Title
A resource of lipidomics and metabolomics data from individuals with undiagnosed diseases.

Journal
Scientific data, 8(1)

ISSN
2052-4463

Authors
Kyle, Jennifer E
Stratton, Kelly G
Zink, Erika M
et al.

Publication Date
2021-04-01

DOI
10.1038/s41597-021-00894-y

Peer reviewed

SCIENTIFIC DATA

Check for updates

OPEN

DATA DESCRIPTOR

A resource of lipidomics and metabolomics data from individuals with undiagnosed diseases

Jennifer E. Kyle[1], Kelly G. Stratton[2], Erika M. Zink[1], Young-Mo Kim[1], Kent J. Bloodsworth[1], Matthew E. Monroe[1], Undiagnosed Diseases Network*, Katrina M. Waters[1], Bobbie-Jo M. Webb-Robertson[1], David M. Koeller[3] ✉ & Thomas O. Metz[1] ✉

Every year individuals experience symptoms that remain undiagnosed by healthcare providers. In the United States, these rare diseases are defined as a condition that affects fewer than 200,000 individuals. However, there are an estimated 7000 rare diseases, and there are an estimated 25–30 million Americans in total (7.6–9.2% of the population as of 2018) affected by such disorders. The NIH Common Fund Undiagnosed Diseases Network (UDN) seeks to provide diagnoses for individuals with undiagnosed disease. Mass spectrometry-based metabolomics and lipidomics analyses could advance the collective understanding of individual symptoms and advance diagnoses for individuals with heretofore undiagnosed disease. Here, we report the mass spectrometry-based metabolomics and lipidomics analyses of blood plasma, urine, and cerebrospinal fluid from 148 patients within the UDN and their families, as well as from a reference population of over 100 individuals with no known metabolic diseases. The raw and processed data are available to the research community so that they might be useful in the diagnoses of current or future patients suffering from undiagnosed disorders.

Background & Summary

Metabolites and lipids can be responsive to both genetic and environmental influences. Variations may occur due to host genes, disease states, lifestyle, diet, medications and the interaction with the gut microbiome[1]. Many rare diseases have genetic origins, but their symptoms can also be impacted by non-inherited causes such as infections, cancers, and other acquired conditions. Metabolomics and lipidomics analyses have been helpful in identifying inborn errors of metabolism, and in characterizing acquired metabolic conditions such as diabetes and metabolic syndrome[2,3]. These conditions are typically associated with a small number of metabolites and/or lipids that are significant outliers, and easily identified as abnormal.

In contrast, the metabolic changes in rare and undiagnosed diseases may be more subtle, consisting of complex patterns of minor changes of a large number of analytes rather than a few significant outliers. Due to the rare nature of these disorders, the number of individuals with a given phenotype is usually limited to one or just a few, precluding the use of the balanced study designs typically used in metabolomics. For these reasons the use of metabolomics and lipidomics analyses in the evaluation of rare and undiagnosed diseases presents many unique challenges.

The NIH Common Fund's Undiagnosed Diseases Network (UDN) was established to accelerate the diagnosis and clinical management of rare or previously unrecognized diseases, and to advance research in disease mechanisms[4]. The UDN is composed of multiple clinical sites around the United States, and multiple research cores including DNA sequencing (whole exome and whole genome), model organisms (e.g., drosophila and zebrafish) and metabolomics[4,5]. As the Metabolomics Core for Phase I of the UDN, our role was to provide comprehensive untargeted measurements to identify qualitative and quantitative changes of metabolites (metabolomics) and

[1]Biological Sciences Division, Earth and Biological Sciences Directorate, Pacific Northwest National Laboratory, Richland, WA, 99352, USA. [2]Computing and Analytics Division, National Security Directorate, Pacific Northwest National Laboratory, Richland, WA, 99352, USA. [3]Molecular and Medical Genetics, School of Medicine, Oregon Health and Science University, Portland, OR, 97239, USA. *A list of authors and their affiliations appears at the end of the paper. ✉e-mail: koellerd@ohsu.edu; thomas.metz@pnnl.gov

