SMfinder: Small Molecules Finder for Metabolomics and Lipidomics analysis

Giuseppe Martano[1,#], Michele Leone[2], Pierluca D'Oro[2], Vittoria Matafora[1], Angela Cattaneo[1], Marco Masseroli[2], Angela Bachi[1,*]

[1] – IFOM, The FIRC Institute of Molecular Oncology, 20139 Milan, Italy.

[2] – Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, 20129 Milan, Italy.

[*] – corresponding author: angela.bachi@ifom.eu

ABSTRACT: Metabolomics and lipidomics studies are becoming increasingly popular but available tools for automated data analysis are still limited. The major issue in untargeted metabolomics is linked to the lack of efficient ranking methods allowing accurate identification of metabolites. Herein, we provide a user friendly open-source software, named SMfinder, for the robust identification and quantification of small molecules. The software introduces a MS2 false discovery rate approach, which is based on single spectral permutation and increases identification accuracy. SMfinder can be efficiently applied to shotgun and targeted analysis in metabolomics and lipidomics without requiring extensive in-house acquisition of standards as it can provide accurate identification by using available MS2 libraries in instrument independent manner. The software, downloadable at www.ifom.eu/SMfinder, is suitable for untargeted, targeted and flux analysis.

Metabolomics and lipidomics are increasingly being applied in various fields from microbial research to clinical studies and are now emerging as high-throughput diagnostic tools[1,2]. Mass spectrometry based approaches in metabolomics are the golden standard in the discipline due to the high sensitivity and specificity that this approach offers. However, the biggest challenge when using mass spectrometry, is the possibility to robustly identify large numbers of compounds[3]. This issue largely restricts the use of metabolomics to targeted analyses, where only a defined subset of metabolites is explored in the specimen of interest. Meanwhile untargeted experiments are often limited to features analysis, thus losing the biological information that this application could provide. In order to overcome these problems, mass spectrometry based data-dependent acquisition (DDA), originally used and optimized to maximize peptide identification in proteomics analysis[4,5], is now consolidating as a method of choice also in metabolomics and lipidomics[6,7]. DDA is suitable for large-scale discovery experiments, but needs bioinformatics tools to mine the large amount of data generated. Existing analysis software for metabolomics data are either commercial or require moderate to high knowledge of computer programming and an advanced analytical background, thus rendering them not easily accessible to the broad scientific community. Differently from proteomics, where few software are widely used and validated by the scientific community, in metabolomics several ad-hoc tools have been developed, but there is not an accepted consensus on which software would be preferable and most reliable. Small molecules identification relies heavily on the comparison with a database containing fragmentation spectra (MS2) of known compounds. One of the main limitations in metabolomics analysis is the lack of public and comprehensive MS2 libraries, as they are usually instrument based[8], home-made or in silico derived[9–11]. This is due to the type of computation, which is performed to match the empirical spectra with those contained in libraries. The calculation is based on the comparison of the spatial distance between empiric MS2 spectra and MS2 spectral libraries acquired from standards. Unfortunately, relying only on spatial distance forces the use of instrument based MS2 libraries as different instruments may generate different fragmentation spectra both in terms of fragments masses and relative intensities. As a result, comparing empirical spectra obtained with a specific instrument with spectral libraries built on a different instrument will return poor scores. A method often used is to compare only the masses of the fragments which are generated at a given collision energy. Similar to selected reaction monitoring (SRM), also in this case, the discrimination efficiency of identification relies on the presence of unique mass fragments i.e. reporter ions. Different fragments are not always presents when multiple isomers are compared. In this case, isomers with the same fragment masses have an identical likelihood to be identified and cannot be distinguished. These issues represent the major obstacle in the generation and efficient use of large empiric MS2 libraries. In practice, each metabolomics laboratory is forced to generate its own library made through acquisition of hundreds of standards with their own instrumentation. Moreover, hampered by the lack of a robust approach to estimate the false discovery rate (FDR)[12], compound annotation is particularly challenging for small molecules. Compared with a more developed field such as proteomics[13], working on small molecules does not allow the same computation algorithms used to calculate the FDR of peptides. Moreover, small molecules include a large variety of compounds from carbohydrates to lipids, which do not share similar fragmentation rules. These characteristics limit the possibility to efficiently predict the fragments that are generated by MS2 in a generalized fashion. With the aim to provide a user-

friendly tool for metabolomics analysis, we developed a software named Small Molecules finder (SMfinder), which enables confident identification and accurate quantification of both metabolites and lipid species, and introduces a FDR approach to estimate identification reliability.

Materials and Methods

Raw data generation for ^{13}C trace analysis using melanoma cells

IGR37 melanoma cell lines were cultured on 12 well plates with 13 mm glass coverslip for 24h in DMEM with glutamine. After 24h the media was switched to a fresh media with U-13C labelled glucose for 5 minutes before quenching and extraction. In control samples, the same procedure was applied by using a standard media composition with naturally labelled glucose. Sample preparation and chromatographic conditions were set as described elsewhere[14]. Briefly, cells were rapidly washed in MilliQ water at 37° and transferred to a new well containing a quenching solution. Metabolites quenching was achieved using cold solution with acetonitrile methanol and water (2:2:1; v,v,v) kept at -20 °C during the quenching procedure. Cells were mechanically detached from the coverslips, and the quenching solution was collect in 1.5 ml Eppendorf tubes. Lyophilized samples were reconstituted in MS grade water and injected for LC-MS analysis. Chromatography was performed on an Expert nanoLC 400 system (ABSciex) equipped with 1 µl loop, and with a chromatographic column prepared in house with 100 µm ID, 100 mm length packed with Kinetex C18-evo core shell particles with 1.7 µm ID (Phenomenex). The LC was coupled online by a nanoESI source with TripleTOF 6600 System (ABSciex).

