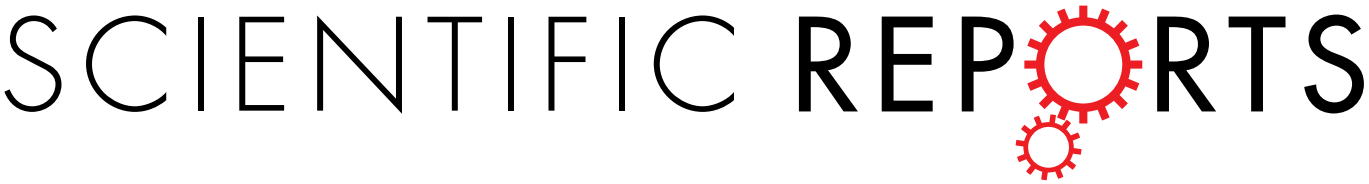

OPEN

Comparative Evaluation of MS-based Metabolomics Software and Its Application to Preclinical Alzheimer's Disease

Received: 1 May 2018
Accepted: 17 May 2018
Published online: 18 June 2018

Ling Hao[1], Jingxin Wang[2], David Page[3], Sanjay Asthana[4], Henrik Zetterberg[5,6,7,8], Cynthia Carlsson[4], Ozioma C. Okonkwo[4] & Lingjun Li[1,9]

Mass spectrometry-based metabolomics has undergone significant progresses in the past decade, with a variety of software packages being developed for data analysis. However, systematic comparison of different metabolomics software tools has rarely been conducted. In this study, several representative software packages were comparatively evaluated throughout the entire pipeline of metabolomics data analysis, including data processing, statistical analysis, feature selection, metabolite identification, pathway analysis, and classification model construction. LC-MS-based metabolomics was applied to preclinical Alzheimer's disease (AD) using a small cohort of human cerebrospinal fluid (CSF) samples (N = 30). All three software packages, XCMS Online, SIEVE, and Compound Discoverer, provided consistent and reproducible data processing results. A hybrid method combining statistical test and support vector machine feature selection was employed to screen key metabolites, achieving a complementary selection of candidate biomarkers from three software packages. Machine learning classification using candidate biomarkers generated highly accurate and predictive models to classify patients into preclinical AD or control category. Overall, our study demonstrated a systematic evaluation of different MS-based metabolomics software packages for the entire data analysis pipeline which was applied to the candidate biomarker discovery of preclinical AD.

Metabolomics is defined as the systematic study of small molecules profile within a biological system. Characterization of the metabolome offers a wealth of information regarding both enzymatic activities and environmental factors[1,2]. Mass spectrometry (MS) has become a foremost technology for the study of metabolites and their dynamic alterations involved in various diseases[3–6]. In particular, hyphenated MS platforms can provide reproducible detection and sensitive measurements for thousands of metabolites in complex samples by online coupling with separation systems prior to MS analysis, including liquid chromatography (LC), gas chromatography (GC), and capillary electrophoresis (CE)[7–12]. Different MS-based analytical platforms provide complementary coverage of the complex metabolome in a given biological sample. LC-MS does not require chemical derivatization and offers large sample loading capacities for small molecules with high degree of structural diversity. GC-MS is highly reproducible and well suited for thermal stable and volatile compounds or compounds that can be derivatized to be volatile. CE-MS has the ability to separate charged or polar molecules based on charge-to-size ratios with high resolution and relatively small amounts of samples.

As hyphenated MS platforms are widely used for metabolomics analysis as well as the application to disease biomarker discovery, significant challenges have arisen with regard to analyzing the large and multidimensional data sets[13,14]. In terms of studies for human diseases, sample size and amount are often very limited with the

[1]School of Pharmacy, University of Wisconsin-Madison, Madison, WI, USA. [2]Baylor College of Medicine, Houston, TX, USA. [3]Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison, Madison, WI, USA. [4]Wisconsin Alzheimer's Disease Research Center, University of Wisconsin-Madison, Madison, WI, USA. [5]Clinical Neurochemistry Laboratory, Sahlgrenska University Hospital Mölndal, Mölndal, Sweden. [6]Institute of Neuroscience and Physiology, Department of Psychiatry and Neurochemistry, the Sahlgrenska Academy at the University of Gothenburg, Mölndal, Sweden. [7]Department of Molecular Neuroscience, UCL Institute of Neurology, London, UK. [8]Dementia Research Institute, London, UK. [9]Department of Chemistry, University of Wisconsin-Madison, Madison, WI, USA. Correspondence and requests for materials should be addressed to L.L. (email: lingjun.li@wisc.edu)

number of compounds far exceeding the number of human subjects. Therefore, sophisticated bioinformatics tools and data-mining technologies are required for automated data analysis and evaluation. With a focus on LC-MS and CE-MS-based metabolomics, a number of freely available and commercial software packages have been developed for data analysis, such as XCMS[15–17], MZmine[18], MetAlign[19], MAVEN[20], SIEVE (Thermo), Progenesis QI (Waters), MetaboScape (Brucker), and most recently, Compound Discoverer (Thermo). Each software package possesses unique advantages in different steps of pre-processing, data analysis, visualization, and interpretation[21,22]. However, systematic comparison of different software packages has rarely been conducted.

In this study, representative software packages (XCMS Online 3.5.1, SIEVE 2.2, and Compound Discoverer 2.0) were selected and comparatively evaluated throughout the entire pipeline of metabolomics data analysis, including data processing, statistical analysis, feature selection, metabolite identification, pathway analysis, and classification model construction. The metabolomics analysis was performed using a small cohort of human cerebrospinal fluid (CSF) samples (N = 30) and further applied to the preclinical Alzheimer's disease (AD).

AD is a neurodegenerative disorder affecting millions of elderly people worldwide. AD patients suffer from progressive syndromes of memory loss, language impairment, and behavioral disturbance. Pathophysiologically, AD is characterized by the presence of extracellular amyloid β (Aβ) plaque and intracellular neurofibrillary aggregates of protein tau in the brain[23]. The concept of AD has evolved to be a continuum over the past decade, with a new disease framework including the preclinical (pre-symptomatic) stage, mild cognitive impairment, and dementia[24]. The pathophysiological process of AD starts years before the emergence of clinical syndrome, yet unequivocal diagnosis and treatment in the early phase of AD is still lacking. It is highly possible that patients could be optimally treated in the preclinical stage of AD before the occurrence of clinical symptoms. CSF circulates within the brain ventricular system, maintains metabolic homeostasis of the brain, and is therefore the most direct and valuable biofluid sample to evaluate the dysfunction of the central nervous system. Discovering the metabolic changes in CSF samples derived from people in the preclinical stage of AD can provide critical insights into disease progression and support the early diagnosis and treatment of AD[25–27].

