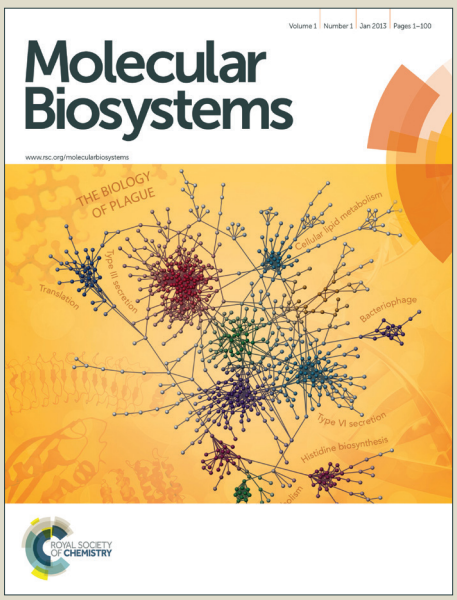
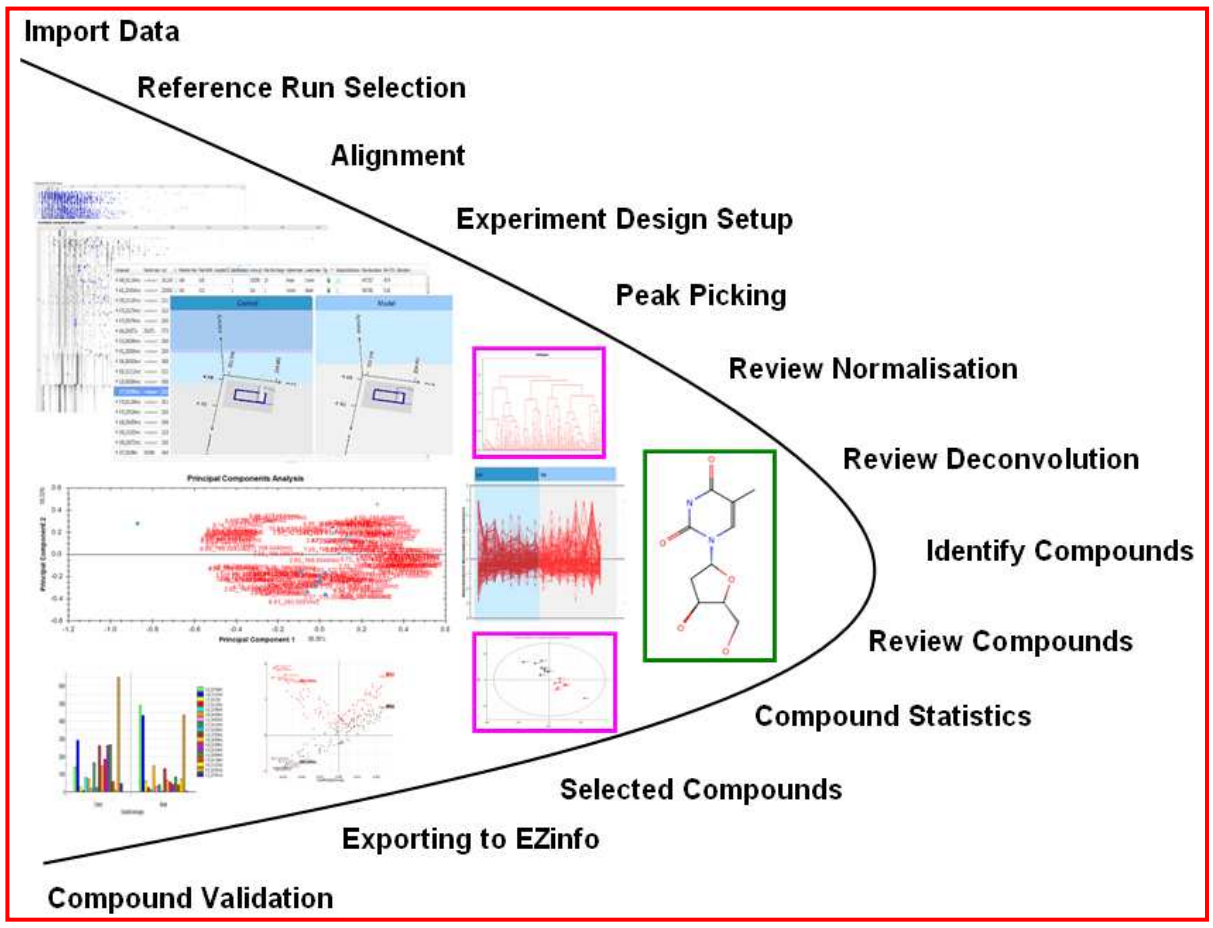
Molecular BioSystems

Accepted Manuscript



ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/molecularbiosystems



Detailed Analysis workflow of TransOmics Informatics for Metabolomics data from large biological data sets

Rapidly Improve Determination of Metabolites from Biological Data Sets Using High-Efficient TransOmics Tool

Aihua Zhang[1], Xiaohang Zhou[1], Hongwei Zhao[2], Yu Guan[1], Shiyu Zhou[2], Guang-li Yan[1], Zhonghua Ma[2], Qi Liu[1], Xijun Wang[1]*