Raw data generation for lipidomics analysis using melanoma cells

WM115 cells were cultured in DMEM+ 10% FBS S.A.+ 2 mM L-Glutamine. Cell pellets from $2x10^6$ cells were resuspended in 200 µl of 150 mM ammonium bicarbonate and passed through a 26G syringe needle to prepare cell lysate. Samples were centrifuged at 10000 g for 10 minutes at 4 °C to eliminate cell debris. An equivalent of 10 µg of proteins was spiked with internal standards and lipids were extracted using a 2 steps extraction protocol with methanol and chloroform in varying proportions[15]. Organic phase fractions were then dried out and resuspended in 50 µL of methanol. The internal standards were 16 combined as follow: PC (12:0/13:0) 40 pmol, PE (12:0/13:0) 52 pmol, PG (12:0/13:0) 7.5 pmol, PS (12:0/13:0) 43 pmol, PI (12:0/13:0) 54 pmol, Cer (d18:1/25:0) 100 pmol, CE(19:0) 100 pmol, GlcCer (d18:1/12:0) 50 pmol, LacCer (d18:1/12:0) 50 pmol, sphinganine (d17:0) 50 pmol, sphingosine-1-P (d17:1) 100 pmol, sphingosine (d17:1) 50 pmol, Galactosyl(ß) Sphingosine-d5 20 pmol, d5-TG ISTD Mix I 20 pmol, d5-DG ISTD Mix I 20 pmol, cholesterol (d7) 800 pmol. 95% phase A (CH3CN:H2O 40:60; 5 mM NH4COOCH3; 0.1% FA) plus 5% phase B (IPA:H2O 90:10; 5 mM NH4COOCH3; 0.1% FA) for subsequent analysis. Lipid extracts were diluted 1:5 in 95% phase A (CH3CN:H2O, 40:60; 5 mM NH4COOCH3; 0.1% FA) plus 5% phase B (IPA:H2O 90:10; 5 mM NH4COOCH3; 0.1% FA) and 1 µL injected on nLC Ekspert nanoLC400 (Eksigent, 5033460C; Singapore) coupled with a Triple TOF 6600 (AB Sciex). Chromatography was a 45 min gradient and MS Acquisition was performed in positive mode (Matafora et al., manuscript in preparation). Proteins were extracted from 20 µL of ammonium bicarbonate resuspended fractions by adding 5 µL of lysis buffer (10% NP40, 2% SDS in PBS) and quantified by BCA protein assay kit (ThermoFisher Scientific, 23225). For untargeted lipidomics analysis in SMfinder the following parameters were used: resolution 15000 with deisotoping option for peak picking; 60 seconds retention time tolerance for "Unique ID"; 50 ppm error with the exclusion of halogenated formulas; blind library with forced association and filter hierarchy based on FDR and isotopic similarity, and minimum count of 3 for filler function.

RESULTS

Overview of Design

The SMfinder primary function is to integrate different analysis pipelines, provide visualization of raw chromatographic data and perform computational analysis depending on the experimental needs. SMfinder is provided with an intuitive graphic interface and installation.exe file, which will automatically install the required dependencies and create the desktop shortcuts. SMfinder source code is not encrypted and can be modified by proficient Python users and also executed directly in the Python environment from the command prompt by using the path stored in the software folder (%AppData%\SMfinder_path.txt).

Framework, Availability and System Requirements

SMfinder is written in Python 2.7[16] using R[17] in the Python environment, eMZed2[18] for interactive data objects visualization, and PyQt4 graphic user interface (GUI) to minimize the user effort and capitalize on robust statistical analysis routines. The R connection provides the essential functions to convert raw chromatographic data into Python usable data object (pyObj) which are stored as metafiles. pyObj can then be visualized and analyzed. SMfinder is freely available at www.ifom.eu/SMfinder in the Download page. On the SMfinder tutorial webpage, example files are provided together with a detailed tutorial, and under the library tab the user can download available MS2 libraries e.g. HMDB library[19] with LC/MS spectra, and the MS2 IFOM curated library. On the same webpage, there is also the possibility to freely upload user libraries which will be released online following quality checks. SMfinder operates on Windows 7 or higher version. The minimal requirements are 4 cpu and 4 GB of RAM.

SMfinder Workflows

Three different workflows are available and are designed to perform untargeted, targeted, or ^{13}C-dynamic tracing profiles with minimal users input beside raw data (Fig. 1). Untargeted analysis is usually performed when there is not prior knowledge on a group of metabolites of interest, and it is a powerful tool for metabotype classification. In this type of analysis, features are extrapolated from raw data based on their spatial characteristics i.e. Gaussian like distribution over time within a mass window. After features detection, identification at MS and

MS2 level is obtained and a chemical formula and compound final annotation is assigned.

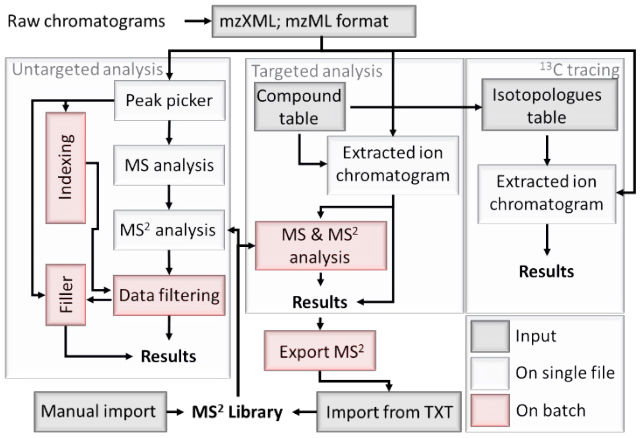

Figure 1 SMfinder workflow for untargeted, targeted and 13C tracing analysis. The arrows' direction represents the input-output of each function, e.g. "Data filtering", the process that returns the best match for each detected feature (Results), requires metadata from indexing and MS2 analysis. The color of each box shows if the process is performed on a single file or on the entire dataset selected by the user. For MS analysis, SMfinder indicates the sum of events in which a molecular formula is assigned to features in untargeted analysis, or superimposed on extracted ion chromatograms in targeted analysis, and then ppm errors and isotopic similarity are calculated. In MS2 analysis, found molecular formulas are matched with the MS2 library, and MS2 score and FDR are computed. Actions performed on batch, e.g. the indexing homogenization (Unique ID in SMfinder) in the dataset, are dependent on the size of the batch selected. When the size of batch is modified by the user, the on-batch functions need to be recomputed while "on single file" functions are not influenced by the batch size.

