



HAL
open science

Untargeted food contaminant detection using UHPLC-HRMS combined with multivariate analysis: feasibility study on tea

Grégoire Delaporte, Mathieu Cladière, Delphine Jouan-Rimbaud Bouveresse,
Valérie V. Camel

► To cite this version:

Grégoire Delaporte, Mathieu Cladière, Delphine Jouan-Rimbaud Bouveresse, Valérie V. Camel. Untargeted food contaminant detection using UHPLC-HRMS combined with multivariate analysis: feasibility study on tea. *Food Chemistry*, 2019, 277, pp.54-62. 10.1016/j.foodchem.2018.10.089. hal-02628598

HAL Id: hal-02628598

<https://hal.inrae.fr/hal-02628598>

Submitted on 9 May 2022

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.

1 **Untargeted food contaminant detection using UHPLC-HRMS combined with multivariate analysis:**
2 **feasibility study on tea**

3 Grégoire Delaporte, Mathieu Cladière, Delphine Jouan-Rimbaud Bouveresse¹, Valérie Camel*

4 *UMR Ingénierie Procédés Aliments, AgroParisTech, Inra, Université Paris-Saclay, 91300 Massy, France*

5 ¹ *present address: UMR Physiologie de la Nutrition et du Comportement Alimentaire, AgroParisTech, Inra,*
6 *Université Paris-Saclay, 75005 Paris, France*

7 * *Corresponding author: AgroParisTech, 16 rue Claude Bernard, F-75005, Paris, France*

8 *Phone: +33 1 44 08 17 25 – email: valerie.camel@agroparistech.fr*

9

10 **Abstract**

11 Powerful data pretreatment strategies inspired from the field of metabolomics were adapted to chemical food
12 safety context to enable samples discrimination by multivariate methods based on low abundance ions. A
13 highly automated workflow was produced. The open-source XCMS package was used and efficient data
14 filtration strategies were set up. Data were treated using Independent Components Analysis, and data mining
15 strategies developed to automatically detect and annotate ions of low abundance by coupling blind data
16 exploration strategies with a broad scale database approach. Our method was efficient in discriminating tea
17 samples based on their contamination levels (even at 10 µg.kg⁻¹) and detecting unexpected impurities in the
18 spiking mix. Several “tracer” contaminants were considered, covering a broad range of physicochemical
19 properties and structural diversity with overall 66% detected and annotated blindly. The methodology was
20 successfully applied to a data set exhibiting only 3 “tracer” contaminants (at 50 µg.kg⁻¹) and more product
21 diversity.

22 **Keywords:** Independent Components Analysis; XCMS; ToF; Chemical food safety; Non-targeted approaches;
23 Unexpected contaminants

24 1. INTRODUCTION

25 In recent years, a new field of food research called “Foodomics”, defined as “a discipline that studies the food
26 and nutrition domains through the application of omics technologies”, has emerged (Cifuentes, 2009). Even
27 though foodomics was first intended for the study of major constituents of food products, related for instance
28 to their authenticity (Ortea et al., 2012), its potential contribution to trace analysis for chemical food safety
29 assessment has also been raised (Antignac et al., 2011). However, chemical food safety assessment faces
30 several challenges since food matrices are highly complex, chemical contaminants are numerous and diverse
31 (pesticides, mycotoxins, process-induced toxicants or migrants from packaging), and often found at trace levels
32 (down to $\mu\text{g}\cdot\text{kg}^{-1}$) (Antignac et al., 2011). Classical targeted analysis strategies are limited since unexpected or
33 unknown contaminants remain non-detected (Tengstrand, Rosén, Hellenäs, & Åberg, 2013). Therefore, there
34 is a rising interest (and demand) for developing untargeted (also referred as “non-targeted”) analyzes of
35 contaminants in food products using a relevant instrumental / analytical strategy combination, as pointed out
36 by numerous reviews (Antignac et al., 2011; Castro-Puyana, Pérez-Míguez, Montero, & Herrero, 2017;
37 Knolhoff & Croley, 2016; Lehotay, Sapozhnikova, & Mol, 2015). There is a consensus on the fact that
38 hyphenated techniques (especially high resolution instrumentation operated in full-scan mode like Ultra-High
39 Pressure Liquid Chromatography coupled to High Resolution Mass Spectrometry, UHPLC-HRMS) are key
40 technologies for this application thanks to their sensitivity and broad analytical scope, as well as the formula
41 information they provide on potential contaminants (Antignac et al., 2011; Castro-Puyana et al., 2017). Our
42 previous study combining generic extraction and UHPLC-HRMS showed its efficiency on analysis of a wide
43 variety of contaminants with a large range of physicochemical properties (Cladière, Delaporte, Le Roux, &
44 Camel, 2018).

45 Among new analytical strategies proposed for global chemical food safety assessment, suspect-screening
46 approaches (Gómez-Ramos, García-Valcárcel, Tadeo, Fernández-Alba, & Hernando, 2016; Gosetti,
47 Mazzucco, Gennaro, & Marengo, 2016) and relevant chemical patterns detection using data mining tools
48 (Cotton et al., 2014) show interesting performances in terms of sensitivity. They do not require any initial
49 analysis of the non-contaminated food product but they both rely on *a priori* hypotheses on the chemical
50 structures of contaminants. On the opposite, untargeted metabolomics-like strategies require the analysis of a
51 reference food product (to compare signals between a control and a suspect group for differences detection),

52 without any *a priori* hypotheses on the structure of potential contaminants. Hence, only untargeted strategies
53 based on tools from the field of metabolomics might enable the real “blind” detection of unknown or
54 unexpected trace molecules in complex food samples. The main characteristic of such untargeted approaches
55 lies in the generation of a very high number of signals (several thousand for a single sample). Therefore,
56 powerful data analysis strategies must be set up to increase the probability to detect contaminated samples.
57 Early results showed their potential in detecting unexpected compounds in food products (Inoue et al., 2015;
58 Knolhoff, Zweigenbaum, & Croley, 2016; Kunzelmann, Winter, Åberg, Hellenäs, & Rosén, 2018; Tengstrand
59 et al., 2013). However, food matrices studied remain relatively simple (orange juice, milk) with either high
60 levels of contamination (near $\text{mg}\cdot\text{kg}^{-1}$) (Tengstrand et al., 2013) or low molecular diversity of chemical
61 contaminants (Inoue et al., 2015; Knolhoff et al., 2016). The latest published paper (Kunzelmann et al., 2018)
62 shows promising results in terms of sensitivity (contamination detection down to $25\ \mu\text{g}/\text{kg}$), but only focused
63 on pesticides. More work is therefore needed to develop such untargeted strategies in the food safety field,
64 especially considering even lower contamination levels (down to $10\ \mu\text{g}\cdot\text{kg}^{-1}$ as frequently required by the
65 European regulation) and a wider contaminants diversity (including migrants from packaging and process-
66 induced toxicants). To that end, the method proposed here relies on the combination of three tools to take full
67 advantage of UHPLC-HRMS data: (i) data filtration based on univariate statistics, (ii) separation of sample
68 groups and highlighting of discriminating ions using Independent Components Analysis (ICA), an
69 unsupervised multivariate method based on source signals decomposition (Rutledge & Jouan-Rimbaud
70 Bouveresse, 2015), (iii) automated data mining-tools to help the annotation of discriminating ions. Thus, our
71 data analysis strategy combines the use of XCMS open-source R package (Smith, Want, O’Maille, Abagyan,
72 & Siuzdak, 2006) and ICA method: to the best of our knowledge, this combination for MS data analysis in
73 untargeted food safety analysis is successfully performed for the very first time. Unless previous untargeted
74 approaches reporting the use of either vendor (Knolhoff et al., 2016) or in-house tools (Tengstrand et al., 2013),
75 that often work as “black boxes”, our approach benefits from using a freely available package that exists for
76 more than 10 years and is supported by a dynamic and worldwide scientific community, which made it become
77 very versatile for MS data analysis. On top of that, it became user friendly thanks to the development of free
78 web-based platforms like XCMS-Online (Tautenhahn, Patti, Rinehart, & Siuzdak, 2012) or
79 Workflow4Metabolomics (Giacomoni et al., 2015).

80 Tea has been chosen as the development foodstuff, for it is the most consumed hot beverage in the world
81 (Chang, 2015). Moreover, in its raw product form (tea leaves), it is classified as a difficult commodity by the
82 European Commission (SANTE/EU, 2015) which makes it very interesting as a methodological development
83 food sample. In addition, tea is frequently produced under remote areas where agricultural and production
84 practices may be less controlled than in Europe. Therefore this food product is the subject of frequent alerts on
85 the European Rapid Alert System for Food and Feed, relative to non-authorized pesticides or contaminant
86 levels above regulated limits (i.e. minimum 10 $\mu\text{g.kg}^{-1}$ for most pesticides). Finally, recent metabolomics-like
87 approaches have been reported on tea but they only focus on quality and authenticity issues (Fraser et al., 2013;
88 Pongsuwan et al., 2008), i.e. on major constituent. The methodology that we propose here focuses on trace
89 compounds, with specific analytical methods and data treatment strategies to be set up to achieve their
90 detection.

91 **2. MATERIAL AND METHODS**

92 **2.1 CHEMICALS AND REAGENTS**

93 Acetonitrile (ACN) (HPLC plus gradient, LC/MS), water, methanol (MeOH) and formic acid (FA) (all LC/MS
94 grade) were purchased from Carlo Erba. Ultrapure water (Milli-Q[®]) was produced by an Integral 3 water
95 purification system from Millipore[®]. The compound used for ToF-MS calibration was Leucine Enkephalin
96 (LC/MS grade), purchased from Waters[®].