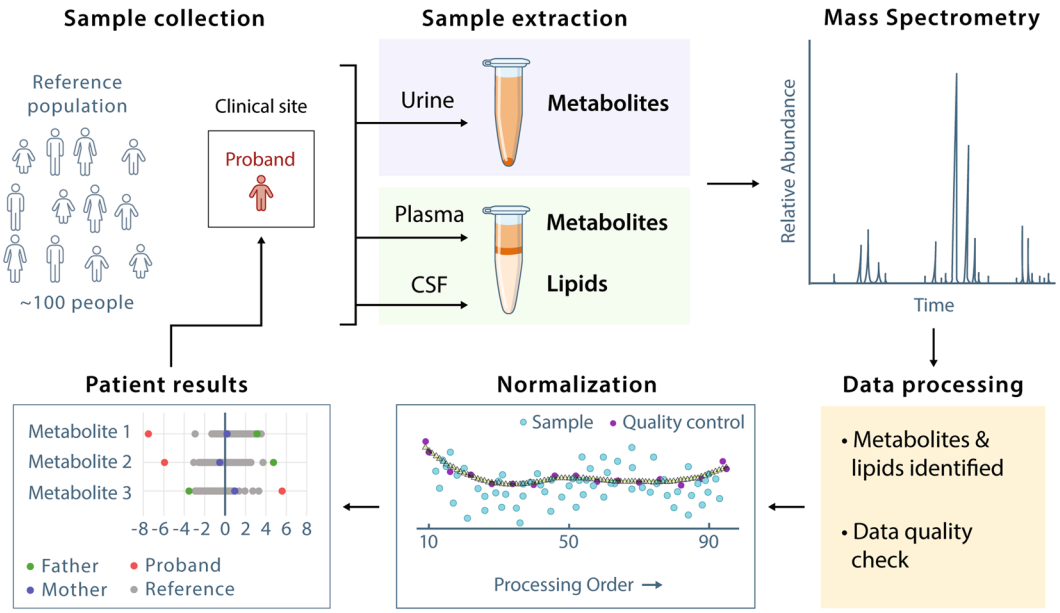

Fig. 1 Overview of the study design. Biofluid samples were collected from probands at the UDN clinical sites and then extracted for metabolomics (urine, plasma, CSF) and lipidomics (plasma and CSF) analyses using chromatography coupled to mass spectrometry (GC-MS for metabolomics and LC-MS/MS for lipidomics). Data were pre-processed, including data quality checks, normalized, and compared against data from the reference population of healthy individuals. Metabolomics and lipidomics results in the form of Z-score, log2 fold change and p-value per metabolite and lipid of the proband (and associated family members, if applicable) were reported back to the respective UDN Clinical Site for diagnostic assistance.

lipids (lipidomics) in biofluids from probands (i.e. individuals with an undiagnosed disease accepted into the UDN) to assist in the evaluation and/or identification of the causes of rare and undiagnosed diseases. Here, we describe in detail the raw and processed metabolomics and lipidomics data from analyses of UDN patient samples and make the data available to the research community so that it might be useful in the diagnoses of current or future patients suffering from undiagnosed disorders. Our previous publication (Webb-Robertson *et al.*[6]) described the detailed statistical approach used for processing this same underlying data set, and so we refer readers to that work for more details on the statistical analyses employed.

Methods

Study design. The identification of metabolite and lipid outliers via metabolomics evaluation of individual probands by untargeted metabolomics required a normal or reference population for comparison. A reference dataset against which metabolomics data from UDN probands and their relatives could be compared was generated by metabolomics analysis of plasma, urine, and CSF from individuals with no known metabolic disease (Fig. 1). Approval for the study of the individuals in the UDN was provided by the National Institutes of Health under protocol number 15-HG-0130. The UDN is registered at ClinicalTrials.gov under identifier NCT02450851.

UDN probands suffer from undiagnosed diseases and thus are typically represented as a sample size of one; therefore, understanding normal variation within a proband's condition is not possible. To address this issue, we performed power analyses of historical plasma and urine data from the Pacific Northwest National Laboratory (PNNL), assuming an uneven study design (e.g. $n = 1$ for probands and $n = \geq(10–150)$ for healthy controls)[6]. This analysis determined that data from 80–120 healthy individuals would be required to perform a well-powered statistical analysis of the data from a UDN proband. This reference dataset is used to understand normal metabolome variation in a population of similar demographics to the UDN population, which is essential for evaluating the metabolome and lipidome data from individual UDN probands and characterizing the pathophysiology and etiology of their undiagnosed disease.