1. National TCM Key Laboratory of Serum Pharmacochemistry, Key Laboratory of Metabolomics and Chinmedomics, Department of Pharmaceutical Analysis, Heilongjiang University of Chinese Medicine, Heping Road 24, Harbin 150040, China

2. Infinitus (China) Company Ltd, Guangzhou, China

*Correspondence

Prof. Xijun Wang

National TCM Key Laboratory of Serum Pharmacochemistry

Key Laboratory of Metabolomics and Chinmedomics

Department of Pharmaceutical Analysis

Heilongjiang University of Chinese Medicine

Heping Road 24

Harbin 150040, China

Tel. & Fax +86-451-82193038

Email: yinghan123456@126.com;xijunwangls@126.com

Abstract:

Metabolomics is a new approach based on the systematic study of the full complement of metabolites in a biological sample. Extracting biomedical information from large datasets is of considerable complexity. Furthermore, the traditional method of assessing metabolomics data is not only time-consuming but is often subjective work. Here we used a sensitive ultra-performance LC-ESI/Q-TOF high-definition mass spectrometry (UPLC-ESI-Q-TOF-MS) in positive ion mode coupled with a new developed software program TransOmics for widely untargeted metabolomics, which incorporates novel nonlinear alignment, deconvolution, matched filtration, peak detection, and peak matching to characterize metabolites as a case study. TransOmics method can facilitate prioritization of the data and greatly increase the probability of identifying metabolites related to the phenotype of interest. By this means, 17 urinary differential metabolites were identified (less than 10 min) involving the key metabolic pathways including tyrosine metabolism, glutathione metabolism, phenylalanine metabolism, ascorbate and aldarate metabolism, arginine and proline metabolism, and so forth. Metabolite identification has also been significantly improved, using the correlation peak patterns in contrast to a reference metabolite panel. It can detect and identify metabolites automatically and remove background noise, and also provides a user-friendly graphical interface to apply principal component analyses, correlation analysis and compound statistics. This investigation illustrates that metabolomics combined with the proposed bioinformatic approach (based on TransOmics) is important to elucidate the developing biomarkers and physiological mechanism of disease, and has opened the door for the development of a new genre of metabolite identification method.

Keywords:

Metabolomics; TransOmics; mass spectrometry; metabolites; biomarkers; ultra-performance liquid chromatography-mass spectrometry

1. Introduction

Metabolomics is based on the dynamic changes of low molecular weight metabolites in biological samples [1]. The metabolites often mirror the end result of genomic and protein perturbations in disease, and they are closely associated with phenotypic changes [2]. Furthermore, the mechanisms of diseases would be elucidated by identifying the metabolites, analyzing the metabolic pathway and so on [3]. It is a powerful tool to advance the diagnosis, treatment and prevention of human diseases [4,5]. Monitoring metabolite level has become an important method to detect early stages in disease [6]. Urine metabolite profiles were used to identify potential biomarkers, which can provide new insights into biological processes [7,8]. By measuring holistic endogenous metabolites in urine, metabolomics can be used for delineating metabolic networks and discovering metabolic markers [9]. Advances in the high-throughput technology and the growing quality and quantity of data put new demands on applied analytical methods [10]. Furthermore, exploration of finally generated and analyzed datasets relies on powerful tools for data mining and visualization. For this purpose, TransOmics software (Waters Corporation, Milford, MA) was chosen for generating the metabolite data and characterizing metabolic changes in biological samples.

The biodata mining from Big data in the biological and biomedical sciences is a huge challenge. With the advantages of high sensitivity and accuracy, wide dynamic range, and the ability to identify metabolites from complex Bio data, mass spectrometry has become the workhorse of metabolomics research [11-13]. As we known, the current UPLC-MS metabolomic technologies require considerable time cost to pick out the interesting peaks correlated to diseases, which restricts the analysis of large amounts of biosamples [14]. Identification and quantification of analytes in biological data sets are critical procedures in metabolomics areas. To solve this problem, we used new software TransOmics to screen and identify low molecular weight metabolites in urine, for the simultaneous analysis of up to several hundred metabolites at high sensitivity, selectivity, and quantitative capability.

Urine has been shown to contain a wealth of metabolic information that may be altered due to underlying disease, and do satisfy the criteria of minimal invasiveness, reasonable cost, or minimal time demand [15,16]. Without using internal standards, the method dynamically identifies hundreds of the metabolites for each sample; the relative intensities are directly obtained and the picking peak and data processing were greatly simplified in our study. Although current mass UPLC/MS method is very powerful, there is a clear need for more rapid, high-throughput approaches for metabolomics studies. Therefore, this is the first paper that was designed to investigate a urine metabolome of kidney-yang deficiency syndrome rats induced by corticosterone, using UPLC/MS combined with TransOmics as a case study, to explore the potential biomarkers, and enhance our understanding of its mechanisms. Potential topics include mining high-dimensional data, computing big data, visualizing big data, integrating diverse data, sharing complex data, etc, for the new frontier of big data in the biological and biomedical sciences.

