De novo transcriptomes of 14 gammarid individuals for proteogenomic analysis of seven taxonomic groups

Yannick Cogne1, Davide Degli-Esosti2, Olivier Pible1, Duarte Gouveia1, Adeline François2, Olivier Bouchez3, Camille Eché3, Alex Ford4, Olivier Geffard2, Jean Armengaud1, Arnaud Chaumot2 & Christine Almunia1

Gammarids are amphipods found worldwide distributed in fresh and marine waters. They play an important role in aquatic ecosystems and are well established sentinel species in ecotoxicology. In this study, we sequenced the transcriptomes of a male individual and a female individual for seven different taxonomic groups belonging to the two genera Gammarus and Echinogammarus: Gammarus fossarum A, G. fossarum B, G. fossarum C, Gammarus wautieri, Gammarus pulex, Echinogammarus berillonii, and Echinogammarus marinus. These taxa were chosen to explore the molecular diversity of transcribed genes of genotyped individuals from these groups. Transcriptomes were de novo assembled and annotated. High-quality assembly was confirmed by BUSCO comparison against the Arthropod dataset. The 14 RNA-Seq-derived protein sequence databases proposed here will be a significant resource for proteogenomics studies of these ecotoxicologically relevant non-model organisms. These transcriptomes represent reliable reference sequences for whole-transcriptome and proteome studies on other gammarids, for primer design to clone specific genes or monitor their specific expression, and for analyses of molecular differences between gammarid species.

Background & Summary
Gammarid amphipods are animals that typically measure a few millimetres long and present in a wide range of aquatic habitats1. In freshwater ecosystems, they are often the most dominant macro-invertebrates, representing a significant proportion of the total biomass, and they also play a central role within food webs. Indeed, they are a prey for many species, but are also predators for many invertebrate species. Amphipods are also scavengers and shredders, and detritivores involved in leaf litter breakdown, playing a central role in the decomposition of organic matter in general. Thus, they modulate the composition of freshwater communities of invertebrates2. Thanks to these essential roles, they have been the subject of many recent studies investigating their sensitivity to pollutants3–7.

Marine and freshwater resources are of the utmost importance for Life. Human-made chemical contaminants released into aquatic environments compromise the quality of water bodies, threatening the resident biodiversity, and the utility of such ecosystems. The quality of these environments should be evaluated not only by measuring the concentrations of pollutants present, but also by monitoring how Life is affected by the bioavailable pollutants present and their synergistic/antagonist effects8. To do this, biomonitoring with caged representative sentinel species has proved to be a valuable tool for efficient ecotoxicological studies9–13. Specific traits such as moult delay, growth impairment, or reproductive defects can be monitored on sensitive animals exposed to toxic environments. These data can be then integrated into a quantitative water quality index that can be used by

1Laboratoire Innovations technologiques pour la Détection et le Diagnostic (Li2D), Service de Pharmacologie et Immonuanalyse (SPI), CEA, INRA, F-30207, Bagnols-sur-Cèze, France. 2Irstea, UR MAL Y Laboratoire d’écotoxicologie, centre de Lyon-Villeurbanne, F-69625, Villeurbanne, France. 3GeT-PlaGe, Genotoul, INRA Auzeville, F-31320, Castanet-Tolosan, France. 4School of Biological Sciences, Institute of Marine Sciences Laboratories, PO4 9LY, Portsmouth, United Kingdom. Correspondence and requests for materials should be addressed to J.A. (email: jean.armengaud@cea.fr)
stakeholders in charge of the aquatic ecosystem and water resource management.14 Because of their abundance and central ecological roles, invertebrates are commonly employed as test organisms in marine and ecotoxicological assessments. Specifically, gammarids have been successfully used as sentinel species for freshwater ecosystems following investigations of their physiological responses to toxicants15–18 and biomonitoring in caging systems19–21. Specific biomarkers have been proposed and can be monitored by innovative methods such as tandem mass spectrometry.22–24. Next-generation proteomics contributed to improving our knowledge of the molecular responses of gammarids to toxicants, and led to the proposal of a broad panel of appropriate biomarkers25–28. This approach was successful after developing a protein sequence database from an RNA-Seq transcriptome translated in all the possible reading frames. This proteogenomics concept was used to establish an extensive catalogue of protein sequences comprising 1873 mass-spectrometry-certified proteins, thus representing a significant amphipod proteomic resource.29