97 Analytical standards solutions (100 $\mu\text{g.mL}^{-1}$ in ACN or MeOH) for 21 pesticides, 4 mycotoxins, 2 process-
98 induced toxicants and labelled compounds acrylamide-d3, dimethoate-d6 and malathion-d6 were purchased at
99 CIL Cluzeau France. Ochratoxin-d5, bisphenols A, F and S, bisphenol A diglycidyl ether (BADGE), bisphenol
100 F diglycidyl ether (BFDGE) and bisphenol A-d14 (purity > 99%) were provided by Sigma Aldrich (Saint-
101 Quentin Fallavier, France). Two pooled stock solutions containing respectively all non-labelled molecules
102 (each at 1 $\mu\text{g.mL}^{-1}$), and all labelled molecules (each at 1 $\mu\text{g.mL}^{-1}$) were prepared in ACN and stored in the
103 fridge. Regularly, target analyzes of these solutions were done to check for their stability.

104

105

2.2 SAMPLE COLLECTION AND STUDY SET-UP

106
107 The goal of this work is to assess the ability of a workflow based on UHPLC-HRMS and chemometrics
108 methods (including multivariate analysis) to blindly detect an unexpected contamination in a food sample. To
109 that end, a study was designed so that this workflow would face two very different situations: (i) a quite
110 homogeneous product (i.e. samples from one brand) contaminated at several levels by a large number of
111 molecules (development data set) (ii) a heterogeneous product (i.e. samples from two brands) contaminated at
112 a single level by only few molecules (validation data set). Raw data sets have been deposited to the EMBL-
113 EBI MetaboLights database (DOI: 10.1093/nar/gks1004. PubMed PMID: 23109552) with the respective
114 identifiers MTBLS752 and MTBLS754 for data set n°1 (development) and n°2 (validation of the approach)
115 (Haug et al., 2013). The complete data sets can be accessed at
116 <https://www.ebi.ac.uk/metabolights/MTBLS752> and <https://www.ebi.ac.uk/metabolights/MTBLS754>.

117 Green teas were purchased at local retailers (Paris, France) and crushed in our laboratory using a mortar and a
118 pestle. Green tea n°1 is a Japanese Bancha tea, and green tea n°2 a Chinese tea. They were used to generate
119 the development data set (green tea n°1) and the validation data set (green teas n°1 & 2).

120 Two spiking mixes were prepared. The first one (mix n°1), intended for the development data set, consists in
121 a pool of 32 chemical contaminants (a detailed list of compounds used can be found in Supplementary data,
122 Table S.1). These target molecules, called “tracers”, were chosen to be representative of potential
123 contaminants, both in terms of chemical structures, source types (mycotoxins, pesticides, process-induced
124 toxicants and migrants from packaging) and analytical behavior (instrumental response, peak width, retention
125 time and adduct / isotopic information). The second one (mix n°2) consists in selected three contaminants from
126 the previous list, chosen for their chemical diversity, namely ochratoxin A (OTA), bisphenol S (BPS) and
127 tolfenpyrad.

128 For the development data set, four samples were considered (each time three sub-samples were collected to
129 obtain triplicates of preparation): three samples spiked with mix n°1 at 10, 50 or 100 µg.kg⁻¹, and a control
130 sample (i.e. spiked only with the ACN solvent). Tea samples were initially analyzed using a classical multi-
131 residue method in order to check for the absence of the “tracers” considered (Cladière et al., 2018).

132 For the validation data set, four samples were considered as well (again three sub-samples were collected each
133 time to get triplicates of preparation): two control samples respectively made of unspiked green tea n°1 and
134 unspiked green tea n°2, and two suspect samples respectively composed of green tea n°1 or green tea n°2
135 spiked at a level of 50 µg.kg⁻¹ with mix n°2.

136 Spiking levels were chosen in accordance to EU regulation No 396/2005 and 1881/2006. Some of the least
137 sensitive compounds (namely deoxynivalenol, bisphenols A and F) were spiked with a magnification factor of
138 5, and for the same reason acrylamide was spiked with a factor of 10. In addition, all sample groups were
139 systematically spiked (at 40 µg.kg⁻¹) with the pool of labelled molecules for analytical quality control purpose.
140 For spiking, samples of 1 g were weighted in centrifuge polypropylene tubes (Corning, New York, USA), and
141 spiking was performed using the lowest possible volume of solution (maximum of 100 µL). After spiking,
142 samples were homogenized using a vortex and allowed to equilibrate for 2 hours at room temperature.

143 The workflow employed (both for sample preparation and data treatment) is shown in **Figure 1**.

144 **2.3 ANALYTICAL METHOD**

145 The generic analytical method is based on previous work (Cladière et al., 2018). Tea samples were extracted
146 using direct solvent extraction with 5 mL of an ACN/MeOH (90/10 v/v) mixture acidified with 0.1% FA, and
147 tubes were agitated upside-down on an agitating plate during 1 h before centrifugation at 3,000 g for 10 min.
148 The supernatant was then collected and an aliquot (2 mL) was evaporated to dryness under a gentle stream of
149 nitrogen. The extract was further reconstituted in 0.2 mL of ACN acidified with 0.1% FA. Then 0.8 mL of
150 ultrapure water with 0.1% FA was added in order to reconstitute 1 mL of final volume, and centrifuged at
151 12,000 g for 10 min. At the end, 0.5 mL of the final extract was sampled and filtered at 0.2 µm using a
152 syringeless filter vial (mini-uniprep G2, Whatman) before analysis. A Quality Control (QC) sample for each
153 data set was prepared by pooling together 0.2 mL of final extract from every sample of the set; an aliquot of
154 0.5 mL was then taken and filtrated at 0.2 µm using a syringeless filter vial.

155 Analyzes were performed on a Waters® Acquity UPLC® H-Class system, composed of a quaternary solvent
156 manager pump, a refrigerated sample manager Flow-Through-Needle and a column oven, coupled to a
157 Waters® high resolution Time-of-Flight mass spectrometer Xevo® G2-S ToF operated in centroid mode
158 (UHPLC/HRMS-ToF) with a mass range from *m/z* 60 to 800. 10 µL of the final extract were injected and

159 separation was performed on a C18-PFP column (150×2.1 mm, 2 μm particles diameter, ACE supplied by AIT
160 France). An electrospray ionization source was used in both positive (ESI⁺) and negative (ESI⁻) modes. ESI⁺
161 and ESI⁻ modes were run separately. For ESI⁺, the mobile phase was composed of water (A) and ACN (B),
162 both acidified with 0.1% FA, and MeOH (C), flowing at 0.4 mL.min⁻¹. Gradient started at 100% A and reached
163 100% B in 10 min, being kept for 6 min before switching to 100% C to rinse the system in 1 min, being hold
164 for 5 min, returning back to 100% A in 1 min and finally equilibrating for 3 min, with a total run duration of
165 26 min. For ESI⁻, the mobile phase was composed of water buffered at pH 6.45 with 10 mM of ammonium
166 formate (A) and MeOH (B) flowing at 0.3 mL.min⁻¹. The gradient started at 100% A and reached 100% B in
167 13 min, holding this condition for 7 min before turning back to 100% A in 1 min and finally equilibrating for
168 3 min, with a total run duration of 24 min. For both chromatographic methods the temperature of the column
169 oven was kept at 30°C. Electrospray parameters have been fixed at the previously reported values (Cladière et
170 al., 2018).

171 The analytical sequence started with injection of 10 mobile phase blanks in order to reach complete equilibrium
172 of UHPLC-HRMS-ToF apparatus. Sample vials were randomized in the analytical sequence, and a blank as
173 well as a QC sample were injected every 10 sample vials. For each ionization mode, all sub-samples were
174 injected either in triplicate (development data set) or in quadruplicate (validation data set).

175 **2.4 DATA TREATMENT**

176 The data treatment workflow was set up to be as much automated as possible. Indeed, only few manual steps
177 are remaining, the main one being the final curation of the automated annotation algorithm. Moreover, no
178 information about the level or nature of contaminants are provided in the workflow, only the group information
179 (ex: “vial n°1 belongs to group n°3”). The term “group” refers here to all injections related to the same sample
180 (i.e. three replicates for sample preparation plus triplicates or quadruplicates of injections each time). In other
181 word, for each sample several raw data are obtained, these being grouped together before data treatment.

182 *Step 1: Building of data matrix from raw data*

183 Vendors (Waters®) raw data files were first converted to the open-source format mzXML using ProteoWizard
184 (Chambers et al., 2012) and then uploaded onto the Workflow4Metabolomics (W4M) platform (Giacomoni et

185 al., 2015). Data matrix building was then achieved using open-source XCMS package (Smith et al., 2006) on
186 this platform.

187 XCMS builds the data matrix from raw data files using the following workflow. First, “xcmsSet” with
188 CentWave method (Tautenhahn, Bottcher, & Neumann, 2008) extracts peaks from the data files. Peaks are
189 then grouped across the samples and aligned using “group”, “retcor” and then “group” functions again. The
190 final step of the algorithm, “fillpeaks”, identifies for each sample the peaks for which this sample has no value:
191 for these peaks, the tool integrates the signal noise in this area to avoid missing values at the end. XCMS
192 parameter values for each step of the workflow were chosen as suggested for UHPLC-Q-ToF instruments by
193 Patti *et al.* (Patti, Tautenhahn, & Siuzdak, 2013), except for the “peak width” parameter in the “xcmsSet” step
194 which was chosen less stringent (5-60 s instead of 5-20 s) to limit data loss. A complete list of XCMS
195 parameters can be found in **Table S.2** of Supplementary material. The XCMS peak extractions were performed
196 separately for ESI⁺ and ESI⁻ sequences. Finally, the data matrix is a table gathering the different peak areas
197 integrated by XCMS sorted by ions in row (combination of *m/z* and retention time) and by samples in column.
198 Data matrix files (.txt) were then imported in Matlab using in-house scripts.