2. Materials and methods

2.1 Materials and reagents

Acetonitrile, HPLC grade, was obtained from Merck (Darmstadt, Germany); methanol (HPLC grade) was purchased from Fisher Scientific Corporation (Loughborough, UK); ultrapure water was prepared by a Milli-Q system (18.2 M, Millipore, MA, USA) and used for the preparation of samples and mobile phase; leucine enkephalin was purchased from Sigma-Aldrich (St. Louis, MO, USA). NSC were provided by Infinitus (China) Company Ltd (Guangzhou, China). Corticosterone was produced from Sigma Ltd. All other reagents were of analytical grade.

2.2 Animal handling procedure.

Male Wistar rats (weighting 220±20g) were supplied by GLP Center of Heilongjiang University of Chinese Medicine (Harbin, China). They had free access to food pellets and tap water under standard conditions of humidity (50 ± 5%), temperature (25 ± 1 °C) and 12 h light-dark cycle. All animals were allowed to acclimatize in metabolism cages for 1 week prior to treatment. All the rats were randomly divided into 2 groups of 8 rats each as follows: control group and model group. After that they were separated randomly into two groups as follows: The model group (kidney-yang deficiency syndrome in chinese) was injected subcutaneously corticosterone at a dose of 1mL/100g once daily for 21 days. Rats in the control group were given saline alone under the same conditions. Animal experiment was carried out in accordance with the Guidelines for Animal Experimentation of Heilongjiang University of Chinese Medicine and the animal study was approved by the Animal Ethics Committee.

2.3 Collection and preparation of biosamples

Rat urine was collected from metabolism cages at ambient temperature throughout the whole procedure and centrifuged at 13,000 rpm at 5◦C for 15 min, and the supernatants were stored frozen at −80 °C until metabolomic analysis. The quality control sample was used to optimize the condition of UPLC-MS, as it contained most information of whole urine samples.

2.4 LC analysis

Chromatographic separation was performed on an ACQUITY UPLC system (Waters Corporation, Milford, MA) with a conditioned autosampler at 4◦C. The column used was an HSS T3 C18 column (100 mm×2.1mm i.d., 1.8μm). Column temperature was maintained at 40 ◦C for all analyses. The mobile phase consisted of a linear gradient system of (A) 0.1% formic acid and (B) 0.1% formic acid in water in acetonitrile. The gradient conditions of the mobile phase were as follows: 0-1 min, 1% A; 1-2.5 min, 1-13% A; 2.5-6.5 min, 13-40% A; 6.5-8.0 min, 40-99% A; 8.0-10.5 min, 99% A; 10.5-11.0 min, 99-1% A; 11-13 min, 1 A. The flow rate was 0.4 mL/min and injection volume was 3μL.

2.5 Mass spectrometric conditions

High-definition mass spectrometry was performed on a Waters Q-TOF equipped with an electrospray ion source in the positive mode (ESI^+) as a case study. The optimal conditions of analysis were as follow: the source temperature was set at 110 °C, desolvation gas temperature was 450 °C, cone gas flow was 50h, desolvation gas flow was 600L/h. In positive ion mode, the capillary voltage was 3.0kV, the sampling cone voltage was 25V, and extraction cone voltage was 3.0V, desolvation gas flow was 600L/h. Data were collected in the centroid mode between m/z 50 and 1000, with a scan time of 0.4 s and interscan time 0.1 s. Dynamic range enhancement was applied throughout the MS experiment to ensure that accurate mass measurements were obtained over a wider dynamic range. All analyses were acquired using the lockspray to ensure accuracy and reproducibility. Leucine–enkephalin was used as the lockmass at a concentration of 0.2 ng/mL and flow rate of 100μl·min-1. The lockspray frequency was set at 10 s, and data were averaged over 10 scans. All the acquisition and analysis of data were controlled by Waters MassLynx v4.1 software.

2.6 TransOmics data processing and determination of low metabolites

The detailed method for the compound identification was described in Fig**.1**. In brief, all the LC–MS raw files were converted to TransOmics program, and subsequently the converted files were calculated for generation of alignment, peak picking, deconvolution, filter data, identify compounds, exporting to EZinfo for compound statistics (principal component analysis (PCA) and orthogonal partial least square discriminant analysis (OPLS-DA)), correlation analysis and compound validation.

2.7 Pathway analysis

Metabolic pathway analysis was performed by MetaboAnalyst tool (http://www.metaboanalyst.ca/) based on the database sources including KEGG (http://www.genome.jp/kegg/), Human Metabolome Database (http://www.hmdb.ca/), and Pubchem (http://www.ncbi.nlm.nih.gov/ pccompound/), etc, to identify the affected metabolic pathways and facilitate further biological interpretation.

3. Results and Discussion

3.1 Metabolomic profling

For UPLC-MS analysis, aliquots were separated using a Waters Acquity UPLC (Waters, Millford, MA) and analyzed using a Q-TOF/HDMS, which consisted of an electrospray ionization source mass analyzer. Total ion chromatograms after retention time correction in positive mode, showed stable retention time with no drift in all of the peaks, demonstrating the robustness of our analytical method (Fig. S1A). The stable profles showed the stability of UPLC-MS analysis and reliability of the metabolomic data. Low molecular mass metabolites could be separated well in the short time of 12 min due to the minor particles (sub-1.7um).