Despite this progress, molecular resources relating to gammarids remain scarce.30 No gammarid whole genome sequence was available until very recently, when a first-draft genome of Gammarus lacustris was released comprising 443,304 scaffolds.32 The genomes of two related amphipods, Parhyale hawaiensis33 and Hyalella azteca34, have also been sequenced. RNA-Seq datasets are now available for P. hawaiensis35–37, Echinogammarus marinus38, Eogammarus possetticus39, Gammarus fossarum40, Gammarus chevreuxi41, Gammarus pulex42, and Gammarus minus43. However, these datasets are not of equal quality in terms of mRNA sequence coverage, which is a crucial parameter for proteogenomics interpretation.44 They are assembled from mRNAs extracted from a pool of several animals or from specific tissues, and in some cases are no longer accessible as it is the case for E. marinus because the repository used no longer exists.45

The data presented in this article consist of assembled transcriptome sequences for 14 different gammarids, seven males and seven females, namely Gammarus fossarum A (Müller type A), G. fossarum B (Müller type B), G. fossarum C (Müller type C), Gammarus wautieri, Gammarus pulex, Echinogammarus berilloni, and Echinogammarus marinus. These transcriptomes were assembled and translated using the same pipelines (full length whole-organism mRNAs), and thus are of equivalent sequencing depth and quality across the different taxa studied. Starting material was extracted from single animals to avoid sequence heterogeneity. The transcriptomes have been annotated to serve as reference protein sequence databases for proteogenomics studies involving these sentinel animals that will be soon conducted to gain more basic knowledge and thus improve how aquatic environmental risks are assessed. For these future studies, an interesting strategy could be to interpret MS/MS shotgun data first on the most appropriate specific single-organism database, and then perform a follow-up search on a multi-organism database. The transcriptomes presented here will also serve in comparative analyses to better define the molecular diversity amongst gammarids and will be a valuable sequence resource for future ecotoxicological studies.

### Methods

#### Experimental design.

Freshwater gammarids were collected in four geographically-distant French rivers (Table 1). One population of Gammarus fossarum was sampled in north-eastern France (Seebach river), which was previously shown to harbour the cryptic type A subspecies according to the three types defined in Müller et al.46, Westram et al.47, and Weiss et al.48. The second river (Pollon River) situated in the mid-eastern area of France, corresponding to a sympatric situation, supplied organisms belonging to Gammarus fossarum type B, type C, and Gammarus pulex species. Gammarus wautieri were collected in the Galaveyson river in the Dauphiné region, and Echinogammarus berilloni organisms from a fourth river in south-western France (Saucats river). These freshwater gammarids were all collected using a hand net following kick-sampling, and subsequently transported to the laboratory. After maintaining them for 1 week in the laboratory – at 12°C with a constant aeration,

