analytical chemistry

pubs.acs.org/ac Article

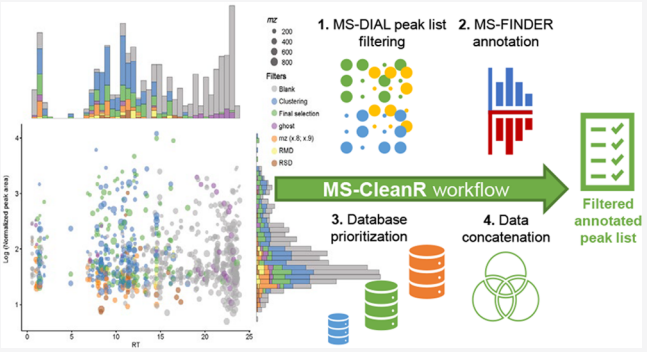
MS-CleanR: A Feature-Filtering Workflow for Untargeted LC−MS Based Metabolomics

Ophélie Fraisier-Vannier, Justine Chervin, Guillaume Cabanac, Virginie Puech, Sylvie Fournier, Virginie Durand, Aurélien Amiel, Olivier André, Omar Abdelaziz Benamar, Bernard Dumas, Hiroshi Tsugawa, and Guillaume Marti*

Cite This: *Anal. Chem.* 2020, 92, 9971−9981 Read Online

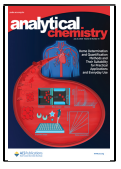
ACCESS | Metrics & More | Article Recommendations | Supporting Information

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. However, this approach still has many limitations; notably, the difficulty of accurately estimating the number of unique metabolites 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 according to the database chosen by the user, which enhance identification accuracy. Application of MS-CleanR to the analysis of *Arabidopsis thaliana* grown in three different conditions fostered class separation resulting from multivariate data analysis and led 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 euteiches*. A group of glycosylated triterpenoids overrepresented in resistant lines were identified as candidate compounds conferring pathogen resistance. MS-CleanR is implemented through a Shiny interface for intuitive use by end-users (available at https://github.com/eMetaboHUB/MS-CleanR).



Untargeted or discovery-based metabolomics have become an essential tool in all biological sciences including clinical research,[1,2] plant science,[3] and natural product mining,[4] among many other applications. Living organisms are estimated to contain more than one million distinct compounds.[5] According to the MetaboLights database (DB), 80% of untargeted metabolomics workflows rely on liquid chromatography−mass spectrometry (LC−MS) (https://www.ebi.ac.uk/metabolights/). Due to its broad coverage of metabolites, LC−MS based metabolomics has become the preferred tool to detect hundreds of compounds encountered in complex biological materials. Many software programs have been developed to turn features (m/z × retention time (RT) pairs) extracted from LC−MS raw data into chromatographic peak lists, including web-based interfaces such as XCMS,[6] Workflow4Metabolomics,[7] local GUI with MZmine,[8] and MS-DIAL.[9] Despite significant progress in feature extraction, it is challenging to accurately estimate the number of unique metabolites in a crude extract profiled by LC−MS.[10] On average, untargeted LC−MS yields hundred to thousands of signals, which may be attributed to either isotopes, contaminants, adducts, dimers, multimers, and heteromeric complexes or artifacts. The feature attribution processing which aims to decipher ion linkages is an essential step prior to metabolite annotation, which refers to tentative metabolite assignment to a given feature. Following these steps, the annotation rate can be calculated as the number of unique annotated metabolites over total features counts. Patti and colleagues[11] used the term "degenerate features" to describe feature relationships between multiple m/z signals arising from in-source phenomena and derived from the same metabolite. Their study demonstrated that feature inflation is highly underestimated in untargeted LC−MS based metabolomics. Additionally, this redundancy trend may have important consequences on metabolite annotation by increasing both false positive results and the number of "unknown" arising from wrongly attributed signals. This is especially true when

Received: April 13, 2020
Accepted: June 26, 2020
Published: June 26, 2020



https://dx.doi.org/10.1021/acs.analchem.0c01594
Anal. Chem. 2020, 92, 9971−9981

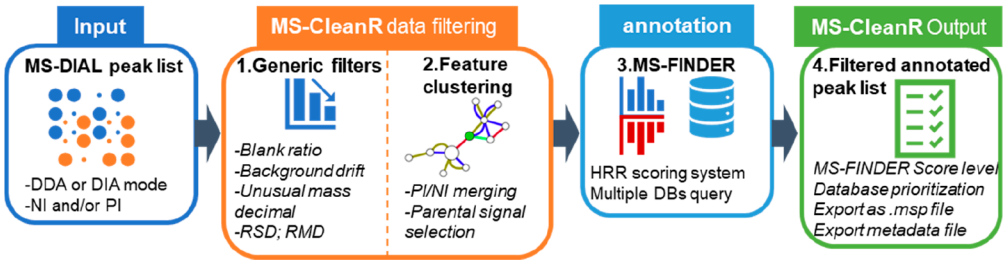


Figure 1. Schematic representation of MS-CleanR workflow.

the metabolite annotation process is based on *in silico* modeling of fragmentation patterns, as are Sirius,[12] MS-FINDER,[13] MetFrag,[14] or CFM-ID,[15] since tandem mass spectrometry (MS/MS) spectra are processed without taking into account feature relationships. As a result, most untargeted metabolomics studies focus on a subset of identified metabolites for which spectral data are easily accessible from public repositories or in-house DBs.

Feature attribution processing has been tackled by several packages using either correlation approach across multiple samples like MSClust,[16] RAMClust,[17] MS-FLO,[18] with additional MS/MS similarity filtering implemented in compMS2Miner,[19] or inferring feature attribution and annotation using metabolic pathway associations in xMSannotator.[20] Other approaches driven by peak shape similarity in a user defined retention time window are applicable to one single chromatogram. These include CAMERA[21] based on Pearson correlation, MetAssign[22] constructed through a Bayesian clustering approach to sort related m/z signals, and CliqueMS[23] which added an estimated adducts frequency to peak similarity for feature relationship deciphering. Within the all-in-one program MS-DIAL, a combination of approaches is gathered by peak character estimation algorithm[24] (MS-DIAL-PCE) providing the ion linkages based on (i) adducts mass differences calculation, (ii) unexpected in-source phenomena using correlated chromatograms within the sample, and (iii) putative ion source fragments candidates based on MS/MS similarity and analogous metabolite profiles among samples. MS-DIAL can handle one single chromatogram or a set of samples providing an aligned peak list (n samples × peak area of detected features) with putative ion linkages data. MS-FINDER is a partner program of MS-DIAL, in which unknown structures can be elucidated from MS/MS spectra by the hydrogen rearrangement rules-(HRR) based scoring system.[13]

Here, we describe a third tool interfaced between these two programs, called MS-CleanR, which aims to obtain an annotated and clean peak list by embedding the results from MS-DIAL and MS-FINDER. Starting from the aligned peak list determined by the MS-DIAL deconvolution process, our R package first removes noise signals by using several generic filters. In the second step, each feature is clustered to construct a graph based on the results of MS-DIAL-PCE and optionally extended by Pearson correlation links. At this point, each graph is submitted to the multilevel optimization of modularity algorithm[25] to extract the most probable parental mass. Optionally, clustered ion features can be merged between positive ionization (PI) and negative ionization (NI) modes and the adduct relationships are corrected accordingly. The cleaned-up feature list is then exported to MS-FINDER for annotation purposes. Finally, the package merges the MS-FINDER annotation output with the cleaned-up peak list and includes the possibility to prioritize identification according to the DBs used for MS-FINDER interrogation. The whole MS-CleanR workflow (Figure 1) is easily accessible through a shiny user interface, and it is available as open source code.