3.2 Discovery and identification of low metabolites

Metabolite profiling obtained from model group induced by corticosterone was performed by UPLC-MS in conjunction with TransOmics data analysis. TransOmics was adopted to identify chromatographic peaks in the total ion chromatogram. To screen the statistically important variables (ions) related to model rats, TransOmics was used to deconvolute and align mass ions from the data files of the samples into a single data set. TransOmics performs feature detection as well as nonlinear retention time alignment and calculates statistics for each feature. The result is a table that contains the m/z and retention time coordinates, p-value and fold change for each feature, and the integrated feature intensities from all aligned samples. The alignment algorithm will generate 'compound ions' in the 2D ion intensity map (Fig. S1B and S1C). In this example we will use Metascope, a flexible search engine which is designed to work with databases that can set thresholds for mass and retention time as required. The data files that were generated by TransOmics were imported into HMDB Database and filtered by Metascope. No restrictions were made on up- or down-regulation. 132 variables with the fold value larger than 2 and p-value less than 0.01 were rapidly identified and selected. Identification of all the low metabolites by TransOmics was shown in Fig. 2. Fig. 2A displays the main table of compounds with identifications as well as those that remain unknown after the identifications have been imported. Fig. 3B displays a compound abundance plot, list of possible identifications, 3D montage, drift time montage (for data collected with drift time), for the current compound highlighted in Fig. 2A. We are going to explore the correlation analysis for all the compounds (with possible identifications) that display a significant 2 fold or greater difference in abundance. TransOmics showed the detailed information for each individual feature including statistics, extracted ion chromatograms, spectrum details, and putative identifications. Correlation analysis enables the grouping of compounds together according to how similar their abundance profiles are. The answer is displayed graphically in the form of an interactive dendrogram where the vertical distance, between each branch can be taken as indicative of how similar the abundance profiles of each cluster of compounds are to each other (Fig. 3). The detailed information for each individual feature was then exporting to EZinfo for compound statistics (PCA, PLS-DA) and compound validation.

3.3 Multivariate statistical analysis of metabolite profiling

Raw data from UPLC/MS were analyzed by the TransOmics was then imported into EZinfo 2.0 software for data analysis. Multivariate data analysis was performed using the score plot of PCA, and there is an obvious separation between the clustering of the model and control groups (see Fig 4.A), suggests that biochemical perturbation significantly happened in model group. The biomarkers were selected from the VIP-plot of PLS-DA (Fig.4B). The VIP plot displayed 17 ions as differentiating metabolites according to their VIP values and considered as potential markers representing the metabolic characteristics. In addition, histogram plots (Fig. 4C) were created on the basis of integrated intensity as exported from EZinfo 2.0 software. According to the protocol detailed above, a total of 17 endogenous metabolites were finally identified as markers and listed in Supplementary Table 1.

3.4 Metabolic pathway and function analysis

Analysis of the relevant pathways was performed by MetaboAnalyst's tool that is a mass translator into pathways. The altered metabolism pathways with higher score were generated using the reference map by searching KEGG. It assigned a total of feature compounds in 14 pathways which were identified together are important for the host response to model rat. The predominant hits were alanine, aspartate and glutamate metabolism, galactose metabolism, arginine and proline metabolism, purine metabolism, amino sugar and nucleotide sugar metabolism, etc (supplementary Table S2). These altered metabolism pathways with higher score had yield satisfactory results and clearly help us to better understand the underlying mechanisms of disease.

Metabolomics combines metabolic profiling and multivariate data analysis to facilitate the high-throughput analysis of metabolites in biological samples [17]. This technique has been developed as a powerful analytical tool and hence has found successful widespread applications in many areas of bioscience [18-20]. In our study, metabolite profiling was performed by LC-MS coupled with TransOmics program. TransOmics can be used for analysis of any high-resolution LC–MS metabolomics data for the purpose of biomarker discovery, drug development, or any other comparative analysis. The introduction of TransOmics program into metabolomics is of significant value as it not only provides an analytical tool for distinguishing disease phenotype, but it also allows for data reduction at the metabolite level. Interestingly, 132 distinct metabolites identified in the urine metabolome, 17 of which are in various stages of disease progress. Furthermore, these metabolites were tightly correlated with the alanine, aspartate and glutamate metabolism, galactose metabolism, arginine and proline metabolism, purine metabolism, amino sugar and nucleotide sugar metabolism. In this tutorial, we had successfully give an illustrative example of the key aspects that any researcher needs to consider how to rapidly improve determination of low metabolites from biological data sets and analyze pathway networks, when working with high-throughput MS data.

4. Conclusions

Metabolomics can provide a powerful approach to discover biomarkers by analyzing global changes in the metabolic profile. In the present study, we developed the integrated the untargeted and targeted analytical approach based on high-resolution UPLC/MS with TransOmics program to profile the metabolic changing in urine samples, to rapidly identify the low metabolites. From our results, the perturbed metabolic pattern in urine and identified metabolic markers associated with model rats. A panel of metabolite markers was selected, which was able to discriminate disease subjects from the controls. Interestingly, 17 metabolites, involved in alanine, aspartate and glutamate metabolism, galactose metabolism, arginine and proline metabolism pathways etc, were identified, all of them were considered as potential biomarkers. TransOmics could give a sensitive detection of all the main peaks, and to obtain complementary structural information for all peaks of interest. It will pave the way for a better understanding of the mechanisms of diseases. Overall, this investigation illustrates the power of the UPLC-MS platform combined with the TransOmics analysis method.

