Journal of Proteomics 232 (2021) 104023

Contents lists available at ScienceDirect

Journal of Proteomics

journal homepage: www.elsevier.com/locate/jprot

MMEASE: Online meta-analysis of metabolomic data by enhanced metabolite annotation, marker selection and enrichment analysis

Qingxia Yang [a,b,1], Bo Li [c,1], Sijie Chen [d,1], Jing Tang [e,1], Yinghong Li [d], Yi Li [a], Song Zhang [a], Cheng Shi [a], Ying Zhang [a], Minjie Mou [a], Weiwei Xue [d], Feng Zhu [a,d,*]

[a] College of Pharmaceutical Sciences, Zhejiang University, Hangzhou, Zhejiang 310058, China
[b] Department of Bioinformatics, Smart Health Big Data Analysis and Location Services Engineering Lab of Jiangsu Province, School of Geographic and Biologic Information, Nanjing University of Posts and Telecommunications, Nanjing 210023, China
[c] College of Life Sciences, Chongqing Normal University, Chongqing, Chongqing 401331, China
[d] School of Pharmaceutical Sciences, School of Big Data and Software Engineering, Chongqing University, Chongqing, Chongqing 401331, China
[e] Department of Bioinformatics, Chongqing Medical University, Chongqing, Chongqing 400016, China

ARTICLE INFO

Keywords:
Large-scale and long-term metabolomics
Data integration
Metabolite annotation
Online tool

ABSTRACT

Large-scale and long-term metabolomic studies have attracted widespread attention in the biomedical studies yet remain challenging despite recent technique progresses. In particular, the ineffective way of experiment integration and limited capacity in metabolite annotation are known issues. Herein, we constructed an online tool MMEASE enabling the integration of multiple analytical experiments with an enhanced metabolite annotation and enrichment analysis (https://idrblab.org/mmease/). MMEASE was unique in capable of (**1**) integrating multiple analytical blocks; (**2**) providing enriched annotation for >330 thousands of metabolites; (**3**) conducting enrichment analysis using various categories/sub-categories. All in all, MMEASE aimed at supplying a comprehensive service for large-scale and long-term metabolomics, which might provide valuable guidance to current biomedical studies.

Significance: To facilitate the studies of large-scale and long-term metabolomic analysis, MMEASE was developed to (**1**) achieve the online integration of multiple datasets from different analytical experiments, (**2**) provide the most diverse strategies for marker discovery, enabling performance assessment and (**3**) significantly amplify metabolite annotation and subsequent enrichment analysis. MMEASE aimed at supplying a comprehensive service for long-term and large-scale metabolomics, which might provide valuable guidance to current biomedical studies.

1. Introduction

Due to its close proximity to the phenotype of the studied system [1], the LC/MS-based metabolomics has emerged to be powerful technique facilitating the understanding of disease etiology [2], discovery of unknown metabolite [3] and prediction of drug response [4]. Among these metabolomic studies, it is observed that there is a clear shift from the short-term studies of relatively small sample size to the "large-scale and long-term" analyses [5,6]. This shift may be attributed to the increasing interests in enhancing the power of statistical analysis [7,8] and analyzing the metabolomic data involving massive samples from multiple analytical experiments [9,10]. For example, the accumulation of samples for a rare disease can be seriously restricted due to the lack of patients, which may thus require a long-term data collection last for many months or even years [11]. In other words, to effectively boost the power of statistical analyses for current biomedical researches, the large-scale and long-term metabolomics that integrates multiple analytical experiments has attracted widespread attention [12–14].

However, the large-scale and long-term metabolomic study is greatly hampered by its ineffective way of experiment integration [6] and limited capacity in metabolite annotation [15]. Particularly, it is still challenging to integrate multiple experiments of thousands of samples and to simultaneously remove unwanted biological and experimental variations [16,17]; less than 2% of the detected MS peaks can be

* Corresponding author at: College of Pharmaceutical Sciences, Zhejiang University, Hangzhou 310058, China.
E-mail address: zhufeng@zju.edu.cn (F. Zhu).
[1] These authors contribute equally to this work

https://doi.org/10.1016/j.jprot.2020.104023
Received 23 December 2019; Received in revised form 12 October 2020; Accepted 22 October 2020
Available online 29 October 2020

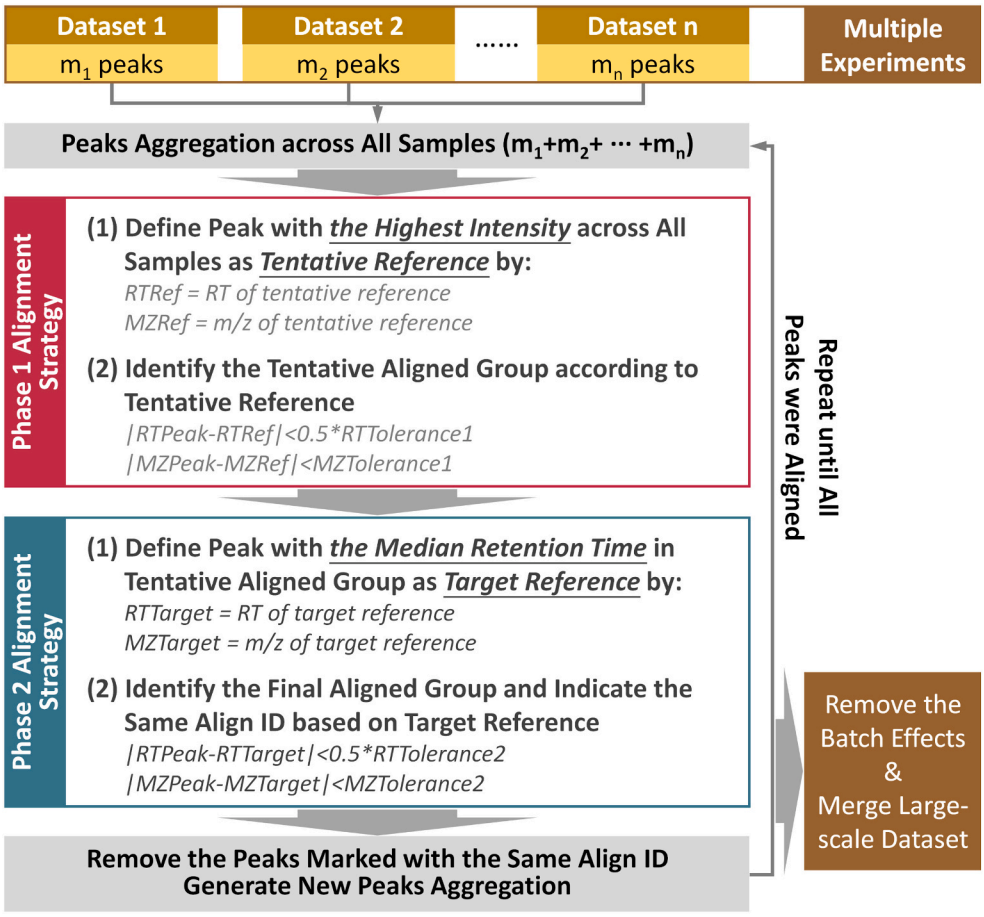

Fig. 1. The detail steps for merging large-scale datasets based on multiple metabolomics experiments in MMEASE. RT: retention time; MZ: mass to charge ratio (m/z).

successfully connected to particular metabolic compound [15] and it is still challenging to annotate these peaks with biologically interpretable information [18–20]. To cope with these problems, quality control (QC) samples [21] and internal standard (IS) [22] are proposed to remove unwanted variations among batches [23], and function annotation is conducted by referring to existing databases [24].

A variety of online tools have been developed to cope with the problems discussed above (experiment integration & metabolite annotation), which included *MetaboAnalyst* [25], *XCMS online* [26], *metaP-server* [27], *metaX* [28], *W4M* [29], *METDAT* [30], *MeltDB* [31] and *Metabolomics Workbench* [32]. On the one hand, *W4M*, *metaP-server* and *metaX* perform data integrations by signal drift corrections and batch effect removal using QC samples, which are considered to be effective only for the multiple batches within single analytical experiment [27–29]. *MetaboAnalyst* and *XCMS online* perform meta-analysis based on late stage data integration [33] and the exact feature names (compound IDs, spectral bins/peaks) [25,26], but this late stage integration may result in loss of intersection [5,6,26]. On the other hand, metabolite annotation using available reference databases still needs to be further enriched [18], and with the rapid accumulation of different metabolite functional databases, it becomes feasible to significantly enhance metabolite annotation. Therefore, it is urgently needed to develop a new tool able to integrate multiple analytical experiments and annotate metabolite functions with significantly enhanced/enriched annotation performance.

Herein, a novel tool MMEASE was developed to provide online integration of multiple metabolomics datasets by the enhanced metabolite annotation and enrichment analysis (https://idrblab.org/mmease/). MMEASE could (**1**) conduct data merging for multiple analytical blocks and remove the batch effect; (**2**) visualize sample separation and identify markers using a very diverse sets of methods or strategies; (**3**) provide enriched functional annotation for >330 thousands of metabolites; (**4**) conduct metabolite enrichment analysis using various categories/sub-categories. Moreover, the originality and usefulness of MMEASE were extensively exhibited by several case studies at the end of this work. All in all, the MMEASE aimed at supplying the comprehensive service for long-term and large-scale metabolomics, which could provide valuable guidance to the biological explanation of metabolites.

2. Materials and methods

2.1. Integration of multiple analytical experiments

To effectively boost the power of statistical analysis and accurately define biological variation, large-scale and long-term studies including hundreds/thousands of samples were highly demanded [34,35]. A large-scale study was frequently broken down to small analytical experiments, which were required to be integrated [36]. Herein, a new strategy based on the unique alignment ID for retention time (RT) and accurate mass (*m/z*) of a certain metabolite peak [6,37] was used to integrate multiple analytical experiments. If both RT and m/z fell into tolerable range, it was applicable to align the corresponding peaks of all samples. The workflow of this integration strategy was illustrated in Fig. 1. As shown, the list of unaligned metabolite peaks was *first* initialized with alignment ID set to 1. *Second*, the peak of the strongest signal was defined as tentative reference (the 1st step in **Phase 1 Alignment Strategy**), and the tentative aligned group composed of the peaks within the tolerable ranges of RT and *m/z* was identified using the tentative reference (the

2nd step in **Phase 1 Alignment Strategy**). *Third*, the peak of median retention time in the groups of tentative alignment was defined as target reference (the 1st step in **Phase 2 Alignment Strategy**). The final aligned group composed of the peaks in the tolerable ranges of RT and m/z was identified based on the target reference, and these peaks were denoted with the same alignment ID (the 2nd step in **Phase 2 Alignment Strategy**). *Fourth*, the aligned peaks were removed from the list of unaligned metabolite peaks with alignment ID adding one. To further reduce the number of peaks within the list of unaligned metabolite peak, the above steps in both phase of the **Alignment Strategy** were repeated until the successful alignment of all peaks.

For large-scale and long-term metabolomics, there were various unwanted variations along with data integration, because the measurements were affected by laboratory conditions and batch effects across various analytical experiments [38–41]. It was essential to remove the unwanted variations after data integration (Fig. 1) and to make preparation for the forthcoming statistical analysis [42]. As shown in **Supplementary Methods**, there were $\geq$ 5 methods applied in MMEASE for removing batch effects in multiple analytical experiments after data integration, which included batch mean-centering, global normalization, cross-platform normalization, etc. Users could choose the most appropriate method to remove batch effects as large as possible by the visualization of boxplots and principal components analysis (PCA) plots of dataset before and after batch effect removal.

