

## Supplementary Material A

Adverse effects of arsenic uptake in rice metabolome and lipidome revealed by untargeted liquid chromatography coupled to mass spectrometry (LC-MS) and regions of interest multivariate curve resolution

Miriam Pérez-Cova<sup>1,2</sup>, Romà Tauler<sup>1</sup>, Joaquim Jaumot<sup>1\*</sup>

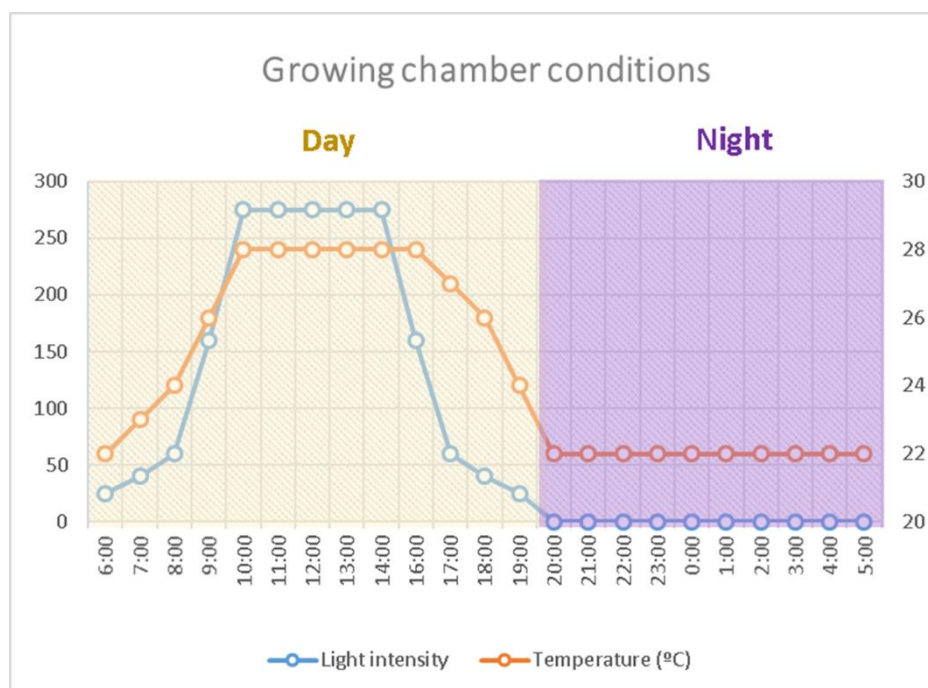
<sup>1</sup>Department of Environmental Chemistry, IDAEA-CSIC, Jordi Girona 18-26, 08034 Barcelona, Spain

<sup>2</sup>Department of Chemical Engineering and Analytical Chemistry, University of Barcelona, Diagonal 647, Barcelona, E08028, Barcelona, Spain

\* Correspondence: joaquim.jaumot@idaea.csic.es

## 1. Experimental conditions

### Conditions of rice growth in chamber MLE-352H



**Figure S1.** Experimental conditions employed for rice growing in the chamber MLE-352H: light intensity (blue) and temperature (orange). Day hours (yellow, left side of the graph), and night hours (purple, right side of the graph) are also distinguished.

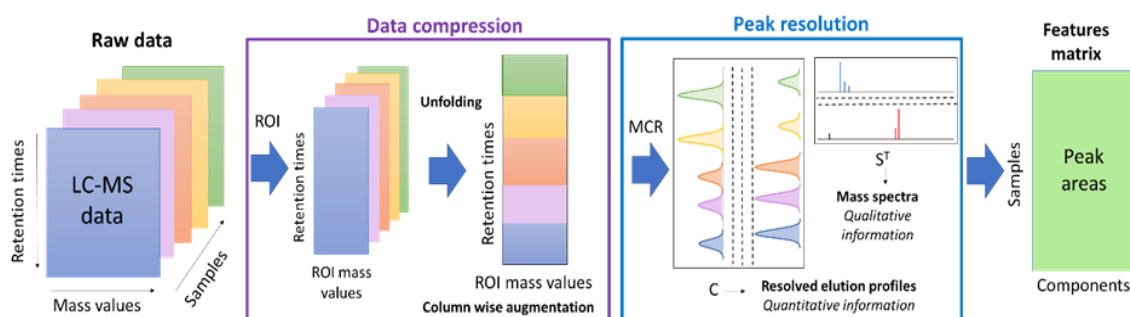
**Table S1.** Summary of the As (V) concentration levels employed in this study.

