Published in final edited form as:

Anal Chem. 2017 January 17; 89(2): 1063–1067. doi:10.1021/acs.analchem.6b01214.

xMSannotator: an R package for network-based annotation of high-resolution metabolomics data

Karan Uppal¹, Douglas I. Walker^{1,2}, and Dean P. Jones^{1,*}

¹Clinical Biomarkers Laboratory, Department of Medicine, Emory University, Atlanta, Georgia 30308, United States

²Department of Civil and Environmental Engineering, Tufts University, Medford, Massachusetts 02153, United States

Abstract

Improved analytical technologies and data extraction algorithms enable detection of >10,000 reproducible signals by liquid chromatography high-resolution mass spectrometry, creating a bottleneck in chemical identification. In principle, measurement of more than one million chemicals would be possible if algorithms were available to facilitate utilization of the raw mass spectrometry data, especially low abundance metabolites. Here we describe an automated computational framework to annotate ions for possible chemical identity using a multistage clustering algorithm in which metabolic pathway associations are used along with intensity profiles, retention time characteristics, mass defect, and isotope/adduct patterns. The algorithm uses high-resolution mass spectrometry data for a series of samples with common properties and publicly available chemical, metabolic and environmental databases to assign confidence levels to annotation results. Evaluation results show that the algorithm achieves an F1-measure of 0.8 for a dataset with known targets and is more robust than previously reported results for cases when database size is much greater than the actual number of metabolites. MS/MS evaluation of a set of randomly selected 210 metabolites annotated using xMSannotator in an untargeted metabolomics human dataset shows that 80% of features with high or medium confidence scores have ion dissociation patterns consistent with the xMSannotator annotation. The algorithm has been incorporated into an R package, xMSannotator, which includes utilities for querying local or online databases such as ChemSpider, KEGG, HMDB, T3DB, and LipidMaps.

Graphical abstract

Corresponding Author: Dr. Dean P. Jones; dpjones@emory.edu; Address for all authors: 615 Michael St., Whitehead Bldg 205P, Atlanta, GA 30322.

Author Contributions: KU and DPJ conceived and coordinated the software design; KU developed the software with advice from DIW and DPJ; DIW generated and analyzed the MS/MS data; KU drafted the manuscript with contributions from DIW and DPJ. All authors read and approved the final manuscript.

Supporting Information Available: This material is available free of charge via the Internet at <http://pubs.acs.org>.

Keywords

Metabolomics; Annotation; Networks; Clustering; High-Resolution Mass Spectrometry

Untargeted metabolomics refers to global profiling of thousands of small molecules in biological samples without any selection bias.¹⁻³ Numerous technological and computational advancements over the last decade have significantly improved the coverage of the metabolome.⁴⁻⁹ However, the resulting increased data complexity has introduced new challenges, especially related to metabolite identification.⁸ Simple database searches for annotation using only the mass-to-charge ratio (m/z) and suspected adduct form information can result in a large number of false positives as a single m/z can match multiple metabolites.⁹ Additional experimental procedures such as MS/MS and confirmation using reference standards are therefore needed for metabolite identification.^{11,12} This is often a laborious process and could be a waste of valuable resources when database matches are pursued for validation simply based on m/z matches. Moreover, experimental validation may not be feasible for metabolites with no commercially available standard or low abundance signals.

As recently reviewed by Uppal et al.¹, several computational approaches for metabolite annotation have been developed that utilize m/z , elution characteristics, correlation, adduct, and isotopic patterns to reduce the number of false positives.¹³⁻¹⁸ These methods require presence of multiple adducts or isotopes to establish confidence in possible chemical identity¹⁶⁻¹⁹; however, metabolites are not always detected as multiple features.²⁰ In our previous work, we have shown that inclusion of intensity-based network structure and incorporating pathway level information can provide new criteria to improve confidence in prediction, even when multiple peaks are not detected for a metabolite.²¹ Use of network and pathway level criteria in addition to existing principles along with incorporation of rules such as abundance ratio checks for isotopes, multiply charged forms and multimers to assign confidence scores to database matches, can make the metabolite annotation and identification process more efficient.^{1,20,22,23} The terms “annotation” and “identification” are used here according to previously defined guidelines, where identification refers to

identity confirmation using at least two independent measures of reference standards (e.g. accurate mass and retention time), and annotation refers to tentative matches in databases based on spectral similarity and/or matches based on physicochemical properties without using authentic standards.^{11,12,17}

