Supporting Material
CliqueMS: A computational tool for annotating in-source metabolite ions from LC-MS untargeted metabolomics data based on a coelution similarity network

Oriol Senan, Antoni Aguilar-Mogas, Miriam Navarro, Jordi Capellades, Luke Noon, Deborah Burks, Oscar Yanes Roger Guimerà, and Marta Sales-Pardo

Table of contents

S1 Spectral data acquisition

Mixture of standards

Materials. MS grade acetonitrile (ACN), ammonium acetate (NH4Ac) and NH4OH were purchased from SDS (Peypin, France). Water was produced in an in-house Milli-Q purification system (Millipore, Molsheim, France). Formic acid was purchased from Sigma-Aldrich (Steinheim, Germany). Standards: (-)riboflavine, 1,2-distearoyl-sn-glycero-3-phosphocholine, biotin, cholic acid, deoxycholic acid, L-methionine sulfoxide, thymine and uracil were purchased from Sigma-Aldrich (Steinheim, Germany).

Standard mix preparation. All standards were pooled to a final concentration of 1ppm in H2O:ACN (5:95) with 0.1% formic acid.

MS1 analyses. MS1 analyses were performed using an UHPLC system (1290 series, Agilent Technologies) coupled to a 6550 ESI-QTOF MS (Agilent Technologies) operated in positive (ESI+) electrospray ionization mode. A vial containing the standard mix was kept at -20^{o}C prior to MS1 analysis. Metabolites were separated using an Acquity UPLC BEH HILIC column (2.1 x 150 mm, 1.8 μm) and the solvent system was A1 = 20mM ammonium acetate and 15 mM NH4OH in water and B1 = 95% ACN and 5% H2O. The linear gradient elution started at 100% B (time 0–2 min) and finished at 75% A (10-15 min). The injection volume was 5 μL. ESI conditions: gas temperature, 150^{o}C; drying gas, 13 L min–1; nebulizer, 35 psig; fragmentor, 400 V; and skimmer, 65 V. The instrument was set to acquire over the m/z range 100–1500 in full-scan mode with an acquisition rate of 4 spectra/sec. MS/MS was performed in targeted mode, and the instrument was set to acquire over the m/z range 50–1000, with a default isolation width (the width half-maximum of the quadrupole mass bandpass used during MS/MS precursor isolation) of 4 m/z. The collision energy was fixed at 20 V (see Supplementary Table S8 for the raw data).

Retina IRS2 KO Samples

Materials. MS1 grade methanol (MeOH) and acetonitrile (ACN) and analytical grade chloroform (CHCl3) were purchased from SDS (Peypin, France). Water was produced in an in-house Milli-Q purification system (Millipore, Molsheim, France). Formic acid and ammonium fluoride were purchased from Sigma-Aldrich (Steinheim, Germany).

Animal model. Irs-2-deficient mice were generated initially on a C57BL6/J: SV129 background [**?**] and then backcrossed to establish a pure C57BL6/J background [**?**]. Thus, the offspring resulting from the breeding of Irs-2(2/2) with RIP-Irs-2 line were C57BL6/J. The generation and genotyping of the Irs-2(2/2) and the RIP-Irs-2(2/2) models have been described previously [**?**, **?**].

Metabolite extraction method. Retinas were first lyophilized and metabolites were extracted adding 190 uL of MeOH and 120 uL of H2O, then vortex during 30 seconds. Afterwards, samples were frozen during 1 min in liquid nitrogen (N2) and thawed by cold sonication during 30 seconds. This step was applied three times. Then 380 uL of chloroform were added and vortexed during 30 seconds. Finally, samples were centrifuged (15000 rpm, 15 min a 4^oC). The aqueous phase was extracted and dried. The sample was suspended in 100 uL of H2O:MeOH (1:1) and stored at -80 ^{o}C until further analysis.

LC-MS1 analyses. LC-MS1 analyses were performed using an UHPLC system (1290 series, Agilent Technologies) coupled to a 6550 ESI-QTOF MS (Agilent Technologies) operated in positive (ESI+) or negative (ESI-) electrospray ionization mode. Vials containing extracted metabolites were kept at -20^oC prior to MS1 analysis. When the instrument was operated in positive ionization mode, metabolites were separated using an Acquity UPLC (HSS T3) C18 reverse phase (RP) column (2.1 x 150 mm, 1.8 μm) and the solvent system was A1 = 0.1% formic acid in water and B1 = 0.1% formic acid in acetonitrile. When the instrument was operated in negative ionization mode, metabolites were separated using an Acquity UPLC (BEH) C18 RP column (2.1 x 150 mm, 1.7 μm) and the solvent system was A2 = 1 mM ammonium fluoride in water and B2 = acetonitrile, as previously reported [**?**]. The linear gradient elution started at 100% A (time 0–2 min) and finished at 100% B (10-15 min). The injection volume was 5 μL. ESI conditions: gas temperature, 150^oC; drying gas, 13 L min–1; nebulizer, 35 psig; fragmentor, 400 V; and skimmer, 65 V. The instrument was set to acquire over the m/z range 100–1500 in full-scan mode with an acquisition rate of 4 spectra/sec. MS/MS was performed in targeted mode, and the instrument was set to acquire over the m/z range 50–1000, with a default isolation width (the width half-maximum of the quadrupole mass bandpass used during MS/MS precursor isolation) of 4 m/z. The collision energy was fixed at 20 V.

S2 Workflow of CliqueMS

Here, we explain the basic use of CliqueMS: its input, its main functions and how to interpret the program output. The nine standards sample will serve as an example for showing the R code. CliqueMS analyzes one sample at a time.

Installation. To install CliqueMS, the two recommended options are either installation from CRAN using ’install.packages(’cliqueMS’)’, or installation from the GitHub repository using ’devtools’ package function ’install_github(’osenan/cliqueMS’)’. In both cases, MacOS users need to have R compiling tools installed (`https://cran.r-project.org/bin/macosx/tools/`). MacOS users that do not have it available, may as well download a precompiled binary version from our GitHub repository (`https://github.com/osenan/cliqueMS/MacOS`). Here are the detailed installation instructions:

1. From a R console, first check that the required CRAN packages for CliqueMS are installed, and if not install them.

```r
if (!requireNamespace("igraph", quietly = TRUE))
   install.packages("igraph")
if (!requireNamespace("qlcMatrix", quietly = TRUE))
   install.packages("qlcMatrix")
if (!requireNamespace("Matrix", quietly = TRUE))
   install.packages("Matrix")
if (!requireNamespace("Rcpp", quietly = TRUE))
   install.packages("Rcpp")
if (!requireNamespace("RcppArmadillo", quietly = TRUE))
   install.packages("RcppArmadillo")
if (!requireNamespace("BH", quietly = TRUE))
   install.packages("BH")
```

2. Install the required packages from the Biconductor repository.

```r
if (!requireNamespace("BiocManager", quietly = TRUE))
   install.packages("BiocManager")
if (!requireNamespace("MSnbase", quietly = TRUE))
   install("MSnbase", version = 3.8)
if (!requireNamespace("xcms", quietly = TRUE))
   install("xcms", version = 3.8)
if (!requireNamespace("CAMERA", quietly = TRUE))
   install("CAMERA", version = 3.8)
```