| Species                  | Code Name | Sex   | River    | City       | Country | GPS             | Number of raw reads | Number of reads after filtering |
|--------------------------|-----------|-------|----------|------------|---------|-----------------|---------------------|-------------------------------|
| Echinogammarus berilloni | EGSF      | Female| Saucats  | Saucats    | France  | 44°39′34″N 0°34′25″W | 80 482 966           | 80 277 434                   |
| Echinogammarus berilloni | EGSF      | Male  | Saucats  | Saucats    | France  | 44°39′34″N 0°34′25″W | 90 372 154           | 90 118 242                   |
| Echinogammarus marinus   | EGUF      | Female| sea coast| Portsmouth | UK      | 50°47′41″N 1°01′50″W | 85 032 246           | 84 652 454                   |
| Echinogammarus marinus   | EGUM      | Male  | sea coast| Portsmouth | UK      | 50°47′41″N 1°01′50″W | 70 768 994           | 70 540 528                   |
| Gammarus fossarum A*     | GFAF      | Female| Saucats  | Fellerier   | France  | 47°53′31″N 6°58′53″E  | 81 959 830           | 81 543 116                   |
| Gammarus fossarum A*     | GFAM      | Male  | Saucats  | Fellerier   | France  | 47°53′31″N 6°58′53″E  | 95 167 986           | 94 695 572                   |
| Gammarus fossarum B*     | GFBF      | Female| Pollon   | Saint-Maurice-de-Rébens | France  | 45°57′21″N 5°15′44″E  | 96 361 300           | 96 093 396                   |
| Gammarus fossarum B*     | GFBM      | Male  | Pollon   | Saint-Maurice-de-Rébens | France  | 45°57′21″N 5°15′44″E  | 85 125 996           | 84 758 816                   |
| Gammarus fossarum C*     | GFCF      | Female| Pollon   | Saint-Maurice-de-Rébens | France  | 45°57′21″N 5°15′44″E  | 78 459 708           | 77 977 148                   |
| Gammarus fossarum C*     | GCFC      | Male  | Pollon   | Saint-Maurice-de-Rébens | France  | 45°57′21″N 5°15′44″E  | 75 598 166           | 75 407 534                   |
| Gammarus pulex           | GPCF      | Female| Pollon   | Saint-Maurice-de-Rébens | France  | 45°57′21″N 5°15′44″E  | 84 202 086           | 83 965 920                   |
| Gammarus pulex           | GP CM     | Male  | Pollon   | Saint-Maurice-de-Rébens | France  | 45°57′21″N 5°15′44″E  | 89 235 492           | 89 024 410                   |
| Gammarus wautieri        | GWF       | Female| Galaveyson| Le Grand Serre| France  | 45°16′27″N 5°07′08″E   | 100 192 262         | 100 695 588                  |
| Gammarus wautieri        | GWM       | Male  | Galaveyson| Le Grand Serre| France  | 45°16′27″N 5°07′08″E   | 63 959 618           | 63 638 482                   |

Table 1. Sampling information and number of reads for each sample before and after filtering by mean quality for the 14 transcriptomes. *Müller type.
under a 16/8-h light/dark photoperiod in buckets containing water sampled from their respective rivers of origin, and with conditioned alder leaves as food source – couples in amplexus were isolated for species determination before RNA extraction. Pairs where the females had well-developed ovaries were selected. Embryos were removed from the marsupial pouch of females for five of these couples. Based on the description of the reproductive cycle in *Gammarus fossarum* [48], for RNA extraction, we were able to select one couple per species in the last stage of the reproductive cycle (pre-moulting stage for the female) by retaining pairs where the females were carrying embryos at the end of their embryonic development stage (stage 4 or 5). For the marine species, *E. marinus* were collected from beneath seaweed in the intertidal zone in Portsmouth, southern England. These species correspond to the same population as used in a previous study [38]. After maintaining them for 1 month in the laboratory – at 10°C under a 12 h light/12 h dark photoperiod in buckets with filtered natural seawater and fed with fucoid seaweed - organisms were transported live in damp seaweed from United Kingdom to France (one-day travel). They were subsequently maintained for a few hours in aquaria containing reconstituted seawater (salinity 30%) before organism selection. For this species, it was not possible to recover couples in amplexus. One free-swimming male and one free-swimming female were isolated from the batch of organisms available. Stage 1 embryos were recovered from the female marsupium, indicating that this female was in a post-moulting stage.

Species were first determined based on morphological criteria [49]. To distinguish between the three cryptic lineages, *A, B, C*, within the *G. fossarum* species, a molecular species assignment was carried out by amplifying the 5’ part of the mtDNA cytochrome c oxidase subunit I (COI) using universal primers (LCO1490 [GGT CAA ATC ATA AAG ATA TTG G] and HCO2198 [TAA ACT TCA GGG TGA CCA AAA AAT CA]) [50]. Briefly, DNA was extracted from one or two pereopods (depending on individual size) cut from organisms before conditioning for RNA extraction. DNA was extracted using the Nucleospin tissue XS kit (Macherey-Nagel), and 10 ng of DNA for each organism was amplified in the PCR mixer (Eppendorf). The PCR condition consisted in 94°C for 30 sec, annealing at 50°C for 30 sec, and elongation at 72°C for 1 min. PCR products were purified by ultrafiltration using the Nucleofast kit (Macherey-Nagel). Purified amplicons were prepared for sequencing using the BigDye Terminator v3.1 kit (ThermoFisher), and then sequenced on a DNA analyzer ABI 3730XL (ThermoFisher).

