

MS-CleanR: A feature-filtering approach to improve annotation rate in untargeted LC-MS based metabolomics

Ophélie Fraisier-Vannier^{1,4}, Justine Chervin^{2,3}, Guillaume Cabanac⁴, Virginie Puech-Pages^{2,3}, Sylvie Fournier^{2,3}, Virginie Durand², Aurélien Amiel^{2,5}, Olivier André^{2,5}, Omar Abdelaziz Benamar^{2,5}, Bernard Dumas², Hiroshi Tsugawa^{6,7} and Guillaume Marti^{1,2,3,4*}

¹ Pharma Dev, Université de Toulouse, IRD, UPS, Toulouse, France

² Laboratoire de Recherche en Sciences Végétales, Université de Toulouse, CNRS, UPS, France

³ Metatoul-AgromiX platform, MetaboHUB, National Infrastructure for Metabolomics and Fluxomics, LRSV, Université de Toulouse, CNRS, UPS, France

⁴ Institut de Recherche en Informatique de Toulouse, Université de Toulouse, UPS, Toulouse, France

⁵ De Sangosse, Bonnel, 47480 Pont-Du-Casse, France

⁶ RIKEN Center for Sustainable Resource Science, Yokohama, Japan

⁷ RIKEN Center for Integrative Medical Science, Yokohama, Japan

Abstract

Untargeted metabolomics using liquid chromatography-mass spectrometry (LC-MS) is currently the gold-standard technique to determine the full chemical diversity in biological samples. This approach still has many limitations, however; notably, the difficulty of estimating accurately the number of unique metabolites being profiled among the thousands of MS ion signals arising from chromatograms. Here, we describe a new workflow, MS-CleanR, based on the MS-DIAL/MS-FINDER suite, which tackles feature degeneracy and improves annotation rates. We show that implementation of MS-CleanR reduces the number of signals by nearly 80% while retaining 95% of unique metabolite features. Moreover, the annotation results from MS-FINDER can be ranked with respect to database chosen by the user, which improves identification accuracy. Application of MS-CleanR to the analysis of *Arabidopsis thaliana* grown in three different conditions improved class separation resulting from multivariate data analysis and lead to annotation of 75% of the final features. The full workflow was applied to metabolomic profiles from three strains of the leguminous plant *Medicago truncatula* that have different susceptibilities to the oomycete pathogen *Aphanomyces*

29 *euteiches*; a group of glycosylated triterpenoids overrepresented in resistant lines were identified as
30 candidate compounds conferring pathogen resistance. MS-CleanR is implemented through a Shiny
31 interface for intuitive use by end-users (available at: <https://github.com/eMetaboHUB/MS-CleanR>).

32 **Keywords:** Untargeted metabolomics, LC-MS, annotation, *Arabidopsis thaliana*, *Medicago*
33 *truncatula*, MS-DIAL, MS-FINDER.

34 Untargeted, or discovery-based metabolomics has become an essential tool in all biological sciences
35 including clinical research^{1,2}, plant science³ and natural product mining⁴, among many other
36 applications. Living organisms are estimated to contain more than one million distinct compounds⁵.
37 According to the MetaboLights database (DB), 80% of untargeted metabolomics workflows rely on
38 liquid chromatography-mass spectrometry (LC-MS) (<https://www.ebi.ac.uk/metabolights/>). Due to its
39 broad coverage of metabolites, LC-MS based metabolomics has become the preferred tool to detect
40 several hundreds of compounds encountered in a complex biological material. Many software
41 programs have been developed to turn features ($m/z \times$ retention time (RT) pairs) extracted from LC-
42 MS raw data into chromatographic peak lists, including web-based interfaces such as XCMS⁶,
43 Workflow4Metabolomics⁷, local GUI with MZmine⁸ and MS-DIAL⁹. Despite significant progress in
44 feature extraction, it remains a challenge to estimate accurately the number of unique metabolites in a
45 crude extract from the profile of one LC-MS experiment¹⁰. On average, untargeted LC-MS profiling
46 yields hundred to thousands of features, which include isotopes, contaminants, adducts, dimers,
47 multimers and heteromeric complexes, and artifacts. Patti and colleagues¹¹ used the term ‘degenerate
48 features’ to describe multiple signals derived from the same metabolite; they demonstrated that feature
49 inflation is highly underestimated and insufficiently addressed in untargeted LC-MS based
50 metabolomics. This may have important consequences by increasing both the false annotation rate and
51 the number of ‘unknown’ features arising from wrongly attributed signals. This is especially true when
52 the annotation process is based on *in silico* modeling of fragmentation patterns, as are Sirius¹², MS-
53 FINDER¹³, MetFrag¹⁴ or CFM-ID¹⁵, since tandem mass spectrometry (MS/MS) spectra are processed
54 without taking into account feature relationships. Thus, most untargeted metabolomics studies focus
55 on a subset of identified metabolites for which spectral data are easily accessible from public
56 repositories or in-house DBs.

57 A few packages have been developed to deal with feature degeneracy: CAMERA¹⁶ is based on adduct
58 relationships; RAMClust¹⁷ correlates features in multiple samples; MS-FLO¹⁸ uses Pearson’s
59 correlation and peak height similarity to identify adducts, duplicate peaks and isotope features of the
60 main monoisotopic ion, and MZunity¹⁹ which confronts adducts or neutral loss index to decipher
61 relationship among the acquired high resolution pseudo-molecular ions list. Deep-learning approaches
62 have also been developed based on LC-MS spectral peak shape filtering^{20,21}. All these packages focus
63 on a single type of degeneracy, however, and they are difficult to implement in a unified workflow.

64 Among the most advanced and versatile methods developed recently for untargeted metabolomics is
65 the tandem MS-DIAL-MS-FINDER suite. MS-DIAL is an all-in-one program for metabolomics and
66 lipidomics that relies on mass spectral libraries such as NIST 14 and MassBank of North America
67 (MoNA) for metabolite annotation. MS-FINDER is a partner program of MS-DIAL, in which
68 unknown structures can be elucidated from MS/MS spectra by the hydrogen rearrangement rules-
69 based scoring system. Here, we describe a third tool in this suite, called MS-CleanR, to remove
70 degenerate features and improve annotation rates from untargeted LC-MS-based metabolomics data.
71 Starting from the aligned peak list files determined by the MS-DIAL deconvolution process, our R
72 package firstly removes noise signals by using generic filters. In the second step, the package groups
73 the ion features based on the results of the MS-DIAL peak character estimation algorithm²² providing
74 the ion linkages of adducts, correlated chromatograms, putative ion source fragments candidates and
75 similar metabolite profiles among samples. In the third step, clustered ion features are merged between
76 positive ionization (PI) and negative ionization (NI) modes and the adduct relationships are corrected
77 accordingly. The cleaned-up feature list can be exported to MS-FINDER for annotation purposes.
78 Finally, the package merges the MS-FINDER annotation output with the cleaned-up peak list and
79 includes the possibility to prioritize identification according to the DBs used for MS-FINDER
80 interrogation. The whole MS-CleanR workflow is easily accessible through a Shiny user interface
81 (Figure 1) and it is available as open source code.

82

83 METHODS

84 Standards

85 Individual solutions of natural products (NPs) compounds (Metasci, Toronto, Canada) were prepared
86 at 100 µg/mL in H₂O or MeOH according to the supplier's recommendations. Mixes of 10 compounds
87 were prepared by pooling 10 µL of each individual solution to a final concentration of 10 µg/mL. We
88 selected 51 NPs eluting from 2-18 minutes as a first test mixture to construct DB-level 1 annotation.

89 IROA Mass Spectrometry Library of Standards (Sigma-Aldrich, Darmstadt, Germany) in 96-well
90 plate format (5 µg per well) were used. The contents of each well were dissolved in 50 µL of solvent
91 (5% MeOH or MeOH/CH₃Cl/H₂O 1:1:0.3), as recommended by the manufacturer, to obtain a
92 concentration of 100 µg/mL. Each plate was then sonicated for 5 minutes. Mixes of up to 12
93 compounds with distinct exact masses were obtained by pooling 20 µL from each well. The final
94 concentration in each mix was 8 µg/mL. We selected 167 standards eluting from 2-18 minutes as a
95 second test mixture.

96

Plant material

97 *A. thaliana* (wild-type Col-O) were grown either in hydroponic culture, in plastic pots (high density),
98 or in Jiffy® pots. For hydroponic culture, seeds were sown in 96 plates in MS liquid medium + 1%
99 sucrose. After 11 days, the medium was replaced by MS medium. After 14 days, seedlings were
100 collected and gently dried on absorbent paper. For culture in plastic pots, seeds were sown densely on
101 soil in plastic pots and cultivated in a growth chamber with a cycle of 16h light-8h dark, at 22°C in the
102 light and 20°C in the dark, and at 80% relative humidity. After 21 days, the aerial parts of the plants
103 were collected. For culture in Jiffy® pots, three seeds were sown per pot and cultivated in a growth
104 chamber, as for the plants in plastic pots. After 32 days, rosette leaves were harvested. For each
105 growing condition 200 mg of plant material per sample were collected, placed in a FastPrep tube (MP
106 Biomedicals Lysing Matrix D, Illkirch, France) and frozen in liquid nitrogen. For extraction, each
107 sample was ground with a Mixer Mill MM 400 grinder (Retsch, Eragny sur Oise, France) by applying
108 two cycles of 30 seconds at 30 m/sec. Biphasic sample extraction was adapted from Salem *et al.*
109 2016²³ by adding two cycles of 20 seconds at 6 m/sec. in the FastPrep-24™ benchtop homogenizer
110 (MP Biomedicals™, Illkirch, France) in 1 mL M1 (methyl tert-butyl ether/methanol, 3:1, v:v)
111 extraction solution. After grinding, FastPrep tube was transferred in glass tube and 5.7 mL of M1 was
112 added with 4.3 mL of M2 (water:methanol, 3:1, v:v) and vortexed for 1 min. The phases were
113 separated by centrifugation at 4°C and 4000 rpm for 5 minutes. The aqueous phases (400 µL) were
114 evaporated under nitrogen and the extracts were resuspended in 750 µL MeOH:H₂O (1:1). Samples
115 were filtered through 0.2 µm PTFE filters (Thermo Scientific™) and transferred to vials. An
116 extraction blank (without plant material) and a QC (Quality Control) sample (aliquot of all samples)
117 were also prepared to validate the LC-MS profiles.