2.2. Sample separation methods and marker identification strategies

After the integration of large-scale metabolomics and the removal of batch effect, sample separations were frequently applied for visualization. In MMEASE, four such methods for the sample separations were provided: hierarchical clustering, self-organizing feature map, *k*-means and principal component analysis (described in **Supplementary Methods**). Among all the relevant metabolomic tools (shown in **Supplementary Table S1**), only the MetaboAnalyst [25] provided all the four methods for sample separation and the number of methods provided in other tools was no more than two.

With the accumulation of feature's dimensionality and the development of computationally expensive methodology, marker identification strategy played critical roles in metabolomics [43–46]. Due to the great variations among the statistic theories and model assumption underlying different strategies, these strategies could lead to contradictory results for the same set of metabolomic data, which made the appropriate application of each strategy heavily dependent on the innate characteristics of the studied dataset [47]. To achieve systematic assessment on different marker identification strategies for both the integrated dataset using integration strategy and one single dataset, 13 popular strategies were offered in MMEASE, which included the fold change, PLS-DA, OPLS-DA, chi-squared test, student's *t*-test, entropy filtering, linear models and empirical *Bayes* method, relief, RF-RFE, correlation-based, SAM, SVM-RFE and Wilcoxon rank sum (**Supplementary Methods**). Among those 13 strategies provided in MMEASE, 5 were not appeared in any of the available metabolomics tools (**Supplementary Table S2**), which made the systematic evaluation of all strategies unique in MMEASE.

In order to identify the most appropriate strategy, two well-established criteria (classification capacity and robustness) were utilized for evaluating the methods and strategies described above. On one hand, three classic measurements (SEN, SPE and ACC for *sensitivity*, *specificity* and *accuracy*, respectively [48,49]) were used to assess the classification ability based on training and test datasets using *Support Vector Machine* (SVM) [50]. On the other hand, to evaluate the stability among the markers identified by different methods/strategies, the consistency score (*CS*) [45,51] among the markers derived from ten sampling sets was calculated and analyzed in this study.

Different from standalone software, the usage of online tools was frequently limited by the bottleneck of network throughput and calculation resource [52,53]. Therefore, various benchmark datasets were collected to measure the functionality and accessibility of MMEASE. The collected datasets included: MTBLS19 [54] (40 hepatocellular carcinoma (HCC) & 49 cirrhosis (CIR) patients, 1295 metabolites) and MTBLS17 [55] (60 HCC & 129 CIR patients, 1810 metabolites). As shown in **Supplementary Table S3**, the time costs of the majority of the assessed methods were less than 3 min with relief and SVM-RFE exceeding 10 min for dataset MTBLS17, which were acceptable consumption of time and resources for the metabolomics-based online tool [23].

2.3. Data collection for enriched metabolite annotation

The MS peaks of interest needed to be annotated with function based on their masses by scanning the reference library, but the traditional annotation was still imperfect in metabolomics especially for the untargeted one [18,56]. Particularly, less than 2% of the detected MS peaks could be annotated using untargeted MS-based metabolomics [15], and the way to convert a raw peak feature into a metabolite with biologically interpretable information remains a big challenge [18,57]. Thus, a total of 338,263 metabolites were *first* accumulated by systematic literature reviews and collected from metabolomic databases (including METLIN [58], HMDB [56], MMCD [59], LMSD [60], MoNA [61], etc.), which resulted in 107,071 endogenous, 124,451 exogenous and 169,352 peptide metabolites. *Then*, detailed annotation data were added for classifying the studied metabolites into biological or functional groups. This was accomplished also by literature reviews and by using information from available metabolite databases (HMDB [56], T3DB [62], ECMDB [63], YMDB [64], PMDB [65], Drugbank [66], CFAM [67], TCMID [68], KEGG [69], etc.), which provided abundant information on the exogenous factors of human life (food, plant, drug, cosmetics, microbe, toxin, environmental pollutant, etc.).

Based on the annotation results, the enrichment analysis could be performed to reveal the aggregation degree of functional role or exogenous source for the studied metabolite list. The over-representation analysis using hypergeometric test was applied for enrichment analyses. There were eight categories provided in MMEASE, which contained a large collection of metabolites for hypergeometric test and enrichment analysis. These categories included: (**1**) 310 metabolic pathways for 5399 compounds in KEGG pathways [69]; (**2**) 736 human metabolic and disease pathways of SMPDB [70]; (**3**) biological function classes reflecting the biological role of metabolite in blood, urine or cerebrospinal tissue; (**4**) structural category for human metabolites based on the structural similarity in Chemical Family [67]; (**5**) food group/source of metabolite in FooDB (http://foodb.ca/); (**6**) therapeutic class of the secondary metabolites from traditional medicine; (**7**) species taxonomy of metabolites from traditional medicine; and (**8**) category of toxins and environmental pollutants for metabolites [62].

2.4. Required data format of MMEASE inputs

For starting MMEASE analysis, the *csv* files containing feature-by-sample matrix should be prepared in advance. Each *csv* file contained five essential columns providing the information of isotope, mass, adduct, intensity and retention time, and different *csv* input files were prepared for different analytical experiments. Particularly, the first two columns of each *csv* file gave the mass and retention time, and samples must be kept in columns with the sample names in the first row. The group label in the second row indicated distinct sample groups such as case and control. Input data values (mass, retention time, intensity) should be numeric, and the blank or "NA" should be adopted to indicate any missing values. An example input file with the corresponding contents separated by comma (*csv*) was provided in the website of MMEASE. Moreover, when performing metabolite annotation and enrichment analysis, a file containing the studied *m/z* features and metabolites should also be properly provided.

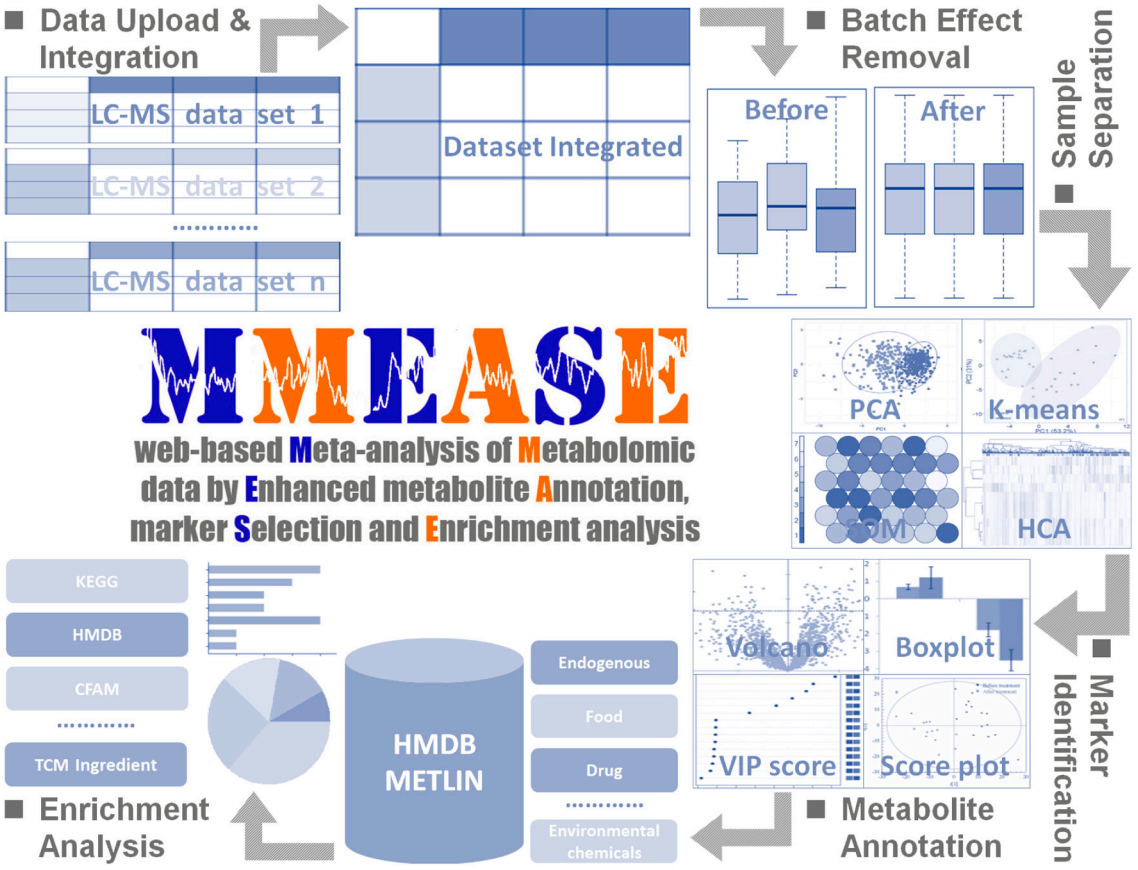

Fig. 2. The general workflow of MMEASE. (A) Data Upload & Integration & Batch Effect Removal; (B) Sample Separation & Marker Selection; (C) Metabolite Annotation & Enrichment Analysis.

3. Results and discussion

3.1. Introduction and operation process of MMEASE

There were five major steps for operating MMEASE: (**S1**) **Dataset Integration**: multiple datasets of different analytical experiments could be uploaded and integrated by MMEASE, and 5 methods were provided to remove unwanted experimental variations; (**S2**) **Sample Separation**: 4 sample separation methods were adopted to visualize the cluster and separation of samples; (**S3**) **Marker Identification**: 13 marker identification strategies were applied to identify metabolic markers for given datasets; (**S4**) **Metabolite Annotation**: metabolites could be annotated using mass spectra based on the adducts and multiply charged ions; (**S5**) **Enrichment Analysis**: a large number of metabolites annotated by eight categories were used for enrichment analysis. The general workflow of MMEASE was illustrated in Fig. 2, and the interactive visualization was provided in responding web page. All graphs could be exported as high-resolution images and the resulting *csv* files could be directly downloaded. Moreover, the Metabolite Annotation (S4) and Enrichment Analysis (S5) are set two independent steps in the beginning of "Analysis" panel. Users could perform metabolite annotation and enrichment analysis to access this page directly.

3.2. Strategy for integrating multiple analytical experiments

In order to effectively boost the power of statistical analysis and accurately define biological variation, multiple analytical experiments were integrated using the strategy shown in Fig. 1. The benchmark dataset MTBLS17 [55] were adopted to evaluate the performance of datasets integration. Particularly, the metabolomic abundance profiles

Table 1
Comparing the performances of biomarker identification based on the data of different experiments or experiment integrations. The datasets of EXP 1, 2 and 3 were collected from MTBLS17. Markers were discovered using OPLS-DA (VIP >1.5) and *t*-test (*p*-value <0.05). The true markers were provided in **Supplementary Table S5**. EF: enrichment factor.

Experiment / Integration	No. of Cases / Controls	No. of Detected Peaks	No. of Metabolites Annotated by Detected Peaks	No. of True Markers Contained in Detected Peaks	No. of Differential Peaks Identified	No. of Metabolites Annotated by Differential Peaks	No. of True Markers Contained in Differential Peaks	EF
EXP 1	60/129	1586	119,934	13	77	12,781	4	2.89
EXP 2	13/50	3230	155,768	13	68	9891	1	1.21
EXP 3	5/5	612	58,182	13	3	4167	1	1.07
EXP 1 & 2	73/179	984	73,970	13	94	12,511	7	3.18
EXP 1 & 3	65/134	256	32,857	11	13	3435	4	3.48
EXP 2 & 3	18/55	225	26,561	11	15	1759	4	5.49
EXP 1, 2 & 3	78/184	220	29,735	9	15	1963	6	10.1

Table 2
Comparing the prediction performances of 13 studied marker identification methods. SEN: sensitivity; SPE: specificity; ACC: accuracy; AVE No.: average number of biomarkers selected by 10 sampling datasets; CS: consistency score; MTBLS19: 40 hepatocellular carcinoma (HCC) patients & 49 cirrhosis (CIR) patients with 1295 metabolites; MTBLS92: 142 and 111 breast cancer patients before and after neoadjuvant chemotherapy with 670 metabolites; MTBLS28: 469 lung cancer patients and 536 healthy controls with 1588 metabolites.