Acknowledgments

This work was supported by grants from the Key Program of Natural Science Foundation of State (Grant No. 90709019, 81173500, 81373930, 81302905, 81102556, 81202639), National Key Technology Research and Development Program of the Ministry of Science and Technology of China (Grant No. 2011BAI03B03, 2011BAI03B06, 2011BAI03B08), National Key Subject of Drug Innovation (Grant No. 2009ZX09502-005).

Competing financial interests

The authors declare no competing financial interests.

References

(1) Fellay J, Thompson AJ, Ge D, Gumbs CE, Urban TJ, Shianna KV, Little LD, Qiu P, Bertelsen AH, Watson M, Warner A, Muir AJ, Brass C, Albrecht J, Sulkowski M, McHutchison JG, Goldstein DB. Nature 2010;464: 405-408.

(2) Hsu CS, Hsu SJ, Chen HC, Tseng TC, Liu CH, Niu WF, Jeng J, Liu CJ, Lai MY, Chen PJ, Kao JH, Chen DS. Proc Natl Acad Sci U S A. 2011;108: 3719-3724.

(3) Wang X, Yang B, Sun H, Zhang A. Anal Chem. 2012;84(1):428-39.

(4) Zhang AH, Sun H, Han Y, Yan GL, Yuan Y, Song GC, Yuan XX, Xie N, Wang XJ. Anal Chem. 2013;85(15):7606-12.

(5) Sun H, Zhang A, Yan G, Piao C, Li W, Sun C, Wu X, Li X, Chen Y, Wang X. Mol Cell Proteomics. 2013;12(3):710-9.

(6) Wang X, Zhang A, Sun H. Hepatology. 2013;57(5):2072-7.

(7) Boughton BA, Callahan DL, Silva C, Bowne J, Nahid A, Rupasinghe T, Tull DL, McConville MJ, Bacic A, Roessner U. Anal Chem. 2011;83:7523-30.

(8) Wang X, Zhang A, Yan G, Sun W, Han Y, Sun H. PLoS One. 2013 Aug 15;8(8):e71403.

(9) Nicholson JK, Lindon JC. Nature. 2008; 455:1054-1056.

(10) Zhang A, Sun H, Han Y, Yan G, Wang X. PLoS One. 2013;8(5):e64381.

(11) Zhang Y, Zhang A, Yan G, Cheng W, Sun H, Meng X, Liu L, Xie N, Wang X. Mol Biosyst. 2014;10(1):65-73

(12) Wang X, Zhang A, Wang P, Sun H, Wu G, Sun W, Lv H, Jiao G, Xu H, Yuan Y, Liu L, Zou D, Wu Z, Han Y, Yan G, Dong W, Wu F, Dong T, Yu Y, Zhang S, Wu X, Tong X, Meng X. Mol Cell Proteomics. 2013;12(5):1226-38.

(13) Spagou K, Wilson ID, Masson P, Theodoridis G, Raikos N, Coen M, Holmes E, Lindon JC, Plumb RS, Nicholson JK, Want EJ. Anal Chem. 2011;83:382-390.

(14) Zhang A, Sun H, Yan G, Han Y, Ye Y, Wang X.Clin Chim Acta. 2013;418:86-90.

(15) Zhang A, Sun H, Wang P, Han Y, Wang X. Analyst. 2012;137(2):293-300.

(16) Zhang A, Sun H, Han Y, Yuan Y, Wang P, Song G, Yuan X, Zhang M, Xie N, Wang X. Analyst. 2012;137(18):4200-8.

(17) Sun H, Zhang S, Zhang A, Yan G, Wu X, Han Y, Wang X. PLoS One. 2014;9(3):e93384.

(18) Yanes O, Tautenhahn R, Patti GJ, Siuzdak G. Anal Chem. 2011;83:2152-2161.

(19) Zhang A, Sun H, Yan G, Wang P, Han Y, Wang X. Cancer Lett. 2014;345(1):17-20.

(20) Wang X, Zhang A, Han Y, Wang P, Sun H, Song G, Dong T, Yuan Y, Yuan X, Zhang M, Xie N, Zhang H, Dong H, Dong W. Mol Cell Proteomics. 2012;11(8):370-80.

