Metabolomics Study of Urine in Autism Spectrum Disorders Using a Multiplatform Analytical Methodology.

Binta Diémé, Sylvie Mavel, Hélène Blasco, Gabriele Tripi, Frédérique Bonnet-Brilhault, Joëlle Malvy, Cinzia Bocca, Christian R Andres, Lydie Nadal-Desbarats, Patrick Emond

Journal of proteome research

Subscriber access provided by UNIV TOURS

Article

Metabolomics study of urine in autism spectrum disorders using a multiplatform analytical methodology

Binta DIEME, Sylvie Mavel, Hélène Blasco, Gabriele Tripi, Frederique Bonnet-Brilhault, Joelle Malvy, Cinzia bocca, Christian R. Andres, Lydie Nadal-Desbarats, and Patrick Emond

J. Proteome Res., **Just Accepted Manuscript** • DOI: 10.1021/acs.jproteome.5b00699 • Publication Date (Web): 05 Nov 2015

Metabolomics study of urine in autism spectrum disorders using a multiplatform analytical methodology

Binta Diémé[a], Sylvie Mavel[a], Hélène Blasco[a,b], Gabriele Tripi[c], Frédérique Bonnet-Brilhault[a,c], Joëlle Malvy [a,c], Cinzia Bocca[a], Christian R Andres[a,b], Lydie Nadal-Desbarats[a], Patrick Emond[a,b,d*]

[a] Université François-Rabelais, INSERM U930 « Imagerie et Cerveau », 37000 Tours, France

[b] Service de biochimie et biologie moléculaire, CHRU de Tours, 37044 Tours, France

[c] Service de pédopsychiatrie, CHRU de Tours, 37044 Tours, France

[d] Service de Médecine Nucléaire In Vitro, CHRU de Tours, 37044 Tours, France

* Corresponding author

ABSTRACT

Autism Spectrum Disorder (ASD) is a neurodevelopmental disorder with no clinical biomarker. Aims of this study were to characterize a metabolic signature of ASD, and to evaluate multi-platform analytical methodologies in order to develop predictive tools for diagnosis and disease follow up.

Urines were analyzed using: ^{1}H- and ^{1} H-^{13}C-NMR-based approaches and LC-HRMS-based approaches (ESI+ and ESI- on a HILIC and C18 chromatography column). Data tables obtained from the six analytical modalities on a training set of 46 urines (22 autistic children and 24 controls) were processed by multivariate analysis (OPLS-DA). Prediction of each of these OPLS-DA models were then evaluated using a prediction set of 16 samples (8 autistic children and 8 controls) and ROC curves. Thereafter, a data fusion block-scaling OPLS-DA model was generated from the 6 best models obtained for each modality. This fused OPLS-DA model showed an enhanced performance (R^2Y(cum)=0.88, Q^2(cum)=0.75) compared to each analytical modality model, as well as a better predictive capacity (AUC=0.91, p-value 0.006). Metabolites that are most significantly different between autistic and control children (p<0.05) are indoxyl sulfate, *N*-(-Acetyl-L-arginine, methyl guanidine and phenylacetylglutamine. This multi-modality approach has the potential to contribute to find robust biomarkers and characterize a metabolic phenotype of the ASD population.

Keywords: metabolomics, Autism Spectrum Disorders, ASD, NMR, LC-HRMS, data fusion

Introduction

Autism spectrum disorder (ASD) refers to a group of complex neurodevelopmental disorders present since early childhood and persisting lifelong.[1] The prevalence of ASD was recently estimated in France to 36.5/10 000 children with a sex ratio of 4.1 boys for 1 girl.[2]

Autism is typically diagnosed before three years of age[3] and ASD children are characterized by deficits in social communication and social interaction, as well as restricted and repetitive behaviors and interests as listed in the Diagnostic and Statistical Manual of Mental Disorders (DSM-V).[4] Diagnosis is made clinically by using different scale tests that evaluate the behavior of the patient[5] and to date, there is no reliable biochemical marker of this disorder. Although etiologies of autism remain unknown, studies have found implication of genetic, environmental and metabolic factors.[6] ASD has several suspected causes, including dysfunctions of the neurologic, immunologic and/or gastrointestinal systems with some markers showing ubiquitous distribution. Due to the social and communication impairments of patients with ASD, identifying gastrointestinal problems remains difficult. However, some studies have linked an imbalance of the gut microbiota with ASD.[7]

Some metabolic disorders have been found to be more frequent in the autistic population compared to the general population. These abnormalities include phenylketonuria, creatine deficiency syndromes, adenylosuccinate lyase deficiency, 5-nucleotidase and metabolism of purine pyrimidine disorders. The part of metabolic abnormalities that contribute to the etiology of autism is still unknown, but these findings suggest that ASD phenotypes may be associated to metabolic pathways imbalance. In order to explore this hypothesis, metabolomics studies which have already shown their potential in biomarker research in central nervous system disorders have been performed.[8]

Metabolomics is the study of the metabolome, which represents the whole content of low-molecular weight compounds present in biological fluids, cells, or tissues.[9] Analytical

platforms most commonly used to identify and quantify metabolites[10–12] are Mass Spectrometry coupled to separation techniques such as gas chromatography (GC-MS) or liquid chromatography (LC-MS)[13] and Nuclear Magnetic Resonance spectroscopy (NMR).

To date, few metabolomics studies have been described for biomarkers exploration in ASD.[14–17] These studies have been primarily based on urine screenings, since urine can be obtained in large quantities by non-invasive sampling. Moreover, repeated sampling is easy to achieve, a major consideration in the case of ASD children. These studies have been performed using a single analytical platform, based on either NMR or MS technologies. Although they have made the proof of concept that exploration of the metabolome allows to classify ASD children compared to control, each analytical platform cannot cover the whole diversity of metabolites in body fluids (molecular diversity and expression levels). Using a single platform also results in partial information and difficulties to confirm identified metabolites as reliable biomarkers. In this study, we would like to take the full advantage of NMR and MS complementarity to explore the urine metabolome, using the combination of ^{1}H-NMR, ^{1}H-^{13}C-HSQC (Heteronuclear single quantum coherence)-NMR and LC-HRMS (Liquid Chromatography coupled to High Resolution Mass Spectrometry). One of the challenges of this approach is to associate data from different analytical platforms in order to generate a statistical model that better represents urinary metabolic differences between children with or without ASD. In order to achieve this, we first built independent supervised multivariate models for each analytical modality in order to select the most discriminant metabolites, which were then concatenated in a new single matrix for a block-scaling model analysis. This multi-platform approach, combined with data fusion, gave us the opportunity to better classify children with or without ASD compared to a unique analytical platform approach.

Experimental section

Patients and Controls

Patients who met ASD diagnostic criteria according to the International Classification of Diseases 10th Edition[18] and the DSM-IV-TR Edition 4th [19] were included in the study after medical consultations at the Regional Center for Autism (CRA) in Tours, between 2011 and 2012. All parents of participants and participants provided informed consent. Urine samples were collected from 30 children with ASD and 32 control children living in France.

The 62 patients urine samples (30 ASD and 32 controls) were split into two sets : (i) a training set of 46 samples (22 ASD and 24 controls), and (ii) an independent validation set of 16 samples (8 ASD and 8 controls). Informations including sex, diagnosis, gastrointestinal disturbances and age at sampling were collected for each participant. Urinary samples were collected in sterile polypropylene tubes untreated with preservatives. After centrifugation at high speed, each urine sample was aliquoted in 1.5 mL sterile Eppendorf tubes and stored at -80°C immediately after collection until analysis.

NMR study

Sample preparation

Urine samples were thawed out at room temperature, and centrifuged at 3000 g for 10 min as described previously. [14,15] The supernatant (500 µL) was then added with 100µL of phosphate buffer (pH = 7.4 ± 0.5) and 100 µL of D_2O solution for 1D analysis or 100 µL of D_2O with internal reference [3-trimethylsilylpropionic acid (TSP), 0.05 wt% in D_2O] for 2D analysis. Samples were then transferred into conventional 5-mm NMR tubes for NMR analysis.

