

This document is confidential and is proprietary to the American Chemical Society and its authors. Do not copy or disclose without written permission. If you have received this item in error, notify the sender and delete all copies.

**Statistically-driven Metabolite and Lipid Profiling of Patients
from the Undiagnosed Diseases Network**

Journal:	<i>Analytical Chemistry</i>
Manuscript ID	ac-2019-03522q.R1
Manuscript Type:	Article
Date Submitted by the Author:	30-Oct-2019
Complete List of Authors:	Webb-Robertson, Bobbie-Jo; Pacific Northwest National Laboratory, Computation Biology & Bioinformatics Stratton, Kelly; Pacific Northwest National Laboratory, Applied Statistics and Computational Modeling Kyle, Jennifer; Pacific Northwest National Laboratory Kim, Young-Mo; Pacific Northwest National Laboratory, Biological Sciences Division Bramer, Lisa; Pacific Northwest National Laboratory, Waters, Katrina; Pacific Northwest National Laboratory Koeller, David; Oregon Health & Science University Metz, Thomas; Pacific Northwest National Laboratory,

SCHOLARONE™
Manuscripts

Statistically-driven Metabolite and Lipid Profiling of Patients from the Undiagnosed Diseases Network

Bobbie-Jo M. Webb-Robertson^{†,‡,*}, Kelly G. Stratton^{‡,‡}, Jennifer E. Kyle[†], Young-Mo Kim[†], Lisa M. Bramer[‡], Katrina M. Waters[†], David M. Koeller[‡], Thomas O. Metz^{†,*}

[†]Biological Sciences Division, Pacific Northwest National Laboratory, Richland WA 99352, USA

[‡]Computing Analytics Division, Pacific Northwest National Laboratory, Richland WA 99352, USA

[‡]Molecular and Medical Genetics, School of Medicine, Oregon Health and Science University, Portland, OR 97239, USA

ABSTRACT: Advancements in molecular separations coupled with mass spectrometry have enabled metabolome analyses for clinical cohorts. A population of interest for metabolome profiling are patients with rare disease for which abnormal metabolic signatures may yield clues into the genetic basis, as well as mechanistic drivers of the disease and possible treatment options. We undertook the metabolome profiling of a large cohort of patients with mysterious conditions characterized through the Undiagnosed Diseases Network (UDN). Due to the size and enrollment procedures, collection of the metabolomes for UDN patients took place over two years. We describe the study design to adjust for measurements collected over a long time-scale and how this enabled statistical analyses to summarize the metabolome of individual patients. We demonstrate the removal of time-based batch effects, overall statistical characteristics of the UDN population, and two case-studies of interest that demonstrate the utility of metabolome profiling for rare diseases.

Untargeted metabolomic and lipidomic analyses offer a methodology to capture a global view of the physiologic state of an individual. The metabolome, which includes both metabolites and lipids, is responsive to multiple factors such as genetics, environmental exposures, lifestyle, and disease. Thus, the metabolome is a highly informative source of information to understand the drivers of metabolic diseases¹⁻³. The application and impacts of global metabolome profiling of study populations with common conditions such as type 2 diabetes, heart disease, as well as rare diseases, have demonstrated important findings⁴⁻⁸. Rare and undiagnosed diseases are particularly challenging because in these cases the phenotype being evaluated is typically limited to only one or a few patients, which does not allow for a balanced study design. Gaining insight into underlying mechanisms driving disease in undiagnosed patients can offer clues for diagnosis and care, but requires analyses of single samples, or “n of 1 studies”⁹.

The Undiagnosed Diseases Network (UDN) offers a unique challenge, where multiple biofluids, such as urine and blood plasma, are collected from many patients with mysterious illnesses¹⁰. These patients live with symptoms for which their healthcare providers are unable to identify the cause through standard metabolic screening^{11, 12}. Thus, a global metabolome profiling of these patients may yield new clues to aide healthcare providers in treatment options. However, there are two core challenges that need to be addressed to successfully perform metabolome profiling of the UDN population. The first challenge is scale and time. The metabolome profiling will take place over the course of years as patient samples are collected, which can introduce batch effects due to instrument drift¹³⁻¹⁵. The second issue is inferring metabolome changes in a single patient with a unique phenotype. A healthy normal reference

population of adequate size is needed to enable statistical analysis and identify specific metabolites or lipids of abnormal abundance for an individual patient. Since samples are collected and sent for metabolome profiling dynamically as patients are enrolled, the statistical analysis must be able to yield reliable results across a broad time scale without regeneration of a reference profile.

Here we establish a strategy for collecting and statistically analyzing metabolomics data from UDN samples under the scenario of single sample statistics. We utilize a standard analytical process to address the batch-effect normalization via incorporation of quality control samples (QC)^{13, 14}. Within this context we studied the utilization of different QC sample types to determine if there are significant differences between using a sample pooled from many individual study samples versus a National Institutes of Standards and Technology (NIST) QC sample for plasma-based analyses¹⁶. In addition to the utilization of QC samples to enable normalization, our metabolome processing incorporated a reference of healthy individuals in the first batch of analyses specifically to establish ranges of normal metabolite and lipid abundance values and enable statistical inference of UDN patients. The analytical process employed is a demonstration of a platform for individual patient metabolome profiling from studies when samples are collected ad hoc rather than collected as a single set for experimental analysis. The methodology is utilized successfully on multiple sample types, including plasma, urine and cerebral spinal fluid (CSF).