Raw data preparation

Before using SMfinder, chromatographic raw data, acquired in centroid mode, have to be converted e.g. using MSConvert[20], in mzXML or mzML format (Fig. 1). Chromatograms need to be acquired in a single polarity mode, either negative or positive, and can contain MS and MS2 or MS spectra only. However, if chromatograms do not contain MS2 spectra, compound name annotation can be achieved only on targeted or flux analysis experiments, but not for untargeted analysis. After conversion, raw files can be uploaded into SMfinder by selecting "Open raw files" under the File tab in the Menu. To illustrate the computational workflow and performance, we used two datasets from literature, one from a metabolic[21] profile analysis and the other from a lipidomics analysis[22], and chromatograms of lipidomics and metabolomics experiments of melanoma samples acquired in-house (Table S1).

Untargeted analysis

Parameters required for this type of analysis need to be specified by the user by typing or selecting the relative values on the GUI under the tab Untargeted. The first parameter is the instrumental resolution. This value varies depending on the used mass spectrometer and it is either imposed by the instrumental limits or selected by the users. The given resolution will be used to extract the peaks from raw data (Figure 2A-B, Supplementary Figure S1A-B). Peak picking is based on the matchFilter algorithm from XCMS for initial peak selection[23]. The choice of a correct resolution value, corresponding to the real instrument resolution, affects the entire analysis; since a fraction of the parameters for feeding the algorithm are extrapolated from the given resolution. For instance, lowering the resolution compared with the real value will result in a loss of spotted peaks (Fig. S1C-D) and it will generate mass windows which are larger than the expected range for the masses of interest. As a consequence, the reconstructed peaks will include larger noise and therefore an overestimation of the peak area (Fig. S1E), and increased integration errors (Fig. S1F). A similar behavior can also be observed if the selected resolution is higher than the real resolution. In this case, the reconstructed peak may contain only a subset of masses which belong to a compound, and therefore the peak area will be underestimated. Peak detection is performed on each single file. It is therefore necessary to cluster all the spotted peaks within the dataset before proceeding with the analysis. In a simplified example, having two raw files and only one spotted peak in each file, we need to assign whether the two peaks may represent the same compound in both samples or not. This process is performed automatically by SMfinder in two steps. First, masses are clustered with clustering criteria defined by the given instrumental resolution which is used to calculate the expected mass accuracy from a theoretical mass at 400 Da. Then, in each generated cluster, the retention times will be sub- clustered within a tolerance range defined by the user. Sub-clusters are enumerated and assigned on each peak. The result of this procedure assign a unique index for each peak that is found in one or more raw files in the dataset. For example, we applied our clustering algorithm to a metabolomics dataset[21], which showed a majority of detected peaks to be shared within the entire dataset (Fig. 2C). Spotted peaks can now be identified, and first, possible molecular formulas are matched with the detected mass of each peak. Possible matches can be pre-filtered by the user by defining the maximum ppm error, expected adducts, and the exclusion of molecular formulas that contain halogens. In this step, a library containing all known chemical formulas from PubChem database[24] is used for calculating all the possible masses based on the chromatographic polarity and the selected adducts options.

Database MS2 matching

Following peak extraction, SMfinder will match all the available MS2 spectra in the library that matches with the empirical formula. In this step SMfinder has a very flexible structure. Matches can be assigned by specifying the type of instrument, the used collision energy and by giving the number of permutations that will be required to compute the FDR. However, in most cases, using this type of setup will limit the database and therefore may reduce the identification coverage. To overcome this issue, two selectable values are included that are: "Use blind library", which will exclude the collision energy and instrumental filter, and "Force library association", which will exclude the adduct filter. During this process, four scores are generated for each feature matching at least one formula and

at least one MS2 spectra in the library, these scores are: ppm error, isotopic similarity, MS2 score, and FDR (as letter described in the computational operations and statistics paragraph). At this stage, a single feature may correspond to multiple potential matches. Data filtering can now be applied, by defining the hierarchy of filtering from the GUI. For example, by selecting FDR plus PPM, SMfinder will search for the match of any single feature with a lower FDR value. In the case of multiple matches with the same value it will select those with higher MS2 score, then with lower PPM error and finally with the higher isotopic similarity score until the best matching compound will be assigned to the detected feature. As an example, we used a raw file from literature (3injections_inj1_POS.mzXML) and ran the untargeted analysis with SMfinder (Table S2). For each peak, the multiplicity of matches was resolved to the best matching in order to return only one match per feature (Fig. 2D). The data filtering is designed to work on the entire dataset. The advantage of using this approach is that MS2 spectra become shared within the dataset. Therefore, for compound identification, it would be sufficient to find the MS2 spectra in only one file in order to identify the compounds also in other files where the MS2 for a give compound is missing (Fig. S1G). The last value that can be specified is the minimum count for the "Filler" function. This function will count the recurrence of each compound in the entire dataset, and if the recurrence is equal or higher of the given value, but lower than the total number of samples, it will search for missing values of those compounds. Results from the untargeted analysis can be now exported as excel file by clicking on Export Results.

