HAL open science

Untargeted analysis of nanoLC-HRMS data by ANOVA-PCA to highlight metabolites in Gammarus fossarum after in vivo exposure to pharmaceuticals

Christelle Bonnefoy, Aurélie Fildier, Audrey Buleté, Claire Bordes, Jeanne Garric, Emmanuelle Vulliet

To cite this version:

Christelle Bonnefoy, Aurélie Fildier, Audrey Buleté, Claire Bordes, Jeanne Garric, et al.. Untargeted analysis of nanoLC-HRMS data by ANOVA-PCA to highlight metabolites in Gammarus fossarum after in vivo exposure to pharmaceuticals. Talanta, 2019, 202, pp.221-229. 10.1016/j.talanta.2019.04.028 . hal-02124467

HAL Id: hal-02124467

https://hal.science/hal-02124467v1

Submitted on 22 Oct 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Version of Record: https://www.sciencedirect.com/science/article/pii/S0039914019304175
Manuscript_66eda4b0a2911ffdcba83e29a2b011a2

Untargeted analysis of nanoLC-HRMS data by ANOVA-PCA to highlight metabolites in Gammarus fossarum after in vivo exposure to pharmaceuticals

C. Bonnefoy[1]*, A. Fildier[1], A. Buleté[1], C. Bordes[1], J. Garric[2], E. Vulliet[1]

*corresponding author

[1]Univ Lyon, CNRS, Université Claude Bernard Lyon 1, Ens de Lyon – Institut des Sciences Analytiques, UMR 5280, 5 rue de la Doua, 69100 Villeurbanne, France

[2]Irstea, UR RIVERLY, centre de Lyon-5 rue de la Doua, Villeurbanne, F-69616 Villeurbanne, France

© 2019 published by Elsevier. This manuscript is made available under the CC BY NC user license
https://creativecommons.org/licenses/by-nc/4.0/

1. Introduction

The contamination of aquatic ecosystems by organic micropollutants notably discharged from wastewater treatment plants is a growing scientific and societal concern that raises several questions regarding the potential for adverse effects on wild organisms and human health [1]. Among the involved chemicals, many pharmaceutical products have been quantified at concentrations below 1 µg L^{-1} [2]. In the mid-1970s, Goldberg *et al*. [3] suggested for the first time the use of biota to monitor levels and trends of chemical contamination in water. Among the so-called integrating matrices implemented in biomonitoring programs, *Gammarus fossarum* is known as a sentinel species for river monitoring [4]. Its use in freshwater biomonitoring and ecotoxicology risk assessment programs has steadily increased over the last few years [5, 6], based on the analysis of a pool of organisms. Very recently, due to analytical improvements, the identification/quantification of contaminants and metabolites in *G. fossarum* was made possible at the scale of one single individual. Indeed, progress in the liquid chromatography coupled to mass spectrometry (LC-MS) systems, instrumental downsizing achieved with the advent of nanoLC-high resolution mass spectrometry (nanoLC-HRMS) [7, 8] and the miniaturization of the sample preparation step thanks to µ-QUECHERS procedure [9] allowed to assess the inter-individual variabilities of contaminants and metabolites in small size organisms. Nevertheless, given the analytical challenge associated with the study of very low biotic matrice masses, and the lack of available data on metabolism of invertebrates, very few studies were interested in the detection and identification of metabolites or biomarkers

in single organism following an exposure to pharmaceutical compounds. The effect of mixture of pharmaceutical compounds on the metabolism, likely to underlie toxic effects, especially in invertebrates, remains a vast field of research.

Besides analytical breakthroughs, the spreading of omic sciences (genomics, proteomics, metabolomics) has also triggered an evolution of the chemometric methods, for many topics including environmental ones. Analysing data coming from designed experiments turned out to be an interesting field of research investigation [10, 11]. Among multivariate methods, ANOVA-Simultaneous Component Analysis (ASCA) [12] and ANOVA-Principal Component Analysis (APCA) [13] were introduced in the second half of the 2000s. They both combine analysis of variance (ANOVA) and Principal Component Analysis (PCA) to get simplified view of the multivariate data while taking into account their designed structure. ANOVA decomposes the data matrix into pure effect submatrices according to the design of experiment and a residuals matrix. Then PCA is performed on each effect submatrix so as to represent information in a reduced-dimensional space. In the case of APCA, experimental submatrices are built by adding the residuals to each pure effect submatrix prior to PCA calculations. To improve regression models, ANOVA decomposition has also been associated to other methods such as PLS [11]. Other authors combined ANOVA with multi-block analysis such as ANOVA ComDim (AComDim) [14, 15] or ANOVA Multiblock OPLS (AMOPLS) [16, 17] to perform one single analysis of the submatrices instead of several PCAs. Due to the ANOVA step, all these methods could provide biased estimators of the factor effects in the case of unbalanced designs. Hence, extensions of ASCA and APCA methods dedicated to unbalanced

designs, called ASCA+ and APCA+, were recently developed [18]. These methods consist in decomposing the data matrix by using the general linear model (GLM) and an appropriate factor coding. GLM allows rewriting ANOVA model as a multiple linear regression making possible the generalization of ANOVA model parameters estimation to unbalanced designs [18, 19].

In this context, the objective of this study was to investigate the effect of an exposure to two benzodiazepine derivatives on the metabolism of male and female *G. fossarum*. Gammarids were exposed to a mix of carbamazepine (widely used as antiepileptic treatment) and oxazepam (used as an anxiolytic and in alcohol withdrawal). Untargeted metabolomic analysis were conducted according to a full factorial design by using nanoLC-HRMS following a µ-QuEChERS extraction. Due to the limited number of individuals, APCA and APCA+ were used for data processing with the aim to investigate the effect of drugs mixture on the metabolome and whether this effect was related to the gender of the individual.

2. Material and methods

2.1. Chemicals and reagents

Carbamazepine and oxazepam standards were purchased from Sigma-Aldrich (St. Quentin Fallavier, France) with purity higher than 99%. Twelve analytical standards of metabolites or degradation products of carbamazepine were used for quality control (QC): 10,11 epoxicarbamazepine, oxacarbamazepine, acridine, 9(10H) acridanone, and 10,11 dihydroxy carbamazepine were supplied by Sigma-Aldrich; licarbazepine, eslicarbamazepine, 11-keto-

oxcarbamazepine, 10,11-dihydro-10,11-dihydroxycarbamazepine and iminodibenzyl by TLC PharmaChem (Mississauga, Ontario, Canada); 2-hydrocarbamazepine and 3-hydroxycarbamazepine by Toronto Research Chemicals (Toronto, Ontario, Canada). Hexakis(2,2-difluoroethoxy)phosphazene used for lock mass calibration was purchased from Santa Cruz Biotechnology, Dallas, (United States) and was dissolved in methanol (MeOH). Ultra-pure water (18.2 MΩ cm^{-1}) was obtained from a Milli-Q A10 water purification system from Millipore (St. Quentin en Yvelines, France). Methanol (MeOH) was a LC-MS quality grade and was supplied by BioSolve (Dieuse, France) and acetic acid was obtained from Fluka (Sigma-Aldrich). Citrate buffer salts from Agilent Technology (Massy, France) were used for QuEChERS extraction.

