Supporting Information

Jane Doe[1,2] John Q. Doe[2]

[1] Acme Corporation
[2] Federation of Planets

Contents

1 Other Algorithms details **1**
1.1 ABC selection . . . 1
1.1.1 Principle . . . 1
1.1.2 Discussion . . . 2
1.2 Network graph presentation . . . 2
1.3 Visualization system . . . 2
1.4 Statistical analysis . . . 2
1.5 Feature detection . . . 2
1.6 Data structure . . . 3
1.7 Reporting system . . . 3
1.8 Code Compatibility . . . 3

2 Other Experimental details **3**
2.1 Evaluation of MCnebula . . . 3
2.1.1 Serum dataset . . . 4
2.1.2 Herbal dataset . . . 4
2.1.3 Data processing . . . 5
2.1.4 Data and code availability . . . 5

3 Figure legend of supplementation **5**

Reference **8**

1 Other Algorithms details

1.1 ABC selection

1.1.1 Principle

The principle of ABC selection algorithm: (1) applied an initial filtering to thousands of chemical classes based on the predicted probability, (2) regarded all ‘features’ as a whole, examined the number and abundance of ‘features’ of each chemical classification (classification at different levels, classification of sub-structure and dominant structure), and then selected representative classes, (3) these chemical classes were followed by goodness assessment (about identification of its classified compounds) and identicality assessment (the extent to which these chemical classes are distinguished from each other in the context of MS/MS spectra). The final chemical classes would serve to the subsequent analysis: visualized as Child-Nebulae and focus on these chemical classes (or Nebulae) for biomarker or chemistry discovery. The top ‘features’ based on statistical analysis could be set as tracer to discover more homology compounds of chemical structure or spectral similarity or chemical class.

1.1.2 Discussion

MCnebula could assist researchers in focusing on potential markers or interesting compounds quickly by combining full-spectrum identification with machine prediction, visualization of sub-nebulae in a multi-dimensional view, and statistical analysis to track top 'features' and find analogues. The ABC selection algorithm can summarize a representative chemical class in a dataset and obtain the features to that class, so the overall direction of the study is unbiased. Also, it is an effective guarantee for statistical analysis to produce top features for tracing analysis in next step. The results of statistical analysis based on feature level may cause bias since the loss of information, filtering on the basis of chemical classes level can prevent the bias in some degree. The Child-Nebula, which mapped on the basis of the chemical classes obtained by the ABC selection algorithm, achieved the goal of visualizing the huge untargeted dataset as a single graph. Above all, the parameters of the ABC selection algorithm were subjectively adjustable and they should be determined according to the richness of the chemical class of the studied object. In general, our default parameters used to acquire the chemical classes that are abundant in variety according to the datasets and filtered out those that were too large or too small classes in conceptual scope.

1.2 Network graph presentation

The features and their annotations are integrated as Nebulae based on the Nebula-Index. These Nebulae are represented as network-type graph data. The feature annotation data includes top candidates for chemical formula and structure. The MS/MS spectral similarity of the features is calculated and used to generate the edge data for the network graph.

1.3 Visualization system

MCnebula integrates various R packages to format data, including the 'ggplot2' package. To make visualization easier for users, we developed the 'ggset' data class, which stores pre-defined ggplot2 plotting functions and parameters for visualizing Nebulae. Users can customize the visualization according to their specific needs or the requirements of the publisher.

1.4 Statistical analysis

MCnebula integrates the functions of the 'limma' package for differential expression analysis of RNA-sequence and microarray data[1], and package them for differential analysis of metabolomics data. The gene expression matrix and feature quantification matrix of LC-MS are similar, both have phenotypic variables (sample information) and dependent variables (gene expression or feature quantification values). Our method can be appropriate for statistical analysis of feature quantification of experimental designs in which explanatory variables are factorial variables and the design matrix is without an intercept[2].