NMR spectroscopy experiments

^{1}H-NMR spectra were obtained as previously described using a Bruker DRX-500 spectrometer (Bruker SADIS, Wissembourg, France) operating at 500 MHz, using a "cpmg" pulse program.

All sensitivity-enhanced ^{1}H-^{13}C-HSQC spectra were adapted from a previously described method on a Bruker DPX Avance spectrometer operating at 300 MHz, using an "hsqcgpphpr" pulse program in the Bruker library.

Data preprocessing for NMR analysis

Spectra were processed using TopSpin version 2.1 software (Bruker Daltonik, Karlsruhe, Germany), then aligned using the work package "speaq"[20] in the R program. The aligned spectra were then displayed in R, and zones with no peaks were removed from file. Finally, intensities of points corresponding to the same peak or nearby peaks were added. These buckets corresponded to either single metabolites or a range of overlapped metabolites. The signal intensity in each bucket was normalized by the total sum of peak intensities, and gathered in the ^{1}H-NMR matrix for further statistical analysis.

2D spectra were processed using MestReNova version 7.1.0 software (Mestrelab Research, S.L., Santiago De Compostela, Spain) as previously described. Each urine spectra was normalized with an external reference TSP that served as a chemical shift reference set at 0 ppm and as a quantitative reference signal. The final 2D matrix contained 677 different ^{1}H-^{13}C cross-peaks between 10 and 150 ppm.

Peaks with low variability (relative standard deviation [RSD] lower than 15 %) were excluded from all data tables since they would not be good predictive biomarkers.[21]

LC-HRMS study

Sample preparation

Samples were prepared from 20 µL of urine, diluted at 1/10 in water or acetonitrile depending on the type of analytical column used (for HILIC or C-18, respectively). After vortexing for

10 min and a centrifugation of 10 min at 10 000g, 150 µL of the supernatant was transferred into a 96 well plate. Quality controls samples (QCs) were obtained from a pooled mixture of equal volumes of all urine samples. QCs followed the same pre-analytic and analytical steps described above. Fifteen QC samples were injected to equilibrate the chromatographic system before each analytic batch. The running order of samples was randomized, and QCs were analyzed every 10 samples.

Liquid Chromatography-High Resolution Mass Spectrometry Analysis

MS1 analysis

LC-HRMS analysis was performed on a UPLC Ultimate 3000 system (Dionex), coupled to a Q-Exactive mass spectrometer (Thermo Fisher Scientific, Bremen, Germany) and operated in positive (ESI+) and negative (ESI-) electrospray ionization modes (one run for each mode). The system was controlled by Xcalibur 2.2 (Thermo Fisher Scientific). Four untargeted LC-HRMS methods were conducted for better metabolome coverage, including C18 and HILIC chromatography coupled to electrospray ionization in both positive and negative ion polarities. Each sample analysis resulted in four separate data acquisitions. Chromatography was carried out with a Phenomenex Kinetex 1.7µm XB – C18 (150 mm×2.10 mm) column with a Waters Cortecs 1.6 µm HILIC (150 mm×2.10 mm) column kept at 40°C. For C18 chromatography, a multi-step gradient (followed by a 2 min equilibration time) had a mobile phase A of 0.1% formic acid in water, and a mobile phase B of acetonitrile (ACN) acidified with 0.1% formic acid; the gradient operated at a flow rate of 0.2 mL/min over a run time of 30 min for both negative and positive modes.

The multi-steps gradient was programmed as follows: 0- 3min: 0 % B; 3-8min: 0-15% B, 8-15min: 15-50 % B; 15-20min: 50-100 % B; 20-25min: 100 % B; 25-28 min: 100-0 % B. For HILIC chromatography, a multi-step gradient (followed by a 2 min equilibration time) had a mobile phase A of ammonium formate in water (10 mM) and a mobile phase B with ACN containing ammonium formate (10 mM); the flow rate was 0.4 mL/min over a run of 22 min

for both positive and negative modes. The multi-steps gradient was programmed as follows: 0- 5min; 0% A, 5-12min; 0-20% A, 12-18.5min; 20-60% A, 18.5-19.5min; 60% A, 19.5-20min; 60-0% A. The autosampler temperature (Ultimate WPS-3000 UHPLC system, Dionex, Germany) was set at 4°C and the injection volume for each sample was 5µL for C18 and 10µL for HILIC.

HESI (heated electrospray ionization) source parameters were, for both modes, a spray voltage of 3 kV, capillary temperature of 380°C or 325°C, heater temperature of 350°C or 325°C, sheath gas flow of 40 arbitrary units (AU) or 35 AU, auxiliary gas flow of 20 AU or 10 AU, spare gas flow of 2 AU or 1 AU, and tube lens voltage of 50 V or 60 V for C18 or HILIC, respectively. During the full-scan acquisition, which ranged from 66.7 to 1000 *m/z*, the instrument operated at 70 000 resolution (*m/z*= 200), with an automatic gain control (AGC) target of 1×10^6 charges and a maximum injection time (IT) of 250 ms.

MS^2 analysis of VIPs

Firstly, a pool sample was injected for molecular ion mass determination at a resolution of 140 000. The most discriminant metabolites (VIPs) obtained from mass spectrometry were then further investigated for their identification. Targeted MS^2 experiments were conducted with an inclusion list of ion mass selected from HILIC and C18 analysis using an isolation window for the quadrupole of 0.5 *m/z* and a resolution of 35 000 (*m/z*=200) for the fragmentation spectrum with an automatic gain control (AGC) target of 2×10^4 charges, a maximum injection time (IT) of 100 ms and a normalized energy collision (Figure S-3).

Data preprocessing for LC-HRMS analysis

XCMS software[22] implanted in the galaxy platform (http://galaxy.workflow4metabolomics.org/) was used to process raw data for peaks alignment and framing. This step produced a table of detected features, characterized by sample retention time, *m/z* ratio, and intensity (i.e. peak area). We normalized each peak area to the total peak area of each chromatogram. For chromatograms obtained with HILIC column in

negative ionization mode, intensities of signals were corrected with Loess[23] after we observed an analytical deviation of signals before normalization to total peak area.

CAMERA package[24] was used to group isotopes and adducts in order to annotate and to identify features.

The stability of signals intensities across batches was evaluated. Extracted ions chromatograms (EICs) were checked in order to review consistency of integration across samples, peak shapes, and to exclude background noise. Variability of features passing this EIC quality review process was then evaluated. QCs variability given by RSD of each feature was assessed. Features with RSD in QCs higher than in samples were excluded. We only kept features with RSD in QCs below 30 % for further multivariate analysis. Features greater than 30 % variance in QC samples were not considered, except if significant variance was observed between groups, meaning that biological variability may exceed analytical variability.[25] Similarly to NMR experiments, we excluded peaks with RSD in samples lower than 15 %.

Data processing

Multivariate data analysis

Quality Control Analysis for LC-HRMS

Clustering of QC samples was assessed by principal component analysis (PCA) according to total peak area data in order to compare analytical variability with biological variability.

Samples batch data analysis

The 62 patients samples (30 ASD and 32 controls) were split into two sets: (i) a training set of 46 samples (22 ASD and 24 controls) for the identification of the most discriminants metabolites between ASD and control urine samples, and (ii) an independent validation set of 16 samples (8 ASD and 8 Controls) to evaluate the performance of the classification models.

This was accomplished by randomizing samples. The training and independent validation sets were matched by age and sex.

The general workflow for the training set is shown in Figure 1.

