OPEN ACCESS DOCUMENT

Information of the Journal in which the present paper is published:

- Wiley. Journal of Biophotonics, (2017) (in press, accepted the 2nd August 2017).

- DOI: 10.1002/jbio.201700089

COMBINING HYPERSPECTRAL IMAGING AND CHEMOMETRICS TO ASSESS AND INTERPRET THE EFFECTS OF ENVIRONMENTAL STRESSORS ON THE ORGANISM AT TISSUE LEVEL

Víctor Olmos[1 (*)], Mònica Marro[2], Pablo Loza-Alvarez[2], Demetrio Raldúa[3], Eva Prats[4], Francesc Padrós[5], Benjamin Piña[3], Romà Tauler[3], Anna de Juan[1].

1. Department of Chemical Engineering and Analytical Chemistry, University of Barcelona. Diagonal 645, 08028 Barcelona, Spain
2. ICFO-Institut de Ciencies Fotòniques, The Barcelona Institute of Science and Technology, Carl Friedrich Gauss 3, 08860 Castelldefels, Spain
3. Department of Environmental Chemistry, Institute of Environmental Assessment and Water Diagnostic (IDAEA-CSIC), Jordi Girona 18, 08034 Barcelona, Spain
4. Research and Development Centre (CID-CSIC), Jordi Girona 18, 08034 Barcelona, Spain
5. Pathological Diagnostic Service in Fish, Universitat Autònoma de Barcelona, 08190 Bellaterra, Spain

ABSTRACT

Changes on an organism by the exposure to environmental stressors may be characterized by hyperspectral images (HSI), which preserve the morphology of biological samples, and suitable chemometric tools. The approach proposed allows assessing and interpreting the effect of contaminant exposure on heterogeneous biological samples monitored by HSI at specific tissue levels. In this work, the model example used consists of the study of the effect of the exposure of chlorpyrifos-oxon on zebra fish tissues. To assess this effect, unmixing of the biological sample images followed by tissue-specific classification models based on the unmixed spectral signatures is proposed. Unmixing and classification are performed by Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS) and Partial Least Squares-Discriminant Analysis (PLS-DA), respectively.

Crucial aspects of the approach are: a) the simultaneous MCR-ALS analysis of all images from one population to take into account biological variability and provide reliable tissue spectral signatures, and b) the use of resolved spectral signatures from control and exposed populations obtained from resampling of pixel subsets analyzed by MCR-ALS multiset analysis as information for the tissue-specific PLS-DA classification models. Classification

results diagnose the presence of a significant effect and identify the spectral regions at a tissue level responsible for the biological change.

Keywords: Raman HSI, zebrafish, MCR-ALS, resampling, PLS-DA

(*) Corresponding author: victor_olmos@ub.edu.

1. INTRODUCTION

In order to characterize the human impact on Earth, the effect of contaminant and drug exposure on life organisms has to be studied. Environmental –omics consists of the characterization, quantification and study of the variation of biological molecules of an organism when exposed to an environmental stress. The most used techniques for this purpose, HPLC-MS, GC-MS, immunoassays, NMR, etc.[1–6], are usually destructive techniques. In this work, hyperspectral images (HSI) are used because they allow preserving the natural morphology of the sample and provide spatial and chemical information. The spatial information preserved by HSI permits a better characterization of the contaminant/drug effect at a specific biological component level (tissue/sub-tissue) in the organism.

HSI may be performed with many different spectroscopic techniques (fluorescence, ultraviolet-visible, infrared, Raman…) as well as mass spectrometry [7–10]. In this work, Raman HSI have been used. Raman spectroscopy is used to observe vibrational and rotational modes of molecule bonds. The so-called "fingerprint region" of Raman spectra (900-1800 cm^{-1}) is usually acquired for biological samples because it contains wide information about different families of biological molecules [9,11–13], e.g. lipids, proteins, DNA… Raman spectroscopy coupled with MCR analysis has been used to monitor different biological molecular components in biochemical processes in tissues [14] or cells [15].

This study proposes a general strategy to assess and interpret the effect of an environmental stressor on an organism by the combined use of HSI and chemometric tools. As a proof of principle we have analyzed the effect of chlorpyrifos-oxon (CPO), the biologically active metabolite of the pesticide chlorpyrifos, on zebrafish (*Danio rerio*) larvae cryosections. Zebrafish is an organism that is increasingly used as a vertebrate model in toxicology, developmental biology and drug discovery [16–19]. Zebrafish is easy to breed, has small size and a high permeability to small external molecules, which are suitable properties for environmental –omics studies [20–24]. CPO is an acetylcholinesterase inhibitor leading to neuronal and muscle toxicity on living organisms. A recent study by Faria et al.[25] described the development of a chemical model of severe acute organophosphorus poisoning (OPP) in zebrafish by CPO exposure. Interestingly, this zebrafish model of severe showed a clear retinotoxicity. Zebrafish retina closely resembles the human retina, exhibiting a similar cell layout with the lens at the top and the retinal pigment epithelium (RPE) at the bottom (Figure 1a) [26]. Therefore, the zebrafish eye has been considered an interesting target tissue to perform the HSI analysis proposed in our work.

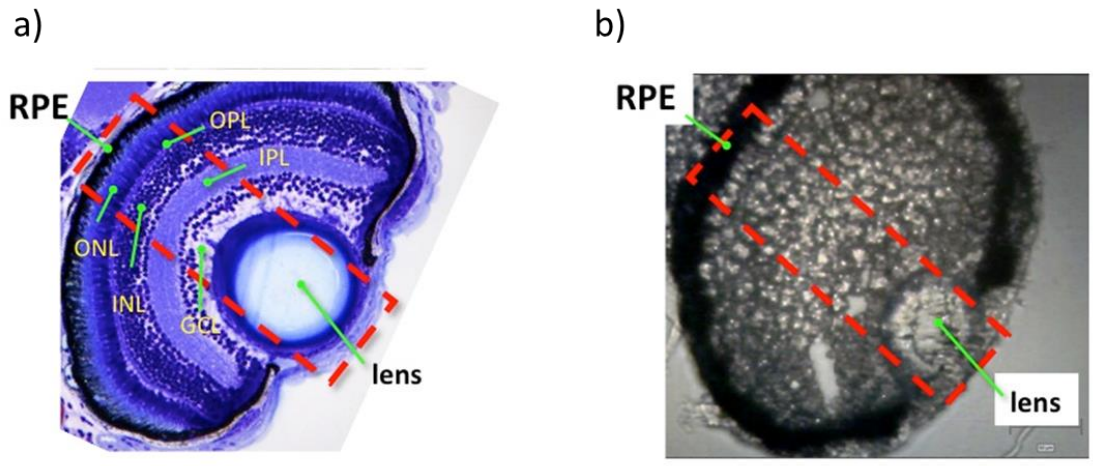


Figure 1: Retinal histology of a representative 8 days post-fertilization zebrafish control larva. a) Transverse plastic semithin section of the eye, where all the retinal layers and the lens are clearly identified; b) Cryosection of the eye. Dashed red rectangle is an approximation of the surface scanned. *RPE*: retinal pigment epithelium; *ONL*, outer nuclear layer; *OPL*, outer plexiform layer; *INL*, inner nuclear layer; *IPL*, inner plexiform layer; *GCL*, granular cell layer.

To start the study of the effect of CPO on zebrafish, two separate sets of Raman HSI are acquired on eye cryosections coming from control and CPO-exposed populations and submitted to analysis by chemometric tools. For an initial assessment of the contaminant effect, two multiset structures, formed by the images of control and CPO-exposed populations, respectively, are analyzed separately by Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS)[27–30]. MCR-ALS analysis of each multiset structure provides a single set of resolved pure spectral signatures, valid for all images in the multiset, and distribution maps related to the biological components in each of the images. In the case of biological samples, resolved components by MCR-ALS are usually formed by the signal of a mixture of molecules, i.e., the fingerprint, normally related to different kinds of tissues (or cell compartments) in the image [31]. Visual comparison of resolved spectral signatures of analogous biological components observed in the multisets of control and exposed populations has been performed in order to obtain a qualitative characterization of the effect of CPO on zebrafish cryosections at a biological component level.