2.2. Sample collection and preparation

To ensure the homogeneity of the physiological conditions of gammarids, calibrated adults (size and state of development) were selected. Both male and female gammarids were exposed 14 days in semi-static conditions. Male or female gammarids were placed into thermostated beakers (12°C) filled either with 200 mL of clean water or of mixture of carbamazepine and oxazepam at a concentration of 1000 ng L^{-1} each. Organisms were continuously oxygenated and fed *ad libitum*.

Five individuals were collected after a 14-days exposure for each of the following condition: female exposed (J14COF), male unexposed (J14M) or female unexposed (J14F). For male exposed (J14COM), only four gammarids were available. A total of 19 samples were therefore collected and immediately freezed and lyophilized.

The extraction of compounds (endogenous or exogenous metabolites) was adapted from the previous work of Garcia-Galan *et al.* [20]. One lyophilized organism was processed at a time. First, it was ground in a centrifuge tube with stainless steel grinding ball for 3 min at 1,000 rpm. Then 500 µL of milliQwater, 500 µL of ACN and 500 mg of citrate buffers were added in the tube before vortexing during 30 s and centrifuging during 2 min at 10,000 rpm. Afterward, 400 µL of supernatant ACN were added to 80 mg of dispersive phase (mixture of primary secondary amine and C18) before vortexing during 30 s and centrifuging during 2 min at 10,000 rpm. Finally, 200 µL of the purified supernatant were transferred into an LC vial, evaporated to dryness under a N_2 stream and reconstituted in 100 µL of H2O/MeOH (90/10) for nanoLC-HRMS analysis.

For the preparation of spiked QC, control gammarids (i.e. unexposed gammarids) were spiked with carbamazepine and oxazepam at 150 ng g^{-1} each and with the twelve analytical standards, then extracted and purified as above.

A blank matrix (BM) was also used, corresponding to control gammarids that have undergone the same sample preparation procedure as the exposed ones.

Pooled QC are detailed in section 2.3.2.

2.3. NanoLC-HRMS analysis

2.3.1. Equipment

Analyses were performed on an Ultimate 3000 Nano Chromatography System (Thermofisher®, Villebon sur Yvette, France) coupled to a Maxis Plus mass spectrometer equipped with a CaptiveSpray source (Bruker Daltonics®). The chromatographic separation was performed on a C18 PepMap 100 column (150 mm × 75 µm ID, 2 µm, 100 Å) (Thermofisher®) with the experimental conditions described by Sordet *et al.* [21]. The ionization was in positive mode with the following settings for the source: capillary voltage of 1500 V, drying gas of 3L/min (N_2), and drying temperature of 150°C. Full scan mass spectra were recorded over the range of 50–1000 m/z, with a scan rate of 0.4 Hz in profile mode with the following transfer parameters: funnel 1 rf of 200 Vpp, multipole rf of 50 Vpp, quadrupole energy of 5 eV, collision energy of 7 eV, stepping basic and a prepulse storage of 5 µs. A lock mass calibration was used to calibrate the data with the infusion of Hexakis(2,2-difluoroethoxy) phosphazene (m/z = 622.0290) into the CaptiveSpray. Hystar®, OToFcontrol 1.9, Data Analysis 4.4 (Bruker Daltonics®) were used for instrument control to perform mass calibration and also to export raw data in netCDF format which is an open one. Hence the data can be further manipulated with online tools or dedicated software.

2.3.2. Sample sequence and operating conditions

A 2^2 full factorial design was used to evaluate the effect of two experimental factors as well as the effect of their two-factor interaction. As illustrated in Figure 1, the first factor (X_1) was the exposure to carbamazepine and oxazepam: the coded levels -1 and +1 corresponded to unexposed and exposed gammarids, respectively. The second factor (X_2) was the gammarid gender, coded -1 for the males and +1 for the females. Each of the four experiments has

been performed five times requiring the characterization of five individuals. From a technical point of view, a single measurement sequence was not possible due to the number of samples to be analysed. Indeed, each sample was planned to be injected in triplicate, and the chromatographic separation was followed by a reconditioning of the column and the cleaning of the pre-column to prevent cross-contamination between samples. So, it was decided to split the 19 samples into two distinct groups to be handled independently: 3 individuals of each operating conditions were gathered in sequence A (except for "male exposed" condition with only two samples) whereas only 2 individuals were characterized per studied condition in sequence B.

One pooled quality control was obtained for each sequence by mixing equal volumes of each of the sequence samples (i.e. 11 samples in the case of sequence A (QCA) and 8 samples in the case of sequence B (QCB)). A blank matrix sample (BM), the pooled QCs (QCA or QCB) and a spiked QC were injected at the beginning and at the end of each analytical sequence, as well as after every 4 injections so as to monitor the instrumental drift. At the end of the sequence A, the mass spectrometer source was cleaned and the mobile phase was renewed before starting sequence B. QCA samples were also regularly injected during sequence B. The full analytical run planned for each analytical sequence is presented in Figure S1. The way QCs were used to correct the data is described in section 2.5.

2.4. Pre-processing and pre-treatment of raw data

As presented by the flowchart in Figure 2, and in accordance with the proposals of Goodacre *et al.* [22], several pre-processing and pre-treatment steps have been applied to the raw

data prior to multivariate data analysis. After mass calibration and net CDF export, data were normalized according to the intensity measured at the lock mass calibration. Then, after a feature detection step, Missing Value Imputation (MVI) was performed. Finally, filtering and QC correction were successively applied prior to APCA/APCA+ in order to highlight the variables responsible for possible exposure or gender effect. Each step of the flowchart is described in more details in the following sections.

2.4.1. Lock mass normalisation

The classical lock mass compound hexakis(2,2-difluoroethoxy)phosphazene was added into the source so as to be ionised continuously throughout the analytical sequences [23].

At each scan, we performed: i) a mass shift correction using the mass of the lock mass compound 622.0290 and ii) a lock mass normalisation by dividing the intensity of each detected signal by the signal intensity of the lock mass compound. This was done for all the samples in both sequences using Matlab® software (Mathworks Inc, Natick MA, USA).

2.4.2. Feature detection

After lock mass normalisation, data were pre-processed on a local cluster thanks to the XCMS R package (Version 1.50.1). The first step was to look for peaks present individually in each sample. To this end, the Centwave algorithm was used with the following parameters for the function xcmsSet: scanrange = c(180,600), ppm = 5, peakwidth = c(10,60), mzdiff = 0.005, snthresh = 6, prefilter = c(6,10000). They were chosen from the results on spiked QCs so as to ensure the recovery of the peaks corresponding to the 14 added compounds (the

two pharmaceuticals and the twelve analytical standards as described in part 2.2). In a second step, the peaks matching the same couple were affected to the same group and identified by the median mass to charge ratio (m/z) and the median retention time (tr) determined from the experimental values. Only those couples present in a half of the total number of samples or in a half of samples prepared under identical operating conditions were retained thereafter (minfrac = 0.5). The other settings for the XCMS group function were: bw = 5, mzwid = 0.015 and max = 100. If a peak was detected in the sample, its height was kept in the resulting table otherwise data was missing and had to be estimated. In the rest of the text, the terms (mz;tr), couple, feature or variable will be used indistinctly.