Treatment applied	Concentration value
Watering Low (WL)	1 $\mu$ M
Soil Low (SL)	15 $\mu$ M
Soil High (SH)	155 $\mu$ M
Watering High (WH)	1000 $\mu$ M

## 2. Chemometric tools: the ROIMCR procedure

ROIMCR procedure is summarized in **Figure S2**. First, regions of interest, ROI, (purple section in **Figure S2**) is employed for compressing LC-MS datasets in the spectral dimension, while creating an augmented matrix which concatenates the different samples in a column-wise manner. Then, multivariate curve resolution alternating least squares, MCR-ALS, (blue section in **Figure S2**) is applied. The compressed LC-MS datasets are resolved, providing separate information about the spectra profiles (qualitative information for identification purposes) and elution profiles (quantitative information, from the resolved peak areas). Each component from the MCR-ALS ideally represents one potential compound, joining isotopic forms and adducts in the same component (i.e. componentization). Therefore, it can be associated with a compound (i.e. lipid or metabolite). Afterwards, multivariate analysis (e.g. statistical, exploratory or classification analysis) is performed on the area matrix obtained from MCR-ALS.

A more detail description of each chemometric method is included below.



**Figure S2.** Scheme of the ROIMCR workflow. Step 1: spectral compression with regions of interest (purple). LC-MS datasets are reduced in the  $m/z$  direction by keeping only relevant  $m/z$  values. Samples are concatenated in a wise-column manner. Step 2: multivariate curve resolution alternating least square (blue). Resolution of the spectral and elution profiles. An area matrix is obtained from the integration of the resolved elution profiles.

### Regions of interest (ROI) as spectral compression

ROI approach selects  $m/z$  values below an intensity threshold established *a priori* by the user according to the signal-to-noise ratio (SN threshold) for each dataset. ROI also takes into account a mass error tolerance, related to the mass accuracy of the mass spectrometer, and a minimum number of occurrences, required for defining a chromatographic peak. A factor can also be set to establish an intensity threshold low, but only considering the features whose intensities are a multiple of this factor (e.g. min max 2, means features kept have intensities at least twice the SN threshold). ROI  $m/z$  values are searched for each retention time, and the final value will be the mean (or the median) of all the values corresponding to the same chromatographic peak. If an  $m/z$  value is detected for some samples but others no, then non-present ROIs will be set to a low random intensity value at the noise level. With this strategy, the original  $m/z$  vector is reduced for all samples simultaneously; the new vector is composed of discrete  $m/z$  values. Hence, only relevant features remain at the end of the procedure, which are considered for further analysis. More information about the MSroi app and the ROI approach can be found elsewhere [1,2]. ROI parameters employed in the analysis of each dataset, according to the platform (lipidomics or metabolomics), the tissue (roots or aerial parts) or the electrospray ionization mode (positive and negative), are included in **Table S1**.

**Table S2.** ROI parameters employed and MCR components obtained for each dataset.

ROI parameters	Lipidomics				Metabolomics			
	Roots ESI (+)	Roots ESI (-)	Aerial parts ESI (+)	Aerial parts ESI (-)	Roots ESI (+)	Roots ESI (-)	Aerial parts ESI (+)	Aerial parts ESI (-)
SN threshold	1.00E+03	3.50E+02	1.00E+03	6.50E+02	2.00E+06	1.00E+07	2.00E+06	1.00E+07
Min max factor	2	2	2	2	2	2	2	2
Mass error (ppm)	30	30	30	30	10	10	10	10
Min occurrences	100	100	100	100	100	100	100	100
Rois calculation	Median	Median	Median	Median	Median	Median	Median	Median
N° of rois obtained	217	297	221	187	347	137	345	212
N° of MCR components	150	100	110	65	150	80	150	100

### Multivariate curve resolution alternating least squares (MCR-ALS) for the resolution of the spectral and elution profiles