Sequence data were analysed using the Sequencer 5.4.6 program (Genecodes). COI sequences (freely available from figshare, YC02_COI sequences and phylogenetic tree [51]) were aligned to build a phylogenetic tree including reference sequences from Weiss et al. [47] and Lagrue et al. [52]. Using this phylogenetic tree (freely available from figshare, YC02_COI sequences and phylogenetic tree [51]) it is possible to position the COI sequences of the *Gammarus* organisms selected for RNA sequencing in relation to the published reference sequences (SeaView software [25]; BioNJ method based on J-C distance). The robustness of the different groupings was evaluated by a bootstrapping procedure (100 iterations). COI sequences were obtained for all *Gammarus* individuals, except for the female *G. fossarum* C as this individual was in pre-copulatory amplexus with the male COI-genotyped as *G. fossarum*. However, in the same location (Pollon River), we also obtained the COI genotypes for 15 additional pairs, all of which were found to be non-heterospecific (4 *G. fossarum* B, 3 *G. fossarum* C, 8 *G. pulex*).

Westraam et al. [53] reported similar findings in the Glovelier river which harbours *G. fossarum* A and B, with only one heterospecific pair for a total of 64 genotyped pairs. Lagrue et al. [52] also observed that mixed pairs are rare in the field for *Gammarus* lineages with a COI distance greater than 4%. Considering that the divergence between the COI-genotyped *G. fossarum* B and C specimens is about 17% in the Pollon River, it is very unlikely that this female does not belong to the *G. fossarum* C species.

**Dataset generation.** Gammarids were placed in RNAlater (Sigma) and stored at 4°C overnight. The RNAlater was then removed, and the organisms were snap frozen in liquid nitrogen and stored at −80°C until RNA was extracted. Organisms were first homogenized in lysis buffer using a bead homogenizer and then RNAs were extracted using the Qiagen fibrous tissue kit (Qiagen). RNA quantity, quality and integrity were assessed by Nanodrop (Thermo Fisher) and Bioanalyzer (Agilent) analysis. RNA-Seq libraries were generated using the TruSeq stranded mRNA Sample Prep kit (Illumina). mRNA was purified using poly-(T) beads from 2 µg of each total RNA sample, then cleaved in segments of 155 bp on average (120–210 bp range). Subsequently, cleaved RNA fragments were primed with random hexamers and reverse-transcribed into first-strand cDNA. A second strand of cDNA was consecutively synthesized, and double-stranded cDNA was purified on beads. The 3’ ends of the blunt fragments obtained were then adenylated. Indexed adapters were ligated to the PCR-enriched cDNA fragments (11 cycles). Libraries were purified and quality-assessed using a Fragment Analyzer (Advanced Analytical Technologies). The 16 libraries were quantified by qPCR using the Kapa Library Quantification Kit (Roche). Their concentrations were normalized, multiplexed in a single pool. Libraries were then sequenced on two lanes of HiSeq3000 (Illumina) using a paired-end read length of 2 × 150 bp with the HiSeq Reagent Kits (Illumina). The two HiSeq lanes produced an average of 40.0 ± 8 million read pairs per library. Quality control of reads was performed by FastQC version V0.11.2 (Babraham Bioinformatics). Detailed results are freely available from figshare (YC02_QC data [31]). The data records are stored in 14 folders, each containing four folders per transcriptome.

**De novo assembly.** For each sample, the forward or reverse reads were merged from two separate lanes. Data were filtered based on the mean Qphred score, with a threshold set at 16.99, and any remaining unpaired reads were removed using a homemade script. The numbers of reads for each sample before and after this filtering step are presented in Table 1. Trinity v2.4 [44] was used to assemble reads for each sample considering pair-end and strand orientation (-SS_lib type RF); all other Trinity parameters were set to their default values, with k set to 25, and minimum contig length to 200 bp.

**Assessing assembly quality.** Transcriptome quality was assessed using Transrate v1.0.1 [55], which generates standard metrics and remapping statistics. No reference protein sequences were used for the assessment with Transrate. The main metrics are shown in Table 2. To validate the quality of all the assemblies, BUSCO v2.0 [56] was
The database used for BUSCO analyses was Arthropoda_odb9 which contains 1066 orthologous genes at the nearest taxon level (i.e., Arthropods) available for *Gammarus*.