Results

Data processing evaluations. Efficient and reliable data processing is the first key step toward successful data analysis and biologically important findings. Herein, we assessed the data processing performance of three widely-used software packages, XCMS Online, SIEVE 2.2, and Compound Discoverer 2.0. Firstly, a pooled mixture of aliquots from all CSF metabolite specimens was prepared as a quality control (QC) sample. QC injections (n = 3) were used to evaluate the consistency, reproducibility, and dynamic range of data processing. Ideally, the results of technical replicates should be perfectly repetitive with zero deviation. But inevitable variations can be introduced from both instrument platform and data analysis platform. Since our previous metabolomics study has demonstrated the excellent reproducibility of the current LC-MS/MS platform[4], data analysis consistency of three software packages can be compared using the peak area relative standard deviation (RSD) of all the compounds detected from QC injections. More than 4000 compounds (combining positive and negative electrospray ionization modes, deisotoped and de-adducted) were detected in the QC sample. Histograms in Fig. 1a indicate excellent consistency and reproducibility of all three software packages. Over 60% of the detected compounds obtained peak area RSDs lower than 0.1. XCMS Online software yielded the most reproducibly integrated peak areas with nearly 2000 compounds' RSD lower than 0.05. The consistency of peak extraction and integration is essential as subsequent statistical analyses are conducted based on the values of peak areas. The different variabilities of peak areas among three software packages could be caused by the intrinsic algorithms used for peak detection and integration and could also be influenced by chromatography alignment[21].

One of the key challenges of metabolomics profiling is the high dynamic range of metabolite concentrations present in biological samples. Different classes of metabolites often present distinct LC affinities and ionization efficiencies on LC-ESI-MS platform. The collective effects of large concentration range and structural diversity result in the high peak area dynamic range of LC-MS-based metabolomics datasets, which was found to be 6 orders of magnitude. The histograms of $\log_2$-transformed peak areas from all detected compounds showed approximately normal distribution, where majority of the detected peaks were within the range of $1e^5$ to $5e^6$ peak area (Fig. 1b). However, the mean of the histogram distribution followed the trend of CD < SIEVE < XCMS Online, and CD software also generated more left-skewed distribution, suggesting that CD software extracted the largest number of low-abundance metabolites. Furthermore, only CD software can extract both ESI + and ESI- spectra simultaneously from data files in a single analysis.

Statistical evaluations. In the preclinical AD vs. control dataset, the total numbers of detected compounds (ESI+ and ESI−) were 4389, 4778, and 4004 using CD, SIEVE, and XCMS Online respectively. Both univariate and multivariate statistical analyses were conducted for the comparative assessment of the three software packages. In the univariate statistical test, the discrimination of individual metabolites between preclinical AD and control groups were represented by *t*-test *p*-values and ratios. *P*-values need to be corrected for multiple hypotheses testing to limit the number of false positives and account for the possible interactions among detected compounds. Both CD and XCMS Online software provide adjusted *p*-value for each quantified mass feature. As illustrated in the histograms in Fig. 1c, ratios from all three software packages followed normal distribution, where the majority of the compounds remained unchanged between preclinical AD vs. control groups. It is worth mentioning that XCMS Online offers more choices of statistical tests than CD and SIEVE software for two-group or multiple group comparisons, such as paired t-test, Mann-Whitney test, Wilcoxon signed-rank test, and ANOVA[17].

For multivariate analysis, PCA is conducted first to visualize metabolite fingerprints in a reduced dimensional subspace. None of the software packages yielded complete separation of preclinical AD and control groups in PCA plots (Fig. 2a). PLS-DA analysis was then carried out to sharpen the separation among groups. Complete

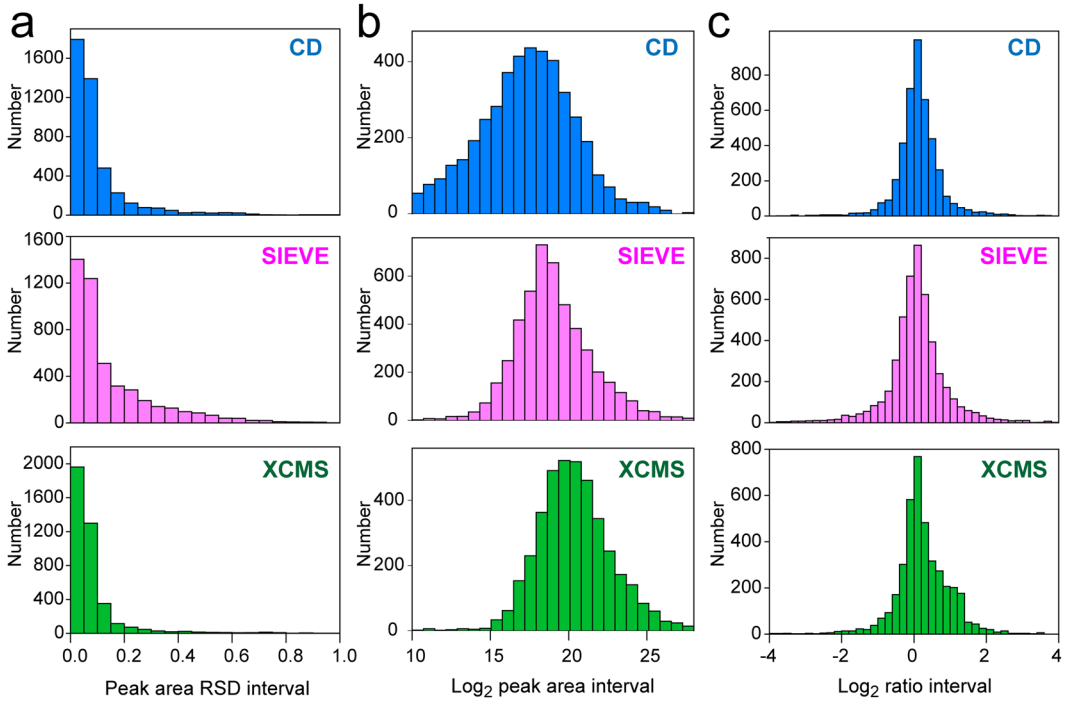

Figure 1. Data processing evaluation of three software packages, Compound Discoverer, SIEVE, and XCMS Online. (**a**) Histograms of peak area relative standard deviations (RSD) of all detected features in QC data set. (**b**) Histograms of $\log_2$-transformed peak areas of all detected features in QC data set. (**c**) Histograms of $\log_2$-transformed ratios of all detected features in preclinical AD vs. control data set.

separation among preclinical AD, control and QC groups were achieved with all three software packages with PLS-DA analysis (Fig. 2b). Compound discoverer software presented the most distinct clustering and separation among different groups. The cross-validated accuracies of the established PLS-DA models were 0.97, 0.96, and 0.93 for CD, SIEVE, and XCMS Online, respectively.