118 Seeds of *Medicago truncatula* strains A17, DZA45.5 and F83005.5 (called F83 hereafter) were
119 scarified with sand paper, sterilized in 3.2% bleach for 2 min then rinsed in water four times before
120 soaking in water for 20 min. Seeds were placed on water agar and placed at 4°C for 4 days then for
121 one night at 25°C to germinate. Germinated seedlings were transferred onto M medium²⁴ then placed
122 in a growth chamber at 22°C and 50% humidity with cycles of 16h light-8h dark for 14 days. The
123 roots were ground with a Mixer Mill MM 400 grinder by applying two cycles of 30 seconds at 300 Hz.
124 One hundred milligrams of ground tissue were placed in 2 mL FastPrep tubes containing 1.4 mm
125 ceramic spheres (Lysing Matrix D) and extracted with 1 mL of acidified aqueous solution of methanol
126 (MeOH/H₂O/HCOOH, 80:19:1). After two cycles of 20 seconds at 6 m/sec. in the FastPrep-24™ (MP
127 Biomedicals™), the samples were centrifuged at 4°C and 10 000 rpm for 10 minutes. The supernatants
128 were transferred into vials. An extraction blank and QC (Quality Control) were also done for
129 extraction and analytical validation.

130 UHPLC-HRMS profiling

131 Ultra High Performance Liquid Chromatography-High Resolution MS (UHPLC-HRMS) analyses
132 were performed on a Q Exactive Plus quadrupole mass spectrometer, equipped with a heated
133 electrospray probe (HESI II) coupled to an U-HPLC Ultimate 3000 RSLC system (Thermo Fisher
134 Scientific, Hemel Hempstead, UK). Samples were separated on a Luna Omega Polar C18 column
135 (150×2.1mm i.d., 1.6µm, Phenomenex, Sartrouville, France) equipped with a guard column. The
136 mobile phase A (MPA) was water with 0.05% formic acid (FA) and mobile phase B (MPB) was
137 acetonitrile with 0.05% FA. The solvent gradient was: 0 min, 100% MPA; 1 min 100% MPA; 22 min,
138 100% MPB; 25 min, 100% MPB, 25.5 min, 100% MPA; 28 min, 100% MPA. The flow rate was
139 0.3 mL/min, the column temperature was set to 40°C, the autosampler temperature was set to 10°C
140 and injection volume fixed to 2 µL for standard mixes and plant extracts. Mass detection was
141 performed in positive ionization (PI) and negative ionization (NI) modes at 30 000 resolving power
142 [full width at half maximum (FWHM) at 400 m/z] for MS1 and 17 500 for MS2 with an automatic
143 gain control (AGC) target of 1e5. Ionization spray voltages were set to 3.5 kV (for PI) and 2.5 kV (for
144 NI) and the capillary temperature was set to 256°C for both modes. The mass scanning range was m/z
145 70-1050 Da for standards and m/z 100-1500 Da for plant extracts. Each full MS scan was followed by
146 data-dependent acquisition of MS/MS data for the six most intense ions.

147 Data processing

148 LC-MS data were first processed with MS-DIAL version 4.12. MS1 and MS2 tolerances were set to
149 0.01 and 0.05 Da, respectively, in centroid mode for each dataset. Peaks were aligned on a quality
150 control (QC) reference file with a RT tolerance of 0.1 min and a mass tolerance of 0.015 Da.
151 Minimum peak height was set to 70% below the observed total ion chromatogram (TIC) baseline for a
152 blank injection. MS-DIAL data was cleaned with MS-CleanR by selecting all filters with a minimum
153 blank ratio set to 0.8, a maximum relative standard deviation (RSD) set to 30 and a relative mass
154 defect (RMD) ranging from 50-3 000. The maximum mass difference for feature relationships
155 detection was set to 0.005 Da and maximum RT difference was set to 0.025 min. The Pearson
156 correlation links were considered only for biological datasets with correlation ≥ 0.8 and statistically
157 significant $\alpha = 0.05$. Two peaks were kept in each cluster: the most intense and the most connected.
158 The kept features were annotated with MS-FINDER version 3.26. The MS1 and MS2 tolerances were
159 set to 5 and 15 ppm, respectively. Formula finder were exclusively processed with C, H, O, N, P and S
160 atoms. DBs based on the genus and the family of the plant species (Table S3, Table S4, Table S7,
161 Table S8) being investigated were constituted with the dictionary of natural product (DNP-CRC press,
162 DNP on DVD v. 28.2) and the internal generic databases used were KNApSAcK, PlantCyc, HMDB,
163 LipidMaps, NANPDB and UNPD. Annotation prioritization was done by ranking genus DB followed
164 by Family DB and then generic DB (internal DB from MS-FINDER).

165 Statistical analysis

166 Statistical analyses were done by using SIMCA-P+ (version 15.0.2, Umetrics, Umea, Sweden). All
167 data were scaled by unit variance (UV) scaling before multivariate analysis. The orthogonal projection
168 to latent structure using discriminant analysis (OPLS-DA) was used to separate data according to *A.*
169 *thaliana* growing conditions. The OPLS regression model used for the *Medicago* datasets was tuned
170 with line resistance as the Y input: the following resistance indices 0, 1 and 2 were respectively
171 indicated for the F83, A17 and DZA45.5 strains. Coefficient scores were used to rank variables
172 according to their class biomarker: a high coefficient indicating a strong correlation with resistance
173 traits.

174 Mass spectral similarity network

175 The .msp NI and metadata files generated at the end of the MS-CleanR workflow were imported into
176 MetGem²⁵ (version 1.2.2). A mass spectral similarity network was created with a cosine score cut off
177 fixed at 0.65, a maximum of ten connections between nodes and at least four matched peaks. The
178 resulting network was then imported into Cytoscape²⁶ (version 3.7.2) to tune visualization. Nodes
179 were thus colored according to their annotated chemical classes and their sizes were indicated relative
180 to the OPLS coefficient score. Edge width was deepened according to their cosine value.

181 RESULTS AND DISCUSSION

182 MS-CleanR Workflow and Implementation

183 Insert Figure 1

184 **Step 1: generic filters.** We first applied several generic filters to pre-clean the feature table of noise.
185 Starting from the alignment result file exported from MS-DIAL, the ratio between the mean of blank
186 samples and quality control (QC) samples (pool of all extracts) was calculated. All features exceeding
187 the user-defined threshold for this 'blank ratio' were removed. The ratio was calculated by using the
188 height of each feature because the normalized height can produce an increase in some blank signals.
189 This filter is also available in MS-DIAL, but MS-CleanR provides additional options for filtering ion
190 features. A second filter, called 'ghost blank peaks', is based on the high background ion drift we
191 observed in blank injections and in other samples that had a significant retention time (RT) shift
192 (Figure S1). These peaks had a low ratio of blank to sample class that excluded them from the usual
193 blank filtering approach. A third generic filter is based on an 'unusual mass decimal'. When singly
194 charged ions of basic organic molecules containing carbon, hydrogen, oxygen, and nitrogen are
195 considered, ion features with a value of more than eight at the first decimal place of m/z (mass to
196 charge ratio) are generally considered to be artifacts: this filter option can be disabled when working
197 with exceptions (e.g., phosphorylated compounds). A fourth generic filtering approach is the 'relative

198 standard deviation' (RSD) among sample classes. A high RSD value highlights poor ionization
199 repeatability. In our implementation, the RSD value was computed for each sample class and features
200 were removed if the RSD values in all sample classes exceeded a user-defined threshold. This
201 approach avoids incorrect feature deletions: in the case of large sample cohorts, for example, repeated
202 QC injections usually result in large RSDs because of a high dilution effect in the samples. Finally, we
203 introduced a fifth filter based on the 'relative mass defect' (RMD) calculation. The RMD is calculated
204 in ppm as $[(\text{mass defect}/\text{measured monoisotopic mass}) \times 10^6]$. It can be used to filter compound
205 classes²³ and it should also be useful to remove artifactual signals. Based on all compounds exported
206 from the Dictionary of Natural Products (DNP; available on DVD v.28.2 from CRC Press), we found
207 that 95% of natural products (NP)s had RMD values of 156.5-969.6 ppm. When this window was
208 extended to 99% of NPs, the range of RMD values was 52.05-2902.9 ppm.

209 **Step 2: Feature clustering.** To improve further the filtering process, we implemented a features
210 clustering function to be applied to those features remaining after the generic-based filtering described
211 above. The main goal of this step is to select the features arising from a unique metabolite signal
212 among each cluster by using the multi-level optimization of modularity algorithm²⁸. Feature clustering
213 is first based on the peak character estimation algorithm computed by MS-DIAL, which aggregates
214 several possible relationships at the same RT range: ion correlation among samples, MS/MS fragments
215 in higher m/z , possible adducts and chromatogram correlations²². Additionally, we also implemented
216 an index of possible neutral loss and a calculation of dimers/heteromers to tag clustered feature
217 relationships. Optionally, Pearson's correlation between features located in the same RT window
218 (typically of 0.025 minutes) can be computed, the strong correlation links being then considered
219 during the clustering process. If the study involves the same set of samples acquired both in PI and NI
220 mode, the MScombine²⁹ tool, incorporated into MS-CleanR, can be used to detect possible links
221 between positive and negative features appearing in the same RT window. This process corrects
222 misidentified relationships to consider observed m/z differences acquired between both ion modes.
223 The package can only treat PI or NI data independently, however. We observed that a unique
224 metabolite signal in each cluster can be selected by: a PI/NI adduct link (e.g. $[M+H]^+/[M-H]^-$,
225 $[M+Na]^+/[M+FA-H]^-$; the most intense peak of the cluster, and the peak with the most relationships to
226 other features (i.e. the highest 'degree' of connection). Among each cluster, one to n features (tunable
227 by the user) can be selected for further annotation: the most intense, the most connected or both. The
228 other features are removed from consideration.

229 **Step 3: Feature annotation.** After the above filtering steps, only a portion of the original features are
230 exported to MS-FINDER, which greatly accelerate the processing time. This software computes
231 feature annotations by querying internal DBs or imported DBs. Several DBs can be used to annotate a

232 single set of features by exporting the results for each DB used. Additionally, a “compound level”
233 column can be added into external DBs to further prioritize annotation within each DB used.

234 **Step 4: Annotated peak list.** This final step selects for each feature the best annotation among match
235 possibilities exported from MS-FINDER. In the case of multiple DB interrogation, the workflow
236 allows compound annotations to be ranked based on MS-FINDER score only or by prioritizing certain
237 DBs, depending on user choices. This latter function can greatly improve the annotation accuracy
238 particularly when dealing with taxonomically defined extracts³⁰. MS-CleanR can also prioritize
239 compounds based on “Compound_level” column tuned by the user in external DBs used for MS-
240 FINDER annotation. Finally, the resulting annotated peaks list can be converted into an .msp format
241 for mass spectral similarity networking as in GNPS³¹ or MetGem²⁵ (for the detailed mathematics of the
242 workflow, see Supporting Information Text 1).