METHODS

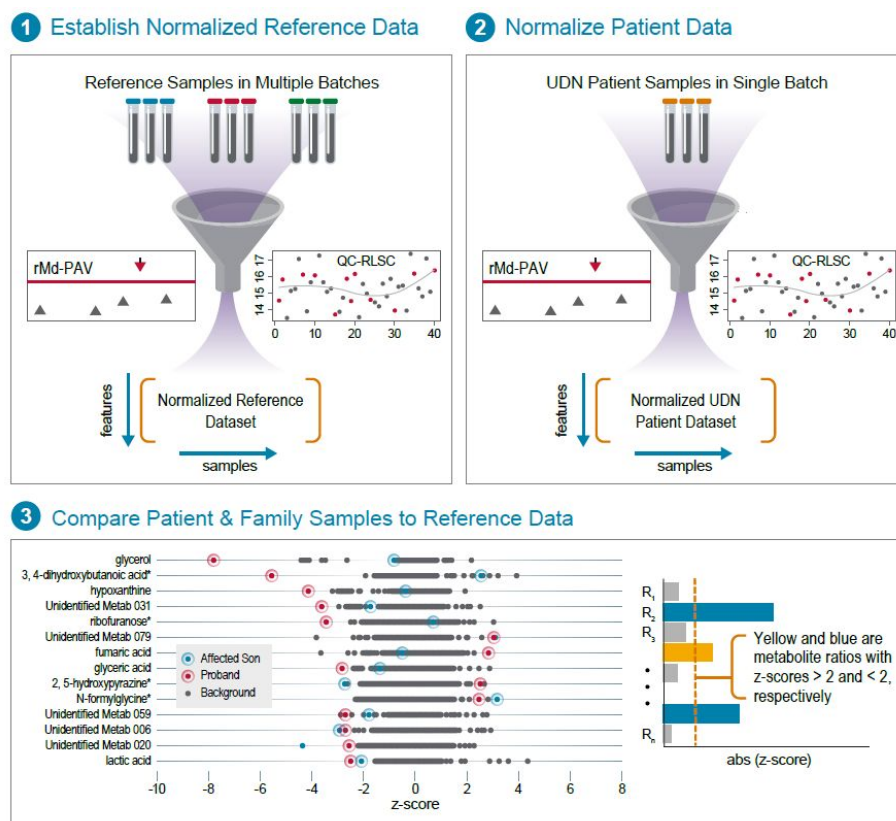


Figure 1: The overall workflow of the UDN sample processing is defined by three core steps. (1) The collection of the normalized reference metabolome dataset for which all probands and associated relatives can be assessed against. (2) Metabolome profiling of a batch of UDN patients utilizing the same quality control samples as the reference dataset to enable normalization. For both (1) and (2) quality control processing to identify outlier samples using robust Mahalanobis distance Panomic Abundance Vectors (rMd-PAV) and Quality Control-based Robust LOESS Signal Correction (QC-RLSC) for normalization are required to attain high quality normalized data for statistical analysis. (3) Statistical comparison of all samples in a batch to the reference data collected in (1). Steps (2) and (3) are repeated as each new UDN batch is collected.

Metabolome profiling of the UDN participants described here was performed on urine metabolites, plasma lipids and CSF lipids. To enable statistical analysis of the UDN participants, two additional sample types were included in the experimental design: 1) QC samples to enable normalization addressing batch effects and 2) a reference normal healthy control background population to identify specific metabolites or lipids that show statistically significant variation, **Figure 1**. There were 281 UDN samples consisting of 148 patients that suffer from unknown diseases (probands) and 133 relatives of the probands. Not all of the 281 participants had adequate collection for all sample types and the total number of QC samples and reference samples varied according to experimental platform and sample types (**Table 1**).

Table 1. Number of analytic runs associated with each sample type

	UDN	Reference	QC
Metabolites	220 (232 ^a)	102	127 (pool)
Plasma Lipids	273 (296 ^a)	136	162 (NIST)
CSF Lipids	6	149	64 (pool)

^aTotal instrument runs, some samples were either evaluated multiple times for data reproducibility or had draws at multiple time points.

UDN Samples

The 281 UDN samples were collected at six different clinical sites within the UDN. Since each sample will be compared to the healthy reference population for statistical analysis all UDN participants were treated as cases as many were identified to have common genetic disease markers as their proband relative. Blood samples for plasma were collected in one 10 ml purple top EDTA Vacutainer® tube that was centrifuged at 2500 RPM (1-1.3 x 10³g) for 10 minutes at 4°C. Three 0.05 ml aliquots were frozen for metabolomics analysis. Urine samples were requested to be the first morning void and were collected in a polypropylene container and centrifuged 100 x g for 5 minutes at 4°C to remove any cells and particulates. Three 0.1 ml aliquots (in Sarstedt Biosphere® SC Micro 0.5ml Tubes) were frozen and saved for metabolomics analysis. CSF was collected by lumbar puncture in the L3/L4 or the L4/L5 inter-space, and three 0.2 ml aliquots were frozen for metabolomics analysis. The samples were shipped to the Pacific

Northwest National Laboratory (PNNL) in dry ice and stored at -70°C . Samples were collected and shipped to PNNL across a broad time period, ranging from September 2016 to October 2018 for metabolomics and lipidomics analyses.

Reference Samples

To enable statistics, a separate set of samples that consisted of similar ethnic, gender and age characteristics to the UDN samples were also evaluated in addition to the 281 UDN samples. The numbers of reference samples were selected based on a power analysis assuming a t-test, a type one error rate of 0.05, a 2-fold detectable change and power set to 0.8 on exemplar data available in-house. Based on the power analysis 102 samples from this reference population were analyzed for the urinary metabolome and 136 for the plasma lipids. Supplemental Figure 1 gives the detailed simulated power curves. Exemplar data was not available for CSF; however, given the power analysis for the urine and plasma the CSF analysis proceeded using all 149 available samples. The data for this reference population in all cases was collected in the first batch of the experiment.

The reference population for adult plasma and urine included samples collected from the Oregon Clinical & Translational Research Institute, Clinical and Translational Research Center, Research Volunteer Registry. The adult CSF samples were provided by the Oregon Health & Science University Layton Aging and Alzheimer's Research Center Biorepository and BioIVT. Juvenile CSF came from the Mayo Clinic, Clinical Biochemical Genetics Laboratory. Juvenile plasma and urine samples were provided by the Vanderbilt University Medical Center. The urine samples were split about 54% female and 47% male with ages ranging from less than 6 months to over the age of 60. The plasma samples were split about 49% female and 51% male with ages ranging from less than 2 years to over the age of 60. Lastly, the CSF was 53% female and 47% male covering the full age range from less than 6 months to over the age of 60.

Quality Control Samples and Batching

As was previously indicated, UDN samples were sent for analyses across approximately two years. Thus, the analyses required data collection in batches. For plasma lipidomics, there were 15 distinct batches of samples for which liquid chromatography tandem mass spectrometry (LC-MS/MS) analyses were performed. For urinary metabolomics, there were 12 batches of samples analyzed by gas chromatography-mass spectrometry (GC-MS). CSF only consisted of one batch due to a limited number of samples. Analytical processing over a long-time window with multiple batches is known to introduce batch effects. A standard approach to manage this issue is to

utilize the same QC samples repeatedly over the analysis to perform normalization to remove these batch effects. All samples were batched and preparation and instrument analysis run orders were randomized based on age, sex, clinical site, and ethnicity (if available) prior to extraction.

QC samples were placed between every 5 to 7 reference or UDN samples^{13, 14}. **Figure 2** gives the overall schema for the utilization of Blanks, QC samples, and a mixture of fatty acid methyl esters (FAMES) during either LC-MS/MS or GC-MS analysis. The blanks minimize carry-over, the FAMES enable retention time alignment between samples and the QC samples allow for QC-RLSC normalization across batches^{13, 14}.