199 *Step 2: Data filtration and reduction*

200 Since the number of output ions generated by XCMS is very high (between 10,000 and 30,000), the data
201 matrices needed to be filtrated to remove as many irrelevant ions as possible. Data cleaning by successive
202 filtration steps is critical since an adequate filtration should enable to clean the data from irrelevant signals
203 while avoiding or minimizing relevant chemical information loss. Data matrices were filtrated using only the
204 group information (i.e. blind to the nature and levels of the spiked molecules). Therefore, ions that do not differ
205 from the blanks or do not vary between samples were filtrated in order to keep only suspect ions and try to
206 highlight a food contamination. This filtration strategy is commonly used in metabolomics approaches
207 (Antignac et al., 2011) but applied here for the first time to non-targeted food safety analysis. It is generally
208 based on statistical tests (t-test) designed to determine significant differences for each ion between samples at
209 a commonly admitted p-value of 0.05 (Gika, Theodoridis, Plumb, & Wilson, 2014; Rubert et al., 2017;
210 Thévenot, Roux, Xu, Ezan, & Junot, 2015). This strategy can be completed by using the fold change of each
211 ion between samples: the common fold change value used is 2, but it is still under discussion for metabolomics
212 purpose (Ortmayr, Charwat, Kasper, Hann, & Koellensperger, 2017).

213 Finally, filtration of data matrices composed initially of around 20,000 ions was done within three successive
214 automated steps plus an automated pre-filtering step:

- 215 a) Pre-filtering step: unusable and unreliable variables that exhibit a poor stability, meaning relative standard
216 deviation (%RSD) on peak area above 100% in every sample group were discarded (about 50 to 100 ions
217 discarded).
- 218 b) First step of filtration: removal of ions that show no significant difference (peak area) between blank runs
219 and any of the sample groups using pairwise t-tests results (blanks vs. sample groups, about 19,000 ions
220 remaining).
- 221 c) Second step of filtration: removal of ions that show no significant difference (peak area) between any
222 sample groups using pairwise t-tests results (sample groups vs. sample groups, about 10,000 ions
223 remaining).
- 224 d) Third step of filtration: removal of ions that show a low fold change among sample groups, to select only
225 ions exhibiting high contrast between sample groups. For each ion, the median value of peak areas of each
226 sample groups (n=9, extraction triplicates, each analyzed in triplicate) was considered. The fold change
227 is then calculated by dividing the highest median value by the lowest one (assumed to be the most
228 concentrated sample divided by the least concentrated one, about 1,000 ions remaining after this step).

229 *Step 3: Normalization and scaling*

230 Missing values and algorithm artifacts such as zero, infinite and negative values in the data matrix were
231 managed according to the guidelines given by Wherens et al.(Wehrens et al., 2016). Briefly, for each ion
232 (named as “ m/z – retention time” combination), any irrelevant value was replaced by the lowest value of this
233 ion (blank excluded).

234 The data matrix was then log- and pareto- scaled (Antignac et al., 2011), and normalized using a median-based
235 Probabilistic Quotient Normalization (PQN) (Dieterle, Ross, Schlotterbeck, & Senn, 2006) using the QC
236 samples. Briefly, the median of each ion in QC samples is computed. Then, for each ion, values in samples are
237 divided by a reference value (here the median of the measurement for the ion in QC samples), leading to a
238 ratio matrix. Then, for each sample, the median of the ratios is computed, and the initial values are divided by
239 the ratio median.

240 *Step 4: Multivariate data analysis*

241 Principal Component Analysis (PCA) and Independent Component Analysis (ICA) were tested in order to
242 discriminate contaminated samples. ICA showed better performance than PCA to resolve complex signal
243 mixtures as already demonstrated for metabolomics data (Liu et al., 2016). Therefore, ICA was used to
244 interpret and visualize data with respect to their source signals, and then try to evidence potential discrimination
245 between sample groups. Indeed, ICA is a blind source separation method, which aims at extracting from mixed
246 signals their original source signals as well as the weights in which they are mixed. Among the different few
247 algorithms enabling to compute ICA models, the JADE algorithm was used here (Rutledge & Jouan-Rimbaud
248 Bouveresse, 2015).

249 The determination of optimal number of Independent Components (ICs) to use is the key step during the
250 building of an ICA model. This optimal number was determined using the random ICA method (Kassouf,
251 Jouan-Rimbaud Bouveresse, & Rutledge, 2018), briefly summarized as follows: the data set is randomly split
252 into two equivalent groups. ICA models with 1 to F components (here, $F = 20$) are calculated in each subset.
253 For each model (i.e., each investigated number of ICs), correlations between all ICs from one subset and all
254 ICs from the other subset are calculated. The idea underlying this procedure is that if an IC is significant, it
255 should be extracted in each subset and therefore, strong correlations should be observed between ICs from
256 each subset. Hence, one looks for the highest number of ICs for which each IC of one subset is highly correlated
257 with one IC of the other subset. However, the repartition of samples into the two subsets being random, there
258 is a possibility that the subsets are not representative, in which case a significant IC might be extracted from
259 one subset only. This is the reason why this procedure has to be repeated, here, 50 times.

260 *Step 5: Annotation and interpretation*

261 For automated annotation, an in-house broad-scale database was built combining data from several databases,
262 namely the Toxin and Toxin-Target Database (T3DB, <http://www.t3db.ca/> (Wishart et al., 2015)), the literature
263 (Gallart-Ayala, Núñez, & Lucci, 2013; Nielsen & Smedsgaard, 2003), and to a lesser extent, the Pesticides
264 Properties Database (Lewis, Tzilivakis, Warner, & Green, 2016).

265 After evidencing a discrimination along one component of the ICA model, the signal matrix, which gives the
266 weight of each ion (“ m/z – retention time” pair) in the component, was analyzed: ions were sorted by

267 descending contribution value along the components explaining the group separation, and the annotation was
268 performed according to the following strategy:

- 269 a) In-house automated tools were developed for isotopic pattern detection (inspired from work by
270 Cotton et al. [14]) and then for in-house toxicant database annotation to highlight suspect ions;
- 271 b) Manual curation of the annotation results was then performed for discriminating ions using
272 information provided by step a) as well as online databases [such as Metlin
273 (<https://metlin.scripps.edu/>), HMDB (<http://www.hmdb.ca/>), mzCloud
274 (<https://www.mzcloud.org/>) and T3DB] and raw data visualization when necessary.

275 With mass spectrometry, and especially electrospray ionization, a single molecule usually produces several
276 observed signals, either fragments, adducts or isotopic peaks. All ions assumed as coming from the same
277 compound (i.e. retention time, correlation, known $\Delta m/z$: M+1, M+2 with relevant intensity ratio) were grouped
278 in “features” during step a), each one representing a single compound. During the same step, adducts were
279 annotated with the database search. Annotation levels nomenclature used is based on guidelines proposed by
280 Sumner et. al. (Sumner et al., 2007).

281 **3. RESULTS AND DISCUSSION**

282 **3.1 MULTIVARIATE ANALYSIS OF DEVELOPMENT DATA SET**

283 All data treatments were developed and performed blindly, meaning without optimizing the parameters for our
284 tracers. The objective of this approach is to evaluate the efficiency of a generic blind untargeted analysis based
285 on multivariate tools to discriminate contaminated samples and annotate ions of potential contaminants.

286 A multivariate exploration of the data was first tried without any (pre)filtration, but it remained unsuccessful
287 since no clear group separation could be observed (see **Figure 2**). The detailed filtration process was thus
288 developed and applied. It appears (Supplementary material - **Figure S.1**) that, even though PCA enables a
289 discrimination for the filtrated data, the one given by ICA is superior both by its quality (better sample
290 separation) and its ability to align chemical phenomenon on a single component. So, the detection of suspect
291 samples and ions is eased thanks to ICA by simply sorting ions based on their weight on the discriminating
292 component.

3.1.1 BLIND DISCRIMINATION OF SAMPLES BY ICA

Optimal number of ICs was determined as 4 for the development data set in both positive and negative modes with the random ICA method. For each ionization mode, score plots were drawn considering the different ICs prone to discriminate sample groups (an illustrative plot is given in **Figure 2**). Interestingly, all sample groups could be discriminated whatever the ionization mode, and each time IC1 was determined as the most probable meaningful component regarding group information. It is clear that group separation along IC1 is related to the level of contamination. It should be emphasized that tea samples contaminated at the low level ($10 \mu\text{g}\cdot\text{kg}^{-1}$) could be distinguished from control tea samples, even with a high chemical diversity of contaminants. This is the first time that contaminated food samples are discriminated from control samples at this level: this opens new perspectives for food safety control, since $10 \mu\text{g}\cdot\text{kg}^{-1}$ is the maximum level authorized for several regulated chemicals, especially pesticides.

The next step is the annotation of discriminating ions, in order to assess if the discrimination observed is really due to the contaminants.

3.1.2 ANNOTATION AND INTERPRETATION OF ICA OUTPUT

Most of the annotation process was automated thanks to database search and data mining scripts. For each ionization mode, the filtered data matrix went through two automated steps: (i) isotopic peaks were first grouped together, and (ii) observed ions were searched through a broad-range toxicants database for testing their potential matching with different adducts. Then, for each ionization mode, ions were sorted along the discriminating component(s) and extracted with their information (m/z , retention time, presence of isotopes, and potential match in the database for different adducts). At the end, the results can be quickly curated by the user, who can then rapidly spot suspect samples and ions. These information can be completed, when needed, by a manual exploration of raw data files.