Studied Methods	Method Type	MetaboLights ID	SEN	SPE	ACC	AVE No.	No. of Markers Co-selected by *n* Sampling Sets (*n* =)										CS	≥5
							10	*9*	*8*	*7*	*6*	*5*	*4*	*3*	*2*	*1*		
		MTBLS19	60.9%	79.0%	69.1%	127	3	8	19	19	21	21	25	48	87	231	4403	0.19
PLS-DA	Wrapper	MTBLS92	65.7%	59.2%	67.2%	43	9	3	1	2	3	10	11	16	38	57	3058	0.19
		MTBLS28	78.1%	69.7%	74.2%	144	50	23	16	20	16	15	28	38	58	88	18,030	0.40
		MTBLS19	69.6%	63.2%	66.7%	55	5	3	7	5	5	11	13	24	30	113	2570	0.17
Student's *t*-test	Filter	MTBLS92	71.6%	50.0%	62.2%	91	16	6	11	5	13	20	21	41	47	94	6309	0.26
		MTBLS28	70.9%	68.4%	69.7%	100	25	12	12	3	7	13	14	38	49	150	9197	0.22
		MTBLS19	34.5%	68.4%	50.0%	39	1	1	4	5	1	14	14	17	17	83	1035	0.17
CFS	Filter	MTBLS92	71.6%	57.7%	65.6%	125	4	4	7	42	40	38	34	26	46	86	4506	0.41
		MTBLS28	59.8%	63.3%	61.4%	20	3	4	1	2	1	2	0	8	16	44	1472	0.16
		MTBLS19	56.5%	68.4%	61.9%	58	2	7	11	15	15	13	12	7	21	38	3019	0.45
Fold Change	Filter	MTBLS92	53.9%	76.9%	63.9%	16	4	2	2	0	2	4	3	5	2	23	1496	0.30
		MTBLS28	75.7%	57.4%	66.7%	100	32	10	12	19	13	9	17	23	25	51	11,267	0.45
		MTBLS19	65.2%	84.2%	73.8%	127	2	10	18	12	29	21	37	42	65	249	4257	0.19
LMEB	Sophisticated filter	MTBLS92	76.1%	65.4%	71.4%	196	12	17	63	36	45	29	52	43	53	76	11,731	0.47
		MTBLS28	76.1%	66.5%	71.6%	100	39	16	11	16	12	10	7	18	15	32	13,599	0.59
		MTBLS19	78.3%	68.4%	73.8%	185	6	17	40	32	27	27	30	72	95	269	8303	0.24
OPLS-DA	Wrapper	MTBLS92	64.2%	57.7%	61.3%	88	5	5	11	13	26	21	27	34	36	60	3836	0.34
		MTBLS28	66.9%	67.4%	67.2%	217	65	21	29	20	31	45	43	40	65	122	22,997	0.44
		MTBLS19	60.9%	57.9%	59.5%	91	0	0	2	10	21	24	19	69	68	161	1258	0.15
SAM	Sophisticated filter	MTBLS92	76.1%	51.9%	65.6%	241	24	25	33	47	49	65	62	61	95	108	14,729	0.43
		MTBLS28	70.1%	66.1%	68.2%	100	28	13	20	14	17	9	9	22	24	48	11,008	0.50
		MTBLS19	56.5%	68.4%	61.9%	149	2	10	12	15	19	41	60	71	97	216	4151	0.18
WRS	Filter	MTBLS92	70.2%	50.0%	61.3%	76	0	0	1	2	14	32	38	48	52	92	908	0.18
		MTBLS28	71.3%	68.4%	69.9%	195	0	6	21	41	47	53	62	84	103	192	5119	0.28
		MTBLS19	61.5%	63.2%	66.7%	10	0	0	0	0	0	0	0	3	5	85	11	0.00
Chi-squared Test	Filter	MTBLS92	61.2%	55.8%	58.8%	4	0	0	0	0	0	0	0	1	1	37	3	0.00
		MTBLS28	63.0%	58.3%	60.8%	22	2	0	1	2	0	4	6	13	20	51	742	0.09
		MTBLS19	60.9%	68.4%	64.3%	100	1	1	82	5	3	1	1	4	60	129	5920	0.32
Entropy-based Filters	Filter	MTBLS92	71.6%	42.3%	58.8%	4	0	0	0	0	0	0	0	1	1	35	3	0.00
		MTBLS28	57.8%	37.2%	48.2%	22	2	0	1	2	0	4	6	13	20	51	742	0.09
		MTBLS19	73.2%	66.8%	70.1%	8	0	0	0	0	0	0	1	6	6	48	22	0.00
RF-RFE	Wrapper	MTBLS92	83.6%	63.5%	74.8%	9	0	1	0	1	0	3	1	3	5	40	199	0.09
		MTBLS28	59.0%	65.6%	62.1%	34	4	3	2	7	5	6	12	11	15	37	1973	0.26
		MTBLS19	73.9%	63.2%	69.1%	100	0	0	0	1	0	7	19	70	169	334	473	0.01
Relief	Sophisticated filter	MTBLS92	67.2%	51.9%	60.5%	100	0	0	0	1	5	15	40	81	149	187	703	0.04
		MTBLS28	64.1%	44.0%	54.8%	100	0	0	0	0	0	6	15	60	155	420	383	0.01
		MTBLS19	73.9%	57.9%	66.7%	100	1	9	17	17	21	30	30	42	38	56	3858	0.36
SVM-RFE	Wrapper	MTBLS92	68.7%	67.3%	68.1%	100	2	7	9	6	18	33	39	51	62	97	3048	0.23
		MTBLS28	72.9%	71.6%	72.3%	100	0	2	6	4	13	12	28	46	128	262	1404	0.07

Table 3
Significantly enriched metabolite annotation in MMEASE based on the specific classes / groups under eight annotation categories together with the number of metabolites under each class / group. A variety of representative classes / groups were listed under each category.

Food Components and Food Additives (by *food class*)					
Animal foods	**153**	Aquatic foods	**267**	Beverages	**557**
Cereals / cereal products	**659**	Cocoa / cocoa products	**59**	Coffee / coffee products	**238**
Fats / oils	**770**	Fruits	**3506**	Herbs / spices	**2722**
Milk / milk products	**120**	Nuts	**247**	Pulses	**807**
Soy	**24**	Teas	**899**	Vegetables	**3273**
Plant Metabolites and Agricultural Chemicals (by *chemical group*)					
Acaricide	**142**	Alkaloid	**57**	Carbohydrate	**8**
Flavonoid	**31**	Fungicide	**332**	Herbicide	**390**
Hormone	**34**	Insecticide	**441**	Pigment	**24**
Plant extract	**10**	Quinone	**12**	Sterol	**9**
Terpenoid	**30**	Plant growth regulator	**94**	Rodenticide	**41**
Small Molecular Drugs and Drug Metabolites (by *disease class*)					
Anti-infection drugs	**423**	Blood-forming organ disease drugs	**19**	Circulatory system disease drugs	**262**
Digestive system disease drugs	**89**	Genitourinary system disease drugs	**74**	Immune system disease drugs	**25**
Musculoskeletal disease drugs	**81**	Nervous system disease drugs	**151**	Respiratory system disease drugs	**117**
Skin disease drugs	**84**	Visual system disease drugs	**52**	Endocrine / metabolic disease drugs	**181**
Neurodevelopmental disorder drugs	**162**	Anti-cancer drugs	**299**	Drug metabolites	**1041**
Metabolites / Secondary Metabolites of Traditional Medicine (by *species taxonomy*)					
Actinobacteria	**297**	Arthropoda	**75**	Ascomycota	**461**
Basidiomycota	**510**	Chlorophyta	**19**	Chordata	**175**
Cnidaria	**27**	Cyanobacteria	**16**	Echinodermata	**9**
Firmicutes	**37**	Mollusca	**58**	Phaeophyceae	**20**
Porifera	**33**	Proteobacteria	**194**	Streptophyta	**11,284**
Metabolites / Secondary Metabolites of Traditional Medicine (by *therapeutic class*)					
Astringent medicinal	**175**	Blood-activating / stasis-resolving	**175**	Dampness-draining diuretic medicinal	**156**
Dampness-resolving medicinal	**33**	Digestant medicinal	**67**	Exterior-releasing medicinal	**171**
Heat-clearing medicinal	**396**	Hemostatic medicinal	**144**	Interior-warming medicinal	**138**
Liver-pacifying / wind-extinguishing	**47**	Orifice-opening medicinal	**36**	Phlegm-resolving / cough-suppressing	**282**
Purgative medicinal	**124**	Qi-regulating medicinal	**205**	Tonifying and replenishing medicinal	**659**
Tranquillizing medicinal	**209**	Wind-dampness-dispelling medicinal	**116**	Worm-killing and itching relieving	**42**
Cosmetic Substances and Ingredients (by *cosmetic functional class*)					
Absorbent	**10**	Antimicrobial	**51**	Antioxidant	**101**
Antistatic	**88**	Astringent	**14**	Binding	**26**
Bleaching	**13**	Buffering	**66**	Bulking	**10**
Chelating	**23**	Cleansing	**30**	Cosmetic colorant	**40**
Denaturant	**27**	Deodorant	**18**	Emollient	**148**
Emulsifying	**76**	Emulsion stabilising	**35**	Film forming	**31**
Foam boosting	**15**	Hair conditioning	**100**	Hair Dyeing	**34**
Humectant	**82**	Masking	**241**	Moisturising	**10**
Opacifying	**28**	Oral care	**22**	Perfuming	**895**
Plasticiser	**29**	Preservative	**32**	Reducing	**13**
Skin conditioning	**490**	Skin protecting	**26**	Solvent	**99**
Surfactant	**59**	UV absorber	**21**	Viscosity controlling	**116**
Toxins, Environmental Pollutants and Microbial Metabolites (by *toxin origin class*)					
Airborne pollutant	**201**	Animal toxin	**240**	Bacterial toxin	**17**
Cigarette toxin	**60**	Drug toxin	**648**	Escherichia metabolite	**4459**
Food toxin	**911**	Fungal toxin	**110**	Household toxin	**409**
Industrial/workplace toxin	**436**	Natural toxin	**720**	Other microbial metabolite	**132**
Pesticide	**448**	Plant toxin	**190**	Pollutant	**277**
Synthetic toxin	**1258**	Uremic toxin	**56**	Yeast metabolite	**2978**
Chemical Family (by *compound chemical class*)					
Adenosine A2A receptor binder	**318**	Adenosine receptor DNA binding nucleoside	**116**	ADP/ATP translocase 1 and *E*-FABP binder	**240**
Aldehyde oxidase inhibitor isovanillin	**324**	ALOX5 and PPAR-gamma binder	**133**	Androgen receptor ligand anabolic steroid	**319**
Cannabinoid receptor 1 ligand marinol	**123**	CB1 ligand oleoylethanolamide	**116**	CI-SDAP and SDHA binder	**163**
COX inhibitor salicylate derivatives	**144**	COX-1 inhibitor gamma-homolinolenic acid	**148**	DD-transpeptidase inhibitor beta-lactams	**559**
DNA binder methoxsalen	**177**	Ethyl icosapentate	**177**	FAAH1 inhibitor 2-arachidonoylglycerol	**129**
FAAH1 binder oxiranylmethyl hexadecenoate	**102**	Factor X inhibitor LMWH fondaparinux	**216**	Ganglioside GM1	**344**
GCR binding corticosteroid	**176**	GABAR ligand l-alanyl-L-glutamine	**166**	HSP-90 inhibitor cromoglicate	**2580**
Ipriflavone	**114**	Lanosterol synthase and TOP2-beta binder	**154**	Neutrophil elastase and PKC-alpha binder	**104**
PKC-alpha inhibitor dioctanoyl-sn-glycerol	**448**	Salicin/schizophyllan	**562**	Sodium/potassium ATPase binder	**322**
Sphingosine kinase 2 inhibitor phenoxodiol	**301**	STAT-3 inhibitor RTA 402	**189**	Steroid 5-reductase inhibitor azelaic acid	**119**
Taste receptor ligand cynarin	**133**	Tyrosinase inhibitor askendoside B	**136**	VDR ligand vitamin D analog	**284**