Here we present a freely available R package, xMSannotator, which incorporates several utilities and an integrative multi-criteria scoring algorithm for improving annotation of high-resolution metabolomics data. The main purpose of the software is to facilitate metabolite identification for untargeted LC/MS data by categorizing annotations based on database matches into different confidence levels, thereby allowing prioritization of metabolites for further validation efforts. xMSannotator builds on existing principles (correlation, coelution) to identify features related to a metabolite, but also includes unique features such as pathway level correlations to enhance confidence level when multiple adduct or isotope forms of a metabolite are not detected. The software takes as input a peak intensity table (a matrix of m/z , time, intensities across samples) and uses a multi-criteria scoring algorithm to automatically associate ions detected by mass spectrometry to known chemicals without using MS/MS. The software uses KEGG²⁴, HMDB²⁵, Toxin and Toxin Target Database (T3DB)²⁶⁻²⁷ and ChemSpider²⁸ for annotation. The R package along with sample input and output files and a user manual is available at: <https://sourceforge.net/projects/xmsannotator/>. The key contribution of xMSannotator is its ability to incorporate both analytical and biological correlations to categorize otherwise thousands of database matches into different levels of confidence for annotation, to filter incorrect matches, and to enhance biological interpretation of untargeted metabolomics data by categorizing related metabolites/features into same network modules as described in the following sections.

Experimental Section

xMSannotator integrative scoring algorithm

xMSannotator uses a multi-step strategy for annotation as described below (Figure 1):

1. Correlation analysis: Pairwise correlation analysis among all measured m/z features in the input peak intensity table, where an m/z feature is a combination of m/z , retention time (RT), and associated intensity, is performed across all samples. Users have the option to use Pearson or Spearman correlation.
2. Network modularity analysis: Data-driven network modules are defined using the WGCNA method.²⁹ Briefly, the algorithm uses the pairwise correlation matrix from step 1 to generate an adjacency matrix, A_{ij} . In the next step, a topological overlap based dissimilarity measure is calculated using the adjacency matrix, which is then used for hierarchical clustering analysis to identify modules of co-expressing m/z features. Each module, M_k , comprises of m/z features that are tightly connected to each other.³⁰ This approach is more robust than simple correlation analysis, which is sensitive to the choice of correlation threshold and could lead to information loss.³⁰ The module membership information is used to filter incorrect matches as described in step 6.

3. Retention-time based clustering: Within each module M_k , kernel density estimation technique is used to identify sub-modules of co-eluting m/z features, M_{kt} .
4. Mass defect analysis: Features within each sub-module, M_{kt} , are grouped based on mass defect, which is calculated from the difference between the measured experimental mass and nominal mass, to identify groups of features that follow an isotopic pattern (+1, +2) or potential adducts or in-source fragments of a metabolite.^{31,32}
5. Database matching: After the features have been grouped in steps 2-4, the m/z features within each sub-module M_{kt} are matched against known metabolites in chemical databases (HMDB, KEGG, T3DB, or LipidMaps) according to user-defined adduct rules and mass search tolerance in ppm.
6. Score assignment: After database matching, a score is assigned to every matching metabolite according to equation 1 if the following conditions are satisfied:
 - I. Do the m/z features matching different adducts/isotopes of a metabolite belong to the same sub-module m_{kt} from steps 1-4? This criterion reduces the risk of incorrect annotation, as it requires that features associated with a metabolite are tightly connected with each other based on their intensity profiles across samples and also have similar retention times.
 - II. If the m/z features are in the same sub-module M_{kt} , are the features also positively correlated with each other? Default threshold: 0.7
 - III. If the m/z features are in the same sub-module M_{kt} and are positively correlated, is the RT range (max-min) across features matching this metabolite within a defined threshold?