2.4.3. Missing value imputation

As detailed in Di Guida et al. [24], missing values in metabolomics datasets may be due to three main reasons: i) some metabolites are not present in all the analysed samples, ii) metabolite concentration is not high enough leading to poor signals similar to the analytical background and iii) no metabolite peak is identified at the feature detection step because the algorithm's criteria are not fully met. Missing values are the result of biological and technical issues and their distribution may be random or systematic. Here the XCMS fillpeaks function, classically used for MVI in metabolomics, was applied to manage missing values prior to the use of chemometric methods. Another method based on means calculation was also tested: a missing value was replaced by the mean intensity calculated from the signals of the corresponding variable (mz;tr) obtained for all the samples subjected to the same experimental conditions. If no peak was detected (i.e. no mean intensity was available), we

used the corresponding mean value calculated from the BMs signals after peak reintegration by the fillpeaks algorithm.

2.4.4. Filtering

A final step consisted in filtering the obtained data matrix. A variable (mz;tr) was kept in the dataset if: i) it was detected in each pooled QC of a sequence with a coefficient of variation less than 30% and ii) it was characterized by a ratio between the mean signals measured from the pooled QCs and the blank samples BMs superior or equal to 10.

2.4.5. QC correction

In each analytical sequence, as described in part 2.3.2, four samples were bracketed by two QCs so as to provide data to use for signal correction within each sequence (Figure S1). The correction for the experimental drift was adapted from the Quality Control-based Robust Loess Signal correction procedure (QC-RLSC) proposed by Dunn *et al.* [25]. A local first degree polynomial was fitted to the QC data with respect to the order of injection. The correction curve thus obtained was interpolated for the whole analytical sequence, to which each metabolic feature was normalized. Signal correction was performed within both analytical sequences A and B on the base of QCA and QCB samples, respectively. Moreover, an additional correction between the two sequences was achieved using QCA samples that were also injected during the sequence B (Figure S2).

2.5. ANOVA-Principal Component Analysis (APCA) and APCA+

APCA starts with the decomposition of the data matrix into pure effects and residuals submatrices according to the ANOVA model:

$$\mathbf{Y} = \mathbf{M_0} + \mathbf{M_1} + \mathbf{M_2} + \mathbf{M_{12}} + \mathbf{E} \qquad \textbf{(1)}$$

where **Y** is the data matrix, $\mathbf{M_0}$ is the mean effect matrix, **E** is the error matrix, $\mathbf{M_1}$, $\mathbf{M_2}$ and $\mathbf{M_{12}}$ are the pure effect (main factor and interaction) submatrices. These matrices are of dimension *n* x *m* where *n* is the total number of samples and *m* is the number of variables.

In APCA, PCA is not directly performed on the different effect matrices as in ASCA but on experimental ones that correspond to pure effect associated with residuals, i.e. ($\mathbf{M_i}$+**E**) matrices. Then, as in ASCA, each factor effect is interpreted through score and loadings plots.

But, as mentioned in introduction, these techniques require a balanced experimental design. And here, a sample was missing in sequence A (see part 2.2) that led us to use APCA+ method to be sure to avoid biased estimations of factor effects. As described in Thiel *et al.* [18], APCA+ is based on the general linear model (GLM) expressed as:

$$\mathbf{Y} = \mathbf{X}\,\boldsymbol{\Theta} + \mathbf{E} \qquad \textbf{(2)}$$

where **Y** is the data matrix of dimension *n* x *m*, **X** is the model matrix of dimension *n* x *p*, $\boldsymbol{\Theta}$ is the parameter matrix of dimension *p* x *m* and **E** is the error matrix of dimension *n* x *m* where *p* is the number of parameters in the model

Here, the used 2^2 full factorial design **X** was made of 19 experiments and is written as:

$$\mathbf{X} = \begin{pmatrix} \mathbf{X_0} & \mathbf{X_1} & \mathbf{X_2} & X_1\mathbf{X_2} \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & -1 & 1 \end{pmatrix} \qquad \textbf{(3)}$$

Unbiased estimators of the model parameters are obtained by ordinary least square method:

$$\widehat{\boldsymbol{\Theta}} = (\mathbf{X'X})^{-1}\mathbf{X'Y} \qquad \textbf{(4)}$$

For each effect *f*, a new model matrix X_f^* is built from **X** keeping only the block $\mathbf{X}_f$ that corresponds to the effect of interest and replacing all the other columns by zeros. The effect matrix $\widehat{\boldsymbol{M}}_{\boldsymbol{f}}$, being obtained by multiplying X_f^* matrix by $\widehat{\boldsymbol{\Theta}}$ matrix, is expressed by :

$$\widehat{\mathbf{M}}_f = \boldsymbol{X}_{\boldsymbol{f}}^*\widehat{\boldsymbol{\Theta}} = \mathbf{X}_f\widehat{\boldsymbol{\Theta}}_f \qquad \textbf{(5)}$$

where $\mathbf{X}_f$ and $\widehat{\boldsymbol{\Theta}}_f$ are the subblocks of matrix **X** and $\widehat{\boldsymbol{\Theta}}$ corresponding to effect *f*.

In APCA and APCA+ methods, the residuals matrix $\boldsymbol{E} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\Theta}}$ is added to each pure effect submatrix $\widehat{\boldsymbol{M}}_{\boldsymbol{f}}$ prior to PCA.

If two groups of samples are clearly identified in the score plot along the first principal component according to the two levels of the experimental factor *f*, then the factor might have a significant effect [14]. By extending the ANOVA F-test to multivariate data, the test statistic $\boldsymbol{F_f}$ allows to test the significance of an effect f and is defined by the following expression [19]:

$$\boldsymbol{F_f} = \frac{\|\boldsymbol{E}_{/f}\|^2 - \|\boldsymbol{E}_{\boldsymbol{full}}\|^2}{\|\boldsymbol{E}_{\boldsymbol{full}}\|^2} \qquad \textbf{(6)}$$

where E_{full}, is the error matrix estimated with all effects taken into account in the model and $E_{/f}$, the residuals matrix estimated without including the effect f in the model. The statistical significance of the test statistic is assessed by performing a permutation test (n_{perm} = 1000) [19, 26]. Here we used the permutation approach described by Anderson *et al.* [26] for fixed effects two-way ANOVA design with interaction (Eq. 1). As an example, to test the significance of the main effect exposure, samples were permuted between the exposure factor levels keeping the gender factor levels constant. For the interaction term, the permutation of raw data were unrestricted.

The test statistic (Eq. 6) is determined for each permutation (re-ordering) (n_{perm} = 1000). Finally, a p-value is obtained that corresponds to the proportion of the values of the statistic obtained under permutation that are equal to or larger than the observed value (the significance level is 0.05).