The decomposition provided by MCR follows a bilinear model:

$$\mathbf{D} = \mathbf{CS}^T + \mathbf{E} \quad \text{Equation (1)}$$

Where  $\mathbf{D}$  is the LC-MS data matrix with the different retention times in the rows and the  $m/z$  values in the columns;  $\mathbf{C}$  is the matrix containing the resolved elution profiles for each of the MCR components (which can ideally be associated with a single chemical compound),  $\mathbf{S}^T$  the matrix containing the spectral profiles for each component too, and  $\mathbf{E}$  the matrix with the residuals not explained by the model. If multiple samples are analyzed simultaneously, then the bilinear model extends accordingly:

$$\mathbf{D}_{\text{aug}} = \begin{bmatrix} \mathbf{D}_1 \\ \dots \\ \mathbf{D}_L \end{bmatrix} = \begin{bmatrix} \mathbf{C}_1 \\ \dots \\ \mathbf{C}_L \end{bmatrix} \mathbf{S}^T + \begin{bmatrix} \mathbf{E}_1 \\ \dots \\ \mathbf{E}_L \end{bmatrix} = \mathbf{C}_{\text{aug}} \mathbf{S}^T + \mathbf{E}_{\text{aug}} \quad \text{Equation (2)}$$

Where  $\mathbf{D}_{\text{aug}}$  is the augmented data matrix including all samples, each of them composed by a data matrix of the chromatographic run ( $\mathbf{D}_1, \dots, \mathbf{D}_L$ ). Samples are concatenated vertically, in a column-wise manner.  $\mathbf{C}_{\text{aug}}$  has the elution profiles of the components present in all analyzed samples, and  $\mathbf{S}^T$  the spectral of these components. Again,  $\mathbf{E}_{\text{aug}}$  is the residual matrix, containing the non-explained variances.

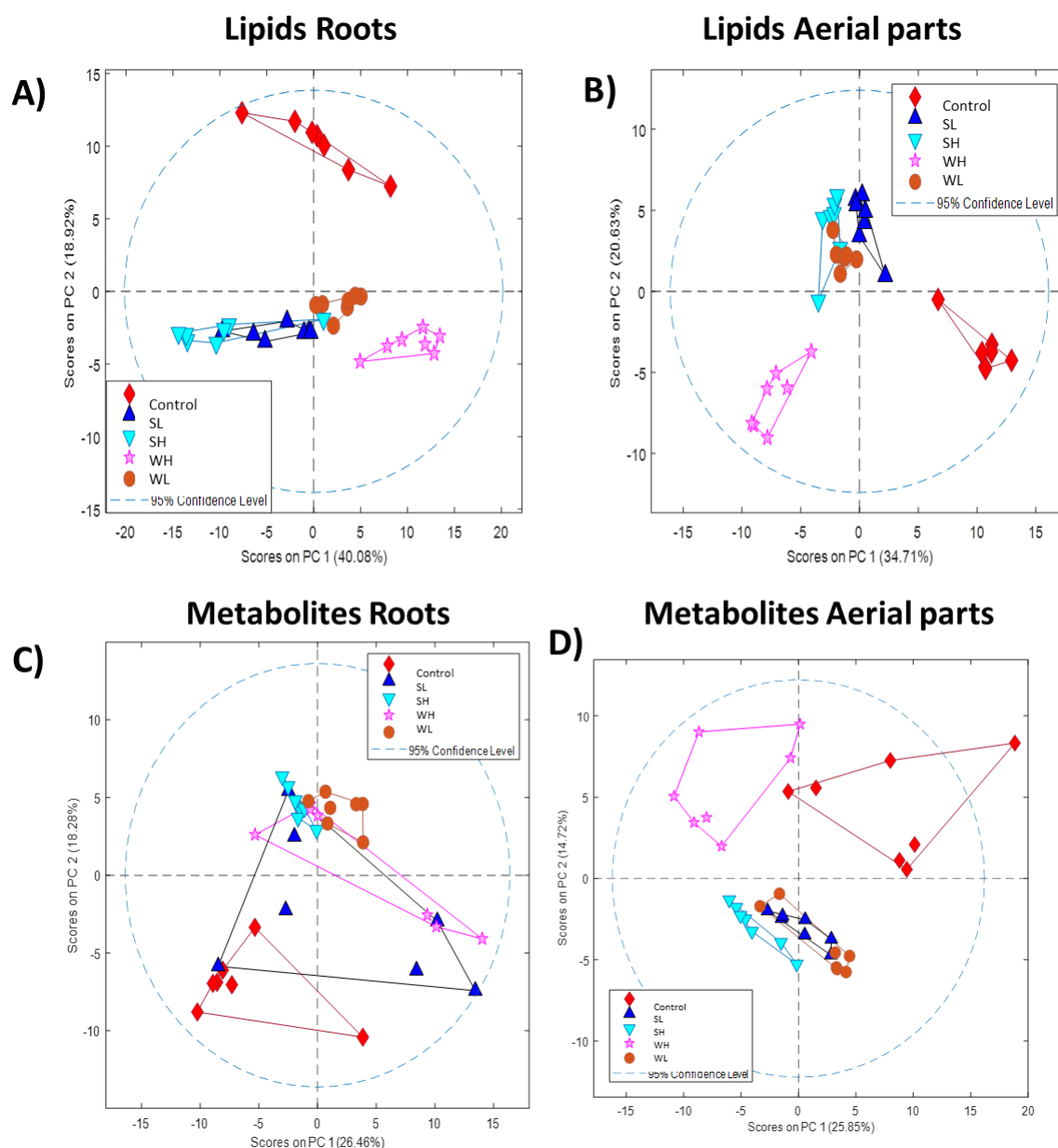
MCR-ALS is a specific version of the MCR method where the optimization step to resolve the component profiles employs an iterative alternating least squares algorithm (ALS). This approach has been described elsewhere [3]. In the first step of the MCR-ALS approach, the number of components is selected, which is initially estimated from the number of singular values, using the singular value decomposition (SVD) algorithm. In the next step, initial estimates (elution or spectral profiles) are obtained for the selected number of components. Then, the ALS iterative optimization begins. Here, optimization was started on initial estimates of pure spectra ( $\mathbf{S}^T$ ), obtained from a purest spectra detection of pure variable detection approach [4]. When convergence criterion is achieved, the process finishes. Constraints are also frequently employed to reduce ambiguities associated with the bilinear model (i.e. it does not assure unique solutions) and provide chemical meaning to the mathematical solutions. In this case, non-negativity constraint was selected for both spectral and elution profiles. Equal height was also applied for spectral normalization. The number of components obtained for each dataset is included at the bottom part of **Table S1**.