of 78 hepatocellular carcinoma (HCC) patients and 184 cirrhotic (CIR) controls were detected from serum tissue using UPLC-QTOF-MS. For this benchmark dataset, the authors generated UPLC-QTOF MS data from sera of 78 hepatocellular carcinoma (HCC) cases and 184 cirrhotic (CIR) controls in three separate experiments (Exp1, Exp2, and Exp3). The three experiments were conducted in May 2010, July 2010, and March 2011, respectively. The first, second, and third experiments in MTBLS17 provided 189 sample profiles (129 CIR controls and 60 HCC patients), 63 samples (50 CIR controls and 13 HCC patients), and 10 samples (5 CIR controls and 5 HCC patients), respectively. The samples of any two or all three experiments were integrated into large dataset using the new integration strategy, and OPLS-DA and student's *t*-test were applied for marker identification. The features of statistic difference between patient and control groups were revealed by p-value <0.05 and VIP >1.5 [71]. These revealed features were subsequently annotated [56] by using 20 ppm as the m/z tolerance [72]. To assess the false discovery

rate, the experimentally validated true markers could be adopted as the golden standard, and the enrichment factor (EF) could be further assessed [73]. EF was an effective indicator measuring the level of increase in the probability of discovering the golden standard from the identified data over the random pick from all metabolites without any identification [74]. The higher the value of EF, the higher the chance to identify true markers [75].

Herein, 13 metabolites that were reported to be differentially expressed between CIR individuals and HCC patients [6] were *first* collected and provided in **Supplementary Table S4**. As shown in Table 1, the number of true markers in the metabolites detected from different datasets varied from 9 to 13. On the one hand, the number of true markers revealed by the single experiment was smaller than that of the integrated dataset. On the other hand, the EF of all single experiments was substantially smaller than that of the integrated datasets. These results demonstrated the powerful performance of the new strategy proposed in MMEASE in controlling the false discovery rate. Moreover, if all three datasets were integrated, the resulting number of the identified true markers (6) and the EF (10.1) were further enhanced, which further illustrated the power of the strategy in controlling false discovery rate.

3.3. Diverse strategies for marker identification in MMEASE

Different marker identification strategies were reported to result in very different outcomes in current metabolomics [76,77]. To assess the performances of different marker identification strategies, three benchmark datasets (MTBLS28 [78], MTBLS92 [79], MTBLS19 [55]) of various sample sizes and metabolites were collected. Particularly, 1588 metabolites were identified in the sample of 536 healthy people and 469 patients of lung carcinoma (MTBLS28); 670 metabolites were detected in the serum of breast cancer patients before (142) and after (111) neoadjuvant chemotherapy (MTBLS92); 1296 metabolites were detected in the serum of 40 HCC patients and 49 CIR controls (MTBLS19).

Two indicators (*classification capacity* and *robustness*) were used to assess the performances of each strategy. (**1**) To evaluate classification capacities, half samples were randomly selected and used for constructing training set, and the remaining half were set as test set for each benchmark dataset. After marker identification, the classification models were trained using SVM, and the test dataset was used for evaluating the performances (SEN, SPE and ACC) of the SVM model. As shown in Table 2, there was significant difference among the classification capacities assessed by different datasets for given strategy. For example, for PLS-DA, the classification accuracy was 69.05%, 67.22% and 74.20% for MTBLS19, MTBLS92 and MTBLS28, respectively. Moreover, classification accuracies were greatly different among strategies as assessed by a given dataset. Specifically, the resulting strategies of the highest ACC assessed by MTBLS19, MTBLS28 and MTBLS92 were LMEB, PLS-DA and RF-RFE, respectively. (**2**) To assess the robustness of marker discovery, ten sub-datasets were generated using half of the whole dataset by 10-time random sampling. Then, ten sets of markers were identified using different strategies for each sub-dataset. As shown in Table 2, there was significant difference in the robustness as assessed by different datasets even using the same strategy. Taking the Student's *t*-test as examples, the number of markers discovered for MTBLS19, MTBLS92 and MTBLS28 was 5, 16 and 25. The proportion of marker discovered by $\geq$5 sub-datasets was 0.17, 0.26 and 0.22, respectively. The calculated CS values were 2570, 6309 and 9197. Furthermore, the robustness was significantly different using different marker identification strategies for the given dataset. The most proportion of markers identified by $\geq$5 sub-datasets for MTBLS19, MTBLS28 and MTBLS92 was 0.45, 0.47 and 0.59 with the CS value of 3019, 13,599 and 11,731 using FC, LMEB and LMEB method, respectively. Therefore, the characteristics of the studied data was critical in selecting the appropriate strategy, and only when the strategy was correctly applied, could the identified biomarkers be capable of answering the studied metabolomic

Table 4
Significantly enhanced annotation by MMEASE compared with HMDB based on the differential metabolites from NASA twins study [86].

Name	HMDB	MMEASE
2-aminophenol sulfate	*Endogenous* *Food*	*Endogenous*: Increased after a 5-week high dietary fiber intake in human plasma (*Anal Bioanal Chem.* 405:4799–809, 2013) *Food*: Observed after rye bread consumption in urinary metabolic profiles (*Am J Clin Nutr.* 99:1286–308, 2014)
2-hydroxyphenylacetate	*N.A.*	*TCM Ingredients*: Chemical constituents from *Averrhoa carambola* L (*Molecules.* 17:12330–40, 2012) *Carcinogenic Potency*: Associated with oxidative stress in post-operative epithelial ovarian cancer (*Sci Rep.* 6:23334. 2016)
3-indolepropionic	*N.A.*	*Cosmetic*: Ingredient for use as antistatic, for hair conditioning and masking (*EFSA Journal*, 12:3826, 2014) *Endogenous*: Clinical association between the metabolite and kidney disease (*Clin Nutr.* doi:https://doi.org/10.1016/j.clnu.2018.11.029, 2018)
3-indoxyl sulfate	*N.A.*	*Food*: Compound of plant growth regulator in food samples (*Food Chem.* 170:123–30, 2015) *Drug*: An extensively studied uremic solute (*Toxins.* 8:E358, 2016) *Endogenous*: A potent endogenous agonist for the human aryl hydrocarbon receptor (*Biochemistry.* 49: 393–400, 2009) *Food*: Decrease in the urinary excretion when the reduced intake of diets (*Am J Nephrol.* 14:207–12, 1994) *Microbe*: The bacterial metabolism of tryptophan in the colon (*Kidney Int Suppl.* 114:S12-9, 2009) *Toxins/Pollutant*: A circulating uremic toxin stimulating glomerular sclerosis process (*J Lab Clin Med.* 124:96–104, 1994)
3-phenylpropionate	*N.A.*	*Drug*: Compound of anti-AIDS agent (*Tetrahedron Lett.* 39:8329–32, 1998) *Drug Metabolite*: Intermediate of anti-AIDS drugs (*Tetrahedron Lett.* 39:8329–32, 1998) *Endogenous*: Protecting alpha-chymotrypsin from denaturation by ethanol / urea (*Proc Soc Exp Biol Med.* 130:1046–7, 1969) *Escherichia coli Metabolite*: Metabolite of the intestinal microflora (*J Mol Microbiol Biotechnol.* 3:127–33, 2001) *Food*: Compound in cultured dairy products and cheese (*Int Dairy J.* 5:227–46, 1995) *Food Additives*: Natural bio-preservatives (*Trends Food Sci Tech.* 33:93–109, 2013) *Microbe*: Metabolite of the intestinal microflora (*J Mol Microbiol Biotechnol.* 3:127–33, 2001)
4-methylcatechol sulfate	*N.A.*	*TCM Ingredients*: Chemical constituents from Galiun verum L. (*Chin J Med Chem.* 4:241–89, 2009) *Carcinogenic Potency*: Carcinogenicity in F344 rats (*Int J Cancer.* 56:146–52, 1994) *Drug*: Induces growth factor synthesis and restores morphological integrity of sensory fiber (*Neurobiol Dis.* 8:626–35,

(*continued on next page*)

Table 4 (*continued*)