243 [Workflow benchmarking on pooled standards](#)

244 To validate our approach, we benchmarked the MS-CleanR workflow by using a mixture of 51 NPs
245 standards profiled in NI and PI modes with a reverse phase column and a 25 minutes gradient. The
246 resulting data were compiled in an in-house DB comprising RT, HRMS and MS/MS fragmentation
247 patterns (DB-level 1 annotation according to the Metabolomics Standards Initiative-MSI³²). To test
248 whether the workflow retained features arising from unique metabolites and removed useless signals,
249 we compared the results obtained by using a combination of MS-DIAL and MS-FINDER and DB-
250 level 1 annotation to those obtained by using MS-CleanR. For the latter, we created another DB of the
251 same metabolite set encompassing accurate mass, molecular formula and SMILES strings (DB-level 2
252 annotation according to the MSI) to reproduce real-case annotation processing. All five generic filters
253 were used and the two most intense and two most connected features within each cluster were
254 exported for annotation by using the ‘formula prediction and structure elucidation by *in silico*
255 fragmentation tool’ in MS-FINDER (Table S1).

256 

257 As anticipated, we observed significant feature inflation in this mixture of 51 NP standards: 869
258 signals from PI and NI acquisition modes were detected (Figure 2). This approximately 95% feature
259 inflation is consistent with a previous report of 10 000-30 000 features detected after injection of 900
260 unique metabolites³³ and with a study that used isotope labeling as a feature filtering approach¹¹. Blank
261 ratio filtering deleted 50% of the features and the other generic filters described above removed 15%
262 of the remaining ones. Feature clustering resulted in a further reduction of 18%, resulting in a total of
263 115 features retained. Overall, the workflow filtered out 80% of all detected signals. By using this
264 approach, there was a remarkable improvement in the annotation rate (unique metabolites/detected

265 features) from 5% to 45% (Figure 2). Consequently, 21 metabolites displayed an isolated m/z -RT
266 signal whereas the others were grouped in clusters of two to eleven features (Figure 3A). Overall, 50
267 metabolites were annotated, 44 of which matched perfectly with level 1 annotation DB (Table S1).
268 The remaining ones were annotated as an isobaric/isomeric match because of prioritization of highest
269 MS-FINDER scoring value (e.g., 4-Aminosalicylic acid and 5-Aminosalicylic acid). In the case of
270 gramine, for example, the major pseudo-molecular ion had an m/z value of 130.06493 at RT 7.75
271 minutes (Figure 3B). By applying feature clustering, we detected an in-source fragment corresponding
272 to the neutral loss of the dimethylamine group at m/z 130.0649. This feature was removed and only the
273 signal at m/z 103.054 and m/z 175.1228 were exported for annotation. Since m/z 175.1228 was the
274 most intense peak, it was retained and annotated as gramine ($\Delta\text{ppm}=0.4$) with a perfect match. The
275 peak detected at RT 11.47 minutes was grouped in a cluster of 11 features, mainly related to similar
276 MS/MS spectra. In this case, the PI and NI clusters were merged according to their detected adduct
277 ($[\text{M}+\text{H}]^+$ and $[\text{M}-\text{H}]^-$, respectively) and the feature with highest MS-Finder annotation score was
278 retained in the final peak list and identified as neohesperidin dihydrochalcone ($\Delta\text{ppm}=0.4$).
279 Formononetin displayed complex adduct relationships in PI and NI modes and successive features
280 with higher m/z 's MS/MS fragment of formononetin in PI mode. The merging of PI and NI modes
281 allowed the main feature in this complex cluster to be selected and provided a perfect match with level
282 1 annotation DB. The only mismatch was encountered for phloridzin due to the neutral loss of a
283 glucose moiety in both in PI and NI modes. Only genine was detected in PI mode, resulting in
284 selection of this signal in the final peak list.

285 Insert Figure 3

286 To model more closely a real biological sample, we standardized our workflow by using a mixture of
287 167 standard compounds from the IROA Mass Spectrometry library (Table S2). As above, we found
288 significant feature inflation: 6732 signals after concatenation of PI and NI datasets (Figure S2). Unlike
289 the standardization with NPs, above, the generic filters removed only 15% of features. The most
290 important improvement was obtained by feature clustering, which filtered out 90% of the detected
291 features leaving 611 signals. Among these, 127 features were identified with a perfect match
292 compared to Level-1 annotation DB and 21 were annotated as an isomeric match (Table S2). Twelve
293 features were removed due to their co-elution with other compounds and four had a significant RT
294 shift due to their poor peak shapes. The final three compounds were not annotated because of neutral
295 loss of the same moiety in PI and NI modes, which led to their misidentification. Overall, the
296 annotation rate with this workflow was 27% (Figure S2) and 90% of unique metabolites were retained.

297 Evaluation of MS-CleanR on biological samples

298 To evaluate the utility of the workflow on a real dataset, we set up an experiment to compare
299 metabolome changes in *Arabidopsis thaliana* plants due to different culture conditions and age of the
300 plants. Three cultural conditions were assessed (low density growth in Jiffy® pots for 32 days, high
301 density growth in plastic pots for 21 days and hydroponic culture in liquid MS medium for 14 days)
302 and 10 biological replicates were analyzed per culture condition. At harvest time, 4 leaves (2
303 cotyledons and 2 leaves) were observed for hydroponic plants, the densely seeding plants showed not
304 more than two small, but completed, developed leaves, while the jiffy growing plants harbored large
305 and well developed rosette leaves. Extracts were made from the aerial parts of the plants grown in pots
306 and from the roots and green tissues of plants in hydroponic culture, and the extracts were profiled by
307 LC-MS. The datasets acquired in PI and NI modes were treated by using the MS-CleanR workflow
308 with default parameters (see Methods). Sequential principal component analysis (PCA) was used to
309 provide an unsupervised overview of the LC-MS fingerprints resulting from the generic filters and
310 feature clustering (Figure 4). The PCA score plot of raw PI and NI mode data displayed 51% of total
311 explained variance using the first two principal components. QC samples appeared in the center of the
312 PCA score plot, demonstrating the reproducibility of the LC-MS analysis. As expected, the youngest
313 plants growing hydroponically were completely separate on the first principal component (PC1) axis
314 from the older plants growing in pots. The plants growing in Jiffy pots and plastic pots could not be
315 distinguished in the raw dataset. After the generic filter step, the data from these latter two conditions
316 formed more distinct clusters, the total explained variance was slightly improved at 58% and the
317 number of features decreased by 35% (Figure 4). After the feature clustering step, the number of
318 features was reduced by 80%. All datasets were annotated with in-lab DB (level 1) and with MS-
319 FINDER (level 2) by reference to external DBs of *Arabidopsis* (Table S3) and Brassicaceae
320 compounds (Table S4) and an internal MS-FINDER plant-related DB (comprising PlantCyc,
321 KNApSAcK, HMDB, LIPID MAPS and UNPD). In the raw PI and NI dataset exported from MS-
322 DIAL (1163 features), 42% of all features were annotated, 26% of them appeared in the *Arabidopsis*
323 DB, 2% in the Brassicaceae DB, 7% in the internal MS-FINDER DBs and 6% with in-lab DB (Figure
324 4); 58% of all features were unidentified. The generic filters removed 15% of all features and
325 increased the annotation rate to 59%. Feature clustering drastically reduced the number of features
326 ($254 m/z \times RT$ pairs) and increased the annotation rate to 73%. Using annotation DB prioritization,
327 53% of retained features were annotated in *Arabidopsis* genus and 13% at level 1 with in-lab DB, only
328 27% remained unidentified. Orthogonal projections to latent structures discriminant analysis (OPLS-
329 DA) of the most highly ranked features identified three amino acids (oxoproline, citrulline and
330 glutamine) that discriminate between growth in pots and hydroponic growth (Table S5). This may be
331 related to differences in nitrogen availability in the hydroponics medium and in potting soil.

332

Insert Figure 4

333

334 Metabolic profiling with MS-CleanR

335 Untargeted metabolomic profiling has emerged as a method of choice to identify metabolic markers
336 associated with beneficial traits in plants, such as resistance to biotic stresses. In this context, the MS-
337 CleanR workflow could greatly improve the results of untargeted metabolomics. To illustrate this
338 point, we used as a model the legume *Medicago truncatula* and the pathogenic oomycete
339 *Aphanomyces euteiches*, a major pathogen of several legume species³⁴. Genome-wide association
340 studies of 179 lines of *M. truncatula* have identified major loci involved in the resistance of the plant
341 to *A. euteiches*. Moreover, genes encoding enzymes involved in the synthesis of antimicrobial
342 metabolites are expressed in uninfected plants³⁵. This suggests that antimicrobial metabolites in
343 uninfected plants may be useful biomarkers with which to select legumes lines resistant to *A.*
344 *euteiches*. To identify these metabolites, we applied the MS-CleanR workflow to analyze the
345 metabolomes of roots from three different strains of *M. truncatula* that have different levels of
346 resistance to *A. euteiches* infection: strain DZA45.5 has the highest level of resistance, A17 an
347 intermediate level, and F83 is the most susceptible³⁶. These three strains were analyzed by LC-MS in
348 NI mode and potential biomarkers were highlighted by multivariate data analysis (Table S6). The
349 metabolites that were differentially produced in the two most resistant strains (A17 and DZA45.5)
350 when compared to the more sensitive one (F83) were identified by OPLS regression.

351

Insert Figure 5

352 After application of the MS-CleanR workflow, the PCA score plot showed a net clustering of the
353 samples from each strain of *M. truncatula*. QC samples were centered on the PCA plot demonstrating
354 very good reproducibility (Figure 5). When annotated by reference to DBs from *Medicago* or the
355 legume family Fabaceae, 60% of the dataset was annotated (Figure 5) and an additional 9% with MS-
356 FINDER DBs. A molecular spectral similarity network was built to highlight common chemical class
357 related to resistance traits (Figure 6). Among all annotated features, flavonoids and terpene glycosides
358 compounds were prevalent. This latter class encompass mostly triterpene sapogenins which appeared
359 to be highly correlated to the resistance traits according to the OPLS regression model. In particular,
360 the four top ranked compounds belonged to two clusters related to sapogenins and one to flavonoids.
361 Our untargeted approach revealed the presence of Apigenin-7-O-glucuronopyranoside (best MS-
362 FINDER score among several possible match in flavonoid class) only in the resistant DZ45.5 strain.
363 This result corroborated a previous study by our group which demonstrated the implication of
364 flavonoid pathway in resistance³⁵. However, other detected flavonoids were not correlated to the
365 resistance contrary to sapogenins class. Among the 151 terpene glycosides annotated in this study, 36
366 were also identified by a large-scale sapogenin profiling study in various ecotypes of *M. truncatula*³⁷
367 (Table S6). Interestingly, the three-top ranked sapogenins by OPLS model (Azukisaponin III,

368 Arjunolic acid 3-glucoside and Soyasaponin I) displayed an isobaric match with tow hederagenin
369 glycoside and a bayogenin derivatives respectively annotated by Sumner and colleagues. These
370 saponin accumulates preferentially in roots than in leaves. These organs, however, have distinct
371 profiles of specific saponins, which may be explained by the adaptation of each ecotype to its biotic
372 environment. A previous study, for example, showed that saponins derived from hederagenin
373 glycoside in *M. truncatula* have antifungal activity³⁸. Our study confirmed a higher level of these
374 compounds in the strains resistant to *A. euteiches* (DZA45.5 and A17) than it is in the sensitive strain
375 F83. Although the relevance of saponins to resistance of *M. truncatula* to *A. euteiches* remains to be
376 confirmed, these findings demonstrate the potential value of applying metabolomics tools to identify
377 biomarkers of plant resistance.