3. After the installation of the required packages, the MacOS precompiled binary package can be installed (please, check that your path to the downloaded file is correct).

```
install.packages('pathtofile/cliqueMS_0.2.3.tgz', repo = NULL)
```

Input. CliqueMS uses a 'xcmsSet-class' or 'XCMSnExp' object as input data. To generate this type of objects, the user needs to use the R package 'xcms' to process the raw LC-MS1 data (any of the available algorithms of 'xcms'can be used in this step). Raw data can be in any format valid for xcms, 'NetCDF', 'mzData', 'mzXML' or 'mzML'. The code in R looks like this:

```
mzfile <- system.file("standards.mzXML", package = "cliqueMS")
msSet <- xcms::xcmsSet(files = mzfile, method = "centWave",
ppm = 15, peakwidth = c(5,20), snthresh = 10)

## Detecting mass traces at 15 ppm ...  OK
## Detecting chromatographic peaks in 2385 regions of interest ...
OK: 276 found.
```

Note that we use 'centWave' as the peak peaking algorithm within xcms that we use to define the vector of intensities associated to each feature (see main text).

Usage Step a - Feature group(clique) identification. The first step is to call the function 'getCliques', which will first generate an 'anClique-class' object, and then it will compute the clique groups. An 'anClique-class' object is a S3 R object that contains several attributes that will store the clique groups, the isotope annotation and the adduct annotation.

Figure S1a shows the details of 'getCliques'. First, it generates an 'anClique-class' object from the 'xcmsSet-class' object. Then, it stores a feature list in the 'anClique-class' object. The feature list is an R 'data.frame'. In this data.frame, each row corresponds to a feature with m/z, retention time, intensity and other values of each feature in the processed LC-MS1 data. Then, if the user sets the 'filter' parameter to 'TRUE' (see next sectionS2 for more details on parameters), features that are almost identical are removed and only one of these two equal features is kept. Most of the times, these identical features are artifacts from processing the raw data. The next step of 'getCliques' is to call the 'createNetwork' function to compute the cosine similarity between features and create the similarity network. This network is also stored in the 'anClique-class' object. The last step of 'getCliques' is to call the function 'computeCliques' and compute the clique groups that maximise the log-likelihood of the probabilistic model. The groups are stored in the 'anClique-class' object.

```
library(cliqueMS)
set.seed(2)
ex.cliqueGroups <- getCliques(msSet, filter = TRUE)

## Creating anClique object
## Creating network
## Features filtered: 1
## Computing cliques
## Beggining value of logl is -1114.05
## Aggregate cliques done, with 386 rounds
## Kernighan-Lin done with 3 rounds
## Finishing value of logl is -253.498
```

Usage Step b - Isotope annotation.

Once the clique groups have been computed, the next step is to obtain the isotope annotation. The user must call the 'getIsotopes' function to perform this task. Figure S1b shows that the first step of this function is to split the feature list into clique groups. Then, within the features of a group, the function 'returnIsotopes' searches for pairs of features that can be isotopes. Features that correspond to isotopic variants of the same metabolite must follow an intensity pattern: the monoisotopic feature has to be more abundant than its isotope. The isotope feature should have a mass value bigger than the monoisotopic feature in the range of the relative error specified by the user. The mass difference corresponds to 1.003355 Da for singly charge ions. Note that because the information we have at this stage is m/z (z being the charge), when scanning for differences in m/z compatible with that of isotopic variants, we have to take into account the possibility that $z > 1$. To this end, for each pair of features the function checks for differences in m/z compatible with the isotopic mass difference for different values of $z = 1, ...,$'maxCharge'. 'maxCharge' is a parameter that can be set by the user, by default it is set to 3 (see Table S2) .

Once the list of isotope features has been obtained, the next step is to build a directed isotope network. Here each node is a feature, and edges go from the heavier isotope to the lighter isotope (or the monoisotopic mass). Now, instead of pairs of features we have chains of isotopes. At this point the 'correctIsotopes' function is applied to correct incompatibilities (e.g. having two isotopes for the same monoisotopic mass). Moreover, this function splits the chain of isotopes if the number of isotopes in the chain is larger than "maxGrade", which is the maximum number of isotopes set by the user. Finally, all group isotope annotations are joined and stored into the 'anClique-class' object as a final step.

```
ex.Isotopes <- getIsotopes(ex.cliqueGroups, ppm = 10)

## Computing isotopes
## Updating anClique object
```

Usage Step c - Adduct annotation. The function 'getAnnotation' annotates pre-defined adducts and in-source fragments and obtains molecular neutral masses of metabolites. To use this function, the user needs to supply a list of adducts and in-source fractions(or use the default list). This list should be a 'data.frame' with information about the putative adducts in the sample. Here is an example of the default list of possible adducts/in-source fractions for positive ionization LC-MS1 spectra:

