson, T DiTommaso, T Tickle, MB Couger, D Payzin-Dogru, et al., A Tissue-Mapped axolotl de novo transcriptome enables identification of limb regeneration factors, Cell Rep. 18 (2017) 762–776 PMID: 28099853 https://doi.org/10.1016/j.celrep.2016.12.063 .

[19] B. Buchfink, C. Xie, D. Huson, Fast and sensitive protein alignment using DIAMOND, Nat. Methods 12 (2015) 59–60 https://doi.org/10.1038/nmeth.3176.

[20] The UniProt Consortium, UniProt: a worldwide hub of protein knowledge, Nucleic Acids Res. 47 (2019) D506–D515 https://doi.org/10.1093/nar/gky1049.

[21] X. Zhang, J. Yuan, Y. Sun, Y.S. Li, Y. Gao, Y. Yu, et al., Penaeid shrimp genome provides insights into benthic adaptation and frequent molting, Nat. Commun. 10 (2019) 356 https://doi.org/10.1038/s41467-018-08197-4.

[22] H.C. Poynton, S. Hasenbein, J.B. Benoit, M.S. Sepulveda, M.F. Poelchau, D.S.T. Hughes, et al., The Toxicogenome of *Hyalella azteca*: a model for sediment ecotoxicology and evolutionary toxicology, Environ. Sci. Technol. 52 (2018) 6009–6022 https://doi.org/10.1021/acs.est.8b00837.

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

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

[25] D. Powell, M. Milton, A. Perry, K. Santos, Degust: Interactive RNA-seq Analysis, 2015 https://doi.org/10.5281/zenodo.3258932.

[26] M.D. Robinson, A.A. Oshlack, Scaling normalization method for differential expression analysis of RNA-seq data, Genome Biol. 11 (2010) R25 https://doi.org/10.1186/gb-2010-11-3-r25.
[27] S. Babicki, D. Arndt, A. Marcu, Y. Liang, J.R. Grant, A. Maciejewski, D.S. Wishart, Heatmapper: web-enabled heat mapping for all, Nucleic Acids Res. 44 (2016) W147–W153 https://doi.org/10.1093/nar/gkw419.
[28] Y. Moriya, M. Itoh, S. Okuda, A.C. Yoshizawa, M. Kanehisa, KAAS: an automatic genome annotation and pathway reconstruction server, Nucleic Acids Res. 35 (2007) W182–W185 https://doi.org/10.1093/nar/gkm321.