METHODS

Standards. Individual solutions of natural products (NPs) compounds (Metasci, Toronto, Canada) were prepared at 100 μg/mL in H_2O or MeOH according to the supplier's recommendations. We selected 51 NPs eluting from 2 to 18 min as a first test mixture to construct DB-level 1 annotation. The second test mixture was set up using 167 standards from the IROA Mass Spectrometry Library of Standards (Sigma-Aldrich, Darmstadt, Germany) (see Supplementary Text S1 for details).

Plant Material. *A. thaliana* (wild-type Col-O) were grown either in hydroponic culture, in plastic pots (high density), or in Jiffy pots. For each growing condition, 200 mg of plant material per sample were collected, placed in a FastPrep tube (MP Biomedicals Lysing Matrix D, Illkirch, France), and frozen in liquid nitrogen. For extraction, each sample was ground with a Mixer Mill MM 400 grinder (Retsh, Eragny sur Oise, France) by applying two cycles of 30 s at 30 m/s. Biphasic sample extraction was adapted from Salem et al., 2016.[26] Samples were filtered through 0.2 μm PTFE filters (Thermo Scientific) and transferred to vials. An extraction blank (without plant material) and a QC (Quality Control) sample (aliquot of all samples) were also prepared to validate the LC−MS profiles.

Germinated seedlings of *Medicago truncatula* were transferred onto M medium[24] and then placed in a growth chamber at 22 °C and 50% humidity with cycles of 16 h light−8 h dark for 14 days. The roots were ground with a Mixer Mill MM 400 grinder by applying two cycles of 30 s at 300 Hz. A total of 100 mg of ground tissue was placed in 2 mL FastPrep tubes containing 1.4 mm ceramic spheres (Lysing Matrix D) and extracted with 1 mL of acidified aqueous solution of methanol ($MeOH/H_2O/HCOOH$, 80:19:1). After two cycles of 20 s at 6 m/s in the FastPrep-24 (MP Biomedicals), the samples were centrifuged at 4 °C and 10 000 rpm for 10 min. The supernatants were transferred into vials. An extraction blank and quality control (QC) were also done for extraction and analytical validation (see Supplementary Text S1 for details).

UHPLC−HRMS Profiling. Ultrahigh-performance liquid chromatography−high-resolution MS (UHPLC−HRMS) analyses were performed on a Q Exactive Plus quadrupole mass spectrometer, equipped with a heated electrospray probe (HESI II) coupled to an U-HPLC Ultimate 3000 RSLC system (Thermo Fisher Scientific, Hemel Hempstead, U.K.). Samples were separated on a Luna Omega Polar C18 column (150 mm × 2.1 mm i.d., 1.6 μm, Phenomenex, Sartrouville,

https://dx.doi.org/10.1021/acs.analchem.0c01594
Anal. Chem. 2020, 92, 9971−9981

France) equipped with a guard column. Mobile phase A (MPA) was water with 0.05% formic acid (FA), and mobile phase B (MPB) was acetonitrile with 0.05% FA. The solvent gradient was 0 min, 100% MPA; 1 min, 100% MPA; 22 min, 100% MPB; 25 min, 100% MPB; 25.5 min, 100% MPA; 28 min, 100% MPA. The flow rate was 0.3 mL/min, and the column temperature was set to 40 °C; the autosampler temperature was set to 10 °C, and the injection volume was fixed to 2 μL for standard mixes and plant extracts. Mass detection was performed in positive ionization (PI) and negative ionization (NI) modes at 30 000 resolving power [full width at half-maximum (fwhm) at 400 m/z] for MS1 and 17 500 for MS2 with an automatic gain control (AGC) target of 1×10^6 for full scan MS1 and 1×10^5 for MS2. Ionization spray voltages were set to 3.5 kV (for PI) and 2.5 kV (for NI), and the capillary temperature was set to 256 °C for both modes. The mass scanning range was m/z 70−1050 for standards and m/z 100−1500 for plant extracts. Each full MS scan was followed by data-dependent acquisition of MS/MS spectra for the six most intense ions using stepped normalized collision energy of 20, 40, and 60 eV.

Data Processing. LC−MS data were first processed with MS-DIAL version 4.12. MS1, and MS2 tolerances were set to 0.01 and 0.05 Da, respectively, in centroid mode for each data set. Peaks were aligned on a quality control (QC) reference file with a RT tolerance of 0.1 min and a mass tolerance of 0.015 Da. Minimum peak height was set to 70% below the observed total ion chromatogram (TIC) baseline for a blank injection. MS-DIAL data was cleaned with MS-CleanR by selecting all filters with a minimum blank ratio set to 0.8, a maximum relative standard deviation (RSD) set to 30, and a relative mass defect (RMD) ranging from 50 to 3 000. The maximum mass difference for feature relationships detection was set to 0.005 Da, and the maximum RT difference was set to 0.025 min. The Pearson correlation links were considered only for biological data sets with correlation ≥0.8 and statistically significant α = 0.05. Two peaks were kept in each cluster: the most intense and the most connected. The kept features were annotated with MS-FINDER version 3.26. The MS1 and MS2 tolerances were set to 5 and 15 ppm, respectively. Formula finder were exclusively processed with C, H, O, N, P, and S atoms. DBs based on the genus and the family of the plant species (Supplementary Table S3, Supplementary Table S4, Supplementary Table S7, Supplementary Table S8) being investigated were constituted with the dictionary of natural product (DNP-CRC press, DNP on DVD v. 28.2) and the internal generic databases used were KNApSAcK, PlantCyc, HMDB, LipidMaps, NANPDB, and UNPD. Annotation prioritization was done by ranking genus DB followed by Family DB and then generic DB (internal DB from MS-FINDER). Statistical analysis and mass spectral similarity network were detailed in Supplementary Text S1.

RESULTS AND DISCUSSION

MS-CleanR Workflow and Implementation. MS-CleanR uses as input a MS-DIAL peak list processed in data dependent analysis (DDA) or data independent analysis (DIA) using either positive ionization mode (PI) or negative ionization mode (NI) or both. First, MS-CleanR applies generic filters encompassing blank injection signal subtraction starting from feature height ratio between QC and blank detected signals. A second generic filter, called "ghost blank peaks" is based on the high-background ion drift removal. Optionally, an unusual mass defect can be filtered out. A fourth generic filtering approach is the application of a "relative standard deviation" (RSD) threshold among sample classes. Finally, we introduced a fifth filter based on the "relative mass defect" (RMD) calculation. The RMD is calculated in ppm as [(mass defect/measured monoisotopic mass) × 10^6)]. It can be used to filter compound classes[27] and may remove signals with unexpected RMD values. All these filters are tunable by the user.

The second step involves a feature clustering method based on MS-DIAL peak character estimation algorithm (MS-DIAL-PCE), which aggregates several possible relationships at the same RT range: ion correlation among samples, MS/MS fragments in higher m/z, possible adducts and chromatogram correlations.[24] Optionally, Pearson's correlation between features located in the same RT window (typically 0.025 min) can be added during the clustering process. At this point, each feature is seen as a node within a graph. First, all neutral losses and dimers/heteromers are detected and taken out of consideration. Adduct attribution is based on MS-DIAL-PCE and extended by calculating m/z differences using the adduct list provided within the MS-CleanR package (tunable by the user). To filter out ambiguous adduct relationships, the frequency of apparition of each identified adduct was computed based on all adducts detected within the data set. Each adduct relationship is represented as an edge linking two nodes in our graph, and the weight is given as the product of its nodes' frequencies. The resulting graph allows one to discard misidentified relationships by keeping only the most probable adduct type for each feature. During this step, MS-CleanR can merge PI and NI mode if both acquisition modes are detected. Among each cluster, one to n features (tunable by the user) can be selected for further annotation: the most intense, the most connected, or both. During this study, we obtained the best results by keeping 2 features per cluster with selection based on both intensity and degree.