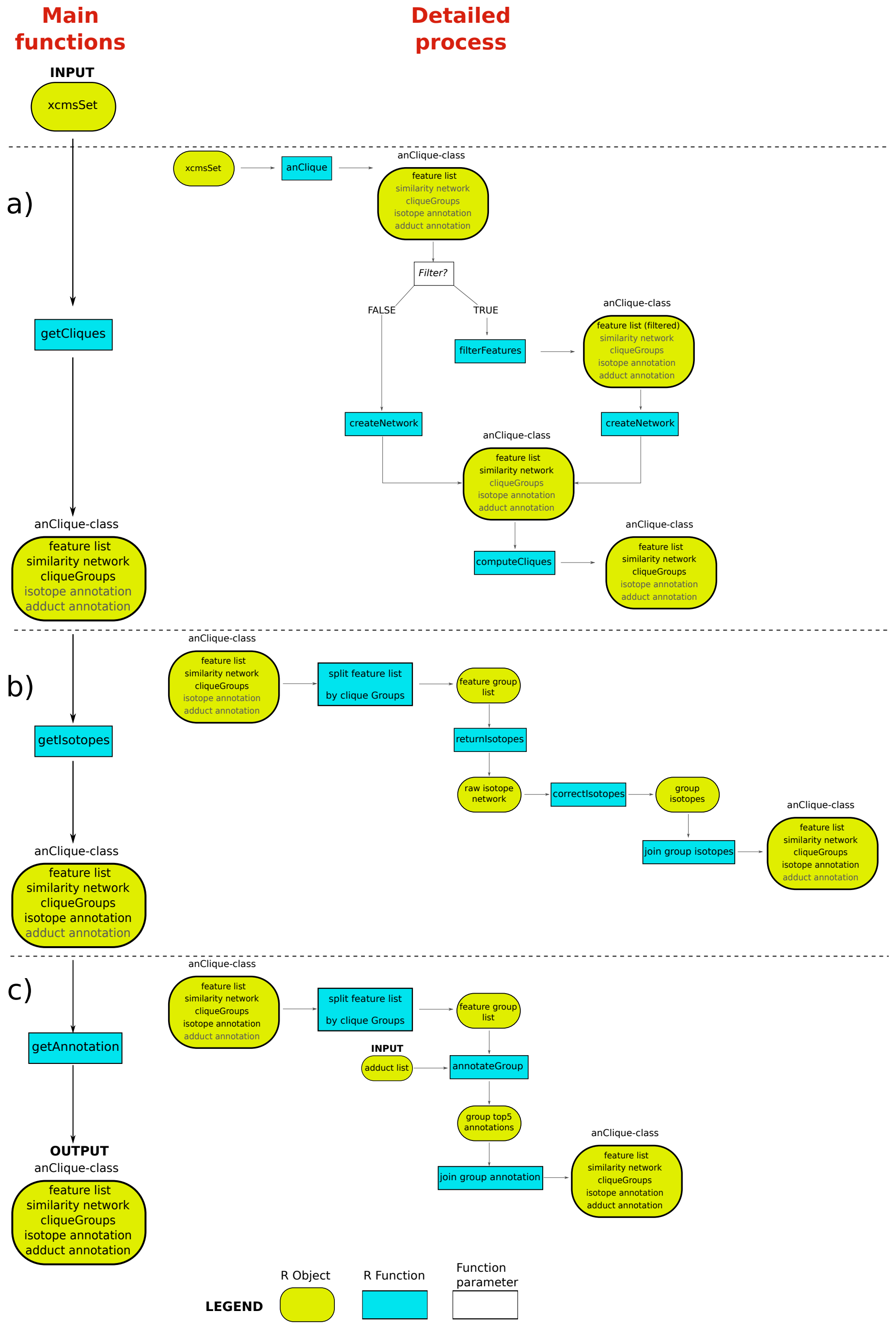


Figure S1: 'CliqueMS' program workflow. a) Feature group (clique) identification. b) Isotope anotation. c) Adduct annotation. The rounded yellow rectangles are R objects, the blue rectangles are R functions and white rectangles are function parameters

```
head(positive.adinfo)

##          adduct log10freq      massdiff nummol charge
## 1 [M+2H-NH3]2+ -3.512904 -15.012016600      1      2
## 2       [Cat]3+ -3.512904  -0.001645737      1      3
## 3       [Cat]2+ -3.512904  -0.001040400      1      2
## 4     [Cat+H]2+ -3.336813   1.006178842      1      2
## 5      [M+2H]2+ -1.813934   2.014552000      1      2
## 6   [M+H+Na]2+ -2.699991  23.996494000      1      2
```

The ’adduct’ column corresponds to a string that describes the adduct. The “log10freq” gives the $\log_{10}$ frequency of observation of that adduct or the log-score associated to that adduct which will be used to compute the annotation score. The “massdiff” column contains the mass associated to that adduct or in-source fraction. The “nummol” column is equal to 1 if the adduct only needs one molecule of metabolite to make the adduct, to 2 if it is a dimerization, to 3 if it is a trimerization and so on and so forth. The ’charge’ column is the charge of the adduct.

The function ’getAnnotation’ (see Fig. S1c) gets annotations for each group (clique) of features. For each feature group, first it removes those features that are not monoisotopic features (based on the previous isotope annotation). Then it produces annotations for each clique or sublclique (following the methodology described in the manuscript), and selects the five top scoring annotations for each group. Finally all group annotations are stored into the ’anClique-class’ object.

```
ex.Adducts <- getAnnotation(ex.Isotopes, ppm = 10,
       adinfo = positive.adinfo, polarity = "positive")

## Computing annotation
## Annotation computed, updating peaklist
```

Output. The final annotations are stored in a ’data.frame’. The columns of this data.frame (using our example) are the following:

```
features.clique6 <- ex.Adducts$cliques[[6]]
colnames(ex.Adducts$peaklist[features.clique6,])

##  [1] "mz"          "mzmin"       "mzmax"       "rt"          "rtmin"
##  [6] "rtmax"       "into"        "intb"        "maxo"        "sn"
## [11] "sample"      "cliqueGroup" "isotope"     "mass1"       "an1"
## [16] "score1"      "mass2"       "an2"         "score2"      "mass3"
## [21] "an3"         "score3"      "mass4"       "an4"         "score4"
## [26] "mass5"       "an5"         "score5"
```

The first 11 columns coorespond to data directly extracted from the ’xcmsSet’ object. Those columns are “mz”, “mzmin”, “mzmax”, “rt”, “rt”, “rtmin”, “rtmax”, “into”, “intb”, “maxo”, “sn”and “sample”.

From column 12 (“cliqueGroup”) to the last colum, they correspond to ’cliqueMS’ output. These columns contain the following information:

```
head(ex.Adducts$peaklist[features.clique6,
  c('cliqueGroup','isotope','mass1','an1',
  'score1')], n = 5)$

##    cliqueGroup isotope    mass1      an1      score1
## 37          38      M0       NA              -34.62174
## 38          38 M1 [12] 484.1610   [M+H]+   -34.62174
## 39          38      M0 484.1610 [M+NH4]+   -34.62174
## 40          38      M0 484.1610   [M+K]+   -34.62174
## 41          38      M0 105.0583 [Cat-H]+ -2524.47662
```

The first column “cliqueGroup” corresponds to the clique group of that feature. The column “isotope” shows the isotopic status of that feature. “M0” indicates features which are a monoisotopic mass. For features which are isotopes of other features we show both the isotope number and the isotope cluster. All the features that have the same isotope cluster are isotopes of the same feature. For instance

```
##    cliqueGroup isotope    mass1      an1      score1
## 38          38 M1 [12] 484.1610   [M+H]+   -34.62174
```

shows that for feature 38 the entry for the column isotope reads M1 [12], indicating that feature 38 is the first isotope (M1) in isotope cluster 12. Therefore, an entry “M*i* [X]” in the isotope column indicates that that feature is the *i*th isotope in isotope cluster X.

Columns “mass$A$ an$A$ score$A$” ($A = 1, .., 5$) correspond to adduct annotations in the top 5 annotations. For annotation A: column “mass$A$” shows the molecular neutral mass of that feature; column “an$A$” shows the adduct annotation of that feature; and, column “score$A$” shows the logarithmic score of the clique group or the score of the subdivision of the clique group.

```
head(ex.Adducts$peaklist[features.clique6,
  c('mass2','an2','score2',
  'mass3','an3','score3')], n = 5)

##       mass2     an2       score2    mass3         an3      score3
## 37 242.0805  [M+K]+    -35.65504 242.0805      [M+K]+   -353.7002
## 38 242.0805 [2M+H]+    -35.65504 502.1711 [M+H-H2O]+   -353.7002
## 39       NA            -35.65504       NA               -353.7002
## 40 242.0805 [2M+K]+    -35.65504 502.1711 [M+K-H2O]+   -353.7002
## 41 208.1014 [Cat]2+ -2918.52977 208.1014     [Cat]2+ -5657.1679

head(ex.Adducts$peaklist[features.clique6,
  c('mass4','an4','score4',
  'mass5','an5','score5')],n = 5)
```

```
##        mass4        an4     score4    mass5        an5     score5
## 37 522.1188  [M+H+K]2+  -353.939 522.1188 [M+H+K]2+  -354.0914
## 38 501.1876 [M+H-NH3]+  -353.939 462.1791   [M+Na]+  -354.0914
## 39 501.1876     [M+H]+  -353.939       NA            -354.0914
## 40 522.1188     [M+H]+  -353.939 522.1188    [M+H]+  -354.0914
## 41 105.0583   [Cat-H]+ -8593.658 105.0583  [Cat-H]+ -8655.6807
```

Function parameters Parameters in ′getCliques′, ′getIsotopes′ and ′getAnnotation′ control all internal functions shown in figure S1.The following tables show the main parameters of these functions and which are the default values we use for the results we present in the manuscript.