The percentage of variance explained by each main effect or interaction was calculated as in [18]:

$$\%\boldsymbol{Var_f} = \frac{\|\boldsymbol{E_{/f}}\|^2 - \|\boldsymbol{E_{full}}\|^2}{\|\boldsymbol{Y} - \boldsymbol{M_0}\|^2} \times \mathbf{100} \qquad \textbf{(7)}$$

All APCA/APCA+ calculations were achieved by using Matlab® software (Mathworks Inc.).

3. Results and discussion

3.1. Influence of the lock mass normalisation

In a first attempt, data were analysed as they were obtained after the conversion to netCDF format. In these conditions, very high variabilities were observed for samples injected in triplicate. This is illustrated in Figure 3a, by the representation of Extracted Ion Chromatograms (EICs) of oxazepam $[M+H]^+$ ion from a female sample.

As internal standards were not included during sample preparation because of the prohibitive price of labelled metabolites and the risk to alter the samples, an original way to normalize the data was found. Ideally, the signal measured for the mass calibrant should remain invariant during run and within runs of a sequence. However, signal may be impacted by both the elution gradient of the mobile phase and the source clogging. Dividing the signal measured at each m/z by the signal at m/z = 622.0290 for each scan made it possible to compare samples without the unwanted variation of the detector sensitivity. After this lock mass normalisation, the repeatability between injection replicates was greatly improved as illustrated in Figure 3c, and was evaluated through the coefficient of variation (CV) determined from the triplicate measurements. As an example, CV characterizing carbamazepine and oxazepam $[M+H]^+$ ions from a male and a female sample exposed to

both pharmaceuticals are reported in Table 1. CV were about 24 to 33% from the raw data while they dropped to values comprised between 5 and 14% after lock mass normalisation, thus indicating the strong benefit to use this pre-processing technique in this metabolomic context.

Another positive impact of this normalisation procedure was on the peak detection step. As illustrated in Figure 3b, the signal for the lock mass compound evolved during the analytical run. As the signal of the lock mass compound was affected in the same manner as other analytes, it was used to correct the drift of the mass spectrometer. This improved the feature detection step in the case of oxazepam since a single peak (Figure 3c) instead of a double one (Figure 3a) was detected using the lock mass normalisation.

3.2. Missing values imputation

After the feature detection step, the data matrix was made of 76 observations (BMs, QCs, spiked QCs and samples) and 10,966 m/z retention pairs and contained 58% of missing intensity values. In metabolomics studies where data are pre-processed by XCMS, missing values are commonly completed by using the Fillpeaks algorithm. Here, this strategy was first applied to fill the data matrix but incoherent values were observed.

As an example to illustrate this point, the EICs of carbamazepine and oxazepam $[M+H]^+$ ions corresponding to blank, unexposed and exposed samples were compared. For this purpose, we calculated the ratio between the mean intensity measured for the unexposed samples and the blank samples (ratio 1) and the ratio between the mean intensity measured for the

exposed and the unexposed samples (ratio 2). Ratio 1 should be low since comparing two signals near the detection limit and ratio 2 should be high.

Calculations were done for both carbamazepine and oxazepam $[M+H]^+$ ions, for male and female samples respectively. Results are gathered in Table 2. In the case of oxazepam $[M+H]^+$ ion, expected results were obtained: no peak was observed from the analysis of blank and unexposed samples leading to low ratio 1 values (<3) while a significant signal was obtained from exposed samples (ratio 2 values > 7). But surprisingly, the exact opposite was observed for carbamazepine indicating the Fillpeaks algorithm failed here. One explanation may be that the carbamazepine $[M+H]^+$ peak is not well defined at an acquisition frequency of 0.4Hz, which was preferred to 1Hz to promote sensitivity (Figure S2). This may lead to errors at the feature detection step that could be propagated in the peak reintegration process. Other examples of peak detection errors that may occur when using XCMS and MZmine2 are explained in details in the work of Myers *et al.* [27]. In particular, these authors highlighted the link between the local noise underestimation and the detection of false positive peaks.

To limit errors of peak reintegration, missing values imputations by mean replacement were preferred. If the peak existed but did not appeared at the feature detection step, this was compensated by taking into account the mean signal from other samples under the same operating conditions.

3.3. QC correction

To illustrate the influence of the QC correction on the results, PCA was performed on the whole dataset including both sequences A and B. The corresponding scores plots (PC1 vs PC2) for the non-corrected and the QC-corrected data are presented in Figure 4a and 4b, respectively. Figure 4a shows a clear separation of the samples along the first principal component PC1 according to the sequence they belonged to. After the QC-based inter-sequences correction, the sequences appeared projected closer onto PC1-PC2 plane (Figure 4b) but the correction was not sufficient enough to make the sequences comparable.

Consequently, from an analytical point of view, results obtained from the two sequences could not be considered together in a whole dataset. The samples from each sequence were thus investigated independently. Sequence A was characterized by an unbalanced full factorial design **X** made of 11 experiments and sequence B by a balanced design with 8 experiments. APCA+ and APCA methods were therefore applied to the results of sequence A and B, respectively. Calculations were done in both cases using the GLM approach since it provides not only identical results to an ANOVA decomposition in the case of balanced designs but also the determination of unbiased estimators in unbalanced ones [19].

3.4. Exposed and unexposed males and females

For each of the 4 operating conditions studied (female exposed, male exposed, female unexposed and male unexposed), three gammarids were analysed in sequence A and two gammarids in sequence B. As indicated in part 2.3.2, each sample was injected in triplicate. The use of ANOVA model (Eq. 1) requiring independent replicated samples, the average of the technical triplicates were first calculated. Then, APCA+ and APCA methods were applied

to the results obtained from sequence A and B, respectively in order to investigate the effect on gammarid metabolome of the two factors (the exposure to pharmaceuticals and the gammarid gender) and their interaction. The data matrix was centred and decomposed into three pure effect submatrices (factor X_1, factor X_2 and interaction X_1X_2) and a residuals matrix (see part 2.5, Eq. 1).

The use of APCA or APCA+ means that the residuals matrix was added to the pure effect submatrix prior to PCA. The resulting score plots PC1 vs PC2 obtained for the exposure and the gender effect, are respectively presented in Figures 5a and 5b for sequence A. A clear separation of the samples depending on their experimental conditions was observed for each factor along the first principal component. Moreover, the percentage of variance explained by each effect in the model was determined to evaluate their impact on the gammarids metabolism (Eq 7). The exposure and the gender effects accounted for 19 and 22% of the total variability respectively, indicating non-negligible effect. To assess the significance of these effects on the gammarids metabolome, a set of 1000 permutations was applied to the data with the restriction that the level of the factor not under consideration was not modified. The significance of both effects was attested by p-values inferior to 0.05.

Similar results were obtained from the analysis of sequence B results: exposure and gender effects explained 15% and 20% respectively of the total variance and the corresponding score plots PC1 vs PC2 exhibited a clear separation of the samples according to the operating conditions (Figures 5c and 5d).