If all three conditions are satisfied, a non-zero score is assigned to metabolite according to the following equation (Eq. 1) that accounts for correlation strength, difference in retention time, number of matching adducts and isotopes, and weights assigned to more probable adducts and isotopes:

$$\text{Score} = (\text{isotope weight}) * (\text{correlation}) * (1 / (\text{RT range})^{\text{sigma}}) * (\# \text{ of matching adducts} + \# \text{ of isotopes}) * 10 * (\sum (\text{adduct weight})),$$

(1)

where,

$$\text{sigma} = \{ 1 \text{ if RT range} < \text{RT threshold}, 10 \text{ otherwise} \}$$

adduct weight = By default, a weight of 1 is assigned to all adduct rules, but higher weight could be assigned to more probable adducts (e.g. 10 for M+H) to reduce the risk of false matching isotope weight = { 100 if a C¹³ or an expected Cl, Br or S is present based on the chemical formula; 1 otherwise }

7. Incorporating biological information: Not all metabolites generate multiple features. A unique feature of the algorithm is that it uses the network modules from step 2 to perform pathway level correlation analysis to enhance confidence in metabolites associated with single features or that were assigned a score of 0 in step 6. Specifically, the score of a metabolite is boosted if there are other metabolites from the same pathway assigned to the same network module M_k , have score greater than 0 in step 6, and have matches for user-defined adducts/ions (eg: M+H). Different scores are assigned to isomers only if they are associated with different pathways.
8. Confidence level assignment: Each chemical is categorized as no confidence (0), low confidence (1), medium confidence, (2) and high confidence (3) based on the following criteria:
 - I. Score is greater than zero
 - II. Presence of required adducts/forms specified by the user for assignment to high confidence categories (e.g. M+H)
 - III. N, O, P, S/C ratio checks²²
 - IV. Hydrogen/Carbon ratio check²²
 - V. Abundance ratio checks for isotopes, multimers (dimers and trimers), and multiply charged adducts with respect to the singly charged adducts and ions according to heuristic rules.

A high confidence match satisfies all criteria, medium confidence is assigned based on the pathway level correlation in step 7, low confidence matches have score greater than 0, but do not satisfy the elemental ratio or abundance ratio checks. No confidence matches have score ≤ 0 ; this should be interpreted as an identity level of 5 as defined by Schymanski et al.¹²

9. Redundancy evaluation and correction: The score assigned to each chemical in steps 6 and 7 is used to reduce the redundancy in matches by assigning the m/z with multiple hits to the chemical with higher overall score.

The individual functions are described in Supplementary Information.

Experimental evaluation

The performance of the scoring algorithm was evaluated using publicly available datasets.

Experiment 1: A standard mixture dataset of 104 metabolites (std1.POS) was used to evaluate the performance of the algorithm on a dataset with known metabolites.¹⁷ F1-measure, which is a harmonic mean of precision (TP/(TP+FP)) and recall (TP/(TP+FN)) was used to evaluate the performance. The recall is based on the 92 detected

metabolites that were also found in HMDB. Sensitivity analysis was performed to evaluate the variation in F1-measure at different parameter settings such as retention time threshold, correlation threshold, and database size at m/z search tolerance of 5 ppm and 10 ppm.

Experiment 2: A plasma dataset from 50 healthy humans from an ongoing healthy aging study (Metabolomics Workbench Study-ID=ST000163) with 15,148 features was used to test the performance of the algorithm for annotation of untargeted metabolomics datasets.^{21,33} The data processing and experimental details are provided in Supplementary Information. The experimental conditions for both LC-MS and LC-MS/MS were identical to the original study.³³

Experiment 3: A plasma dataset from 50 common marmosets selected from ongoing aging studies (Metabolomics Workbench StudyID=ST000163) with 5,852 features was used to evaluate the performance of xMSannotator for targeted annotation or suspect screening of ketamine, a commonly used anesthetic in non-human primates, and its metabolites.²¹ Additional data processing and experimental details are provided in Supplementary Information.