Then, all selected features are exported to MS-FINDER program for *in silico*-based annotation using hydrogen rearrangement rules (HRR) scoring system. During this step, multiple databases can be queried and each annotation result will be handled by MS-CleanR (see Supplementary Table S3 and Supplementary Table S4 for user-defined DB models).

The final step will merge annotation results with the filtered peak list by prioritizing database annotation depending on user choice. This latter function can greatly improve the annotation accuracy particularly when dealing with taxonomically defined extracts.[28] Optionally, all results can be exported as .msp files for mass spectral similarity networking purposes. For the detailed workflow, see Supplementary Text S1.

Workflow Benchmarking on Pooled Standards. To validate our approach, we benchmarked the MS-CleanR workflow by using a mixture of 51 NPs standards profiled in NI and PI modes with a reverse phase column and a 25 min gradient. The resulting data were compiled in an in-house DB comprising RT, HRMS, and MS/MS fragmentation patterns (DB-level 1 annotation according to the Metabolomics Standards Initiative-MSI[29]). To test whether the workflow retained features arising from unique metabolites and removed useless signals, we compared the results obtained by using a combination of MS-DIAL and MS-FINDER and DB-level 1 annotation to those obtained by using MS-CleanR. For the latter, we created another DB of the same metabolite set including accurate mass, molecular formula, and SMILES

https://dx.doi.org/10.1021/acs.analchem.0c01594
Anal. Chem. 2020, 92, 9971−9981

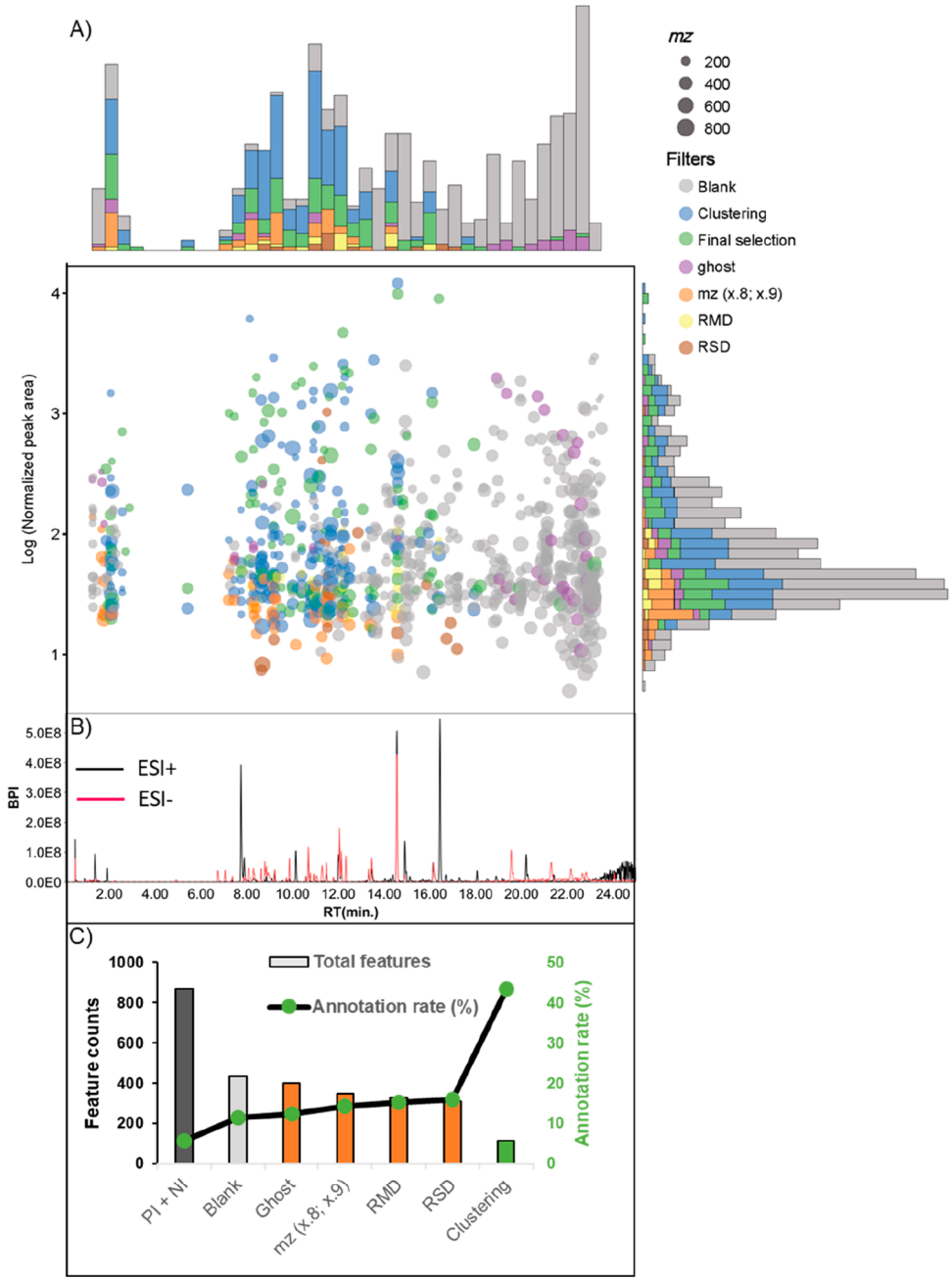


Figure 2. Feature filtering of the LC−MS data set from 51 NPs standards. Generic filters and the feature clustering algorithm were applied to the initial PI + NI mode data set. (A) Scatter plot of LC−MS features (x = RT in minute, y = log of normalized peak area, size = m/z value) colored according to filtering steps. Margin histogram plots display feature count along RT (x) and peak area (y). (B) Base peak chromatogram of LC−MS from 51 NPs standards compounds. (C) Bar plot displays feature counts after successive filters. The line plot displays annotation rate (unique metabolites/feature counts in %).

strings (DB-level 2 annotation according to the MSI) to reproduce real-case annotation processing. All generic filters were used, and the first two features within each cluster (n = 2), either most intense or most connected, were exported for annotation using the "formula prediction and structure elucidation by *in silico* fragmentation tool" in MS-FINDER (Supplementary Table S1).

The behavior of MS-CleanR filtering approach was benchmarked using a combination of both ionization modes acquired from a mixture of 51 NPs standards. The base peak intensity (BPI) chromatogram of the PI and NI modes displayed little overlap as well as a higher signal thickness between 6 and 18 min (Figure 2B). As shown in Figure 2A, the blank filtering process was the most effective after 16 min, within a sprinkling peak area. This result could be explained by lower competition for ion currents which induces higher signals for blank samples. The background ion drift signal filter (namely, "ghost") followed the behavior of blank filters. Unexpected mass signals (namely, "mz (x.8; x.9)") were mainly spread over low intensity signals, e.g., with m/z repeating signal at 158.964 (TFA contaminant) in PI or 112.982 in NI (formic acid dimer contaminant). This filter is particularly useful in