For each sample, the transcripts were annotated using the Trinotate v3.1.1 annotation pipeline54. The Swissprot database was used as the main database, and amphipod proteins referenced on Uniref were used as a custom database. Similarity searches were performed with Blastx and Blastp, with an e-value cutoff set at 1e-2. Results from these searches were then used to generate the annotation report with the same e-value cutoff.

### Data Records

**Reads.** Read sequences for each sample were deposited in the NCBI Sequence Reads Archive under accession Numbers SRR808972057, SRR8089722–SRR8089725, and SRR8089727–SRR8089735, as indicated in Table 3 alongside the corresponding Bioproject and Biosample codes. The FastQC results for the 14 samples are freely available from figshare (YC02_QC data)51. The data records are stored as 14 folders, each of which contain four folders per transcriptome.

**Transcriptomes.** Transcriptome assemblies were deposited in the NCBI Transcriptome Shotgun Assembly Sequence Database. These data have been deposited in GenBank under identifiers GHCN0100000071, GHCP0100000072–74, GCTO01000000–GHCZ01000000, GHA01000000–GHD01000000, as indicated in Table 3 alongside the corresponding Bioproject and Biosample codes.

**Proteogenomics databases.** Translations of coding sequence regions were produced for each transcriptome from stop to stop codons by Transdecoder v3.0.154, analysing only the top strand. The 500 longest ORFs

| Code | Name | Transcriptome accession | Read accession | BioProject | BioSample |
|------|------|--------------------------|----------------|------------|-----------|
| EGSF | GHCT01000000 | SRR8089732 | PRJNA497972 | SAMN10259946 |
| EGSM | GHCU01000000 | SRR8089733 | PRJNA497972 | SAMN10259947 |
| EGUF | GHCW01000000 | SRR8089734 | PRJNA497972 | SAMN10259948 |
| EGUM | GHC01000000 | SRR8089735 | PRJNA497972 | SAMN10259949 |
| GFAF | GHCH01000000 | SRR8089727 | PRJNA497972 | SAMN10259934 |
| GFAM | GHCT01000000 | SRR8089728 | PRJNA497972 | SAMN10259935 |
| GFBF | GHCD01000000 | SRR8089729 | PRJNA497972 | SAMN10259936 |
| GFBM | GHA01000000 | SRR8089722 | PRJNA497972 | SAMN10259944 |
| GFCF | GHD01000000 | SRR8089723 | PRJNA497972 | SAMN10259938 |
| GFCM | GHB01000000 | SRR8089724 | PRJNA497972 | SAMN10259939 |
| GPCF | GHCP01000000 | SRR8089725 | PRJNA497972 | SAMN10259940 |
| GPCM | GHCQ01000000 | SRR8089720 | PRJNA497972 | SAMN10259941 |
| GWF | GHCR01000000 | SRR8089730 | PRJNA497972 | SAMN10259944 |
| GWM | GHCN01000000 | SRR8089731 | PRJNA497972 | SAMN10259945 |

**Table 2.** Assembly quality metrics. *RMBT means Reads Mapping Back on the Transcriptome; G-RMBT means Good Reads Mapping Back on the Transcriptome. *Score calculated by Transrate.

**Table 3.** Accessions for the 14 transcriptomes.

used. The database used for BUSCO analyses was Arthropoda_odb9 which contains 1066 orthologous genes at the nearest taxon level (i.e., Arthropods) available for *Gammarus*.

**Annotation.** For each sample, the transcripts were annotated using the Trinotate v3.1.1 annotation pipeline54. The Swissprot database was used as the main database, and amphipod proteins referenced on Uniref were used as a custom database. Similarity searches were performed with Blastx and Blastp, with an e-value cutoff set at 1e-2. Results from these searches were then used to generate the annotation report with the same e-value cutoff.

**Data Records**

**Reads.** Read sequences for each sample were deposited in the NCBI Sequence Reads Archive under accession Numbers SRR808972057, SRR8089722–SRR8089725, and SRR8089727–SRR8089735, as indicated in Table 3 alongside the corresponding Bioproject and Biosample codes. The FastQC results for the 14 samples are freely available from figshare (YC02_QC data)51. The data records are stored as 14 folders, each of which contain four folders per transcriptome.