Results and Discussion

For the standard mixture dataset, out of the 88 total matches, 65 were assigned as high confidence and 13 as medium confidence (84.7%) (Table S-2). For the high confidence matches, m/z features associated with the same metabolite were assigned to the same network module and retention time cluster (Module_RTclust), and satisfied all criteria in step 8 of the scoring algorithm. The medium confidence matches were assigned based on network and pathway level associations as the standard mixture included endogenous metabolites. As shown in Figure S-1, the F1-measure for the high confidence matches varied from 0.69 to 0.8 and 0.68 to 0.81 at different parameter settings using m/z threshold of 5 and 10 ppm, respectively. The variation in performance as a result of change in database size was greater than variation due to change in other parameters such as correlation threshold, retention time threshold, and m/z search threshold. For instance, the F1 measure only varied by +/- 0.2 as the retention time threshold was varied from 3-10 seconds at m/z threshold of 5ppm (Figure S-1A). On average, 54% of the database matches (40%-65% for different database sizes) were filtered.

Overall, the F1 scores were better as compared to previously reported results by Daly et al.¹⁷ (MetAssign=0.5, CAMERA=0.27, mzMatch=0.13) for the same dataset even at database size of 1000. The performance of the integrative multi-criteria approach was also evaluated against annotation using single criteria at m/z search threshold of 5 ppm and database size of 1000. F1 measures of 0.52 and 0.65 were achieved using only presence of two or more adducts and one or more isotopes as criteria for annotation, respectively. These values were lower than the score achieved using the integrative scoring algorithm, F1=0.72. These results suggest that the integrative multi-criteria clustering is more stable and reduces the risk of incorrect annotations.

For experiment 2, 709 m/z features were annotated using a database of 3,075 HMDB metabolites as described in Supplementary Information. Out of the 210 medium-to-high confidence matches that were randomly selected for annotation by MSMS, 129 (61%) were detected in the pooled plasma sample (Table S3). The remaining 81 matches were either not detected or the abundance level was too low for MS/MS. Out of the 129 detected metabolites, annotation was consistent with the xMSannotator results for 103 (80%) metabolites based upon confidence levels 1-3, as described by Schymanski et al.¹² (Table S-3).

The same dataset from Experiment 2 was also annotated using all 41,514 metabolites in HMDB (Figure 2). 2,846 m/z features were annotated. As shown in Figure 2A, the algorithm dramatically decreases the number of false matches (>67%) and allows prioritization of metabolites for further validation by assigning them into different confidence levels based on the multi-criteria algorithm. The algorithm uses the correlation-based network information and biological information to enhance confidence in metabolites with only single adduct matches (Figure 2B). In this case, a subset of features in module 28 matched to metabolites involved in Phenylalanine metabolism pathway illustrating how network modularity analysis also helps with biological interpretation. The network and pathway level connectivity was used to assign these matches to medium confidence level.

We also evaluated abundance levels of base peaks across different confidence levels (Figure S-2). The results show that overall the abundance levels are higher in high confidence group compared to other groups. Interestingly, the abundance levels ranged from 10^3 to 10^8 even for the high confidence group indicating that the multi-criteria clustering approach works for low abundance metabolites as well.

For experiment 3, ketamine, norketamine, and (4-, 5-, or 6-) hydroxyketamine were classified as high confidence matches based on the multi-criteria scoring algorithm (Figure S-3A). The algorithm identified m/z features corresponding to M+Na and M+H-H₂O adducts of ketamine, M+23 and M-17, respectively, and their isotopes. Similarly, hydroxyketamine and norketamine were classified as high confidence, but the algorithm could not differentiate between isomers of hydroxyketamine. The identity of ketamine was confirmed using MS/MS (Figure S-3B).