https://dx.doi.org/10.1021/acs.analchem.0c01594
Anal. Chem. 2020, 92, 9971−9981

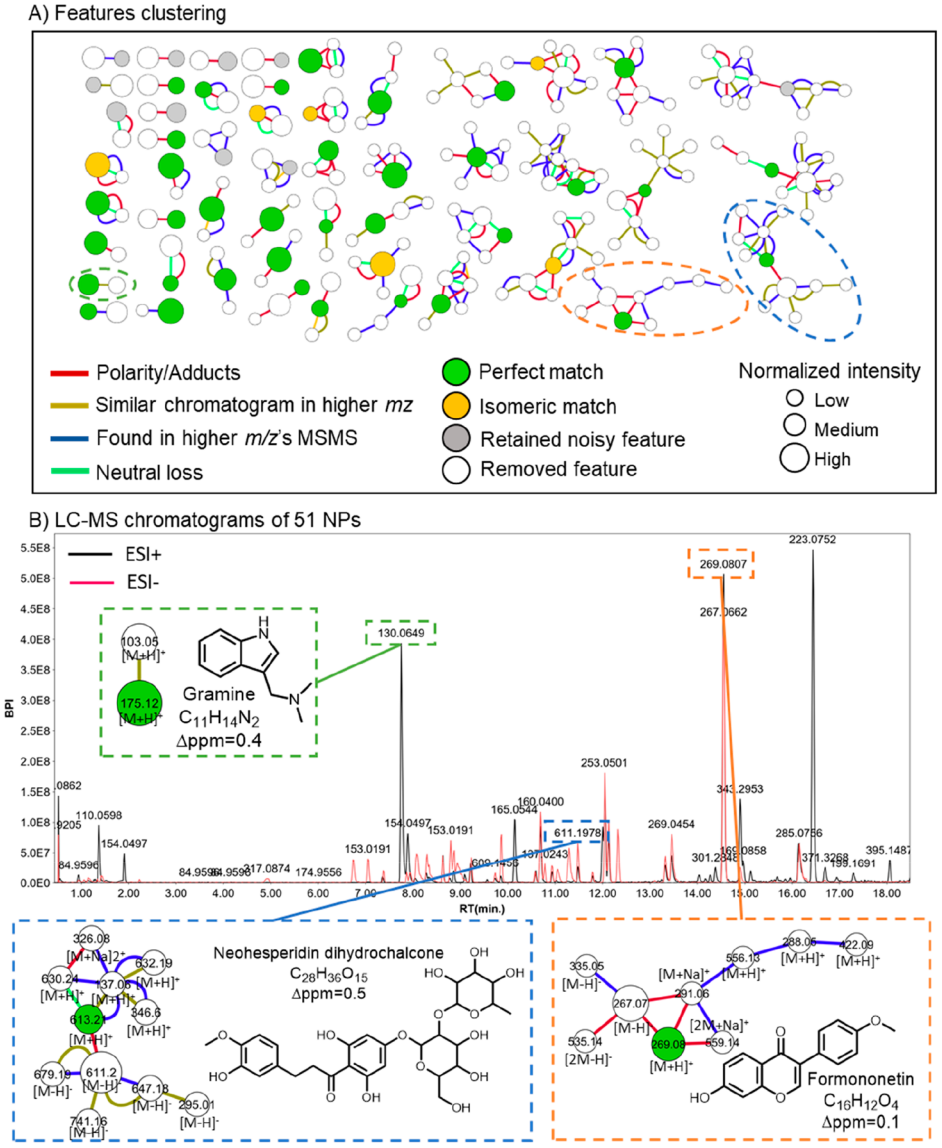


Figure 3. MS-CleanR feature clustering of 51 NPs. Clustering was based on the peak character estimation and multilevel optimization of modularity algorithms. (A) Cluster plot of the whole data set excluding size one clusters. (B) UHPLC−HRMS base peak intensity (BPI) chromatogram of the standard mixture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

signal-rich areas such as the injection peak observed between 1 and 2 min. The RMD filter was kept with default parameter of MS-CleanR which encompass most NPs mass defect values (between 50 and 3000 ppm). The RSD filter was also kept with a default value of 30%, which is generally the deviation value accepted between two consecutive injections of the same sample. Both RMD and RSD were spread between medium to low intensity peaks and removed a low number of signals. Interestingly, the final clustering-based feature selection displayed a repartition from low to high intensity signals all along the chromatogram. According to our observations, the parental feature arising from a unique metabolite signal among each cluster was highlighted by either a PI/NI adduct link in the case where both ionization modes were used (e.g., $[M + H]^+/[M - H]^-$, $[M + Na]^+/[M + FA - H]^-$); the most intense peak of the cluster and/or the peak with the most relationships to other features (i.e., the highest "degree" of connection). A selection based on the most intense and connected feature per cluster avoids the risk of retaining only the most intense signals of each cluster (as shown by the peak pattern at 5 min on Figure 2A). We advise combining both filtering options for optimal feature selection.

As anticipated, we observed a significant feature inflation in this mixture of 51 NPs standards: 869 signals from the PI and NI acquisition modes were detected (Figure 2C). This approximately 95% feature inflation is consistent with a previous report of 10 000−30 000 features detected after injection of 900 unique metabolites,[30] and with a study that used isotope labeling as a feature filtering approach.[11] Blank ratio filtering deleted 50% of the features, and the other generic filters described above removed 15% of the remaining features. Feature clustering caused a further reduction of 18%, resulting in a total of 115 features retained. Overall, the workflow filtered out 80% of all detected signals. By using this approach, there was a remarkable improvement in the annotation rate

https://dx.doi.org/10.1021/acs.analchem.0c01594
Anal. Chem. 2020, 92, 9971−9981

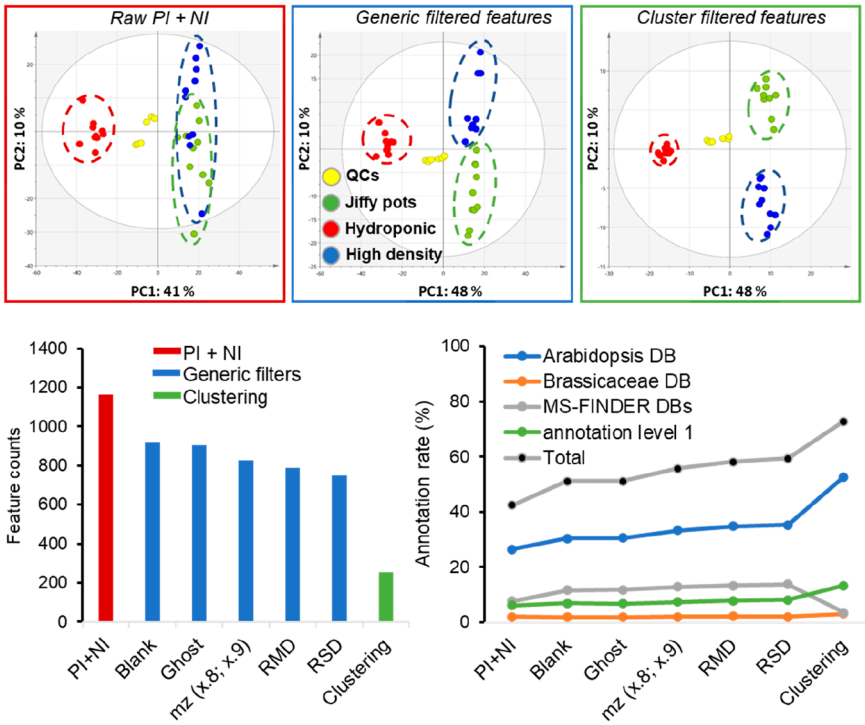


Figure 4. LC−MS data set processing of the metabolomes of *A. thaliana* plants growing in different conditions. Top: Sequential PCA score plots of raw PI and NI modes data set and after applying generic filters and feature clustering. Dotted circles indicate biological sample type distribution (yellow, 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.

(unique metabolites/detected features) from 5% to 45% (Figure 2C).