In order to obtain a statistical assessment of the significance of the effect of the contaminant, partial least squares-discriminant analysis (PLS-DA) [32,33] has been performed based on the use of MCR resolved spectral signatures of analogous components in both control and CPO-exposed populations. This new approach applies PLS-DA to HSI information taking advantage of the representative and component-specific information enclosed on the resolved spectral signatures provided by MCR multiset analysis. To do so, a suitable resampling strategy

has been applied to obtain many multisets formed by different representative pixel subsets of all images belonging to the same population. All resampled multisets (from control and CPO-exposed populations) are separately analyzed by MCR-ALS. For each biological component, the sets of resolved signatures for both control and CPO-exposed resampled multisets are submitted to build a component-specific PLS-DA model. The classification parameters provide a reliable assessment of the significance of the contaminant effect and, when significant, the variable importance in projection (VIP) identifies the spectral features changing because of the contaminant exposure. Doing it in this way, the effect of CPO exposure is statistically assessed and interpreted at a biological component (tissue or sub-tissue) level.

2. EXPERIMENTAL

2.1. Fish husbandry and larvae production

Adult wild-type zebrafish were maintained in fish water [reverse-osmosis purified water containing 90 µg/ml of Instant Ocean (Aquarium Systems, Sarrebourg, France) and 0.58 mM $CaSO_4 \cdot 2H_2O$] at 28 ± 1°C in the Research and Development Centre of the Spanish Research Council (CID-CSIC) facilities under standard conditions. Embryos were obtained by natural mating and maintained in fish water at 28.5°C. Larvae were not fed during the experimental period. All procedures were conducted in accordance with the institutional guidelines under a license from the local government (DAMM 7669) and were approved by the Institutional Animal Care and Use Committee at the Spanish Research Council.

2.2. Stressor exposure and samples collection

Chlorpyrifos-oxon (CPO) (CAS#5598-15-2, 98.1% purity) was purchased from Chem Service (West Chester, USA, PA). For the severe acute OP intoxication model generation, zebrafish larvae were transferred to 48-well plates (1 larva per well) at 7 days post-fertilization (dpf) and exposed for 24 h to 3 µM CPO, in a dark incubator at 28.5°C. Control larvae were exposed to the same concentration of the carrier (0.1% DMSO) under identical conditions. The zebrafish model was characterized by a compacted head with areas of opacification at the gross morphological level. At the end of the experiment, control and treated larvae were mounted with TissueTek (O.C.T), plunge frozen in liquid nitrogen and the head was cryosectioned at 10 µm in a Leica CM30505 cryostat microtome (Leica Biosystems, Nussloch, Germany). CaF_2 optical windows were used as a support for the cryosections. Since the retina is a multilayer organ and the precise location of one specific layer may be difficult to identify in the cryosections, high quality semithin sections of the central eye of representative control and CPO-treated larvae

were obtained following the protocol described by Faria et al. and are used as a support information for result interpretation [25].

2.3. Image acquisition

Raman HSI were acquired at the Institut de Ciències Fotòniques (ICFO) by an inVia Raman Microscope spectrometer (Renishaw, Gloucestershire, UK). A 532 nm laser beam focused through a 20X objective (NA=0.4) was used as a light source. A continuous point mapping (StreamHR^{TM}) for fast imaging has been performed (1.5 s for each pixel position). The studied spectral range goes from 450 cm^{-1} to 1745 cm^{-1}, with a spectral resolution of 2 cm^{-1} and pixel size of 5×5 μm^2. For this study, eight eye cryosections of control fish and ten cryosections of CPO-exposed fish have been analyzed. All images correspond to the central part of the eyes, containing from the RPE of the retina to the lens (Figure 1b).

3. DATA TREATMENT

The data treatment includes the preprocessing of the Raman HSI, the application of MCR-ALS analysis to the multisets of the different biological sample populations separately and the use of PLS-DA for a statistical assessment and interpretation of the effect of the environmental stressor at a biological component level. All these steps are described below in detail.

Data treatment has been mainly performed using in-house made routines under MATLAB platform (MathWorks Inc., Natick, MA, USA). A graphical user interface for MCR-ALS was proposed by Jaumot et al.[29] and can be downloaded from the MCR webpage [34]. PLS-DA[32,33] analysis has been performed using the PLS-toolbox software (Eigenvector Research Inc., Manson, WA, USA).

3.1. Data preprocessing

HSI data can be described as a cube structure, with two dimensions related to pixel coordinates and one dimension to the spectral information. To perform the HSI data treatment, the cube is unfolded to create a **D** matrix with all spectra of the image one under the other. The data preprocessing applied to the images involves the following steps:

1. Elimination of irrelevant and anomalous pixels, i.e., pixels with low intensity signal that do not contain relevant information for the analysis, e.g., pixels from sample support and pixels with saturated signal. To do so, a small threshold value is set by

visual inspection (pixels removed are not used for further analysis). Saturated pixels can be easily recognized because the Raman intensity of the spectrum suddenly drops to zero. All valid pixels have Raman intensities clearly above the null signal, even if they have small values at some Raman shifts. In this way, a gap is created between low Raman intensities and null values, where it is easy to set a threshold value. Figure S1 in supplementary material shows an example of raw and preprocessed data and displays the threshold used to remove saturated pixels.

2. Spectra smoothing by a Savitzky-Golay filter with a 2nd order polynomial and 11 point spectral channels window width [35].
3. Baseline correction by Asymmetric Least Squares (AsLS) [36]. This method is based on a recursive fitting of the whole spectrum using a baseline, which is afterwards subtracted. To do so, two parameters are used to control the baseline fitting (see equation 1), one associated with the smoothness of the fit (λ) and the other with the penalty imposed to the spectral readings related to channels providing positive residuals, i.e., signal above the fitted baseline (*p*). The error function, S, minimized is shown below:

$$S = \sum_i \omega_i (y_i - z_i)^2 + \lambda \sum_i (\Delta^2 z_i)^2 \qquad \text{Eq. 1}$$

where ***y*** is the signal to correct and ***z*** the fitted baseline, ω_i=p if $y_i > z_i$ or otherwise ω_i=1-p, and $\Delta^2 z_i = z_i - 2z_{i-1} + z_{i-2}$.

The AsLS parameters have been optimized using the median spectrum of the dataset as a reference of spectrum to be baseline corrected. The parameters are adjusted until they are suitable to generate a baseline that fits the median spectrum (checked by visual inspection). Then the correction is applied to all spectra of the dataset.

3.2. Multivariate Curve Resolution-Alternating Least Squares (MCR-ALS)

Once the image preprocessing is performed, two column-wise augmented multiset structures that contain the hyperspectral images of the control and the CPO-exposed population, respectively, are built. Augmented data matrices **D** in a multiset contain different submatrices $\mathbf{D_i}$ (in this case, each $\mathbf{D_i}$ submatrix contains the pixel spectra of an image of a particular population). The control multiset is formed by pixel spectra of eight images and the CPO-exposed multiset contains pixel spectra of ten images. The multiset structures are submitted separately to MCR-ALS analysis. Information provided by multiset structures reflects appropriately the biological variability among samples and cryosections within the same population and, hence, allows for a more reliable recovery of representative spectral signatures of the existing biological components.