Regarding the interaction effect, the explained variance was about 9% for sequence A and the performed permutations test indicated non statistically significant effect (p-value ~0.30). Conclusions were less clear for sequence B since the variance explained by the interaction effect was about 15% indicating a non-negligible interaction between the gender and the exposure factors.

Nevertheless, owing to the agreement of the results obtained from both analytical sequences concerning the significativity of the main effects, an in-depth investigation of the variables involved in these effects was achieved by studying the principal components loadings. Due to their high number (1372 couples for sequence A and 1084 couples for sequence B), only the variables whose contribution (loadings) belonged to the fourth quartile were selected as discriminating couples. For the remaining variables (343 and 274 couples for sequences A and B, respectively), the extracted ion chromatograms were individually examined. Couples were considered important: i) in the exposure effect if detected in all samples exposed to the pharmaceuticals or in all samples non exposed, ii) in the gender effect if detected in all female gammarids or in all male gammarids. This led to only a few couples that seemed to be responsible for a discrimination between male and female organisms or between exposed and unexposed gammarids. These variables are listed in Table 3.

The peak (m/z 170.0554; tr = 23.1 min) was proposed as a possible marker of the gender effect. The corresponding extracted ion chromatograms presented in Figure 6 showed that

the signal intensity was reduced in the male gammarids samples compared to the female ones for both exposed and unexposed gammarids whatever the sequence.

The peak (m/z 318.3002; tr = 22.8) min was proposed as a possible marker of the exposure effect.

4. Conclusion

In this untargeted metabolomics study performed on aquatic microinvertebrates at the individual scale, we put in evidence the necessity to normalize the data and raised mathematical problems with the Fillpeaks algorithm at the missing values imputation step. This last issue was solved with a means method that was used in complement to the Fillpeaks algorithm. Furthermore, to our knowledge, it was the first time that the lock mass normalisation was proposed to help in normalising the data.

Thanks to an efficient combination of nanoLC-HRMS with the recent APCA or APCA+ method, a difference in the behaviour of *Gammarus fossarum* males and females exposed to a mixture of two pharmaceutical compounds was shown. Discriminant variables were highlighted to explain part of the exposure and gender effects thus attesting the strategy was adapted to experiments with a limited number of individuals exposed to the same operating conditions.

Further analytical strategies (based on MS/MS fragmentation, mass spectral database interrogation) should now be considered to identify the nature of the potential biomarkers.

Acknowledgments

The authors gratefully acknowledge the support by the French EC2CO program (INSU-CNRS).

Compliance with ethical standards

Conflict of interest

The authors declare that there is no conflict of interest.

Figure 1: Representation of the four experiments designed according to a 2^2 full factorial design

Figure 2: Pre-processing (in orange), pre-treatment (in yellow) and processing (in green) steps applied to the raw data

Figure 3: EIC of oxazepam $[M+H]^+$ ion from an exposed female sample injected in triplicate measured before a) and after c) lock mass normalisation. The corresponding EIC of the lock mass compound is given in b)

Figure 4: ASCA+ scores plots PC1 vs PC2 obtained from the analysis of both sequences A (full dots) and B (empty dots) data before (a) and after (b) inter-sequences correction

Figure 5: Scores plots PC1 vs PC2 obtained from PCA of (M1+E) matrix (a) for sequence A and c) for sequence B) and (M2+E) matrix (b) for sequence A and d) for sequence B) to determine the effects of the exposure factor and the gender factor, respectively

Figure 6: EIC obtained for the variable (m/z 170.0556; tr 23.1 min) detected in sequences A (a and b) and B (c and d) for male (M) or female (F) samples exposed to the pharmaceuticals (J14CO) or not (J14)

[1] M. Gavrilescu, K. Demnerova, J. Aamand, S. Agathos, F. Fava, Emerging pollutants in the environment: present and future challenges in biomonitoring ecological risks and bioremediation, New Biotechnology 32 (2015) 147-156

[2] F. Botta, V. Dulio, Etude sur les contaminants émergents dans les eaux françaises, Convention ONEMA-INERIS (2014) 1-139

[3] E.D. Goldberg, The mussel watch A first step in global marine monitoring, Marine Pollution Bulletin 6 (1975) 11

[4] P.Y. Kunz, C. Kienle, A. Gerhardt, Gammarus spp. In aquatic Ecotoxicology and water quality assessment: toward integrated multilevels tests. Reviews of Environmental Contamination and Toxicology 205 (2010) 1-76

[5] N. Urien, J.D.Lebrun, L.C. Fechner, E. Uher, A. François, H. Quéau, M. Coquery, A. Chaumot, O. Geffard, Environmental relevance of laboratory-derived kinetic models to predict trace metal bioaccumulation in gammarids: Field experimentation at a large spatial scale (France). Water Research 95 (2016) 330-339

[6] J.P. Besse., M. Coquery, C. Lopes, A. Chaumot, H.Budzinski, P. Labadie, O. Geffard, 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 (2013) 650-660

[7] A.J. Chetwynd, A.David, A review of nanoscale LC-ESI for metabolomics and its potential to enhance the metabolome coverage. Talanta 182 (2018) 380-390

[8] A. Berlioz-Barbier, A. Buleté, A. Fildier, J. Garric, E. Vulliet, Non-targeted investigation of benthic invertebrates (Chironomus riparius) exposed to wastewater treatment plant

effluents using nanoliquid chromatography coupled to high resolution mass spectrometry, Chemosphere 196 (2018) 347-353

[9] A. Berlioz-Barbier, R. Baudot, L. Wiest, M. Gust, J. Garric, C. Cren-Olivé, A. Buleté, MicroQuEChERS-nanoliquid chromatography, nanospray-tandem mass spectrometry for the detection and quantification of trace pharmaceuticals in benthic invertebrates. Talanta 132 (2015) 796-802

[10] F. Marini, D. de Beer, E. Joubert, B. Walczak Analysis of variance of designed chromatographic data sets: The analysis of variance-target projection approach. Journal of Chromatography A 1405 (2015) 94-102

[11] E.G. Angelina, E.M. Qannari, T. Moyon, M.C. Alexandre-Gouabau, AOV-PLS: a new method for the analysis of multivariate data depending on several factors. Electronic Journal of Applied Statistical Analysis 8 (2015) 214-235

[12] A.K. Smilde, J.J.Jansen, H.C.J. Hoelsfoot, R.J.A.N. Lamers, J. van der Greef, M.E.Timmerman, ANOVA-simultaneous component analysis (ASCA): a new tool for analyzing designed metabolomics data. Bioinformatics 13 (2005) 3043-3048

[13] P de B. Harrington, N.E. Vieira, J. Espinoza, J.K. Nien, R. Romero, A.L. Yergey, Analysis of variance-principal component analysis: a soft tool for proteomic discovery. Analytica Chimica Acta 544 (2005) 118-127

[14] D. Jouan-Rimbaud Bouveresse, R. Climaco Pinto, L.M. Schmidtke, N. Locquet, D.N. Rutledge, Identification of significant factors by an extension of ANOVA-PCA based on multiblock analysis. Chemometrics and Intelligent Laboratory Systems 106 (2011) 173-182