Application to biomarker discovery of preclinical AD. In order to select a panel of most significantly altered features as candidate biomarkers, we have designed a hybrid method combining traditional statistical tests and machine learning feature selection as illustrated in Fig. 3a. The process of biomarker selection and metabolite identification resulted 90, 86, and 81 candidate metabolite biomarkers from CD, SIEVE, and XCMS Online, respectively. As shown in Fig. 3b, the three software packages provided complementary coverage of candidate biomarkers with over 75% shared metabolites between at least two platforms. All 142 metabolites were identified in three software packages, but some of them did not reach statistical significance in one or two software due to the different results of peak integration generated from different software packages. The complete peak list of these 142 metabolites is provided in Supplemental Table S1. The heatmaps of representative metabolite biomarkers (fold change >1.2 among shared metabolites) are displayed in Fig. 4. Each shade of color represents the relative expression level of certain metabolite in an individual human CSF sample. The overall trends of color distributions are similar among the three software packages, indicating consistent and effective analyses of the CSF data set by all three packages.

Candidate metabolite biomarkers generated through three software packages were further evaluated by building binary classification models to differentiate preclinical AD and control patients. Ideally for datasets with a large number of samples, classification models can be constructed on a training set and tested on an independent test set to avoid biased estimates of classifier accuracy. However, for many omics studies, clinical human samples are often very limited, especially for precious samples like human CSF. Leave-one-patient-out cross-validation was therefore used in this study for an estimate of the binary classification model accuracy. The SVM algorithm was selected for machine learning classification, which has gained great success in analyzing gene expression data and handling noisy data in omics studies[4,28–31]. In order to minimize the over-fitting of the classification model, the process of biomarker selection was repeated on every fold of the cross-validation to generate the cross-validated receiver operating characteristic (ROC) curve and its area under the curve (AUC) value[4]. As compared in Table 1, all three software packages achieved highly accurate classification results with sensitivity, specificity, precision and AUC ROC greater than 0.93. If the established model can be validated in the future with a larger sample size (N > 100), it will possess great potential to be used for developing objective diagnostic test of preclinical AD via CSF sampling.

Functional enrichment of the selected 142 metabolites was performed with MBROLE 2.0 tool based on metabolites' categorical annotations in multiple databases[32]. Categorical annotations include metabolic pathways, biofunctions, chemical classifications, and possible association with diseases. HMDB and LIPID MAPS taxonomies

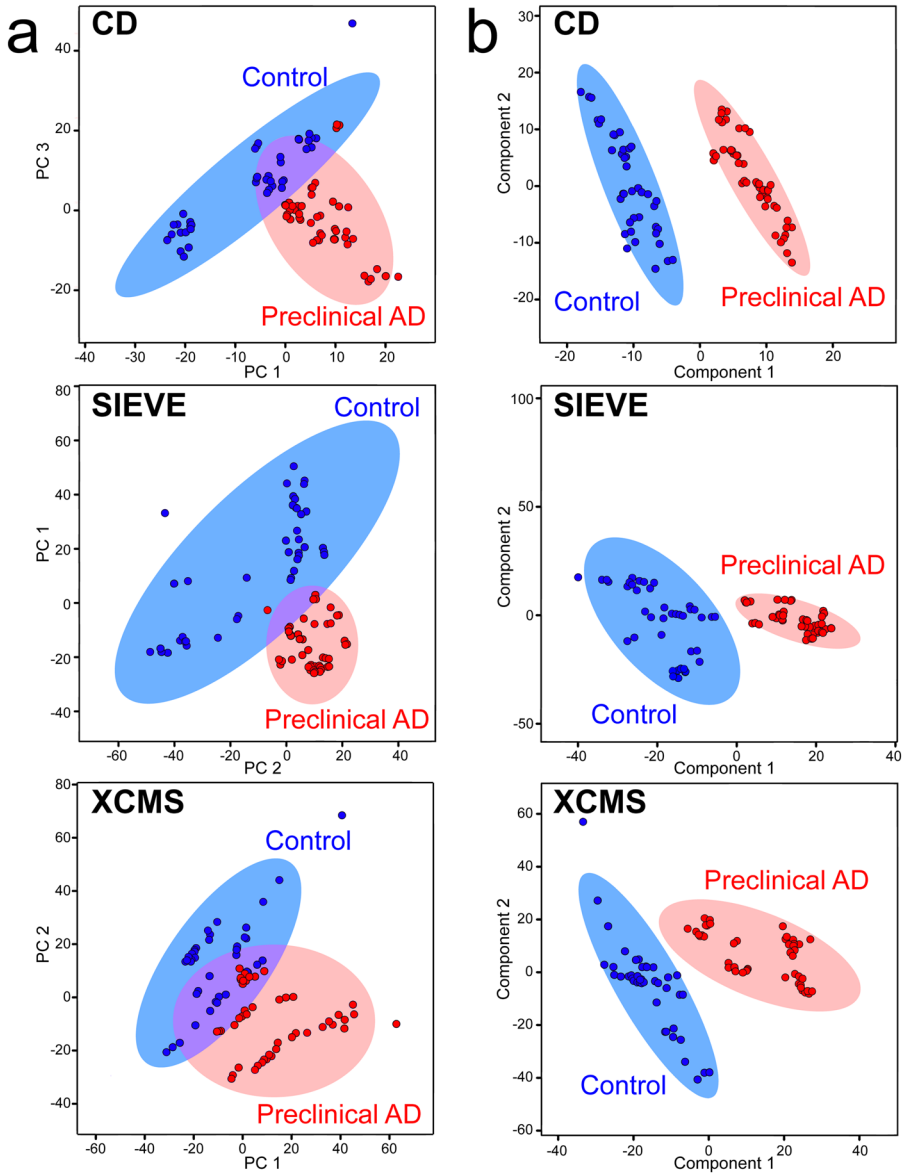

Figure 2. Multivariate statistical analyses of human CSF samples in preclinical AD vs. control vs. QC groups. (**a**) Principal component analysis. (**b**) Partial least squares-discriminant analysis.

indicated that the most enriched chemical groups of candidate biomarkers were carboxylic acids, amino acids, fatty acyls, fatty acids and conjugates, pyrimidines, and nucleosides and analogues. The most enriched cellular locations of candidate biomarkers were cytoplasm, nucleus, mitochondria, endoplasmic reticulum, lysosome, golgi apparatus, peroxisome, and membrane. Six metabolites were directly assigned to AD classification in HMDB based on literature record, including glutamate, tryptophan, glucose, tyrosine, glycerophosphocholine, and homocitric acids. Metabolic pathway analysis was conducted in both MBROLE and Compound Discoverer software through KEGG database. *P*-values and FDR corrected *p*-value of metabolic pathways were calculated by weighing the number of compounds in the set against in the background in MBROLE[32]. Fourteen metabolic pathways were significantly dysregulated in preclinical AD patient vs. control groups (Fig. 5).