378 Insert Figure 6

379 CONCLUSIONS

380 The main goal of LC-MS-based untargeted metabolomics is to convert chromatographic profiles of
381 complex biological extracts into a comprehensive metabolite list. Professor Ian Wilson summarized
382 the challenge thus: “LC-MS includes everything, which means you see everything. Thus, the challenge
383 is to take oceans of data, and make rivers of information, and finally puddles of knowledge.” (NIH
384 Metabolomics symposium, August 2013). We demonstrate here that feature degeneracy - the ocean of
385 data - has a great impact on the final annotated peak list information, thus impacting the biological
386 knowledge mined from untargeted metabolomic studies. We estimate, based on analysis of standard
387 mixtures, that feature inflation is close to 95%, in agreement with other studies^{33,11}. Our package MS-
388 CleanR, with its a point-and-click software on a Shiny interface, is a new component in the suite of
389 tools comprising the GUI software MS-DIAL and the annotation capabilities of MS-FINDER which
390 altogether provide a comprehensive workflow, from raw data to final annotated peaklist. MS-CleanR
391 can reduce the number of features by 80-90% and keep most unique metabolite signals without
392 compromising the final data structure. The opportunity to rank the annotation results with reference to
393 in-house databases narrows down the final identification possibilities. Additionally, the package is
394 able to combine both PI and NI mode (*A. thaliana* experiment) or to treat only one mode (*M.*
395 *truncatula* study) depending of the study objectives. We demonstrate the utility of this workflow by
396 analyzing secondary metabolites levels in three *M. truncatula* strains with different susceptibilities to a
397 pathogenic oomycete. We could annotate 70% of the dataset with 60% at the genus or family level
398 using DBs prioritization. The resulting mass spectral similarity network further supports annotation
399 results since most clusters gathered the same metabolite chemical class. Still, our approach was unable
400 to keep only unique metabolite features regarding the annotation rate comprising between 24 and 45%
401 for standard mixtures. A limitation of our filtering process is its dependence to chromatographic

402 resolution, which can seriously impair the final results by clustering several unique metabolites
403 together. In the present study, we chose a twenty minutes gradient, like those generally applied in most
404 untargeted metabolomics studies. Extending the elution time might improve the chromatographic
405 resolution but is difficult to apply in day-to-day work, especially for high-throughput experiments.
406 These challenges will be addressed in future developments of MS-CleanR.

407

408 ASSOCIATED CONTENT

409 The Supporting Information is available free of charge on...

- 410 • figure S1. Alignment spot screenshot in ESI PI and NI ionization modes showing repeated
411 blank pseudomolecular ions detected massively in QCs samples with a retention time shift.
412 (PDF)
- 413 • Figure S2. Features filtering of LC-MS dataset from 167 IROA-MS standards library
414 according to generic filters and clustering algorithm. Barplot display feature counts after
415 successive filters. Line plot display annotation rate (unique metabolites/feature counts in %).
416 (PDF)
- 417 • Sup Table S1: Excel table with cluster annotation, result summary and database for level 2
418 annotation imported in MS-FINDER for the 51 NPs dataset (XLSX)
- 419 • Sup Table S2: Excel table with cluster annotation, result summary and database for level 2
420 annotation imported in MS-FINDER for the 167 IROA MS standard library dataset (XLSX)
- 421 • Sup Table S3: Arabidopsis DB used as input for level 2 annotation in MS-FINDER (TXT)
- 422 • Sup Table S4: Brassicaceae DB used as input for level 2 annotation in MS-FINDER (TXT)
- 423 • Sup Table S5: Arabidopsis dataset treated by MS-CleanR and OPLS-DA top ranked features
424 (XLSX)
- 425 • Sup Table S6: Medicago dataset treated by MS-CleanR and OPLS regression coefficient for
426 feature ranking (XLSX)
- 427 • Sup Table S7: Medicago DB used as input for level 2 annotation in MS-FINDER (TXT)
- 428 • Sup Table S8: Fabaceae DB used as input for level 2 annotation in MS-FINDER (TXT)
- 429 • Raw data from *Arabidopsis* and *Medicago* LC-MS profiling available on Zenodo using the
430 DOI: 10.5281/zenodo.3744480

431 AUTHOR INFORMATION

432 Corresponding author
433 Phone: (+33) 534 32 38 31; mail: guillaume.marti@univ-tlse3.fr
434 ORCID : <https://orcid.org/0000-0002-6321-9005>

435

436 ACKNOWLEDGMENTS

437 We thank Dr. Stephane Bertani for providing us standards from the IROA-MS Library. Financial
438 support from The French national infrastructure for metabolomics and fluxomics, MetaboHUB-ANR-
439 11-INBS-0010 and PSPC Solstice project supported by Bpifrance (SOLutionS for Integrates
440 Treatments under Environmental Management). We thank E Amblard, N. Jariais and C. Jacquet for *M.*
441 *truncatula* cultures and A. Haouy for *A. thalina* cultures and sample preparations. We also
442 acknowledge Carol Featherstone of Plume Scientific Communication Services for professional
443 scientific editing.

444 References

445

- 446 (1) Zierer, J.; Jackson, M. A.; Kastenmüller, G.; Mangino, M.; Long, T.; Telenti, A.;
447 Mohney, R. P.; Small, K. S.; Bell, J. T.; Steves, C. J.; Valdes, A. M.; Spector, T. D.;
448 Menni, C. The Fecal Metabolome as a Functional Readout of the Gut Microbiome. *Nat.*
449 *Genet.* **2018**, *50* (6), 790-795. <https://doi.org/10.1038/s41588-018-0135-7>.
- 450 (2) Li, H.; Ning, S.; Ghandi, M.; Kryukov, G. V.; Gopal, S.; Deik, A.; Souza, A.; Pierce,
451 K.; Keskula, P.; Hernandez, D.; Ann, J.; Shkoza, D.; Apfel, V.; Zou, Y.; Vazquez, F.;
452 Barretina, J.; Pagliarini, R. A.; Galli, G. G.; Root, D. E.; Hahn, W. C.; Tsherniak, A.;
453 Giannakis, M.; Schreiber, S. L.; Clish, C. B.; Garraway, L. A.; Sellers, W. R. The
454 Landscape of Cancer Cell Line Metabolism. *Nat. Med.* **2019**, *25* (5), 850-860.
455 <https://doi.org/10.1038/s41591-019-0404-8>.
- 456 (3) Gargallo-Garriga, A.; Sardans, J.; Pérez-Trujillo, M.; Oravec, M.; Urban, O.; Jentsch,
457 A.; Kreyling, J.; Beierkuhnlein, C.; Parella, T.; Peñuelas, J. Warming Differentially
458 Influences the Effects of Drought on Stoichiometry and Metabolomics in Shoots and
459 Roots. *New Phytol.* **2015**, *207* (3), 591-603. <https://doi.org/10.1111/nph.13377>.
- 460 (4) Mohimani, H.; Gurevich, A.; Shlemov, A.; Mikheenko, A.; Korobeynikov, A.; Cao, L.;
461 Shcherbin, E.; Nothias, L.-F.; Dorrestein, P. C.; Pevzner, P. A. Dereplication of
462 Microbial Metabolites through Database Search of Mass Spectra. *Nat. Commun.* **2018**,
463 *9* (1). <https://doi.org/10.1038/s41467-018-06082-8>.
- 464 (5) Wang, S.; Alseekh, S.; Fernie, A. R.; Luo, J. The Structure and Function of Major Plant
465 Metabolite Modifications. *Mol. Plant* **2019**, *12* (7), 899-919.
466 <https://doi.org/10.1016/j.molp.2019.06.001>.
- 467 (6) Huan, T.; Forsberg, E. M.; Rinehart, D.; Johnson, C. H.; Ivanisevic, J.; Benton, H. P.;
468 Fang, M.; Aisporna, A.; Hilmers, B.; Poole, F. L.; Thorgersen, M. P.; Adams, M. W.
469 W.; Krantz, G.; Fields, M. W.; Robbins, P. D.; Niedernhofer, L. J.; Ideker, T.;
470 Majumder, E. L.; Wall, J. D.; Rattray, N. J. W.; Goodacre, R.; Lairson, L. L.; Siuzdak,
471 G. Systems Biology Guided by XCMS Online Metabolomics. *Nat Meth* **2017**, *14* (5),
472 461-462.
- 473 (7) Giacomoni, F.; Le Corguille, G.; Monsoor, M.; Landi, M.; Pericard, P.; Petera, M.;
474 Duperier, C.; Tremblay-Franco, M.; Martin, J.-F.; Jacob, D.; Goulitquer, S.; Thevenot,