Parameter	Value	Default	Usage
filter	TRUE/FALSE	TRUE	If «TRUE», filter features that have cosine similarity > 0.99 and equal m/z, retention time and intensity value
mzerror	numeric	$5*10^{-6}$	If m/z relative error is below this value features are considered with the same m/z value
intdiff	numeric	$1*10^{-4}$	If intensity relative error is below this value features are considered with the same m/z value
rtdiff	numeric	$1*10^{-4}$	If retention time relative error is below this value features are considered with the same m/z value
tol	numeric	$1*10^{-5}$	Minimum relative increase in log-likelihood to do a new round of log-likelihood maximisation

Table S1: Main parameters for ′getCliques′

Parameter	Value	Default	Usage
maxCharge	numeric	3	Maximum charge considered when comparing pairs of features
maxGrade	numeric	2	Maximum number of isotopes in an isotope cluster, without counting the monoisotopic mass
ppm	numeric	10	Relative error in ppm to consider that two features have the mass difference of an isotope
isom	numeric	1,003355	Mass difference of an isotope

Table S2: Main parameters for ′getIsotopes′

Parameter	Value	Default	Usage
polarity	«positive/negative»		Polarity of the adducts
ppm	numeric	10	Relative error in ppm under which we consider two or more features compatible with a neutral mass and two or more adducts from the adduct list
emptyS	numeric	$1*10^{-6}$	Score given to non annotated features, used to compute the group score

Table S3: Main parameters for ′getAnnotation′

S3 Detail of the analysis of the different datasets

S3.1 Detail of the analysis of single sample datasets

For single samples (standards and IRS2 KO), we compared the annotation of CliqueMS with that of CAMERA, as the other annotation tools used in this work are not suited for single samples.

In the three samples, we used CliqueMS with default parameters (see Tables S1, S2, and S3) except for the parameter ′maxGrade′ in function ′getIsotopes′ where we set ′maxGrade′ = 3.

CAMERA For these three samples, we used CAMERA version 1.2.6 with R version 3.2.3. We used function ′groupFWHM′ with default parameters. For isotope annotation in function ′getIsotopes′ we set ′maxCharge′ = 2, ′maxiso′ = 3, ppm = 10, mzabs = 0, ′maxo′ for parameter ′intval′ and used a custom isotope matrix with isotope mass difference 1.003355. For grouping we used the following parameters with function ′groupCorr′: ′calcIso′ = TRUE, ′calcCiS′ = TRUE, and default values for the other parameters. For adduct annotation, we used a custom adduct list with the same adducts as CliqueMS. In function ′findAdduct′ we set parameter ′ppm′ = 10 and parameter ′mzabs′ = 0.

S3.2 Detail of analysis of MBTLS103 dataset

To test the suitability of CliqueMS for multiple samples, we consider a subset of spectra with positive ionization. These datasets, called ′HILIC′ (13 samples) and ′C18′ (18 samples) are found in MBTLS103 (`https://www.ebi.ac.uk/metabolights/MTBLS103`) dataset, all samples belong to the control group. Reported metabolites in the original study were manually inspected for other adducts in the raw data, which were used for validation of the results. The annotation of CliqueMS was compared with that obtained from running the same datasets with CAMERA, MS-FLO and xMSAnnotator.

Since CliqueMS is designed for single sample annotation, files were annotated independently, obtaining a different annotation list for each sample. We used default parameters for ′getCliques, ′getIsotopes′ and ′getAnnotation′ functions. On the other hand, MS-FlO and xMSAnnotator are suited for numerous samples, thus for the latter three methods the inputs were the matrices that result from performing a complete XCMS workflow analysis that includes: peak picking, alignment and grouping (parameters set as reported in the original study [**?**]). Importantly, this meant that the feature lists used by these methods were different than the CliqueMS feature list, thus this was taken into account when analysing the outputs.

xMSAnnotator. We used xMSAnnotator package version 1.3.2 in R 3.3.3. For obtaining annotations we used ′multilevelannotation()′ function with default parameters except that we set ′min_ions_perchem′ = 2 and used an ′adduct_weights′ table to mimic CliqueMS adduct frequency table. The function ′multilevelannotation′ performs a multistage clustering algorithm based on intensity profiles, retention time characteristics, mass defect, isotope/adduct patterns and correlation with signals for metabolic precursors and products.

We must note that not all adducts available in CliqueMS are found in xMSannotator. Therefore, and for a better comparison, we selected all the overlapping options. From all adducts reported by CliqueMS in the 6 metabolites of dataset MTBLS103 Hilic and in the 9 metabolites of dataset MTBLS103 C18, only adducts (M+H-NH3)+ and (M-2H+3Na)+ could not be selected for xMSAnnotator.

MS-FLO. We also used the online tool MS-FLO (v1.7)for performance comparison against CliqueMS. We used the XLSX input format, which consists of the modified matrix obtained by XCMS grouping. The isotope and adduct annotation were enabled, using an RT window of 0.05 min (default is 0.02 min).

Since the adduct list in MS-Flo is customizable but only suitable for pairs of single charged adducts, a total of 171 unique pairs of these adducts available in the CliqueMS adduct list were used (nummol = 1 and charge = 1).

CAMERA For this dataset we used version 1.36.0 of CAMERA package with R version 3.5.0. For function 'groupFWHM' we used default parameters. For isotope annotation in function 'getIsotopes' we set 'maxCharge' = 2, 'maxiso' = 3, ppm = 10, mzabs = 0, 'maxo' for parameter 'intval' and used a custom isotope matrix with isotope mass difference 1.003355. For grouping we used the following parameters with function 'groupCorr': 'calcIso' = TRUE, 'calcCiS' = TRUE, 'calcCaS' = TRUE, 'graphMethod' = 'lpc' with default correlation and p.value thresholds. For adduct annotation, we used a custom adduct list with the same adducts as CliqueMS. In function 'findAdduct' we set parameter 'ppm' = 10 and parameter 'mzabs' = 0.