Please insert figure 1

Each analytical method generated a data table with detected features presented in columns, and urine samples presented in lines. The preprocessed data sets were used as input for Simca P+ version 13.0 (Umetrics, Umeå, Sweden) and data analysis was preceded by log transformation and Unit variance (UV) scaling. The training set of these six data sets were tested individually in order to find the best orthogonal partial least square discriminant analysis (OPLS-DA) model. Model development was performed in order to: (i) select a minimum set of predictive metabolites (VIP, value > 1.0) that are the most implicated in the difference between ASD and control urine samples and (ii) test performance of the optimal model with Receiver Operating Characteristic (ROC) curve analysis and the validation set data. The SIMCA prediction score (Ypred) on the independent validation set was used to build the ROC curves. The main benefit of OPLS-DA compared to PLS-DA is its ability to separate the systemic variation in variables X into two parts: variation related to class membership to variation unrelated to class membership (orthogonal). This partitioning of the X-data will facilitate model interpretation and prediction of new samples.[26] OPLS-DA models performed on training set were then evaluated by cross-validation by withholding 1/7 of the samples in seven successive simulations, so that each sample was omitted once in order to prevent against overfitting. The set of multiple models resulting from cross-validation was used to calculate Jackknife uncertainty measures.[27] We set a maximum number of iterations at 200 in order to ensure convergence of the OPLS algorithm.[28] A Pearson correlation test was performed between discriminant variables for each analytical method to remove the redundancy of information due to VIPs that correspond to the same metabolites. Each of the

six OPLS-DA models performances were assessed using a validation set data and ROC curves analysis.

A data fusion block-scaling model was then generated by combining data tables coming from the previous six OPLS-DA models. Block scaling, in the context of data fusion provides a way to balance influence of blocks of variables in relation to their size.[29] Block scaling allows each group of variables to be considered independently as an entity with a specific variance.[30] This final data fusion block-scaling OPLS-DA model was also evaluated using a ROC curve analysis.

ROC curves were performed using GraphPad Prism version 6.00 for Windows, La Jolla California USA, www.graphpad.com.

Univariate Data Analysis

Univariate analysis focused on the Variable Importance in Projection (VIP>1) obtained from the data fusion block-scaling model using Wilcoxon test. A statistical correction for multiple tests was applied in order to adjust the *p*-value for significance by accounting for the number of metabolites evaluated (Bonferonni adjustment). So differences were deemed significant when $p<0.05/n$ (n= number of VIP). Statistical analyses were performed with JMP statistical software version 7.0.2 (SAS Institute, Cary, NC).

Variable Importance Parameter Annotation

Most discriminant VIPs in multivariate analyses were investigated to be annotated or identified. The VIP assignments for molecular formula elucidation were made with the help of seven golden rules in mass spectrometry.[31] From molecular formula, free access databases queries: Chemspider (http:/www.chemspider.com/), Human Metabolome Database

(http:/www.hmdb.ca/), and MassBank (http:/www.massbank.jp/), were used to annotate compounds. Finally, each VIP was analyzed at high resolution and by MS^2. From HRMS (ion mass and isotopic abundances) one or several parent's structures are generated. To determine the structure of these compounds, Mass Frontier software's fragment comparator gives the opportunity to compare fragments derived from different compounds. Firstly, Mass Frontier software predicted comprehensive pathways based on a set of general ionization, fragmentation and rearrangement rules. Fragmentation spectra were then compared to *ab initio* fragmentation spectra of potential identified metabolites using Mass Frontier.[32] Experimental MS^2 mass spectra with fragment structure annotations are given in the supplementary data (Figure S-3) which illustrates this compound annotation workflow.

For VIPs obtained from 2D-NMR (^1H-^{13}C-HSQC), chemical shifts were submitted to Metabominer database (http://wishart.biology.ualberta.ca/metabominer/). In MetaboMiner, a compound is considered to be present if the requirements of minimal signatures are met. A minimal signature is defined as the minimum peak set of that can uniquely identify a compound from all others in a given spectral library[33]. A compound is considered potentially identified when a minimal signature is highlighted.

For VIPs obtained from ^1H-NMR, we used the Chenomx database (http:/ http://www.chenomx.com/) . These queries resulted in metabolites propositions that were considered for identification if all chemical shifts of theoretical spectra matched our experimental spectra. Chemical shifts of NMR VIPs are given in supplementary information (Table S-1).

RESULTS AND DISCUSSION

The 62 patient urine samples (30 ASD and 32 controls) were split into two sets: (i) a training set of 46 samples (22 ASD [mean 8.64 years] and 24 controls [mean 8.08 years]) and an independent validation set of 16 samples (8 ASD [mean 9.24 years] and 8 controls [mean 9.37 years]). Informations collected for each participant, including age, sex, age at diagnosis, medication and age at sampling, are shown in table 1. Differences of age and sex between groups of training sets are not significant (p-value=0.52 for age and p-value=1 for sex) likewise for validation set (p-value=0.73 for age and p-value=1 for sex).

Please insert Table 1

NMR Experiments

^{1}H-NMR

^{1}H-NMR spectroscopy, a rapid, robust and reliable analytical tool with high reproducibility, has shown its potential to explore the urine metabolome in ASD patients.[34,35]

^{1}H-NMR spectra (0 to 9.5 ppm) were divided in 147 buckets. After removing buckets with low variability in patients (RSD <15 %), the data table of the training set (138 buckets) was analyzed by OPLS-DA. Performances of OPLS-DA internal cross-validated model obtained with 8 discriminant variables, were R^2Y(cum)=0.52, Q^2 =0.37 (Table 2). The sensitivity (percentage of ASD children_correctly identified) obtained from the OPLS-DA model using the training set was 81.8% and specificity (percentage of healthy children correctly identified) was 91.7%. OPLS-DA model p-value was significant (p-value= 6.10^{-4}). This OPLS-DA model was then externally evaluated using the independent validation set and ROC curve analysis. Area under curve (AUC) was 0.83 and p-value was 0.03 (Supplementary material S-2). The sensitivity and specificity obtained from the ROC curve analysis using the independent validation set were lower, with respective values of 62.5% and 87.5% (table 2).

This result shows that the internal validation during OPLS-DA model construction overvalued prediction's capacity due to overfitting of data. Assessing classification performances of each model by an independent set appears to be necessary in order to minimize this overvaluation.

Please insert Table 2

^{1}H-^{13}C-NMR

^{1}H-NMR signals can be disturbed by spectral overlap, resulting in a potential lack in spectral resolution and making difficult the identification step. To improve the resolution of urine components, a two-dimensional NMR (^{1}H-^{13}C-HSQC-NMR) acquisition may be valuable. ^{1}H-^{13}C-HSQC-based NMR avoids spectral overlap by dispatching the overall information in two dimensions. Comparatively to the data table of 138 buckets of ^{1}H-NMR, ^{1}H-^{13}C-HSQC-NMR a list contained 677 different ^{1}H-^{13}C cross-peaks between 10 and 150 ppm was established as previously described due to less overlap of the signals.

The model obtained by ^{1}H-^{13}C-HSQC-NMR gave a better information on prediction with a Q^2=0.51 (Table 2) obtained from 7 cross peaks in the optimized OPLS-DA model compared to ^{1}H-NMR OPLS-DA model (Q^2=0.37). The OPLS-DA model sensitivity for the training set was found higher compared to the ^{1}H-NMR (86.36% *vs* 81.82%). However, specificity was found lower compared to the ^{1}H-NMR (75% *vs* 91.7%). p-value of the ^{1}H-^{13}C-HSQC-NMR OPLS-DA model was significant (*p-value*= 2.10^{-7}).

Regarding the validation set, AUC ROC curve gives similar results compared to ^{1}H-NMR (0.84 *vs* 0.83) and p-value of ^{1}H-^{13}C-HSQC-NMR ROC curve was 0.03.

The sensitivity of ROC curve analysis for independent validation was found better compared to the ^{1}H-NMR (75% *vs* 62.5%) with an equal specificity between them (87.5%). There again, these results show that overfitting of data in OPLS-DA models could be attenuated by the use

of a validation set and a better spectral separation results in a better sensitivity in discriminant analysis.

LC-HRMS analysis

Since only a few untargeted metabolomics studies by LC-MS have been realized with urines of ASD patients, and since they have been done using a single chromatographic column type (C18),[36] we decided, in order to better cover the urine metabolome, to analyze each sample using LC-HRMS coupled to a HILIC or C18 column in both positive (ESI+) and negative (ESI-) ionization modes. Four independent models were built depending on column separation and ionization mode.

Quality control

Fifteen QC samples were injected to equilibrate the chromatographic system before each analytical batch.

PCA scatter plot of QCs and samples were analyzed in order to compare analytical and biological variabilities for each batch. As shown in supplementary data, figure S-1, the clustering of QCs compared to samples clearly shows that biological variability exceeds the analytical one. This QCs step validates all batches series.