- 475 E. A.; Caron, C. Workflow4Metabolomics: A Collaborative Research Infrastructure for
476 Computational Metabolomics. *Bioinformatics* **2015**, *31* (9), 1493-1495.
477 <https://doi.org/10.1093/bioinformatics/btu813>.
- 478 (8) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Orešič, M. MZmine 2: Modular Framework
479 for Processing, Visualizing, and Analyzing Mass Spectrometry-Based Molecular
480 Profile Data. *BMC Bioinformatics* **2010**, *11* (1), 1.
- 481 (9) Tsugawa, H.; Cajka, T.; Kind, T.; Ma, Y.; Higgins, B.; Ikeda, K.; Kanazawa, M.;
482 VanderGheynst, J.; Fiehn, O.; Arita, M. MS-DIAL: Data-Independent MS/MS
483 Deconvolution for Comprehensive Metabolome Analysis. *Nat. Methods* **2015**, *12* (6),
484 523-526. <https://doi.org/10.1038/nmeth.3393>.
- 485 (10) Patti, G. J.; Yanes, O.; Siuzdak, G. Metabolomics: The Apogee of the Omics Trilogy.
486 *Nat. Rev. Mol. Cell Biol.* **2012**, *13* (4), 263-269. <https://doi.org/10.1038/nrm3314>.
- 487 (11) Mahieu, N. G.; Patti, G. J. Systems-Level Annotation of a Metabolomics Data Set
488 Reduces 25 000 Features to Fewer than 1000 Unique Metabolites. *Anal. Chem.* **2017**,
489 *89* (19), 10397-10406. <https://doi.org/10.1021/acs.analchem.7b02380>.
- 490 (12) Dührkop, K.; Fleischauer, M.; Ludwig, M.; Aksenov, A. A.; Melnik, A. V.; Meusel,
491 M.; Dorrestein, P. C.; Rousu, J.; Böcker, S. SIRIUS 4: A Rapid Tool for Turning
492 Tandem Mass Spectra into Metabolite Structure Information. *Nat. Methods* **2019**, *16*
493 (4), 299-302. <https://doi.org/10.1038/s41592-019-0344-8>.
- 494 (13) Tsugawa, H.; Kind, T.; Nakabayashi, R.; Yukihira, D.; Tanaka, W.; Cajka, T.; Saito,
495 K.; Fiehn, O.; Arita, M. Hydrogen Rearrangement Rules: Computational MS/MS
496 Fragmentation and Structure Elucidation Using MS-FINDER Software. *Anal. Chem.*
497 **2016**. <https://doi.org/10.1021/acs.analchem.6b00770>.
- 498 (14) Ruttkies, C.; Schymanski, E. L.; Wolf, S.; Hollender, J.; Neumann, S. MetFrag
499 Relaunch: Incorporating Strategies beyond in Silico Fragmentation. *J.*
500 *Cheminformatics* **2016**, *8* (1), 3. <https://doi.org/10.1186/s13321-016-0115-9>.
- 501 (15) Djoumbou-Feunang, Y.; Pon, A.; Karu, N.; Zheng, J.; Li, C.; Arndt, D.; Gautam, M.;
502 Allen, F.; Wishart, D. S. CFM-ID 3.0: Significantly Improved ESI-MS/MS Prediction
503 and Compound Identification. *Metabolites* **2019**, *9* (4), 72.
504 <https://doi.org/10.3390/metabo9040072>.
- 505 (16) Kuhl, C.; Tautenhahn, R.; Böttcher, C.; Larson, T. R.; Neumann, S. CAMERA: An
506 Integrated Strategy for Compound Spectra Extraction and Annotation of Liquid
507 Chromatography/Mass Spectrometry Data Sets. *Anal. Chem.* **2012**, *84* (1), 283-289.
508 <https://doi.org/10.1021/ac202450g>.
- 509 (17) Broeckling, C. D.; Afsar, F. A.; Neumann, S.; Ben-Hur, A.; Prenni, J. E. RAMClust: A
510 Novel Feature Clustering Method Enables Spectral-Matching-Based Annotation for
511 Metabolomics Data. *Anal. Chem.* **2014**, *86* (14), 6812-6817.
512 <https://doi.org/10.1021/ac501530d>.
- 513 (18) DeFelice, B. C.; Mehta, S. S.; Samra, S.; Čajka, T.; Wancewicz, B.; Fahrman, J. F.;
514 Fiehn, O. Mass Spectral Feature List Optimizer (MS-FLO): A Tool To Minimize False
515 Positive Peak Reports in Untargeted Liquid Chromatography-Mass Spectroscopy (LC-
516 MS) Data Processing. *Anal. Chem.* **2017**, *89* (6), 3250-3255.
517 <https://doi.org/10.1021/acs.analchem.6b04372>.

- 518 (19) Mahieu, N. G.; Spalding, J. L.; Gelman, S. J.; Patti, G. J. Defining and Detecting
519 Complex Peak Relationships in Mass Spectral Data: The Mz.Unity Algorithm. *Anal.*
520 *Chem.* **2016**. <https://doi.org/10.1021/acs.analchem.6b01702>.
- 521 (20) Kantz, E.; Tiwari, S.; Watrous, J. D.; Cheng, S.; Jain, M. Deep Neural Networks for
522 Classification of LC-MS Spectral Peaks. *Anal. Chem.* **2019**, *acs.analchem.9b02983*.
523 <https://doi.org/10.1021/acs.analchem.9b02983>.
- 524 (21) Melnikov, A. D.; Tsentalovich, Y. P.; Yanshole, V. V. Deep Learning for the Precise
525 Peak Detection in High-Resolution LC-MS Data. *Anal. Chem.* **2020**, *92* (1), 588-592.
526 <https://doi.org/10.1021/acs.analchem.9b04811>.
- 527 (22) Tsugawa, H.; Nakabayashi, R.; Mori, T.; Yamada, Y.; Takahashi, M.; Rai, A.;
528 Sugiyama, R.; Yamamoto, H.; Nakaya, T.; Yamazaki, M.; Kooke, R.; Bac-Molenaar, J.
529 A.; Oztolan-Erol, N.; Keurentjes, J. J. B.; Arita, M.; Saito, K. A Cheminformatics
530 Approach to Characterize Metabolomes in Stable-Isotope-Labeled Organisms. *Nat.*
531 *Methods* **2019**, *16* (4), 295-298. <https://doi.org/10.1038/s41592-019-0358-2>.
- 532 (23) Salem, M. A.; Jüppner, J.; Bajdzienko, K.; Giavalisco, P. Protocol: A Fast,
533 Comprehensive and Reproducible One-Step Extraction Method for the Rapid
534 Preparation of Polar and Semi-Polar Metabolites, Lipids, Proteins, Starch and Cell Wall
535 Polymers from a Single Sample. *Plant Methods* **2016**, *12* (1), 45.
536 <https://doi.org/10.1186/s13007-016-0146-2>.
- 537 (24) BÉCARD, G.; FORTIN, J. A. Early Events of Vesicular-Arbuscular Mycorrhiza
538 Formation on Ri T-DNA Transformed Roots. *New Phytol.* **1988**, *108* (2), 211-218.
539 <https://doi.org/10.1111/j.1469-8137.1988.tb03698.x>.
- 540 (25) Olivon, F.; Elie, N.; Grelier, G.; Roussi, F.; Litaudon, M.; Touboul, D. MetGem
541 Software for the Generation of Molecular Networks Based on T-SNE Algorithm. *Anal.*
542 *Chem.* **2018**. <https://doi.org/10.1021/acs.analchem.8b03099>.
- 543 (26) Shannon, P. Cytoscape: A Software Environment for Integrated Models of
544 Biomolecular Interaction Networks. *Genome Res.* **2003**, *13* (11), 2498-2504.
545 <https://doi.org/10.1101/gr.1239303>.
- 546 (27) Ekanayaka, E. A. P.; Celiz, M. D.; Jones, A. D. Relative Mass Defect Filtering of Mass
547 Spectra: A Path to Discovery of Plant Specialized Metabolites. *Plant Physiol.* **2015**,
548 pp.114.251165. <https://doi.org/10.1104/pp.114.251165>.
- 549 (28) Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; Lefebvre, E. Fast Unfolding of
550 Communities in Large Networks. *J. Stat. Mech. Theory Exp.* **2008**, *2008* (10), P10008.
551 <https://doi.org/10.1088/1742-5468/2008/10/p10008>.
- 552 (29) Calderón-Santiago, M.; Fernández-Peralbo, M. A.; Priego-Capote, F.; Luque de Castro,
553 M. D. MSCombine: A Tool for Merging Untargeted Metabolomic Data from High-
554 Resolution Mass Spectrometry in the Positive and Negative Ionization Modes.
555 *Metabolomics* **2016**, *12* (3). <https://doi.org/10.1007/s11306-016-0970-4>.
- 556 (30) Rutz, A.; Dounoue-Kubo, M.; Ollivier, S.; Bisson, J.; Bagheri, M.; Saesong, T.;
557 Ebrahimi, S. N.; Ingkaninan, K.; Wolfender, J.-L.; Allard, P.-M. Taxonomically
558 Informed Scoring Enhances Confidence in Natural Products Annotation. *Front. Plant*
559 *Sci.* **2019**, *10*, 1329. <https://doi.org/10.3389/fpls.2019.01329>.
- 560 (31) Nothias, L. F.; Petras, D.; Schmid, R.; Dührkop, K.; Rainer, J.; Sarvepalli, A.;
561 Protsyuk, I.; Ernst, M.; Tsugawa, H.; Fleischauer, M.; Aicheler, F.; Aksenov, A.; Alka,
562 O.; Allard, P.-M.; Barsch, A.; Cachet, X.; Caraballo, M.; Da Silva, R. R.; Dang, T.;