[15] R.Korifi, S. Amat, C. Rébufa, V. Labed, D.N. Rutledge, N. Dupuy, AComDim as a multivariate tool to analyse experimental design application to γ-irradiated and leached ion exchange resins. Chemometrics and Intelligent Laboratory Systems 141 (2015) 12-23

[16] J. Boccard, S. Rudaz, Exploring Omics data from designed experiments using analysis of variance multiblock Orthogonal Partial Least Squares. Analytica Chimica Acta 920 (2016) 18-28

[17] V.Gonzalez-Ruiz, J. Pezzatti, A. Roux, L. Stoppini, J. Boccard, S. Rudaz, Unravelling the effects of multiple experimental factors in metabolomics, analysis of human neural cells with hydrophilic interaction liquid chromatography hyphenated to high resolution mass spectrometry. Journal of Chromatography A 1527 (2017) 53-60

[18] M. Thiel, B. Féraud, B. Govaerts, ASCA+ and APCA+: Extensions of ASCA and APCA in the analysis of unbalanced multifactorial designs. Journal of Chemometrics 31 (2017) e2895 1-13

[19] S. Guisset, M. Martin, B. Govaerts, Comparison of PARAFASCA, AComDim, and AMOPLS approaches in the multivariate GLM modelling of multi-factorial designs, Chemometrics and Intelligent Laboratory Systems 184 (2019) 44-63

[20] M.J.Garcia-Galan, M. Sordet, A. Buleté, J. Garric, E. Vulliet, Evaluation of the influence of surfactants in the bioaccumulation kinetics of sulfamethoxazole and oxazepam in benthic invertebrates. Science of the Total Environment 592 (2017) 554-564

[21] M. Sordet, A. Berlioz-Barbier, A. Buleté, J. Garric, E. Vulliet, Quantification of emerging pollutants in an amphipod crustacean by nanoliquid chromatography coupled to

mass spectrometry using multiple reaction monitoring cubed mode. Journal of Chromatography A 1456 (2016) 217-225

[22] R. Goodacre, D. Broadhurst, A.K. Smilde, B.S.Kristal, J.D.Baker, R. Beger, Bessant C., S. Connor, G. Capuani, A. Craig, T. Ebbels, D.B. Kell, C. Manetti, J. Newton, G. Paternostro, R. Somorjai, M. Sjöström, J. Trygg, F. Wulfert, Proposed minimum reporting standards for data analysis in metabolomics. Metabolomics 3 (2007) 231-241

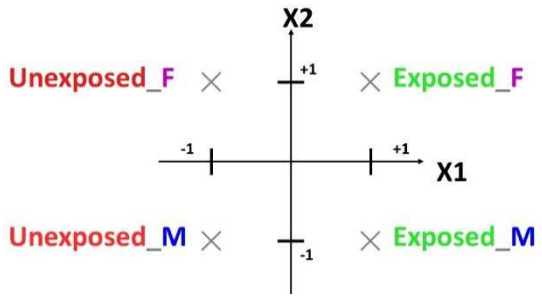
[23] G. Gabant, A. Boyer, M. Cadene, SSPaQ : A subtractive segmentation approach for the exhaustive parallel quantification of the extent of protein modification at every possible site, Journal of The American Society for Mass Spectrometry 27 (2016) 1328-1343

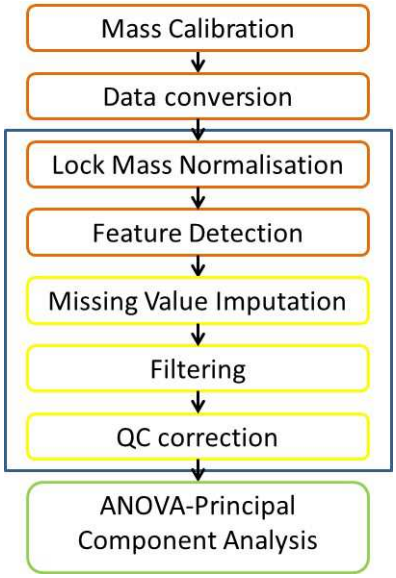
[24] R. Di Guida, J. Engel, J.W. Allwood, R.J.M. Weber, M.R. Jones, U. Sommer, M.R. Viant, W.B. Dunn, Non-targeted UHPLC-MS metabolomics data processing methods : a comparative investigation of normalisation, missing value imputation, transformation and scaling, Metabolomics 12 (2016) 93

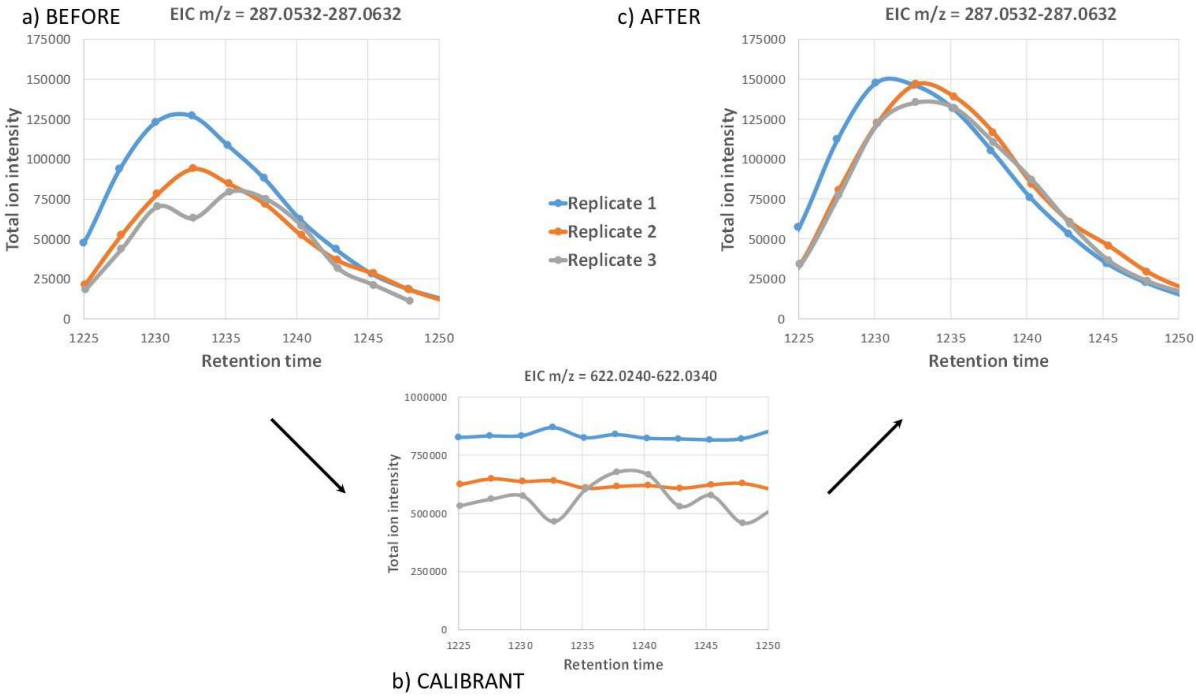
[25] WB. Dunn, D. Broadhurst, P. Begley, E. Zelena, S. Francis-McIntyre, N. Anderson, M. Brown, J.D. Knowles, A. Halsall, J.N. Haselden, A.W.Nicholls, I.D.Wilson, D.B. Kell, R. Goodacre, the Human Serum Metabolome (HUSERMET) Consortium, Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. Nature Protocols 6 (2011) 1060-1083