That way, for positive mode, the 69 first discriminating ions were putatively annotated or characterized and grouped into 20 “features” (see **Table 1a**). Over those 20 “features”, 14 were attributed to our “tracers”. Three others (#13, 18 and 19) were not expected to be present in the samples (since they were not found after a targeted analysis of the control samples, nor reported in the analysis certificate of standards used). Raw chemical formulas could be proposed, that show very strong similarities with some of our “tracers”, so they

320 were putatively annotated as impurities from the initial standard solution (such result has been confirmed by
321 an *a posteriori* classical targeted analysis of the standard solution). This clearly underlines the potential of our
322 developed method to detect unexpected compounds at trace levels since those impurities were not expected
323 before the analysis. Three other features remained unknown.

324 Putative annotation and characterization of the 69 first ions in negative mode was also achieved, and these ions
325 were grouped into 14 “features” (see **Table 1b**). For these features, raw molecular formulas hypotheses were
326 made based on information given by the automated annotation step, enabling a putative characterization to be
327 achieved. Thus, 12 of our “tracers” could be recovered. Two “features” (corresponding to eight ions) remained
328 unresolved after annotation attempt (using both automated scripts and manual exploration of raw data), but
329 they were characterized as being halogenated compounds thanks to the isotopic peaks found during the
330 automated data mining step.

331 **3.2 METHOD PERFORMANCE**

332 Performance was assessed based on blind detection rates of “tracer” contaminants. In positive mode, 44% of
333 our “tracers” were successfully putatively annotated, and 38% in negative mode. When considering both
334 modes, the overall detection rate is 66% (since some molecules were detected in both modes, e.g. diuron and
335 ochratoxin A). By comparing with detection performance of a dedicated targeted multi-residue method on the
336 same samples (Cladière et al., 2018), it appears that the “tracers” not successfully annotated using the
337 untargeted approach were also the most difficult to analyze with a targeted approach (i.e. showing high ion
338 suppression and therefore low signal/noise ratio, high relative standard deviation of the signals and poor
339 recoveries). A manual exploration of the raw chromatograms reveals that these molecules give very noisy
340 peaks, which are not even extracted during the pretreatment step with XCMS. In fact, to date, it is likely that
341 there is no algorithm that can achieve exhaustive peak extraction from raw data in untargeted LC-MS study
342 (Coble & Fraga, 2014). In our case, XCMS managed to extract 75% of our “tracers” (i.e. 24 over 32) from the
343 raw data files, which is still a good score even though the noisiest peaks are missed. We tested the only existing
344 optimization algorithm for XCMS parameters (IPO) (Libiseller et al., 2015), but with no improvement.
345 Unsurprisingly, the response factor of a compound has been found to be the main factor affecting its
346 detectability.

347 Thus, the data treatment methodology applied after XCMS treatment (including filtration of the data matrix,
348 preprocessing and multivariate analysis coupled with data mining) successfully annotated 21 tracers over the
349 24 extracted by XCMS (i.e. 88%). This is very satisfactory regarding the wide diversity of molecules studied,
350 both in terms of chemical structure and response factor in LC-MS, as well as regarding the trace levels studied.
351 Last but not least, the detection and putative annotation of unexpected impurities coming from the spiking mix
352 highlights the ability of our untargeted approach to detect potentially unknown or unexpected trace
353 contaminants in food products and to propose the user annotation hypotheses. It should be spotted that a
354 molecule generating a high number of ions (adducts, isotopes, fragments, etc.) will be more easily annotated
355 than a molecule generating only few signals.

356 For each annotated compound, a limit of detection (LOD) was estimated based on the calculated fold change
357 (calculated between the group with the highest level -100 $\mu\text{g.kg}^{-1}$ - and the control group). Briefly, a rule of
358 three was made to figure out what concentration would lead to a fold change of 3, which is the most commonly
359 used signal/noise ratio for LOD determination. For compounds annotated as impurities, assumption was made
360 that they come from standards of similar families (i.e. atrazine for simazine, diflufenzuron or diuron for
361 fenuron, and acid herbicides for 2,4-D isopropyl ester) and a LOD was then estimated for each by taking the
362 standard purity into account. As shown in **Table 1a and 1b**, estimated LODs are relevant against EU regulation
363 since they are in the range 10 $\mu\text{g.kg}^{-1}$ or below for almost every annotated compound. Again, our untargeted
364 approach proves to have quite similar performance in terms of sensitivity as compared to our dedicated targeted
365 multi-residue method [18], having the additional asset to detect unexpected molecules.

366 **3.3 APPLICATION ON VALIDATION DATA SET**

367 The developed methodology (including filtration parameters, pretreatment steps and multivariate method) was
368 blindly applied to the validation data set, obtained based on the analysis of two different types of green tea,
369 either non-spiked (controls) or spiked with a mix composed of only three contaminants in order to offer a much
370 more challenging discrimination between blanks and contaminated samples.

371 Data from the two types of tea were treated simultaneously. Data matrix was filtrated using same parameters
372 as for development set, and then the optimal number of ICs was determined for this data set. It was calculated
373 as 6 both for positive and negative modes. Unsurprisingly, since the data set is more heterogeneous, the
374 filtration led to a smaller reduction of ion number than for the development data set. Still, the number of ions

375 dropped from 23,391 and 17,269 (respectively for positive and negative mode) to 9,789 and 9,409 thanks to
376 the filtration. For each ionization mode, two discriminating ICs were clearly observed (**Figure 3**). IC1
377 separates samples based on their brand, and IC6 separates control and contaminated samples. It should be
378 emphasized that in ICA, ICs are not ordered by descending contribution like in PCA, meaning in other words
379 that IC1 does not necessarily explain more variability than IC6.

380 Annotation of the data matrix was done as described for the development data set. IC6 was determined as
381 bearing the separation due to contaminants thank to information provided in the automated annotation step.
382 The three “tracers” were successfully annotated within the first 10 ions of IC6 with this methodology in at
383 least one ionization mode. This results is of prime interest in our case since ICA, as employed here, shows its
384 main assets which is to separate independent phenomenon. Indeed, we can see that in our case, the “natural”
385 variability of the product is well separated from the variability brought by the spiking (these two phenomenon
386 are likely to be mixed in less powerful methods). Moreover, the number of ions generated in a complex data
387 set such as the validation one is very high despite of filtration strategy applied (in our case ~9,000 per ionization
388 mode). Therefore, the use of multivariate methods enables the reduction of the dimensionality of the data and
389 the achievement of suspect samples and ions detection. Thanks to ICA, the annotation of only 10 ions per
390 ionization mode was sufficient to underscore a contamination of tea.

391 This highlight the ability of ICA to resolve complex signal mixtures and simplify annotation of relevant ions,
392 even in cases where the information is bore by few, low intensity ions. Our proposed approach thus has a
393 strong potential in detecting food contaminants at low levels in complex and rather heterogeneous data sets.
394 Its applicability to other food matrices should be feasible if reference samples are available.

395 **4. CONCLUSION**

396 This work shows some important methodological features for untargeted approach development for food
397 chemical contaminants detection. It gives evidence that the blind untargeted detection of contaminants in
398 complex food matrices is feasible thanks to high resolution methods coupled to powerful data analysis
399 strategies. A widely spread, well-known, freely available and easy to use tool (i.e. XCMS run on W4M
400 platform) was used for peaks extraction from raw data. Then, an efficient automated strategy was set up for
401 data filtration, using well-known easy to use tools (t-tests and fold change). Samples were separated using a

402 multivariate method (ICA), and discriminating ions putatively annotated by the help of automated data mining
403 methods.

404 Thanks to this strategy, 66% of the “tracers” considered were successfully putatively annotated. This detection
405 rate rises to 88% if brought back to tracers actually in the data matrix (after the peak extraction step). This
406 shows the power of our developed data treatment strategy to detect potential food contaminants in a data
407 matrix. In addition to the known “tracers”, some unexpected molecules were detected and putatively annotated
408 in the samples, which clearly highlights the potential of this approach. Method LODs were roughly estimated
409 for each putatively annotated compound (both expected and unexpected), with values below or near $10 \mu\text{g}\cdot\text{kg}^{-1}$
410 ¹ for most of them, which compares favorably with a targeted multi-residue method. The approach, developed
411 on a rather simple case, has been validated on a more complex and realistic situation, where the contamination
412 is brought by a low number of molecules, and in which different brands of the same food product are considered
413 simultaneously. This study opens new perspectives in the development of truly untargeted approaches based
414 on tools and strategies from metabolomics (particularly HRMS and chemometrics) for food chemical safety
415 assessment. Such approaches may constitute, in a near future, a major complement to targeted methods in a
416 view of rapidly screening possibly contaminated food products. A next step towards a routine use of these
417 approaches would be to implement them on even more complex cases like the following of a production batch.
418 Interestingly, since our developed methodology is rather generic, it could be applied with only few
419 development on any other UHPLC-HRMS data set, or even to other applications such as origin or authenticity
420 issues to complement existing approaches. As a conclusion, no doubt that these results will encourage new
421 developments on this analytical issue, both on the methodology and the tools, especially on the improvement
422 of existing peak extraction methods or the handling of the product variability.

423 **FUNDING**

424 This work was supported by Paris Institute of Technology for Life, Food and Environmental Sciences
425 (AgroParisTech), the French National Institute for Agricultural Research (INRA) and the French Ministry of
426 Higher Education and Research.

427 **ACKNOWLEDGEMENTS**

428 The authors wish to thank Céline Dalle for her advices on W4M platform and Even Leroux for his technical
429 help with MS instrumentation.