Discussion

In this study, different MS-based metabolomics software platforms were compared and evaluated throughout the entire data analysis workflow and disease biomarker discovery. Recognizing a variety of open-source or commercial software available for metabolomics data analysis, we selected three representative and widely-used software packages, XCMS Online 3.5.1, SIEVE™ 2.2, and Compound Discoverer™ 2.0. Each software package possesses unique characteristics for data analysis. For instance, XCMS online offers more options for statistical tests than other software but normalization function is not available during data processing. SIEVE presents separate view of peak alignment (both chromatography and alignment scores) allowing optima peak alignment to be performed before feature detection but has no MS/MS matching function for metabolite identification. Compound

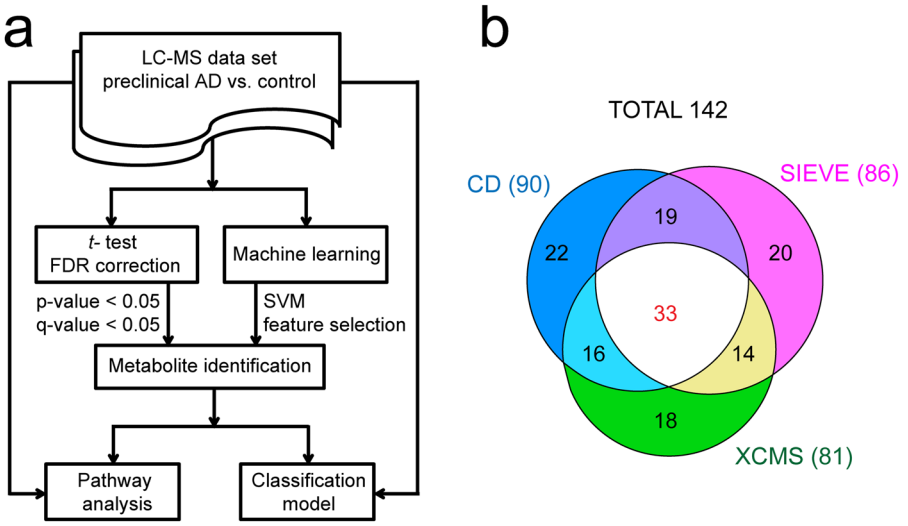

Figure 3. Data analysis flowchart (**a**) and the overlapping candidate biomarkers of preclinical AD resulted from three software packages (**b**).

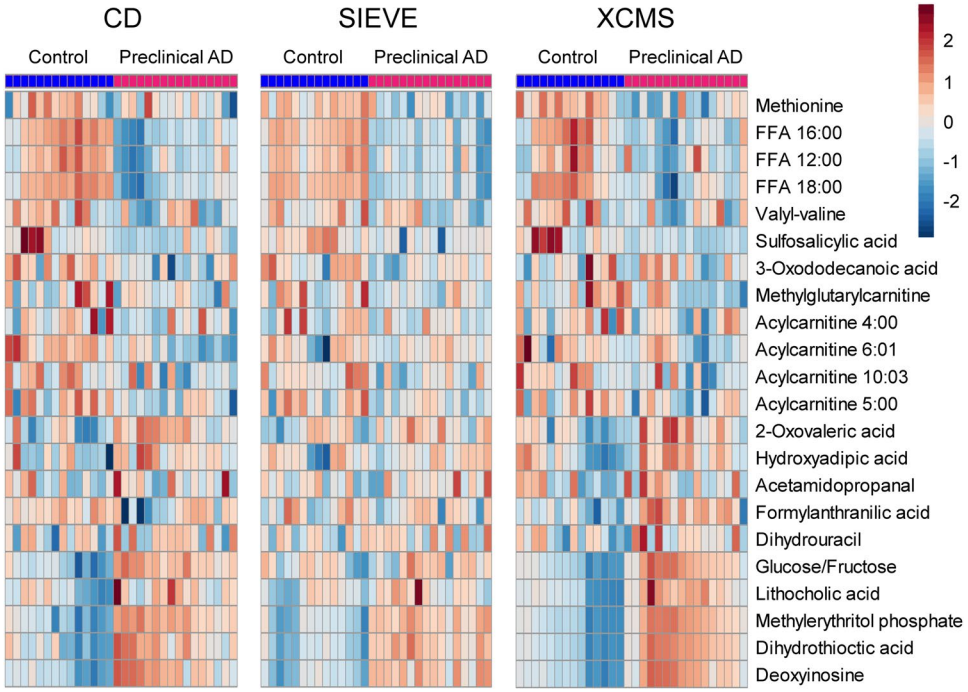

Figure 4. Heatmaps of representative candidate biomarkers among shared metabolites from three software packages. The data was log transformed and auto-scaled.

Software	Feature number	Sensitivity	Specificity	Precision	ROC area
CD	90	0.967	0.962	0.969	0.964
SIEVE	86	0.967	0.962	0.969	0.964
XCMS Online	81	0.933	0.933	0.933	0.933

Table 1. Binary classification performance of candidate biomarkers to differentiate preclinical AD and control groups using three software packages.

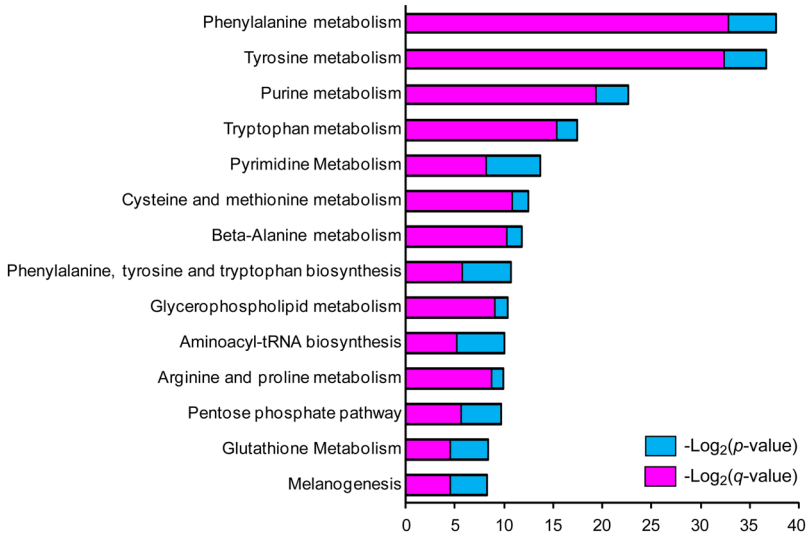

Figure 5. Dysregulated Metabolic pathways in human CSF of preclinical AD vs. control patients. *P*-values and corrected *p*-values of metabolic pathways were calculated by weighing the number of compounds in the set against in the background in MBROLE software.