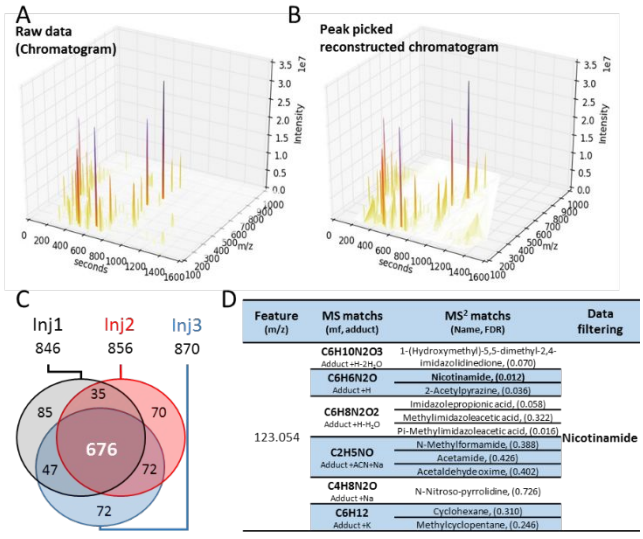

Figure 2 main SMfinder operations in untargeted analysis. (A) Graphic representation of a raw chromatogram from metabolomics analysis[21]. (B) Reconstructed chromatogram generated from the spotted peaks after peak picking. (C) Venn diagram of the spotted peaks detected in inj1, inj2 and inj3 upon clustering. Out of mo 856 peaks detected in each injection, 676 (79%) are found in all replicates, and 154 on two out of three replicates. Clusterization was performed with a retention time tolerance of 30 seconds. (D) Nicotinamide identification by SMfinder. Among all possible assigned compounds, nicotinamide (identified in the original paper by standard comparison) shows the lowest FRD score. For this analysis, the HMDB database was used as MS2 library, and the maximum number of permutation for FDR evaluation was set to 500. For data filtering, the higher hierarchy was given to FDR.

Targeted analysis

Another available workflow is designed to perform targeted analysis. The first step consists in the generation of a reference table listing the compounds of interest. The list can be generated directly on SMfinder by using the interacting windows or by uploading a CSV file containing all the required information: the molecular formula, chemical adducts and expected retention time. The retention time can be checked before starting the analysis by clicking on the homologous button on the GUI. As an example, we chose a list of lipids which were identified in a published article[22] and imported them as CSV into SMfinder (Table S4). Similar to the untargeted workflow, the instrumental resolution and peak width must be specified before starting the analysis. The resolution will be used to define the mass windows for the generation of the extracted ion chromatograms (EIC) from raw files (Fig 3A-B). EICs are integrated and the area is calculated for each peak. Results of the targeted analysis can be exported at this stage or the findings can be further validated by calculating MS and MS2 parameters similar to untargeted analysis. When this option is selected, the ppm error and isotopic similarity will be calculated at MS level, while SMfinder will match the formulas from the reference table with the library and evaluate the MS2 score and FDR. Validated results can be exported as excel file by clicking on the "Export Results" in the GUI.

MS2 data mining from raw data

When performing targeted analysis, compounds of interest are commonly acquired as standards prior to analysis of the experimental samples. In this case, the MS2 spectra recorded from standards can be extracted from the chromatogram and exported as a txt file by clicking on "Export txt for library". In our example of lipidomics analysis from literature[22], we can use the same principle to retrieve MS2 spectra from real samples. Indeed, when the targeted validation is performed, MS2 spectra from the raw chromatograms are bound to the EICs of each precursor ion at MS level within the retention time window of each EICs (Fig. 3C). This process can operate on a single file or on the entire batch. In batch mode, the generated quality and quantity of MS2 spectra is obviously higher as some MS2 spectra might have been acquired in a subset of chromatograms. For compounds which have more than one MS2 spectra within the dataset, the spectra with the highest summed intensity of all the fragments will be selected (Fig. 3D).

^{13}C dynamic tracing

This workflow follows the same principle of targeted analysis. It can be used in combination with targeted or untargeted analysis or used as stand-alone module. The first step consists in the generation of a reference table which is suitable for 13C dynamic tracing experiment. The reference table contains all the possible isotopologues for the compounds of interest. The isotopologues are generated from previously acquired unlabeled samples analyzed either with the untargeted or

targeted workflow. In this case, the information regarding the number of metabolites to be analyzed and their retention time are retrieved from the previous analysis. However, the reference table can be also newly generated starting from a compound table suitable for targeted analysis without the need of having a reference raw file. Once the reference table has been generated, it can be used for the 13C dynamic tracing analysis. The analysis can now be performed and the results will indicate the number of labelled carbons and the relative intensity of each isotopologue (Fig S2). Notably, the analysis of the unlabeled samples which is usually the starting point of the analysis will return the natural isotopic distribution as isotopologues abundance. This value can be used to estimate the starting level of labelling as far as no isotopic abundance correction is performed by SMfinder.