430 REFERENCES

431 Antignac, J. P., Courant, F., Pinel, G., Bichon, E., Monteau, F., Elliott, C., & Le Bizec, B. (2011). Mass
432 spectrometry-based metabolomics applied to the chemical safety of food. *TrAC - Trends in Analytical*
433 *Chemistry*, 30(2), 292–301. <http://doi.org/10.1016/j.trac.2010.11.003>

434 Castro-Puyana, M., Pérez-Míguez, R., Montero, L., & Herrero, M. (2017). Application of mass spectrometry-
435 based metabolomics approaches for food safety, quality and traceability. *TrAC - Trends in Analytical*
436 *Chemistry*, 93, 102–118. <http://doi.org/10.1016/j.trac.2017.05.004>

437 Chambers, M. C., MacLean, B., Burke, R., Amodei, D., Ruderman, D. L., Neumann, S., ... Mallick, P. (2012).
438 A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology*, 30(10), 918–920.
439 <http://doi.org/10.1038/nbt.2377>

440 Chang, K. (2015). *World tea production and trade Current and future development. Food and Agriculture*
441 *Organisation*.

442 Cifuentes, A. (2009). Food analysis and foodomics. *Journal of Chromatography A*, 1216(43), 7109.
443 <http://doi.org/10.1016/j.chroma.2009.09.018>

444 Cladière, M., Delaporte, G., Le Roux, E., & Camel, V. (2018). Multi-class analysis for simultaneous
445 determination of pesticides, mycotoxins, process-induced toxicants and packaging contaminants in tea.
446 *Food Chemistry*, 242, 113–121. <http://doi.org/10.1016/j.foodchem.2017.08.108>

447 Coble, J. B., & Fraga, C. G. (2014). Comparative evaluation of preprocessing freeware on
448 chromatography/mass spectrometry data for signature discovery. *Journal of Chromatography A*, 1358,
449 155–164. <http://doi.org/10.1016/j.chroma.2014.06.100>

450 Cotton, J., Leroux, F., Broudin, S., Marie, M., Corman, B., Tabet, J. C., ... Junot, C. (2014). High-resolution
451 mass spectrometry associated with data mining tools for the detection of pollutants and chemical
452 characterization of honey samples. *Journal of Agricultural and Food Chemistry*, 62(46), 11335–11345.
453 <http://doi.org/10.1021/jf504400c>

- 454 Dieterle, F., Ross, A., Schlotterbeck, G., & Senn, H. (2006). Probabilistic quotient normalization as robust
455 method to account for dilution of complex biological mixtures. Application in ¹H NMR metabonomics.
456 *Analytical Chemistry*, 78(13), 4281–4290. <http://doi.org/10.1021/ac051632c>
- 457 Fraser, K., Lane, G. A., Otter, D. E., Hemar, Y., Quek, S. Y., Harrison, S. J., & Rasmussen, S. (2013). Analysis
458 of metabolic markers of tea origin by UHPLC and high resolution mass spectrometry. *Food Research*
459 *International*, 53(2), 827–835. <http://doi.org/10.1016/j.foodres.2012.10.015>
- 460 Gallart-Ayala, H., Núñez, O., & Lucci, P. (2013). Recent advances in LC-MS analysis of food-packaging
461 contaminants. *TrAC - Trends in Analytical Chemistry*, 42, 186–204.
462 <http://doi.org/10.1016/j.trac.2012.09.017>
- 463 Giacomoni, F., Le Corguillé, G., Monsoor, M., Landi, M., Pericard, P., Pétéra, M., ... Caron, C. (2015).
464 Workflow4Metabolomics: A collaborative research infrastructure for computational metabolomics.
465 *Bioinformatics*, 31(9), 1493–1495. <http://doi.org/10.1093/bioinformatics/btu813>
- 466 Gika, H. G., Theodoridis, G. A., Plumb, R. S., & Wilson, I. D. (2014). Current practice of liquid
467 chromatography-mass spectrometry in metabolomics and metabonomics. *Journal of Pharmaceutical and*
468 *Biomedical Analysis*, 87, 12–25. <http://doi.org/10.1016/j.jpba.2013.06.032>
- 469 Gómez-Ramos, M. M., García-Valcárcel, A. I., Tadeo, J. L., Fernández-Alba, A. R., & Hernando, M. D.
470 (2016). Screening of environmental contaminants in honey bee wax comb using gas chromatography–
471 high-resolution time-of-flight mass spectrometry. *Environmental Science and Pollution Research*, 23(5),
472 4609–4620. <http://doi.org/10.1007/s11356-015-5667-0>
- 473 Gosetti, F., Mazzucco, E., Gennaro, M. C., & Marengo, E. (2016). Contaminants in water: non-target
474 UHPLC/MS analysis. *Environmental Chemistry Letters*, 14(1), 51–65. [http://doi.org/10.1007/s10311-](http://doi.org/10.1007/s10311-015-0527-1)
475 [015-0527-1](http://doi.org/10.1007/s10311-015-0527-1)
- 476 Haug, K., Salek, R. M., Conesa, P., Hastings, J., De Matos, P., Rijnbeek, M., ... Steinbeck, C. (2013).
477 MetaboLights - An open-access general-purpose repository for metabolomics studies and associated
478 meta-data. *Nucleic Acids Research*, 41(D1), 781–786. <https://doi.org/10.1093/nar/gks1004>
- 479 Inoue, K., Tanada, C., Sakamoto, T., Tsutsui, H., Akiba, T., Min, J. Z., ... Toyo’Oka, T. (2015). Metabolomics
480 approach of infant formula for the evaluation of contamination and degradation using hydrophilic

- 481 interaction liquid chromatography coupled with mass spectrometry. *Food Chemistry*, *181*, 318–324.
482 <http://doi.org/10.1016/j.foodchem.2015.02.117>
- 483 Kassouf, A., Jouan-Rimbaud Bouveresse, D., & Rutledge, D. N. (2018). Determination of the optimal number
484 of components in independent components analysis. *Talanta*, *179*, 538–545.
485 <http://doi.org/10.1016/j.talanta.2017.11.051>
- 486 Knolhoff, A. M., & Croley, T. R. (2016). Non-targeted screening approaches for contaminants and adulterants
487 in food using liquid chromatography hyphenated to high resolution mass spectrometry. *Journal of*
488 *Chromatography A*, *1428*, 86–96. <http://doi.org/10.1016/j.chroma.2015.08.059>
- 489 Knolhoff, A. M., Zweigenbaum, J. A., & Croley, T. R. (2016). Nontargeted Screening of Food Matrices:
490 Development of a Chemometric Software Strategy to Identify Unknowns in Liquid Chromatography-
491 Mass Spectrometry Data. *Analytical Chemistry*, *88*(7), 3617–3623.
492 <http://doi.org/10.1021/acs.analchem.5b04208>
- 493 Kunzelmann, M., Winter, M., Åberg, M., Hellenäs, K.-E., & Rosén, J. (2018). Non-targeted analysis of
494 unexpected food contaminants using LC-HRMS. *Analytical and Bioanalytical Chemistry*, 1–10.
495 <http://doi.org/10.1007/s00216-018-1028-4>
- 496 Lehotay, S. J., Sapozhnikova, Y., & Mol, H. G. J. (2015). Current issues involving screening and identification
497 of chemical contaminants in foods by mass spectrometry. *TrAC - Trends in Analytical Chemistry*, *69*,
498 62–75. <http://doi.org/10.1016/j.trac.2015.02.012>
- 499 Lewis, K. A., Tzilivakis, J., Warner, D. J., & Green, A. (2016). An international database for pesticide risk
500 assessments and management. *Human and Ecological Risk Assessment*, *22*(4), 1050–1064.
501 <http://doi.org/10.1080/10807039.2015.1133242>
- 502 Libiseller, G., Dvorzak, M., Kleb, U., Gander, E., Eisenberg, T., Madeo, F., ... Magnes, C. (2015). IPO: a tool
503 for automated optimization of XCMS parameters. *BMC Bioinformatics*, *16*(1), 118.
504 <http://doi.org/10.1186/s12859-015-0562-8>
- 505 Liu, Y., Smirnov, K., Lucio, M., Gougeon, R. D., Alexandre, H., & Schmitt-Kopplin, P. (2016). MetICA:
506 Independent component analysis for high-resolution mass-spectrometry based non-targeted
507 metabolomics. *BMC Bioinformatics*, *17*(1), 114. <http://doi.org/10.1186/s12859-016-0970-4>