Consequently, 10 metabolites exhibited an isolated *mz*-RT signal whereas 41 metabolites displayed multiple signals grouped into clusters of the same RT range (±0.05 min, depending on MS-DIAL parameters used for RT windows) from 2 to 11 features (Figure 3A). The parental feature arising from a unique metabolite (green dots, Figure 3A) displayed either the most intense signal and/or highest degree of connection to other features among each cluster. It was also noticed that if a PI/NI linkage between two pseudomolecular ions was detected, the parental signal was one of them. The application of these rules allowed the annotation of 50 metabolites, 44 of which matched perfectly with level 1 annotation DB (Supplementary Table S1). The remaining ones were annotated as an isobaric/isomeric match due to prioritization of the highest MS-FINDER scoring value (e.g., 4-aminosalicylic acid and 5-aminosalicylic acid). In the case of gramine, for example, the major pseudomolecular ion had an m/z value of 130.0649 at RT 7.75 min (Figure 3B). By applying feature clustering, we detected an in-source fragment corresponding to the neutral loss of the dimethylamine group at m/z 130.0649. This feature was removed and only the signal at m/z 103.054 and m/z 175.1228 were retained. Since the signal at m/z 103.054 displayed a similar chromatogram to the most intense signal at m/z 175.1228, this last was exported and annotated as gramine (Δppm = 0.4) with a perfect match. The detected signals at 11.47 min; m/z 611.1983 $[M - H]^-$ and 11.48 min; m/z 613.2122 $[M + H]^+$ in NI and PI modes, respectively, were grouped in a cluster of 11 features. Feature relationships among this cluster were mainly related to similar MS/MS patterns among multiple features at the same RT range in the PI mode while similar chromatograms were mainly detected in NI mode (Figure 3B). In this case, feature selection was driven by the PI/NI adduct relationship and the feature with the highest MS-Finder annotation score was retained in the final peak list and identified as neohesperidin dihydrochalcone (Δppm = 0.4). Formononetin displayed a pseudomolecular ion $[M + H]^+$ in PI (RT = 14.55; m/z at 269.0806) and $[M - H]^-$ in NI (RT = 14.55; m/z at 267.0662) modes. The formononetin cluster encompassed complex adduct relationships in the NI mode and multiple similar MS/MS patterns in the PI mode. The detection of the most intense and connected features among this cluster added to the PI/NI relationship has conducted the selection of signals at m/z 269.0806 for annotation, which provided a perfect match with level 1 annotation DB. The only mismatch was encountered for phloridzin due to the neutral loss of a glucose moiety in both PI and NI modes. Only genine was detected in PI mode, resulting in selection of this signal in the final peak list.

To more closely model a real biological sample, we standardized our workflow by using a mixture of 167 standard compounds from the IROA Mass Spectrometry library (Supplementary Table S2). As above, we found significant feature inflation: 6732 signals after concatenation of PI and NI data sets (Figure S2). Unlike the standardization with NPs, above, the generic filters removed only 15% of features. The most important improvement was obtained by feature clustering, which filtered out 90% of the detected features leaving 611 signals. Among these, 127 features were identified as a perfect match compared to Level-1 annotation DB, and 21 were annotated as an isomeric match (Supplementary Table S2). Twelve features were removed due to their coelution with

https://dx.doi.org/10.1021/acs.analchem.0c01594
Anal. Chem. 2020, 92, 9971−9981

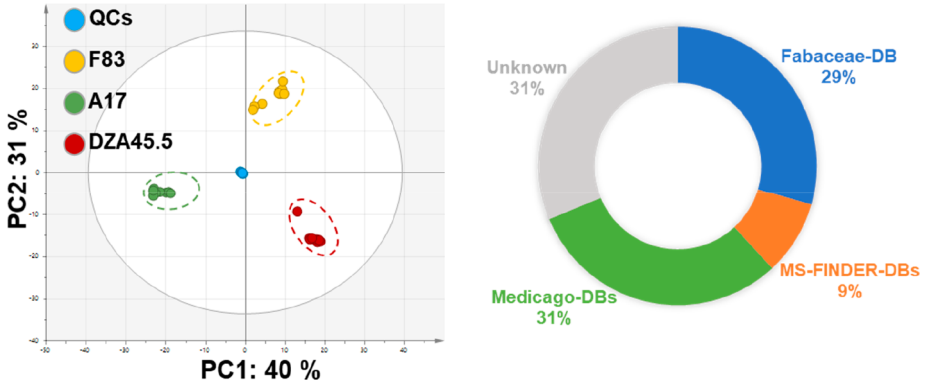


Figure 5. LC−MS NI data set processing of the metabolome of roots from three strains of *M. truncatula*. Left: PCA score plot after applying the MS-CleanR workflow. Dotted circles enclose samples from each plant strain. Right: Circular plot of the proportions of features annotated with reference to the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

other compounds, and four had a significant RT shift due to their poor peak shapes. The final three compounds were not annotated because of neutral loss of the same moiety in PI and NI modes, which led to their misidentification. Overall, the annotation rate with this workflow was 27% (Figure S2), and 90% of unique metabolites were retained.

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

Metabolic Profiling with MS-CleanR. Untargeted metabolomic profiling has emerged as the method of choice to identify metabolic markers associated with beneficial traits in plants, such as resistance to biotic stresses. In this context, the MS-CleanR workflow could greatly improve the results of untargeted metabolomics. To illustrate this point, we used as models the legume *Medicago truncatula* and the pathogenic oomycete *Aphanomyces euteiches*, a major pathogen of several legume species.[31] Genome-wide association studies of 179 lines of *M. truncatula* have identified major loci involved in the resistance of the plant to *A. euteiches*.[32] Moreover, genes encoding enzymes involved in the synthesis of antimicrobial metabolites are expressed in uninfected plants.[33] This suggests that antimicrobial metabolites in uninfected plants may be useful biomarkers with which to select legume lines resistant to *A. euteiches*. To identify these metabolites, we applied the MS-CleanR workflow to analyze the metabolomes of roots from three different strains of *M. truncatula* that have different levels of resistance to *A. euteiches* infection: strain DZA45.5 has the highest level of resistance, A17 is an intermediate level, and F83 is the most susceptible. These three strains were analyzed by LC−MS in NI mode and potential biomarkers were highlighted by multivariate data analysis (Supplementary Table S6). The metabolites that were differentially produced in the two most resistant strains (A17 and DZA45.5) when compared