HILIC analysis

HILIC provides a good separation for polar metabolites that are of major abundance in urine. This is the first study performed with HILIC chromatography in order to explore ASD urine metabolome. After data preprocessing (see section II.3.3), data tables contained 1067 and 845 features in ESI+ and ESI-. Multivariate analysis gave the similar ability to separate the two

populations of urine samples with slightly better statistical values (R^2Y(cum)=0.51 and Q^2 =0.47) for ESI+ compared to ESI- (R^2Y(cum)=0.4 and Q^2 =0.34).

OPLS-DA model's sensitivities (built with 9 and 4 variables from ESI+ and ESI- respectively) were 86.4% and 77.3% for ESI+ and ESI- respectively. Likewise, OPLS-DA model's specificity were 91.7% and 79.2% for of ESI+ and ESI- respectively. AUCs curves for the validation set (ESI+ and ESI- respectively) were 0.70 and 0.64 ((Supplementary material S-2). Model's curves were not found to be significant (p>0.05), unlike p-values of OPLS-DA models that were significant (table 2). Reasons for these discrepancies could be an OPLS-DA models overfitting and/or a too small number of samples for the validation set.

C18 analysis

Reverse-phase separations are ideal for relatively non-polar metabolites. 1555 and 1226 features in ESI+ and ESI-, respectively, were selected as repeatable. From C18 ESI- data, an OPLS-DA model built with 11 discriminant variables, gave statistical parameters R^2Y(cum)= 0.48 and Q^2= 0.39 (table 2). The sensitivity and specificity of this OPLS-DA model were 86.36% and 75% respectively with a significant p-value of 2.10^{-5}. Performances of this model were also evaluated using ROC curve and we found similar sensitivity and specificity compared to the OPLS-DA model (table 2). AUC was found to be 0.83 with a significant p-value 0.03 (Supplementary material S-2). The significant p-value of ROC curve confirms here that the OPLS-DA model is able to classify ASD patients and controls with minimum overfitting.

The OPLS-DA model obtained with 9 discriminant variables from C18-ESI+ data gave better information on the class descriptor with a R^2Y(cum)=0.64. Moreover, compared to other LC-HRMS modalities, ESI+ model gave highest predictive ability with a Q^2=0.53 and a p-value of 3.10^{-6}. OPLS-DA model sensitivity and specificity were 90.9% and 91.7% respectively. Despite these models results, the ROC curve AUC value was not found to be significant (p-value 0.06).

Multivariate Statistical Analysis of the combined ^{1}H-NMR, ^{1}H-^{13}C-HSQC-NMR, ESI±C18 column and ESI±HILIC column

The fusion and extraction of information from multiple data tables has become a decisive issue.[37] Several strategies can be used to associate data from different analytical platforms. The simplest approach is to concatenate the different data sets in a low-level fusion approach where the data matrix results in a fused data table that is used for multivariate analysis.[38] This approach is greatly affected by disparate signal intensities range and size of data matrix obtained from the different analytical platforms. In order to resolve this issue, the selection of the most relevant/predictive variables from each data tables may solve the problem of dimensionality. Such strategies are called intermediate or mid-level data fusion.[39] The model responses are combined in order to produce a final 'fused' response that provides a meaningful synthesis. Following a mid-level data fusion strategy, the most discriminant variables of each analytical platform (^{1}H-NMR, ^{1}H-^{13}C-HSQC, C18 ESI±, and HILIC ESI±) were selected and combined in order to build a data fusion block-scaling model. Since a subset of the most valuable variables can be selected from each data sources, the prediction performance can be increased when compared with individual analysis. While some signals are expected to be common to different blocks, the remaining would be specific and should bring complementary information. Since a data fusion model could be dominated by the largest matrices for numerical reasons, the fairness between blocks was ensured by UV scaling normalization.

We first built six independent OPLS-DA models in order to select the most discriminant variable (VIP >1.0) for each analytical platforms: 8, 7, 7, 11, 9 and 4 variables from ^{1}H-NMR, ^{1}H-^{13}C-HSQC, C18 ESI+, C18 ESI- , HILIC ESI+ and HILIC ESI-, respectively. An equal weight was then assigned to each block corresponding to each analytical platform in the data table.

The data fusion block-scaling model was obtained from X=46 features giving performances parameters R^2Y(cum)=0.88 Q^2=0.75, p-value=9.10^{-12} and 100% of specificity and sensibility (table 2, figure 2).

When we performed the ROC curve analysis, an AUC of 0.91 was obtained with a significant p-value 0.006 (table 2, figure 2). In addition, sensitivity was found to be improved when compared to each separate analytical platform.

Please insert figure 2

Besides its classification capacities, OPLS-DA model gives the opportunity to highlight metabolites that are the most involved with this classification.

Table 3 shows the most discriminant metabolites (VIP>1), their variations in ASD group and univariate p-value associated.

Please insert table 3

Firstly, we found higher levels of *N*-acetylarginine in the ASD group compared to controls. Arginine has been described to be higher in plasma samples of ASD children.[40] Excessive arginine is thought to induce oxidative stress via NO production.[41] We also found perturbations of guanidinosuccinic acid and methylguanidine which are produced from oxidation of argininosuccinic acid and creatine respectively by free radicals.[42,43] The exact relationship between arginine pathway and oxidative stress in neuropsychiatric disorders remains unclear; however, a common susceptibility gene for ASD and schizophrenia, NOS1, has been suggested to be involved in the arginine-NO pathway.[44,45] This metabolic profile is consistent with impaired oxidative stress in children with autism.[46] The exact relationship between guanidino compounds pathway and autism needs further investigation.

Dihyroxy-1H-indole glucuronide I and desaminotyrosine, which are related to tyrosine metabolism, were also found in the most discriminant metabolites. These results, associated to perturbations of level of valine[47,48] and N-acetylasparagine, are in agreement with perturbations of amino acids levels published by Tu et al[48] and may be interpreted as abnormal amino acids metabolism affecting neurotransmitters levels such as dopamine, noradrenaline and epinephrine as observed in ASD children.[48]

Level of dihydrouracil, an intermediate breakdown product of uracil, was found altered in ASD children's metabotypes. Purines and pyrimidines disorders such as dihydropyrimidine dehydrogenase or dihydropyrimidinase deficiencies, adenylosuccinate lyase or adenosine deaminase deficiencies, have been linked to autistic features.

Deficiencies of dihydropyrimidine dehydrogenase, the enzyme which catalyzes the conversion of uracil to dihydrouracil have been linked to ASD. Indeed, symptoms of this pyrimidine disorder include psychomotor retardation, epileptic encephalopathy and autistic features.

Recent studies have shown that an antipurinergic therapy could reverse the behavioral and metabolic disturbances in the maternal immune activation mouse model[49,50] and in the Fragile X (Fmr1 knockout) mouse model. From these two mechanically distinct examples of ASD mouse models, the purinergic pathway is a neurochemical hypothesis that triggers the evolutionarily conserved cell danger to stress that may be associated with ASD.

Indoxyl and indoxyl sulfate, which are produced by tryptophan metabolism in gut bacteria, were found here as metabolite candidates (from different analytical modality: NMR 1D and 2D and HILIC ESI+) and confirmed our previous results.[14] Since indoxyl sulfate has been found as a VIP in more than one analytical platform, it may be assumed that this metabolite expression is modified between our two groups of children.

Desaminotyrosine, which can also be the result of deamination of tyrosine by intestinal microflora[51,52], was found increased in the ASD group. Moreover, a recent study by Noto et al using GC/MS as an analytical platform points out a gut microbiota dysfunction, including tyrosine metabolism perturbation.[53] Our study also points out perturbations of phenylacetylglutamine (PAG) and p-cresol sulfate concentrations, which are also produced by the microbiota respectively from tyrosine[54] and phenylalanine.[55] These results which confirmed those previously reported[15,55] underline the importance of mammalian-microbial cometabolites in ASD [16,17,53], supporting emerging evidence for a gut-brain connection in autism, wherein gastrointestinal microbiota may contribute to the ASD symptoms.[56] It has to be noticed that half of our autistic cohort is clinically diagnosed as suffering of gastrointestinal disturbances that include diarrhea, constipation and colitis.