Discoverer is most recently developed and able to simultaneously process polarity switching data format, yet the comprehensive user-defined workflow requires sufficient knowledge of each node for the best design. In addition, XCMS online requires data upload to the website which could take a couple hours, while SIEVE and CD are installed locally but have system requirements (e.g. Memory, processor) for the local computer to process large scale omics data. XCMS online uses METLIN[33] database for metabolite identification (MS and MS/MS level), SIEVE and CD use ChemSpider for accurate mass matching, and CD can also search MS and MS/MS data against mzCloud database (https://www.mzcloud.org). We recognize that software packages undergo continuous update with new versions and it is also unrealistic to perform data analysis using multiple software platform in the routine metabolomics analysis, our study offers informative reference and starting point for selecting the most appropriate software platforms based on specific needs for metabolomics data analysis.

LC-MS-based human CSF metabolomics analysis was applied to the candidate biomarker discovery of preclinical AD using a small cohort of patient samples (14 Control and 16 Preclinical). three software packages provided complementary selection of candidate biomarkers and were highly predictive to classify patients into diseased or control groups. Ideally, the resulted candidate biomarkers should be independent on the software used for data analysis, but due the intrinsic differences of algorithms in these software packages and the data processing variations, the results (the total of 142 metabolites) shared about 75% candidate biomarkers between two software platforms. As inter-laboratory reproducibility is one of the major challenges in the omics field, data analysis variations due to different software used across labs must be carefully considered. Using multiple metabolomics software packages for data processing and only considering the overlapping features for subsequent analysis could be beneficial to rule out false positive features, and more likely lead to the discovery of more robust biomarkers. This aspect should be taken into consideration for studies aimed at biomarker discovery. Proper S/N cutoff thresholds should also be carefully determined to ensure the best data quality for metabolomics studies.

Many dysregulated metabolites and metabolic pathways have been reported to potentially correlate with neurodegenerative diseases; particularly their functional roles involved in neurotransmission, oxidative stress, and neuroinflammation[34–36]. Tryptophan metabolism pathway was altered in CSF of preclinical AD patients with a total of 29 identified metabolites. The major route of tryptophan metabolism, kynurenine pathway, is a key regulator of both neuroprotective and neurotoxic compounds and has found to be disturbed in neurological diseases including AD[37]. Tryptophan, kynurenine, and kynurenic acid were all decreased in CSF of preclinical AD patients, which was in agreement with another study suggested the correlation between cognitive function and kynurenine pathway metabolites[26,38–40]. Tryptophan also plays fundamental roles in the synthesis of neurotransmitters like serotonin, melatonin, and tryptamine. Ample evidences indicated that cognitive and memory impairment in early stage of AD patients begins with the inefficiency of hippocampal synaptic functions involved in neurotransmitter systems[41]. Both serotonin and its major metabolite 5-hydroxyindoleacetate were significantly decreased in CSF of preclinical AD patients. In consistent with the dysfunction of glutamatergic synapse reported in early stage of AD[41], both glutamate and glutamine were significantly changed and selected as candidate biomarkers in CSF of preclinical AD patients. Moreover, aminoacyl-tRNAs are important translation substrates for protein synthesis, and aminoacyl-tRNA synthetases (AARS) participate in the biosynthesis of signaling molecules, dinucleotide polyphosphates, which can stimulate GABA release in CNS systems[42]. A member of AARS, TrypRS, has also been reported to be a potential marker for AD pathology in a transgenic mouse model of the disease[43]. The dysregulation of AARS pathway agrees well with another plasma and CSF metabolomics study of AD[26]. Higher level of homocysteine, caused by the deficiencies in homocysteine re-methylation cofactors vitamin B12 and B9 (folate), has found to be directly associated with decreased cognitive performance in the elderly[44], which is consistent with

our results. In addition, glutathione metabolism was found to be disturbed in preclinical AD which contributes to oxidative stress related to AD[45]. The increased oxidative stress during AD progression may also contribute to neuroinflammation, mitochondrial dysfunction, and synaptic dysfunction[34]. Glycerophospholid-derived lipid mediators play important roles in these bioprocesses and were found to be disturbed in the present study[36,46]. Although exact mechanisms underlying these disturbed metabolic pathways in AD are still unclear, the identified key metabolites and pathways provide key molecular targets for future mechanistic studies.

Methods

Participants. Thirty enrollees in the Wisconsin Alzheimer's Disease Research Center (WADRC) participated in this study. They comprised sixteen Stage 3 preclinical AD individuals and fourteen age-matched control individuals who were not in the AD pathway. Control subjects were cognitively normal individuals determined by comprehensive neuropsychological assessments and enrolled in the WADRC without family history of AD. Detailed inclusion/exclusion criteria for patient recruitment are described elsewhere[24,47]. The average age of the total sample was 61 ± 6 years. The University of Wisconsin Institutional Review Board (IRB) approved all study procedures, and each enrollee provided a signed informed consent form before participation. All methods were performed in accordance with the IRB guideline and regulation. The clinical information of recruited human subjects is provided in Supplemental Table S2.

CSF sample collection and preparation. Human CSF samples were collected by lumbar puncture at L3/4 or L4/5 following local anesthesia in the morning after 12-hour fast. Each CSF sample was collected via a syringe into a sample collection tube, gently mixed to avoid gradient effect, and centrifuged at 2000 g for 10 min. The supernatant was collected, sub-aliquoted, and stored at −80 °C until use. One milliliter of each CSF sample was thawed on ice and metabolite extraction was achieved by using 3 kDa molecular weight cut-off (MWCO) ultracentrifugation (Millipore Amicon Ultra, MA). The flow-through fraction was collected as CSF metabolite fraction.