For the urine metabolomics, the QC sample was a pool of the 102 reference samples, and for CSF lipidomics a pool of the 149 reference samples (**Table 1**). For the plasma lipidomics the final QC used was the NIST SRM 1950, which is composed of plasma from 100 healthy individuals between 40-50 years old, of equal numbers of men and women and consists of a race distribution of the US population (QC-NIST). This was selected after comparison with the standard approach, which was a pool of the 136 reference samples (QC-POOL). The QC-NIST is a commercially available reference material (certified until year 2023) and preferred due to the multi-year nature of this study and the limited quantity of reference sample plasma. For urine and CSF, as no similar reference material was commercially available, only the pooled QC samples from the reference population were used.

Metabolome Generation

Single aliquots of plasma (50 μL), urine (100 μL), and CSF (200 μL) were thawed and extracted for analysis by MS. Metabolites and lipids were extracted from the same samples using the MPLEX method¹⁷. To the urine 50 μL of GC-MS internal standards (malonic acid-d4, fructose 13C6, L-tryptophan-d5, lysine-d4, alanine-d7, stearic acid-d35, benzoic acid-d5, octanoic acid-d15, octanoic acid-d15 at a final concentration of 1 $\mu\text{g}/\mu\text{L}$ each) were added. For the plasma and CSF 10 μL of LC-MS internal standards (SPLASHTM Lipidomix[®] Mass Spec Standard, Avanti Polar Lipids, Inc.) were added. For CSF lipids, 10 μL of LC-MS internal standards ((PC(17:0/14:1) at 1 $\mu\text{g}/\mu\text{L}$, LPC(19:0) at 0.01 $\mu\text{g}/\mu\text{L}$, and TG(17:0/17:1/17:0)-d5 at 0.01 $\mu\text{g}/\mu\text{L}$) were added.

Extracted metabolites were dried in vacuo and stored at -20°C until analysis. Chemical derivatization of metabolites followed previously published standard approaches¹⁸. Metabolites were analyzed with an Agilent 7890A GC coupled with a single quadrupole Mass Selective Detector 5975C. The GC-MS analysis was described previously¹⁸. Metabolite identification was performed using the Metabolite Detector software (Version 2.0.6 beta)¹⁹.

Extracted lipids were stored at -20°C in 2:1 chloroform/methanol until LC-tandem MS (MS/MS) analysis. Prior to analysis, total lipid extracts (TLEs) were again dried in vacuo, and then plasma TLEs were reconstituted in 200 μL of methanol²⁰ and CSF TLEs in 50 μL of methanol. LC-electrospray ionization (ESI)-MS/MS analysis was performed with a Waters Acquity Ultra Performance LC H class system interfaced with a Velos-Electron Transfer Dissociation (ETD) Orbitrap mass spectrometer, operated in both positive and negative mode. Lipids were identified using the Lipid Quantification and Identification (LIQUID)²¹ software. All lipid identifications made using LIQUID were verified by manual evaluation of MS/MS spectra. Lipids data from both the

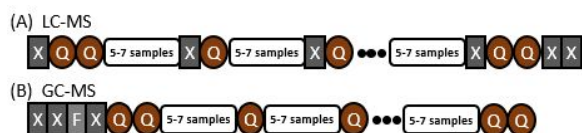


Figure 2: The structure of blanks (X), FAMES (F) and QC (Q) samples distribution within a batch for (A) LC-MS analysis of the plasma and CSF lipids and (B) urine metabolomics. The QC sample placement is representative of the QC sample placement used for normalization for each sample type. Under the pilot project comparing the pooled and NIST-based QC samples they were placed adjacent to one another in the design.

positive and negative mode are processed separately through subsequent statistical analysis.

Data Analysis

Normalization was performed on a per-batch basis using Quality Control Robust Loess Signal Correction (QC-RLSC)^{13, 14}. In this approach, a LOESS curve of first or second degree is fitted to the QC samples with respect to the order of injection for one feature at a time. The LOESS curve is fitted using weighted least squares and smoothing and polynomial degree parameters optimized with leave-one-out cross-validation. A feature-specific normalization curve for the entire batch is then interpolated using cubic splines from the fitted LOESS model and data for the respective feature (lipid or metabolite) is normalized to the estimated LOESS model.

From the normalized data for each metabolite or lipid feature the statistical significance for a UDN proband or relative was computed as a standard *z-score*²² relative to the reference population. That is, the mean abundance for a specific metabolite or lipid for the reference samples was subtracted from the observed abundance for the UDN sample and lastly divided by the standard deviation of the reference samples. If a metabolite or lipid feature does not have adequate coverage it cannot be analyzed because each UDN proband and relative are being evaluated singly against the reference population. The presence or absence from one or the other group does not offer adequate information to assign statistical significance. Thus, all features that are not observed in at least 10% of the samples are removed. For the urine metabolomics this reduced the number of metabolites from 128 to 116 and from 457 to 402 and 210 to 152 for the plasma and CSF lipids, respectively. For the remainder of the features that included missing values these were ignored in the computation of the parameters from the reference population to compute the *z-score*. The percent of missing data ranged for each dataset from about 4 to 8 percent.

RESULTS AND DISCUSSION

The metabolome profiling of the UDN subjects required a design to adjust for measurements collected over a long time-scale and enable statistical analyses to summarize the metabolome of an individual patient. We present results on the removal of time-based batch effects, statistical summaries of the UDN population, and several case-studies that demonstrate the impact of the metabolite and lipid profiles on several exemplar probands from the UDN.

Establishing QC for Reference Baseline

The choice of the QC sample for normalization purposes can affect the ability to both reduce variability and to profile the full identifiable metabolome¹⁴. In addition, the availability of the same QC sample throughout the entire experiment is necessary and ideally is a pool of the samples being analyzed. The combination of a large experiment and a sample of limited quantity, such as plasma, makes this a challenge and therefore the identification of a commercially available QC, such as the QC-NIST, is attractive. To determine the potential effect of the QC sample on the UDN data processing, two types of QC samples were analyzed in the initial batch of reference samples; QC-NIST and QC-POOL. In this first batch, 50 QC samples of each type were processed according to the schema in **Figure 2** where each (Q) in the figure has a QC-POOL and a QC-NIST adjacent for this first batch.