- 563 Garg, N.; Gauglitz, J. M.; Gurevich, A.; Isaac, G.; Jarmusch, A. K.; Kamenik, Z.;
564 Kang, K. B.; Kessler, N.; Koester, I.; Korf, A.; Gouellec, A. L.; Ludwig, M.; Christian,
565 M. H.; McCall, L.-I.; McSayles, J.; Meyer, S. W.; Mohimani, H.; Morsy, M.; Moyne,
566 O.; Neumann, S.; Neuweger, H.; Nguyen, N. H.; Nothias-Esposito, M.; Paolini, J.;
567 Phelan, V. V.; Pluskal, T.; Quinn, R. A.; Rogers, S.; Shrestha, B.; Tripathi, A.; van der
568 Hooft, J. J. J.; Vargas, F.; Weldon, K. C.; Witting, M.; Yang, H.; Zhang, Z.; Zubeil, F.;
569 Kohlbacher, O.; Böcker, S.; Alexandrov, T.; Bandeira, N.; Wang, M.; Dorrestein, P. C.
570 *Feature-Based Molecular Networking in the GNPS Analysis Environment*; preprint;
571 Bioinformatics, 2019. <https://doi.org/10.1101/812404>.
- 572 (32) Creek, D. J.; Dunn, W. B.; Fiehn, O.; Griffin, J. L.; Hall, R. D.; Lei, Z.; Mistrik, R.;
573 Neumann, S.; Schymanski, E. L.; Sumner, L. W.; Trengove, R.; Wolfender, J.-L.
574 Metabolite Identification: Are You Sure? And How Do Your Peers Gauge Your
575 Confidence? *Metabolomics* **2014**, *10* (3), 350-353. [https://doi.org/10.1007/s11306-014-](https://doi.org/10.1007/s11306-014-0656-8)
576 0656-8.
- 577 (33) Li, Z.; Lu, Y.; Guo, Y.; Cao, H.; Wang, Q.; Shui, W. Comprehensive Evaluation of
578 Untargeted Metabolomics Data Processing Software in Feature Detection,
579 Quantification and Discriminating Marker Selection. *Anal. Chim. Acta* **2018**, *1029*, 50-
580 57. <https://doi.org/10.1016/j.aca.2018.05.001>.
- 581 (34) Gaulin, E.; Jacquet, C.; Bottin, A.; Dumas, B. Root Rot Disease of Legumes Caused by
582 *Aphanomyces Euteiches*. *Mol. Plant Pathol.* **2007**, *8* (5), 539-548.
583 <https://doi.org/10.1111/j.1364-3703.2007.00413.x>.
- 584 (35) Badis, Y.; Bonhomme, M.; Lafitte, C.; Huguet, S.; Balzergue, S.; Dumas, B.; Jacquet,
585 C. Transcriptome Analysis Highlights Preformed Defences and Signalling Pathways
586 Controlled by the *PrAe1* Quantitative Trait Locus (QTL), Conferring Partial Resistance
587 to *Aphanomyces Euteiches* in *Medicago Truncatula*: Molecular Mechanisms Controlled
588 by the *PrAe1* QTL. *Mol. Plant Pathol.* **2015**, *16* (9), 973-986.
589 <https://doi.org/10.1111/mpp.12253>.
- 590 (36) Bonhomme, M.; André, O.; Badis, Y.; Ronfort, J.; Burgarella, C.; Chantret, N.;
591 Prospero, J.-M.; Briskine, R.; Mudge, J.; Debéllé, F.; Navier, H.; Miteul, H.; Hajri, A.;
592 Baranger, A.; Tiffin, P.; Dumas, B.; Pilet-Nayel, M.-L.; Young, N. D.; Jacquet, C.
593 High-Density Genome-Wide Association Mapping Implicates an F-Box Encoding
594 Gene in *Medicago Truncatula* Resistance to *Aphanomyces Euteiches*. *New Phytol.*
595 **2014**, *201* (4), 1328-1342. <https://doi.org/10.1111/nph.12611>.
- 596 (37) Lei, Z.; Watson, B. S.; Huhman, D.; Yang, D. S.; Sumner, L. W. Large-Scale Profiling
597 of Saponins in Different Ecotypes of *Medicago Truncatula*. *Front. Plant Sci.* **2019**, *10*.
598 <https://doi.org/10.3389/fpls.2019.00850>.
- 599 (38) Abbruscato, P.; Tosi, S.; Crispino, L.; Biazzini, E.; Menin, B.; Picco, A. M.; Pecetti, L.;
600 Avato, P.; Tava, A. Triterpenoid Glycosides from *Medicago Sativa* as Antifungal
601 Agents against *Pyricularia Oryzae*. *J. Agric. Food Chem.* **2014**, *62* (46), 11030-11036.
602 <https://doi.org/10.1021/jf5049063>.

603

604

605

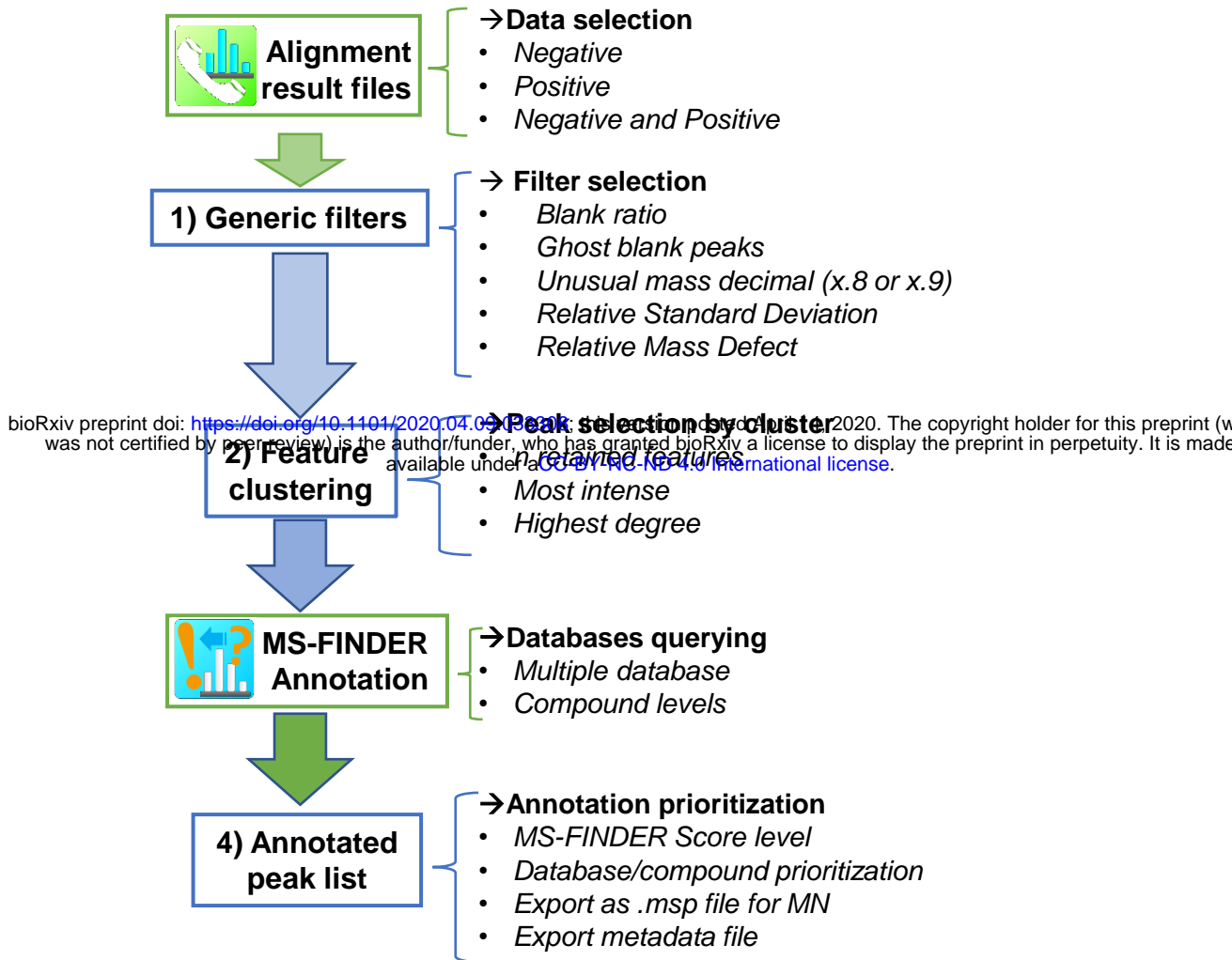


Figure 1: MS-CleanR workflow. Description of each step in the Shiny user interface workflow and the options available at each step. (MN, mass spectral similarity networking)

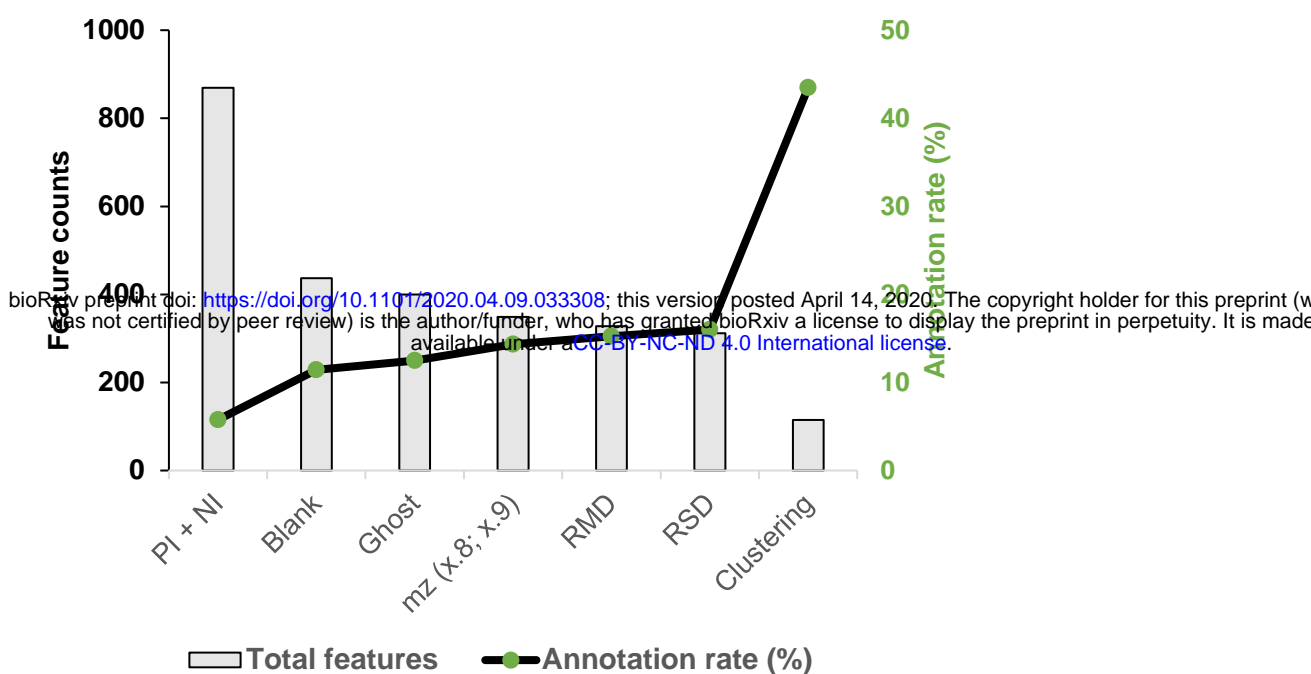
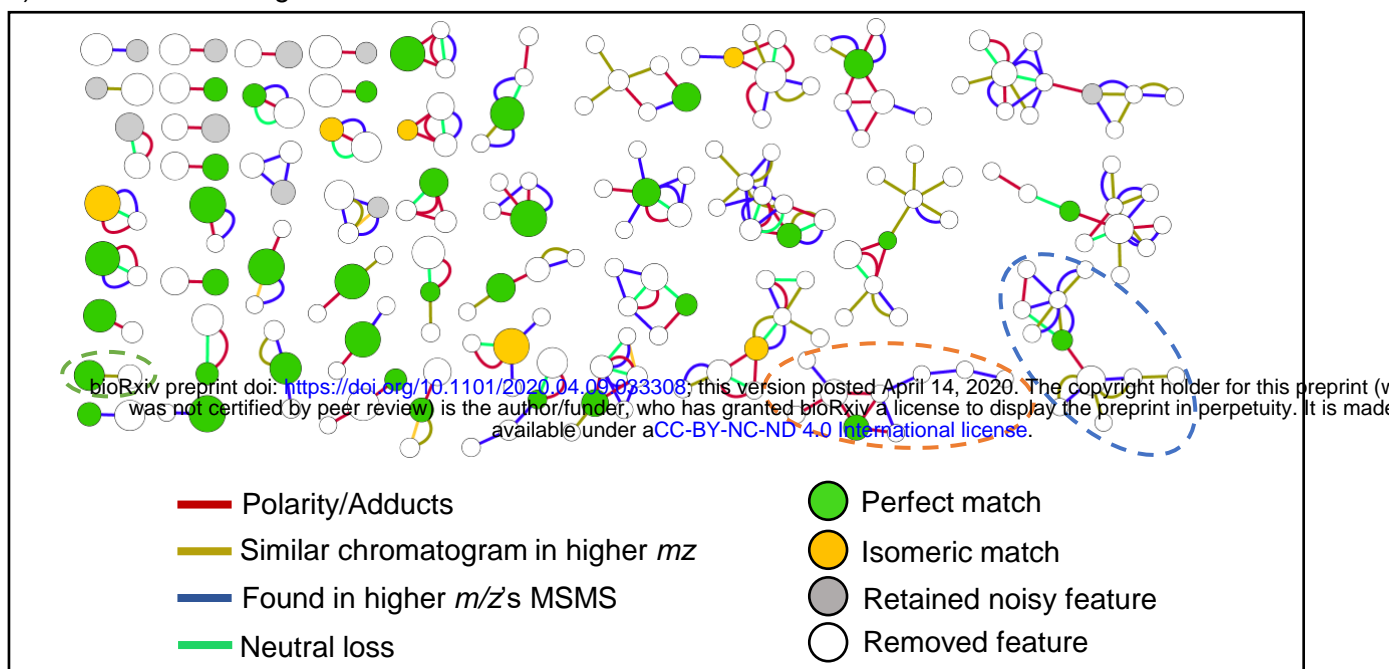