Whether it is the cause or the consequence of autism's physiopathology, our results confirm that gut microbiome seems to be associated with this disorder.

It has been recently shown that there are Gender differences in emotional and sociability in children with autism spectrum disorders[57]. This suggests that, in addition to phenotypic differences, there are metabotype differences linked to gender. One should keep in mind that the statistical power of our study is related to the small size of our groups of patients (training and control sets), emphasized by the small number of patients enrolled. Since only a few females have been included in this study, this tends to decrease the biological significance of our conclusions. However, significant metabolic differences found in this work between ASD and control children may underline a typical general metabolic signature of ASD.

Conclusion:

Discovery of biomarkers for an early diagnosis of ASD, as well as a follow up of its evolution for an improved patient care, is still a challenge to achieve. Because of its non-invasive accessibility and since it has already shown its potential for containing discriminative metabolites, we decided to explore the children's urines metabolome as deep as possible,

using six complementary analytical platforms.: ^{1}H-NMR, ^{1}H-^{13}C NMR, ESI+ ESI- LC-MS with C18 and HILIC chromatographic support.

Firstly, our results highlighted and confirmed that several metabolic pathways belonging to amino acids including tyrosine, asparagine, phenylalanine, tryptophan and arginine— seem to be involved in ASD. In addition, as other studies have previously shown, a gut dysbiosys is likely to be associated with ASD. It has to be noticed that all these metabolism perturbations could be observed in a single study, thanks to the use of a multiplatform strategy that enabled a deep exploration of the metabolome. Secondly, our results show that OPLS-DA model construction based on data fusion and block scaling by combining the most discriminant variables from multiple analytical methods, results in a valuable model of prediction. Using the complementarity of analytical methods associated to the selection of the most relevant variables, our OPLS-DA model has a raised the predictive power with an AUC of 91 % in the validation set, which confirms the promise of combining multiple analytical methods by multivariates analyzes.

Our results need to be validated within a larger cohort of patients, firstly in order to confirm this metabolites panel as a potential clinical tool, but also in order to explore the metabotype variability associated with ASD phenotype's heterogeneity especially when considering the gender.

Associated content

Supporting information

Two figures showing PCA of QC samples, and ROC curve analysis for each analytical platform. This material is available free of charge via the Internet at http://pubs.acs.org.

Contents:

Figure S-1: PCA scatter plot of quality controls (QCs) and controls injected during each batch analysis. ASD patients, controls and QCs are respectively colored in blue, green and grey.

a) scatter plot for HILIC ESI-,b) scatter plot for HILIC ESI+, c) scatter plot for C18 ESI- and d) scatter plot for C18 ESI+

Figure S-2: ROC curves for independent validation set for each analytical modality.

a) ^{1}H-NMR, b) ^{1}H-^{13}C-NMR, c) HILIC ESI+, d) HILIC ESI-, e) C18 ESI+ and f) C18 ESI-

Figure S-3: MS^2 spectra of most discriminant metabolites putatively annotated in multivariate analysis: a) dihydroxy-1H-indole glucuronide I, b) Dihydrouracil, c) *N*-α-Acetyl-L-arginine, d) *N*-acetylasparagine, e) Desaminotyrosine, f) Guanidinosuccinic acid, g) indoxyl, h) *N*-α-Phenylacetyl-L-glutamine, i) p-cresol sulfate

Table S-1: putatively annotated metabolites detected by NMR (VIP>1).

Author information

Corresponding Author

* Corresponding author: Université François Rabelais, INSERM U930, 10 Bv Tonnellé, 37044 Tours, France Tel: + 33 2 47 36 61 53 ; e-mail: patrick.emond@univ-tours.fr

Acknowledgements

The authors thank Odette Viaud who has contributed to this work with drive and dedication.

The authors thank Anne Wick who has corrected the English language of this work.

This work was supported by the “Institut National de la Santé et de la Recherche” INSERM, by the University François-Rabelais de Tours. We thank “La Région Centre” for a PhD graduate grant. We thank the “Département de métabolomique et d’Analyses Chimiques, ” PPF-ASB, Tours, France website : http://ppf.med.univ-tours.fr for chemical analyses.

Notes

The authors declare no competing financial interest.

References

(1) Baio, J. *Prevalence of Autism Spectrum Disorders: Autism and Developmental Disabilities Monitoring Network, 14 Sites, United States, 2008.* ; Centers for Disease Control and Prevention, 2012; Vol. 61.

(2) Van Bakel, M. M. E.; Delobel-Ayoub, M.; Cans, C.; Assouline, B.; Jouk, P.-S.; Raynaud, J.-P.; Arnaud, C. Low but Increasing Prevalence of Autism Spectrum Disorders in a French Area from Register-Based Data. *J. Autism Child. Schizophr.* **2015**, 1–7.

(3) Kolvin, I. Studies in the childhood psychoses: I. Diagnostic criteria and classification. *Br. J Psychiat* **1971**, *118*, 381–384.

(4) American Psychiatric Association. *Diagnostic and statistical manual of mental disorders: DSM-5*, 5th ed.; American Psychiatric Association Publishing: Arlington, 2013.

(5) Lord, C.; Rutter, M.; Le Couteur, A. Autism Diagnostic Interview-Revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J. Autism Child. Schizophr.* **1994**, *24* (5), 659–685.

(6) Goldani, A. A.; Downs, S. R.; Widjaja, F.; Lawton, B.; Hendren, R. L. Biomarkers in autism. *Front Psychiatry* **2014**, *5*, 100.

(7) Wang, L.; Conlon, M. A.; Christophersen, C. T.; Sorich, M. J.; Angley, M. T. Gastrointestinal microbiota and metabolite biomarkers in children with autism spectrum disorders. *Biomark Med* **2014**, *8* (3), 331–344.

(8) Trushina, E.; Mielke, M. M. Recent advances in the application of metabolomics to Alzheimer's Disease. *Biochim Biophys Acta* **2014**, *1842* (8), 1232–1239.

(9) Nicholson, J. K.; Lindon Jc Fau - Holmes, E.; Holmes, E.; Xenobiotica. "Metabonomics": understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *xenobiotica* **1999**, *29*, 1181–1189.

(10) Gebregiworgis, T.; Powers, R. Application of NMR metabolomics to search for human disease biomarkers. *Comb Chem High T Scr* **2012**, *15* (8), 595–610.

(11) Nicholson, G.; Rantalainen, M.; Maher, A. D.; Li, J. V; Malmodin, D.; Ahmadi, K. R.; Faber, J. H.; Hallgrímsdóttir, I. B.; Barrett, A.; Toft, H. Human metabolic profiles are stably controlled by genetic and environmental variation. *Mol Syst Biol* **2011**, *7* (1), 525.

(12) Patti, G. J.; Yanes, O.; Siuzdak, G. Innovation: Metabolomics: the apogee of the omics trilogy. *Nat Rev Mol Cell Bio* **2012**, *13* (4), 263–269.

(13) Żurawicz, E.; Kałużna-Czaplińska, J.; Rynkowski, J. Chromatographic methods in the study of autism. *Biomed Chromatogr* **2013**, *27* (10), 1273–1279.

(14) Mavel, S.; Nadal-Desbarats, L.; Blasco, H.; Bonnet-Brilhault, F.; Barthelemy, C.; Montigny, F.; Sarda, P.; Laumonnier, F.; Vourc'h, P.; Andres, C. R.; et al. 1H-13C NMR-based urine metabolic profiling in autism spectrum disorders. *Talanta* **2013**, *114*, 95–102.

(15) Nadal-Desbarats, L.; Aïdoud, N.; Emond, P.; Blasco, H.; Filipiak, I.; Sarda, P.; Bonnet-Brilhault, F.; Mavel, S.; Andres, C. R. Combined 1 H-NMR and 1 H–13 C HSQC-NMR to improve urinary screening in autism spectrum disorders. *Analyst* **2014**, *139* (13), 3460–3468.