Wozniak et al. used a joint approach of Ensemble Feature Selection (EFS) and Mann-Whitney U (MWU) tests to screen top metabolites[3]. When we compared the 50 top 'features' obtained by the 'binary comparison' method integrated in MCnebula with the top 50 metabolites (top 50 of EFS and 50 of MWU) obtained by the joint method of W et al., a total of 37 overlapped metabolites were screened out, including the key metabolite of L-Thyroxine in the reference study. Top 'features' were usually different according to the feature selection algorithm. The reliability of the 'binary comparison' method was verified again by our ranked results comparing with those of Wozniak et al.

1.5 Feature detection

Feature detection is a kind of algorithm for detecting peaks from mass data file, and most mass spectrometry processing tools have a similar function. Users can implement this process with any tool, but to access the MCnebula workflow, .mgf (long list file containing MS/MS information) and .csv files (or other formatted table file of feature quantification) were required for output. In this study, all processing of Feature detection were implemented in MZmine2 (version 2.53). But now, the R package MCnebula2 has provided some convenient tools which integrated XCMS (R package) methods or functions for Feature detection. See details in: https://mcnebula.org/.

1.6 Data structure

MCnebula was primarily developed using the R S4 system of object-oriented programming. All data including 'features' annotation data and visualization data is stored in a single object (class 'mcnebula'), which simplifies the application, makes data management and analysis easier to perform and repeat.

1.7 Reporting system

MCnebula includes a reporting system that enables the analysis process to be output as a PDF document or in other formats. The reporting system is based on the 'report' data class, which stores each step of the analysis as a section and can be easily modified according to the user's requirements. Furthermore, the 'rmarkdown' R package[4] is incorporated in the reporting system to generate reports.

1.8 Code Compatibility

MCnebula performs downstream analysis by extracting data from the pre-computed SIRIUS project, which is the primary data source for MCnebula 2. The SIRIUS software is continually updated and enhanced. From SIRIUS version 4 to version 5 (https://github.com/boecker-lab/sirius), the data structure and attribute names in the project directory have been modified. To ensure that MCnebula is not affected by version problems, we have designed its application programming interface (API) for various SIRIUS versions.

2 Other Experimental details

2.1 Evaluation of MCnebula

The details of simulation of noise were as following:

- A global mass shift was simulated by drawing a random number δ^* from $N\left(0, \sigma_{mb}^2\right)$ (Normal distribution) and then shifting every peak mass m by $\delta^* m$. The standard deviation σ_{mb} was chosen as $\sigma_{mb} = (10/3) \times 10^{-6}$ (medium noise) or $\sigma_{mb} = (15/3) \times 10^{-6}$ (high noise), so that the $3\sigma_{mb}$ interval represents a 10-ppm shift for medium noise and a 15-ppm shift for high noise.

- Individual mass deviations was simulated by drawing, for each peak with mass m individually, a random number δ from $N\left(0, \sigma_{md}^2\right)$ and shifting the peak by δm. The standard deviation σ_{md} was chosen so that the $3\sigma_{md}$ interval represents a 10-ppm shift for medium noise and a 20-ppm shift for high noise.

- Intensity variations were simulated in the spectrum. Each peak intensity was multiplied by an individual random number ϵ drawn from $N\left(1, \sigma_{id}^2\right)$. Variance was chosen as $\sigma_{id}^2 = 1$ for medium noise and $\sigma_{id}^2 = 2$ for high noise. 0.03 times the maximum peak intensity of the spectrum was subtracted from each peak intensity. If a peak intensity fell below the threshold of one thousands of the maximum intensity in the spectrum, the peak was discarded.

- Additional 'noise peaks' were added to the spectrum. In processing of origin dataset, a pool of 'noise peaks' was gathered from the fragmentation spectra, using all peaks that did not have a molecular sub-formula decomposition of the known molecular formula of the precursor. For each spectrum, αn of these 'noise peaks' were added to the spectrum, where n is the number of peaks in the spectrum and $\alpha = 0.2$ for medium noise and $\alpha = 0.4$ for high noise. Intensities of 'noise peaks' were adjusted for maximum peak intensities in the contributing and receiving spectrum. 'Noise peaks' were randomly drew from the pool of 'noise peaks' and added to the spectrum.