LC-MS/MS analysis. Human CSF metabolite samples were analyzed using a Dionex UltiMate 3000 LC system coupled with a Q-Exactive™ Orbitrap mass spectrometer (San Jose, CA), operated on both positive and negative ESI mode. Mobile phase A was 0.1% formic acid in water and mobile phase B was 0.1% formic acid in methanol. Metabolites were separated on a biphenyl column (Phenomenex, 75.1 μm × 150 mm, 1.7 μm, 100 Å) at an LC flow rate of 0.3 ml/min and a column temperature of 35 °C. The 15 min gradient for positive ESI mode was set as follows: 0–5 min, 0–2% solvent B; 5–10 min, 2–50% solvent B; 10–11 min, 50–90% solvent B; 11–13 min, 90% solvent B; 13–15 min, 0% solvent B. The 11 min gradient for negative ESI mode was set as follows: 0–4 min, 0–2% solvent B; 4–7 min, 2–90% solvent B; 7–9 min, 90% solvent B; 9–11 min, 0% solvent B. The injection volume was 5 μL, and each sample was injected in triplicates. Injection order was randomized, and the group information was blinded for LC-MS analysis. Full MS scans were acquired from *m/z* 70–1000 at a resolution of 70 K. Automatic gain control (AGC) target was 1×10^6 and maximum injection time (max IT) was 100 ms. The targeted LC-MS/MS experiments were conducted with an inclusion list of accurate masses and retention times for the purpose of metabolite identification. Resolution was 17.5 K, AGC target was 5×10^5, max IT was 50 ms, isolation window was 1 *m/z*, and normalized collision energy was 30% with higher-energy collisional dissociation fragmentation. LC-MS instrument was controlled by Thermo Scientific Xcalibur 2.2 software.

Data processing by three software packages. Raw data files were independently processed by SIEVE™ 2.2, XCMS Online 3.5.1, and Compound Discoverer™ 2.0 software for metabolomics data analysis. A blank sample was used for background subtraction and noise removal during the pre-processing step.

Commercial SIEVE software was developed by Thermo Scientific as a differential analysis software for both label-free metabolomics and proteomics data analyses. Raw metabolomics data files of preclinical AD, control, and QC groups were processed by SIEVE™ 2.2 with peak alignment and framing algorithm and the experimental type of Control Compare Trend. A QC data file was used as the reference file for peak alignment. The frame time width was 2 min and *m/z* width was 5 ppm. The intensity threshold for component extraction was $1e^6$, and signal-to-noise ratio was 3. ICIS algorithm was used for peak detection. The maximum retention time shift for peak alignment was 0.2 min. Total ion current (TIC) normalization embedded in SIEVE was conducted to reduce instrumental variation before statistical analysis.

XCMS Online is a freely available platform, developed by the Scripps Center for untargeted metabolomics data analysis (http://xcmsonline.scripps.edu). Raw datasets were converted to.mzXML format using the Proteowizard MSConvert tool, and uploaded to XCMS Online for data processing. (Note that the current version of XCMS Online also accept raw data files from several vendors) Multigroup analysis job was created for preclinical AD, control, and QC groups. The predefined parameter set for orbitrap instrument was used and adjusted to be similar as the SIEVE parameters listed above, if possible. CentWave algorithm was used for feature detection and orbitrap settings for retention-time correction. Since normalization function is not available in XCMS Online, normalization to the sum is performed in an Excel file after data processing.

Compound Discoverer (CD) is a relatively new software released by Thermo Scientific this year for targeted and untargeted metabolomics data analysis. Custom designed workflow was established for spectra alignment, compound detection, grouping, metabolite identification, and pathway analysis. The detailed workflow is provided in Supplemental Figure S1. Chromatographic alignment, compound detection, and accurate mass identification parameters were set as the same with SIEVE. Compound peak areas were normalized to the constant sum using embedded function before statistical analysis.

Statistical analysis. Data processing of each software package yielded a multi-dimensional peak table including accurate *m/z*, molecular weight, retention time, compound formula, peak area and other statistical

information. Positive and negative ESI data were separately processed and peak tables were combined into an Excel file. Features with >50% missing values were removed. Compound peak areas were $\log_2$-transformed for subsequent statistical analyses.

Univariate Student's *t*-test was then performed to generate a *p*-value for each detected compound, together with an average fold change, between preclinical AD and control groups, together with an average fold change. The *p*-values were ranked to calculate FDR corrected *p*-value by the Benjamini-Hochberg procedure as implemented in R package. The threshold of statistical significance was set at both *p*-value and corrected *p*-value lower than 0.05. Multivariate statistical analyses, principal component analysis (PCA) and partial least squares-discriminant analysis (PLS-DA), were also carried out using the web-based software MetaboAnalyst 3.0 after log transformation and auto-scaling[48].

Metabolite identification and pathway analysis. Metabolite identification was achieved by accurate mass matching, MS/MS matching, and standard confirmation, following the designed flowchart as described previously[4]. All three software packages provide accurate mass matching functions for tentative metabolite identification, where *m/z* tolerance was set at 10 ppm. Kyoto Encyclopedia of Genes and Genomes (KEGG)[49], Human Metabolome Database (HMDB)[50], Madison Metabolomics Consortium Database (MMCD)[51], and LIPID MAPS[52] were selected for database searching through ChemSpider for both CD and SIEVE software. Metabolite identification for XCMS Online was achieved by searching against METLIN[33] metabolite database. An in-house lipid library with more than 2000 entries was used to assist with lipid identification with an accurate mass tolerance of 10 ppm. For MS/MS matching, XCMS Online and CD software have the capability to search MS/MS data against METLIN and mzCloud database, respectively. Additionally, LC-MS/MS spectra were searched against a web-based MetFrag software to provide complementary coverage[53]. Metabolite IDs were also confirmed with an in-house metabolite library and available metabolite standards by comparing their LC retention time, precursor m/z and MS/MS spectra. Metabolic pathway analysis function is available in the CD software where metabolite IDs were mapped into the KEGG pathways. The current version of XCMS Online also support pathway analysis and visualization through the new Pathway Cloud Plot. The pathway function in SIEVE can be optionally licensed but was not purchased in the present study.

Machine learning feature selection and classification. Machine learning feature selection was achieved by WEKA 3.8 software[54], specifically employing the support vector machine algorithm to evaluate features' contributions to the task of separating preclinical AD vs. control groups. Chromatographic peak areas of detected compounds were averaged across technical replicates and input into WEKA. We ran the SVM with a linear kernel, which maps the original feature space into a high dimensional feature space capturing feature interactions in additional to the original features. We employed our previously described method[4] combining SVM feature selection and traditional statistical test for feature selection. Briefly, all input features were ranked based on their contributions to separate two groups and the top 200 ranked features were then overlapped with statistically significant features (both *p*-value and corrected *p*-value lower than 0.05) for subsequent metabolite identification. The process of biomarker selection was conducted independently for the data set generated from each software package.