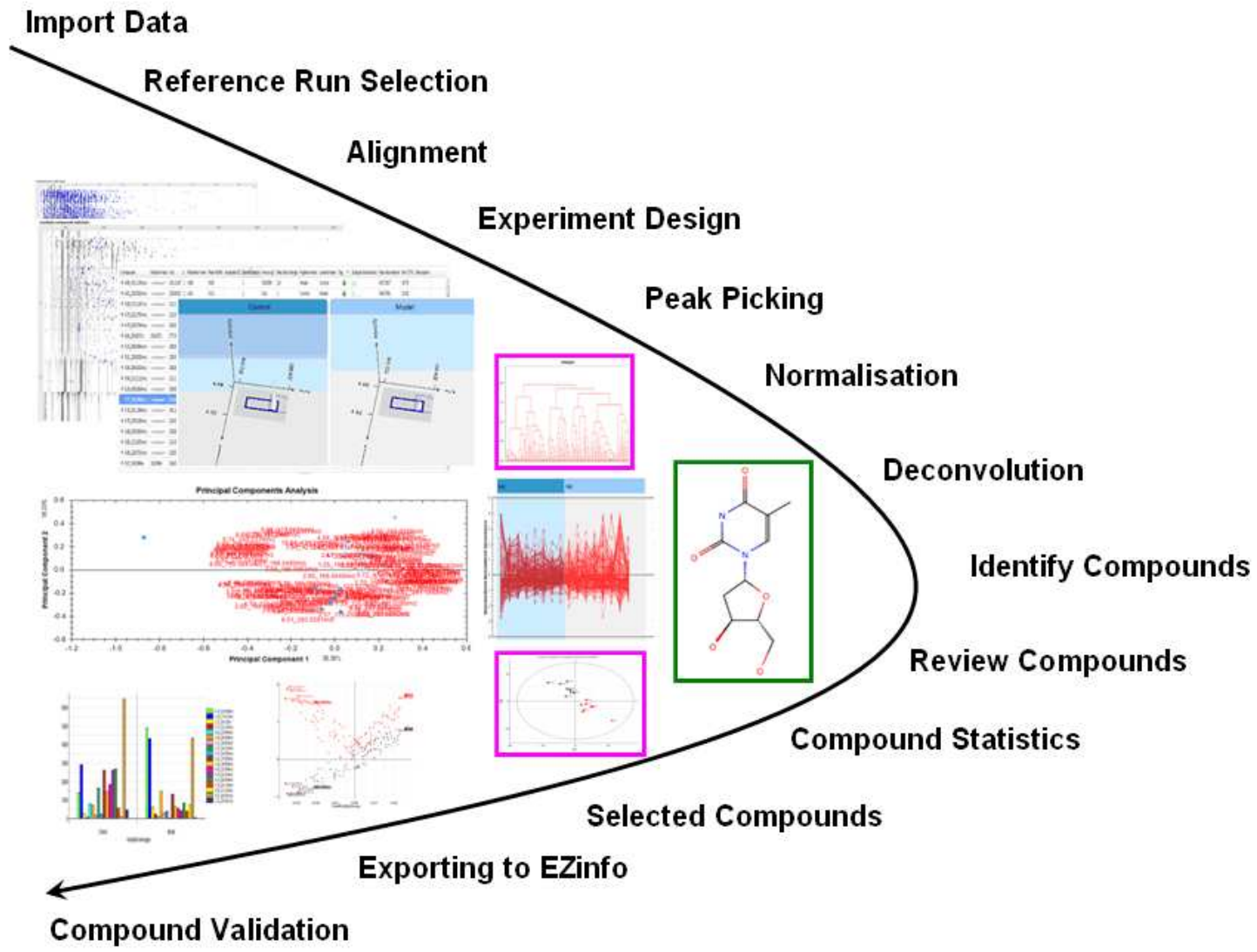


Fig. 1 Detailed analysis workflow of TransOmics informatics for metabolomics data from large biological data sets