Name	HMDB	MMEASE
Betaine	*Endogenous* *Food* *Plant*	2001) *Endogenous*: Metabolites characteristic of gut bacteria metabolism of polyphenols were increased (*PLoS One*. 8:e72215, 2013) *Food*: Found in many foods such as coffee, beer, arabica coffee and cocoa powder (*Appl Environ Microbiol*. 70: 1804–10, 2004) *TCM Ingredients*: Metabolites of quercetin, a potential drug for neuroprotection / anticancer (*Cell Biol Int*. 39:770–4, 2015) *Cosmetic*: Compound of hair conditioning agent; humectant & skin-conditioning agent (*Canada Gazette Part I*, 132:5, 1998) *Drug*: Therapy for patients with NASH (*Am J Gastroenterol*. 96:2711–7, 2001) *Endogenous*: Betaine was reported to be decreased in schizophrenia (*EBioMedicine*. doi:https://doi.org/10.1016/j.ebiom.2019.05.062, 2019) *Escherichia Coli Metabolite*: Metabolite of Escherichia coli from glycine betaine (*Appl Environ Microbiol*. 80:4745–56, 2014) *Food*: Compounds commonly found in the western diet (*Food Chem*. 83:197–204, 2003) *Microbe*: Compatible solutes necessary for osmoregulation in microbes (*Am J Bot*. 100:1692–705, 2013) *Plant*: Compounds in a variety of plants (*Environ Exp Bot*. 59:206–16, 2007) *TCM Ingredients*: Chemical constituents of Cistanche species (*J Pharm Biomed Anal*. 162:16–27, 2019) *Yeast Metabolite*: Metabolite of Saccharomyces cerevisae (*J Chromatogr A*. 1217:8161–6, 2010) *Carcinogenic Potency*: Carcinogenicity in Syrian hamsters (*J Steroid Biochem*. 24:353–6, 1896) *Cosmetic*: Compounds in leave-on products (*Food Chem Toxicol*. 41:885–95, 2003) *Drug*: Compounds of catechol drugs (*J Neurosci Res*. 5:587–98, 1980) *Endogenous*: Catechol sulfate is an endogenous metabolite (*EBioMedicine*. 17:57–66, 2017)
Catechol sulfate	*N.A.*	*Escherichia Coli Metabolite*: Metabolite from benzoates by Escherichia coli cells (*Appl Environ Microbiol*. 50:1409–13, 1985) *Food*: Compound from various food sources (*Am J Clin Nutr*. 79:727–47, 2004) *Microbe*: Metabolite from benzoates by Escherichia coli cells (*Appl Environ Microbiol*. 50:1409–13, 1985) *TCM Ingredients*: Compound exists in some traditional Chinese medicines (*Analyst*. 138:1141–8, 2013) *Toxins/Pollutant*: Toxic to bacteria and higher organisms (*Environ Toxicol Chem*. 20:239–47, 2001)
Chenodeoxycholic acid	*Endogenous* *Food*	*Carcinogenic Potency*: Alter cell cycle in colonic crypts undergoing neoplastic changes (*Cancer Lett*. 76:101–7, 1994) *Endogenous*: An endogenous antagonist of the G-coupled formyl peptide receptors (*Inflamm Res*. 67:21–30, 2018) *Food*: Compounds in feed (*Am J Dig Dis*. 19:877–86, 1974) *Herbal Ingredients* In-Vivo *Metabolism*: Metabolite of Qingkailing (*Biochem Pharmacol*. 63:533–41, 2002) *Endogenous*: A metabolite biomarker in intrahepatic cholestasis of pregnancy (*Clin Chim Acta*. 471:292–7, 2017)
Chenodeoxycholic acid 3-sulfate	*Endogenous* *Food* *Microbe*	*Food*: Compounds in feed (*Am J Dig Dis*. 19:877–86, 1974) *Microbe*: Metabolite of microbial flora of the colon (*Clin Chim Acta*. 436:207–16, 2014)
Chenodeoxycholic acid glycine conjugate	*Endogenous* *Food* *Microbe*	*Endogenous*: Serum metabolite biomarkers for hepatocellular carcinoma diagnosis (*World J Gastroenterol*. 19:3423–32, 2013) *Food*: Compounds in feed (*Am J Dig Dis*. 19:877–86, 1974) *Microbe*: Metabolite of microbial flora of the colon (*Clin Chim Acta*. 436:207–16, 2014)
Cholic acid	*Endogenous* *Food*	*Drug*: Therapy for hereditary defects of primary bile acid synthesis (*Biochem Pharmacol*. 63:533–41, 2002) *Endogenous*: Endogenous cholic acid in liver disease due to cystic fibrosis (*Hepatology*, 39:1673–82, 2004) *Escherichia Coli Metabolite*: Metabolite of the intestinal microflora (*Anal Chem*. 80:2939–48, 2008) *Food*: Compounds in the diet (*Eur J Biochem*. 267:4272–80, 2000) *Food Additives*: Surfactant permitted as additives in food (*J Chromatogr A*. 847:369–75, 1999) *Microbe*: Metabolite of the intestinal microflora (*Anal Chem*. 80:2939–48, 2008) *TCM Ingredients*: Chemical constituents of Niuhuang (*Biochem Pharmacol*. 63:533–41, 2002) *Toxins/Pollutant*: Potent toxic to membrane disruption (*Am J Physiol Gastrointest Liver Physiol*. 284:G349-56, 2003)
Dihydro-coumaric acid	*N.A.*	*Food*: A food component in tomato (*Int J Biol Macromol*. 133:284–93, 2019) *Endogenous*: A key role for glycocholic acid in human liver cancer (*Clin Chim Acta*. 418:86–90, 2013) *Food*: Food additives and ingredients (*Steroids*. 86:62–8, 2014)
Glycocholic acid	*Endogenous* *Food* *Microbe*	*Food Additives*: A food additive and an emulsifying agent (*Steroids*. 86:62–8, 2014) *Microbe*: Glycocholic acid had lowest binding energy from Lactobacillus gasseri FR4 (*Front Microbiol*. 8:1004, 2017) *TCM Ingredients*: Major bioactive constituents in traditional Chinese medicines (*J Chromatogr A*. 1218:107–17, 2011) *Cosmetic*: Originated from cosmetics (*Indoor Air*, 7:17–32, 1997) *Drug Metabolite*: Metabolite generate form benzoic acid (*J Pharm Sci*. 61:1278–84, 1972)
Hippuric acid	*Endogenous* *Food* *Plant* *Microbe*	*Endogenous*: The presence of hippuric acid endogenously in all individuals (*Bull Environ Contam Toxicol*, 63:1–8, 1999) *Food*: Metabolite of dietary components (*Bosn J Basic Med Sci*. 8:38–43, 2008) *Herbal Ingredients* In-Vivo *Metabolism*:

(*continued on next page*)

Table 4 (*continued*)

Name	HMDB	MMEASE
Hydroxyhippurate	*N.A.*	Metabolism of some nephrotoxic herbs (*Curr Med Chem.* 20:2812–9, 2013) *Microbe*: Antibacterial action (*J Lab Clin Med.* 54:881–8, 1959) *Plant*: Simple phenols in fruits (*Bosn J Basic Med Sci.* 8:38–43, 2008) *TCM Ingredients*: Chemical constituent of Tongxinluo (*J Pharm Biomed Anal.* 56:86–92, 2011) *Toxins/Pollutant*: Uremic syndrome (*Semin Nephrol.* 16:167–82, 1996) *Drug*: Directly inhibited palmitate oxidation in control and Reye's syndrome cells (*Biochim Biophys Acta.* 1454:115–25, 1999) *Endogenous*: Contribution to impaired ligand binding by plasma in azotemic humans (*Biochem Pharmacol.* 36:4215–20, 1987) *Food*: The strongest urinary markers of consuming a mix of red wine and grape juice (*J Agric Food Chem.* 60:3078–85, 2012) *Microbe*: The main microbe metabolites formed from the catechin diet (*J Nutr.* 133:461–7, 2003) *Endogenous*: Interference on routine serum (*Clin Biochem.* doi:https://doi.org/10.1016/j.clinbiochem.2019.06.010, 2019)
Hydroxyphenyllactic acid	*Endogenous* *Plant* *Microbe*	*Food*: A chemical compound found in olive oil and beer (*Food Chem.* 205:212–20, 2016) *Microbe*: One of bacterial metabolites (*J Microb Pathophysiol Pathogenesis.* 2:003, 2016) *Plant*: Component in the drying leaves of *Perilla frutescens* (*Food Sci Nutr.* 7:1494–501, 2019) *Yeast Metabolite*: Found in phenol-induced cells of yeast Candida maltosa SBUG 700 (*J Basic Microbiol.* 36:239–43, 1996) *Endogenous*: Production from L-tryptophan by species of Phytophthora (*Can J Microbiol.* 14:595–600, 1968)
Indolelactic acid	*Endogenous* *Microbe*	*Food*: Indolelactic acid characterize average Danish diet (*Mol Nutr Food Res.* 63:e1800215, 2019) *Microbe*: Converted by intestinal bacteria such as Bacteroides, Clostridia, and *E. coli* (*Front Cell Infect Microbiol.* 8:13, 2018) *Cosmetic*: Estrogenic potential and possible risks in foods and cosmetics (*Toxicol Lett.* 209:146–53, 2012)
Methyl-4-hydroxybenzoate sulfate	*N.A.*	*Food*: The predominant constituent of oil of wintergreen (*Crit Rev Toxicol.* 47:98–120, 2017) *Food Additives*: A flavoring agent in products such as chewing gum and berries (*Crit Rev Toxicol.* 47:98–120, 2017) *Microbe*: As an anti-Microbe agents (*Biomed Res Int.* 2018:8308640, 2018) *TCM Ingredients*: Induce plant defense against insect herbivores and Microbe pathogens (*J Chem Ecol.* 33:1089–103, 2007) *Toxins/Pollutant*: Toxicity levels of an acceptable daily intake of 0.5 mg/kg/d (*Crit Rev Toxicol.* 47:98–120, 2017)
Phenyllactate	*N.A.*	*Drug*: Antifungal activity against molds isolated from bakery products (*Appl Environ Microbiol.* 69:634–40, 2003) *Endogenous*: The disease is elevated

Table 4 (*continued*)

Name	HMDB	MMEASE
Phenol sulphate	*Endogenous*	levels of phenylalanine metabolites (*Anal Biochem.* 280:242–9, 2000) *Escherichia Coli Metabolite*: A component in an Escherichia coli whole-cell biocatalyst (*Bioresour Technol.* 287:121423, 2019) *Microbe*: The production in an Escherichia coli whole-cell biocatalyst (*Bioresour Technol.* 287:121423, 2019) *Plant*: Useful compound from pyrolytic lignins (*Molecules.* 22:E372, 2017) *Yeast Metabolite*: A component in co-substrates (glucose, yeast extract, and glycerol) (*Bioresour Technol.* 287:121423, 2019) *Endogenous*: Increased in rat plasma gavaged with purified cranberry procyanidins (*Mol Nutr Food Res.* 59:2107–18, 2015)
p-cresol sulfate	*Endogenous* *Microbe*	*Escherichia Coli Metabolite*: Phenolic acids inhibited Escherichia coli (*Appl Environ Microbiol.* 67:1063–9, 2001) *Microbe*: Increased excretion in the bacterially modified metabolite (*Sci Rep.* 7:43326, 2017) *Endogenous*: Biomarker of serum and urine in overweight / obese young men (*Asia Pac J Clin Nutr.* 27:1067–76, 2018) *Food*: Acting on food components that escape absorption in the small bowel (*Clin J Am Soc Nephrol.* 7:982–8, 2012) *Microbe*: In the gut microbiome of rats (*Am J Physiol Renal Physiol.* 310:F857-71, 2016) *Toxins/Pollutant*: Cause renal damage and dysfunction (*Sci Rep.* 9:3207, 2019)
p-cresol glucuronide	*Endogenous*	*Endogenous*: Urine potential biomarker in general Chinese pregnant women (*J Chromatogr A.* 1479:145–52, 2017)
Sulfolithocholyl glycine	*Endogenous* *Food* *Microbe*	*Endogenous*: Cholylglycine and sulfolithocholylglycine in focal and diffuse hepatopathies (*Med Clin.* 85:653–5, 1985) *Microbe*: A secondary bile acid in microbial flora (*J Chromatogr B Analyt Technol Biomed Life Sci.* 846:69–77, 2007)
Taurocholic acid	*Endogenous* *Food*	*Food*: Postprandial metabolite higher than fasting level in controls and all liver diseases (*Korean J Intern Med.* 1:37–42, 1986) *Drug*: Cholagogues and Choleretics (*Rev Hosp Clin Fac Med Sao Paulo.* 51:116–20, 1996) *Endogenous*: Active promoting factor and biomarker of progression of liver cirrhosis (*BMC Gastroenterol.* 18:112, 2018) *Food*: Emulsifier in foods (*Food Chem.* 179:270–7, 2015) *Food Additives*: Diet supplement (*Nature.* 487:104–8, 2012) *Microbe*: Intestinal bacteria stimulants (*Gut Microbes.* 7:201–15, 2016) *TCM Ingredients*: Chemical constituents of Si Jun Zi Tang (*Biomed Pharmacother.* 111:1132–40, 2019) *Yeast Metabolite*: Metabolite of sulfur compound in yeast (*Plant Physiol.* 1:337–47, 1926)
Tauroursodeoxycholic acid	*Endogenous* *Food*	*Endogenous*: one of Endogenous bile acids (*Trends Mol Med.* 14: 54–62, 2008) *Drug*: FDA-approved for the treatment

(*continued on next page*)

Table 4 (*continued*)

Name	HMDB	MMEASE
		of certain cholestatic liver diseases (*Glob Adv Health Med.* 3:58–69, 2014) *Food*: Bear bile contains large amounts (*Microb Cell Fact.* 18:34, 2019)

problem.

3.4. Significantly enriched metabolite annotation in MMEASE

Table 3 gave the significantly enriched metabolite annotations in MMEASE. By the metabolite annotation, more clues and information about function and source of metabolites (such as drug metabolite, microbial metabolites, *etc*) [80,81] could be shown in result page for particular study. Particularly, there were eight annotation categories which included chemical family, cosmetic substance and ingredient, food component and food additive, plant metabolite and agricultural chemical, small molecular drug and drug metabolite, toxin, environmental pollutant and microbial metabolite, and species taxonomy and therapeutic class of secondary metabolites for traditional medicines. Taking the category of "cosmetic substance and ingredient" as example, there were 2960 metabolites which could be further classified into 36 sub-categories according to the cosmetic functional classes, and the sub-category "Perfuming" and "Skin conditioning" were among the sub-categories of the largest number of metabolites.