(16) Yap, I. K. S.; Angley, M.; Veselkov, K. A.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. Urinary metabolic phenotyping differentiates children with autism from their unaffected siblings and age-matched controls. *J Proteome Res* **2010**, *9* (6), 2996–3004.

(17) Ming, X.; Stein, T. P.; Barnes, V.; Rhodes, N.; Guo, L. Metabolic perturbance in autism spectrum disorders: a metabolomics study. *J Proteome Res* **2012**, *11* (12), 5856–5862.

(18) World Health Organization. *International statistical classification of diseases and related health problems*; World Health Organization: Geneva, 2004; Vol. 1.

(19) American Psychiatric Association. *Diagnostic and Statistical Manual of Mental Disorders, 4th Edition, Text Revision (DSM-IV-TR)*; American Psychiatric Association Publishing: Arlington, 2000.

(20) Vu, T. N.; Valkenborg, D.; Smets, K.; Verwaest, K. A.; Dommisse, R.; Lemière, F.; Verschoren, A.; Goethals, B.; Laukens, K. An integrated workflow for robust alignment and simplified quantitative analysis of NMR spectrometry data. *BMC Bioinformatics* **2011**, *12* (1), 405.

(21) Xia, J.; Broadhurst, D. I.; Wilson, M.; Wishart, D. S. Translational biomarker discovery in clinical metabolomics: an introductory tutorial. *Metabolomics* **2013**, *9* (2), 280–299.

(22) 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* **2006**, *78* (3), 779–787.

(23) Dunn, W. B.; Broadhurst, D.; Begley, P.; Zelena, E.; Francis-McIntyre, S.; Anderson, N.; Brown, M.; Knowles, J. D.; Halsall, A.; Haselden, J. N.; et al. Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nat Protoc* **2011**, *6* (7), 1060–1083.

(24) Kuhl, C.; Tautenhahn, R.; Bottcher, C.; Larson, T. R.; Neumann, S. CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets. *Anal Chem* **2012**, *84* (1), 283–289.

(25) Want, E. J.; Wilson, I. D.; Gika, H.; Theodoridis, G.; Plumb, R. S.; Shockcor, J.; Holmes, E.; Nicholson, J. K. Global metabolic profiling procedures for urine using UPLC–MS. *Nat Protoc* **2010**, *5* (6), 1005–1018.

(26) Eriksson, L.; Byrne, T.; Johansson, E.; Trygg, J.; Vikström, C. *Multi-and megavariate data analysis basic principles and applications*; Umetrics Academy: Umea, 2013.

(27) Blasco, H.; Corcia, P.; Pradat, P. F.; Bocca, C.; Gordon, P. H.; Veyrat-Durebex, C.; Mavel, S.; Nadal-Desbarats, L.; Moreau, C.; Devos, D.; et al. Metabolomics in cerebrospinal fluid of patients with amyotrophic lateral sclerosis: an untargeted approach via high-resolution mass spectrometry. *J Proteome Res* **2013**, *12* (8), 3746–3754.

(28) Westerhuis, J. A.; van Velzen, E. J.; Hoefsloot, H. C.; Smilde, A. K. Multivariate paired data analysis: multilevel PLSDA versus OPLSDA. *Metabolomics* **2010**, *6* (1), 119–128.

(29) Boccard, J.; Rudaz, S. Harnessing the complexity of metabolomic data with chemometrics. *J Chemom.* **2014**, *28* (1), 1–9.

(30) Biais, B.; Allwood, J. W.; Deborde, C.; Xu, Y.; Maucourt, M.; Beauvoit, B.; Dunn, W. B.; Jacob, D.; Goodacre, R.; Rolin, D. 1H NMR, GC− EI-TOFMS, and Data Set Correlation for Fruit Metabolomics: Application to Spatial Metabolite Analysis in Melon. *Anal Chem* **2009**, *81* (8), 2884–2894.

(31) Kind, T.; Fiehn, O. Seven golden rules for heuristic filtering of molecular formulas obtained by accurate mass spectrometry. *BMC Bioinformatics* **2007**, *8* (1), 105.

(32) Huang, Y.; Herbold, R.; Nakamura, D. Identification of Metabolites from Maropitant Using a Dual-Pressure Linear Ion Trap and Mass Frontier Software. In *Drug Metabolism Reviews*; Taylor & Francis: Philadelphia, 2010; Vol. 42, pp 223–224.

(33) Xia, J.; Bjorndahl, T. C.; Tang, P.; Wishart, D. S. MetaboMiner–semi-automated identification of metabolites from 2D NMR spectra of complex biofluids. *BMC Bioinformatics* **2008**, *9* (1), 507.

(34) Ryan, D.; Robards, K.; Prenzler, P. D.; Kendall, M. Recent and potential developments in the analysis of urine: a review. *Anal Chim Acta* **2011**, *684* (1-2), 8–20.

(35) Zhang, T.; Creek, D. J.; Barrett, M. P.; Blackburn, G.; Watson, D. G. Evaluation of coupling reversed phase, aqueous normal phase, and hydrophilic interaction liquid chromatography with Orbitrap mass spectrometry for metabolomic studies of human urine. *Anal Chem* **2012**, *84* (4), 1994–2001.

(36) Fei, F.; Bowdish, D. M. E.; McCarry, B. E. Comprehensive and simultaneous coverage of lipid and polar metabolites for endogenous cellular metabolomics using HILIC-TOF-MS. *Anal Bioanal Chem* **2014**, *406* (15), 3723–3733.

(37) Boccard, J.; Rutledge, D. N. A consensus orthogonal partial least squares discriminant analysis (OPLS-DA) strategy for multiblock Omics data fusion. *Anal Chim Acta* **2013**, *769*, 30–39.

(38) Roussel, S.; Bellon-Maurel, V.; Roger, J.-M.; Grenier, P. Authenticating white grape must variety with classification models based on aroma sensors, FT-IR and UV spectrometry. *J. Food Eng* **2003**, *60* (4), 407–419.

(39) Blanchet, L.; Smolinska, A.; Attali, A.; Stoop, M. P.; Ampt, K. A.; van Aken, H.; Suidgeest, E.; Tuinstra, T.; Wijmenga, S. S.; Luider, T.; et al. Fusion of metabolomics and proteomics data for biomarkers discovery: case study on the experimental autoimmune encephalomyelitis. *BMC Bioinformatics* **2011**, *12*, 254.

(40) Kuwabara, H.; Yamasue, H.; Koike, S.; Inoue, H.; Kawakubo, Y.; Kuroda, M.; Takano, Y.; Iwashiro, N.; Natsubori, T.; Aoki, Y.; et al. Altered metabolites in the plasma of autism spectrum disorder: a capillary electrophoresis time-of-flight mass spectroscopy study. *PLoS One* **2013**, *8* (9), e73814.

(41) Delwing, D.; Delwing, D.; Bavaresco, C. S.; Wyse, A. T. S. Protective effect of nitric oxide synthase inhibition or antioxidants on brain oxidative damage caused by intracerebroventricular arginine administration. *Brain Res* **2008**, *1193*, 120–127.

(42) De Jonge, W. J.; Marescau, B.; D'Hooge, R.; De Deyn, P. P.; Hallemeesch, M. M.; Deutz, N. E. P.; Ruijter, J. M.; Lamers, W. H. Overexpression of arginase alters circulating and tissue amino acids and guanidino compounds and affects neuromotor behavior in mice. *J Nutr* **2001**, *131* (10), 2732–2740.

(43) Tachikawa, M.; Hosoya, K. Transport characteristics of guanidino compounds at the blood-brain barrier and blood-cerebrospinal fluid barrier: relevance to neural disorders. *Fluids Barriers CNS* **2011**, *8* (1), 13.

(44) Kim, H.-W.; Cho, S.-C.; Kim, J.-W.; Cho, I. H.; Kim, S. A.; Park, M.; Cho, E. J.; Yoo, H.-J. Family-based association study between NOS-I and -IIA polymorphisms and autism spectrum disorders in Korean trios. *Am J Med Genet B* **2009**, *150B* (2), 300–306.