3.2.1. Image analysis by MCR-ALS and qualitative interpretation of contaminant exposure

Image resolution by MCR-ALS [28,30,31] allows the decomposition of a data matrix **D** formed by an individual image or by a set of images in a multiset structure into the distribution maps (**C**) and the pure spectra ($\mathbf{S^T}$) of the constituents present in the image. MCR-ALS is based on recovering the underlying spectroscopic bilinear model of the data, as shown in Equation 2, where **D** is the matrix of raw pixel spectra, **C** contains the concentration profiles of the pure components, $\mathbf{S^T}$ the related pure spectra and **E** the experimental error contained in the raw measurement and unexplained by the MCR model.

$$\mathbf{D} = \mathbf{CS^T} + \mathbf{E} \quad \text{Eq. 2}$$

For column-wise augmented multisets, the matrix **D** is formed by submatrices containing the pixel spectra of several images ($\mathbf{D_i}$). The MCR model arising from a multiset structure is formed by a single set of pure spectral signatures resolved ($\mathbf{S^T}$), common to all the images of the multiset, and a column-wise augmented concentration matrix, which can be divided in small submatrices $\mathbf{C_i}$ that correspond to each image of the multiset ($\mathbf{D_i}$) (see equation 3). The previous knowledge of the 2D geometry of each image is used to refold the $\mathbf{C_i}$ matrices into the distribution maps of the components resolved on each one of the images (see Figure 2).

$$\boldsymbol{D} = \begin{pmatrix} \boldsymbol{D_1} \\ \boldsymbol{D_2} \\ \dots \\ \boldsymbol{D_n} \end{pmatrix} = \begin{pmatrix} \boldsymbol{C_1} \\ \boldsymbol{C_2} \\ \dots \\ \boldsymbol{C_n} \end{pmatrix} \boldsymbol{S^T} + \begin{pmatrix} \boldsymbol{E_1} \\ \boldsymbol{E_2} \\ \dots \\ \boldsymbol{E_n} \end{pmatrix} \quad \text{Eq. 3}$$

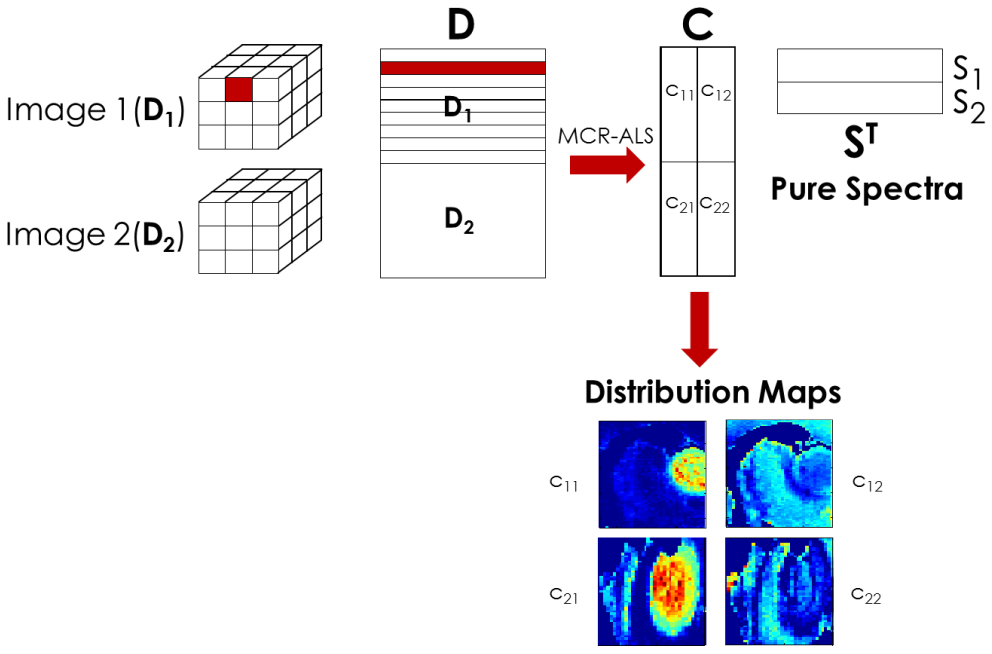


Figure 2: MCR-ALS multiset resolution for image analysis. **D** contains the unfolded images of the multiset structure, **C** the concentration profiles for each one of the images and $\mathbf{S^T}$ the pure signatures, which are common to the whole multiset. **C** can be refolded into the distribution maps of the resolved components.

MCR-ALS performs the decomposition of the raw data set **D** using an iterative alternating least-squares algorithm. Some constraints can be applied during the resolution to obtain chemically meaningful resolved profiles and to decrease the ambiguity in the final solutions. The diversity and optional application of constraints makes MCR-ALS suitable to tackle a high variety of data sets (images, processes, environmental data…). In this work, the constraints applied to deal with Raman spectra from HSI have been non-negativity on $\mathbf{S^T}$ and **C** profiles and correspondence among species, which allows setting presence/absence of components in each $\mathbf{C_i}$ submatrix of the dataset. The main steps to perform MCR-ALS are the following:

1. Determination of the number of chemical contributions in the raw data (**D**)
2. Generation of initial estimates of the $\mathbf{S^T}$ matrix using a method to select the purest image spectra based on SIMPLISMA [37].
3. Calculation of **C** and $\mathbf{S^T}$ iteratively by alternating least squares under constraints until convergence is achieved.

The number of chemical contributions has been estimated using singular value decomposition (SVD)[38]. Then, the initial estimates of the $\mathbf{S^T}$ matrix have been generated by a SIMPLISMA-based approach[37]. Finally, the least squares algorithm is applied and involves the operations $\mathbf{C} = \mathbf{DS}(\mathbf{S^TS})^{-1}$ and $\mathbf{S^T} = (\mathbf{C^TC})^{-1}\mathbf{C^TD}$ alternatingly in each iterative cycle and the corrections of the profiles according to the constraints selected. The convergence criterion is achieved when the original data is well reproduced by the bilinear model and there is no significant variation in the model fit among two consecutive iterative cycles. The parameters used to assess the quality of the model are the percentage of lack of fit (see Equation 4) and the variance explained, r^2 (see Equation 5).

$$Lof(\%) = 100 \times \sqrt{\frac{\sum_{i,j} e_{ij}^2}{\sum_{i,j} d_{ij}^2}} \qquad \text{Eq. 4}$$

$$r^2(\%) = 100 \times \left(1 - \frac{\sum_{i,j} {e_{ij}}^2}{\sum_{i,j} {d_{ij}}^2}\right) \text{Eq. 5}$$

Where d_{ij} is the *ijth* element of the original data set and e_{ij} the residual associated with the reproduction of this value by the MCR model.

The application of the correspondence among species constraint requires information of the composition of the different samples (images). When prior knowledge does not exist, as usually happens in biological samples, a previous MCR analysis on individual images helps to identify which components can be present or absent in each of the images forming the multiset [28,39,40].

MCR-ALS analysis is performed on both CPO-exposed and control image multisets. For the qualitative interpretation of the results obtained, analogous components resolved in both multisets, detected by similar morphology of the distribution maps (precise location at histological level) and similar resolved spectral signatures, are identified. The interpretation of the CPO exposure effect is mainly done by the comparison of analogous resolved spectral signatures in both populations. Relevant changes in Raman spectral features can be associated with changes in the biological compounds related to the identified bands. In case of presence of specific resolved components on only one of the multisets, these components should be considered a direct effect of the CPO exposure, whether they appear or disappear in the CPO-exposed population when compared with the control.