- 508 Nielsen, K. F., & Smedsgaard, J. (2003). Fungal metabolite screening: Database of 474 mycotoxins and fungal
509 metabolites for dereplication by standardised liquid chromatography-UV-mass spectrometry
510 methodology. *Journal of Chromatography A*, 1002(1–2), 111–136. [http://doi.org/10.1016/S0021-](http://doi.org/10.1016/S0021-9673(03)00490-4)
511 9673(03)00490-4
- 512 Ortea, I., Pascoal, A., Cañas, B., Gallardo, J. M., Barros-Velázquez, J., & Calo-Mata, P. (2012, August). Food
513 authentication of commercially-relevant shrimp and prawn species: From classical methods to
514 Foodomics. *Electrophoresis*. <http://doi.org/10.1002/elps.201100576>
- 515 Ortmayr, K., Charwat, V., Kasper, C., Hann, S., & Koellensperger, G. (2017). Uncertainty budgeting in fold
516 change determination and implications for non-targeted metabolomics studies in model systems. *The*
517 *Analyst*, 142(1), 80–90. <http://doi.org/10.1039/C6AN01342B>
- 518 Patti, G. J., Tautenhahn, R., & Siuzdak, G. (2013). Meta-Analysis of Untargeted Metabolomic Data:
519 Combining Results from Multiple Profiling Experiments. *Nature Protocols*, 7(3), 508–516.
520 <http://doi.org/10.1038/nprot.2011.454>.Meta-Analysis
- 521 Pongsuwan, W., Bamba, T., Harada, K., Yonetani, T., Kobayashi, A., & Fukusaki, E. (2008). High-throughput
522 technique for comprehensive analysis of Japanese green tea quality assessment using ultra-performance
523 liquid chromatography with time-of-flight mass spectrometry (UPLC/TOF MS). *Journal of Agricultural*
524 *and Food Chemistry*, 56(22), 10705–10708. <http://doi.org/10.1021/jf8018003>
- 525 Rubert, J., Righetti, L., Stranska-Zachariasova, M., Dzuman, Z., Chrpova, J., Dall'Asta, C., & Hajslova, J.
526 (2017). Untargeted metabolomics based on ultra-high-performance liquid chromatography–high-
527 resolution mass spectrometry merged with chemometrics: A new predictable tool for an early detection
528 of mycotoxins. *Food Chemistry*, 224, 423–431. <http://doi.org/10.1016/j.foodchem.2016.11.132>
- 529 Rutledge, D. N., & Jouan-Rimbaud Bouveresse, D. (2015). Corrigendum to “Independent Components
530 Analysis with the JADE algorithm.” *TrAC - Trends in Analytical Chemistry*, 67, 220.
531 <http://doi.org/10.1016/j.trac.2015.02.001>
- 532 SANTE/EU. (2015). Guidance document on analytical quality control and method validation procedures for
533 pesticides residues analysis in food and feed. Retrieved from
534 https://ec.europa.eu/food/sites/food/files/plant/docs/pesticides_mrl_guidelines_wrkdoc_11945.pdf

535 Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R., & Siuzdak, G. (2006). XCMS: Processing mass
536 spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification.
537 *Analytical Chemistry*, 78(3), 779–787. <http://doi.org/10.1021/ac051437y>

538 Sumner, L. W., Amberg, A., Barrett, D., Beale, M. H., Beger, R., Daykin, C. A., ... Viant, M. R. (2007).
539 Proposed minimum reporting standards for chemical analysis: Chemical Analysis Working Group
540 (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics*, 3(3), 211–221.
541 <http://doi.org/10.1007/s11306-007-0082-2>

542 Tautenhahn, R., Bottcher, C., & Neumann, S. (2008). Highly sensitive feature detection for high resolution
543 LC/MS. *BMC Bioinformatics*, 9, 16. <http://doi.org/10.1186/1471-2105-9-504>

544 Tautenhahn, R., Patti, G. J., Rinehart, D., & Siuzdak, G. (2012). XCMS online: A web-based platform to
545 process untargeted metabolomic data. *Analytical Chemistry*, 84(11), 5035–5039.
546 <http://doi.org/10.1021/ac300698c>

547 Tengstrand, E., Rosén, J., Hellenäs, K. E., & Åberg, K. M. (2013). A concept study on non-targeted screening
548 for chemical contaminants in food using liquid chromatography-mass spectrometry in combination with
549 a metabolomics approach. *Analytical and Bioanalytical Chemistry*, 405(4), 1237–1243.
550 <http://doi.org/10.1007/s00216-012-6506-5>

551 Thévenot, E. A., Roux, A., Xu, Y., Ezan, E., & Junot, C. (2015). Analysis of the Human Adult Urinary
552 Metabolome Variations with Age, Body Mass Index, and Gender by Implementing a Comprehensive
553 Workflow for Univariate and OPLS Statistical Analyses. *Journal of Proteome Research*, 14(8), 3322–
554 3335. <http://doi.org/10.1021/acs.jproteome.5b00354>

555 Wehrens, R., Hageman, J. A., van Eeuwijk, F., Kooke, R., Flood, P. J., Wijnker, E., ... de Vos, R. C. H. (2016).
556 Improved batch correction in untargeted MS-based metabolomics. *Metabolomics*, 12(5), 88.
557 <http://doi.org/10.1007/s11306-016-1015-8>

558 Wishart, D., Arndt, D., Pon, A., Sajed, T., Guo, A. C., Djoumbou, Y., ... Rappaport, S. M. (2015). T3DB: The
559 toxic exposome database. *Nucleic Acids Research*, 43(D1), D928–D934.
560 <http://doi.org/10.1093/nar/gku1004>

562 **Table 1a:** Discriminating features for positive ionization mode and respective putative annotation.

POSITIVE IONIZATION MODE									
Feature # ^a	Cumulative weight of the feature	Number of ions in the feature	Experimental exact mass of adduct ^b	Adduct	Proposed raw formula	Proposed putative annotation	Mono-isotopic mass ^c	Delta (ppm)	Estimated LOD ($\mu\text{g}\cdot\text{kg}^{-1}$)
1	23.76	8	331.0435	[M+H] ⁺	C ₁₀ H ₁₉ O ₆ PS ₂	Malathion	330.0361	0.48	1.1
2	20.44	12	607.2926	[M+H] ⁺	N/A ^d	Unknown	N/A	N/A	N/A
3	18.32	7	230.0076	[M+H] ⁺	C ₅ H ₁₂ NO ₃ PS ₂	Dimethoate	228.9996	3.04	0.9
4	13.83	5	233.0248	[M+H] ⁺	C ₉ H ₁₀ Cl ₂ N ₂ O	Diuron	232.0170	2.32	1.0
5	12.22	5	404.0894	[M+H] ⁺	C ₂₀ H ₁₈ ClNO ₆	Ochratoxin A	403.0823	-0.34	4.9
6	9.37	4	229.0416	[M+K] ⁺	C ₇ H ₁₄ N ₂ O ₂ S	Aldicarb	190.0776	4.41	15.1
7	9.23	3	311.0398	[M+H] ⁺	C ₁₄ H ₉ ClF ₂ N ₂ O ₂	Diflubenzuron	310.0321	1.45	1.9
8	9.09	3	220.9537	[M+H] ⁺	C ₄ H ₇ Cl ₂ O ₄ P	Dichlorvos	219.9459	2.47	3.6
9	8.56	3	384.1471	[M+H] ⁺	C ₂₁ H ₂₂ ClN ₃ O ₂	Tolfenpyrad	383.1401	-0.67	1.3
10	8.00	3	228.1283	[M+H] ⁺	C ₉ H ₁₇ N ₅ S	Ametryn	227.1205	2.59	1.5
11	7.86	3	216.1010	[M+H] ⁺	C ₈ H ₁₄ ClN ₅	Atrazine	215.0938	-0.14	2.9
12	6.01	2	306.1041	[M+H] ⁺	C ₁₁ H ₂₀ N ₃ O ₃ PS	Pirimiphos methyl	305.0963	1.56	0.4
13	4.69	2	263.0243	[M+H] ⁺	C ₁₁ H ₁₂ Cl ₂ O ₃	<i>2,4-D Isopropyl Ester</i> ^e	<i>262.0163</i>	<i>2.73</i>	<i>0.5</i>
14	4.69	1	256.0604	[M+H] ⁺	C ₉ H ₁₀ ClN ₅ O ₂	Imidacloprid	255.0523	3.19	16.3
15	4.36	2	621.2713	N/A	N/A	Unknown	N/A	N/A	N/A
16	4.29	2	623.2868	N/A	N/A	Unknown	N/A	N/A	N/A
17	2.73	1	251.0380	[M+H] ⁺	C ₁₂ H ₁₀ O ₄ S	Bisphenol S	250.0300	3.08	12.1
18	2.24	1	202.0855	[M+H] ⁺	C ₇ H ₁₂ ClN ₅	<i>Simazine</i> ^e	<i>201.0781</i>	<i>0.68</i>	<i>0.7</i>
19	2.03	1	182.1282	[M+NH ₄] ⁺	C ₉ H ₁₂ N ₂ O	<i>Fenuron</i> ^e	<i>164.0950</i>	<i>-3.57</i>	<i>0.9</i>
20	1.69	1	335.1254	[M+Na] ⁺	C ₁₉ H ₂₀ O ₄	Bisphenol F diglycidyl Ether	312.1362	0.12	57.0

563 ^a Features sorted by descending cumulative weight of ions in the discriminating IC564 ^b Mass measured for the ion having the highest weight in the discriminating IC565 ^c Electron mass used: $5.485\cdot 10^{-4}$ Da566 ^d Not Applicable567 ^e Found in the spiking mix, may be considered as impurities