Figure 2. Feature filtering of the LC-MS dataset from 51 NPs standards. Generic filters and the feature clustering algorithm were applied to the initial PI + NI mode dataset. The bar plot displays feature counts after successive filters. The line plot displays annotation rate (unique metabolites/feature counts in %).

A) Features clustering



B) LC-MS chromatograms of 51 NPs

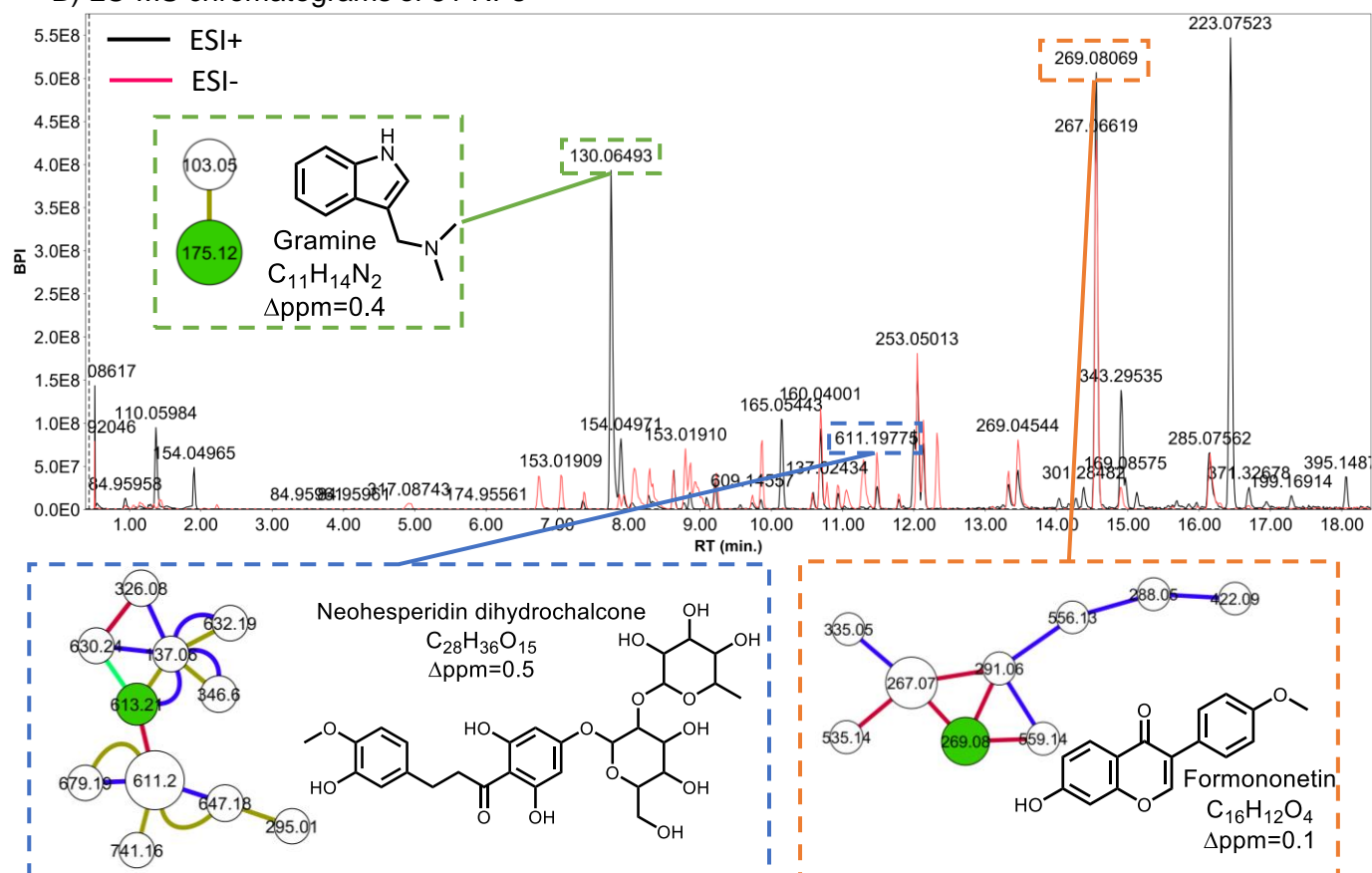


Figure 3. MS-CleanR feature clustering of 51 NPs. Clustering was based on the peak character estimation and multi-level optimization of modularity algorithms. A) Cluster plot of the whole dataset excluding size one clusters. B) UHPLC-HRMS base peak intensity (BPI) chromatogram of the standards mixture containing 51 NPs. Three representative compounds and their respective clusters are indicated.

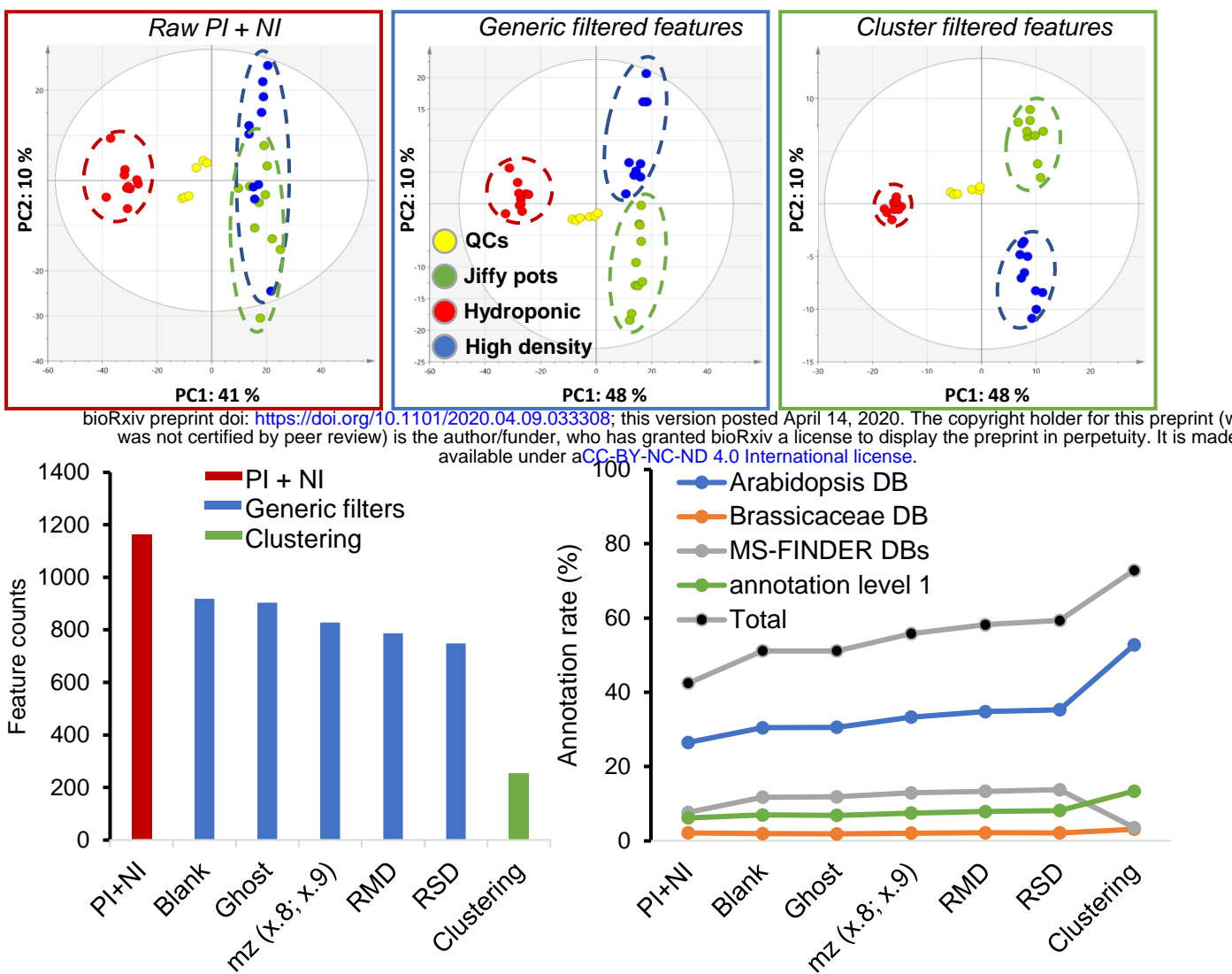


Figure 4. LC-MS dataset processing of the metabolomes of *A. thaliana* plants growing in different conditions. Top: Sequential PCA score plots of raw PI and NI mode data and the data after applying generic filters and feature clustering. Dotted circles indicate biological sample type distribution (yellow, QC injections; green, plants growing in Jiffy pots at low density; blue, plants growing in plastic pots at high density; red plants in hydroponic culture). Bottom: The bar plot shows the feature counts after successive filtering steps. The line plot displays the annotation rate (unique metabolites/feature counts expressed as %) after successive filtering steps using annotation DBs prioritization.