For another issue, the spectra collection did not possess isotopes pattern. In real LC-MS processing (feature detection), isotope peaks were grouped and merged, which favorable for SIRIUS to detect some specific element[5]. To simulate isotopes pattern, we used function of 'get.isotopes.pattern' within 'rcdk' R package to get isotope mass and its abundance[6]. Furthermore, these mass were considered for the adduct type to increase or decrease exact mass. For the 'intensity' of these isotope pattern, we simulated as relative intensity, i.e., the abundance of isotopes multiply by 100 by value. These 'isotope peaks' were merged into MS^1 list of its compounds. All the spectra collections were formatted to fit with input of MCnebula workflow or benchmark method (.mgf file and feature quantified table .csv).

2.1.1 Serum dataset

We re-analyzed 245 LC-MS/MS data (.mzML) from MASSIVE (id no. MSV000079949) (blanks, controls and samples)[3]. MZmine2 (version 2.53) was performed for feature detection. The detection workflow mainly involves **1)** Automated Data Analysis Pipeline (ADAP) for peak detection and deconvolution[7], **2)** isotopes peak finder, **4)** parallel samples join alignment, **5)** gap filling algorithm. While we exported MS/MS spectra (.mgf) for SIRIUS 4 software computation, spectra were merged across samples into one fragmentation list with 30% Peak Count threshold filtering. The feature detection workflow was referred to FBMN preprocessing and SIRIUS computational prerequisites. The output .mgf was run with SIRIUS 4 software (version 4.9.12) for computation with SIRIUS[8], ZODIAC[9], CSI:fingerID[10], CANOPUS[11]. In particular, SIRIUS was customized set to detect Iodine element while predicting formula. MCnebula package were used for subsequent data analysis. All subsequent analysis have been organized into concise code and exported as reports (see section of Data and code availability).

Heat map analysis was performed with ‘Acyl carnitines’ (ACs), ‘Lysophosphatidylcholines’ (LPCs) and ‘Bile acids, alcohols and derivatives’ (BAs). The ‘features’ are select by either in infection groups versus control groups or HM versus HS group: Q-value < 0.05, $|\log2(FC)| \geq 0.3$.

Kyoto Encyclopedia of Genes and Genomes (KEGG) metabolic pathway enrichment analysis was performed with LPCs and BAs, respectively. We used the identified InChIKey planar of structures to hit compounds in metabolic pathway. In detail, firstly, in order to avoid the identified structural deviations due to stereoisomerism, the InChIKey planar was used to obtain all possible InChIKeys via PubChem API. In this step, PubChem CID of those compounds were also obtained. The R package of MetaboAnalystR was used for converting PubChem CID to KEGG ID[12]. Many compounds were not related to metabolic pathway so those were filtered out. The R package of FELLA was used for KEGG enrichment with ‘pagerank’ algorithm[13]. The above methods have been integrated as functions to interface with the MCnebula workflow. These functions are available in the ‘exMCnebula2’ package.

2.1.2 Herbal dataset

Material and processing. *E. ulmoides* dried bark was obtained from company of ZheJiang ZuoLi Chinese Medical Pieces LTD. Raw-Eucommia and Pro-Eucommia were prepared as followed: (1) Raw-Eucommia: The shreds or blocks of *E. ulmoides* dried bark were took, powdered and passed through 80-mesh sieves for further process. (2) Pro-Eucommia: The shreds or blocks of *E. ulmoides* dried bark were sprayed with saline water (the amount of salt is 2% of *E. ulmoides*, add 10 fold of water to dissolve), and smothered in airtight for 30 min. Then, the barks were dried in oven at 60 °C, followed by baking at 140 °C for 60 min. Finally, the baked barks were powdered and passed through 80-mesh sieves for further process.