### 3. Multivariate analysis

#### Statistical assessment and exploratory analysis of arsenic exposure

**Table S3.** Statistical results from ASCA.

ASCA results	Lipidomics				Metabolomics			
	Roots ESI (+)	Roots ESI (-)	Aerial parts ESI (+)	Aerial parts ESI (-)	Roots ESI (+)	Roots ESI (-)	Aerial parts ESI (+)	Aerial parts ESI (-)
<i>C vs W vs S (treatment)</i>	1	1	1	1	1	1	1	1
<i>C vs WVL vs SL vs SM vs WH (concentration)</i>	0.0001	0.0001	0.0001	0.003	0.0001	0.0405	0.0086	0.0253
<i>Interaction</i>	1	1	1	1	1	1	1	1
<i>C vs WVL</i>	0.0001	0.0001	0.0001	0.0001	0.0001	0.0001	0.2368	0.0062
<i>C vs WH</i>	0.0001	0.0002	0.0001	0.0001	0.0001	0.0001	0.0247	0.0036
<i>C vs WVL vs WH</i>	0.0001	0.0001	0.0008	0.0013	0.0001	0.0001	0.0045	0.0205
<i>C vs SL</i>	0.0001	0.0001	0.0001	0.0001	0.0014	0.3679	0.2867	0.0013
<i>C vs SM</i>	0.0001	0.0004	0.0016	0.0001	0.0001	0.0001	0.062	0.0346
<i>C vs SL vs SM</i>	0.0001	0.0001	0.0029	0.0003	0.0001	0.1478	0.0978	0.0094