S4 Supporting Figures

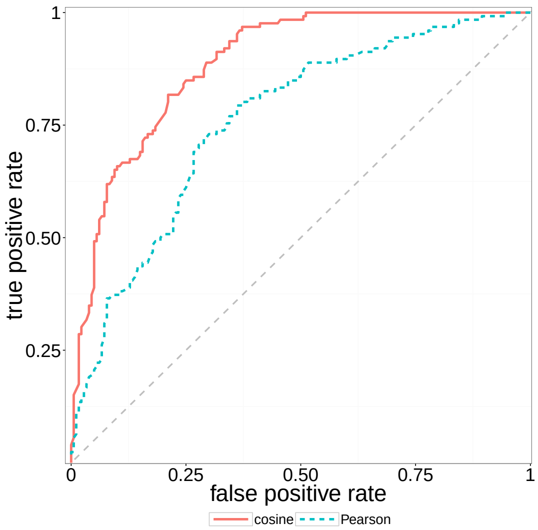


Figure S2: We compare the power of Pearson correlation and cosine similarity to distinguish pairs of features that are adducts/fragments of the same metabolite (true positives) from pairs of features belonging to different metabolites (false positives) in simulated coelution. To simulate coelution, we use 43 features from the LC-MS1 spectrum of a mixture of 9 standards that which were easy to manually identify due to the differences in retention times and the m/z values (see main 2.2). To simulate coelution we manually shifted features corresponding to different metabolites along the retention time axis, so that the maximum intensities of all of the features are aligned. For each pair of features we then compute Pearson correlation and cosine similarity. To compare the discriminatory power of these two similarity metrics, we use receiving operating characteristic curves (ROC) [**?**] and its area under the curve (AUC) for both methods (Fig. S3). The AUC value is the probability that a random pair of features corresponding to the same metabolite has a larger similarity than a random pair of features corresponding to different metabolites. Therefore, the larger the AUC value the larger the discriminatory power. To draw the ROC curve for the cosine similarity we first select threshold $T \in [0, 1]$. Pairs of features with a similarity $\geq T$ are considered 'positives', that is to come from the same metabolite; the remaining pairs are considered as 'negatives.' The true positive rate (y axis) is the fraction of positive pairs that truly come from the same metabolite. The false positive rate (x axis) is the fraction of positive pairs that do not come from the same metabolite. We sweep over all possible values of T to produce each curve. For the Pearson correlation $T \in [-1, 1]$. The cosine similarity (red, area under the ROC curve (AUC) = 0.887) has a higher discriminatory power than that of the Pearson correlation (blue, AUC = 0.760). Total number of random correlations: 180 Total number of real correlations: 126

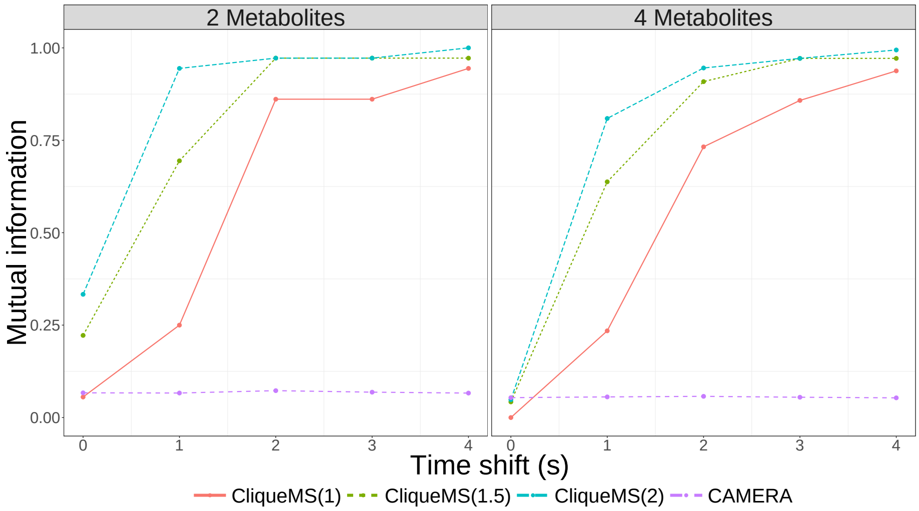
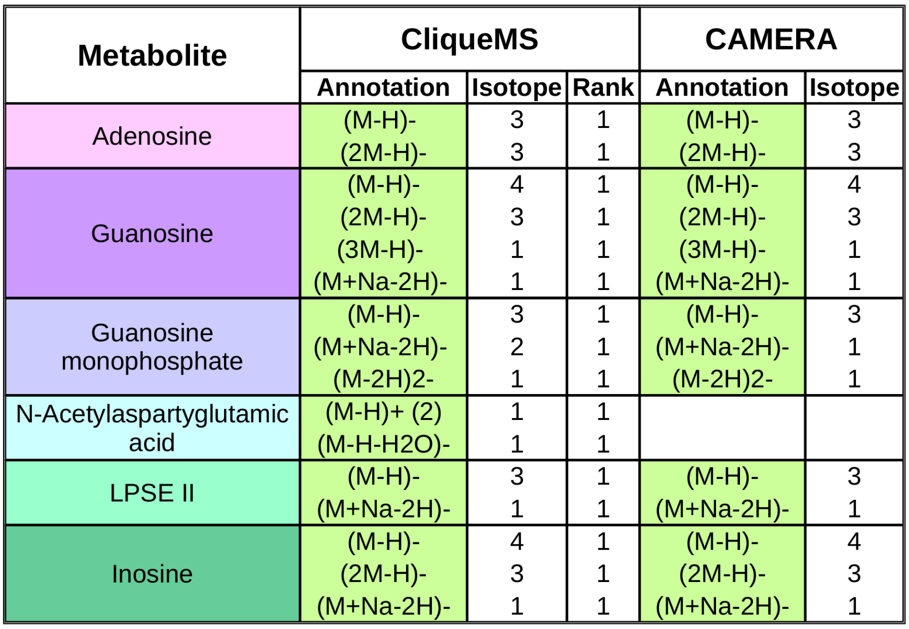
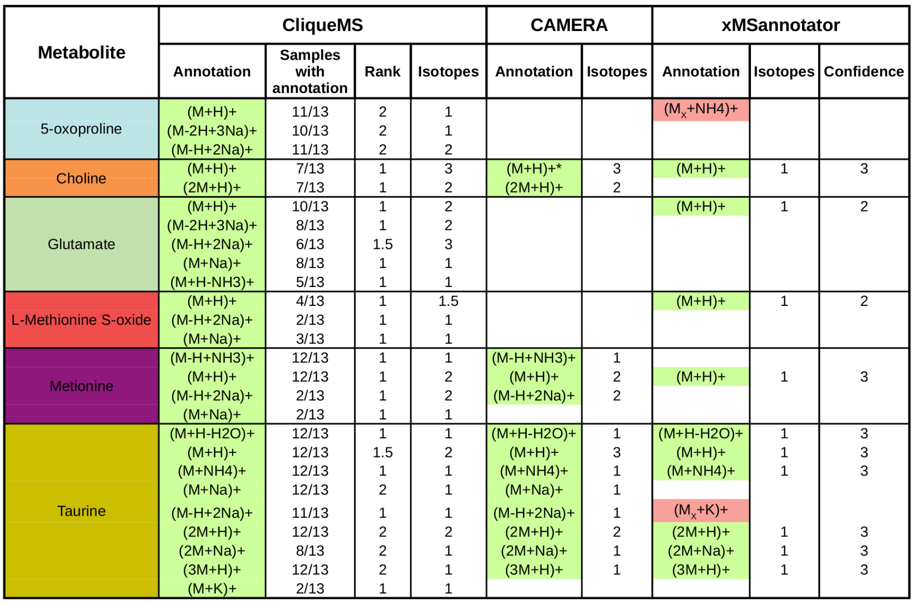


Figure S3: Identification of groups of features with similar coelution patterns. We simulate the coelution of 2 and 4 metabolites at different time shifts. The time shift is the time difference between the most intense feature of each metabolite. We run our clique identification algorithm for different choices of $\alpha = 1, 1.5, 2$ to define $p(\sigma_i = \sigma_j | c_{ij})$ as shown in Eq. 3 in the main text. We also run the group identification algorithm in CAMERA, which is also network based, for reference. To asses the accuracy of the group label assignments produced by each algorithm with respect to the known groupings, we use the adjusted mutual information (AMI). We show the AMI versus the time shift for the different algorithms we consider. The grouping algorithm within CliqueMS returns grouping closer to the nominal ones for all the time shifts considered. We find that for CliqueMS overall the best group assignments correspond to $\alpha = 2$.