3.2.2. PLS-DA analysis based on MCR-ALS results

In order to perform a classification model based on PLS-DA [32,33], a sufficiently large and representative set of spectra per class and a balanced number of spectra among classes is required. Many options can be suggested for doing it, but some of them are not suitable for hyperspectral images coming from biological samples, e.g.:

a) Taking all pixel spectra of CPO-exposed and control images to build a PLS-DA model is not a good option because the sample surface (cryosection) is heterogeneous and formed by different biological components. Very often, the spectral differences among different biological components within a cryosection are far more significant than the spectral differences among populations for a particular biological component due to contaminant exposure.
b) Using average image spectra for PLS-DA models would not be a good approach either. This would lead at best to a general information about the significance of the effect of the contaminant at a general organism level, something that can be better achieved using other analytical methods. In addition, even small changes among cryosections may introduce spectral changes more related to the different representation of organism tissues in each section than to the exposure to the contaminant.

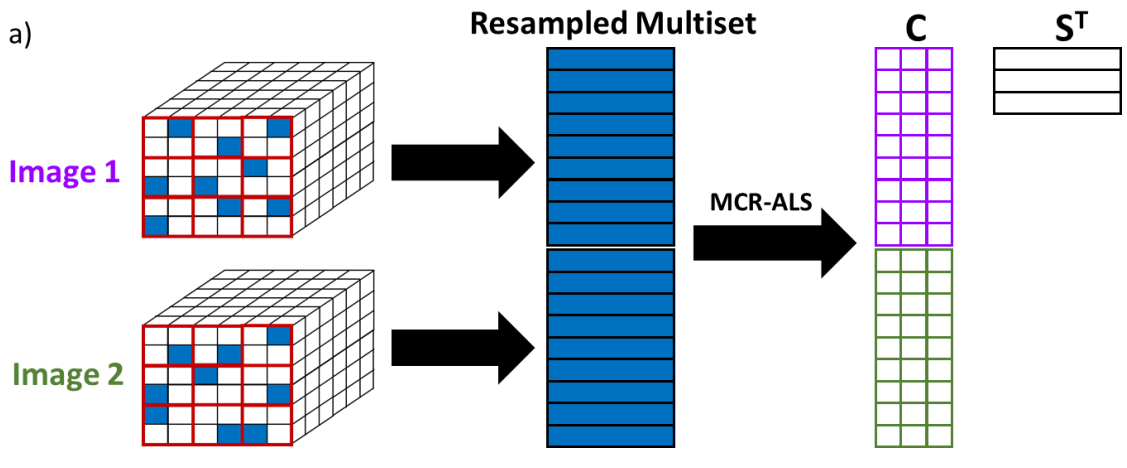
Instead, the use of resolved spectral signatures of analogous components in the compared populations obtained by MCR multiset analysis is a good strategy to alleviate all the

problems mentioned above. On the one hand, the variability due to the different individuals imaged and to the non-equivalence of cryosections within a population is suppressed because the single set of resolved spectral signatures is representative of all images in the multiset. On the other hand, the unmixing provided by MCR allows the study of the effect of the contaminant exposure at a biological component level instead of on the total organism.

At this stage, the only problem resides in the fact that MCR analysis on the complete control and CPO-exposed multisets provides a single set of resolved spectral signatures per population (i.e., a single matrix $\mathbf{S}^T$), which is insufficient to build a PLS-DA model. The resampling strategy proposed in this work is focused on obtaining a sufficient number of different spectra per class to build a PLS-DA model, keeping the advantages linked to the information provided by MCR multiset analysis. For the resampling operation, all images within the same population are divided in 2×2 pixel blocks according to the original 2D sample surface structure and the spectrum of one pixel is selected randomly within each of the blocks (see Figure 3a). This procedure is performed until the complete multiset is resampled. This local random resampling is done to ensure that representative information of all sample surface, i.e., of all biological components, is preserved, as opposed to what could happen if random pixel selection on the global image was carried out. The selected pixel subset, representative of all images in the original multiset, is submitted to MCR analysis maintaining the constraints of the original resolution (explained in section 3.2.1). A set of pure resolved spectral signatures is obtained from the resolution and is stored as well as the resolution quality parameters. This procedure is repeated as many times as necessary and a new set of pure resolved spectral signatures is obtained every time. In this way, all the spectra obtained will be related to the original dataset but will present small differences because different pixel subsets have been selected in each resampling run. Resampling done in this way provides:

a) A good estimate of the variability of each resolved spectral signature, since the pixel spectra subsets used in the resampled multisets refer to different real parts of the sample surface of images and yet always maintain representative information on the original images, and

b) A sufficient number of spectra (as many of resampled runs) per class to build a good balanced PLS-DA model.



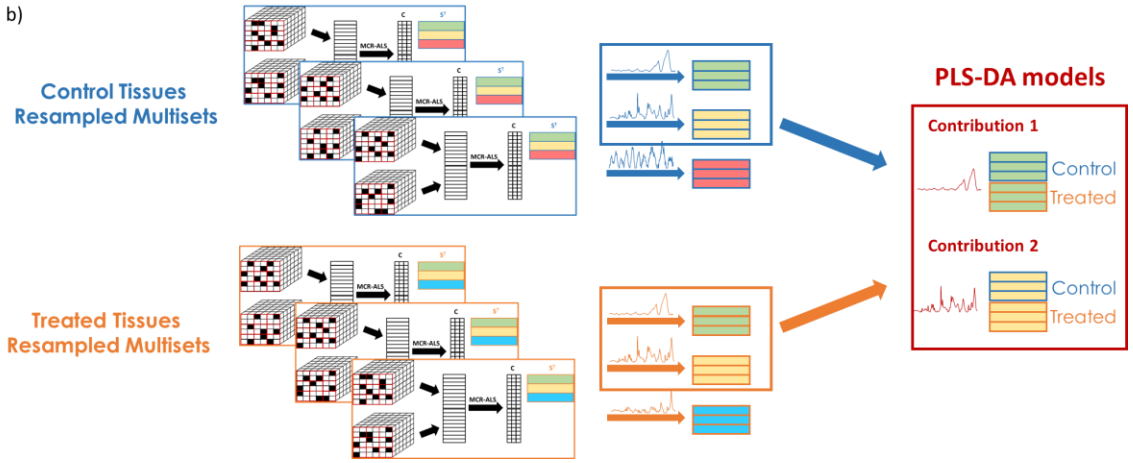


Figure 3: Scheme of image resampling strategy. a) Image multisets are divided in 2x2 pixel blocks (□) and for each block one of the pixels is randomly selected (■). The selected pixels are analyzed by MCR-ALS obtaining a single set of resolved spectral signatures ($\mathbf{S}^T$). b) Several resampled multisets from each population are analyzed by MCR-ALS. A set of resolved spectral signatures containing as many spectra as resampled multisets is obtained for each MCR contribution. A PLS-DA model is built for each analogous MCR contribution on both populations.

Once the resampling is finished, a number of PLS-DA models equal to the number of analogous resolved biological components in both population multisets is created. In each of these models, all resolved spectral signatures related to the same biological component for both control and CPO-exposed populations are grouped together to create the PLS-DA model of that specific MCR component, as shown in Figure 3b. For all PLS-DA models, 300 spectra have been used (150 from control and 150 from CPO-exposed populations). 200 of them have been used as training set for building the classification model and the other 100 as external validation set.

PLS-DA [32,33] is a classification method based on the PLS regression between the **X** block, which contains Raman spectra of the two populations studied, and the **Y** block, which is formed by the class membership information about control and CPO-exposed spectra. A venetian blinds cross-validation method (10 splits and samples assigned one by one alternatingly to each split) has been performed to decide the number of components of the models. The number of components providing the best classification rate has been adopted.

In order to avoid overfitting and to assess the reliability of the classification models, cross-validation, the use of the external validation set and permutation tests have been used. Permutation tests consist of testing the chance to obtain the same quality in classification results when building models using the original **X** and **Y** data or when using multiple data sets generated by using the **X** matrix and randomly reordering the class membership information block (**Y**). All models are generated in the same conditions as the original model i.e.