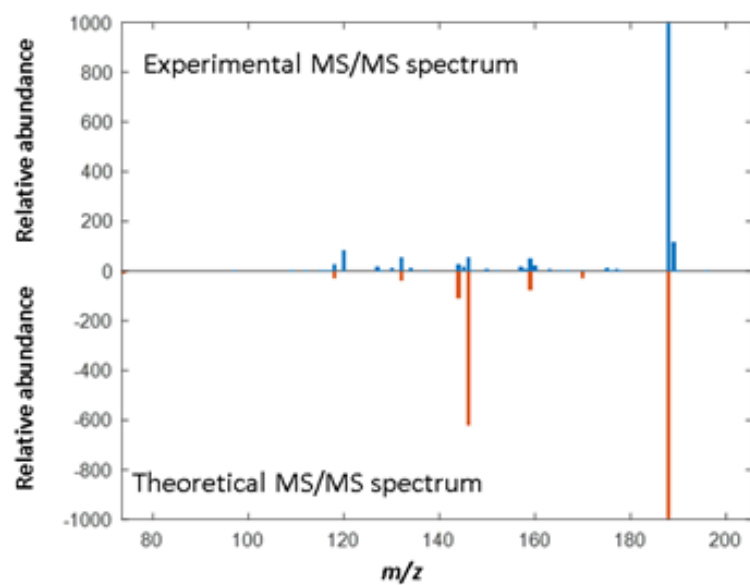
**Figure S3.** PCA score plots are shown for positive ionization mode obtained for lipidomic results for roots (A) and aerial parts (B), as well as metabolomic results for roots (C) and aerial parts (D). Both lipids and metabolites from root tissue are more affected by treatment rather than concentration level, whereas aerial parts have the opposite scenario.

## Feature selection and annotation

**Table S4.** PLS-DA results:  $n^\circ$  of Variables Important in Projection  $> 1.0$  and Matthew Correlation Coefficients.

PLS-DA results		Lipidomics				Metabolomics			
		Roots ESI (+)	Roots ESI (-)	Aerial parts ESI (+)	Aerial parts ESI (-)	Roots ESI (+)	Roots ESI (-)	Aerial parts ESI (+)	Aerial parts ESI (-)
<b>PLSDA Vips &gt; 1.0</b>	<i>C vs WH</i>	72	48	61	37	62	35	69	46
	<i>C vs WV</i>	72	47	57	35	68	44	50	43
	<i>C vs SM</i>	74	47	58	37	65	37	59	53
	<i>C vs SL</i>	70	55	57	38	60	23	53	46
<b>MCC</b>	<i>C vs WH</i>	1.0	1.0	1.0	1.0	0.9	0.9	1.0	0.7
	<i>C vs WV</i>	1.0	1.0	1.0	1.0	0.9	1.0	1.0	0.7
	<i>C vs SM</i>	1.0	1.0	1.0	1.0	0.7	1.0	1.0	0.9
	<i>C vs SL</i>	1.0	1.0	1.0	1.0	0.7	0.7	1.0	0.9





**Figure S4.** Experimental MS/MS spectrum of L-tryptophan compared to a theoretical MS/MS spectrum for the same compound in aerial tissues with ESI (+).