Sample preparation. 2 g of Raw-Eucommia powder and Pro-Eucommia powder were weighed, respectively, added 50 mL of methanol/water (1:1, v/v) followed by ultrasonic (20 kHz for 40 min). After ultrasonic, the mixture was filtered to obtain filtrate and residue. The residue was added with 50 mL of methanol/water (1:1, v/v) and extracted with ultrasonic (40 kHz, 250 W for 20 min) again. The mixture was filtered. Then, the filtrate of the two extracts was combined, the solvent was evaporated. Methanol/water (1:1, v/v) was added to redissolve the extract and the volume was fixed to 5 mL. Finally, the supernatant was obtained by centrifugation (12,000 r.p.m. for 10 min) for further LC−MS analysis.

LC-MS experiments. LC−MS analysis was performed using a Dionex Ultimate 3000 UHPLC system (Dionex, Germany) coupled with a high-resolution Fourier-transform mass spectrometer (Orbitrap Elite, Thermo Fisher Scientific, Germany) using a Waters Acquity HSS T3 column (1.8 mum, 100 mm × 2.1 mm, Waters Corporation, Milford, MA, USA). Solvent A, formic acid/water (0.1:99, v/v), and solvent B, formic acid/acetonitrile (0.1:99, v/v), were used as the mobile phase. The gradient profile for separation was as follows: 2% of solvent B at 0min, 5% of solvent B at 2 min, 15% of solvent B at 10 min, 25% of solvent B at 15 min, 50% of solvent B at 18 min, 100% of solvent B at 23 min, 2% of solvent B at 25 min, and 2% of solvent B at 30 min. The flow rate was 0.3 mL/min. The column temperature was set at 40°C.

Mass spectrometric analysis was performed using an Orbitrap Elite instrument equipped with an ESI source (Thermo FisherScientific, Germany) that operated in the negative ionization mode. The ESI source was operated at 50 °C with a capillary temperature of 275 °C, an ionization voltage of 3.5 kV, and a sheath gas flow rate of 35 L/min. The survey scans were conducted in the Orbitrap mass analyzer operating at a 120,000 (full width at half-maximum) resolving power. A mass range of 100−1500 m/z and a normalized collision energy of 30 eV were used for survey scans. The analysis method was set to analyze the top 10 most intense ions from the survey scan, and a dynamic exclusion was enabled for 15s.

MCnebula Workflow. E.ulmoides dataset was preprocessed with MZmine2 for feature detection, followed by SIRIUS software computation. The subsequent analysis was similar to serum metabolic dataset and also been organized as a report (see section of Data and code availability).

2.1.3 Data processing

Raw data (.raw) were converted to m/z extensible markup language (mzML, i.e., .mzml format data) in centroid mode using MSConvert ProteoWizard[14]. The .mzml files were processed with MZmine2 (v.2.53) and followed by SIRIUS 4 in Pop!-OS (Ubuntu) 22.04 LTS 64-bits workstation (Intel Core i9-10900X, 3.70GHz × 20, 125.5 Gb of RAM)[15]. MCnebula (MCnebula2) and other R packages were executed in Pop!_OS (Ubuntu) 22.04 LTS 64-bits PC (Intel Core i7-1065G7, 1.3 GHz × 8, 16 Gb of RAM).

2.1.4 Data and code availability

The source code of MCnebula was available at https://github.com/Cao-lab-zcmu/MCnebula. The source code of exMCnebula2 was available at https://github.com/Cao-lab-zcmu/exMCnebula2. The code for all the analysis in this study can be found in the internal data directory (‘inst/extdata/’) of the ‘exMCnebula2’ package. In addition, .mgf files (msms spectra) and .csv files (feature quantification) and SIRIUS output files (use MCnebula function to filter and compress tens of GB of data to just a few tens of MB) and analysis report of serum and herbal dataset were compressed and stored in the exMCnebula2 package. By downloading and installing MCnebula package and exMCnebula2 package, all the analyses of this study can be reproduced by executing R codes.