The evaluation results for the show that xMSannotator reduces the number of false matches and can be used for prioritization of metabolites for further validation using MS/MS and reference standards as well as suspect screening of drugs or other chemical exposures. In addition, isotopic and mass defect information can be used to improve prediction accuracy of chemical structure, molecule functional groups and presence of homologous series. Limitations and future work: Currently, unless discriminated by pathway associations, the algorithm cannot distinguish between isomers or metabolites with the same chemical formula. Also, even though the multi-level scoring scheme reduces the number of false positives as compared to simple m/z and single criteria based annotation; additional work is needed to improve the precision of the algorithm. Future work will focus on addressing these limitations and extending the functionality of the algorithm for identification of biotransformations.

Conclusion

xMSannotator facilitates metabolite identification by using an integrative multi-criteria scoring algorithm to categorize database matches into different confidence levels. The software allows prioritization of metabolites for confirmation using MS/MS and reference standards. The evaluation results show that the multi-criteria scoring algorithm is more robust than other computational approaches when the database size is much greater than the true number of metabolites. The results show that xMSannotator filters incorrect matches and allows biological interpretation of untargeted metabolomics data by not only grouping related features, but also linking them with metabolite names and organizing them into correlation-based network modules. The package can be used in both a suspect screening and untargeted annotation framework, and is compatible with peak intensity tables generated using any data extraction software that provides m/z, time, and intensity information across all samples.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors want to thank Dr. Frederick Strobel for critical input during algorithm development. This project was funded in part by federal funds from the US National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services under contract # HHSN272201200031C. This research was also supported by National Institutes of Health award numbers ES023485, HL113451, AG038746, ES019776, P01 HL 086773.

References

1. Uppal K, Walker DI, Liu K, Li S, Go YM, Jones DP. Computational Metabolomics: A Framework for the Million Metabolome. *Chem Res Toxicol*. 2016
2. Patti GJ, Yanes O, Siuzdak G. Innovation: Metabolomics: the apogee of the omics trilogy. *Nat Rev Mol Cell Biol*. 2012; 13(4):263–9. [PubMed: 22436749]
3. Walker, DI., Go, YM., Liu, K., Pennell, K., Jones, DP. Population Screening for Biological and Environmental Properties of the Human Metabolic Phenotype: Implications for Personalized Medicine. Vol. 7. Elsevier; Amsterdam, The Netherlands: 2016.
4. Tautenhahn R, Bottcher C, Neumann S. Highly sensitive feature detection for high resolution LC/MS. *BMC bioinformatics*. 2008; 9:504. [PubMed: 19040729]
5. Scalbert A, Brennan L, Fiehn O, Hankemeier T, Kristal BS, van Ommen B, Pujos-Guillot E, Verheij E, Wishart D, Wopereis S. Mass-spectrometry-based metabolomics: limitations and recommendations for future progress with particular focus on nutrition research. *Metabolomics*. 2009; 5(4):435–458. [PubMed: 20046865]
6. Yu T, Park Y, Johnson JM, Jones DP. apLCMS---adaptive processing of high-resolution LC/MS data. *Bioinformatics*. 2009; 25(15):1930–6. [PubMed: 19414529]
7. Johnson JM, Yu T, Strobel FH, Jones DP. A practical approach to detect unique metabolic patterns for personalized medicine. *The Analyst*. 2010; 135(11):2864–70. [PubMed: 20838665]
8. Dunn WB, Broadhurst D, Begley P, Zelena E, Francis-McIntyre S, Anderson N, Brown M, Knowles JD, Halsall A, Haselden JN, Nicholls AW, Wilson ID, Kell DB, Goodacre R, Human Serum Metabolome C. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nat Protoc*. 2011; 6(7): 1060–83. [PubMed: 21720319]