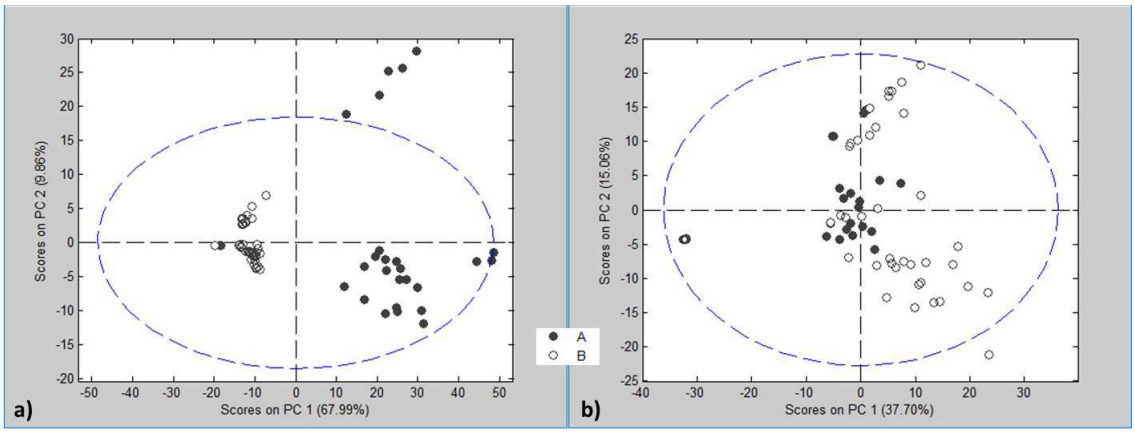
[26] M.J. Anderson, C.J.F. Ter Braak, Permutation tests for multi-factorial analysis of variance, Journal of Statistical Computation and Simulation 73 (2003) 85-113

[27] O.D. Myers, S.J. Sumner, S. Li, S. Barnes, X. Du, Detailed investigation and comparison of the XCMS and Mzmine 2 chromatogram construction and chromatographic peak detection methods for preprocessing mass spectrometry metabolomics data. Analytical Chemistry 89 (2017) 8689-8695

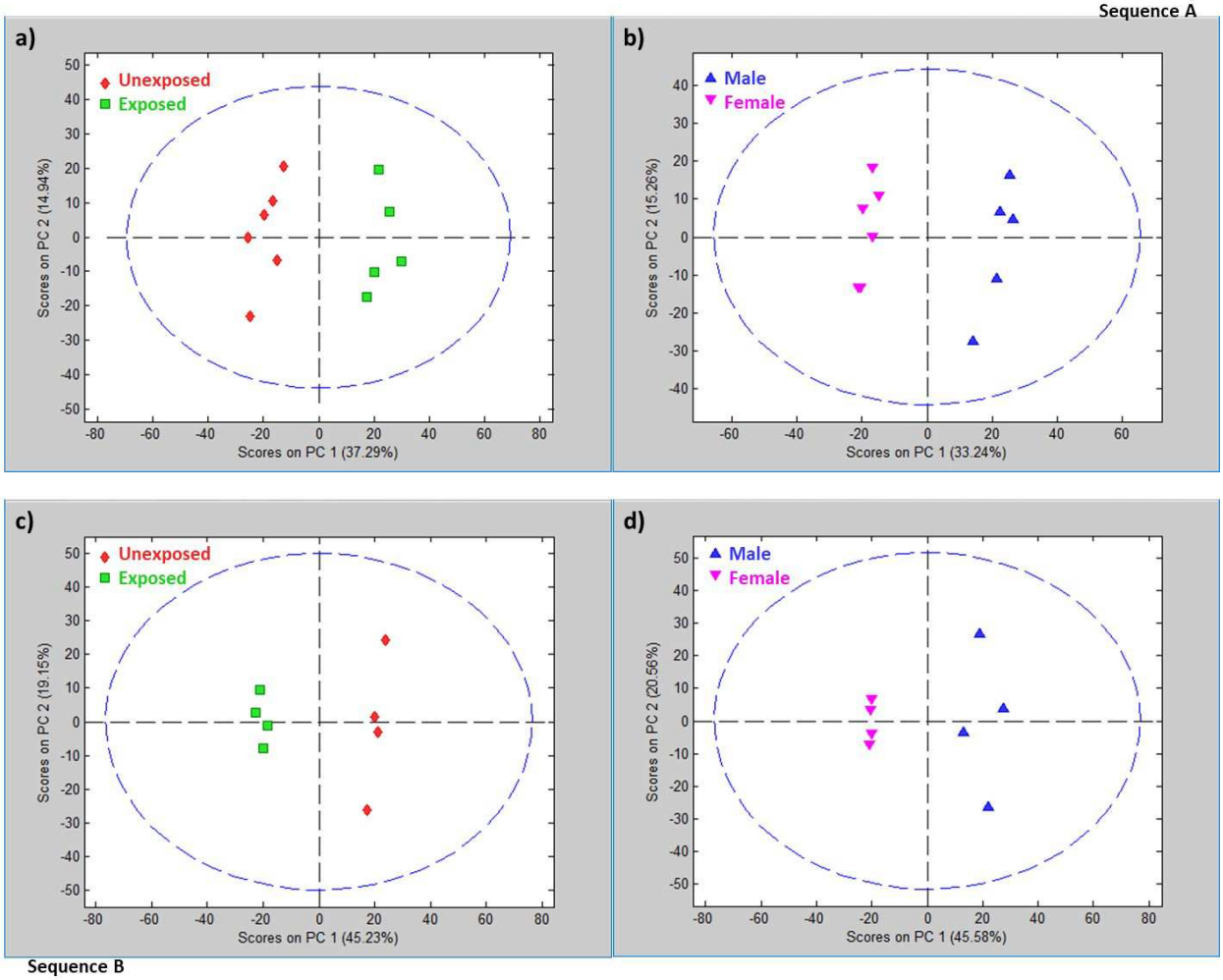

X2
Unexposed_F
+1
Exposed_F
-1
+1
X1
Unexposed_M
-1
Exposed_M


Mass Calibration
Data conversion
Lock Mass Normalisation
Feature Detection
Missing Value Imputation
Filtering
QC correction
ANOVA-Principal Component Analysis


a) BEFORE
EIC m/z = 287.0532-287.0632
Total ion intensity
Retention time
c) AFTER
EIC m/z = 287.0532-287.0632
Total ion intensity
Retention time
Replicate 1
Replicate 2
Replicate 3
EIC m/z = 622.0240-622.0340
Total ion intensity
Retention time
b) CALIBRANT


a)
Scores on PC 2 (9.86%)
Scores on PC 1 (67.99%)
A
B
b)
Scores on PC 2 (15.06%)
Scores on PC 1 (37.70%)