The serum dataset were available at MassIVE web service (id no. MSV000079949). The submission job in GNPS of evaluation dataset are available:

1) original dataset:

FBMN: https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=05f492249df5413ba72a1def76ca973d. MolnetEnhancer: https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=9d9c7f83fa2046c2bf615a3dbe35ca62;

2) medium noise dataset:

FBMN: https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=c65abe76cd9846c99f1ae47ddbd34927; MolnetEnhancer: https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=7cc8b5a2476f4d4e90256ec0a0f94ca7;

3) high noise dataset:

FBMN: https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=62b25cf2dcf041d3a8b5593fdbf5ac5e; MolnetEnhancer: https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=f6d08a335e814c5eac7c97598b26fb80.

3 Figure legend of supplementation

Fig. S1 | Parent-Nebula of serum metabolomics dataset. In **Parent-nebula**, ‘features’ are mapped as nodes in network graph. The edges illustrated the spectral similarity of adjacent ‘features’. Not all ‘features’ are shown in the Parent-Nebula, as the isolated nodes are removed.

Fig. S2 | Child-Nebulae of serum metabolomics dataset. The Child-Nebulae are created according to chemical classes in Nebula-Index. The classified ‘features’ of chemical classes are mapped into corresponding Child-Nebulae.

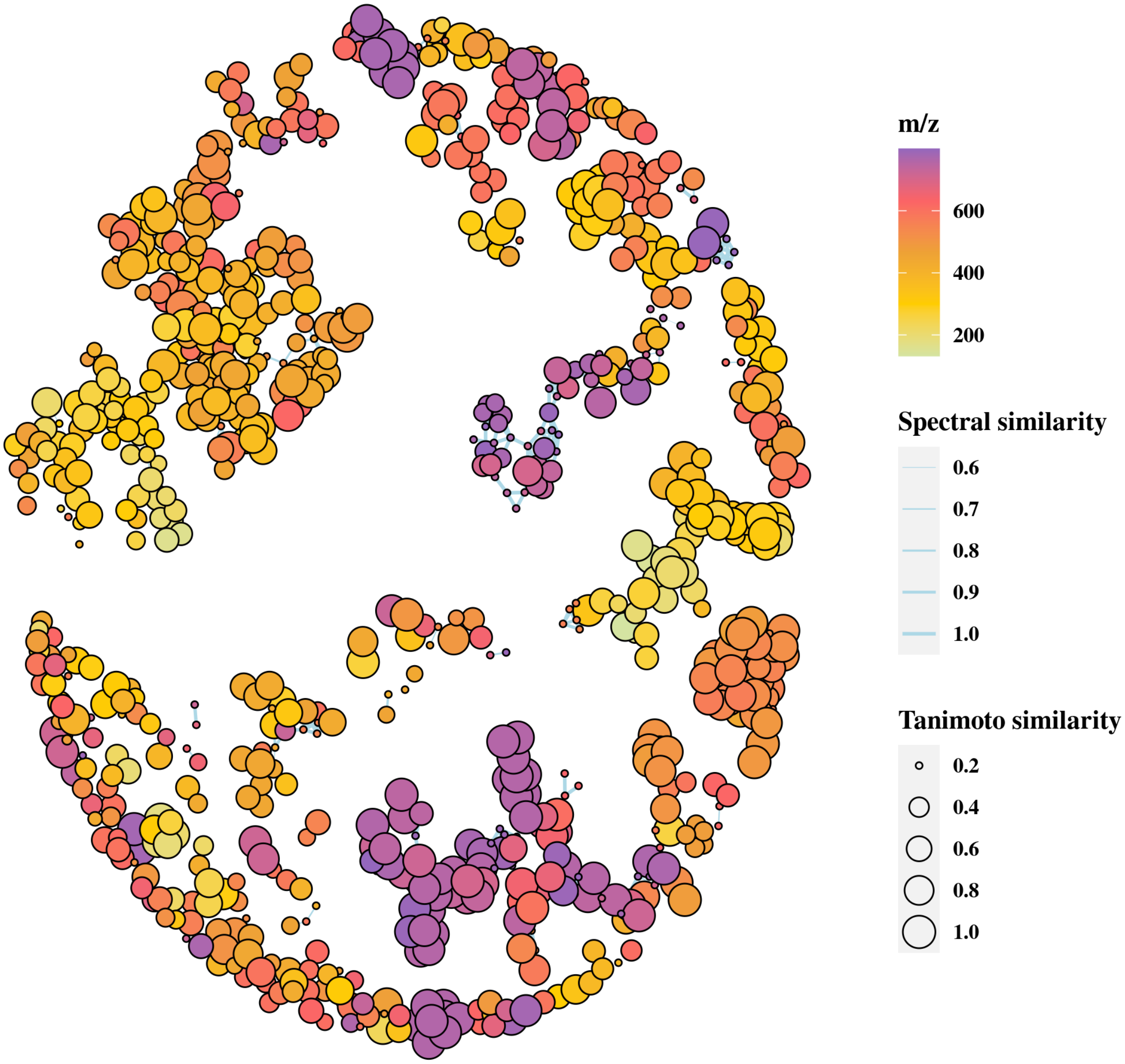

Fig. S1: Parent-Nebula of serum metabolomics dataset.