Integration accuracy

In order to verify the integration accuracy of SMfinder, we selected three lipid extracts from the same melanoma cell line, each acquired in technical duplicate. We choose this dataset because it should be homogenous, containing approximately the same number, and the same intensities of lipids while the chromatographic noise should partially differ between chromatograms i.e. instrumental noise. The samples, analyzed with the commercial software LipidView, returned the identification

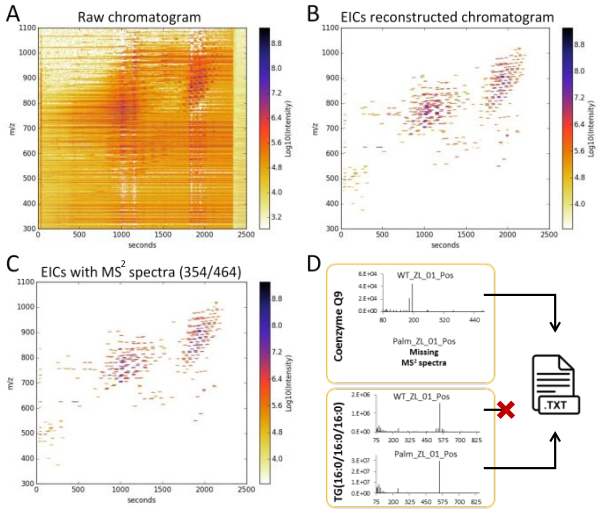

Figure 3 (A) Scatter plot of the raw file WT_ZL_01_Pos from a lipidomics experiment[22]. (B) Reconstructed scatter plot using the EICs with 464 lipids selected for targeted analysis. (C) Reconstruction of the 354 lipids for which MS2 spectra were found. (D) Example of MS2 spectra mining and selection from WT_ZL_01_Pos, and Palm_ZL_01_Pos files.

181 lipids (Table S3), which were found to be present in all the six chromatograms. For each of three biological samples, we compared the technical replicates against each other to check the reproducibility of the samples (Fig 4A). We used the list of detected lipids to build a reference table and ran the targeted analysis on the same samples using SMfinder. Results were compared by plotting orthogonally the peak area of each lipid found in the technical replicates for each of the biological samples. All the samples fit into a linear distribution, and the observed variation was within the expected instrumental reproducibility (Fig. 4B) demonstrating integration consistencies for peak area calculation within multiple injection of the same sample.

PPM and Isotopic similarity

In order to assign a chemical formula to the detected peaks, matches between chemical formulas and m/z values are computed. This is done by setting a tolerance range to the empiric m/z values that is based on the part per million (ppm) error provided by the user. The mass distance between the empiric mass value and the theoretical exact mass of each formula matching the previous criteria are then reported as ppm error values in SMfinder. Similarly, isotopic similarity is calculated by generating the exact mass of the natural isotopic distribution of each matching chemical formula. The theoretical and empiric distribution are then compared by using the Kolmogorov goodness of fit.

MS2 Score and FDR calculation

Depending on the instrument used, chemical formulas for each feature can be assigned, based solely upon molecular weight, with a reasonable accuracy. However, each chemical formula may correspond to a large number of isomers and stereoisomers which cannot be discriminated with MS information alone. To resolve isomers complexity, features with MS2 spectra that match at list one chemical formula, are compared with the MS2 library. First, features and library spectra are matched by chemical formulas. Then, the MS2 score is calculated starting from two matrices $x_{m,i}$ and $X_{m,i}$, respectively from the empiric MS2 spectra and MS2 spectra deposited in the library, with m and i that representing the masses and signal intensities respectively.

$$X_{m,i} = \begin{Bmatrix} X_{m_0} & X_{m_1}\dots & X_{m_n} \\ X_{i_0} & X_{i_1}\dots & X_{i_n} \end{Bmatrix};\ x_{m,i} = \begin{Bmatrix} x_{m_0} & x_{m_1}\dots & x_{m_n} \\ x_{i_0} & x_{i_1}\dots & x_{i_n} \end{Bmatrix}$$

Xm and xm are matched by considering xm a subgroup of Xm, and a merged matrix (Mx,y), with the relative intensities of X and x, is generated.

$$X_m \subseteq x_m \rightarrow M_{x,y} = \begin{Bmatrix} \frac{X_{i_0}}{\sum X_i} & \frac{X_{i_1}}{\sum X_i}\dots & \frac{X_{i_n}}{\sum X_i} \\ \frac{x_{i_0}}{\sum x_i} & \frac{x_{i_1}}{\sum x_i}\dots & \frac{x_{i_n}}{\sum x_i} \end{Bmatrix}$$

Following, cosine similarity is calculated based on the intensities.

$$MS^2\ score = \mathrm{Cos}(\theta) = \frac{M_x * M_y}{\|M_x\|\|M_y\|}$$

Similarly, FDR is calculated by re-computing N times, with N equal to the number of selected permutation, Cos(θ) against a generated library entry, where the intensities from the library (M_x) have been permuted (ε)

$$FDR = \frac{1 + \sum_0^N p}{1 + N}$$

where $p = 1$ if $\frac{M_x * M_y}{\|M_x\|\|M_y\|} \geq \frac{M_{\varepsilon x} * M_y}{\|M_{\varepsilon x}\|\|M_y\|}$ otherwise $p = 0$

We tested the FDR performance by analyzing the identification of UDP-glucose from the raw file 3injections_inj1_NEG from literature. Identification of compounds with sugar residues, such as UDP-glucose, is particularly challenging since MS2 spectra fragments of carbohydrates isomers are nearly or completely identical, with only minor differences in intensities. We evaluated UDP-glucose against MS2 spectra generated at -20 eV of UDP-glucose and UDP-galactose, manually imported into the SMfinder library from the Metlin database[8]. First, the MS2 score was computed as described above between MS2 spectra present in the database and the empiric spectra, and resulted to be 0.82 for UDP-glucose and 0.80 for UDP-galactose (Fig. 4C-D). Then, the FDR was calculated using between 50 to 10,000 permutations per test and each test was repeated 10 times (Fig. 4E).