9. Uppal K, Soltow QA, Strobel FH, Pittard WS, Gernert KM, Yu T, Jones DP. xMSAnalyzer: automated pipeline for improved feature detection and downstream analysis of large-scale, non-targeted metabolomics data. *BMC bioinformatics*. 2013; 14:15. [PubMed: 23323971]
10. Kind T, Fiehn O. Metabolomic database annotations via query of elemental compositions: mass accuracy is insufficient even at less than 1 ppm. *BMC bioinformatics*. 2006; 7:234. [PubMed: 16646969]
11. Sumner LW, Amberg A, Barrett D, Beale MH, Beger R, Daykin CA, Fan TW, Fiehn O, Goodacre R, Griffin JL, Hankemeier T, Hardy N, Harnly J, Higashi R, Kopka J, Lane AN, Lindon JC, Marriott P, Nicholls AW, Reily MD, Thaden JJ, Viant MR. Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics*. 2007; 3(3):211–221. [PubMed: 24039616]
12. Schymanski EL, Jeon J, Gulde R, Fenner K, Ruff M, Singer HP, Hollender J. Identifying small molecules via high resolution mass spectrometry: communicating confidence. *Environmental science & technology*. 2014; 48(4):2097–8. [PubMed: 24476540]
13. Rogers S, Scheltema RA, Girolami M, Breitling R. Probabilistic assignment of formulas to mass peaks in metabolomics experiments. *Bioinformatics*. 2009; 25(4):512–8. [PubMed: 19095699]
14. Brown M, Dunn WB, Dobson P, Patel Y, Winder CL, Francis-McIntyre S, Begley P, Carroll K, Broadhurst D, Tseng A, Swainston N, Spasic I, Goodacre R, Kell DB. Mass spectrometry tools and metabolite-specific databases for molecular identification in metabolomics. *The Analyst*. 2009; 134(7):1322–32. [PubMed: 19562197]
15. Alonso A, Julia A, Beltran A, Vinaixa M, Diaz M, Ibanez L, Correig X, Marsal S. AStream: an R package for annotating LC/MS metabolomic data. *Bioinformatics*. 2011; 27(9):1339–40. [PubMed: 21414990]
16. Kuhl C, Tautenhahn R, Bottcher C, Larson TR, Neumann S. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Analytical chemistry*. 2012; 84(1):283–9. [PubMed: 22111785]
17. Daly R, Rogers S, Wandy J, Jankevics A, Burgess KE, Breitling R. MetAssign: probabilistic annotation of metabolites from LC-MS data using a Bayesian clustering approach. *Bioinformatics*. 2014; 30(19):2764–71. [PubMed: 24916385]
18. Silva RR, Jourdan F, Salvanha DM, Letisse F, Jamin EL, Guidetti-Gonzalez S, Labate CA, Vencio RZ. ProbMetab: an R package for Bayesian probabilistic annotation of LC-MS-based metabolomics. *Bioinformatics*. 2014; 30(9):1336–7. [PubMed: 24443383]
19. Broeckling CD, Afsar FA, Neumann S, Ben-Hur A, Prenni JE. RAMClust: a novel feature clustering method enables spectral-matching-based annotation for metabolomics data. *Analytical chemistry*. 2014; 86(14):6812–7. [PubMed: 24927477]
20. Dunn WB, Erban A, Weber RJM, Creek DJ, Brown M, Breitling R, Hankemeier T, Goodacre R, Neumann S, Kopka J, Viant MR. Mass appeal: Metabolite identification in mass spectrometry-focused untargeted metabolomics. *Metabolomics*. 2013; 9:S44–S66.
21. Uppal K, Soltow QA, Promislow DE, Wachtman LM, Quyyumi AA, Jones DP. MetabNet: An R Package for Metabolic Association Analysis of High-Resolution Metabolomics Data. *Front Bioeng Biotechnol*. 2015; 3:87. [PubMed: 26125020]
22. Kind T, Fiehn O. Seven Golden Rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC bioinformatics*. 2007; 8:105. [PubMed: 17389044]
23. Li S, Park Y, Duraisingham S, Strobel FH, Khan N, Soltow QA, Jones DP, Pulendran B. Predicting network activity from high throughput metabolomics. *PLoS computational biology*. 2013; 9(7):e1003123. [PubMed: 23861661]
24. Kanehisa M. The KEGG database. *Novartis Found Symp*. 2002; 247:91–101. discussion 101-3, 119-28, 244-52. [PubMed: 12539951]
25. Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, Liu Y, Djombou Y, Mandal R, Aziat F, Dong E, Bouatra S, Sinelnikov I, Arndt D, Xia J, Liu P, Yallou F, Bjorn Dahl T, Perez-Pineiro R, Eisner R, Allen F, Neveu V, Greiner R, Scalbert A. HMDB 3.0--The Human Metabolome Database in 2013. *Nucleic acids research*. 2013; 41(Database issue):D801–7. [PubMed: 23161693]