**Transcriptomes.** Transcriptome assemblies were deposited in the NCBI Transcriptome Shotgun Assembly Sequence Database. These data have been deposited in GenBank under identifiers GHCN0100000071, GHCP0100000072–74, GCTO01000000–GHCZ01000000, GHA01000000–GHD01000000, as indicated in Table 3 alongside the corresponding Bioproject and Biosample codes.

**Proteogenomics databases.** Translations of coding sequence regions were produced for each transcriptome from stop to stop codons by Transdecoder v3.0.154, analysing only the top strand. The 500 longest ORFs
were used for training, retaining 600-bp ORFs and only proteins with a minimum length of 50 amino acids. The 14 translations are freely available for download as FASTA files from figshare (YC02_Transcriptome translated ORFs51).

Annotation. Annotations of each assembly are freely available for download as Excel files from figshare (YC02_Transcript annotations51). The folder contains 14 Excel files.

Technical Validation

Transrate. Transrate analyses showed good remapping of results, with more than 80% of reads remapped and most assemblies with more than 70% were classed as well mapped. Raw results from Transrate are freely available through figshare (YC02_Transrate results)51.

BUSCO. A high level of single-copy ortholog retrieval was noted for the 14 assemblies, with at least a 75% ratio, as shown in Fig. 1. Furthermore, fewer than 8% of orthologs were missing in the worst case, and fewer than 5% were missing in 11 transcriptomes.

Code Availability
Filtering before assembly was performed with an in-house Pythonv2.7 script, which is freely available (https://github.com/YannickCogne/Qfiltering). The script was automated with a bash script for each sample.