*Table S5. MS-DIAL parameters used in metabolomic annotation.*

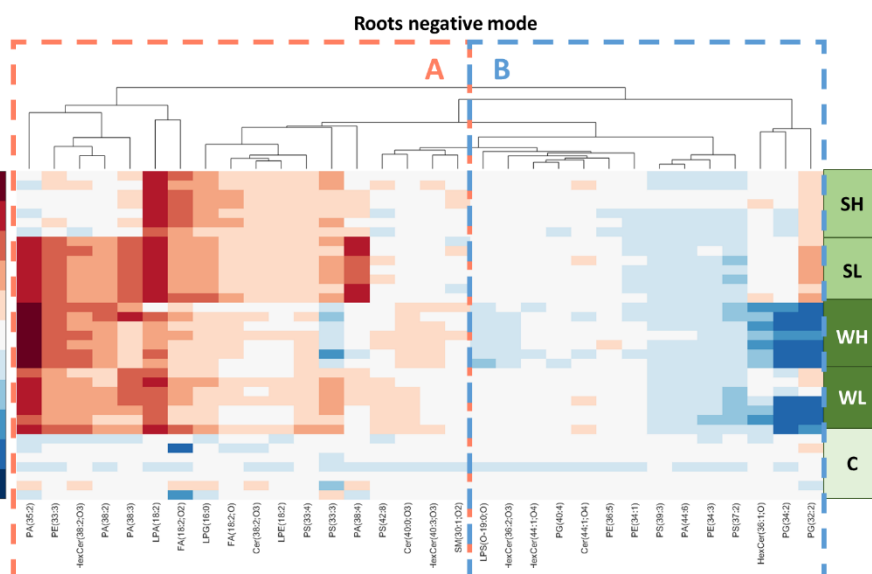
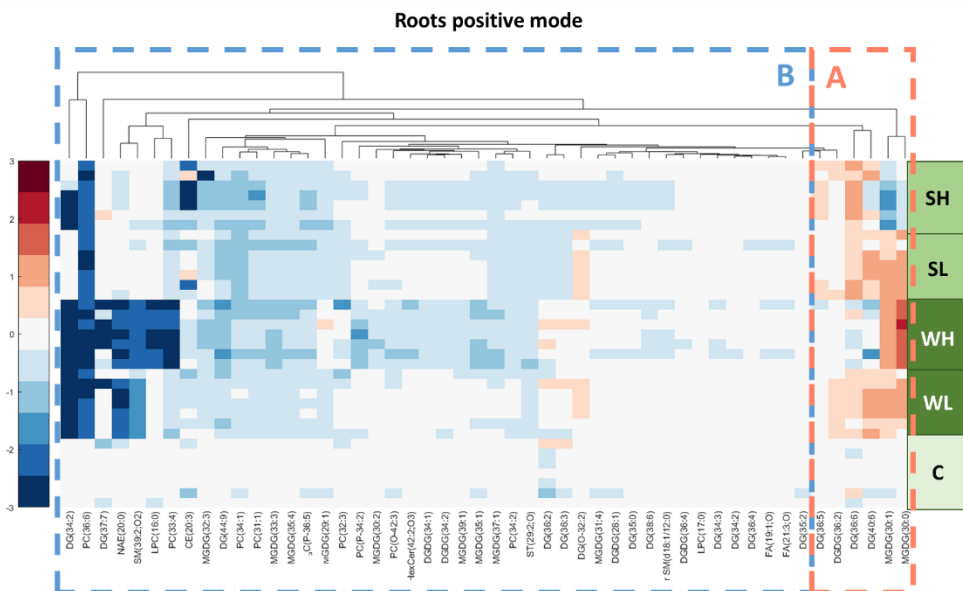
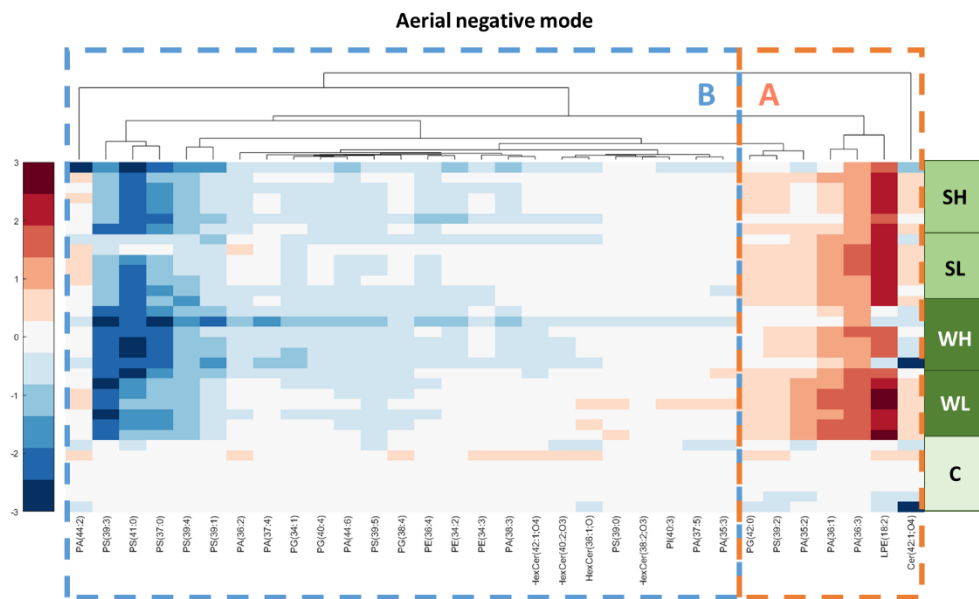
Start up a project	HILIC-HRMS method ESI(+)	HILIC-HRMS method ESI(-)
<b>Ionization type</b>	Soft ionization	Soft ionization
<b>Separation type</b>	Chromatography (LC)	Chromatography (LC)
<b>Method type</b>	SWATH-MS or conventional All-ions method	SWATH-MS or conventional All-ions method
<b>Data type (MS1)</b>	Profile	Profile
<b>Data type (MS/MS)</b>	Profile	Profile
<b>Ion mode</b>	Positive ion mode	Negative ion mode
<b>Target omics</b>	Metabolomics	Metabolomics
<b>Data collection</b>		
<b>MS1 tolerance</b>	0.01	0.01
<b>MS2 tolerance</b>	0.1	0.1
<b>Retention time begin</b>	0	0
<b>Retention time end</b>	20	20
<b>Mass range begin</b>	90	90
<b>Mass range end</b>	1000	1000
<b>Maximum charged number</b>	2	2
<b>Consider Cl and Br elements</b>	Unchecked	Unchecked
<b>Number of threads</b>	20	20
<b>Execute retention time corrections</b>	Unchecked	Unchecked
<b>Peak detection</b>		
<b>Minimum peak height</b>	500	500
<b>Mass slice width</b>	0.1	0.1
<b>Smoothing method</b>	Linear weighted moving average	Linear weighted moving average
<b>Smoothing level</b>	3	3
<b>Minimum peak width</b>	5	5
<b>Exclusion mass list (tolerance: 0.01Da)</b>	Not used	Not used
<b>MS2Dec</b>		
<b>Sigma window value</b>	0.5	0.5
<b>MS2Dec amplitude cut off</b>	100	100
<b>Exclude after precursor</b>	Checked	Checked
<b>Keep isotope until</b>	0.5	0.5
<b>Keep the isotopic ion w/o MS2Dec</b>	Unchecked	Unchecked
<b>Identification</b>		
<b>Retention time tolerance</b>	0.5	0.5