(45) O'donovan, M. C.; Craddock, N.; Norton, N.; Williams, H.; Peirce, T.; Moskvina, V.; Nikolov, I.; Hamshere, M.; Carroll, L.; Georgieva, L. Identification of loci associated with schizophrenia by genome-wide association and follow-up. *Nat Genet* **2008**, *40* (9), 1053–1055.

(46) James, S. J.; Cutler, P.; Melnyk, S.; Jernigan, S.; Janak, L.; Gaylor, D. W.; Neubrander, J. A. Metabolic biomarkers of increased oxidative stress and impaired methylation capacity in children with autism. *Am J Clin Nutr* **2004**, *80* (6), 1611–1617.

(47) Evans, C.; Dunstan, H. R.; Rothkirch, T.; Roberts, T. K.; Reichelt, K. L.; Cosford, R.; Deed, G.; Ellis, L. B.; Sparkes, D. L. Altered amino acid excretion in children with autism. *Nutr. Neurosci.* **2008**, *11* (1), 9–17.

(48) Tu, W.-J.; Chen, H.; He, J. Application of LC-MS/MS analysis of plasma amino acids profiles in children with autism. *J. Clin. Biochem. Nutr.* **2012**, *51* (3), 248.

(49) Naviaux, J. C.; Schuchbauer, M. A.; Li, K.; Wang, L.; Risbrough, V. B.; Powell, S. B.; Naviaux, R. K. Reversal of autism-like behaviors and metabolism in adult mice with single-dose antipurinergic therapy. *Transl Psychiat* **2014**, *4* (6), e400.

(50) Naviaux, R. K.; Zolkipli, Z.; Wang, L.; Nakayama, T.; Naviaux, J. C.; Le, T. P.; Schuchbauer, M. A.; Rogac, M.; Tang, Q.; Dugan, L. L. Antipurinergic therapy

corrects the autism-like features in the poly (IC) mouse model. *PLoS One* **2013**, *8* (3), e57380.

(51) Holmes, E.; Li, J. V; Athanasiou, T.; Ashrafian, H.; Nicholson, J. K. Understanding the role of gut microbiome–host metabolic signal disruption in health and disease. *Trends Microbiol* **2011**, *19* (7), 349–359.

(52) Lambert, M. A.; Moss, C. W. Production of p-hydroxyhydrocinnamic acid from tyrosine by Peptostreptococcus anaerobius. *J. Clin Microbiol* **1980**, *12* (2), 291–293.

(53) Noto, A.; Fanos, V.; Barberini, L.; Grapov, D.; Fattuoni, C.; Zaffanello, M.; Casanova, A.; Fenu, G.; De Giacomo, A.; De Angelis, M.; et al. The urinary metabolomics profile of an Italian autistic children population and their unaffected siblings. *J. Matern. Neonatal Med.* **2014**, *27* (S2), 46–52.

(54) Patel, K. P.; Luo, F. J.-G.; Plummer, N. S.; Hostetter, T. H.; Meyer, T. W. The production of p-cresol sulfate and indoxyl sulfate in vegetarians versus omnivores. *Clin J Am Soc Nephro* **2012**, *7* (6), 982–988.

(55) Gabriele, S.; Sacco, R.; Cerullo, S.; Neri, C.; Urbani, A.; Tripi, G.; Malvy, J.; Barthelemy, C.; Bonnet-Brihault, F.; Persico, A. M. Urinary p-cresol is elevated in young French children with autism spectrum disorder: a replication study. *Biomarkers* **2014**, *19* (6), 463–470.

(56) Hsiao, E. Y. Gastrointestinal issues in autism spectrum disorder. *Harvard Rev Psychiat* **2014**, *22* (2), 104–111.

(57) Head, A. M.; McGillivray, J. A.; Stokes, M. A. Gender differences in emotionality and sociability in children with autism spectrum disorders. *Mol Autism* **2014**, *5* (1), 19.

Table 1: ASD and controls groups characteristics

	Training set		Independent validation set	
	ASD *(n=22)*	Controls (*n*=24)	ASD (*n=8)*	Controls *(n=8)*
age, years [a]	8.64 ± 3.62	8.08 ± 3.67	9.24±3.79	9.37±4.07
no. of males (%) [a]	19 (86.36%)	21 (87.50%)	7 (87.50%)	6 (75%)
no. of females (%) [a]	3 (13.64%)	3 (12.50%)	1 (12.50%)	2 (25%)
Diagnostic				
Autism disorder	9	0	2	0
Asperger Syndrome	2	0	0	0
PDD-NOS	11	0	6	0
Neurotypical	0	24	0	8
Gastro-intestinal disturbance				
yes	9	NA	5	NA
no	12	NA	3	NA
NA	1	24	0	8

[a] Data are expressed as mean

Table 2: Summary of statistical values of OPLS-DA models and ROC curve analysis of training set and independent validation set obtained with the different methodologies. The different cumulated modeled variations in Y (R^2Y(cum)] matrices on spectral datasets and predictability of the model (Q^2) are given [observations (N)=46].

Models	Training set (OPLS-DA model)						validation set (ROC curves)			
	X variables	*R2Y* (cum)	*Q2* (cum)	Sensitivity[a]	Specificity[b]	p-value[c]	AUC	*p*-value[c]	Sensitivity[a]	Specificity[b]
^{1}H-NMR	8	0.52	0.37	81.8%	91.7%	6.10^{-4}	0.83	0.02	62.5%	87.5%
^{1}H-^{13}C-NMR	7	0.59	0.51	86.4 %	75.0 %	2.10^{-7}	0.84	0.03	75.0 %	87.5 %
HILIC ESI+	9	0.51	0.47	86.4 %	91.7%	1.10^{-6}	0.70	0.17	100 %	62.5%
HILIC ESI-	4	0.40	0.34	77.3 %	79.2 %	1.10^{-3}	0.64	0.34	62.5 %	75.0 %
C18 ESI+	7	0.64	0.53	90.9 %	91.7%	3.10^{-6}	0.78	0.06	62.5 %	87.5 %
C18 ESI-	11	0.48	0.39	86.4 %	75.0 %	2.10^{-5}	0.83	0.03	87.5 %	75.0 %

	Training set (OPLS-DA model)						validation set (ROC curve)			
	X variables	*R2Y* (cum)	*Q2* (cum)	Sensitivity[a]	Specificity[b]	*p*-value[c]	AUC	*p*-value[c]	Sensitivity[a]	Specificity[b]
Block Model	46	0.88	0.75	100 %	100 %	9.10^{-12}	0.91	0.006	100 %	75.0%

[a] Sensitivity (S_n = the number of diseased subjects that are correctly identified as diseased).

[b] Specificity (S_p =the number of healthy subjects that are correctly identified as healthy) on training set and validation set.

[c] *p*-value of ROC curve analysis

Table 3: putatively annotated metabolites of data fusion OPLS-DA model (VIP>1)

Analytical platform	Potential assignment	*p*-value[a]	Differentiation for ASD samples[b]
C18 ESI-	Dihyroxy-1H-indole glucuronide I	0.005	↑
C18 ESI+	Dihydrouracil	0.03	↓
C18 ESI+	N-a-Acetyl-L-arginine	0.009	↑
C18 ESI+	Unknown	0.009	↓
C18 ESI+	Unknown	0.006	↑
C18 ESI+	N-Acetylasparagine	0.03	↑
C18 ESI+	Desaminotyrosine	0.006	↓
C18 ESI+	Guanidinosuccinic acid	0.007	↓
HILIC ESI+	Indoxyl	0.01	↑
HILIC ESI+	**Unknown**	**0.003***	↓
HILIC ESI+	Valine, norvaline, 5-aminopentanoic acid	0.007	↑
HILIC ESI-	**Unknown**	**0.0004***	↑
HILIC ESI-	Alpha-*N*-Phenylacetyl-L-glutamine	0.004	↑
HILIC ESI-	p-cresol sulfate	0.02	↑
RMN 1D	Indoxyl sulfate	0.01	↑
RMN 1D	**Methylguanidine**	**0.003***	↓
RMN 2D	Valine	0.06	↑
RMN 2D	**Indoxyl sulfate**	**0.003***	↑
RMN 2D	Glucuronic acid	0.009	↑

[a] *p*-value of univariate analysis (Wilcoxon test)

[b] ↑ denotes higher level in ASD urines, ↓ denotes lower level in ASD urines

* significant *p*-value after Bonferroni correction ($p<0.003$)

Figure 1: workflow of treatment of data tables from NMR and LC-HRMS analysis

Figure 2: OPLS-DA model and ROC Curve analysis obtained from training set and independent validation set. a) data fusion block-scaling of OPLS-DA model on training set [R2Y(cum)=0.88 Q2=0.75, p =9.10^{-12}], b) ROC curve analysis of independent validation set (8 ASD and 8 controls).