References
1. MacNeil, C., Dick, J. T. A. & Elwood, R. W. The trophic ecology of freshwater Gammarus Spp. (crustacea:amphipoda): problems and perspectives concerning the functional feeding group concept. Biological Reviews 72, 349–364 (1997).
2. Kelly, D. W., Dick, J. T. A. & Montgomery, W. I. The functional role of Gammarus (Crustacea, Amphipoda): shredders, predators, or both? Hydrobiologia 485, 199–203 (2002).
3. Arce-Funck, J. et al. High stoichiometric food quality increases moulting organism vulnerability to pollutant impacts: An experimental test with Gammarus fossarum (Crustacea: Amphipoda). Sci Total Environ 645, 1484–1495 (2018).
4. Ganser, B. et al. Wastewater alters feeding rate but not vitellogenin level of Gammarus fossarum (Amphipoda). Sci Total Environ 657, 1246–1252 (2019).
5. Konemann, S. et al. Combination of In Situ Feeding Rate Experiments and Chemical Body Burden Analysis to Assess the Influence of Micropollutants in Wastewater on Gammarus pulex. Int J Environ Res Public Health 16, 883 (2019).
6. Munz, N. A., Fu, Q., Stamm, C. & Hollender, J. Internal Concentrations in Gammarus pulex (Gammaridae) as a robust tool for the characterization of bioavailable contamination levels in continental waters: towards the determination of threshold values. Water Res 47, 650–660 (2013).
7. Chaumot, A., Geffard, O., Armengaud, J. & Maltby, L. In Aquatic Ecotoxicology - Advancing tools for dealing with emerging risks (eds Amiard-Triquet, C., Amiard, J.-C. & Mouneyrac, C.) 253–280 (Academic Press, London, 2015).
8. Ciliberti, A. et al. Caged Gammarus as biomonitoring tools identifying thresholds of toxic metal bioavailability that affect gammarid densities at the French national scale. Water Res 118, 131–140 (2017).
51. Cogne, Y.
49. Piscart, C. & Bollache, L.
42. Carlini, D. B. & Fong, D. W. The transcriptomes of cave and surface populations of Gammarus minus (Crustacea: Amphipoda)
41. Gismondi, E. & Thome, J. P. Transcriptome of the freshwater amphipod Gammarus pulex hepatopancreas.
39. Chen, J., Liu, H., Cai, S. & Zhang, H. Comparative transcriptome analysis of Eogammarus possjeticus at different hydrostatic
38. Jayatilaka, N. & Mead, C. R. F. T. Two-dimensional gel electrophoresis of the protein components of Gammarus minus (Crustacea: Amphipoda)
36. Westram, A. M., Jokela, J., Baumgartner, C. & Keller, I. Spatial distribution of cryptic species diversity in european freshwater
35. Jones, M. & Blaxter, M. afterParty: turning raw transcriptomes into permanent resources.
33. Kao, D.
30. Trapp, J.
32. Jin, S.
31. Armengaud, J.
26. Simon, R.
25. Gouveia, D.
18. Felten, V. et al. Physiological and behavioural responses of Gammarus pulex (Crustacea: Amphipoda) exposed to cadmium. Aquat. Environ. Pollut 86, 413–425 (2008).
16. Chaumot, A., Gos, P. et al. Additive vs non-additive genetic components in lethal cadmium tolerance of Gammarus fossarum. Aquat. Toxicol 112–113, 72–82 (2012).
14. Coulaud, R. et al. In situ feeding assay with Gammarus fossarum (Crustacea): Modelling the influence of confounding factors to improve water quality biomonitoring. Water Res 45, 6417–6429 (2011).
13. Trapp, J., Armengaud, J., Salvador, A., Chaumot, A. & Geffard, O. Next-generation proteomics: toward customized biomarkers for environmental biomonitoring. Environ Sci Technol 48, 13560–13572 (2014).
12. Colaud, R. et al. In situ feeding assay with Gammarus fossarum (Crustacea): Modelling the influence of confounding factors to improve water quality biomonitoring. Water Res 45, 6417–6429 (2011).
11. Barros, S. et al. Chronic effects of triclofoncarb in the amphipod Gammarus locusta: Behavioural and biochemical impairment. Ecotoxicol Environ Saf 135, 276–283 (2017).
10. Chaumot, A., Gos, P., Garric, J. & Geffard, O. Additive vs non-additive genetic components in lethal cadmium tolerance of Gammarus (Crustacea): novel light on the assessment of the potential for adaptation to contamination. Aquat Toxicol 94, 294–299 (2009).
9. Correia, A. D., Lima, G., Costa, M. H. & Livingstone, D. R. Studies on biomarkers of copper exposure and toxicity in the marine amphipod Gammarus locusta (Crustacea): I. Induction of metallothionein and lipid peroxidation. Biomarkers 7, 422–437 (2002).
8. Felten, V. et al. Physiological and behavioural responses of Gammarus pulex (Crustacea: Amphipoda) exposed to cadmium. Aquat. Environ. Pollut 86, 413–425 (2008).
7. Jubeaux, G. et al. Vitellogenin-like proteins in the freshwater amphipod Gammarus fossarum (Koch, 1835): functional characterization throughout reproductive process, potential for use as an indicator of oocyte quality and endocrine disruption biomarker in males. Aquat Toxicol 112–113, 72–82 (2012).
6. Kohler, S. A., Parker, M. O. & Ford, A. T. Species-specific behaviours in amphipods highlight the need for understanding baseline behaviours in ecotoxicology. Aquat Toxicol 202, 173–180 (2018).
5. Malbry, L. & Crane, M. Responses of Gammarus pulex (Amphipoda, Crustacea) to metalloferrous effluents: identification of toxic components and the importance of interpopulation variation. Environ Pollut 84, 45–52 (1994).
4. Xuereb, B., Chaumot, A., Mons, R., Garric, J. & Geffard, O. Acetylcholinesterase activity in Gammarus fossarum (Crustacea Amphipoda) Intrinsic variability, reference levels, and a reliable tool for field surveys. Aquat Toxicol 93, 225–233 (2009).
3. Xuereb, B., Noury, P., Felten, V., Garric, J. & Geffard, O. Cholinesterase activity in Gammarus pulex (Crustacea Amphipoda): characterization and effects of chlorpyrifos. Toxicology 236, 178–189 (2007).
2. Gouveia, D. et al. Ecotoxicology of Gammarus fossarum to document the reproductive system of amphipods. Mol Cell Proteomics 13, 3612–3625 (2014).
1. Trapp, J. et al. High-throughput proteome dynamics for discovery of key proteins in sentinel species: Unsuspected vitellogenins diversity in the crustacean Gammarus fossarum. J Proteome Res 16, 207–214 (2016).