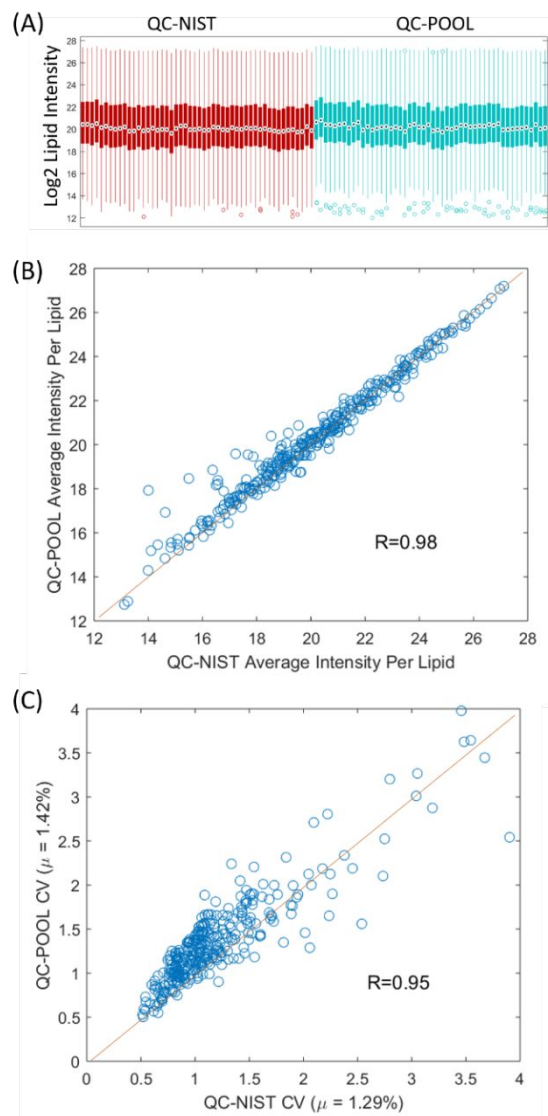


Figure 3: The first batch of samples included both QC-NIST and QC-POOL samples that showed similar (A) abundance profiles visually, (B) average abundance per lipid, and (C) CVs per lipid. For the abundance and CV similarity is shown as the positive coefficient of correlation (R).

Figure 3A shows the overall distribution of each of the 50 QC-NIST and QC-POOL samples as a boxplot for the positive mode lipidomics data as returned from the first batch. The similarity of these two QCs was of interest. As previously mentioned, the QC-NIST is preferable because there is not enough plasma from individual samples to generate the QC-POOL, which is the standard approach, for the full experimental design. From **Figure 3A**, they have very similar overall abundance profiles of the 342 lipids represented in this data. To determine if there was an appreciable difference between the QC types, we compared the average intensity of each lipid across the QC samples of each type. As seen in **Figure 3B** the abundances estimated for the two types of QCs are highly correlated, with a Pearson correlation of > 0.98 . The QC-POOL samples are slightly higher in average abundance, but they also have higher variability in relation to the mean as measured by the coefficient of variation (CV), **Figure 3C**. The CVs for the

QC-POOLS are statistically larger ($p < 2.4E-15$; paired t-test); however, the median CVs are small, 1.06% and 1.26% for the QC-NIST and QC-POOL samples, respectively. The maximum CV was also similar, 11.16% and 10.52%, respectively, for the QC-NIST and QC-POOL samples. The negative mode lipidomics data yielded similar results, with a correlation of 0.95 for the mean intensity and 0.96 for the CVs. Again, the average CVs for the QC-NIST are significantly smaller ($p < 5.8E-4$), with a median of 2.23% and 2.29% for the QC-NIST and QC-POOL, respectively. The lower CV values for the NIST could be due to many factors, such as sample preparation or the underlying sample population. However, given the similarity of the abundance and variability profiles the QC-NIST samples are utilized for the full plasma-based lipid analysis.

Normalization

A major challenge with multi-batch metabolomics studies is data normalization¹³⁻¹⁵. The collection of the metabolomics and lipidomics data for all UDN samples spanned nearly two years. The samples were processed in batches (B), but the QC-RLSC normalization approach allowed data from UDN samples to be processed throughout the study rather than only at the end. The plasma data were collected in 15 batches, denoted as B1-B15 in **Figure 4**. The urine samples were analyzed over 12 batches, and the CSF data was limited to only 1. All data were subjected to sample-level quality assurance²³ processing prior to normalization. Any sample identified as having a potential

quality issue was followed up with manual inspection. For both the lipid and metabolite measurements, QC-RLSC-based normalization¹³ was utilized to adjust for instrument drift. Subsequently total abundance adjustments were made with standard median based centering within each sample.

Figure 4 shows an example of the pre- and post-normalization batch effects visualized through a Principal Components Analysis (PCA) plot of the plasma log₂-transformed lipidomics data collected in negative mode. Similar effects are observed for the lipidomics data collected in positive mode and the urinary metabolomics data. Multiple CSF projects were not possible; however, the inclusion of the QC samples allows later evaluation and comparison of more CSF samples. In **Figure 4** there is a clear separation on the first and second principal components based on when the samples were processed. **Figure 4A** and **Figure 4B** show the PCA plot of the QC samples pre- and post- normalization. As expected after normalization the QC samples are in a tight cluster in the center of the PCA plot. **Figure 4C** and **Figure 4D** show the pre- and post-normalization for the UDN samples with the reference samples used to compute the z-score highlighted as well. Again prior to normalization there is a clear separation by time and batch effects are visible as well as the reference samples clustering with the first batch of UDN samples, with which they were collected. After normalization there is no clear visible batch-effect observed and the reference samples now separate