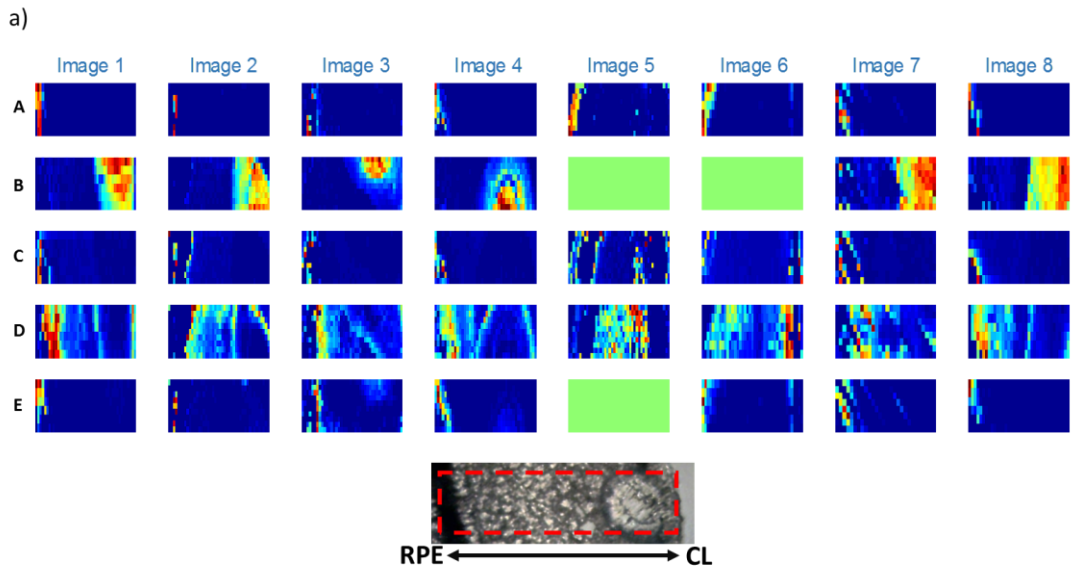
preprocessing and number of components. Each test run compares the original model with a model with wrong **Y** assignments and determines the probability that the predictions for both models are significantly different. The probabilities are calculated using pairwise Wilcoxon signed rank test, pairwise signed rank test and a randomization t-test[41,42] and the results are given at 5% significance level This procedure is performed for a number of runs (100 in this case) to help to assess the correctness of the results. The final probability values presented are obtained as the mean of the probabilities (p) obtained in each individual test run. If the average probability values are below a certain threshold (5% in our case), it means that the classification model obtained is significantly better than one coming from random correlation chance and, hence, reliable. The error in classification rate for cross-validated and external validation set is also used to assess the quality of the models.

For the qualitative interpretation of the effect of the exposure to the contaminant, the spectral information related to the discrimination of the control and CPO-exposed populations in the PLS-DA model built for each biological component has been studied. For this purpose, the variable importance in the projection (VIP) indicator is used, which provides the most relevant spectroscopic variables responsible for the discrimination among the classes. Identification of the Raman bands affected by the exposure to the contaminant for each one of the biological components resolved will help to interpret the CPO exposure effect on the zebrafish samples.

4. RESULTS AND DISCUSSION

4.1. Qualitative interpretation of CPO effect by MCR multiset image analysis

Eight images of control zebrafish and ten images of CPO-exposed zebrafish have been acquired with Raman microscopy as explained in section 2.3 and two separate multisets formed by all images of each biological population have been built and subsequently analyzed by MCR-ALS, as explained in section 3.2.1 [28,43].



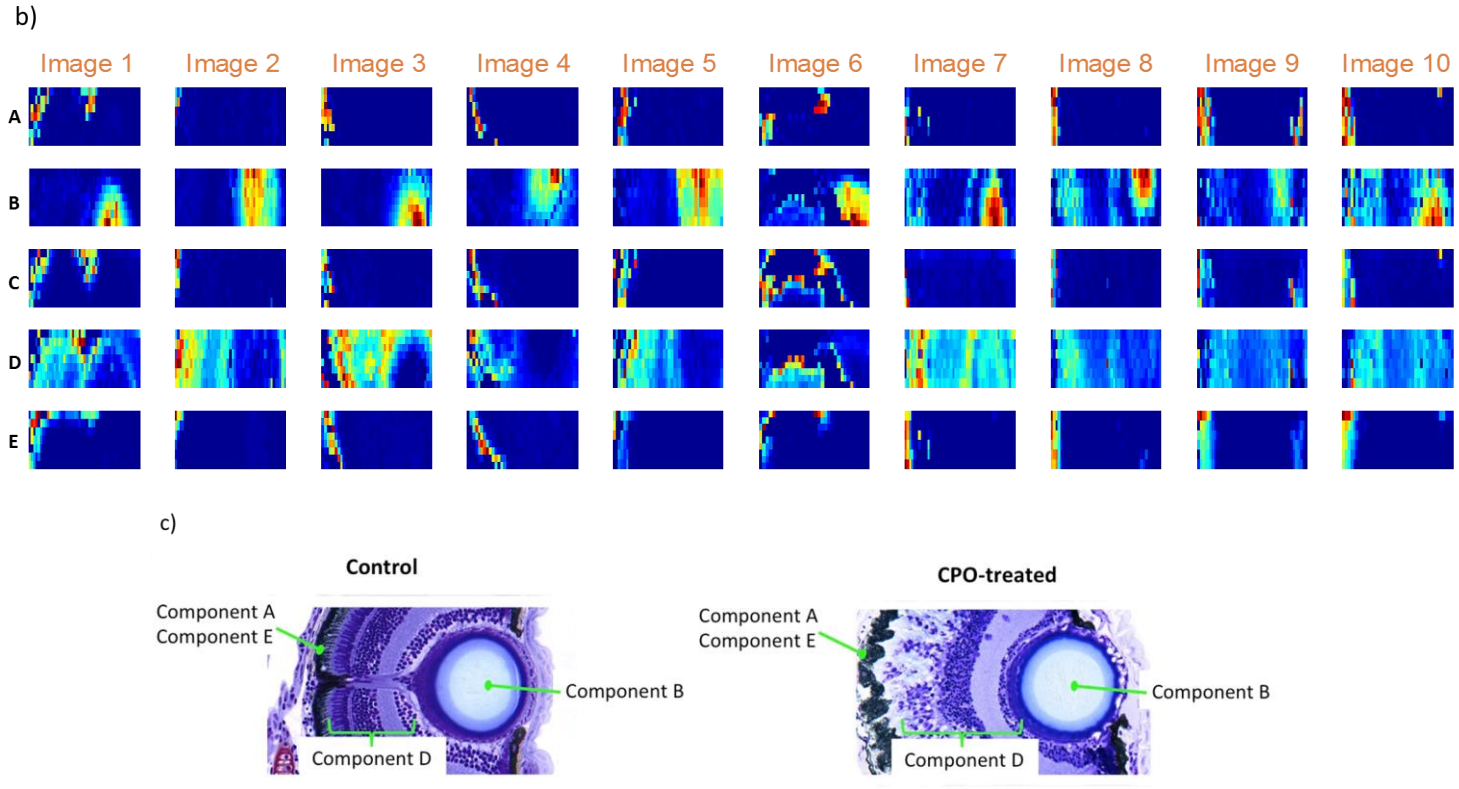


Figure 4: Distribution maps from MCR resolutions of HSI multisets. a, b) Distribution maps of control (a) and CPO-exposed (b) zebrafish larvae eyes are shown (RPE on the left and lens on the right). Green distribution maps are related to absent species. Concentration scale goes from blue color (low concentrations) to red (high concentrations). Pixel size of distribution maps has been slightly distorted to facilitate visualization of component morphological structure. C) Topographic histology of the different components in a semithin section of an eye from a representative control and CPO-exposed zebrafish larva.