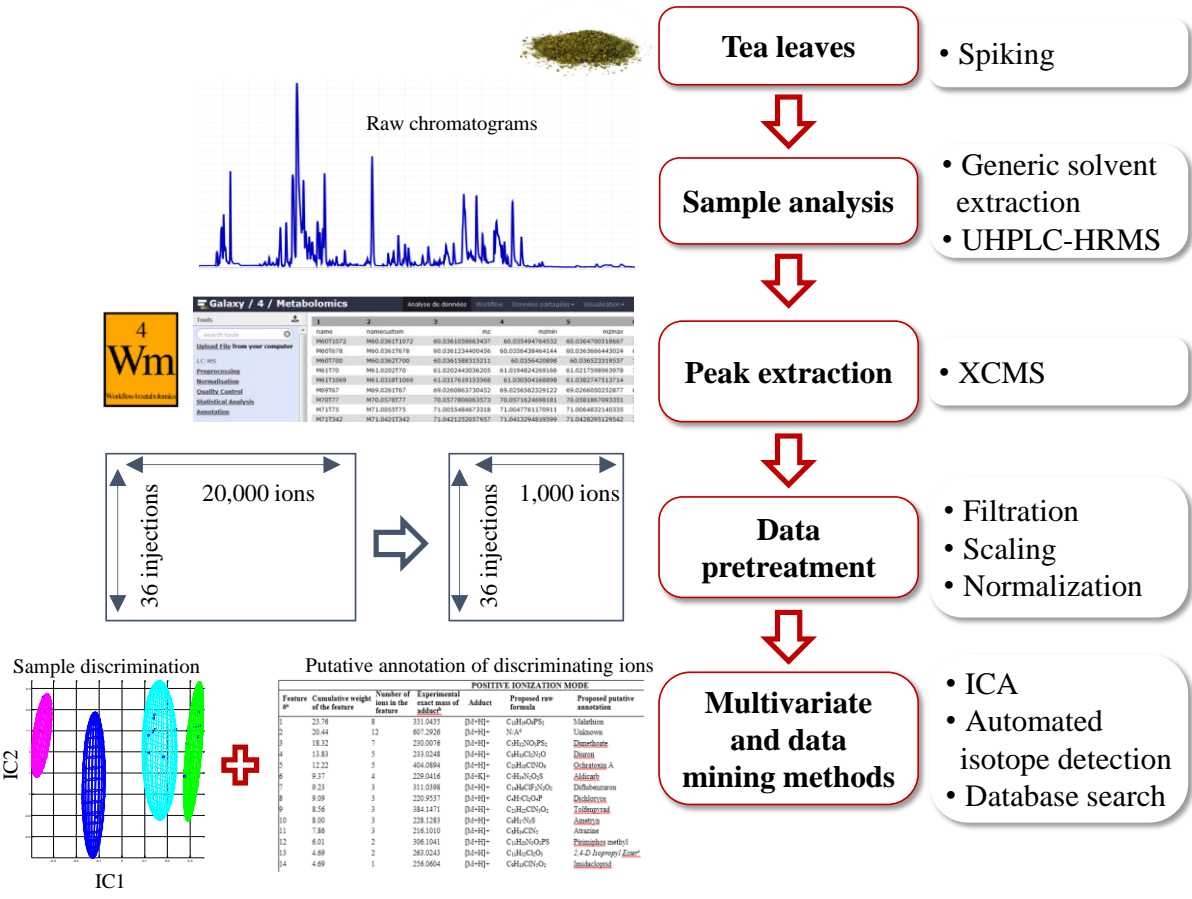
568
569

Table 1b: Discriminating features for negative ionization mode and respective putative annotation.

NEGATIVE IONIZATION MODE									
Feature # ^a	Cumulative weight of the feature	Number of ions in the feature	Experimental exact mass of adduct ^b	Adduct	Proposed raw formula	Proposed putative annotation	Mono-isotopic mass ^c	Delta (ppm)	Estimated LOD ($\mu\text{g}\cdot\text{kg}^{-1}$)
1	27.50	10	266.9385	[M-H]-	C ₉ H ₇ Cl ₃ O ₃	Fenoprop	267.9461	-1.02	2.0
2	22.07	8	252.9227	[M-H]-	C ₈ H ₅ Cl ₃ O ₃	2,4,5-T	253.9304	-1.90	3.0
3	21.62	8	309.0249	[M-H]-	C ₁₄ H ₉ ClF ₂ N ₂ O ₂	Diflubenzuron	310.0321	0.40	2.6
4	16.59	6	231.0091	[M-H]-	C ₉ H ₁₀ Cl ₂ N ₂ O	Diuron	232.0170	-2.97	0.9
5	14.32	8	232.9771	[M-H]-	C ₉ H ₈ Cl ₂ O ₃	Dichlorprop	233.9850	-2.75	5.9
6	13.48	5	213.0313	[M-H]-	C ₁₀ H ₁₁ ClO ₃	MCPD	214.0397	-4.98	2.5
7	11.68	5	199.0152	[M-H]-	C ₉ H ₉ ClO ₃	MCPA	200.024	-7.72	5.2
8	8.05	6	204.9217	N/A	N/A	Unknown chlorinated	N/A	N/A	N/A
9	7.75	3	249.0223	[M-H]-	C ₁₂ H ₁₀ O ₄ S	Bisphenol S	250.0300	-1.53	6.0
10	5.37	2	239.0668	[M-H]-	C ₁₀ H ₁₂ N ₂ O ₅	Dinoseb	240.0746	-2.08	10.4
11	4.61	2	402.0749	[M-H]-	C ₂₀ H ₁₈ ClNO ₆	Ochratoxin A	403.0823	-0.13	10.3
12	3.88	2	254.0444	[M-H]-	C ₉ H ₁₀ ClN ₅ O ₂	Imidacloprid	255.0523	-2.29	25.9
13	2.62	2	218.9611	[M-H]-	C ₈ H ₆ Cl ₂ O ₃	2,4-D	219.9694	-4.74	15.5
14	2.36	2	771.1431	N/A	N/A	Unknown chlorinated	N/A	N/A	N/A

570

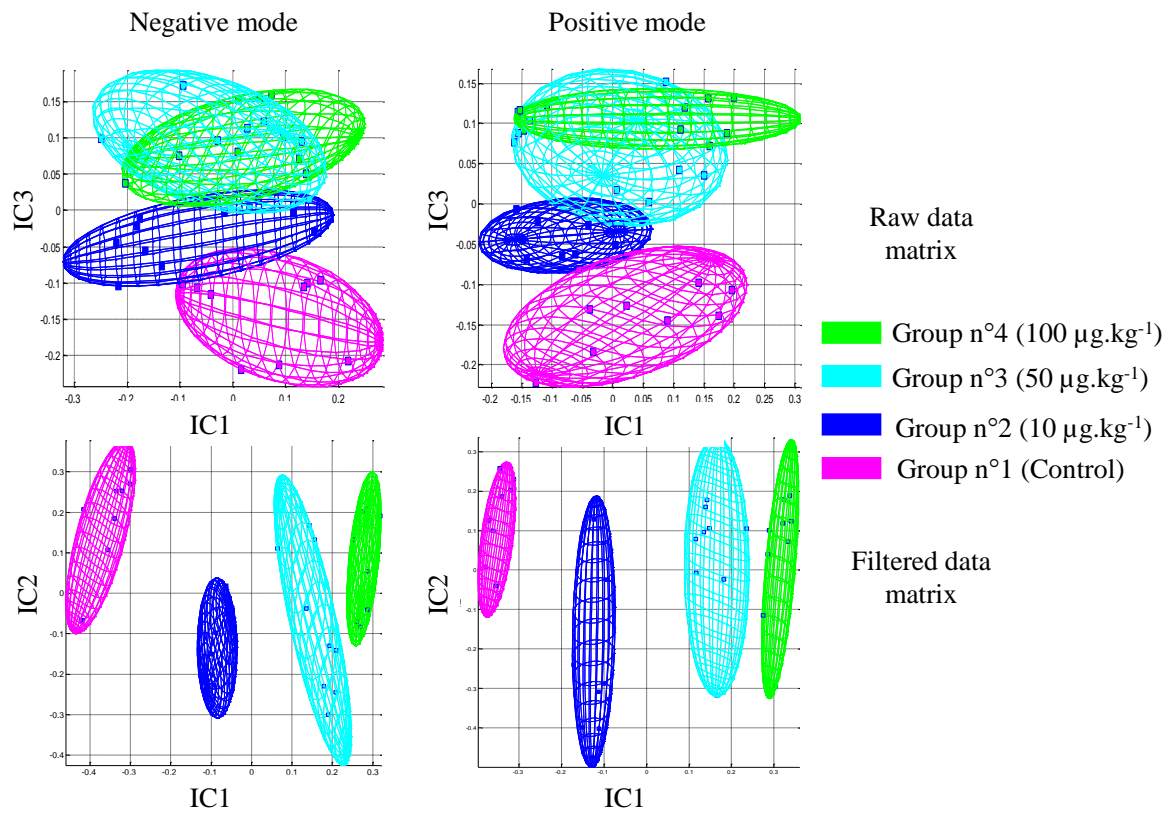
571



572

573 **Figure 1:** Workflow developed for untargeted contaminants detection at trace levels in tea.

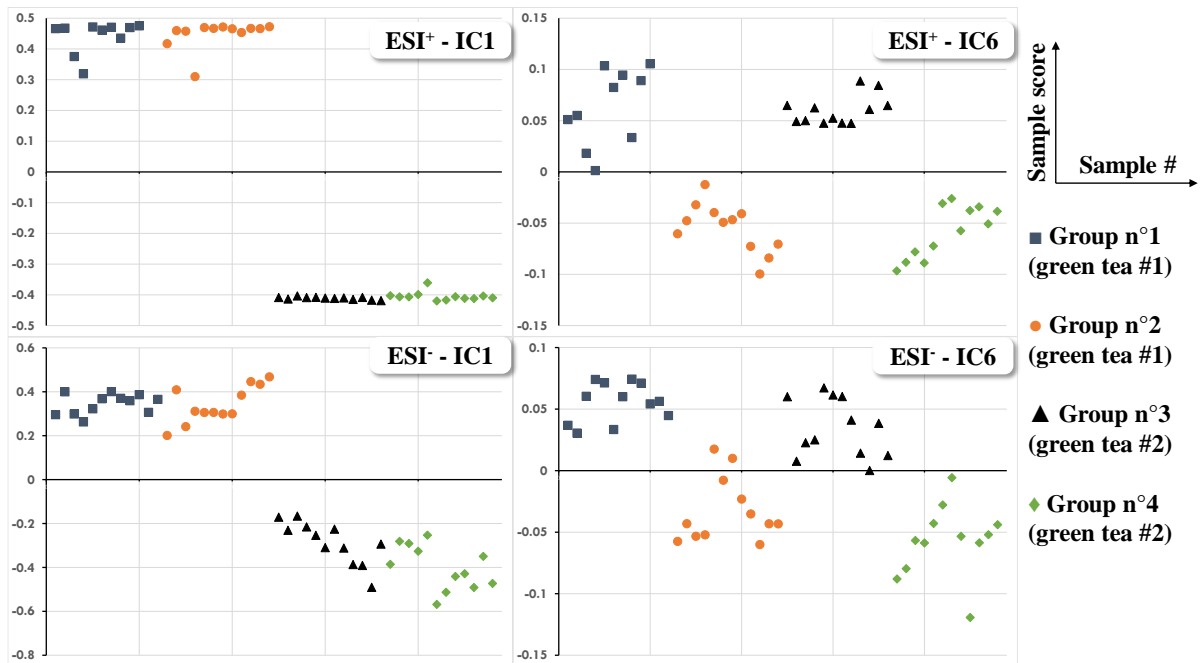
574



575

576 **Figure 2:** ICA score plots for filtered and unfiltered data matrix for both ionization modes on the
 577 development data set.