Reference population. The composition of the reference population (approximately 50% children (<18 years of age) and approximately 50% female) was selected to represent the demographics of the participants enrolled in the Undiagnosed Diseases Program, an NIH intramural program upon which the UDN is based[7]. Biofluids for the reference population included samples collected from the Oregon Clinical & Translational Research Institute Biolibrary (adult plasma and urine), the Oregon Health & Science University Layton Aging and Alzheimer's Research Center (adult CSF), the Vanderbilt University Metabolic Screening Laboratory (paediatric plasma), the Mayo Clinic Biochemical Genetics Laboratory (paediatric and young adult urine and CSF), and BioIVT (adult CSF)[6] (Table 1; ***figshare***[8] **('Demographic information for reference population')**). The individuals composing the reference dataset also consented to sample collection under Institutional Review Boards (IRB) at the respective institutions.

	Plasma (n = 136)	Urine (n = 102)	CSF (n = 149)
Sex			
Female	67 (49%)	55 (54%)	79 (53%)
Male	69 (51%)	47 (46%)	70 (47%)
Age group (years)			
0–0.5	0 (0%)	12 (12%)	16 (11%)
0.6–1.9	23 (17%)	4 (4%)	4 (3%)
2–10	53 (39%)	26 (25%)	25 (17%)
11–17	6 (4%)	8 (8%)	11 (7%)
18–30	23 (17%)	21 (21%)	4 (3%)
31–60	22 (16%)	22 (22%)	29 (19%)
>60	9 (7%)	9 (9%)	60 (40%)

Table 1. Demographics of reference population per biofluid type.

MS analysis	Plasma	Urine	CSF
Reference			
Metabolite (GC-MS)	136	102	149
Lipid (LC-MS/MS)	272	204	298
UDN			
Metabolite (GC-MS)	295	239	6
Lipid (LC-MS/MS)	590	478	12
Total biofluid MS analyses	***1293***	***1023***	***465***

Table 2. Summary of mass spectrometry (MS) sample analyses per biofluid for the reference population and UDN cohort.

For the paediatric and young adult CSF, due to the limited volumes available (100 µl), samples were pooled to reach the required volume of 200 µl for metabolomics analysis. Each CSF paediatric and adolescent reference sample is thus composed of two individuals of the same sex and similar age (e.g., 2 years old combined with 3 months old, and 14 years old combined with 16 years old).

Biofluid collection for UDN participants and sample management. Biofluids were collected from UDN probands by the UDN clinical site at which the individual was evaluated. Written consent from all UDN participants and/or legal guardians was provided prior to sample collection and approved IRB. For each sample, the collection time, fasting state and duration, symptoms, diet supplements, and medications were documented. (***figshare***[8] **('Listing of metabolomics and lipidomics raw data files')**. To assist in determination of potential genetic and environmental influences on the metabolomics findings, when it was possible samples were also collected from unaffected family members. Metabolomics and lipidomics analyses were conducted on 281 UDN participants, including 148 probands and 133 family members (101 unaffected, 25 affected, 7 unknown) (***figshare***[8] **('Listing of metabolomics and lipidomics raw data files')**. This comprised 540 biofluid samples for analysis (295 plasma, 239 urine, and 6 CSF) (Table 2). Combining the reference population and the UDN participants, mass spectrometry analyses were conducted on 2781 biofluid samples. UDN probands with diagnoses are available (***figshare***[8] **('UDN probands with available diagnoses')**).

Blood samples for plasma were collected in purple top EDTA Vacutainer® tubes. The blood was centrifuged at 10 000 × *g* for 10 minutes at 4 °C. Three 50 µl aliquots of plasma were transferred into 0.5 mL Sarstedt Biosphere® SC Micro Tubes. Samples were flash frozen in liquid nitrogen or quick frozen in dry ice/ethanol prior to storage in either a −80 °C or liquid nitrogen freezer with appropriate labels (ID, sample type, and collection date).

Urine samples were requested to be the first morning void and were collected in a polypropylene container. The urine was centrifuged at 1000 × *g* for 5 minutes at 4 °C to remove any cells and particulates. Three 100 µl aliquots were transferred into 0.5 mL Sarstedt Biosphere® SC Micro Tubes and flash frozen in liquid nitrogen or quick frozen in dry ice/ethanol prior to storage in either a −80 °C or liquid nitrogen freezer with appropriate labels (ID, sample type, and collection date).