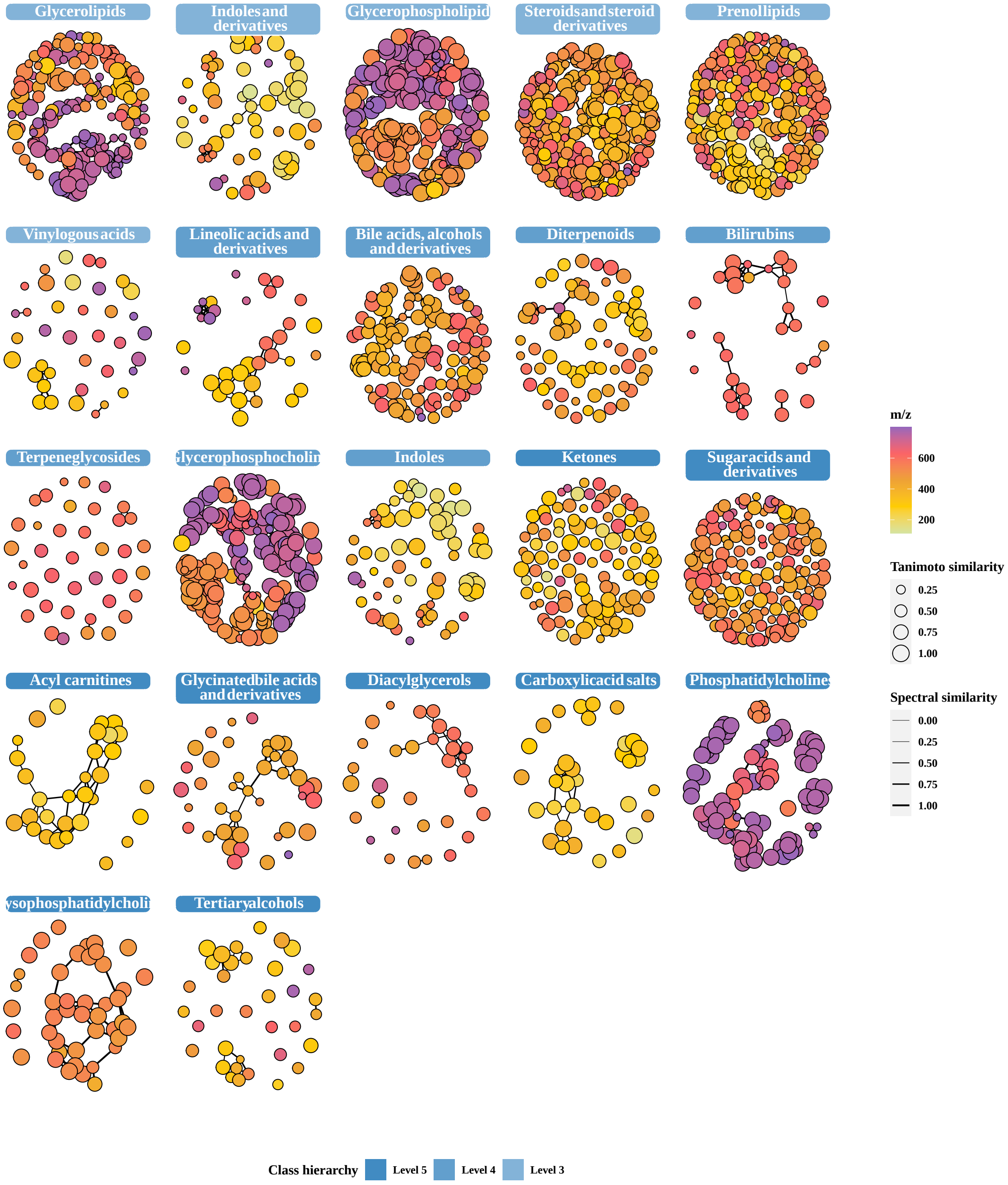

Fig. S2: Child-Nebulae of serum metabolomics dataset.

Reference

1. Smyth, G. K. Limma: Linear Models for Microarray Data. in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor* (eds. Gentleman, R., Carey, V. J., Huber, W., Irizarry, R. A. & Dudoit, S.) 397–420 (Springer-Verlag, 2005). doi:10.1007/0-387-29362-0_23.

2. Law, C. W. *et al.* A guide to creating design matrices for gene expression experiments. *F1000Research* **9**, 1444 (2020).

3. Wozniak, J. M. *et al.* Mortality Risk Profiling of Staphylococcus aureus Bacteremia by Multi-omic Serum Analysis Reveals Early Predictive and Pathogenic Signatures. *Cell* **182**, 1311–1327.e14 (2020).

4. Xie, Y., Dervieux, C. & Riederer, E. *R markdown cookbook.* (Chapman and Hall/CRC, 2020).

5. Böcker, S., Letzel, M. C., Lipták, Z. & Pervukhin, A. SIRIUS: Decomposing isotope patterns for metabolite identification†. *Bioinformatics* **25**, 218–224 (2009).

6. Guha, R. Chemical informatics functionality in R. *Journal of Statistical Software* **18**, (2007).

7. Myers, O. D., Sumner, S. J., Li, S., Barnes, S. & Du, X. One Step Forward for Reducing False Positive and False Negative Compound Identifications from Mass Spectrometry Metabolomics Data: New Algorithms for Constructing Extracted Ion Chromatograms and Detecting Chromatographic Peaks. *Analytical Chemistry* **89**, 8696–8703 (2017).