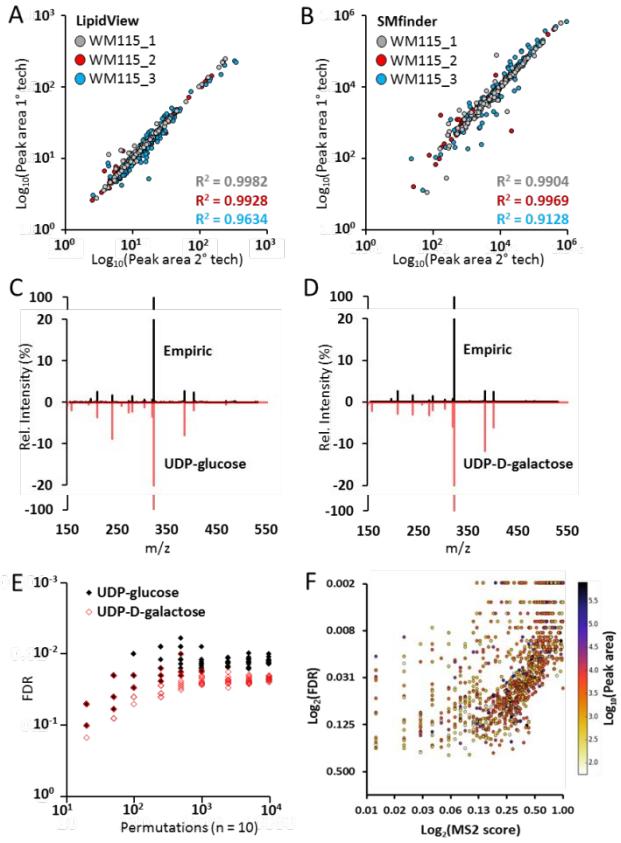

Figure 4 (A-B) Linear regression of 3 lipidomics samples acquired in tech-nical duplicates with integrated area calculated by LipidView and SMfinder respectively. Each dot represents the peak area of a lipid in the two technical replicates. (C-D) MS2 spectral comparison between empiric spectra with molecular formula matching for UPD-glucose and UDP-galactose manually imported from Metlin10 database. The two spectra are nearly identical with a minor difference in fragment peaks intensities. (E) Evaluation of FDR values against the number of permutations for the discrimination between UDP-glucose and UDP-galactose. The graph represents the distribution of empiric MS2 spectra in black, and of the MS2 spectra of UDP-glucose and galactose in red. (F) Distribution of MS2 score, FDR and peak area of the 1245 compounds generated upon data-filtering returning the best match for each annotatable feature from lipidomics analysis. The analysis was performed by using 500 permutations.

Applying the FDR, we were able to demonstrate that using a threshold of at least 1000 permutations per test, we could efficiently discriminate between the two compounds, supporting the validity of a permutation based FDR strategy to discriminate between false and positive identification. Next, we evaluated if our FDR method could be used as a stand-alone parameter for data annotation or if it is intrinsically mirroring the MS2 score. For this test, we ran an untargeted analysis on the lipidomics dataset from the melanoma cells and compared the correlation between MS2 score and FDR. We found that the two distributions are not statistically correlated.

Moreover, by comparing the abundance of the investigated compound, we observed that MS2 score tends to be more highly influenced by the compound abundance than the FDR, although both correlations are not significant (Fig 4F). Distribution of MS2 score, FDR and peak area of the 1245 compounds generated upon data-filtering returning the best match for each annotatable feature from lipidomics analysis. The analysis was performed by using 500 permutations.

Gap Filler

Often in untargeted analysis, not all features are detected in all samples. Compared with other omics disciplines, missing metabolites are less common due to the fact that the majority of metabolites are highly conserved within and between taxonomies, while major differences are found in metabolite concentration within the specimen of interest. Missing values can occur for two main reasons: either the compound is missing in a particular sample or the instrumental signal is instable, or too low, and is not selected for fragmentation, which occurs stochastically between different acquisitions (Fig. 5A). SMfinder includes a "Filler" function designed to search for missing peaks within the experimental dataset. To do so, a matrix containing all the compounds identified within the sample set is generated. The generated matrix is clustered by the index which was previously assigned to peaks during clusterization. For each compound, the result will be 3D array with masses, retention times and intensities for all the detected peaks (Fig. 5B). When the array is constituted by a subgroup of files, i.e. when a compound is not detected in the entire dataset, the array is used to calculate the mass range and retention time windows, and then used as a parameter to search for the missing peaks in the files in which those peaks were not detected. The parameters used for each peak i.e. m/z and retention time windows, will be assigned from their relative composite 3D array taking the maximum and minimum values for both variables. As example, we used the previous untargeted analysis of lipidomics samples, in which 3597 peaks were detected. For graphic representation purposes only, the areas of missing values were set to 1 (Fig. 5C). We then applied the "Filler" function and found that all the missing peaks were reintegrated

in the dataset (Fig. 5D). This is reasonable as we are analyzing three biological replicates from the same specimen against their technical counterparts. Moreover, the recomputed missing values could be coherently integrated with the expected distribution as observed by comparing the R-square values before and after the "filler" application.

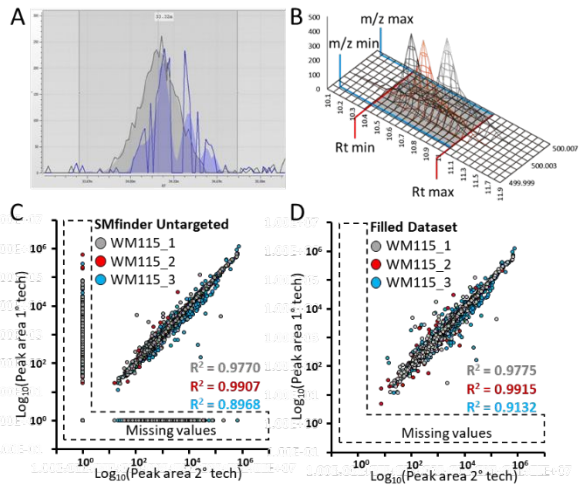

Figure 5 (A) Example of an undetected peak by the peak picker (blue) from the sample WM115_2 and the same peak from the corresponding technical replicate WM115_2r (gray). The peak was identified as triacyl-glycerol (TAG 58:8) in 5 out of 6 samples. (B) Graphic art representing how parameters are calculated to search missing values within the dataset. The composite matrix is constituted by the overlaying peaks with the same index from different files. The composite matrix is then used to calculate the mass range, i.e. the minimum and maximum mass values, and retention time range. (C) Linear regression of discovered features from untargeted analysis of 3 lipidomics samples; (D) features values recovered by the "Filler" function.