For machine learning classification, the peak table containing peak areas of identified candidate biomarkers from each software package was directed into WEKA software. Binary classification model was constructed by leave-one-patient-out cross-validation via linear SVM algorithm. Thirty folds of cross-validation were carried out, where each fold involves withholding one patient as a test set and the rest of 29 patients as a training set. The sensitivity, specificity, and precision of the predictive model were evaluated for classifying patients into preclinical AD or control group.

Supporting data is available. Supplementary data includes: the complete list of 142 identified dysregulated metabolites of preclinical AD from three software packages, clinical information of recruited human subjects, and custom workflow in Compound Discoverer software. Raw LC-MS/MS data are available upon request.

References

1. Patti, G. J., Yanes, O. & Siuzdak, G. Metabolomics: the apogee of the omics trilogy. *Nature Reviews Molecular Cell Biology* **13**, 263–269 (2012).
2. Nicholson, J. K. & Lindon, J. C. Systems biology - Metabonomics. *Nature* **455**, 1054–1056 (2008).
3. Weiss, R. H. & Kim, K. Metabolomics in the study of kidney diseases. *Nat Rev Nephrol* **8**, 22–33 (2011).
4. Hao, L. et al. In-Depth Characterization and Validation of Human Urine Metabolomes Reveal Novel Metabolic Signatures of Lower Urinary Tract Symptoms. *Sci Rep-Uk* **6** (2016).
5. Zang, X., Monge, M. E., McCarty, N. A., Stecenko, A. A. & Fernandez, F. M. Feasibility of Early Detection of Cystic Fibrosis Acute Pulmonary Exacerbations by Exhaled Breath Condensate Metabolomics: A Pilot Study. *J Proteome Res* **16**, 550–558 (2017).
6. Liu, F. *et al*. PKM2 methylation by CARM1 activates aerobic glycolysis to promote tumorigenesis. *Nat Cell Biol* **19**, 1358–1370 (2017).
7. Hao, L., Zhong, X. F., Greer, T., Ye, H. & Li, L. J. Relative quantification of amine-containing metabolites using isobaric N,N-dimethyl leucine (DiLeu) reagents via LC-ESI-MS/MS and CE-ESI-MS/MS. *Analyst* **140**, 467–475 (2015).
8. Zhong, X. F. *et al*. Quantitative analysis of serotonin secreted by human embryonic stem cells-derived serotonergic neurons via pH-mediated online stacking-CE-ESI-MRM. *Electrophoresis* **37**, 1027–1030 (2016).
9. Hao, L. *et al*. Mass Defect-Based N,N-Dimethyl Leucine Labels for Quantitative Proteomics and Amine Metabolomics of Pancreatic Cancer Cells. *Anal Chem* **89**, 1138–1146 (2017).
10. Edwards, J. L., Chisolm, C. N., Shackman, J. G. & Kennedy, R. T. Negative mode sheathless capillary electrophoresis electrospray ionization-mass spectrometry for metabolite analysis of prokaryotes. *Journal Of Chromatography A* **1106**, 80–88 (2006).
11. Hao, L., Li, H. & Lin, J. M. Fractional factorial design based microwave-assisted extraction for the determination of organophosphorus and organochlorine residues in tobacco by using gas chromatography-mass spectrometry. *J Sep Sci* **40**, 542–549 (2017).