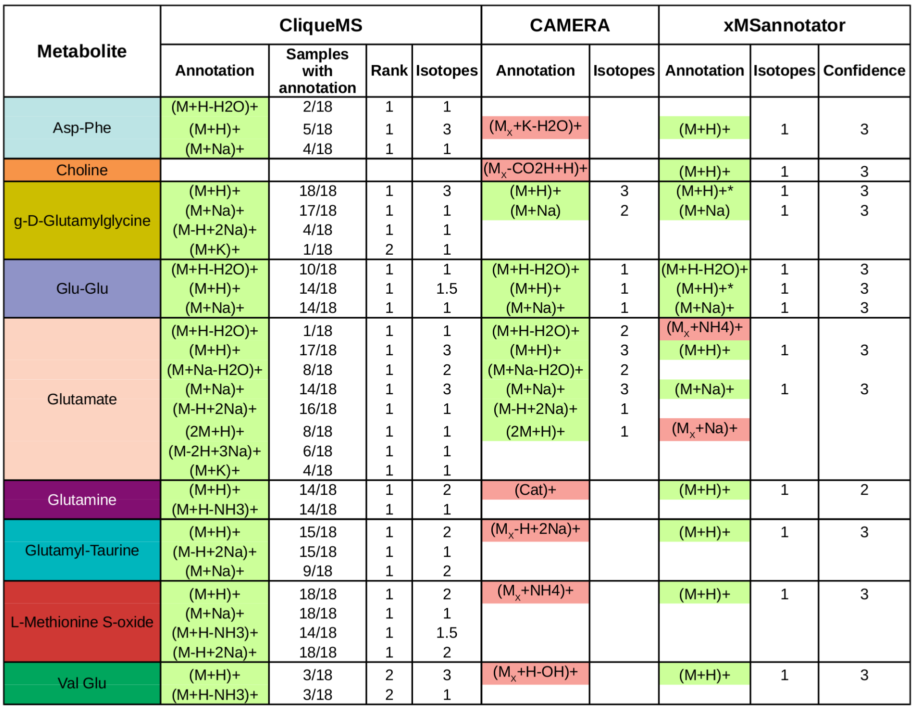
Metabolite	CliqueMS			CAMERA	
	Annotation	Isotope	Rank	Annotation	Isotope
Adenosine	(M-H)- (2M-H)-	3 3	1 1	(M-H)- (2M-H)-	3 3
Guanosine	(M-H)- (2M-H)- (3M-H)- (M+Na-2H)-	4 3 1 1	1 1 1 1	(M-H)- (2M-H)- (3M-H)- (M+Na-2H)-	4 3 1 1
Guanosine monophosphate	(M-H)- (M+Na-2H)- (M-2H)2-	3 2 1	1 1 1	(M-H)- (M+Na-2H)- (M-2H)2-	3 1 1
N-Acetylaspartylglutamic acid	(M-H)+ (2) (M-H-H2O)-	1 1	1 1		
LPSE II	(M-H)- (M+Na-2H)-	3 1	1 1	(M-H)- (M+Na-2H)-	3 1
Inosine	(M-H)- (2M-H)- (M+Na-2H)-	4 3 1	1 1 1	(M-H)- (2M-H)- (M+Na-2H)-	4 3 1

Figure S4: Detail of the adducts and in-source fragments annotated by CliqueMS and CAMERA for the retina samples of IRS2 deficient mice (- ionization). For each molecule we show the different adducts and in-source fragments annotated and the total number of isotopic variants of that particular adduct/in-source fragment. Correctly annotated features are shown in green; incorrectly annotated features are shown in red, with M_X indicating that the associated parental mass was incorrect; non-annotated features are shown in white. For CliqueMS we also show the ranking of the feature annotation that matches manual annotation. For CAMERA the * indicates those features for which the algorithm returned two possible annotations (see Supplementary Table S5 for the complete results obtained for this sample using CliqueMS and Table S7 for the complete list of manually annotated metabolites)



Metabolite	CliqueMS				CAMERA		xMSannotator		
	Annotation	Samples with annotation	Rank	Isotopes	Annotation	Isotopes	Annotation	Isotopes	Confidence
5-oxoproline	(M+H)+	11/13	2	1			(M$_x$+NH4)+		
	(M-2H+3Na)+	10/13	2	1					
	(M-H+2Na)+	11/13	2	2					
Choline	(M+H)+	7/13	1	3	(M+H)+*	3	(M+H)+	1	3
	(2M+H)+	7/13	1	2	(2M+H)+	2			
Glutamate	(M+H)+	10/13	1	2			(M+H)+	1	2
	(M-2H+3Na)+	8/13	1	2					
	(M-H+2Na)+	6/13	1.5	3					
	(M+Na)+	8/13	1	1					
	(M+H-NH3)+	5/13	1	1					
L-Methionine S-oxide	(M+H)+	4/13	1	1.5			(M+H)+	1	2
	(M-H+2Na)+	2/13	1	1					
	(M+Na)+	3/13	1	1					
Metionine	(M-H+NH3)+	12/13	1	1	(M-H+NH3)+	1			
	(M+H)+	12/13	1	2	(M+H)+	2	(M+H)+	1	3
	(M-H+2Na)+	2/13	1	2	(M-H+2Na)+	2			
	(M+Na)+	2/13	1	1					
Taurine	(M+H-H2O)+	12/13	1	1	(M+H-H2O)+	1	(M+H-H2O)+	1	3
	(M+H)+	12/13	1.5	2	(M+H)+	3	(M+H)+	1	3
	(M+NH4)+	12/13	1	1	(M+NH4)+	1	(M+NH4)+	1	3
	(M+Na)+	12/13	2	1	(M+Na)+	1			
	(M-H+2Na)+	11/13	1	1	(M-H+2Na)+	1	(M$_x$+K)+		
	(2M+H)+	12/13	2	2	(2M+H)+	2	(2M+H)+	1	3
	(2M+Na)+	8/13	2	1	(2M+Na)+	1	(2M+Na)+	1	3
	(3M+H)+	12/13	2	1	(3M+H)+	1	(3M+H)+	1	3
	(M+K)+	2/13	1	1					

Figure S5: Detail of the adducts and in-source fragments annotated by CliqueMS, CAMERA and xMSannotator for the MTBLS103 HILIC dataset. For each molecule we show the different adducts and in-source fragments annotated and the total number of isotopic variants of that particular adduct/in-source fragment. Correctly annotated features are shown in green; incorrectly annotated features are shown in red, with M_X indicating that the associated parental mass was incorrect; non-annotated features are shown in white. For CliqueMS, we also show the number of samples (out of 13) in which the correct annotation was among the top 5 annotations and the median ranking across samples of the feature annotation that matches manual annotation. For CAMERA the * indicates those features for which the algorithm returned two possible annotations. For xMSannotator we show the confidence level assigned to each annotation. A confidence level of 3 corresponds to high confidence; a confidence level of 0 corresponds to no confidence.