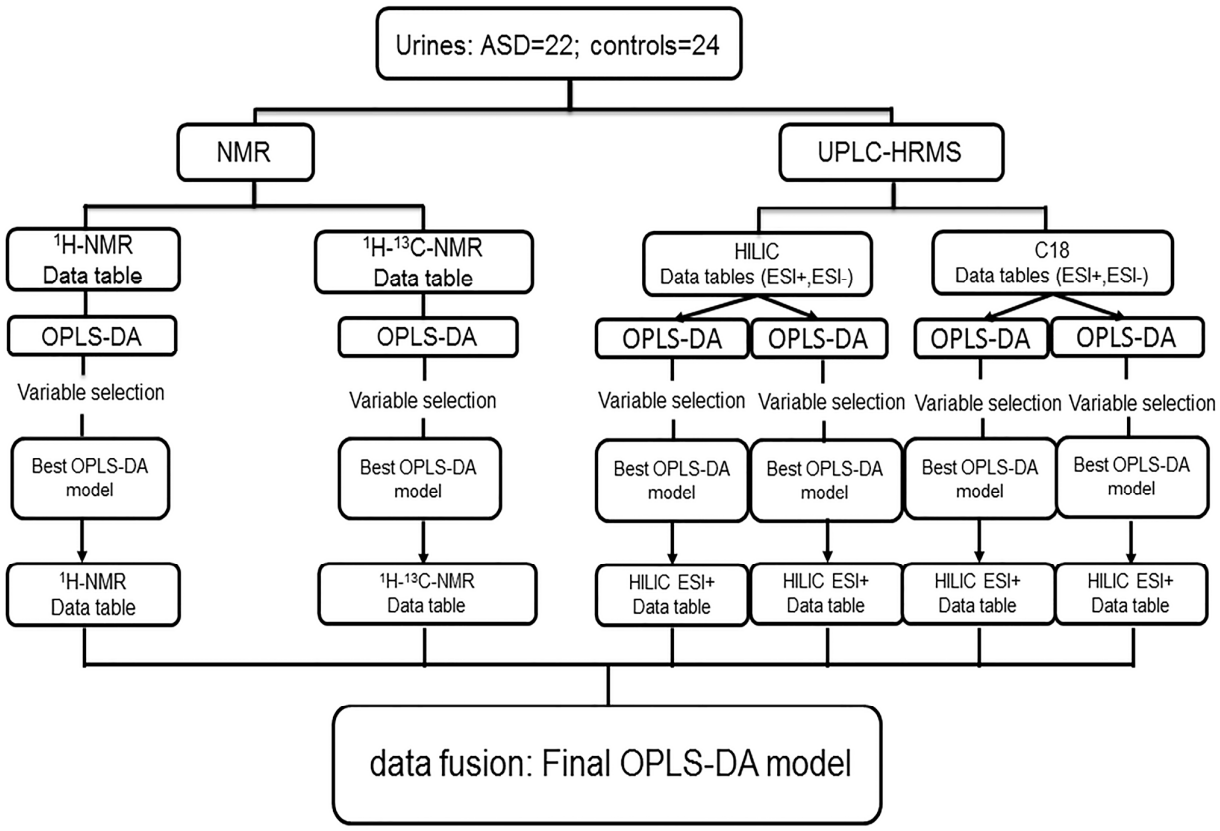

Urines: ASD=22; controls=24
NMR
UPLC-HRMS
1H-NMR Data table
1H-13C-NMR Data table
HILIC Data tables (ESI+,ESI-)
C18 Data tables (ESI+,ESI-)
OPLS-DA
OPLS-DA
OPLS-DA
OPLS-DA
OPLS-DA
OPLS-DA
Variable selection
Variable selection
Variable selection
Variable selection
Variable selection
Variable selection
Best OPLS-DA model
Best OPLS-DA model
Best OPLS-DA model
Best OPLS-DA model
Best OPLS-DA model
Best OPLS-DA model
1H-NMR Data table
1H-13C-NMR Data table
HILIC ESI+ Data table
HILIC ESI+ Data table
HILIC ESI+ Data table
HILIC ESI+ Data table
data fusion: Final OPLS-DA model

Figure 1

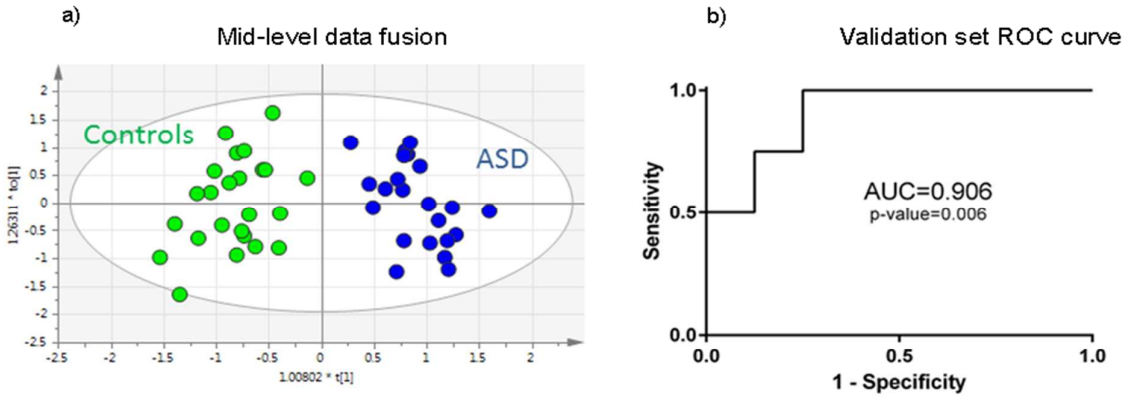
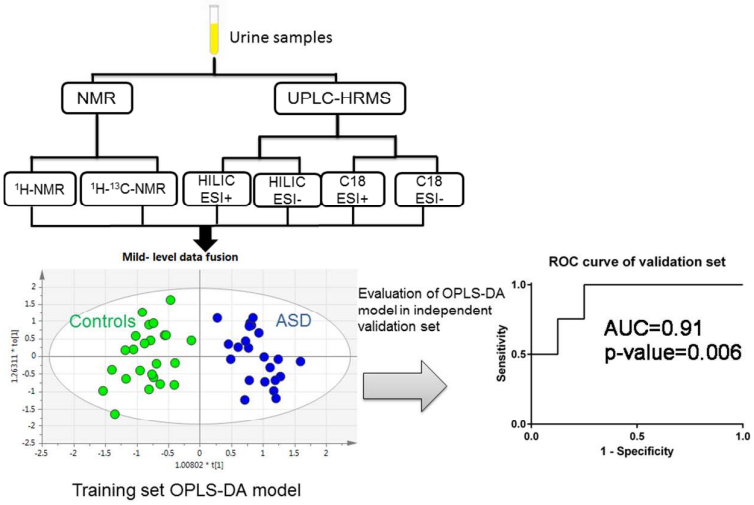

a)
Mid-level data fusion
Controls
ASD
1.00802 * t[1]
b)
Validation set ROC curve
AUC=0.906
p-value=0.006
Sensitivity
1 - Specificity

Figure 2

For TOC only


Urine samples
NMR
UPLC-HRMS
¹H-NMR
¹H-¹³C-NMR
HILIC ESI+
HILIC ESI-
C18 ESI+
C18 ESI-
Mild- level data fusion
Controls
ASD
Training set OPLS-DA model
Evaluation of OPLS-DA model in independent validation set
ROC curve of validation set
AUC=0.91
p-value=0.006
Sensitivity
1 - Specificity