https://dx.doi.org/10.1021/acs.analchem.0c01594
Anal. Chem. 2020, 92, 9971−9981

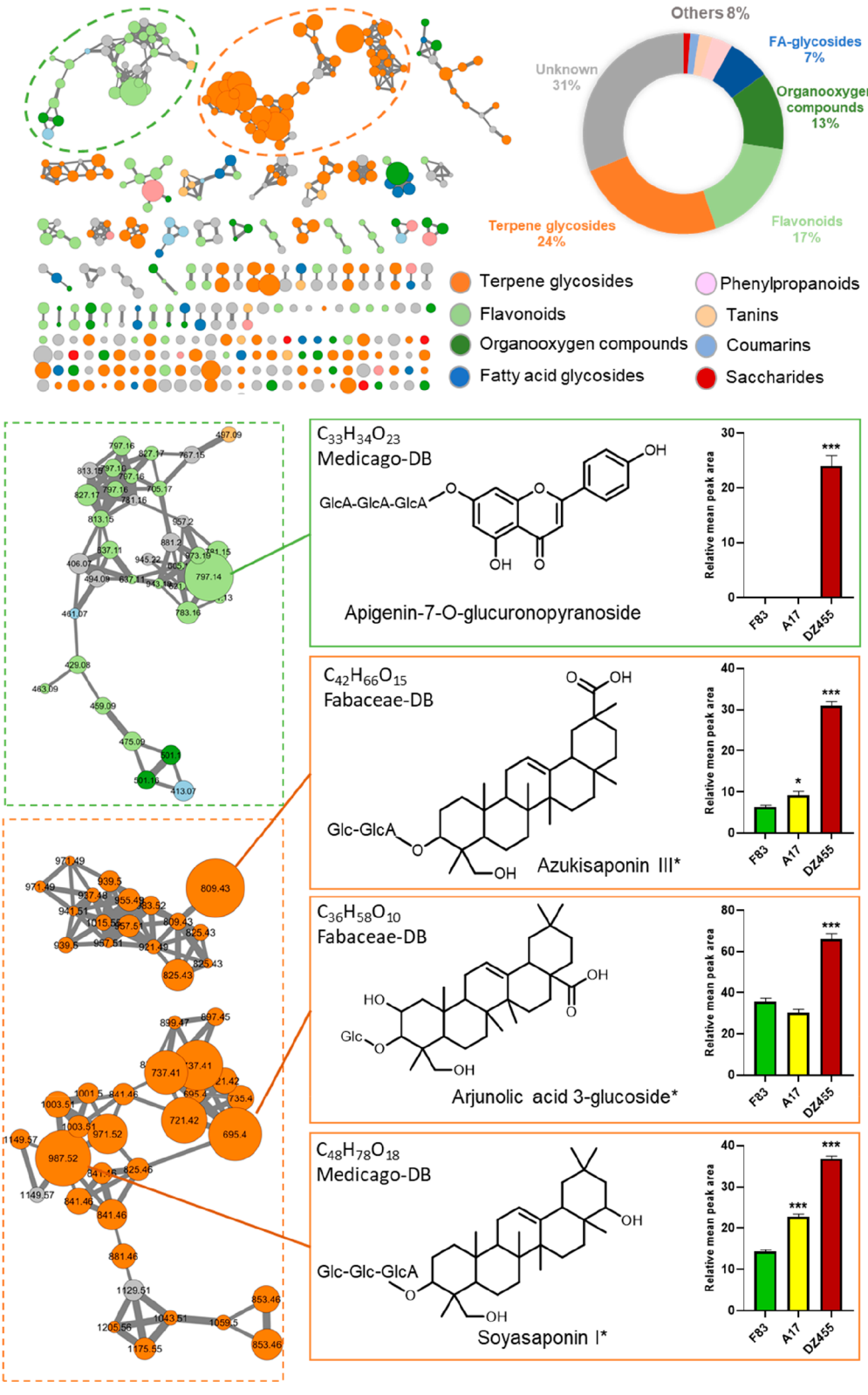


Figure 6. Mass spectral similarity network of *M. truncatula* NI data set (cosine ≥ 0.8). Nodes are colored according to their chemical classes and sized relative to their OPLS regression coefficient score (see text for details). The edge width is proportional to the cosine value. The pie chart displays the annotated chemical class ratio in the LC−MS NI data set (others include the coumarin derivatives, tannins, and saccharides chemical classes). Bar plots display normalized mean peak areas for the four most highly ranked structures by OPLS-regression modeling (Supplementary Table S6). One-way ANOVA and Dunnett's posthoc 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 an asterisk indicate an isobaric annotation match with ref 34 (Glc, glucoside; GlcA, glucuronopyranoside).

https://dx.doi.org/10.1021/acs.analchem.0c01594
Anal. Chem. 2020, 92, 9971−9981

to the more sensitive one (F83) were identified by OPLS regression.

After application of the MS-CleanR workflow, the PCA score plot showed a net clustering of the samples from each strain of *M. truncatula*. QC samples were centered on the PCA plot, demonstrating very good reproducibility (Figure 5). When annotated by reference to DBs from *Medicago* or the legume family Fabaceae, 60% of the data set was annotated (Figure 5) and an additional 9% with MS-FINDER DBs. A molecular spectral similarity network was built to highlight common chemical classes related to resistance traits (Figure 6). Among all annotated features, flavonoids and terpene glycosides compounds were prevalent. This latter class included primarily triterpene sapogenins which appeared to be highly correlated to the resistance traits according to the OPLS regression model. In particular, the four top ranked compounds belonged to two clusters related to sapogenins and one to flavonoids. Our untargeted approach revealed the presence of apigenin-7-*O*-glucuronopyranoside (best MS-FINDER score among several possible match in flavonoid class) only in the resistant DZ45.5 strain. This result corroborated a previous study by our group which demonstrated the implication of flavonoid pathways in resistance.[33] However, other detected flavonoids were not correlated to the resistance contrary to sapogenins class. Among the 151 terpene glycosides annotated in this study, 36 were also identified by a large-scale sapogenin profiling study in various ecotypes of *M. truncatula*[34] (Supplementary Table S6). Interestingly, the three-top ranked sapogenins by the OPLS model (azukisaponin III, arjunolic acid 3-glucoside, and soyasaponin I) displayed an isobaric match with two hederagenin glycosides and a bayogenin derivative, respectively, annotated by Sumner and colleagues. These sapogenins accumulate primarily in the roots rather than the leaves. These organs, however, have distinct profiles of specific saponins, which may be explained by the adaptation of each ecotype to its biotic environment. A previous study showed that saponins derived from hederagenin glycoside in *M. truncatula* have antifungal activity.[35] Our study confirmed a higher level of these compounds in the strains resistant to *A. euteiches* (DZA45.5 and A17) than in the sensitive strain F83. Although the relevance of saponins to resistance of *M. truncatula* to *A. euteiches* has not been confirmed, these findings demonstrate the potential value of applying metabolomics tools to identify biomarkers of plant resistance.

CONCLUSIONS

The main goal of LC−MS-based untargeted metabolomics is to convert chromatographic profiles of complex biological extracts into a comprehensive metabolite list. We demonstrate here that feature degeneracy has a great effect on the final annotated peak list, thus impacting biological knowledge mined from untargeted metabolomic studies. We estimate, based on analysis of standard mixtures, that feature inflation is close to 95%, in agreement with other studies.[11,30] Our package MS-CleanR, with its point-and-click software on a Shiny interface, is a new component in the suite of tools comprising the GUI software MS-DIAL and the annotation capabilities of MS-FINDER. Together, these provide a comprehensive workflow, from raw data to final annotated peak list. MS-CleanR can reduce the number of features by 80−90% and keep most unique metabolite signals without compromising the final data structure. Several parameters are tunable by the user, including the generic filters values and the selection method after feature clustering. The default values displayed in the MS-CleanR interface are a good starting point, but we advise careful inspection of the raw data to set the RSD value as well as the blank filtering threshold. The RMD window can be tuned based on the matrices component knowledge under study. The addition of all generic filters and Pearson correlation clustering extension will provide sharp data set filtering, thus focusing on high quality peaks. Conversely, a wider filtered data set could be obtained by removing some or all of the filters. (Additional tips can be found in the tutorials available at https://github.com/eMetaboHUB/MS-CleanR). The package is also able to combine the PI and NI modes (*A. thaliana* experiment) or to treat only one mode independently (*M. truncatula* study) depending on the study objectives. The opportunity to rank the annotation results with reference to user-defined databases narrows down the final identification possibilities. Each MS-CleanR step takes from a few seconds to a few minutes, even for the clustering and feature selection stage which use a computationally fast heuristic approach.[25] Another advantage of feature filtering prior to MS-FINDER annotation is the drastic reduction of processing time which is divided by several orders of magnitude. We demonstrate the utility of this workflow by analyzing secondary metabolites levels in three *M. truncatula* strains with different susceptibilities to a pathogenic oomycete. We annotated 70% of the data set with 60% at the genus or family level using DBs prioritization. The resulting mass spectral similarity network further supports annotation results as most clusters gathered the same metabolite chemical class. Still, our approach was unable to keep only unique metabolite features regarding the annotation rate comprising between 24 and 45% for standard mixtures. A limitation of our filtering process is its dependence on chromatographic resolution, which can seriously impair the final results by clustering several unique metabolites together. In the present study, we chose a 20 min gradient, like those generally applied in most untargeted metabolomics studies. Extending the elution time might improve the chromatographic resolution but is difficult to apply in day-to-day work, especially for high-throughput experiments. These challenges will be addressed in future developments of MS-CleanR.

ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge at https://pubs.acs.org/doi/10.1021/acs.analchem.0c01594.

Supplementary Text S1, detailed workflow of MS-CleanR package and materials used in this study (PDF)

Figure S1, alignment spot screenshot in ESI PI and NI ionization modes showing repeated blank pseudomolecular ions detected in QCs samples with a retention time shift; Figure S2, feature filtering of LC−MS data set from 167 IROA-MS standards library according to generic filters and clustering algorithm (PDF)

Supplementary Table S1, Excel table with cluster annotation, result summary and database for level 2 annotation imported in MS-FINDER for the 51 NPs data set (XLSX)

Supplementary Table S2, Excel table with cluster annotation, result summary and database for level 2

https://dx.doi.org/10.1021/acs.analchem.0c01594
Anal. Chem. 2020, 92, 9971−9981

annotation imported in MS-FINDER for the 167 IROA MS standard library data set (XLSX)

Supplementary Table S3, *Arabidopsis* DB used as input for level 2 annotation in MS-FINDER (ZIP)

Supplementary Table S4, *Brassicaceae* DB used as input for level 2 annotation in MS-FINDER (ZIP)

Supplementary Table S5, *Arabidopsis* data set treated by MS-CleanR and OPLS-DA top ranked features (XLSX)

Supplementary Table S6, *Medicago* data set treated by MS-CleanR and OPLS regression coefficient for feature ranking (XLSX)

Supplementary Table S7, *Medicago* DB used as input for level 2 annotation in MS-FINDER (ZIP)

Supplementary Table S8, *Fabaceae* DB used as input for level 2 annotation in MS-FINDER (ZIP)

AUTHOR INFORMATION

Corresponding Author

Guillaume Marti − *Pharma Dev, Université de Toulouse, IRD, UPS, 31400 Toulouse, France; Laboratoire de Recherche en Sciences Végétales and Metatoul-AgromiX Platform, MetaboHUB, National Infrastructure for Metabolomics and Fluxomics, LRSV, Université de Toulouse, CNRS, UPS, 31400 Toulouse, France; Institut de Recherche en Informatique de Toulouse, Université de Toulouse, UPS, Toulouse 31400, France;* orcid.org/0000-0002-6321-9005; Phone: (+33) 534 32 38 31; Email: guillaume.marti@univ-tlse3.fr

Authors

Ophélie Fraisier-Vannier − *Pharma Dev, Université de Toulouse, IRD, UPS, 31400 Toulouse, France; Institut de Recherche en Informatique de Toulouse, Université de Toulouse, UPS, Toulouse 31400, France*

Justine Chervin − *Laboratoire de Recherche en Sciences Végétales and Metatoul-AgromiX Platform, MetaboHUB, National Infrastructure for Metabolomics and Fluxomics, LRSV, Université de Toulouse, CNRS, UPS, 31400 Toulouse, France*

Guillaume Cabanac − *Institut de Recherche en Informatique de Toulouse, Université de Toulouse, UPS, Toulouse 31400, France;* orcid.org/0000-0003-3060-6241

Virginie Puech − *Laboratoire de Recherche en Sciences Végétales and Metatoul-AgromiX Platform, MetaboHUB, National Infrastructure for Metabolomics and Fluxomics, LRSV, Université de Toulouse, CNRS, UPS, 31400 Toulouse, France*

Sylvie Fournier − *Laboratoire de Recherche en Sciences Végétales and Metatoul-AgromiX Platform, MetaboHUB, National Infrastructure for Metabolomics and Fluxomics, LRSV, Université de Toulouse, CNRS, UPS, 31400 Toulouse, France*

Virginie Durand − *Laboratoire de Recherche en Sciences Végétales, Université de Toulouse, CNRS, UPS, 31400 Toulouse, France*

Aurélien Amiel − *Laboratoire de Recherche en Sciences Végétales, Université de Toulouse, CNRS, UPS, 31400 Toulouse, France; De Sangosse, Bonnel, 47480 Pont-Du-Casse, France*

Olivier André − *Laboratoire de Recherche en Sciences Végétales, Université de Toulouse, CNRS, UPS, 31400 Toulouse, France; De Sangosse, Bonnel, 47480 Pont-Du-Casse, France*

Omar Abdelaziz Benamar − *Laboratoire de Recherche en Sciences Végétales, Université de Toulouse, CNRS, UPS, 31400 Toulouse, France*

Bernard Dumas − *Laboratoire de Recherche en Sciences Végétales, Université de Toulouse, CNRS, UPS, 31400 Toulouse, France*

Hiroshi Tsugawa − *RIKEN Center for Sustainable Resource Science, Yokohama 230-0045, Japan; RIKEN Center for Integrative Medical Science, Yokohama 230-0045, Japan;* orcid.org/0000-0002-2015-3958

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.analchem.0c01594

Notes

The authors declare no competing financial interest.

Raw data from *Arabidopsis* and *Medicago* LC−MS profiling available on Zenodo.org using the DOI 10.5281/zenodo.3744480.

ACKNOWLEDGMENTS

We thank Dr. Stephane Bertani for providing us the standards from the IROA-MS Library. Financial support was received from the French National Infrastructure for Metabolomics and Fluxomics, Grant MetaboHUB-ANR-11-INBS-0010, and the PSPC SOLSTICE Project (SOLutionS pour des Traitements Intégrés dans une Conduite Environnementale) managed by Belchim Crop Protection and partly funded by the French state within the framework of the Programme d'Investissements d'Avenir. We thank E. Amblard, N. Jariais, and C. Jacquet for the *M. truncatula* cultures and A. Haouy for the *A. thalina* cultures and sample preparations. We also acknowledge Carol Featherstone of Plume Scientific Communication Services for the professional scientific editing.

REFERENCES

(1) Zierer, J.; Jackson, M. A.; Kastenmüller, G.; Mangino, M.; Long, T.; Telenti, A.; Mohney, R. P.; Small, K. S.; Bell, J. T.; Steves, C. J.; Valdes, A. M.; Spector, T. D.; Menni, C. *Nat. Genet.* **2018**, *50* (6), 790−795.

(2) Li, H.; Ning, S.; Ghandi, M.; Kryukov, G. V.; Gopal, S.; Deik, A.; Souza, A.; Pierce, K.; Keskula, P.; Hernandez, D.; Ann, J.; Shkoza, D.; Apfel, V.; Zou, Y.; Vazquez, F.; Barretina, J.; Pagliarini, R. A.; Galli, G. G.; Root, D. E.; Hahn, W. C.; Tsherniak, A.; Giannakis, M.; Schreiber, S. L.; Clish, C. B.; Garraway, L. A.; Sellers, W. R. *Nat. Med.* **2019**, *25* (5), 850−860.

(3) Gargallo-Garriga, A.; Sardans, J.; Pérez-Trujillo, M.; Oravec, M.; Urban, O.; Jentsch, A.; Kreyling, J.; Beierkuhnlein, C.; Parella, T.; Peñuelas, J. *New Phytol.* **2015**, *207* (3), 591−603.

(4) Mohimani, H.; Gurevich, A.; Shlemov, A.; Mikheenko, A.; Korobeynikov, A.; Cao, L.; Shcherbin, E.; Nothias, L.-F.; Dorrestein, P. C.; Pevzner, P. A. *Nat. Commun.* **2018**, *9*, 4035 DOI: 10.1038/s41467-018-06082-8.

(5) Wang, S.; Alseekh, S.; Fernie, A. R.; Luo, J. *Mol. Plant* **2019**, *12* (7), 899−919.