8. Dührkop, K. *et al.* SIRIUS 4: A rapid tool for turning tandem mass spectra into metabolite structure information. *Nature Methods* **16**, 299–302 (2019).

9. Ludwig, M. *et al.* Database-independent molecular formula annotation using Gibbs sampling through ZODIAC. *Nature Machine Intelligence* **2**, 629–641 (2020).

10. Dührkop, K., Shen, H., Meusel, M., Rousu, J. & Böcker, S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proceedings of the National Academy of Sciences* **112**, 12580–12585 (2015).

11. Dührkop, K. *et al.* Systematic classification of unknown metabolites using high-resolution fragmentation mass spectra. *Nature Biotechnology* **39**, 462–471 (2021).

12. Pang, Z., Chong, J., Li, S. & Xia, J. MetaboAnalystR 3.0: Toward an optimized workflow for global metabolomics. *Metabolites* (2020) doi:10.3390/metabo10050186.

13. Picart-Armada, S., Fernandez-Albert, F., Vinaixa, M., Yanes, O. & Perera-Lluna, A. FELLA: An R package to enrich metabolomics data. *BMC Bioinformatics* **19**, 538 (2018).

14. Chambers, M. C. *et al.* A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology* **30**, 918–920 (2012).

15. Pluskal, T., Castillo, S., Villar-Briones, A. & Orešič, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **11**, 395 (2010).