Both multisets have been described by five MCR components (labeled A-E). The distribution maps for the control and CPO-treated larvae, as well as the topographic histology of the different identified components are shown in Figure 4. Figure 5 shows overlapped resolved spectral signatures of both multisets for analogous components to facilitate comparison. The lack of fit and the variance explained for the control multiset were 14.16 % and 97.99 %, respectively, and for the CPO-exposed multiset were 13.35 % and 98.22 %.

The interpretation of the MCR components has been performed by analyzing the histological topography of the distribution maps and by comparison with spectral signatures in the literature. Components A and E have been identified as two types of melanin located in the RPE, the outermost layer of the retina [44,45]. In Figure 4, it is shown that component B is located close to the center of the lens. The intense band at 1000 cm^{-1} indicates that component B presents a high amount of proteins, which also agrees with the crystalline lens composition.

Component D location corresponds with most of the retinal area, from the outer nuclear layer (ONL) to the granular cell layer (GCL), corresponding with photoreceptors, amacrine, bipolar, horizontal and ganglion cells of the retina. As Beattie et al. reported in their study of porcine eyes, this area is probably composed by a combination of fatty acids and an oxidation product of melatonin[46]. Finally, the spectral signature of component C does not look like the Raman spectra of a chemical component and its location is near the borders of other components. The fringe-pattern of this component suggests that it is an interference effect caused by light scattering in the CCD camera in the spectrometer (see supplementary material 2 for clarification). This component was necessary for an appropriate resolution of the other four components but it can be defined as a residual background of the spectral measurement and has been excluded from PLS-DA analysis. MCR allows separating biological and non-biological contributions and, as a consequence, only the relevant biological components are used for further interpretation. This advantage cannot be ensured when using methods providing bilinear decompositions based on orthogonality (PCA) or statistical independence (ICA), since most often biological and non-biological contributions are mixed in the components obtained.

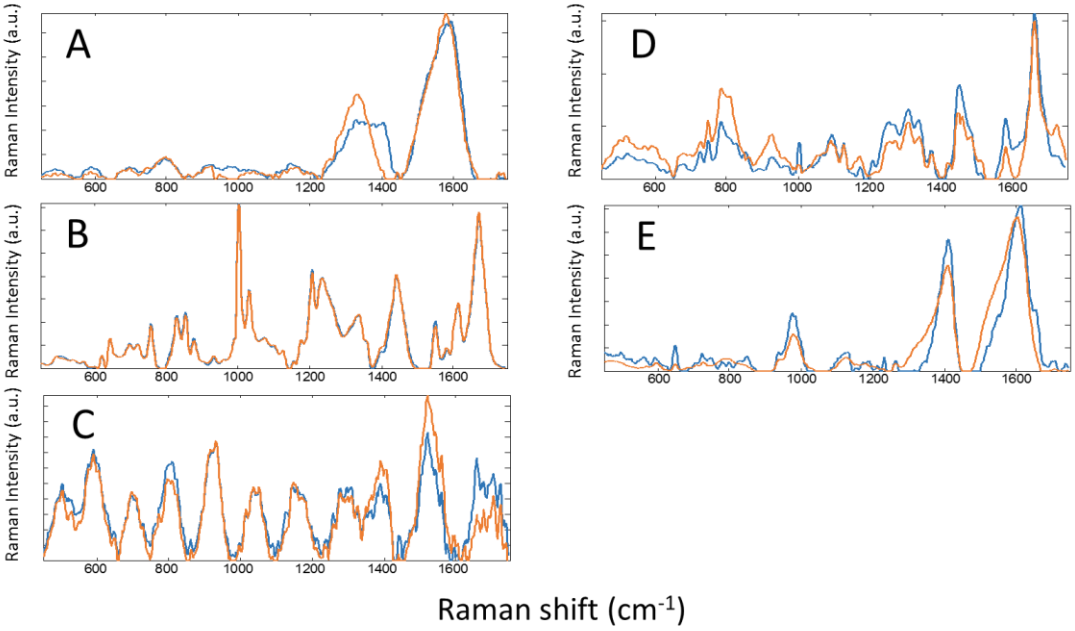


Figure 5: Pure resolved spectra in MCR multiset analysis. Blue spectra belong to the control multiset and orange spectra to the CPO-exposed multiset.

Figure 5 shows paired spectral signatures from analogous components in the multisets of both populations. A visual comparison among resolved analogous components shows that differences in component A are mainly placed in one of the broad bands (around 1200-1450 cm^{-1}). Component B, related to the lens, presents very slight differences at 1400 and 1550 cm^{-1}. Component D shows many changes between control and CPO-exposed multiset, the main ones being the rise of a band at 800 cm^{-1}, the decrease of a band at 1000 cm^{-1} and the growth of bands

from 1200 to 1600 cm^{-1}. Finally, for component E, bands at 600, 1000, 1400 and 1600 cm^{-1} have been decreased by the effect of the CPO. By simple visual inspection, it can be suggested that the CPO effect on the lens component is very subtle, whereas the effect on the two melanin spectra and the internal tissue are more important among the compared populations.

4.2. Statistical assessment of CPO effect by PLS-DA

PLS-DA[32,33] was performed to obtain a statistical assessment of the significance of the effect of CPO on each of the biological components resolved by MCR-ALS. The spectral signatures used in each one of the PLS-DA models come from the results of the MCR-ALS analysis of resampled pixel subsets obtained using the original data in each population (see section 3.2.2).

Resampling and subsequent MCR-ALS analysis of the pixel subset obtained has been repeated 150 times per population. All the resampled multisets obtained are representative of the original multisets. Hence, when resolved by MCR, the variability of the resolved spectral signatures per each resolved component is well described. For each analogous biological component in both sample populations (A, B, D and E in Figure 5), 200 resolved spectra (100 from the control and 100 from the CPO-exposed population) are submitted to build the related PLS-DA model (see Figure 6a). The remaining 100 resolved spectra (50 from control and 50 CPO-exposed) are used as external validation set to test the PLS-DA models built. As indicated above, no specific PLS-DA model was built for component C, since it is related to an instrumental artifact and, hence, does not add any valuable biological information.

Therefore, four PLS-DA models have been built on the raw data sets for each biological component resolved, as described in section 3.2.2. All models needed one latent variable to be built and the classification rate both in the calibration and the external validation set was equal to 100%. The results of permutation tests indicate that all models at a 95 % confidence level are distinguishable from models created with a random **Y**-matrix and classification results are, hence, reliable.