Sequence A



Sequence B

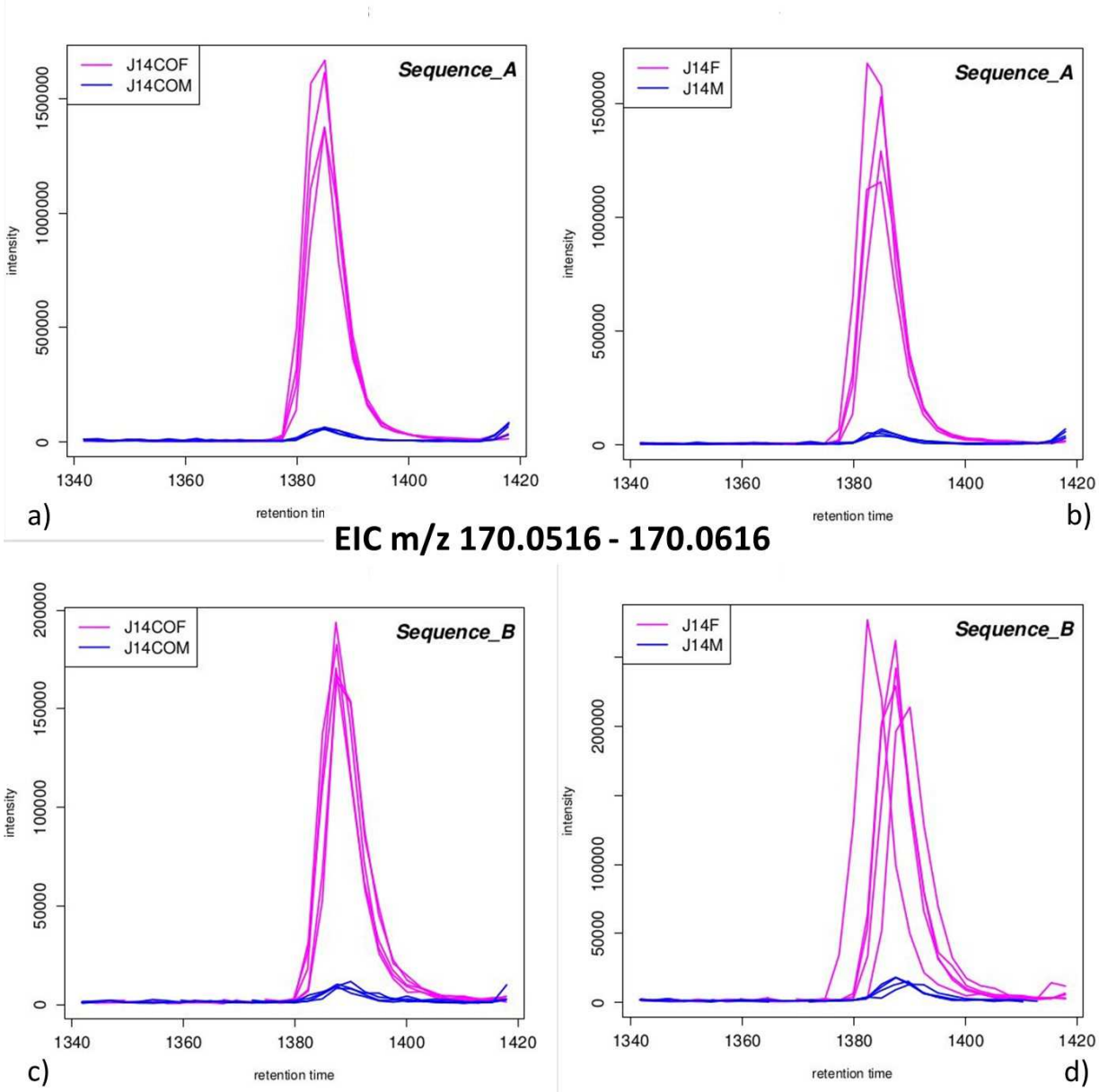

J14COF
J14COM
Sequence_A
intensity
retention tin
a)
J14F
J14M
Sequence_A
intensity
retention time
b)
EIC m/z 170.0516 - 170.0616
J14COF
J14COM
Sequence_B
intensity
retention time
c)
J14F
J14M
Sequence_B
intensity
retention time
d)

Table 1 Inter-replicates CV measured from a male and a female samples before and after lock mass normalisation

Inter-replicates CV	Before normalisation		After normalisation	
	male	female	male	female
Carbamazepine	30%	33%	14%	9%
Oxazepam	29%	24%	10%	5%

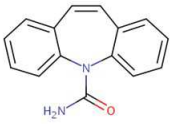
Table 2 Mean ratios measured between blank and unexposed samples (Ratio 1) and between unexposed and exposed samples (Ratio 2) for both carbamazepine and oxazepam $[M+H]^+$ ions after the Fillpeaks reintegration obtained from male and female gammarids

Means Ratio	Unexposed Samples vs Blank Ratio 1		Exposed Samples vs Unexposed Ratio 2	
	male	female	male	female
Carbamazepine	10	9	1	1
Oxazepam	3	2	7	13

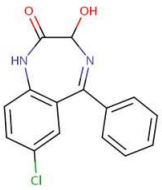
Table 3 Variables responsible for the exposure, gender and interaction effects in the gammarid metabolism

m/z	**t_R in min**	**effect**
170.0556	23.1	gender
172.9555	23.5	gender
182.1254	20.5	gender
277.2164	24.9	gender
213.1485	22.2	exposure
277.1910	20.1	exposure
318.3002	23.0	exposure
522.2086	10.5	exposure

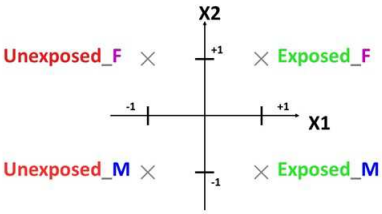
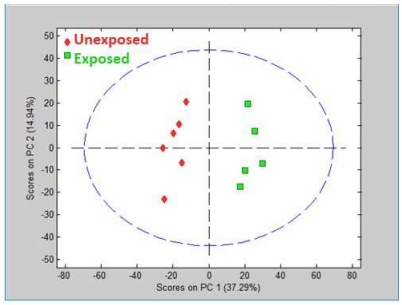
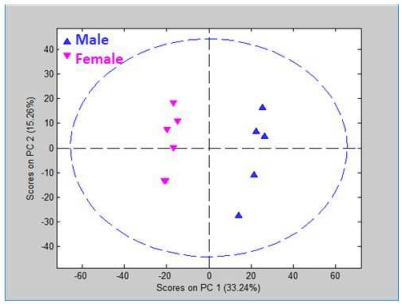
Gammarus fossarum



Carbamazepine



Oxazepam



Mass Calibration
↓
Data conversion
↓
Lock Mass Normalisation
↓
Feature Detection
↓
Missing Value Imputation
↓
Filtering
↓
QC correction
↓
ANOVA-Principal Component Analysis