CSF was collected by lumbar puncture in the L3/L4 or the L4/L5 inter-space. If the samples were not blood contaminated, the sample tubes were placed on ice (or dry ice if available), and then transferred to a −80 °C freezer. If the samples were blood contaminated, the samples were centrifuged immediately (prior to freezing) and the clear CSF transferred to new tubes. Three 200 µl aliquots were transferred into 0.5 mL Sarstedt Biosphere® SC Micro Tubes and flash frozen in liquid nitrogen or quick frozen in dry ice or ethanol prior to storage in a −80 °C freezer with appropriate labels (ID, sample type, and collection date)

All biofluid samples were shipped to the Pacific Northwest National Laboratory on dry ice and stored in −70 °C freezers until sample processing for mass spectrometry (MS) analysis.

Analysis Batch	Metabolomics (GC-MS)			Lipidomics (LC-MS/MS)	
	Plasma	Urine	CSF	Plasma	CSF
Reference_population	06-Jun-16	07-Dec-16	31-Jan-17	22-Sept-16	03-Mar-17
UDN_Pilot1	22-Jun-16	—	—	03-Oct-16	—
UDN_Pilot2	05-Oct-16	—	—	04-Oct-16	—
UDN_Project01	01-Dec-16	21-Dec-16	26-Sep-17	08-Dec-16	10-Oct-17
UDN_Project02	27-Feb-17	22-Dec-16	—	27-Mar-17	—
UDN_Project03	27-Sep-17	02-Oct-17	—	06-Oct-17	—
UDN_Project04	20-Sep-17	03-Oct-17	—	07-Oct-17	—
UDN_Project05	29-Sep-17	04-Oct-17	—	08-Oct-17	—
UDN_Project06	21-Sep-17	05-Oct-17	—	09-Oct-17	—
UDN_Project07	24-Oct-17	25-Oct-17	—	24-Oct-17	—
UDN_Project08	12-Dec-17	21-Feb-18	—	11-Dec-17	—
UDN_Project09	20-Feb-18	16-May-18	—	03-Feb-18	—
UDN_Project10	14-May-18	17-May-18	—	21-May-18	—
UDN_Project11	15-May-18	01-Oct-18	—	22-May-18	—
UDN_Project12	02-Oct-18	—	—	18-Oct-18	—

Table 3. Batch start date for the mass spectrometry analysis of the reference population and subsequent UDN participant samples per instrument type and biofluid.

Quality control samples. The NIST SRM 1950 was used as a plasma QC[9,10]. The NIST QC is composed of 100 healthy individuals between 40–50 years old, an equal number of men and women, and a race distribution representative of the US population. The NIST QC is a commercially available reference material (certified until year 2023) and was chosen due to the multi-year nature of this study. For urine and CSF, as no commercially available reference materials were identified, pools were generated from the reference population for each respective biofluid and used as QCs.

Sample batches. Sample batches were formed based on the number of analyses that could be performed in approximately one day (~33 analyses). Randomized run orders were generated based on sex, age, ethnicity (if provided), family association, and clinical site (if samples from more than 1 clinical site were available at the time of batching) (see Technical Validation section) prior to extraction, sample preparation, chemical derivatization of metabolites, and instrument analysis runs.

The instrument run order included a batch structure that enabled data normalization via Quality Control (QC)-based Robust Locally Estimated Scatterplot Smoothing (LOESS) Signal Correction (QC-RLSC)[11] to specifically account for batches of samples that were not analysed back to back but dispersed over a longer timeframe[6] (Table 3). The Pilot batches were the initial batches to be analysed and were used to confirm the normalization approach. Both Pilot and Project batches were processed using the same methodologies. This normalization method required a batch structure with specific placement of QC samples. For GC-MS analyses, the batch began with 2 blanks, 1 fatty acid methyl ester (FAME), 1 blank, 3 QCs, samples with evenly dispersed single QCs, and ending with 2 QCs. LC-MS/MS batches were similar except there was no FAME and a blank was run after the first 3 QCs, the middle QC, and at the very end to assess carryover[6].