Accurate mass tolerance (MS1)	0.01	0.01
Accurate mass tolerance (MS2)	0.1	0.1
Identification score cut off	70	70
Use retention time for scoring	Unchecked	Unchecked
Use retention time for filtering	Unchecked	Unchecked
Postidentification	Not used	Not used
<b>Adduct</b>		
Molecular species	[M+H] <sup>+</sup> , [M+NH <sub>4</sub> ] <sup>+</sup> , [M+H-H <sub>2</sub> O] <sup>+</sup>	[M-H] <sup>-</sup> , [M+Hac-H] <sup>-</sup> , [M-H-H <sub>2</sub> O] <sup>-</sup>
<b>Alignment</b>		
Retention time tolerance	0.2	0.2
MS1 tolerance	0.02	0.02
Retention time factor	0	0
MS1 factor	1	1
Peak count filter	0	0
N% detected in at least one group	0	0
Remove feature based on blank information	Unchecked	Unchecked
Sample average / blank average	5	5
Keep "reference matched" metabolite features	Checked	Checked
Keep "suggested (w/o MS2)" metabolite features	Unchecked	Unchecked
Keep removable features and assign the tag	Checked	Checked
Gap filling by compulsion	Checked	Checked
<b>Isotope tracking</b>		
	Not used	Not used

**Table S6.** Experiment file used in MS method type section from start a project window in MS-DIAL

<b>Experiment file</b>			
<i>Experiment</i>	<i>MS Type</i>	<i>Min m/z</i>	<i>Max m/z</i>
<b>0</b>	SCAN	90	1000
<b>1</b>	MSMS	90	1000

## Discussion of lipidomic results



**Figure S5.** Hierarchical clustering heatmaps applied to the logarithm of the fold changes of the annotated lipids for aerial in positive mode, roots in positive and negative modes, respectively. A color intensity bar is included on the left side of the figure, indicating the relative abundance of the lipid regarding control samples (higher abundance in red, lower abundance in blue). Two clusters are differentiated in all cases, with an increasing abundance (A) and a diminishing abundance (B).



## References

1. Pérez-Cova, M.; Bedia, C.; Stoll, D.R.; Tauler, R.; Jaumot, J. MSroi: A Pre-Processing Tool for Mass Spectrometry-Based Studies. *Chemometrics and Intelligent Laboratory Systems* **2021**, *215*, doi:10.1016/j.chemolab.2021.104333.
2. Gorrochategui, E.; Jaumot, J.; Tauler, R. ROIMCR: A Powerful Analysis Strategy for LC-MS Metabolomic Datasets. *BMC Bioinformatics* **2019**, *20*, 1–17, doi:10.1186/s12859-019-2848-8.
3. de Juan, A.; Jaumot, J.; Tauler, R. Multivariate Curve Resolution (MCR). Solving the Mixture Analysis Problem. *Analytical Methods* **2014**, *6*, 4964–4976, doi:10.1039/c4ay00571f.
4. Windig, W.; Stephenson, D.A. Self-Modeling Mixture Analysis of Second-Derivative Near-Infrared Spectral Data Using the Simplisma Approach. **1992**, *64*, 2735–2742.