Metabolite	CliqueMS				CAMERA		xMSannotator		
	Annotation	Samples with annotation	Rank	Isotopes	Annotation	Isotopes	Annotation	Isotopes	Confidence
Asp-Phe	(M+H-H2O)+	2/18	1	1					
	(M+H)+	5/18	1	3	(M_x+K-H2O)+		(M+H)+	1	3
	(M+Na)+	4/18	1	1					
Choline					(M_x-CO2H+H)+		(M+H)+	1	3
g-D-Glutamylglycine	(M+H)+	18/18	1	3	(M+H)+	3	(M+H)+*	1	3
	(M+Na)+	17/18	1	1	(M+Na)	2	(M+Na)	1	3
	(M-H+2Na)+	4/18	1	1					
	(M+K)+	1/18	2	1					
Glu-Glu	(M+H-H2O)+	10/18	1	1	(M+H-H2O)+	1	(M+H-H2O)+	1	3
	(M+H)+	14/18	1	1.5	(M+H)+	1	(M+H)+*	1	3
	(M+Na)+	14/18	1	1	(M+Na)+	1	(M+Na)+	1	3
Glutamate	(M+H-H2O)+	1/18	1	1	(M+H-H2O)+	2	(M_x+NH4)+		
	(M+H)+	17/18	1	3	(M+H)+	3	(M+H)+	1	3
	(M+Na-H2O)+	8/18	1	2	(M+Na-H2O)+	2			
	(M+Na)+	14/18	1	3	(M+Na)+	3	(M+Na)+	1	3
	(M-H+2Na)+	16/18	1	1	(M-H+2Na)+	1			
	(2M+H)+	8/18	1	1	(2M+H)+	1	(M_x+Na)+		
	(M-2H+3Na)+	6/18	1	1					
	(M+K)+	4/18	1	1					
Glutamine	(M+H)+	14/18	1	2	(Cat)+		(M+H)+	1	2
	(M+H-NH3)+	14/18	1	1					
Glutamyl-Taurine	(M+H)+	15/18	1	2	(M_x-H+2Na)+		(M+H)+	1	3
	(M-H+2Na)+	15/18	1	1					
	(M+Na)+	9/18	1	2					
L-Methionine S-oxide	(M+H)+	18/18	1	2	(M_x+NH4)+		(M+H)+	1	3
	(M+Na)+	18/18	1	1					
	(M+H-NH3)+	14/18	1	1.5					
	(M-H+2Na)+	18/18	1	2					
Val Glu	(M+H)+	3/18	2	3	(M_x+H-OH)+		(M+H)+	1	3
	(M+H-NH3)+	3/18	2	1					

Figure S6: Detail of the adducts and in-source fragments annotated by CliqueMS, CAMERA and xMSannotator for the MTBLS103 C18 dataset. For each molecule we show the different adducts and in-source fragments annotated and the total number of isotopic variants of that particular adduct/in-source fragment. Correctly annotated features are shown in green; incorrectly annotated features are shown in red, with M_X indicating that the associated parental mass was incorrect; non-annotated features are shown in white. For CliqueMS, we also show the number of samples (out of 18) in which the correct annotation was among the top 5 annotations and the median ranking across samples of the feature annotation that matches manual annotation. For CAMERA the * indicates those features for which the algorithm returned two possible annotations. For xMSannotator we show the confidence level assigned to each annotation. A confidence level of 3 corresponds to high confidence; a confidence level of 0 corresponds to no confidence.

S5 Description of Supporting Tables/Files

Adduct/In-source Fragment input tables for CliqueMS
Each file has the format:
Adduct/In-source Fragment, mass difference, number of molecules, charge

Supplementary Table 1 (adducts.positive.cliqueMS.csv) – Table listing the mass difference, charge, and frequency of adducts and fragments we used as input to CliqueMS to obtain the results we report in the manuscript for positive ionization spectra.
Supplementary Table 2 (adducts.negative.CliqueMS.csv) – Table listing the mass difference, charge, and frequency of adducts and fragments we used as input to CliqueMS to obtain the results we report in the manuscript for negative ionization spectra.

Table S4 lists the remaining online files and their content.
Files peaklist.XXX.cliqueMS.csv show CliqueMS output for sample XXX as detailed in Table S4. CliqueMS files have the following columns for each feature in the XCMS file:
`[mz,mzmin,mzmax,rt,rtmin,rtmax,into,intb,maxo,sn,sample,cliqueGroup, isotope,mass1,an1,score1,mass2,an2,score2,mass3,an3,score3,mass4,an4, score4,mass5,an5,score5]`

`[mz,mzmin,mzmax,rt,rtmin,rtmax,into,intb,maxo,sn]` are directly obtained from the XCMS output and are useful for feature identification