52. Lagrue, C. et al. Confrontation of cryptic diversity and mate discrimination within Gammarus pulex and Gammarus fossarum species complexes. *Freshwater Biology* **59**, 2555–2570 (2014).
53. Gouy, M., Guindon, S. & Gascuel, O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* **27**, 221–224 (2010).
54. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* **29**, 644–652 (2011).
55. Smith-Unna, R., Boursnell, C., Patro, R., Hibberd, J. M. & Kelly, S. TransRate: reference-free quality assessment of *de novo* transcriptome assemblies. *Genome Res* **26**, 1134–1144 (2016).
56. Seppey, M., Manni, M. & Zdobnov, E. M. BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol Biol* **1962**, 227–245 (2019).
57. NCBI Sequence Read Archive, http://identifiers.org/ncbi/sra:SRR8089720 (2019).
58. NCBI Sequence Read Archive, http://identifiers.org/ncbi/sra:SRR8089722 (2019).
59. NCBI Sequence Read Archive, http://identifiers.org/ncbi/sra:SRR8089723 (2019).
60. NCBI Sequence Read Archive, http://identifiers.org/ncbi/sra:SRR8089724 (2019).
61. NCBI Sequence Read Archive, http://identifiers.org/ncbi/sra:SRR8089725 (2019).
62. NCBI Sequence Read Archive, http://identifiers.org/ncbi/sra:SRR8089727 (2019).
63. NCBI Sequence Read Archive, http://identifiers.org/ncbi/sra:SRR8089728 (2019).
64. NCBI Sequence Read Archive, http://identifiers.org/ncbi/sra:SRR8089729 (2019).
65. NCBI Sequence Read Archive, http://identifiers.org/ncbi/sra:SRR8089730 (2019).
66. NCBI Sequence Read Archive, http://identifiers.org/ncbi/sra:SRR8089731 (2019).
67. NCBI Sequence Read Archive, http://identifiers.org/ncbi/sra:SRR8089732 (2019).
68. NCBI Sequence Read Archive, http://identifiers.org/ncbi/sra:SRR8089733 (2019).
69. NCBI Sequence Read Archive, http://identifiers.org/ncbi/sra:SRR8089734 (2019).
70. NCBI Sequence Read Archive, http://identifiers.org/ncbi/sra:SRR8089735 (2019).
71. GenBank, https://identifiers.org/ncbi/insdc:GHCN01000000 (2019).
72. GenBank, https://identifiers.org/ncbi/insdc:GHC01000000 (2019).
73. GenBank, https://identifiers.org/ncbi/insdc:GHCQ01000000 (2019).
74. GenBank, https://identifiers.org/ncbi/insdc:GHR01000000 (2019).
75. GenBank, https://identifiers.org/ncbi/insdc:GHT01000000 (2019).
76. GenBank, https://identifiers.org/ncbi/insdc:GHCU01000000 (2019).
77. GenBank, https://identifiers.org/ncbi/insdc:GVC01000000 (2019).
78. GenBank, https://identifiers.org/ncbi/insdc:GHCW01000000 (2019).
79. GenBank, https://identifiers.org/ncbi/insdc:GHCX01000000 (2019).
80. GenBank, https://identifiers.org/ncbi/insdc:GHCY01000000 (2019).
81. GenBank, https://identifiers.org/ncbi/insdc:GHCZ01000000 (2019).
82. GenBank, https://identifiers.org/ncbi/insdc:GD01000000 (2019).
83. GenBank, https://identifiers.org/ncbi/insdc:GDA01000000 (2019).
84. GenBank, https://identifiers.org/ncbi/insdc:GHD01000000 (2019).

**Acknowledgements**

We thank the Institut National de Recherche en Sciences et Technologies pour l’Environnement et l’Agriculture (France), the Commissariat à l’Énergie Atomique et aux Énergies Alternatives (France), and the Agence Nationale de la Recherche “ProteoGam” program (ANR-14-CE21-0006-02) for financial support.

**Author Contributions**

Y.C., D.D.E., O.P., J.A., A.C. and C.A. performed the experimental work. Y.C., D.D.E., O.P., J.A., A.C. and C.A. analysed the data. Y.C. and J.A. wrote the manuscript with input from all co-authors.

**Additional Information**

**Competing Interests:** The authors declare no competing interests.

**Publisher’s note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.