Moreover, 26 metabolites were found changed in small-molecule microbiome-derived metabolomics data, which suggested that the microbiome undergone the functional change in response to spaceflight from NASA twins study [69]. Based on the metabolite library in MMEASE with the detailed category information, these 26 metabolites were fully annotated. As shown in Table 4, compared with HMDB, significantly enhanced annotations for these 26 metabolites were achieved by MMEASE. Particularly, 16 out of these 26 metabolites were successfully annotated by both HMDB and MMEASE (with the annotation results of MMEASE significantly enriched by both category and sub-category data). For instance, betaine was annotated to be endogenous, food component and plant metabolite using HMDB, while the MMEASE could provide much more enriched clues (cosmetic substance, drug, endogenous, microbial metabolite, food component and traditional medicine ingredient). Moreover, 10 metabolites could be only annotated by the MMEASE (shown in Table 4). For example, the catechol sulfate was annotated as carcinogenic compound, cosmetic substance, drug, endogenous chemical, microbe, food component, traditional medicine ingredient and toxins/pollutant.

Moreover, enrichment analyses could be performed by including 8 functions of 338,263 metabolites. The metabolite enrichment analysis is a separate step, a list of metabolites known to users could be enriched for MMEASE. The names or compound IDs (such as PubChem, CAS and KEGG ID) of qualitative metabolites are adopted as input in this module to guarantee the accuracy of results. Then, hyper-geometric test is used for enrichment of a category (e.g., a pathway or a chemical family) in a metabolite signature. Herein, the enrichment of KEGG pathway and Chemical Family was performed using a list of differential metabolites correlated with the human age in urine samples in the study of Thevenot et al [82]. The chord diagram (Fig. 3A) using *GOChord* function of *GOplot* package and bar plot (Fig. 3B) using *ggplot* function of *ggplot2* package presented the first ten KEGG pathways and Chemical Families based on the *p*-value in the hyper-geometric test. Through the enrichment analysis in MMEASE, users could get more clues of function and structure about the differential metabolites in the further study. For example, hsa00232 (Caffeine metabolism) and CFFHM189 (Imidazopyrimidines Metabolite 1,3-Dimethyluric acid Family) were enriched, and both were supported by previous studies about age [83–85].

4. Conclusions

To facilitate the studies of large-scale and long-term metabolomic analysis, MMEASE was developed to enable the integration of multiple analytical experiments. The MMEASE was unique in providing (**1**) a strategy to integrate multiple analytical experiments, (**2**) diverse marker identification strategies and (**3**) significantly enriched metabolite annotation.

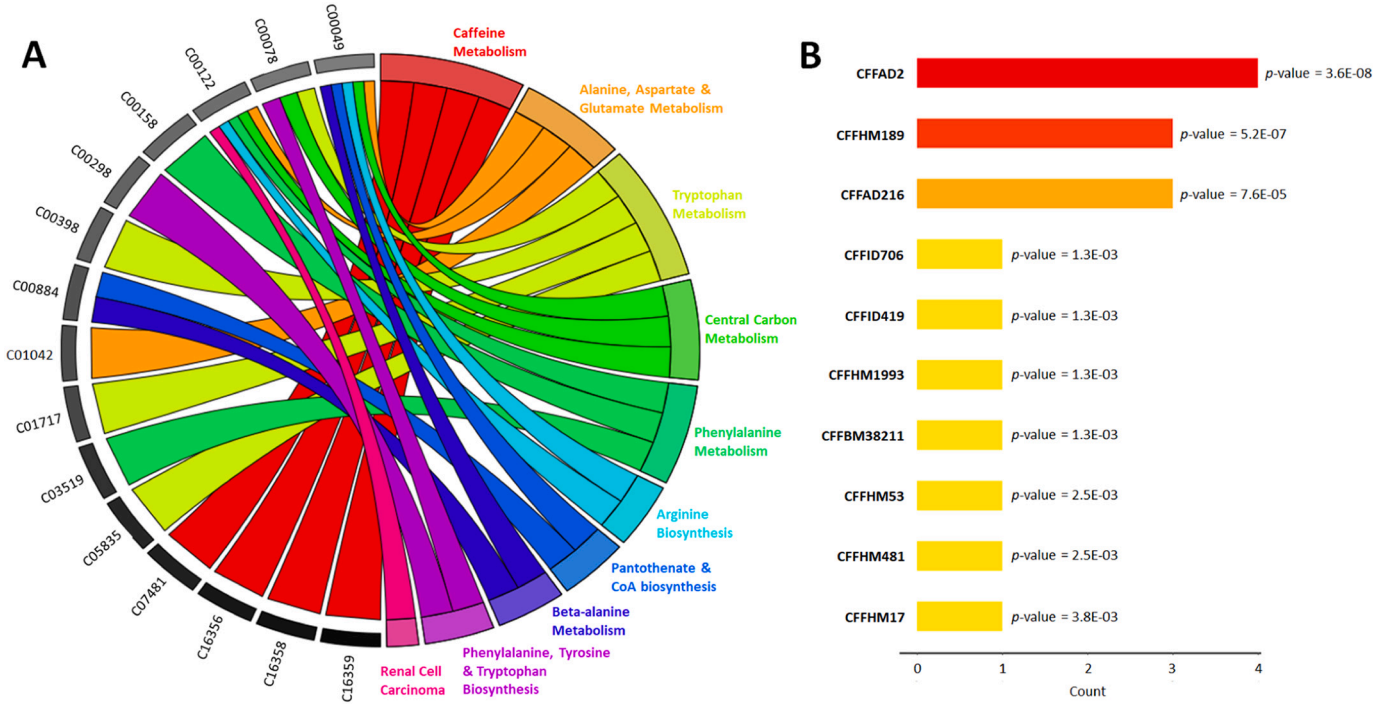

Fig. 3. Enrichment results of pathways and chemical family according to the differential metabolites identified from the benchmark dataset [82].

CRediT authorship contribution statement

Qingxia Yang: Conceptualization, Methodology, Software, Writing - original draft. **Li Bo:** Methodology, Software. **Sijie Chen:** Methodology, Software. **Jing Tang:** Methodology, Visualization. **Li Yinghong:** Visualization. **Li Yi:** Methodology. **Song Zhang:** Investigation. **Cheng Shi:** Investigation. **Ying Zhang:** Validation. **Minjie Mou:** Validation. **Weiwei Xue:** Supervision, Validation. **Feng Zhu:** Writing - review & editing.

Declaration of Competing Interest

The authors declare no competing interests.

Data availability

Example data is provided in the web-server (https://idrblab.org/mmease/).

Acknowledgements

Funded by National Key Research and Development Program of China (2018YFC0910500), National Natural Science Foundation of China (81872798), Fundamental Research Fund for Central University (2018QNA7023, 10611CDJXZ238826, 2018CDQYSG0007 & CDJZR14468801), and Innovation Project on Industrial Generic Key Technologies of Chongqing (cstc2015zdcy-ztzx120003).

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jprot.2020.104023.