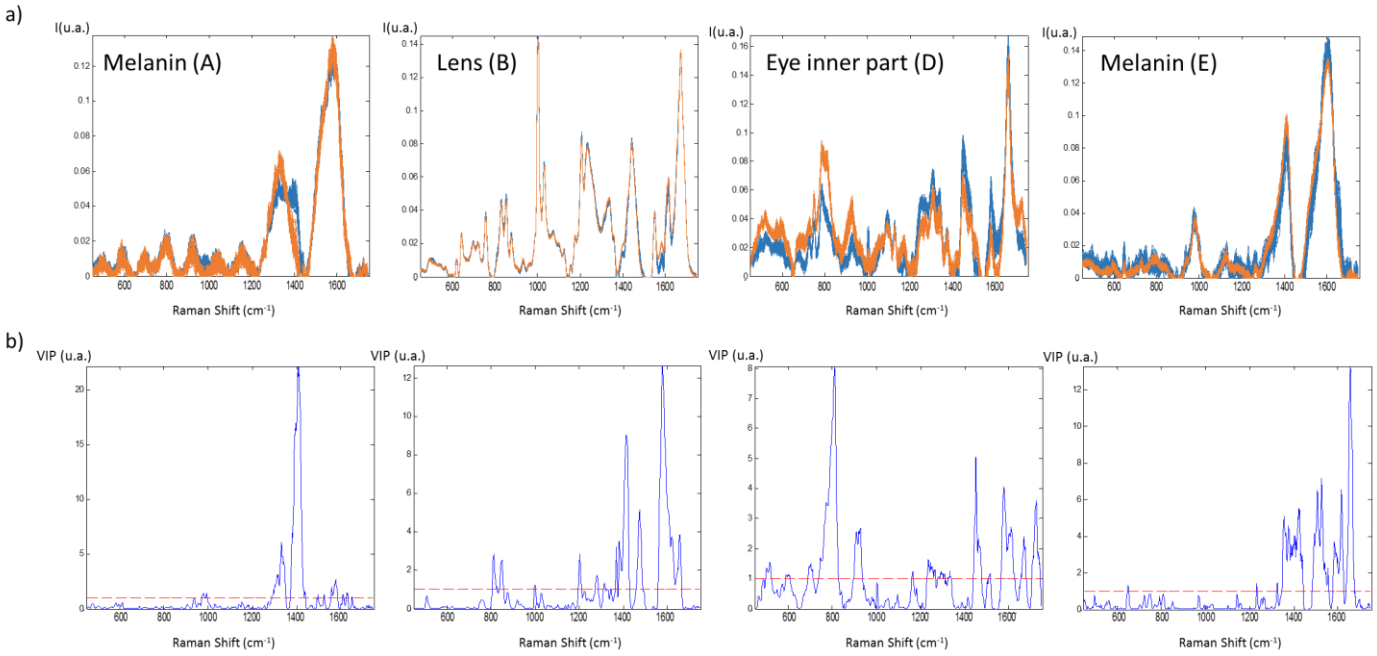


Figure 6: a) Spectral signatures of biological components obtained from MCR applied to the resampled multisets. Spectral signatures obtained from control (blue) and CPO-exposed (orange) populations related to the same biological component have been grouped together. b) Variable importance in the projection (VIP) from PLS-DA models. Variables over threshold are relevant for discrimination between control and CPO-exposed populations.

Therefore, the main conclusion is that the zebrafish exposure to CPO produces significant alterations across the retina and lens. Such a conclusion could be visually expected in components A, E (RPE) or D (from ONL to GCL), where the difference in shape among spectral signatures is evident, but was not clear for component B (lens). This last component shows a very low variability in the resolved signatures by resampling within a population and, hence, small differences in some spectroscopic features between populations, difficult to appreciate at naked eye, may be diagnosed as statistically significant.

Finally, variable importance in the projection (VIP) allows characterizing the effect of CPO exposure by identifying the spectral features more relevant in the discrimination among control and CPO-exposed populations (see Figure 6b). The Raman shifts with higher VIP values for the different components have been compared with the differences found among the spectra resolved from control and CPO-exposed images in section 4.1. (shown in Figure 5). Many spectral bands found by visual inspection are confirmed by the VIP indicator as important bands for discrimination, but additional Raman shifts are provided.

In the case of component B (lens), most of the Raman shifts that are relevant to describe the effect of contaminant exposure are not visible by visual inspection of the ordinary multiset resolutions. The opposite effect happens in the internal eye component (D), where many bands

seemed to vary by visual inspection, but only some of them were actually relevant to discriminate among classes.

Interpretation of relevant spectral features for discrimination according to the VIP parameter provides a real understanding of the contaminant effect. In biological samples, some molecular vibrations described by Raman bands can be related to a specific group of biomolecules. Component A and E have been considered types of melanin because of the broad bands presented at 1400 and 1600 cm^{-1}[44] related to the stretching in-plane of the aromatic rings and the stretching C-C within the rings respectively. VIP results show that variations in component A are mainly in the 1400 cm^{-1} band and in component E in both bands. Melanin is a pigment located at the RPE of the retina, which previous studies using different techniques found to be affected after the CPO exposure[25]. The Raman shifts related to discrimination of component B (lens) are probably related to variations in the protein composition of the crystalline; bands around 800 and 1400 cm^{-1} are related to amino acids, and bands at 1200, 1480 and 1660 cm^{-1} are probably linked to amides (III, II and I respectively, see [13] for a review on the significance on Raman bands in biological issues). Component D (retinal tissue) with high VIP values around 800 cm^{-1} may be related to effects on RNA, bands from 1450-1470 cm^{-1} are probably due to CH_2 vibrations from lipids or proteins, and bands from 1700 to 1730 cm^{-1} can be linked to stretching of C=O bonds [13].

5. CONCLUSIONS

The methodology presented in this work proves that the use of hyperspectral images in combination with MCR-ALS analysis and a classification method, such as PLS-DA, allows a qualitative interpretation and statistical assessment of the effect of a contaminant in an organism at a biological component level. The methodology lies on the joint power of images to preserve the morphology of biological samples and tissues and provide very massive and rich spectral information and on the use of multiset MCR-ALS analysis on sets of images representing each population (control and exposed) for a proper characterization of each biological component through its related spectral fingerprint.

Thus, MCR-ALS analysis of multiset structures collecting several images from the same population encloses biological and chemometric advantages. From a biological point of view, the different images in a multiset structure can include representative biological variability of organisms and tissues through the diversity of cryosections and the massive amount of spectra acquired. From a chemometric point of view, MCR-ALS takes advantage of the diversity of information to provide less ambiguous results and more reliable spectral signatures.

Using resolved spectral signatures from multiset analysis of HSI to build PLS-DA models allows working with compressed and reliable spectral information and at a biological component level. The sets of spectral signatures obtained by HSI resampling and subsequent MCR-ALS analysis are both representative and useful to express the biological variability in the populations compared. They are representative because all resampled pixel subsets contain spectra from all images and from all areas within an image (local random resampling). At the same time, they show properly the biological variability because every pixel subset resampled refers to different material parts of the samples analyzed. Besides, the use of PLS-DA allows for the assessment of the statistical significance of the exposure to a contaminant and for the qualitative interpretation of the spectral biomarkers most related to the change suffered in each biological component.

The methodology offered is general and implies two main contributions: a) the possibility to perform –omic studies investigating effects at a biological component level instead of on a global organism and b) the possibility to solve classification problems when hyperspectral images come from heterogeneous samples by using consecutive steps of resampling and unmixing by MCR-ALS and use of resolved spectral signatures as seeding information for component-specific PLS-DA models.

6. ACKNOWLEDGEMENTS

The research leading to these results has received funding from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013) / ERC Grant Agreement n. 32073 (CHEMAGEB project). The authors of this work belong to the network of recognized research groups by the Catalan government (2014 SGR 1106) and acknowledge the support of the Spanish government through project CTQ2015-66254-C2-2-P.

ICFO would like to acknowledge financial support from Laserlab-Europe (EU-H2020 654148), the Spanish MINECO (Severo Ochoa grant SEV-2015-0522), Marató de TV3 (20142030), and the National Institute of Health (NIH, grant 5R21CA187890-02). The research conducted at ICFO's Super Resolution Light Microscopy and Nanoscopy Facility has been partially supported by Fundació Cellex Barcelona.

SUPPLEMENTARY MATERIAL.

Figure S1: a) Raw data of control image 1 and b) Preprocessed data of control image 1. A zoom of the threshold used for elimination of irrelevant pixels is shown.

Figure S2: Spectra related to pixels from three different compositions of control image 1 are shown. A zoom of the spectral range where the interference is more present has been performed

for a better visualization of the effect. The fringe interference is present before and after baseline correction and, hence, is not an artefact of the AsLS method.