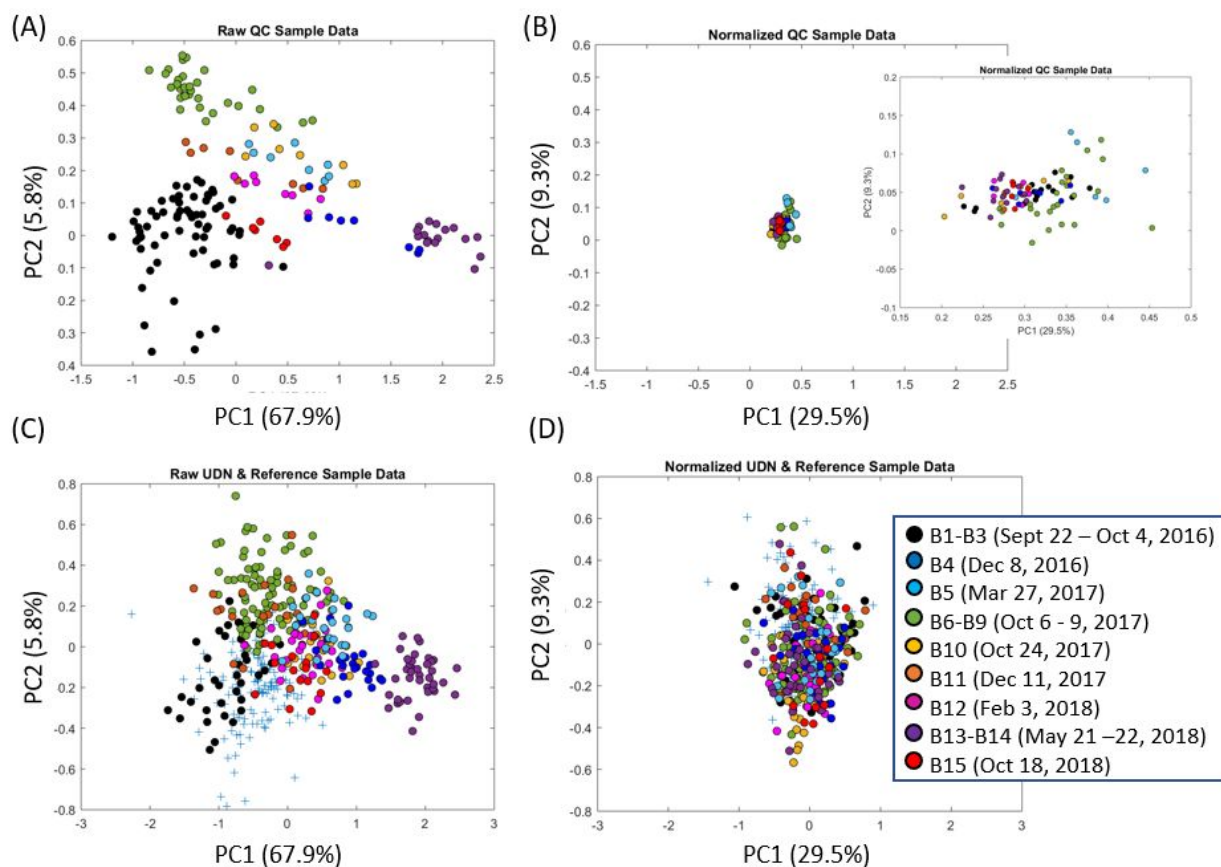


Figure 4: Pre- and post-normalization PCA plot shows dramatic decrease in batch effects on the (A) QC samples before and (B) after normalization. The UDN samples show a similar batch effect in (C) prior to normalization, which is no longer visible after normalization in (D). Also shows in the the reference population (+), which clusters with the first batch in (C), but is dispersed across all samples after normalization in (D) removing the visible batch effect.

across the entire space rather than clustering with the initial batch.

To evaluate the removal of the batch effect a standard Analysis of Variance (ANOVA) was performed for each of the plasma lipids and urine metabolites to test the hypothesis of equal abundance of that molecule in the QC samples across the projects. The p-values were highly skewed to the very small side, especially in the raw data prior to normalization; thus the visualization of the p-value results again from the lipidomics data (Figure 5) is on the log10 scale. The log scale improved the overall distributional properties of the p-values, but they still are not normally distributed and therefore a non-parametric sign-rank test was utilized to determine if the p-values for the normalized data were statistically smaller than the raw data. From Figure 5 it is clear the Raw data had significant batch effect compared to the normalized data and over 92% of the markers showed a decrease in bias for the normalized data.

UDN Sample Analyses

The lipidomics data included results from one UDN proband that was analyzed 9 times across three distinct batches. In addition, the mother of this proband was also analyzed multiple times (4 total) across two batches. This offered the opportunity to evaluate intra-subject and intra-family variability. The Pearson correlation of the 9 biological replicates of the proband sample to one another were computed, as well as the correlation of the 9 replicates with the mother's four samples. Finally, each of the 9 replicate samples were also correlated to five randomly-chosen samples from the entire UDN cohort. These three sets of correlations were compared using ANOVA ($p < 3.4E-18$) and a Tukey's post-hoc test was employed to determine which of the groups of correlations were statistically significant from

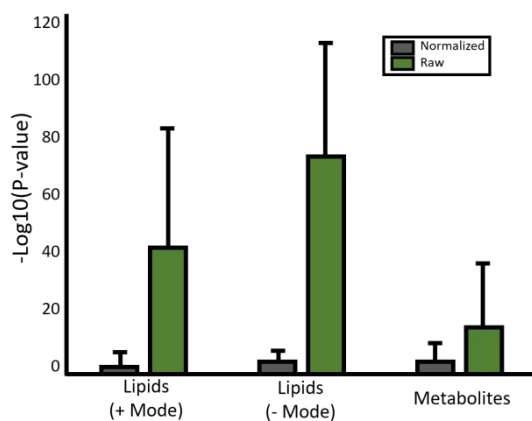


Figure 5: P-values testing for project effects within the lipids and metabolite data show a clear increase in p-values.

one another. Figure 6 demonstrates that the intra-subject correlation is significantly higher than both the intra-familial (proband-mother) correlation, as well as the correlation with random samples (all significant at $p < 0.05$).

UDN Case Study Insights

Statistics were computed for all of the UDN patients by calculating a z-score as the difference between the observed abundance for an individual lipid or metabolite to the reference mean divided by the reference standard deviation²². This

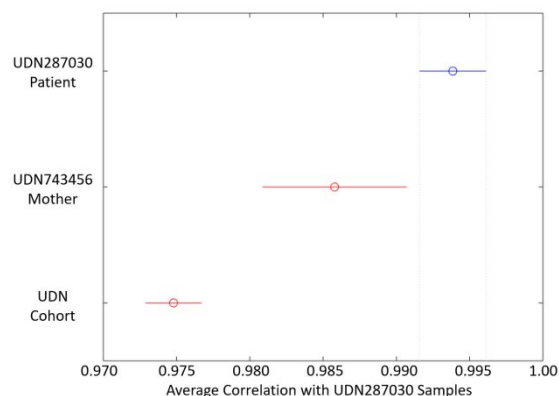


Figure 6: Tukey's post-hoc test demonstrates that the correlation between replicates from the same subject are significantly larger than intra-family and random sample relationships.

generated thousands of z-scores, one per UDN subject/marker. The majority, over 90% for all three datasets, are not significant ($|z\text{-score}| < 2$). This is expected, since a large majority significant would indicate a bias comparing back to the reference data generated at the start of the study. However, the extreme z-scores can identify metabolites or lipids of potential relevance to the patient's symptoms and/or diagnosis that can be further explored for individual patients

The first UDN case we highlight was a 17-year-old Caucasian female with symptoms that included a peripheral neuropathy, retinopathy, and muscle weakness. Urine metabolomics analysis identified an isolated, extreme elevation of urine lysine (un-corrected z-score ~ 4.5 , p-value $< 6.7E-6$). Figure 7 shows the distribution of z-scores for urine lysine level for all UDN subjects based on a Benjamini-Hochberg correction of the p-values²⁴. This proband's urine lysine level (black bar) was higher than all other UDN subjects, representing a q-value of less than 0.01. This finding supported the suspected diagnosis of 2,4-dienoyl coenzyme A reductase (DECR) deficiency, which is known to be associated with elevation of urine lysine^{25, 26}, which was then confirmed via

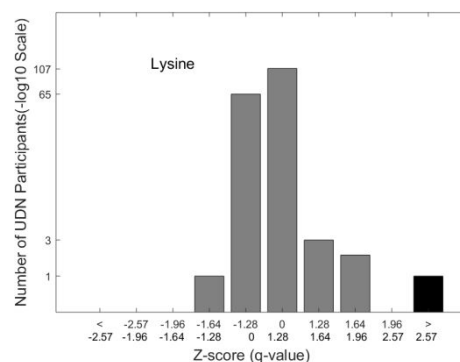


Figure 7: Bar graphs showing the number of patients in each z-score bin for the metabolite lysine where the bin represents the number of subjects with the defined z-score range. The z-scores $\pm 1.28, \pm 1.64, \pm 1.96, \pm 2.57$ correspond approximately to q-values of 0.2, 0.1, 0.05 and 0.01, respectively. Only one UDN patient (black bar) has an extreme elevated amount of lysine in comparison to the full population.