References

[1] R.K. Naviaux, J.C. Naviaux, K. Li, A.T. Bright, W.A. Alaynick, L. Wang, A. Baxter, N. Nathan, W. Anderson, E. Gordon, Metabolic features of chronic fatigue syndrome, Proc. Natl. Acad. Sci. U. S. A. 113 (37) (2016) E5472–E5480.
[2] S. Yachida, S. Mizutani, H. Shiroma, S. Shiba, T. Nakajima, T. Sakamoto, H. Watanabe, K. Masuda, Y. Nishimoto, M. Kubo, F. Hosoda, H. Rokutan, M. Matsumoto, H. Takamaru, M. Yamada, T. Matsuda, M. Iwasaki, T. Yamaji, T. Yachida, T. Soga, K. Kurokawa, A. Toyoda, Y. Ogura, T. Hayashi, M. Hatakeyama, H. Nakagama, Y. Saito, S. Fukuda, T. Shibata, T. Yamada, Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer, Nat. Med. 25 (6) (2019) 968–976.
[3] E.M. Llufrio, K. Cho, G.J. Patti, Systems-level analysis of isotopic labeling in untargeted metabolomic data by X(13)CMS, Nat. Protoc. 14 (7) (2019) 1970–1990.
[4] C.M. McGrath, S.P. Young, Can metabolomic profiling predict response to therapy? Nat. Rev. Rheumatol. 15 (3) (2019) 129–130.
[5] Y. Zhao, Z. Hao, C. Zhao, J. Zhao, J. Zhang, Y. Li, L. Li, X. Huang, X. Lin, Z. Zeng, X. Lu, G. Xu, A novel strategy for large-scale metabolomics study by calibrating gross and systematic errors in gas chromatography-mass spectrometry, Anal. Chem. 88 (4) (2016) 2234–2242.
[6] X. Cui, Q. Yang, B. Li, J. Tang, X. Zhang, S. Li, F. Li, J. Hu, Y. Lou, Y. Qiu, W. Xue, F. Zhu, Assessing the effectiveness of direct data merging strategy in long-term and large-scale Pharmacometabonomics, Front. Pharmacol. 10 (2019) 127.
[7] C.I. Li, D.C. Samuels, Y.Y. Zhao, Y. Shyr, Y. Guo, Power and sample size calculations for high-throughput sequencing-based experiments, Brief. Bioinform. 19 (6) (2018) 1247–1255.
[8] Q. Yang, Y. Wang, Y. Zhang, F. Li, W. Xia, Y. Zhou, Y. Qiu, H. Li, F. Zhu, NOREVA: enhanced normalization and evaluation of time-course and multi-class metabolomic data, Nucleic Acids Res. 48 (W1) (2020) W436–W448.
[9] J.R. Mayers, C. Wu, C.B. Clish, P. Kraft, M.E. Torrence, B.P. Fiske, C. Yuan, Y. Bao, M.K. Townsend, S.S. Tworoger, S.M. Davidson, T. Papagiannakopoulos, A. Yang, T. L. Dayton, S. Ogino, M.J. Stampfer, E.L. Giovannucci, Z.R. Qian, D.A. Rubinson, J. Ma, H.D. Sesso, J.M. Gaziano, B.B. Cochrane, S. Liu, J. Wactawski-Wende, J. E. Manson, M.N. Pollak, A.C. Kimmelman, A. Souza, K. Pierce, T.J. Wang, R. E. Gerszten, C.S. Fuchs, M.G. Vander Heiden, B.M. Wolpin, Elevation of circulating branched-chain amino acids is an early event in human pancreatic adenocarcinoma development, Nat. Med. 20 (10) (2014) 1193–1198.
[10] D.S. Wishart, Emerging applications of metabolomics in drug discovery and precision medicine, Nat. Rev. Drug Discov. 15 (7) (2016) 473–484.
[11] J. Kinkorova, Biobanks in the era of personalized medicine: objectives, challenges, and innovation: overview, EPMA J. 7 (2015) 4.
[12] E. Deutsch, C. Chargari, L. Galluzzi, G. Kroemer, Optimising efficacy and reducing toxicity of anticancer radioimmunotherapy, Lancet Oncol. 20 (8) (2019) e452–e463.
[13] K. Shameer, M.A. Badgeley, R. Miotto, B.S. Glicksberg, J.W. Morgan, J.T. Dudley, Translational bioinformatics in the era of real-time biomedical, health care and wellness data streams, Brief. Bioinform. 18 (1) (2017) 105–124.
[14] F. Li, Y. Zhou, X. Zhang, J. Tang, Q. Yang, Y. Zhang, Y. Luo, J. Hu, W. Xue, Y. Qiu, Q. He, B. Yang, F. Zhu, SSizer: determining the sample sufficiency for comparative biological study, J. Mol. Biol. 432 (11) (2020) 3411–3421.
[15] R.R. da Silva, P.C. Dorrestein, R.A. Quinn, Illuminating the dark matter in metabolomics, Proc. Natl. Acad. Sci. U. S. A. 112 (41) (2015) 12549–12550.
[16] P. Luo, P. Yin, W. Zhang, L. Zhou, X. Lu, X. Lin, G. Xu, Optimization of large-scale pseudotargeted metabolomics method based on liquid chromatography-mass spectrometry, J. Chromatogr. A 1437 (2016) 127–136.
[17] A. Cambiaghi, M. Ferrario, M. Masseroli, Analysis of metabolomic data: tools, current strategies and future challenges for omics data integration, Brief. Bioinform. 18 (3) (2017) 498–510.
[18] X. Domingo-Almenara, J.R. Montenegro-Burke, H.P. Benton, G. Siuzdak, Annotation: a computational solution for streamlining metabolomics analysis, Anal. Chem. 90 (1) (2018) 480–489.
[19] Q. Yan, X. Zhu, L. Jiang, M. Ye, L. Sun, J.S. Terblanche, R. Wu, A computing platform to map ecological metabolism by integrating functional mapping and the metabolic theory of ecology, Brief. Bioinform. 18 (1) (2017) 137–144.
[20] W. Ma, L. Zhang, P. Zeng, C. Huang, J. Li, B. Geng, J. Yang, W. Kong, X. Zhou, Q. Cui, An analysis of human microbe-disease associations, Brief. Bioinform. 18 (1) (2017) 85–97.
[21] M.A. Kamleh, T.M. Ebbels, K. Spagou, P. Masson, E.J. Want, Optimizing the use of quality control samples for signal drift correction in large-scale urine metabolic profiling studies, Anal. Chem. 84 (6) (2012) 2670–2677.
[22] A.D. Dane, M.M. Hendriks, T.H. Reijmers, A.C. Harms, J. Troost, R.J. Vreeken, D. I. Boomsma, C.M. van Duijn, E.P. Slagboom, T. Hankemeier, Integrating metabolomics profiling measurements across multiple biobanks, Anal. Chem. 86 (9) (2014) 4110–4114.
[23] B. Li, J. Tang, Q. Yang, S. Li, X. Cui, Y. Li, Y. Chen, W. Xue, X. Li, F. Zhu, NOREVA: normalization and evaluation of MS-based metabolomics data, Nucleic Acids Res. 45 (W1) (2017) W162–W170.
[24] A.D. Southam, R.J. Weber, J. Engel, M.R. Jones, M.R. Viant, A complete workflow for high-resolution spectral-stitching nanoelectrospray direct-infusion mass-spectrometry-based metabolomics and lipidomics, Nat. Protoc. 12 (2) (2016) 310–328.
[25] J. Chong, O. Soufan, C. Li, I. Caraus, S. Li, G. Bourque, D.S. Wishart, J. Xia, MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis, Nucleic Acids Res. 46 (W1) (2018) W486–W494.
[26] H. Gowda, J. Ivanisevic, C.H. Johnson, M.E. Kurczy, H.P. Benton, D. Rinehart, T. Nguyen, J. Ray, J. Kuehl, B. Arevalo, P.D. Westenskow, J. Wang, A.P. Arkin, A. M. Deutschbauer, G.J. Patti, G. Siuzdak, Interactive XCMS online: simplifying advanced metabolomic data processing and subsequent statistical analyses, Anal. Chem. 86 (14) (2014) 6931–6939.
[27] G. Kastenmuller, W. Romisch-Margl, B. Wagele, E. Altmaier, K. Suhre, metaP-server: a web-based metabolomics data analysis tool, J. Biomed. Biotechnol. 2011 (2011) 839862.
[28] B. Wen, Z. Mei, C. Zeng, S. Liu, metaX: a flexible and comprehensive software for processing metabolomics data, BMC Bioinformatics 18 (1) (2017) 183.
[29] Y. Guitton, M. Tremblay-Franco, G. Le Corguille, J.F. Martin, M. Petera, P. Roger-Mele, A. Delabriere, S. Goulitquer, M. Monsoor, C. Duperier, C. Canlet, R. Servien, P. Tardivel, C. Caron, F. Giacomoni, E.A. Thevenot, Create, run, share, publish, and reference your LC-MS, FIA-MS, GC-MS, and NMR data analysis workflows with the Workflow4Metabolomics 3.0 Galaxy online infrastructure for metabolomics, Int. J. Biochem. Cell Biol. 93 (2017) 89–101.
[30] A. Biswas, K.C. Mynampati, S. Umashankar, S. Reuben, G. Parab, R. Rao, V. S. Kannan, S. Swarup, MetDAT: a modular and workflow-based free online pipeline for mass spectrometry data processing, analysis and interpretation, Bioinformatics 26 (20) (2010) 2639–2640.
[31] N. Kessler, H. Neuweger, A. Bonte, G. Langenkamper, K. Niehaus, T. W. Nattkemper, A. Goesmann, MeltDB 2.0-advances of the metabolomics software system, Bioinformatics 29 (19) (2013) 2452–2459.
[32] M. Sud, E. Fahy, D. Cotter, K. Azam, I. Vadivelu, C. Burant, A. Edison, O. Fiehn, R. Higashi, K.S. Nair, S. Sumner, S. Subramaniam, Metabolomics workbench: an international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools, Nucleic Acids Res. 44 (D1) (2016) D463–D470.
[33] J. Taminau, C. Lazar, S. Meganck, A. Nowe, Comparison of merging and meta-analysis as alternative approaches for integrative gene expression analysis, ISRN Bioinform. 2014 (2014) 345106.
[34] E. Zelena, W.B. Dunn, D. Broadhurst, S. Francis-McIntyre, K.M. Carroll, P. Begley, S. O'Hagan, J.D. Knowles, A. Halsall, H. Consortium, I.D. Wilson, D.B. Kell, Development of a robust and repeatable UPLC-MS method for the long-term metabolomic study of human serum, Anal. Chem. 81 (4) (2009) 1357–1364.
[35] M. Krzywinski, N. Altman, Points of significance power and sample size, Nat. Methods 10 (12) (2013) 1139–1140.
[36] W.B. Dunn, D. Broadhurst, P. Begley, E. Zelena, S. Francis-McIntyre, N. Anderson, M. Brown, J.D. Knowles, A. Halsall, J.N. Haselden, A.W. Nicholls, I.D. Wilson, D. B. Kell, R. Goodacre, C. Human Serum Metabolome, Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry, Nat. Protoc. 6 (7) (2011) 1060–1083.
[37] W. Zhang, Z. Lei, D. Huhman, L.W. Sumner, P.X. Zhao, MET-XAlign: a metabolite cross-alignment tool for LC/MS-based comparative metabolomics, Anal. Chem. 87 (18) (2015) 9114–9119.