FDR performance on published datasets

To further test the performance of FDR, we downloaded from the MetaboLights repository ctwo distinct datasets generated in two different laboratories,. For both datasets, the original analysis was carried out in two steps by the respective authors[25,26]. First, the features detection was performed using MZmine, followed by compound identification by manual comparison of the MS2 spectra generated in house against commercial available standards. In our case, the analysis was performed using SMfinder and the HMDB as database, which contains MS2 spectra derived from more than 40 different mass spectrometers, and by selecting 500 permutations for FDR evaluation. Compounds without MS2 empiric spectra or without MS2 spectra in HMDB were excluded from the comparison. Identification performance is summarized in the Figure S3, which shows that the majority of metabolites were correctly assigned without the need of analyzing the standards. SMfinder also includes the MS2 evaluation also for targeted analysis. In this case, once the MS2 spectra of interest have been added to the library, the identification of the empiric spectra can be easily achieved by computing MS score and FDR for each sample in the batch. Those parameters are often missing when the comparison is performed manually. We provide here the identification of deuterated standards included in the lipidomics dataset as an example of this strategy. MS2 spectra, acquired in-house, were added into the SMfinder library. Identification confidence of standards in lipidomics samples, which were spiked-in prior to extraction and sample preparation, is reported in Table S4 showing an homogeneous MS score and FDR values along the entire dataset, indicating that standards were found within all the samples as expected. Moreover, this indicates that the detected peaks represent the spiked standards and not potential isobaric compounds/contaminations, which may occur in real samples.

Discussion

Herein, we demonstrate how SMfinder can be efficiently exploited for various types of small molecules analyses, including untargeted, targeted, and 13C trace analysis. Due to the flexible software architecture, the MS2 library of SMfinder is easily scalable by importing one or more available database in txt format, and by the possibility to efficiently mine large number of MS2 spectra from previously validated chromatograms such as those acquired with commercial standards or from published datasets. The introduction of the false discovery rate score based on single spectra permutation provides a novel identification parameter that can be efficiently applied for compound identification. FDR is not dependent on the size of the selected library, and can also be used for post-validation of targeted analysis. This is due to the computational architecture of the FDR in which, each possible match between MS2 spectra empiric or from libraries is computed independently. However, it is important to consider that large libraries are more accurate. As in the example of the characterization of UDP-glucose, both scores are below the FDR or 0.05. In this case, only the presence of multiple spectra from different isomers allows for the correct annotation of the compound of interest while a library without isomers may generate annotation errors. The software is fully embedded in intuitive GUI and does not require programming knowledge to be efficiently used. The GUI parameters are fully customizable and the setup can be stored or exported and later imported to efficiently trace back SMfinder parameters, for extensive automated batch analysis. SMfinder is executed in Python environment, and the entire source code, including the GUIs, is available without encryption upon installation. This gives proficient python users the possibility to fully customize SMfinder based on their own needs. SMfinder is capable of robustly identifying and quantifying large numbers of compounds from raw data, making it suitable for routine metabolomics and lipidomics studies.

ASSOCIATED CONTENT

Supporting Information

Supplementary figures S1-3.
Supplementary tables S1-4.

AUTHOR INFORMATION

Corresponding Author

* angela.bachi@ifom.eu

Present Address

[#]Institute of Neuroscience, CNR, 20129 Milan, Italy.

Author Contributions

G.M. and A.B. designed the research. G.M. developed the SMfinder program. M.L. and P.D. contributed to the improvement of SMfinder program. G.M. and V.M. analysed the samples. V.M. and A.C. prepared the samples for lipidomics analysis. G.M., V.M., M.M. and A.B. wrote the manuscript.

ACKNOWLEDGMENT

This work was supported by the Italian Association of Cancer Research (AIRC), grant number AIRC-IG-18607. The authors wish to thank Dr. Cinzia Villa for web support.

REFERENCES

(1) Collins, M. E.; Sweeney, S. R.; Tiziani, S. Isolation of Synergistic Natural Products Targeting Metabolic Dysfunction in Pediatric Pre-B Cell Acute Lymphoblastic Leukemia Using High-Throughput Screening and Metabolomics. *FASEB J.* **2019**, *33* (1_supplement), lb225–lb225.

(2) Tokarz, J.; Haid, M.; Cecil, A.; Prehn, C.; Artati, A.; Möller, G.; Adamski, J. Endocrinology Meets Metabolomics: Achievements, Pitfalls, and Challenges. *Trends Endocrinol. Metab.* **2017**, *28* (10), 705–721.

(3) Zamboni, N.; Saghatelian, A.; Patti, G. J. Defining the Metabolome: Size, Flux, and Regulation. *Mol. Cell* **2015**, *58* (4), 699–706.

(4) Aebersold, R.; Mann, M. Mass-Spectrometric Exploration of Proteome Structure and Function. *Nature* **2016**, *537* (7620), 347–355.

(5) Bateman, N. W.; Goulding, S. P.; Shulman, N. J.; Gadok, A. K.; Szumlinski, K. K.; MacCoss, M. J.; Wu, C. C. Maximizing Peptide Identification Events in Proteomic Workflows Using Data-Dependent Acquisition (DDA). *Mol. Cell. Proteomics* **2014**, *13* (1), 329–338.