26. Lim E, Pon A, Djoumbou Y, Knox C, Shrivastava S, Guo AC, Neveu V, Wishart DS. T3DB: a comprehensively annotated database of common toxins and their targets. *Nucleic acids research*. 2010; 38(Database issue):D781–6. [PubMed: 19897546]
27. Wishart D, Arndt D, Pon A, Sajed T, Guo AC, Djoumbou Y, Knox C, Wilson M, Liang Y, Grant J, Liu Y, Goldansaz SA, Rappaport SM. T3DB: the toxic exposome database. *Nucleic acids research*. 2015; 43(Database issue):D928–34. [PubMed: 25378312]
28. Pence HE, Williams A. ChemSpider: An Online Chemical Information Resource. *J Chem Educ*. 2010; 87:1123–1124.
29. Langfelder P, Horvath S. WGCNA: an R package for weighted correlation network analysis. *BMC bioinformatics*. 2008; 9:559. [PubMed: 19114008]
30. Zhang B, Horvath S. A general framework for weighted gene co-expression network analysis. *Stat Appl Genet Mol Biol*. 2005; 4 Article17.
31. Zhang H, Zhang D, Ray K, Zhu M. Mass defect filter technique and its applications to drug metabolite identification by high-resolution mass spectrometry. *J Mass Spectrom*. 2009; 44(7): 999–1016. [PubMed: 19598168]
32. Xu YF, Lu W, Rabinowitz JD. Avoiding misannotation of in-source fragmentation products as cellular metabolites in liquid chromatography-mass spectrometry-based metabolomics. *Analytical chemistry*. 2015; 87(4):2273–81. [PubMed: 25591916]
33. Go YM, Walker DI, Liang Y, Uppal K, Soltow QA, Tran V, Strobel F, Quyyumi AA, Ziegler TR, Pennell KD, Miller GW, Jones DP. Reference Standardization for Mass Spectrometry and High-resolution Metabolomics Applications to Exposome Research. *Toxicological sciences : an official journal of the Society of Toxicology*. 2015; 148(2):531. [PubMed: 26358001]

Step 1: Global network modularity analysis **Figure 1.**

xMSannotator integrative scoring algorithm. xMSannotator uses a multi-stage scoring algorithm to assign database matches into different confidence levels of annotation. In step one, data-driven network modules are derived using the intensities across all samples. In step two, each module, M_k , is further sub-grouped based on retention time, M_{kt} . In steps three and four, isotopes and adduct patterns are used for database matching within each sub-module, M_{kt} . In step five, pathway information from KEGG or HMDB is incorporated to enrich confidence level of low confidence matches if there is a high confidence match in the same pathway and module, M_k . High confidence match: all criteria in steps 1-4 are satisfied; Medium confidence match: step 5 criteria is satisfied.

Figure 2.

xMSannotator reduces the number of incorrect matches and assigns confidence level to each database match. A. Distribution of database matches based on different confidence levels. The algorithm filtered over 67% of matches and identified 533 high confidence and 10,704 medium confidence matches. B. Illustration of how the data-driven network information is integrated with pathway information using Phenylalanine and its metabolites as an example. These matches were assigned to the same network module (module 28). 12 metabolites with single adduct matches were assigned to medium confidence level based on the pathway and network level associations, and 5 were assigned to no confidence level.