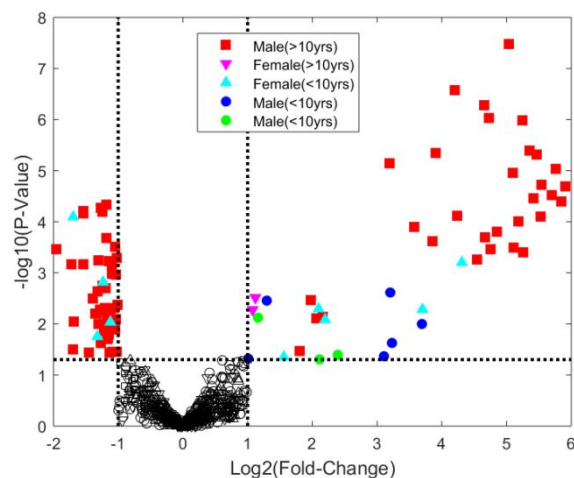


Figure 8: Volcano plot of CSF lipids from 5 UDN probands. Only lipids with both a significant change (p -value < 0.05) and a large fold-change ($\geq \pm 2$ -fold relative to the reference data) are shown.

whole genome sequencing, which identified two mutations in the Mitochondrial NAD kinase 2 (*NADK2*) gene²⁷. DECR deficiency results from mutations in the *NADK2* gene and is inherited as an autosomal recessive disorder.

Our second example was a UDN proband who was found to have a unique profile of CSF lipids, with over 8-times more significantly increased or decreased lipid species than any other UDN subject (**Figure 8**, red squares). This included a generalized decrease in diacyl-phosphoglycerocholines (PCs), alkyl-PCs (PCO), and sphingomyelins (SMs), and a surprising increase in triglycerides (TGs). A corresponding plasma signature of these lipids was not observed, highlighting the uniqueness of the CSF lipid signature. The proband's CSF was collected while fasted, and symptomatic with tremors and dysarthria. Clinical MRI spectroscopy noted a decrease in N-acetylaspartate and an increase in choline, consistent with a loss of normal neural tissue. A specific evaluation of brain lipids was not performed. PC and SM lipids have been noted to be associated with neurological disease markers identified in CSF²⁸⁻³⁰. This finding along with the high amount of TGs, which are rarely identified in CSF, suggest a potential problem with the blood brain barrier; however, further analysis will be required to confirm this hypothesis and define a medical designation.

CONCLUSIONS

The presence of metabolic outliers may provide significant clues regarding the pathophysiology and etiology of rare or undiagnosed diseases. Metabolome profiling presents a possible solution to the identification of these changes in a rapid and robust manner. We have demonstrated a protocol for evaluating such patients in an individual manner via the inclusion of both QC samples to manage batch effect issues and a reference population for identification of significant changes for individual metabolites and lipids within a single patient. We demonstrated both the overall analytical strategy as well as two case studies that returned interesting findings, enabling clinicians to hypothesize potential mechanisms associated with the rare disease

AUTHOR INFORMATION

Corresponding Authors

*E-mail for B.M.W.; bj@pnnl.gov

*E-mail for T.O.M.; thomas.metz@pnnl.gov

Author Contributions

[†] These authors contributed equally.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

We thank the Undiagnosed Diseases Network investigators and are grateful for the participation of the patients and family members and their referring clinicians. This work was funded by the Undiagnosed Diseases Network (1U01TR001395) supported by the National Institutes of Health (NIH) Common Fund. The authors thank the following individuals, institutions and funding sources for providing or making available the reference population samples: Dr. Mary Samuels at Oregon Health Science University (via the Oregon Clinical and Translational Research Institute, supported by the National Center for Advancing Translational Sciences of the NIH under award number UL1TR0000128); Dr. Joseph Quinn at Oregon Health Science University (via the Oregon Alzheimer's Disease Center Biorepository, supported by grant number NIA-AG008017 from the NIH National Institute on Aging); Dr. Rizwan Hamid at Samples were courtesy of courtesy of Dr. Rizwan Hamid at the Vanderbilt University Medical Center; and Dr. Devin Oglesbee at Mayo Clinic (with support by Mayo Clinic's Department of Laboratory Medicine and Pathology). Mass spectrometry analyses were performed in the Environmental Molecular Sciences Laboratory, a national scientific user facility sponsored by the Department of Energy (DOE) Office of Biological and Environmental Research and located at Pacific Northwest National Laboratory (PNNL). We also thank Mr. Michael Perkins at PNNL for graphic design. PNNL is operated by Battelle Memorial Institute for the DOE under contract DEAC05-76RLO1830.

REFERENCES

- Haijes, H. A.; Willemsen, M.; Van der Ham, M.; Gerrits, J.; Pras-Raves, M. L.; Prinsen, H.; Van Hasselt, P. M.; De Sain-van der Velden, M. G. M.; Verhoeven-Duif, N. M.; Jans, J. J. M., Direct Infusion Based Metabolomics Identifies Metabolic Disease in Patients' Dried Blood Spots and Plasma. *Metabolites* **2019**, *9* (1).
- Vinayavekhin, N.; Homan, E. A.; Saghatelian, A., Exploring disease through metabolomics. *ACS Chem Biol* **2010**, *5* (1), 91-103.
- Yan, M.; Xu, G., Current and future perspectives of functional metabolomics in disease studies-A review. *Anal Chim Acta* **2018**, *1037*, 41-54.
- Cheng, S.; Shah, S. H.; Corwin, E. J.; Fiehn, O.; Fitzgerald, R. L.; Gerszten, R. E.; Illig, T.; Rhee, E. P.; Srinivas, P. R.; Wang, T. J.; Jain, M.; American Heart Association Council on Functional, G.; Translational, B.; Council on, C.; Stroke, N.; Council on Clinical, C.; Stroke, C., Potential Impact and Study Considerations of Metabolomics in Cardiovascular Health and Disease: A Scientific Statement From the American Heart Association. *Circ Cardiovasc Genet* **2017**, *10* (2).
- Liu, J.; Semiz, S.; van der Lee, S. J.; van der Spek, A.; Verhoeven, A.; van Klinken, J. B.; Sijbrands, E.; Harms, A. C.; Hankemeier, T.; van Dijk, K. W.; van Duijn, C. M.; Demirkan,