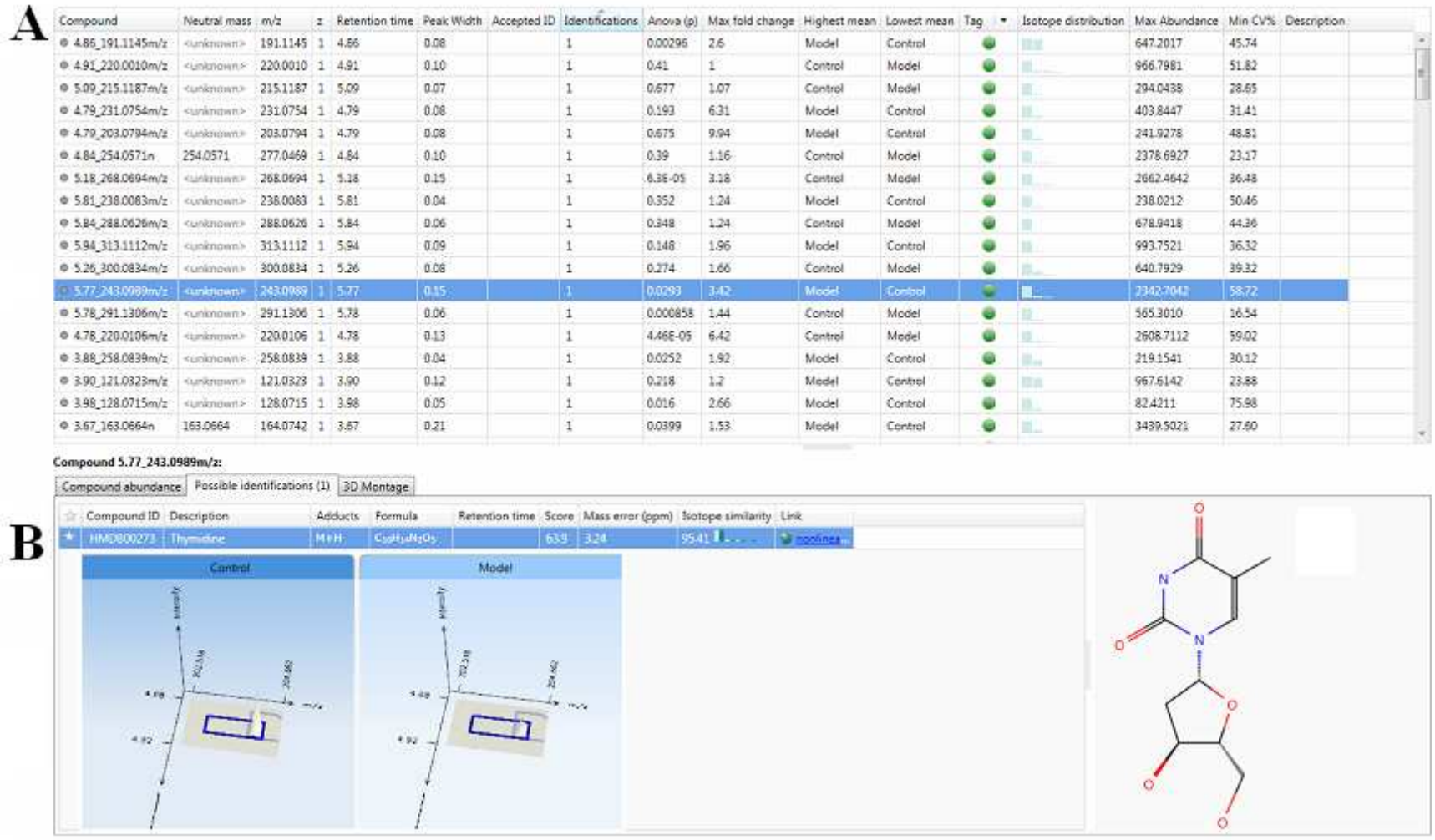


Fig. 2 The detailed information for identification of metabolites by TransOmics online.

Window A displays the main table of compounds with identifications as well as those that remain unknown after the identifications have been imported. Window B displays a compound abundance plot, list of possible identifications, 3D montage, drift time montage (for data collected with drift time), for the current compound highlighted in Window A.