REFERENCES

[1] C. Bedia, N. Dalmau, J. Jaumot, R. Tauler, Environ. Res. 140 (2015) 18–31.

[2] F. Puig-Castellví, I. Alfonso, B. Piña, R. Tauler, Sci. Rep. 6 (2016) 30982–94.

[3] E. Gorrochategui, S. Lacorte, R. Tauler, F.L. Martin, Chem. Res. Toxicol. 29 (2016) 924–932.

[4] E. Garreta-Lara, B. Campos, C. Barata, S. Lacorte, R. Tauler, Metabolomics 12 (2016) 86–100.

[5] R. Jordão, B. Campos, M.F.L. Lemos, A.M.V.M. Soares, R. Tauler, C. Barata, Aquat. Toxicol. 175 (2016) 132–143.

[6] M. Navarro-Reig, J. Jaumot, A. García-Reiriz, R. Tauler, Anal. Bioanal. Chem. 407 (2015) 8835–8847.

[7] M. Bennet, A. Akiva, D. Faivre, G. Malkinson, K. Yaniv, S. Abdelilah-Seyfried, P. Fratzl, A. Masic, Biophys. J. 106 (2014) 17–19.

[8] C.R. Flach, D.J. Moore, Int. J. Cosmet. Sci. 35 (2013) 125–135.

[9] C. Zhang, D. Zhang, J.-X. Cheng, Annu. Rev. Biomed. Eng. 17 (2015) 415–45.

[10] J. Jaumot, R. Tauler, Analyst 140 (2015) 837–46.

[11] R. Salzer, G. Steiner, H.H. Mantsch, J. Mansfield, E.N. Lewis, Fresenius. J. Anal. Chem. 366 (2000) 712–6.

[12] C. Krafft, J. Popp, Anal. Bioanal. Chem. 407 (2015) 699–717.

[13] A.C.S. Talari, Z. Movasaghi, S. Rehman, I.U. Rehman, Appl. Spectrosc. Rev. 50 (2015) 46–111.

[14] M. Marro, A. Taubes, A. Abernathy, S. Balint, B. Moreno, B. Sanchez-Dalmau, E.H. Martínez-Lapiscina, I. Amat-Roldan, D. Petrov, P. Villoslada, J. Biophotonics 7 (2014) 724–734.

[15] M. Marro, C. Nieva, R. Sanz-Pamplona, A. Sierra, Biochim. Biophys. Acta 1843 (2014) 1785–95.

[16] M.C. Fishman, Science 294 (2001) 1290–1291.

[17] D.M. Parichy, M.R. Elizondo, M.G. Mills, T.N. Gordon, R.E. Engeszer, Dev. Dyn. 238 (2009) 2975–3015.

[18] C.B. Kimmel, W.W. Ballard, S.R. Kimmel, B. Ullmann, T.F. Schilling, Dev. Dyn. 203 (1995) 253–310.

[19] D.G. Howe, Y.M. Bradford, T. Conlin, A.E. Eagle, D. Fashena, K. Frazer, J. Knight, P. Mani, R. Martin, S.A.T. Moxon, H. Paddock, C. Pich, S. Ramachandran, B.J. Ruef, L. Ruzicka, K. Schaper, X. Shao, A. Singer, B. Sprunger, C.E. Van Slyke, M. Westerfield, Nucleic Acids Res. 41 (2013) 854–60.

[20] A. V. Kalueff, D.J. Echevarria, S. Homechaudhuri, A.M. Stewart, A.D. Collier, A.A. Kaluyeva, S. Li, Y. Liu, P. Chen, J. Wang, L. Yang, A. Mitra, S. Pal, A. Chaudhuri, A. Roy, M. Biswas, D. Roy, A. Podder, M.K. Poudel, D.P. Katare, R.J. Mani, E.J. Kyzar, S. Gaikwad, M. Nguyen, C. Song, Aquat. Toxicol. 170 (2015) 297–309.

[21] U. Gündel, S. Kalkhof, D. Zitzkat, M. von Bergen, R. Altenburger, E. Küster, Ecotoxicol. Environ. Saf. 76 (2012) 11–22.

[22] D. Raldúa, B. Piña, Expert Opin. Drug Metab. Toxicol. 10 (2014) 685–697.

[23] S.R. Mesquita, J. Dachs, B.L. van Drooge, J. Castro-Jiménez, L. Navarro-Martín, C. Barata, N. Vieira, L. Guimarães, B. Piña, Sci. Total Environ. 545–546 (2016) 163–170.

[24] S.R. Mesquita, B.L. Van Drooge, C. Reche, L. Guimarães, J.O. Grimalt, C. Barata, B. Piña, Environ. Pollut. 184 (2014) 555–562.

[25] M. Faria, N. Garcia-Reyero, F. Padrós, P.J. Babin, D. Sebastián, J. Cachot, E. Prats, M. Arick Ii, E. Rial, A. Knoll-Gellida, G. Mathieu, F. Le Bihanic, B.L. Escalon, A. Zorzano, A.M.V.M. Soares, D. Raldúa, Sci. Rep. 5 (2015) 15591–15.

[26] P. Goldsmith, W.A. Harris, Semin. Cell Dev. Biol. 14 (2003) 11–18.

[27] A. de Juan, R. Tauler, Crit. Rev. Anal. Chem. 36 (2006) 163–176.

[28] R. Tauler, M. Maeder, A. De Juan, in:, S.D. Brown, R. Tauler, B. Walczak (Eds.), Compr. Chemom., Elsevier B.V., 2009, pp. 473–505.

[29] J. Jaumot, A. de Juan, R. Tauler, Chemom. Intell. Lab. Syst. 140 (2015) 1–12.

[30] A. De Juan, S. Piqueras, M. Maeder, T. Hancewicz, L. Duponchel, R. Tauler, in:, Infrared Raman Spectrosc. Imaging Second Ed., 2014, pp. 57–110.

[31] S. Piqueras, L. Duponchel, R. Tauler, A. De Juan, Anal. Chim. Acta 705 (2011) 182–192.

[32] S. Wold, M. Sjöström, L. Eriksson, Chemom. Intell. Lab. Syst. 58 (2001) 109–130.

[33] M. Barker, W. Rayens, J. Chemom. 17 (2003) 166–173.

[34] Http://www.mcrals.info (2016).

[35] A. Savitzky, M.J.E. Golay, Anal. Chem. 36 (1964) 1627–1639.

[36] P.H.C. Eilers, H.F.M. Boelens, Life Sci. (2005) 1–24.

[37] W. Windig, J. Guilment, Anal. Chem. 63 (1991) 1425–1432.

[38] G.H. Golub, C. Reinsch, Numer. Math. 14 (1970) 403–420.

[39] R. Tauler, D. Barceló, Trends Anal. Chem. 12 (1993) 319–327.

[40] V. Olmos, L. Benítez, M. Marro, P. Loza-Alvarez, B. Piña, R. Tauler, A. de Juan, TrAC Trends Anal. Chem. 94 (2017) 130–140.

[41] H. van der Voet, Chemom. Intell. Lab. Syst. 25 (1994) 313–323.

[42] E. V. Thomas, J. Chemom. 17 (2003) 653–659.

[43] R. Tauler, D. Barceló, TrAC Trends Anal. Chem. 12 (1993) 319–327.

[44] Z. Huang, H. Lui, X.K. Chen, A. Alajlan, D.I. McLean, H. Zeng, J. Biomed. Opt. 9 (2004) 1198–205.

[45] S. V Saenko, J. Teyssier, D. van der Marel, M.C. Milinkovitch, BMC Biol. 11 (2013) 105–122.

[46] J.R. Beattie, S. Brockbank, J.J. McGarvey, W.J. Curry, Mol. Vis. 13 (2007) 1106–13.