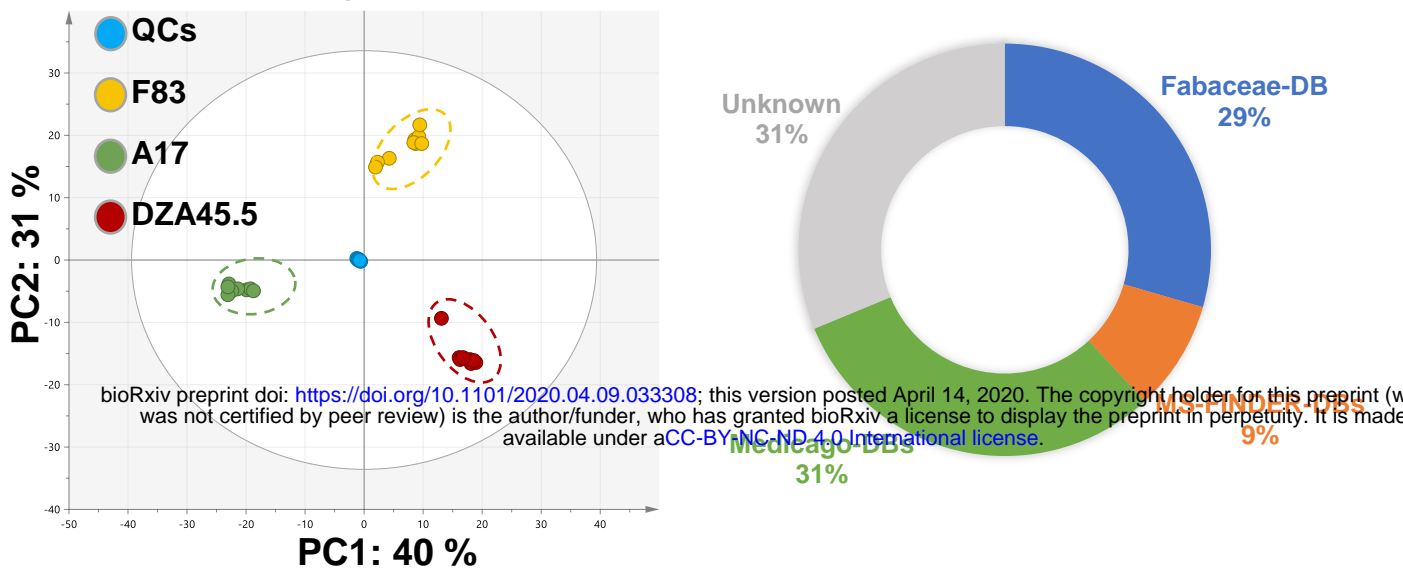


Figure 5. LC-MS NI dataset processing of the metabolome of roots from three strains of *M. truncatula*. Left: PCA score plot after applying the MS-CleanR workflow. Dotted circles enclose samples from each plant strain. Right: Circular plot of the proportions of features annotated with reference to the indicated databases (DB).

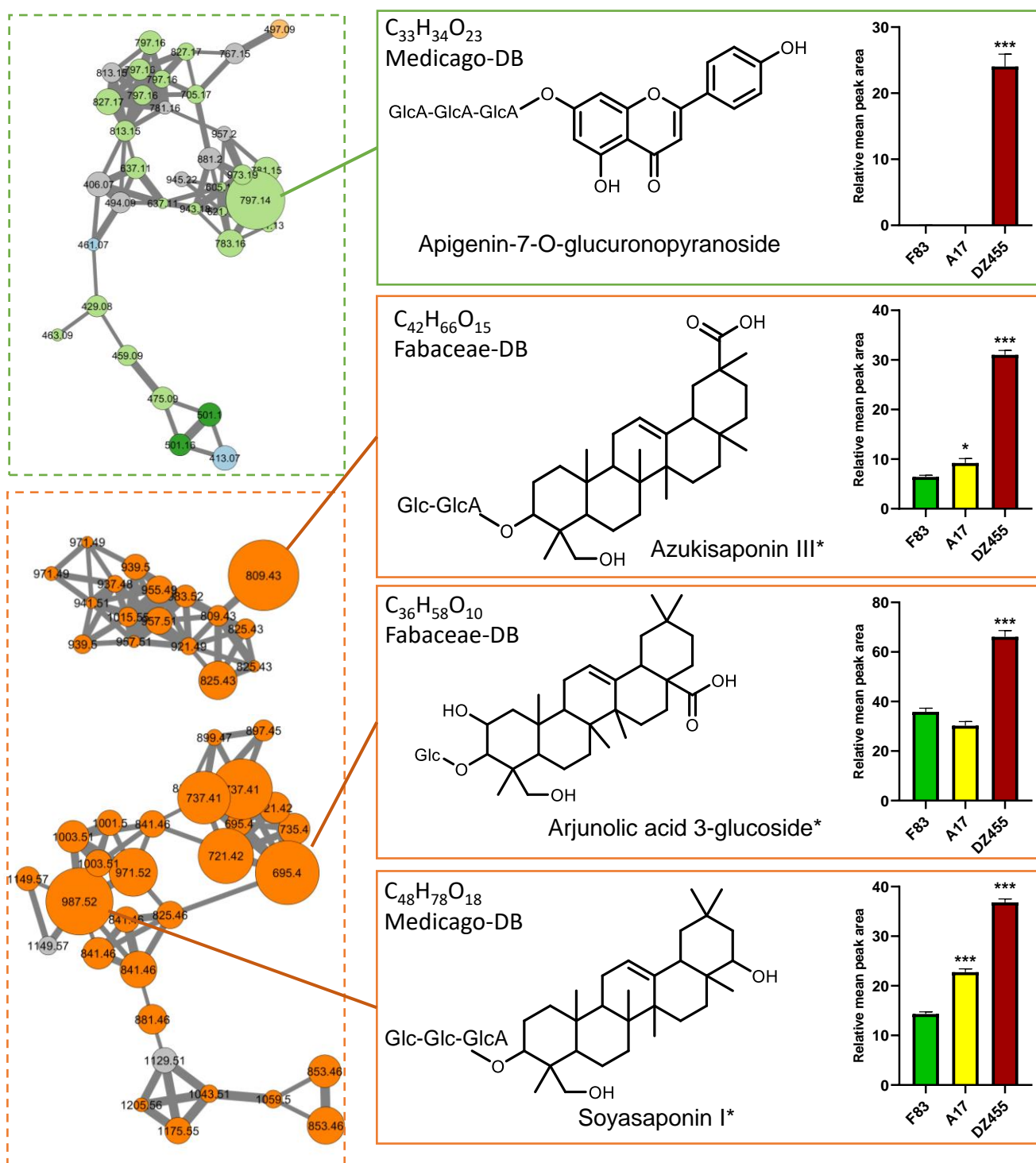
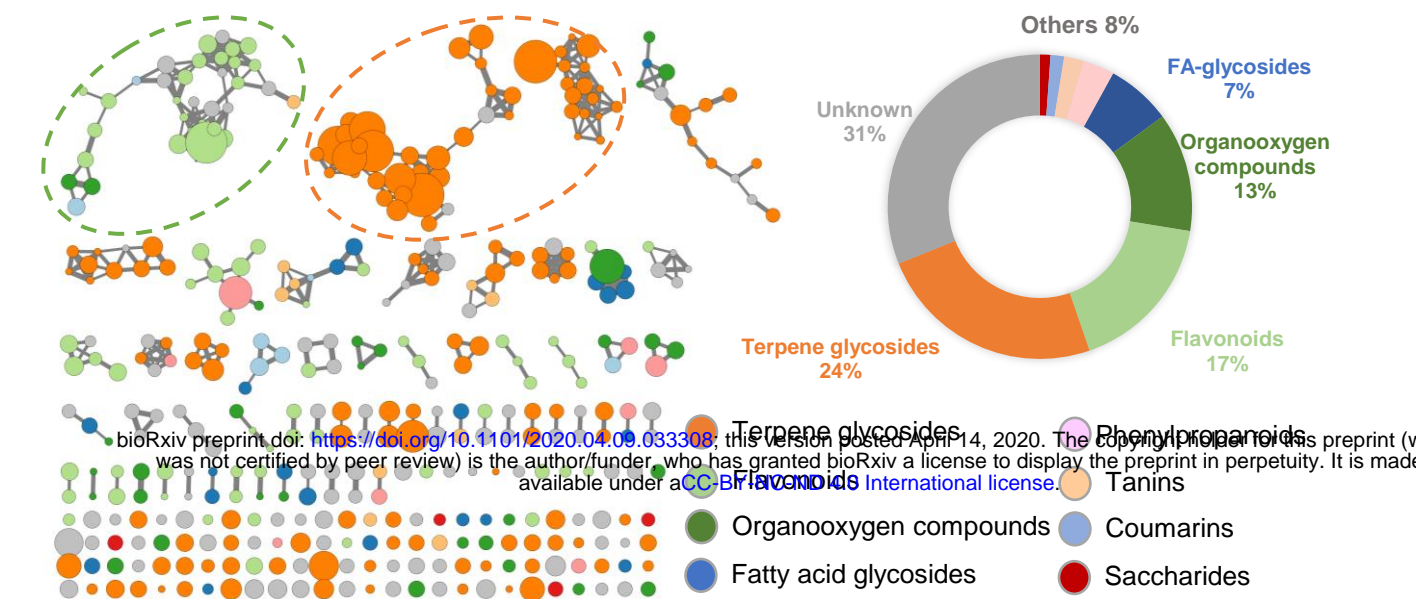


Figure 6. Mass spectral similarity network of *M. truncatula* NI dataset (cosine ≥ 0.8). Nodes are colored according to their chemical classes and sized relative to their OPLS regression coefficient score (See text for details). Edge width is proportional to cosine value. Pie chart display annotated chemical class ratio in LC-MS NI dataset (Others include coumarins derivatives, tanins and saccharides chemical classes). Bar plots display normalized mean peak areas for the four most highly ranked structures by OPLS-regression modeling (Table S6). One-way ANOVA and Dunnett's post-hoc test ($p \leq 0.05$) were used to assess differences between the sensitive (F83) and resistant (A17 and DZA45.5) *M. truncatula* strains ($p \leq 0.05$: *; $p \leq 0.01$: **; $p \leq 0.001$: ***). Compound names with asterisk indicate an isobaric annotation match with ref 37. (Glc: Glucoside, GlcA: Glucuronopyranoside)