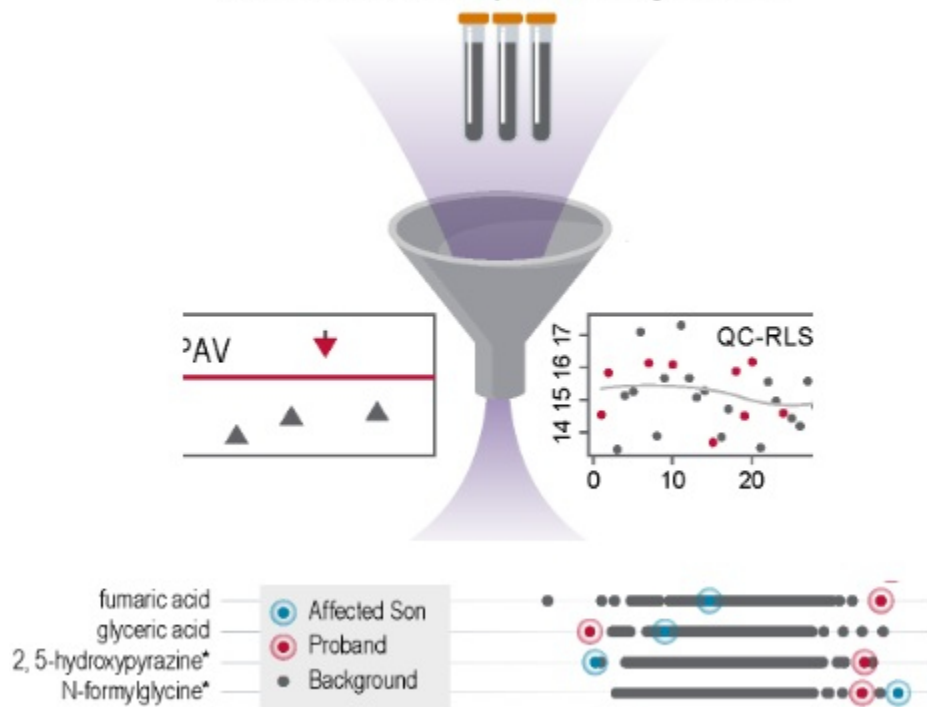
- A., Metabolomics based markers predict type 2 diabetes in a 14-year follow-up study. *Metabolomics* **2017**, *13* (9), 104.
6. Graham, E.; Lee, J.; Price, M.; Tarailo-Graovac, M.; Matthews, A.; Engelke, U.; Tang, J.; Kluijtmans, L. A. J.; Wevers, R. A.; Wasserman, W. W.; van Karnebeek, C. D. M.; Mostafavi, S., Integration of genomics and metabolomics for prioritization of rare disease variants: a 2018 literature review. *J Inherit Metab Dis* **2018**, *41* (3), 435-445.
7. Mussap, M.; Zaffanello, M.; Fanos, V., Metabolomics: a challenge for detecting and monitoring inborn errors of metabolism. *Ann Transl Med* **2018**, *6* (17), 338.
8. Pampols, T., Inherited metabolic rare disease. *Adv Exp Med Biol* **2010**, *686*, 397-431.
9. Lillie, E. O.; Patay, B.; Diamant, J.; Issell, B.; Topol, E. J.; Schork, N. J., The n-of-1 clinical trial: the ultimate strategy for individualizing medicine? *Per Med* **2011**, *8* (2), 161-173.
10. Ramoni, R. B.; Mulvihill, J. J.; Adams, D. R.; Allard, P.; Ashley, E. A.; Bernstein, J. A.; Gahl, W. A.; Hamid, R.; Loscalzo, J.; McCray, A. T.; Shashi, V.; Tift, C. J.; Undiagnosed Diseases, N.; Wise, A. L., The Undiagnosed Diseases Network: Accelerating Discovery about Health and Disease. *Am J Hum Genet* **2017**, *100* (2), 185-192.
11. Bower, A.; Imbard, A.; Benoist, J. F.; Pichard, S.; Rigal, O.; Baud, O.; Schiff, M., Diagnostic contribution of metabolic workup for neonatal inherited metabolic disorders in the absence of expanded newborn screening. *Sci Rep* **2019**, *9* (1), 14098.
12. Villoria, J. G.; Pajares, S.; Lopez, R. M.; Marin, J. L.; Ribes, A., Neonatal Screening for Inherited Metabolic Diseases in 2016. *Semin Pediatr Neurol* **2016**, *23* (4), 257-272.
13. Dunn, W. B.; Broadhurst, D.; Begley, P.; Zelena, E.; Francis-McIntyre, S.; Anderson, N.; Brown, M.; Knowles, J. D.; Halsall, A.; Haselden, J. N.; Nicholls, A. W.; Wilson, I. D.; Kell, D. B.; Goodacre, R.; Human Serum Metabolome, C., 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), 1060-83.
14. Dunn, W. B.; Wilson, I. D.; Nicholls, A. W.; Broadhurst, D., The importance of experimental design and QC samples in large-scale and MS-driven untargeted metabolomic studies of humans. *Bioanalysis* **2012**, *4* (18), 2249-64.
15. Fan, S.; Kind, T.; Cajka, T.; Hazen, S. L.; Tang, W. H. W.; Kaddurah-Daouk, R.; Irvin, M. R.; Arnett, D. K.; Barupal, D. K.; Fiehn, O., Systematic Error Removal Using Random Forest for Normalizing Large-Scale Untargeted Lipidomics Data. *Anal Chem* **2019**, *91* (5), 3590-3596.
16. Simon-Manso, Y.; Lowenthal, M. S.; Kilpatrick, L. E.; Sampson, M. L.; Telu, K. H.; Rudnick, P. A.; Mallard, W. G.; Bearden, D. W.; Schock, T. B.; Tchekhovskoi, D. V.; Blonder, N.; Yan, X.; Liang, Y.; Zheng, Y.; Wallace, W. E.; Neta, P.; Phinney, K. W.; Remaley, A. T.; Stein, S. E., Metabolite profiling of a NIST Standard Reference Material for human plasma (SRM 1950): GC-MS, LC-MS, NMR, and clinical laboratory analyses, libraries, and web-based resources. *Anal Chem* **2013**, *85* (24), 11725-31.
17. Nakayasu, E. S.; Nicora, C. D.; Sims, A. C.; Burnum-Johnson, K. E.; Kim, Y. M.; Kyle, J. E.; Matzke, M. M.; Shukla, A. K.; Chu, R. K.; Schepmoes, A. A.; Jacobs, J. M.; Baric, R. S.; Webb-Robertson, B. J.; Smith, R. D.; Metz, T. O., MPLEx: a Robust and Universal Protocol for Single-Sample Integrative Proteomic, Metabolomic, and Lipidomic Analyses. *mSystems* **2016**, *1* (3).
18. Hallaian, K. A.; Zhang, Q.; Madupu, R.; Waters, K. M.; Metz, T. O., A Statistical Analysis of the Effects of Urease Pre-treatment on the Measurement of the Urinary Metabolome by Gas Chromatography-Mass Spectrometry. *Metabolomics* **2014**, *10* (5), 897-908.
19. Hiller, K.; Hangebrauk, J.; Jager, C.; Spura, J.; Schreiber, K.; Schomburg, D., MetaboliteDetector: comprehensive analysis tool for targeted and nontargeted GC/MS based metabolome analysis. *Anal Chem* **2009**, *81* (9), 3429-39.