578



579

580 **Figure 3:** Score plots of discriminating ICs (both ionization modes) for the validation data set.

581

582

584 **Table S.1:** Information on “tracer” contaminants used in this study.

N°	Name	Class	Chemical family	log K _{ow}	Ionization mode	Adduct - ESI+/ESI-
1	(2,4-dichlorophenoxy)acetic acid 2,4-D	Pesticide (herbicide)	Auxinic herbicide	-0.82	+/-	[M+H] ⁺ /[M-H] ⁻
2	(4-chloro-2-methylphenoxy)acetic acid MCPA	Pesticide (herbicide)	Auxinic herbicide	-0.81	-	ND/[M-H] ⁻
3	2-(4-chloro-2-methylphenoxy)propanoic acid MCPP	Pesticide (herbicide)	Auxinic herbicide	-0.19	-	ND/[M-H] ⁻
4	2,4,5-trichlorophenoxyacetic acid 2,4,5-T	Pesticide (herbicide)	Auxinic herbicide	2.88	-	ND*/[M-H] ⁻
5	4-(2,4-dichlorophenoxy)butanoic acid 2,4-DB	Pesticide (herbicide)	Auxinic herbicide	1.35	+/-	[M+H] ⁺ /Frag**
6	Acetamiprid	Pesticide (insecticide)	Neonicotinoid	0.8	+	[M+H] ⁺ /ND
7	Acrylamide	Process-induced	Amide	-0.67	+	[M+H] ⁺ /ND
8	Aldicarb	Pesticide (acaricide)	Carbamate	1.15	+	[M+Na] ⁺ /ND
9	Ametryn	Pesticide (herbicide)	Triazine	2.63	+	[M+H] ⁺ /ND
10	Atrazine	Pesticide (herbicide)	Triazine	2.7	+	[M+H] ⁺ /ND
11	Bisphenol A	Migrant from packaging	Bisphenol	3.3	-	ND/[M-H] ⁻
12	Bisphenol A diglycidyl ether BADGE	Migrant from packaging	Diglycidyl ether	3.84	+	[M+Na] ⁺ /ND
13	Bisphenol F	Migrant from packaging	Bisphenol	1.65	-	ND/[M-H] ⁻
14	Bisphenol F diglycidyl ether BFDGE	Migrant from packaging	Diglycidyl ether	Not available	+	[M+Na] ⁺ /ND
15	Bisphenol S	Migrant from packaging	Bisphenol	2.91	+/-	[M+H] ⁺ /[M-H] ⁻
16	Deoxynivalenol	Mycotoxin	Trichothecene	0.29	+/-	[M+H] ⁺ /[M-H] ⁻
17	Dichloprop	Pesticide (herbicide)	Auxinic herbicide	2.29	-	ND/[M-H] ⁻
18	Dichlorvos	Pesticide (acaricide)	Organochlorinated	1.9	+	[M+H] ⁺ /ND
19	Diflubenzuron	Pesticide (insecticide)	Benzoylurea	3.89	+/-	[M+H] ⁺ /[M-H] ⁻
20	Dimethoate	Pesticide (acaricide)	Organophosphate	0.7	+	[M+H] ⁺ /ND
21	Dinoseb	Pesticide (herbicide)	Dinitrophenol	2.29	-	ND/[M-H] ⁻

22	Diuron	Pesticide (herbicide)	Phenylurea	2.87	+/-	[M+H] ⁺ /[M-H] ⁻
23	Fenoprop 2,4,5-TP	Pesticide (herbicide)	Auxinic herbicide	2.84	-	ND/Frag**
24	Fumonisin B1	Mycotoxin	Fumonisin	-0.5	+	[M+H] ⁺ /ND
25	Fumonisin B2	Mycotoxin	Fumonisin	1.2	+	[M+H] ⁺ /ND
26	Hydroxymethylfurfural	Process- induced	Furan	-0.09	+	[M+H] ⁺
27	Imidacloprid	Pesticide (insecticide)	Neonicotinoid	0.57	+/-	[M+H] ⁺ /[M-H] ⁻
28	Malathion	Pesticide (insecticide)	Organophosphate	2.75	+	[M+H] ⁺ /ND
29	Ochratoxin A	Mycotoxin	Ochratoxin	4.74	+/-	[M+H] ⁺ /[M-H] ⁻
30	Pirimiphos-methyl	Pesticide (insecticide)	Organophosphate	3.9	+	[M+H] ⁺ /ND
31	Propargite	Pesticide (acaricide)	Organosulfite	5.7	+	Frag**/ND
32	Tolfenpyrad	Pesticide (insecticide)	Pyrazole	5.61	+/-	[M+H] ⁺ /[M-H] ⁻

585 *ND: Not Detected

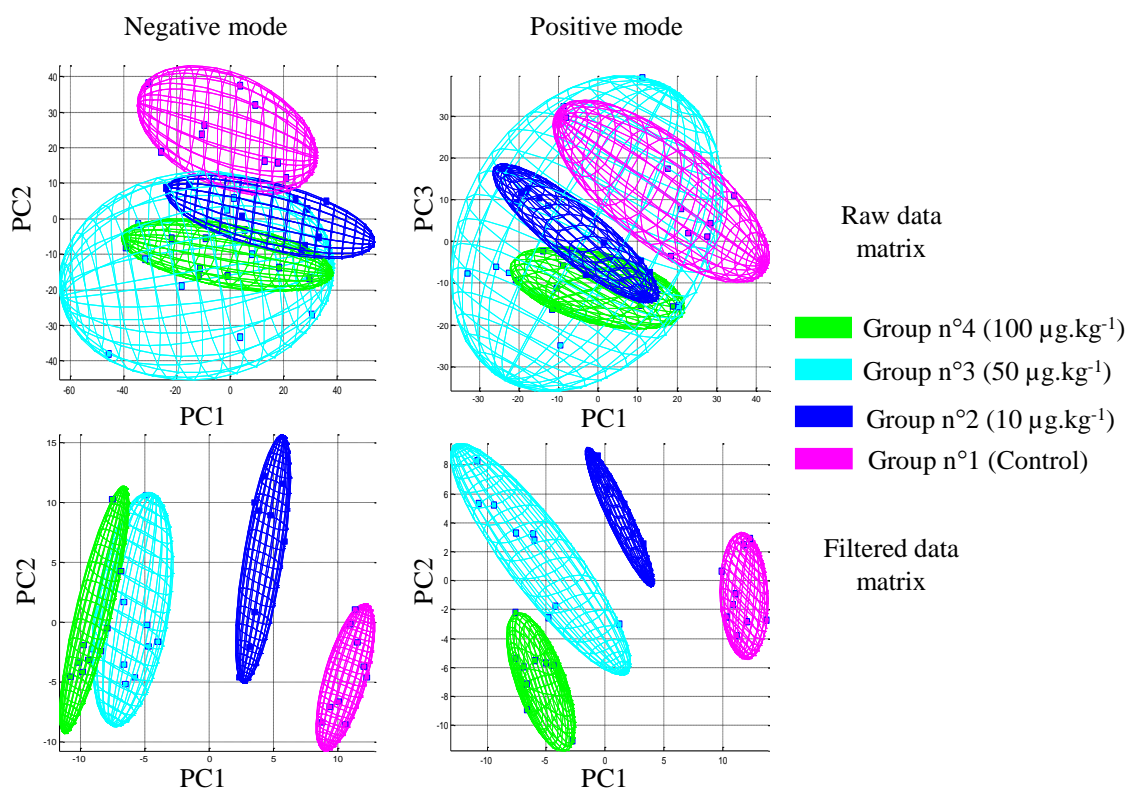
586 ** Frag=Fragment

587

588 **Table S.2:** Full parameters and their corresponding values for peak extraction using XCMS.

Step	Parameter	Value
xcmsSet	scanrange	180-2400
	nSlaves	1
	method	centWave
	ppm	15
	peakwidth	5-60
	mzdiff	-0.001
	snthresh	10
	integrate	1
	noise	0
	prefilter	0
group - A	method	density
	minfrac	0.5
	bw	2
	mzwid	0.015
	sleep	0.001
retcor	method	peakgroups
	smooth	loess
	extra	1
	missing	1
	span	0.2
	family	gaussian
group - B	method	density
	minfrac	0.5
	bw	2
	mzwid	0.015
	sleep	0.001
	max	50
	fillPeaks	method
convertRTMinute		FALSE
numDigitsMZ		4
numDigitsRT		2
intval		into
annotatediff	nSlaves	4
	sigma	6
	perfwHM	0.6
	ppm	15
	mzabs	0.015
	maxcharge	1
	maxiso	4
	minfrac	0.5
	quick	TRUE
	convertRTMinute	FALSE
	numDigitsMZ	4
	numDigitsRT	0
	intval	into

590



591

592 **Figure S.1:** PCA score plots for filtered and unfiltered data matrix for both ionization modes on the
593 development data.

594