[38] G. Tomasi, F. van den Berg, C. Andersson, Correlation optimized warping and dynamic time warping as preprocessing methods for chromatographic data, J. Chemom. 18 (5) (2004) 231–241.
[39] A. Nordstrom, G. O'Maille, C. Qin, G. Siuzdak, Nonlinear data alignment for UPLC-MS and HPLC-MS based metabolomics: quantitative analysis of endogenous and exogenous metabolites in human serum, Anal. Chem. 78 (10) (2006) 3289–3295.
[40] J. Tang, J. Fu, Y. Wang, B. Li, Y. Li, Q. Yang, X. Cui, J. Hong, X. Li, Y. Chen, W. Xue, F. Zhu, ANPELA: analysis and performance assessment of the label-free quantification workflow for metaproteomic studies, Brief. Bioinform. 21 (2) (2019) 621–636.
[41] Q. Yang, J. Hong, Y. Li, W. Xue, S. Li, H. Yang, F. Zhu, A novel bioinformatics approach to identify the consistently well-performing normalization strategy for current metabolomic studies, Brief. Bioinform. (2019) pii: bbz137.
[42] A.M. De Livera, D.A. Dias, D. De Souza, T. Rupasinghe, J. Pyke, D. Tull, U. Roessner, M. McConville, T.P. Speed, Normalizing and integrating metabolomics data, Anal. Chem. 84 (24) (2012) 10768–10776.
[43] Z. Li, Y. Lu, Y. Guo, H. Cao, Q. Wang, W. Shui, Comprehensive evaluation of untargeted metabolomics data processing software in feature detection, quantification and discriminating marker selection, Anal. Chim. Acta 1029 (2018) 50–57.
[44] Q. Yang, Y. Wang, S. Zhang, J. Tang, F. Li, J. Yin, Y. Li, J. Fu, B. Li, Y. Luo, W. Xue, F. Zhu, Biomarker discovery for immunotherapy of pituitary adenomas: enhanced robustness and prediction ability by modern computational tools, Int. J. Mol. Sci. 20 (1) (2019) 151.
[45] Q. Yang, B. Li, J. Tang, X. Cui, Y. Wang, X. Li, J. Hu, Y. Chen, W. Xue, Y. Lou, Y. Qiu, F. Zhu, Consistent gene signature of schizophrenia identified by a novel feature selection strategy from comprehensive sets of transcriptomic data, Brief. Bioinform. 21 (3) (2020) 1058–1068.
[46] J. Tang, M. Mou, Y. Wang, Y. Luo, F. Zhu, MetaFS: Performance assessment of biomarker discovery in metaproteomics, Brief. Bioinform. (2020) pii: bbaa105.
[47] C. Christin, H.C. Hoefsloot, A.K. Smilde, B. Hoekman, F. Suits, R. Bischoff, P. Horvatovich, A critical assessment of feature selection methods for biomarker discovery in clinical proteomics, Mol. Cell. Proteomics 12 (1) (2013) 263–276.
[48] C.Y. Yu, X.X. Li, H. Yang, Y.H. Li, W.W. Xue, Y.Z. Chen, L. Tao, F. Zhu, Assessing the performances of protein function prediction algorithms from the perspectives of identification accuracy and false discovery rate, Int. J. Mol. Sci. 19 (1) (2018) 183.
[49] J. Tang, J. Fu, Y. Wang, Y. Luo, Q. Yang, B. Li, G. Tu, J. Hong, X. Cui, Y. Chen, L. Yao, W. Xue, F. Zhu, Simultaneous improvement in the precision, accuracy and robustness of label-free proteome quantification by optimizing data manipulation chains, Mol. Cell. Proteomics 18 (8) (2019) 1683–1699.
[50] S. Min, B. Lee, S. Yoon, Deep learning in bioinformatics, Brief. Bioinform. 18 (5) (2017) 851–869.
[51] Q.X. Yang, Y.X. Wang, F.C. Li, S. Zhang, Y.C. Luo, Y. Li, J. Tang, B. Li, Y.Z. Chen, W. W. Xue, F. Zhu, Identification of the gene signature reflecting schizophrenia's etiology by constructing artificial intelligence-based method of enhanced reproducibility, CNS Neurosci. Ther. 25 (9) (2019) 1054–1063.
[52] D.Y. Lee, R. Saha, F.N. Yusufi, W. Park, I.A. Karimi, Web-based applications for building, managing and analysing kinetic models of biological systems, Brief. Bioinform. 10 (1) (2009) 65–74.
[53] J. Leipzig, A review of bioinformatic pipeline frameworks, Brief. Bioinform. 18 (3) (2017) 530–536.
[54] J.F. Xiao, R.S. Varghese, B. Zhou, M.R. Nezami Ranjbar, Y. Zhao, T.H. Tsai, C. Di Poto, J. Wang, D. Goerlitz, Y. Luo, A.K. Cheema, N. Sarhan, H. Soliman, M. G. Tadesse, D.H. Ziada, H.W. Ressom, LC-MS based serum metabolomics for identification of hepatocellular carcinoma biomarkers in Egyptian cohort, J. Proteome Res. 11 (12) (2012) 5914–5923.
[55] H.W. Ressom, J.F. Xiao, L. Tuli, R.S. Varghese, B. Zhou, T.H. Tsai, M.R. Ranjbar, Y. Zhao, J. Wang, C. Di Poto, A.K. Cheema, M.G. Tadesse, R. Goldman, K. Shetty, Utilization of metabolomics to identify serum biomarkers for hepatocellular carcinoma in patients with liver cirrhosis, Anal. Chim. Acta 743 (2012) 90–100.
[56] D.S. Wishart, Y.D. Feunang, A. Marcu, A.C. Guo, K. Liang, R. Vazquez-Fresno, T. Sajed, D. Johnson, C. Li, N. Karu, Z. Sayeeda, E. Lo, N. Assempour, M. Berjanskii, S. Singhal, D. Arndt, Y. Liang, H. Badran, J. Grant, A. Serra-Cayuela, Y. Liu, R. Mandal, V. Neveu, A. Pon, C. Knox, M. Wilson, C. Manach, A. Scalbert, HMDB 4.0: the human metabolome database for 2018, Nucleic Acids Res. 46 (D1) (2018) D608–D617.
[57] C. Frainay, F. Jourdan, Computational methods to identify metabolic sub-networks based on metabolomic profiles, Brief. Bioinform. 18 (1) (2017) 43–56.
[58] C. Guijas, J.R. Montenegro-Burke, X. Domingo-Almenara, A. Palermo, B. Warth, G. Hermann, G. Koellensperger, T. Huan, W. Uritboonthai, A.E. Aisporna, D. W. Wolan, M.E. Spilker, H.P. Benton, G. Siuzdak, METLIN: a technology platform for identifying knowns and unknowns, Anal. Chem. 90 (5) (2018) 3156–3164.
[59] Q. Cui, I.A. Lewis, A.D. Hegeman, M.E. Anderson, J. Li, C.F. Schulte, W.M. Westler, H.R. Eghbalnia, M.R. Sussman, J.L. Markley, Metabolite identification via the Madison metabolomics Consortium database, Nat. Biotechnol. 26 (2) (2008) 162–164.
[60] M. Sud, E. Fahy, D. Cotter, A. Brown, E.A. Dennis, C.K. Glass, A.H. Merrill Jr., R. C. Murphy, C.R. Raetz, D.W. Russell, S. Subramaniam, LMSD: LIPID MAPS structure database, Nucleic Acids Res. 35 (D1) (2007) D527–D532.
[61] H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M.Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito, T. Nishioka, MassBank: a public repository for sharing mass spectral data for life sciences, J. Mass Spectrom. 45 (7) (2010) 703–714.
[62] D. Wishart, D. Arndt, A. Pon, T. Sajed, A.C. Guo, Y. Djoumbou, C. Knox, M. Wilson, Y. Liang, J. Grant, Y. Liu, S.A. Goldansaz, S.M. Rappaport, T3DB: the toxic exposome database, Nucleic Acids Res. 43 (D1) (2015) D928–D934.
[63] T. Sajed, A. Marcu, M. Ramirez, A. Pon, A.C. Guo, C. Knox, M. Wilson, J.R. Grant, Y. Djoumbou, D.S. Wishart, ECMDB 2.0: A richer resource for understanding the biochemistry of *E. coli*, Nucleic Acids Res. 44 (D1) (2016) D495–D501.
[64] M. Ramirez-Gaona, A. Marcu, A. Pon, A.C. Guo, T. Sajed, N.A. Wishart, N. Karu, Y. Djoumbou Feunang, D. Arndt, D.S. Wishart, YMDB 2.0: a significantly expanded version of the yeast metabolome database, Nucleic Acids Res. 45 (D1) (2017) D440–D445.
[65] M. Udayakumar, D. Prem Chandar, N. Arun, J. Mathangi, K. Hemavathi, R. Seenivasagam, PMDB: plant metabolome database-a Metabolomic approach, Med. Chem. Res. 21 (1) (2012) 47–52.
[66] D.S. Wishart, Y.D. Feunang, A.C. Guo, E.J. Lo, A. Marcu, J.R. Grant, T. Sajed, D. Johnson, C. Li, Z. Sayeeda, N. Assempour, I. Iynkkaran, Y. Liu, A. Maciejewski, N. Gale, A. Wilson, L. Chin, R. Cummings, D. Le, A. Pon, C. Knox, M. Wilson, DrugBank 5.0: a major update to the DrugBank database for 2018, Nucleic Acids Res. 46 (D1) (2018) D1074–D1082.
[67] C. Zhang, L. Tao, C. Qin, P. Zhang, S. Chen, X. Zeng, F. Xu, Z. Chen, S.Y. Yang, Y. Z. Chen, CFam: a chemical families database based on iterative selection of functional seeds and seed-directed compound clustering, Nucleic Acids Res. 43 (D1) (2015) D558–D565.
[68] L. Huang, D. Xie, Y. Yu, H. Liu, Y. Shi, T. Shi, C. Wen, TCMID 2.0: a comprehensive resource for TCM, Nucleic Acids Res. 46 (D1) (2018) D1117–D1120.
[69] M. Kanehisa, M. Furumichi, M. Tanabe, Y. Sato, K. Morishima, KEGG: new perspectives on genomes, pathways, diseases and drugs, Nucleic Acids Res. 45 (D1) (2017) D353–D361.
[70] T. Jewison, Y. Su, F.M. Disfany, Y. Liang, C. Knox, A. Maciejewski, J. Poelzer, J. Huynh, Y. Zhou, D. Arndt, Y. Djoumbou, Y. Liu, L. Deng, A.C. Guo, B. Han, A. Pon, M. Wilson, S. Rafatnia, P. Liu, D.S. Wishart, SMPDB 2.0: big improvements to the Small Molecule Pathway Database, Nucleic Acids Res. 42 (D1) (2014) D478–D484.
[71] S.C.S. H, R. Abdul Wahab, C. Cuparencu, L.O. Dragsted, L. Brennan, A Metabolomics Approach to the Identification of Urinary Biomarkers of Pea Intake, Nutrients 10 (12) (2018) 1911.
[72] J. Peng, L. Li, Liquid-liquid extraction combined with differential isotope dimethylaminophenacyl labeling for improved metabolomic profiling of organic acids, Anal. Chim. Acta 803 (2013) 97–105.
[73] J. Cai, J. Zhang, Y. Tian, L. Zhang, E. Hatzakis, K.W. Krausz, P.B. Smith, F. J. Gonzalez, A.D. Patterson, Orthogonal comparison of GC-MS and (1)H NMR spectroscopy for short chain fatty acid quantitation, Anal. Chem. 89 (15) (2017) 7900–7906.
[74] T. Liu, J. Diao, S. Di, Z. Zhou, Stereoselective bioaccumulation and metabolite formation of triadimefon in Tubifex tubifex, Environ. Sci. Technol. 48 (12) (2014) 6687–6693.
[75] W. Zhang, C. Zhao, S. Wang, C. Fang, Y. Xu, H. Lu, P. Yang, Coating cells with cationic silica-magnetite nanocomposites for rapid purification of integral plasma membrane proteins, Proteomics 11 (17) (2011) 3482–3490.
[76] D. Grissa, M. Petera, M. Brandolini, A. Napoli, B. Comte, E. Pujos-Guillot, Feature selection methods for early predictive biomarker discovery using untargeted Metabolomic data, Front. Mol. Biosci. 3 (2016) 30.
[77] J. Fu, J. Tang, Y. Wang, X. Cui, Q. Yang, J. Hong, X. Li, S. Li, Y. Chen, W. Xue, F. Zhu, Discovery of the consistently well-performed analysis chain for SWATH-MS based Pharmacoproteomic quantification, Front. Pharmacol. 9 (2018) 681.
[78] E.A. Mathe, A.D. Patterson, M. Haznadar, S.K. Manna, K.W. Krausz, E.D. Bowman, P.G. Shields, J.R. Idle, P.B. Smith, K. Anami, D.G. Kazandjian, E. Hatzakis, F. J. Gonzalez, C.C. Harris, Noninvasive urinary metabolomic profiling identifies diagnostic and prognostic markers in lung cancer, Cancer Res. 74 (12) (2014) 3259–3270.
[79] M. Hilvo, S. Gade, T. Hyotylainen, V. Nekljudova, T. Seppanen-Laakso, M. Sysi-Aho, M. Untch, J. Huober, G. von Minckwitz, C. Denkert, M. Oresic, S. Loibl, Monounsaturated fatty acids in serum triacylglycerols are associated with response to neoadjuvant chemotherapy in breast cancer patients, Int. J. Cancer 134 (7) (2014) 1725–1733.
[80] J. Yin, W. Sun, F. Li, J. Hong, X. Li, Y. Zhou, Y. Lu, M. Liu, X. Zhang, N. Chen, X. Jin, J. Xue, S. Zeng, L. Yu, F. Zhu, VARIDT 1.0: variability of drug transporter database, Nucleic Acids Res. 48 (D1) (2020) D1042–D1050.
[81] Y. Wang, S. Zhang, F. Li, Y. Zhou, Y. Zhang, Z. Wang, R. Zhang, J. Zhu, Y. Ren, Y. Tan, C. Qin, Y. Li, X. Li, Y. Chen, F. Zhu, Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics, Nucleic Acids Res. 48 (D1) (2020) D1031–D1041.
[82] E.A. Thevenot, A. Roux, Y. Xu, E. Ezan, C. Junot, Analysis of the human adult urinary metabolome variations with age, body mass index, and gender by implementing a comprehensive workflow for univariate and OPLS statistical analyses, J. Proteome Res. 14 (8) (2015) 3322–3335.
[83] R.D. Prediger, L.C. Batista, R.N. Takahashi, Caffeine reverses age-related deficits in olfactory discrimination and social recognition memory in rats. Involvement of adenosine A1 and A2A receptors, Neurobiol. Aging 26 (6) (2005) 957–964.

[84] I. Bonnacker, D. Berdel, R. Suverkrup, A. von Berg, Renal clearance of theophylline and its major metabolites: age and urine flow dependency in paediatric patients, Eur. J. Clin. Pharmacol. 36 (2) (1989) 145–150.
[85] D.S. Sardina, S. Alaimo, A. Ferro, A. Pulvirenti, R. Giugno, A novel computational method for inferring competing endogenous interactions, Brief. Bioinform. 18 (6) (2017) 1071–1081.
[86] F.E. Garrett-Bakelman, M. Darshi, S.J. Green, R.C. Gur, L. Lin, B.R. Macias, M. J. McKenna, C. Meydan, T. Mishra, J. Nasrini, B.D. Piening, L.F. Rizzardi, K. Sharma, J.H. Siamwala, L. Taylor, M.H. Vitaterna, M. Afkarian, E. Afshinnekoo, S. Ahadi, A. Ambati, M. Arya, D. Bezdan, C.M. Callahan, S. Chen, A.M.K. Choi, G. E. Chlipala, K. Contrepois, M. Covington, B.E. Crucian, I. De Vivo, D.F. Dinges, D. J. Ebert, J.I. Feinberg, J.A. Gandara, K.A. George, J. Goutsias, G.S. Grills, A. R. Hargens, M. Heer, R.P. Hillary, A.N. Hoofnagle, V.Y.H. Hook, G. Jenkinson, P. Jiang, A. Keshavarzian, S.S. Laurie, B. Lee-McMullen, S.B. Lumpkins, M. MacKay, M.G. Maienschein-Cline, A.M. Melnick, T.M. Moore, K. Nakahira, H. H. Patel, R. Pietrzyk, V. Rao, R. Saito, D.N. Salins, J.M. Schilling, D.D. Sears, C. K. Sheridan, M.B. Stenger, R. Tryggvadottir, A.E. Urban, T. Vaisar, B. Van Espen, J. Zhang, M.G. Ziegler, S.R. Zwart, J.B. Charles, C.E. Kundrot, G.B.I. Scott, S. M. Bailey, M. Basner, A.P. Feinberg, S.M.C. Lee, C.E. Mason, E. Mignot, B.K. Rana, S.M. Smith, M.P. Snyder, F.W. Turek, The NASA Twins Study: A multidimensional analysis of a year-long human spaceflight, Science 364 (6436) (2019) 144.