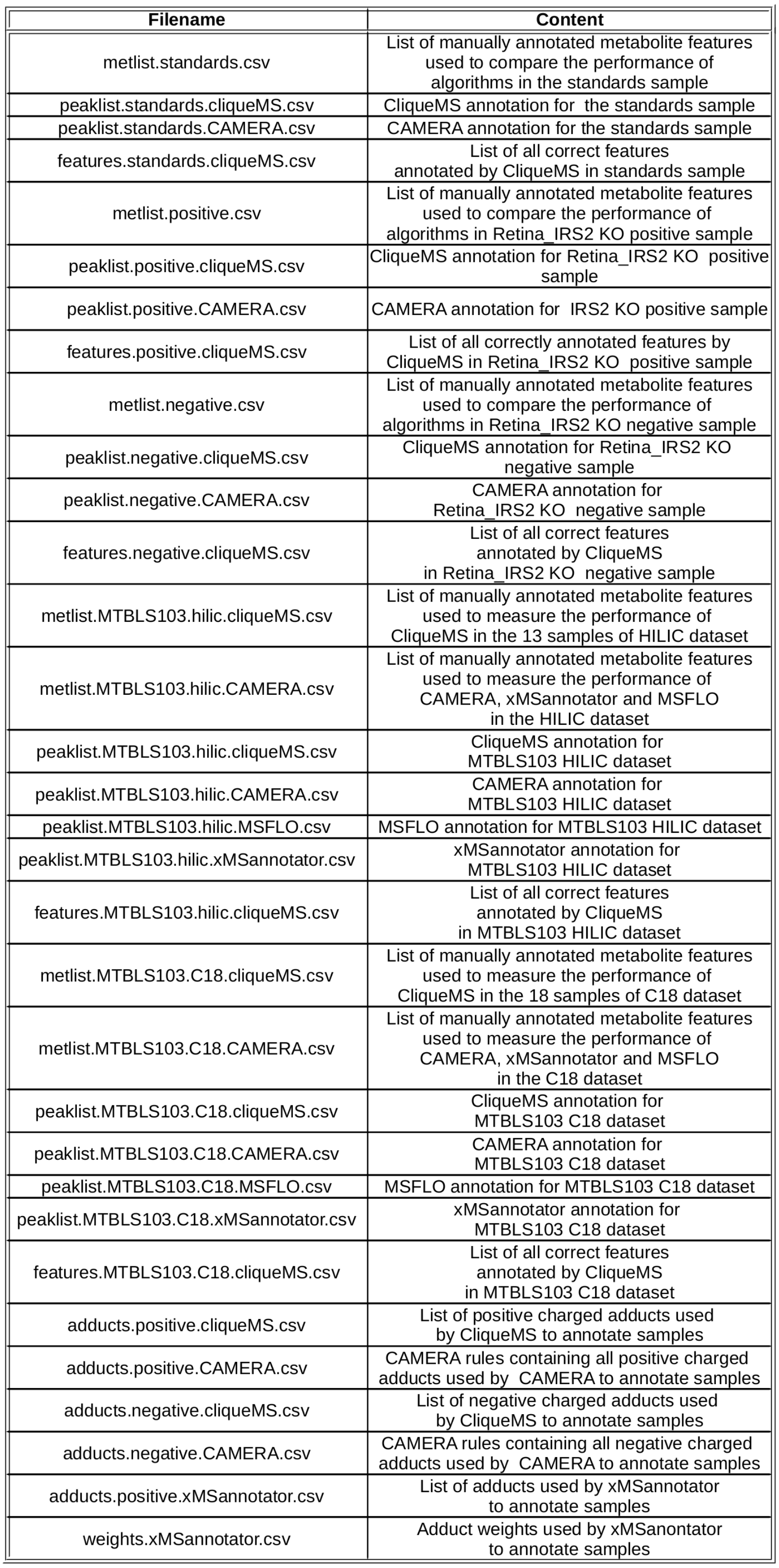
`mz`: average m/Z value for this feature.
`mzmin/mzmax`: minimum and maximum values of m/Z.
`rt`: average retention time for this feature.
`rtmin/rtmax`: time at which the intensity starts/stops being different from zero.
`maxo`: maximum intensity value for that feature.
`into`: total intensity area.
`intb`: total intensity area corrected by the peak baseline.
`sn`: signal to noise ratio.
`sample`: sample number.

`[cliqueGroup,isotope,(mass$i,an$i,score$i $i ∈ {1,2,3,4,5})]` are generated by CliqueMS.

`cliqueGroup`: clique id to which this feature belongs.
`isotope`: it reports whether that feature is an isotope or not; M0 - C12, M1 - 1C13,M2 - 2C13, M3-3C13 (...). M1[7] means that this feature is an isotope with one C13 of another feature within clique 7.
`mass$i`: neutral mass assigned to this feature in annotation $i ∈ 1 to 5.

`an$i`: adduct/in-source fragment assigned to this feature in annotation $i.
`score$i`: score associated to annotation $i for clique cliqueGroup.

Note that as we explain in the main text, to speed up the process we split large cliques into smaller groups. We then provide an annotation and score for such smaller groups. The seller groups can be tracked by looking for the different unique scores associated to a cliqueGroup id. The second and subsequent annotations are constructed with the second highest ranked (or the pertinent order) annotations for each clique or subgroup.



Filename	Content
metlist.standards.csv	List of manually annotated metabolite features used to compare the performance of algorithms in the standards sample
peaklist.standards.cliqueMS.csv	CliqueMS annotation for the standards sample
peaklist.standards.CAMERA.csv	CAMERA annotation for the standards sample
features.standards.cliqueMS.csv	List of all correct features annotated by CliqueMS in standards sample
metlist.positive.csv	List of manually annotated metabolite features used to compare the performance of algorithms in Retina_IRS2 KO positive sample
peaklist.positive.cliqueMS.csv	CliqueMS annotation for Retina_IRS2 KO positive sample
peaklist.positive.CAMERA.csv	CAMERA annotation for IRS2 KO positive sample
features.positive.cliqueMS.csv	List of all correctly annotated features by CliqueMS in Retina_IRS2 KO positive sample
metlist.negative.csv	List of manually annotated metabolite features used to compare the performance of algorithms in Retina_IRS2 KO negative sample
peaklist.negative.cliqueMS.csv	CliqueMS annotation for Retina_IRS2 KO negative sample
peaklist.negative.CAMERA.csv	CAMERA annotation for Retina_IRS2 KO negative sample
features.negative.cliqueMS.csv	List of all correct features annotated by CliqueMS in Retina_IRS2 KO negative sample
metlist.MTBLS103.hilic.cliqueMS.csv	List of manually annotated metabolite features used to measure the performance of CliqueMS in the 13 samples of HILIC dataset
metlist.MTBLS103.hilic.CAMERA.csv	List of manually annotated metabolite features used to measure the performance of CAMERA, xMSannotator and MSFLO in the HILIC dataset
peaklist.MTBLS103.hilic.cliqueMS.csv	CliqueMS annotation for MTBLS103 HILIC dataset
peaklist.MTBLS103.hilic.CAMERA.csv	CAMERA annotation for MTBLS103 HILIC dataset
peaklist.MTBLS103.hilic.MSFLO.csv	MSFLO annotation for MTBLS103 HILIC dataset
peaklist.MTBLS103.hilic.xMSannotator.csv	xMSannotator annotation for MTBLS103 HILIC dataset
features.MTBLS103.hilic.cliqueMS.csv	List of all correct features annotated by CliqueMS in MTBLS103 HILIC dataset
metlist.MTBLS103.C18.cliqueMS.csv	List of manually annotated metabolite features used to measure the performance of CliqueMS in the 18 samples of C18 dataset
metlist.MTBLS103.C18.CAMERA.csv	List of manually annotated metabolite features used to measure the performance of CAMERA, xMSannotator and MSFLO in the C18 dataset
peaklist.MTBLS103.C18.cliqueMS.csv	CliqueMS annotation for MTBLS103 C18 dataset
peaklist.MTBLS103.C18.CAMERA.csv	CAMERA annotation for MTBLS103 C18 dataset
peaklist.MTBLS103.C18.MSFLO.csv	MSFLO annotation for MTBLS103 C18 dataset
peaklist.MTBLS103.C18.xMSannotator.csv	xMSannotator annotation for MTBLS103 C18 dataset
features.MTBLS103.C18.cliqueMS.csv	List of all correct features annotated by CliqueMS in MTBLS103 C18 dataset
adducts.positive.cliqueMS.csv	List of positive charged adducts used by CliqueMS to annotate samples
adducts.positive.CAMERA.csv	CAMERA rules containing all positive charged adducts used by CAMERA to annotate samples
adducts.negative.cliqueMS.csv	List of negative charged adducts used by CliqueMS to annotate samples
adducts.negative.CAMERA.csv	CAMERA rules containing all negative charged adducts used by CAMERA to annotate samples
adducts.positive.xMSannotator.csv	List of adducts used by xMSannotator to annotate samples
weights.xMSannotator.csv	Adduct weights used by xMSanontator to annotate samples

Table S4: List of supplementary files and content description available online