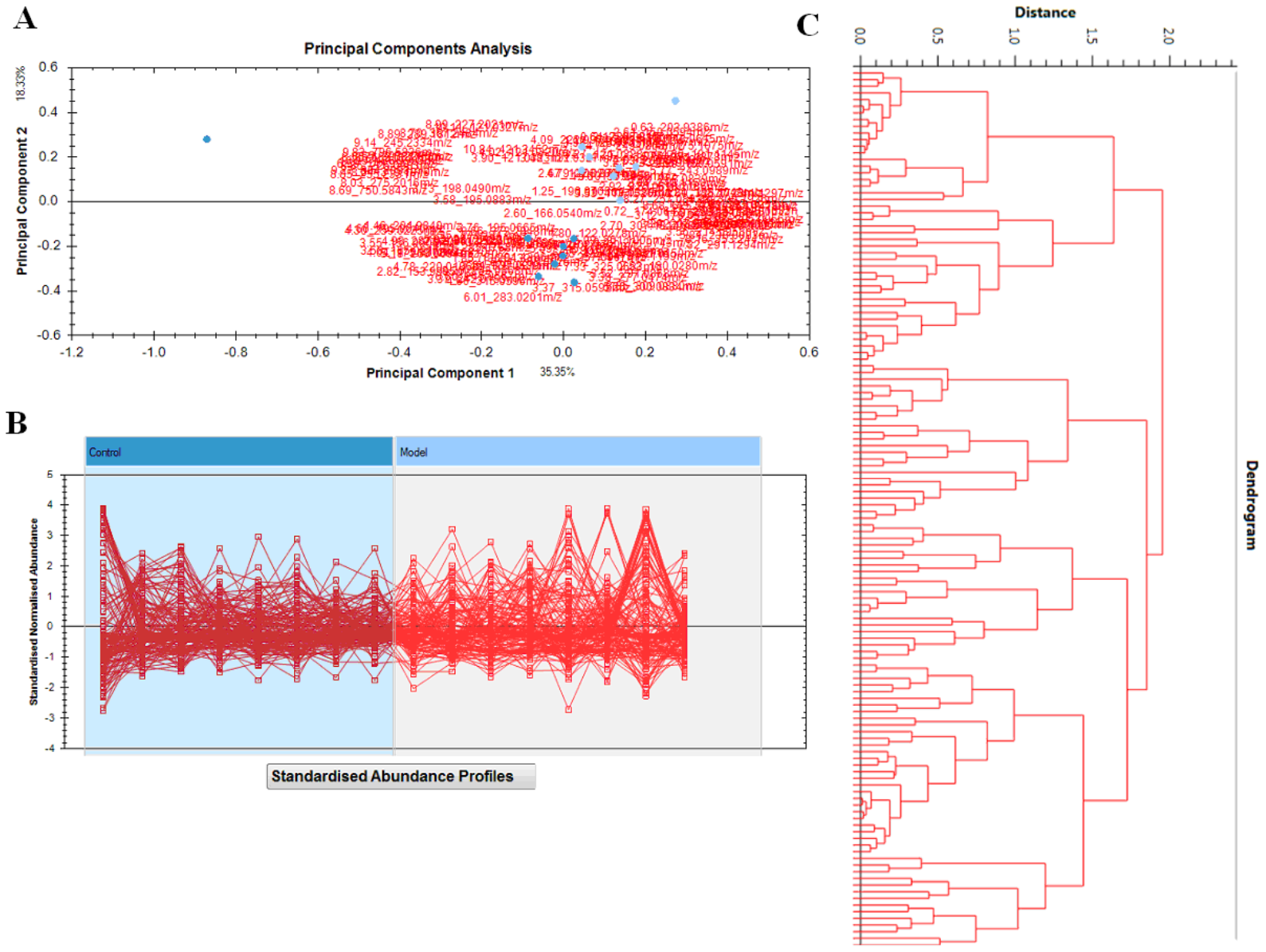


Fig. 3 Compound Statistics. Using Principal Components Analysis (PCA) to produce a simplified graphical representation of the multidimensional data (A). PCA can be used to determine whether there are any outliers in the data and also look at how well the samples group. Their abundance profiles will appear in the lower panel B. Correlation analysis for all the compounds (with possible identifications) that display a significant fold or greater difference in abundance (C). Correlation analysis enables the grouping of compounds together according to how similar their abundance profiles are.

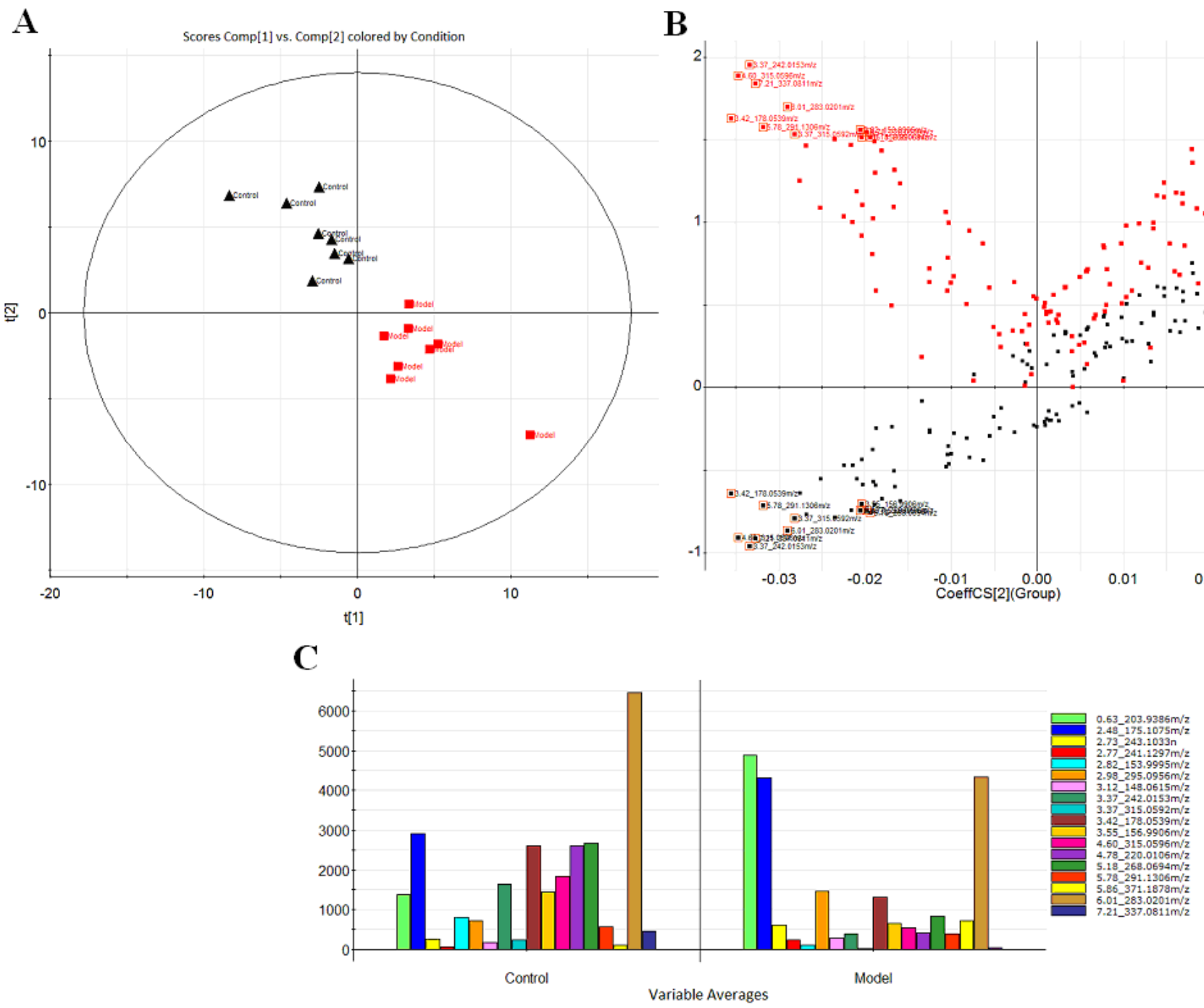


Fig. 4 Metabolomic profiling of model and matched control in positive ionization mode using EZinfo system. PCA model results in positive mode (A). Panel B shows the VIP-score plot constructed from the supervised OPLS analysis of urine (ESI^+ mode). Selected compounds in the Panel B will highlight the compounds on the ' Histogram plot' and their abundance profiles will appear in the lower panel C.