12. Jiang, S., Liang, Z. D., Hao, L. & Li, L. J. Investigation of signaling molecules and metabolites found in crustacean hemolymph via *in vivo* microdialysis using a multifaceted mass spectrometric platform. *Electrophoresis* **37**, 1031–1038 (2016).
13. Smith, R. D. Mass Spectrometry in Biomarker Applications: From Untargeted Discovery to Targeted Verification, and Implications for Platform Convergence and Clinical Application. *Clinical Chemistry* **58**, 528–530 (2012).
14. Hawkridge, A. M. & Muddiman, D. C. Mass Spectrometry-Based BiomarkerDiscovery: Toward a Global Proteome Index of Individuality. *Annual Review Of Analytical Chemistry* **2**, 265–277 (2009).
15. Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R. & Siuzdak, G. XCMS: Processing mass spectrometry data for metabolite profiling using Nonlinear peak alignment, matching, and identification. *Anal Chem* **78**, 779–787 (2006).
16. Tautenhahn, R., Patti, G. J., Rinehart, D. & Siuzdak, G. XCMS Online: A Web-Based Platform to Process Untargeted Metabolomic Data. *Anal Chem* **84**, 5035–5039 (2012).
17. Gowda, H. *et al*. Interactive XCMS Online: Simplifying Advanced Metabolomic Data Processing and Subsequent Statistical Analyses. *Anal Chem* **86**, 6931–6939 (2014).
18. Pluskal, T., Castillo, S., Villar-Briones, A. & Oresic, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *Bmc Bioinformatics* **11** (2010).
19. Lommen, A. MetAlign: Interface-Driven, Versatile Metabolomics Tool for Hyphenated Full-Scan Mass Spectrometry Data Preprocessing. *Anal Chem* **81**, 3079–3086 (2009).
20. Melamud, E., Vastag, L. & Rabinowitz, J. D. Metabolomic Analysis and Visualization Engine for LC-MS Data. *Anal Chem* **82**, 9818–9826 (2010).
21. Castillo, S., Gopalacharyulu, P., Yetukuri, L. & Oresic, M. Algorithms and tools for the preprocessing of LC-MS metabolomics data. *Chemometr Intell Lab* **108**, 23–32 (2011).
22. Sugimoto, M., Kawakami, M., Robert, M., Soga, T. & Tomita, M. Bioinformatics Tools for Mass Spectroscopy-Based Metabolomic Data Processing and Analysis. *Curr Bioinform* **7**, 96–108 (2012).
23. Mattson, M. P. Pathways towards and away from Alzheimer's disease. *Nature* **430**, 631–639 (2004).
24. Sperling, R. A. *et al*. Toward defining the preclinical stages of Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. *Alzheimers Dement* **7**, 280–292 (2011).
25. Wishart, D. S. *et al*. The human cerebrospinal fluid metabolome. *J Chromatogr B* **871**, 164–173 (2008).
26. Trushina, E., Dutta, T., Persson, X. M. T., Mielke, M. M. & Petersen, R. C. Identification of Altered Metabolic Pathways in Plasma and CSF in Mild Cognitive Impairment and Alzheimer's Disease Using Metabolomics. *Plos One* **8** (2013).
27. Guo, K., Bamforth, F. & Li, L. Qualitative metabolome analysis of human cerebrospinal fluid by 13C-/12C-isotope dansylation labeling combined with liquid chromatography Fourier transform ion cyclotron resonance mass spectrometry. *J Am Soc Mass Spectrom* **22**, 339–347 (2011).
28. Mahadevan, S., Shah, S. L., Marrie, T. J. & Slupsky, C. M. Analysis of metabolomic data using support vector machines. *Anal Chem* **80**, 7562–7570 (2008).
29. Nanni, L., Lumini, A. & Brahnam, S. Advanced machine learning techniques for microarray spot quality classification. *Neural Computing & Applications* **19**, 471–475 (2010).
30. Schilling, J. *et al*. Compartment Proteomics Analysis of White Perch (Morone americana) Ovary Using Support Vector Machines. *Journal Of Proteome Research* **13**, 1515–1526 (2014).
31. Gaul, D. A. *et al*. Highly-accurate metabolomic detection of early-stage ovarian cancer. *Sci Rep-Uk* **5** (2015).
32. Lopez-Ibanez, J., Pazos, F. & Chagoyen, M. MBROLE 2.0-functional enrichment of chemical compounds. *Nucleic Acids Res* **44**, W201–W204 (2016).
33. Smith, C. A. *et al*. METLIN - A metabolite mass spectral database. *Therapeutic Drug Monitoring* **27**, 747–751 (2005).
34. Gonzalez-Dominguez, R., Garcia-Barrera, T. & Gomez-Ariza, J. L. Using direct infusion mass spectrometry for serum metabolomics in Alzheimer's disease. *Anal Bioanal Chem* **406**, 7137–7148 (2014).
35. Wang, G. *et al*. Plasma metabolite profiles of Alzheimer's disease and mild cognitive impairment. *J Proteome Res* **13**, 2649–2658 (2014).
36. Gonzalez-Dominguez, R., Garcia, A., Garcia-Barrera, T., Barbas, C. & Gomez-Ariza, J. L. Metabolomic profiling of serum in the progression of Alzheimer's disease by capillary electrophoresis-mass spectrometry. *Electrophoresis* **35**, 3321–3330 (2014).
37. Lovelace, M. D. *et al*. Recent evidence for an expanded role of the kynurenine pathway of tryptophan metabolism in neurological diseases. *Neuropharmacology* **112**, 373–388 (2017).
38. Gulaj, E., Pawlak, K., Bien, B. & Pawlak, D. Kynurenine and its metabolites in Alzheimer's disease patients. *Adv Med Sci-Poland* **55**, 204–211 (2010).
39. Ibanez, C. *et al*. A new metabolomic workflow for early detection of Alzheimer's disease. *Journal Of Chromatography A* **1302**, 65–71 (2013).
40. Kaddurah-Daouk, R. *et al*. Metabolomic changes in autopsy-confirmed Alzheimer's disease. *Alzheimers Dement* **7**, 309–317 (2011).
41. Selkoe, D. J. Alzheimer's disease is a synaptic failure. *Science* **298**, 789–791 (2002).
42. Miras-Portugal, M. T., Gualix, J. & Pintor, J. The neurotransmitter role of diadenosine polyphosphates. *Febs Lett* **430**, 78–82 (1998).
43. Solomon, B. Immunotherapeutic strategies for prevention and treatment of Alzheimer's disease. *DNA Cell Biol* **20**, 697–703 (2001).
44. Duthie, S. J. *et al*. Homocysteine, B vitamin status, and cognitive function in the elderly. *Am J Clin Nutr* **75**, 908–913 (2002).
45. Liu, H. L., Wang, H., Shenvi, S., Hagen, T. M. & Liu, R. M. Glutathione metabolism during aging and in Alzheimer disease. *Ann Ny Acad Sci* **1019**, 346–349 (2004).
46. Frisardi, V., Panza, F., Seripa, D., Farooqui, T. & Farooqui, A. A. Glycerophospholipids and glycerophospholipid-derived lipid mediators: A complex meshwork in Alzheimer's disease pathology. *Prog Lipid Res* **50**, 313–330 (2011).
47. Wang, J. *et al*. Label-free quantitative comparison of cerebrospinal fluid glycoproteins and endogenous peptides in subjects with Alzheimer's disease, mild cognitive impairment, and healthy individuals. *PROTEOMICS-Clinical Applications* **10**, 1225–1241 (2016).
48. Xia, J. G., Mandal, R., Sinelnikov, I. V., Broadhurst, D. & Wishart, D. S. MetaboAnalyst 2.0-a comprehensive server for metabolomic data analysis. *Nucleic Acids Res* **40**, W127–W133 (2012).
49. Ogata, H. *et al*. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **27**, 29–34 (1999).
50. Wishart, D. S. *et al*. HMDB 3.0-The Human Metabolome Database in 2013. *Nucleic Acids Res* **41**, D801–D807 (2013).
51. Cui, Q. *et al*. Metabolite identification via the Madison Metabolomics Consortium Database. *Nature Biotechnology* **26**, 162–164 (2008).
52. Fahy, E., Sud, M., Cotter, D. & Subramaniam, S. LIPID MAPS online tools for lipid research. *Nucleic Acids Res* **35**, W606–W612 (2007).
53. Wolf, S., Schmidt, S., Muller-Hannemann, M. & Neumann, S. In silico fragmentation for computer assisted identification of metabolite mass spectra. *Bmc Bioinformatics* **11** (2010).
54. Hall, M. & Frank, E. Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA Data Mining Software: An Update. *SIGKDD Explorations* **11**, 11 (2009).

Acknowledgements

The authors wish to thank Prof. Xianlin Han from Sanford-Burnham Medical Research Institute for generously providing the custom lipid library to assist with lipid identification. We also acknowledge Dr. Matthew Glover in the Li research group for helpful discussions, as well as the Analytical Instrumentation Center in School of Pharmacy for instruments access. This work was financially supported in part by National Institutes of Health through Grants 1P50AG033514, R01AG052324, 1P20 DK097826, U54 DK104310. The Q-Exactive Orbitrap instrument was purchased through the support of an NIH shared instrument grant (NIH-NCRR S10RR029531). LL acknowledges a Vilas Distinguished Achievement Professorship with funding provided by the Wisconsin Alumni Research Foundation and University of Wisconsin-Madison School of Pharmacy.

Author Contributions

L.H. and J.W. conducted the study under the mentorship of L.L. D.P. provided insights for statistical analysis and machine learning classification. O.C.O. and C.C. recruited the patients and collected the C.S.F. samples. L.H. analyzed the data and wrote the manuscript with input from J.W., D.P., S.A., H.Z., O.C.O., and L.L. All authors provided revisions on the manuscript and approved the final version.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-27031-x.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.