(6) Huan, T.; Forsberg, E. M.; Rinehart, D.; Johnson, C. H.; Ivanisevic, J.; Benton, H. P.; Fang, M.; Aisporna, A.; Hilmers, B.; Poole, F. L.; Thorgersen, M. P.; Adams, M. W. W.; Krantz, G.; Fields, M. W.; Robbins, P. D.; Niedernhofer, L. J.; Ideker, T.; Majumder, E. L.; Wall, J. D.; Rattray, N. J. W.; Goodacre, R.; Lairson, L. L.; Siuzdak, G. *Nat. Methods* **2017**, *14* (5), 461−462.

(7) Giacomoni, F.; Le Corguille, G.; Monsoor, M.; Landi, M.; Pericard, P.; Petera, M.; Duperier, C.; Tremblay-Franco, M.; Martin, J.-F.; Jacob, D.; Goulitquer, S.; Thevenot, E. A.; Caron, C. *Bioinformatics* **2015**, *31* (9), 1493−1495.

(8) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Orešič, M. *BMC Bioinf.* **2010**, *11* (1), 395.

(9) Tsugawa, H.; Cajka, T.; Kind, T.; Ma, Y.; Higgins, B.; Ikeda, K.; Kanazawa, M.; VanderGheynst, J.; Fiehn, O.; Arita, M. *Nat. Methods* **2015**, *12* (6), 523−526.
(10) Patti, G. J.; Yanes, O.; Siuzdak, G. *Nat. Rev. Mol. Cell Biol.* **2012**, *13* (4), 263−269.
(11) Mahieu, N. G.; Patti, G. J. *Anal. Chem.* **2017**, *89* (19), 10397−10406.
(12) Dührkop, K.; Fleischauer, M.; Ludwig, M.; Aksenov, A. A.; Melnik, A. V.; Meusel, M.; Dorrestein, P. C.; Rousu, J.; Böcker, S. *Nat. Methods* **2019**, *16* (4), 299−302.
(13) Tsugawa, H.; Kind, T.; Nakabayashi, R.; Yukihira, D.; Tanaka, W.; Cajka, T.; Saito, K.; Fiehn, O.; Arita, M. *Anal. Chem.* **2016**, *88* (16), 7946−7958.
(14) Ruttkies, C.; Schymanski, E. L.; Wolf, S.; Hollender, J.; Neumann, S. *J. Cheminf.* **2016**, *8* (1), 3.
(15) Djoumbou-Feunang, Y.; Pon, A.; Karu, N.; Zheng, J.; Li, C.; Arndt, D.; Gautam, M.; Allen, F.; Wishart, D. S. *Metabolites* **2019**, *9* (4), 72.
(16) Tikunov, Y. M.; Laptenok, S.; Hall, R. D.; Bovy, A.; de Vos, R. C. H. *Metabolomics* **2012**, *8* (4), 714−718.
(17) Broeckling, C. D.; Afsar, F. A.; Neumann, S.; Ben-Hur, A.; Prenni, J. E. *Anal. Chem.* **2014**, *86* (14), 6812−6817.
(18) DeFelice, B. C.; Mehta, S. S.; Samra, S.; Čajka, T.; Wancewicz, B.; Fahrmann, J. F.; Fiehn, O. *Anal. Chem.* **2017**, *89* (6), 3250−3255.
(19) Edmands, W. M. B.; Petrick, L. M.; Barupal, D. K.; Scalbert, A.; Wilson, M.; Wickliffe, J.; Rappaport, S. M. *Anal. Chem.* **2017**, *89*, 3919.
(20) Uppal, K.; Walker, D. I.; Jones, D. P. *Anal. Chem.* **2017**, *89* (2), 1063−1067.
(21) Kuhl, C.; Tautenhahn, R.; Böttcher, C.; Larson, T. R.; Neumann, S. *Anal. Chem.* **2012**, *84* (1), 283−289.
(22) Daly, R.; Rogers, S.; Wandy, J.; Jankevics, A.; Burgess, K. E. V.; Breitling, R. *Bioinformatics* **2014**, *30* (19), 2764−2771.
(23) Senan, O.; Aguilar-Mogas, A.; Navarro, M.; Capellades, J.; Noon, L.; Burks, D.; Yanes, O.; Guimerà, R.; Sales-Pardo, M. *Bioinformatics* **2019**, *35* (20), 4089−4097.
(24) Tsugawa, H.; Nakabayashi, R.; Mori, T.; Yamada, Y.; Takahashi, M.; Rai, A.; Sugiyama, R.; Yamamoto, H.; Nakaya, T.; Yamazaki, M.; Kooke, R.; Bac-Molenaar, J. A.; Oztolan-Erol, N.; Keurentjes, J. J. B.; Arita, M.; Saito, K. *Nat. Methods* **2019**, *16* (4), 295−298.
(25) Blondel, V. D.; Guillaume, J.-L.; Lambiotte, R.; Lefebvre, E. *J. Stat. Mech.: Theory Exp.* **2008**, *2008* (10), P10008.
(26) Salem, M. A.; Jüppner, J.; Bajdzienko, K.; Giavalisco, P. *Plant Methods* **2016**, *12*, 45 DOI: 10.1186/s13007-016-0146-2.
(27) Ekanayaka, E. A. P.; Celiz, M. D.; Jones, A. D. *Plant Physiol.* **2015**, *167* (4), 1221−1232.
(28) Rutz, A.; Dounoue-Kubo, M.; Ollivier, S.; Bisson, J.; Bagheri, M.; Saesong, T.; Ebrahimi, S. N.; Ingkaninan, K.; Wolfender, J.-L.; Allard, P.-M. *Front. Plant Sci.* **2019**, *10*, 1329.
(29) Creek, D. J.; Dunn, W. B.; Fiehn, O.; Griffin, J. L.; Hall, R. D.; Lei, Z.; Mistrik, R.; Neumann, S.; Schymanski, E. L.; Sumner, L. W.; Trengove, R.; Wolfender, J.-L. *Metabolomics* **2014**, *10* (3), 350−353.
(30) Li, Z.; Lu, Y.; Guo, Y.; Cao, H.; Wang, Q.; Shui, W. *Anal. Chim. Acta* **2018**, *1029*, 50−57.
(31) Gaulin, E.; Jacquet, C.; Bottin, A.; Dumas, B. *Mol. Plant Pathol.* **2007**, *8* (5), 539−548.
(32) Bonhomme, M.; André, O.; Badis, Y.; Ronfort, J.; Burgarella, C.; Chantret, N.; Prosperi, J.-M.; Briskine, R.; Mudge, J.; Debéllé, F.; Navier, H.; Miteul, H.; Hajri, A.; Baranger, A.; Tiffin, P.; Dumas, B.; Pilet-Nayel, M.-L.; Young, N. D.; Jacquet, C. *New Phytol.* **2014**, *201* (4), 1328−1342.
(33) Badis, Y.; Bonhomme, M.; Lafitte, C.; Huguet, S.; Balzergue, S.; Dumas, B.; Jacquet, C. *Mol. Plant Pathol.* **2015**, *16* (9), 973−986.
(34) Lei, Z.; Watson, B. S.; Huhman, D.; Yang, D. S.; Sumner, L. W. *Front. Plant Sci.* **2019**, *10*, 850 DOI: 10.3389/fpls.2019.00850.
(35) Abbruscato, P.; Tosi, S.; Crispino, L.; Biazzi, E.; Menin, B.; Picco, A. M.; Pecetti, L.; Avato, P.; Tava, A. *J. Agric. Food Chem.* **2014**, *62* (46), 11030−11036.

https://dx.doi.org/10.1021/acs.analchem.0c01594
Anal. Chem. 2020, 92, 9971−9981