20. Kyle, J. E.; Casey, C. P.; Stratton, K. G.; Zink, E. M.; Kim, Y. M.; Zheng, X.; Monroe, M. E.; Weitz, K. K.; Bloodworth, K. J.; Orton, D. J.; Ibrahim, Y. M.; Moore, R. J.; Lee, C. G.; Pedersen, C.; Orwoll, E.; Smith, R. D.; Burnum-Johnson, K. E.; Baker, E. S., Comparing identified and statistically significant lipids and polar metabolites in 15-year old serum and dried blood spot samples for longitudinal studies. *Rapid Commun Mass Spectrom* **2017**, *31* (5), 447-456.
21. Kyle, J. E.; Crowell, K. L.; Casey, C. P.; Fujimoto, G. M.; Kim, S.; Dautel, S. E.; Smith, R. D.; Payne, S. H.; Metz, T. O., LIQUID: an open source software for identifying lipids in LC-MS/MS-based lipidomics data. *Bioinformatics* **2017**, *33* (11), 1744-1746.
22. Miller, M. J.; Kennedy, A. D.; Eckhart, A. D.; Burrage, L. C.; Wulff, J. E.; Miller, L. A.; Milburn, M. V.; Ryals, J. A.; Beaudet, A. L.; Sun, Q.; Sutton, V. R.; Elsea, S. H., Untargeted metabolomic analysis for the clinical screening of inborn errors of metabolism. *J Inherit Metab Dis* **2015**, *38* (6), 1029-39.
23. Matzke, M. M.; Waters, K. M.; Metz, T. O.; Jacobs, J. M.; Sims, A. C.; Baric, R. S.; Pounds, J. G.; Webb-Robertson, B. J., Improved quality control processing of peptide-centric LC-MS proteomics data. *Bioinformatics* **2011**, *27* (20), 2866-72.
24. Noble, W. S., How does multiple testing correction work? *Nat Biotechnol* **2009**, *27* (12), 1135-7.
25. Houten, S. M.; Denis, S.; Te Brinke, H.; Jongejan, A.; van Kampen, A. H.; Bradley, E. J.; Baas, F.; Hennekam, R. C.; Millington, D. S.; Young, S. P.; Frazier, D. M.; Gucsavas-Calikoglu, M.; Wanders, R. J., Mitochondrial NAD(P)H deficiency due to a mutation in NADK2 causes dienoyl-CoA reductase deficiency with hyperlysinemia. *Hum Mol Genet* **2014**, *23* (18), 5009-16.
26. Splinter, K.; Adams, D. R.; Bacino, C. A.; Bellen, H. J.; Bernstein, J. A.; Cheate-Jarvela, A. M.; Eng, C. M.; Esteves, C.; Gahl, W. A.; Hamid, R.; Jacob, H. J.; Kikani, B.; Koeller, D. M.; Kohane, I. S.; Lee, B. H.; Loscalzo, J.; Luo, X.; McCray, A. T.; Metz, T. O.; Mulvihill, J. J.; Nelson, S. F.; Palmer, C. G. S.; Phillips, J. A., 3rd; Pick, L.; Postlethwait, J. H.; Reuter, C.; Shashi, V.; Sweetser, D. A.; Tift, C. J.; Walley, N. M.; Wangler, M. F.; Westerfield, M.; Wheeler, M. T.; Wise, A. L.; Worthey, E. A.; Yamamoto, S.; Ashley, E. A.; Undiagnosed Diseases, N., Effect of Genetic Diagnosis on Patients with Previously Undiagnosed Disease. *N Engl J Med* **2018**, *379* (22), 2131-2139.
27. Pomerantz, D. J.; Ferdinandusse, S.; Cogan, J.; Cooper, D. N.; Reimschisel, T.; Robertson, A.; Bican, A.; McGregor, T.; Gauthier, J.; Millington, D. S.; Andrae, J. L. W.; Tschannen, M. R.; Helbling, D. C.; Demos, W. M.; Denis, S.; Wanders, R. J. A.; Newman, J. N.; Hamid, R.; Phillips, J. A., 3rd; Collaborators of, U. D. N., Clinical heterogeneity of mitochondrial NAD kinase deficiency caused by a NADK2 start loss variant. *Am J Med Genet A* **2018**, *176* (3), 692-698.
28. Capodivento, G.; Visigalli, D.; Garnerio, M.; Fancellu, R.; Ferrara, M. D.; Basit, A.; Hamid, Z.; Pastore, V. P.; Garibaldi, S.; Armirotti, A.; Mancardi, G.; Serrati, C.; Capello, E.; Schenone, A.; Nobbio, L., Sphingomyelin as a myelin biomarker in CSF of acquired demyelinating neuropathies. *Sci Rep* **2017**, *7* (1), 7831.
29. Fonteh, A. N.; Ormseth, C.; Chiang, J.; Cipolla, M.; Arakaki, X.; Harrington, M. G., Sphingolipid metabolism

1 correlates with cerebrospinal fluid Beta amyloid levels in
2 Alzheimer's disease. *PLoS One* **2015**, *10* (5), e0125597.
3 30. Mulder, C.; Wahlund, L. O.; Teerlink, T.; Blomberg,
4 M.; Veerhuis, R.; van Kamp, G. J.; Scheltens, P.; Scheffer, P.
5 G., Decreased lysophosphatidylcholine/phosphatidylcholine ratio

in cerebrospinal fluid in Alzheimer's disease. *J Neural Transm*
(*Vienna*) **2003**, *110* (8), 949-55.

6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

UDN Patient Samples in Single Batch



TOC