(6) Schwudke, D.; Oegema, J.; Burton, L.; Entchev, E.; Hannich, J. T.; Ejsing, C. S.; Kurzchalia, T.; Shevchenko, A. Lipid Profiling by Multiple Precursor and Neutral Loss Scanning Driven by the Data-Dependent Acquisition. *Anal. Chem.* **2006**, *78* (2), 585–595.

(7) Xiao, J. F.; Zhou, B.; Ressom, H. W. Metabolite Identification and Quantitation in LC-MS/MS-Based Metabolomics. *TrAC Trends Anal. Chem.* **2012**, *32*, 1–14.

(8) Smith, C. A.; O'Maille, G.; Want, E. J.; Qin, C.; Trauger, S. A.; Brandon, T. R.; Custodio, D. E.; Abagyan, R.; Siuzdak, G. METLIN: A Metabolite Mass Spectral Database. *Ther. Drug Monit.* **2005**, *27* (6), 747–751.

(9) Ruttkies, C.; Schymanski, E. L.; Wolf, S.; Hollender, J.; Neumann, S. MetFrag Relaunched: Incorporating Strategies beyond in Silico Fragmentation. *J. Cheminform.* **2016**, *8* (1), 3.

(10) Tsugawa, H.; Kind, T.; Nakabayashi, R.; Yukihira, D.; Tanaka, W.; Cajka, T.; Saito, K.; Fiehn, O.; Arita, M. Hydrogen Rearrangement Rules: Computational MS/MS Fragmentation and Structure Elucidation Using MS-FINDER Software. *Anal. Chem.* **2016**, *88* (16), 7946–7958.

(11) Wang, Y.; Kora, G.; Bowen, B. P.; Pan, C. MIDAS: A Database-Searching Algorithm for Metabolite Identification in Metabolomics. *Anal. Chem.* **2014**, *86* (19), 9496–9503.

(12) Schrimpe-Rutledge, A. C.; Codreanu, S. G.; Sherrod, S. D.; McLean, J. A. Untargeted Metabolomics Strategies—Challenges and Emerging Directions. *J. Am. Soc. Mass Spectrom.* **2016**, *27* (12), 1897–1905.

(13) Cox, J.; Mann, M. MaxQuant Enables High Peptide Identification Rates, Individualized Ppb-Range Mass Accuracies and Proteome-Wide Protein Quantification. *Nat. Biotechnol.* **2008**, *26* (12), 1367–1372.

(14) Martano, G.; Delmotte, N.; Kiefer, P.; Christen, P.; Kentner, D.; Bumann, D.; Vorholt, J. A. Fast Sampling Method for Mammalian Cell Metabolic Analyses Using Liquid Chromatography-Mass Spectrometry. *Nat. Protoc.* **2015**, *10* (1). https://doi.org/10.1038/nprot.2014.198.

(15) Folch, J.; Lees, M.; Stanley, G. H. S. A Simple Method for the Isolation and Purification of Total Lipides from Animal Tissues. *J. Biol. Chem.* **1957**, *226* (1), 497–509.

(16) Van Rossum, G. Python Programming Language. In *USENIX annual technical conference*; 2007; Vol. 41, p 36.

(17) Ihaka, R.; Gentleman, R. R: A Language for Data Analysis and Graphics. *J. Comput. Graph. Stat.* **1996**, *5* (3), 299–314.

(18) Kiefer, P.; Schmitt, U.; Vorholt, J. A. EMZed: An Open Source Framework in Python for Rapid and Interactive Development of LC/MS Data Analysis Workflows. *Bioinformatics* **2013**, *29* (7), 963–964.

(19) Wishart, D. S.; Feunang, Y. D.; Marcu, A.; Guo, A. C.; Liang, K.; Vázquez-Fresno, R.; Sajed, T.; Johnson, D.; Li, C.; Karu, N. HMDB 4.0: The Human Metabolome Database for 2018. *Nucleic Acids Res.* **2018**, *46* (D1), D608–D617.

(20) Adusumilli, R.; Mallick, P. Data Conversion with ProteoWizard MsConvert. In *Proteomics*; Springer, 2017; pp 339–368.

(21) Chaleckis, R.; Murakami, I.; Takada, J.; Kondoh, H.; Yanagida, M. Individual Variability in Human Blood Metabolites Identifies Age-Related Differences. *Proc. Natl. Acad. Sci.* **2016**, *113* (16), 4252–4259.

(22) Li, Z.; Lai, Z. W.; Christiano, R.; Gazos-Lopes, F.; Walther, T. C.; Farese, R. V. Global Analyses of Selective Insulin Resistance in Hepatocytes Caused by Palmitate Lipotoxicity. *Mol. Cell. Proteomics* **2018**, *17* (5), 836–849.

(23) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Anal. Chem.* **2006**, *78* (3), 779–787.

(24) Kim, S.; Thiessen, P. A.; Bolton, E. E.; Chen, J.; Fu, G.; Gindulyte, A.; Han, L.; He, J.; He, S.; Shoemaker, B. A. PubChem Substance and Compound Databases. *Nucleic Acids Res.* **2016**, *44* (D1), D1202–D1213.

(25) Pluskal, T.; Hayashi, T.; Saitoh, S.; Fujisawa, A.; Yanagida, M. Specific Biomarkers for Stochastic Division Patterns and Starvation-induced Quiescence under Limited Glucose Levels in Fission Yeast. *FEBS J.* **2011**, *278* (8), 1299–1315.

(26) Teruya, T.; Chaleckis, R.; Takada, J.; Yanagida, M.; Kondoh, H. Diverse Metabolic Reactions Activated during 58-Hr Fasting Are Revealed by Non-Targeted Metabolomic Analysis of Human Blood. *Sci. Rep.* **2019**, *9* (1), 1–11.