Extraction of metabolites and lipids from plasma and CSF. For plasma and CSF, 50 µL and 200 µL, respectively, were used for metabolite and lipid extraction using a modified Folch extraction[12], the MPLEx protocol[13]. Prior to extraction, samples were transferred to MµlTI SafeSeal Sorenson microcentrifuge tubes. To the plasma, 50 µL of GC-MS internal standards (malonic acid-d4, succinic acid-d4, glycine-d5, citric acid-d4, fructose 13C6, L-tryptophan-d5, lysine-d4, alanine-d7, stearic acid-d35, benzoic acid-d5, octanoic acid-d15 at a final concentration of 1 µg/µL each)[11] and 10 µL of LC-MS internal standards (SPLASH™ Lipidomix® Mass Spec Standard, Avanti Polar Lipids, Inc.) were added. To the CSF, 50 µL of GC-MS internal standards (fructose 13C6, L-tryptophan-d5, lysine-d4, alanine-d7, stearic acid-d35, benzoic acid-d5, octanoic acid-d15 at a final concentration of 1 µg/µL each) and 10 µL of LC-internal standards ((PC(17:0/14:1) at 1 µg/µL, LPC(19:0) at 0.01 µg/µL, and TG(17:0/17:1/17:0)-d5 at 0.01 µg/µL) were added. Cold (−20 °C) chloroform/methanol (2:1, v/v) was added in a 4-fold excess to the sample volume. Samples were vortexed for 10 seconds to facilitate mixing of samples and solvent, allowed to sit on ice for 5 minutes, and then vortexed again for 10 seconds. Then, the samples were centrifuged to facilitate separation of a top hydrophilic layer containing polar metabolites and a bottom hydrophobic layer containing lipids. The hydrophilic layers were transferred into new 2.0 mL glass autosampler vials, evaporated to dryness *in vacuo*, and stored dry at −20 °C until chemical derivatization (see below). The lower hydrophobic layers containing the total lipid extract (TLE) were transferred into new 1.7 mL glass autosampler vials, evaporated to dryness *in vacuo*, and stored at −20 °C in 500 µL of chloroform/methanol (2:1, v/v) until instrument analysis.

Extraction of metabolites from urine. For urine, 100 µL was used for metabolite extraction, as previously described[14]. Samples were transferred to MµlTI SafeSeal Sorenson microcentrifuge tubes to which 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 at a final concentration of 1 μg/μL each) and 100 μL of a 1 mg/mL solution of urease prepared in water were added. The samples were incubated for 30 minutes at 37 °C with mild shaking to deplete urea. Metabolites were then extracted with concomitant protein precipitation by addition of 1 mL of cold (−20 °C) methanol. Samples were vortexed for 30 seconds and precipitated proteins were isolated by centrifugation. The supernatants were transferred to glass autosampler vials and then dried *in vacuo*. Metabolite extracts were stored dry at −20 °C until chemical derivatization (see below).

Chemical derivatization of metabolites. Polar metabolites were chemically derivatized prior to metabolomics analysis. Two post-extraction standards (pentadecanoic acid-d3 and 3-hydroxymyristic acid-d5 at 1 μg/μL final concentration) were added to monitor instrument performance. Chemical derivatization of metabolites was previously detailed[14]. To protect carbonyl groups and reduce the number of tautomeric isomers, 20 μL of methoxyamine in pyridine (30 mg/mL) was added to each sample, followed by vortexing for 30 seconds and incubation at 37 °C with generous shaking for 90 minutes. To derivatize hydroxyl and amine groups to trimethylsilylated (TMS) forms, 80 μL of N-methyl-N-(trimethylsilyl)trifluoroacetamide (MSTFA) with 1% trimethylchlorosilane (TMCS) was added to each vial, followed by vortexing for 10 seconds and incubation at 37 °C with shaking for 30 minutes. The samples were allowed to cool to room temperature and were analysed the same day.

GC-MS analysis. An Agilent GC 7890 A coupled with a single quadrupole MSD 5975 C was used to analyze chemically derivatized metabolites. GC-MS analysis was previously detailed[14]. Briefly, 1 μL of each sample was injected onto a HP-5MS column (30 m × 0.25 mm × 0.25 μm; Agilent Technologies, Inc). The injection port temperature was held at 250 °C throughout the analysis. The GC oven was held at 60 °C for 1 minute after injection then increased to 325 °C by 10 °C/min, followed by a 5-minute hold at 325 °C. Total analysis time was 34 minutes per injection. The helium gas flow rates were determined by the Agilent Retention Time Locking function based on analysis of deuterated myristic acid. Data were collected over the mass range 50–550 m/z. A mixture of fatty acid methyl esters (C8-C28) was analysed once per day at the beginning of each batch together with the samples for retention index alignment purposes during subsequent data analysis.

LC-MS analysis. Stored plasma TLEs were dried *in vacuo* (30 min) and reconstituted in 200 μL of methanol containing post-extraction internal standards (PE(17:0/14:1) and PI(17:0/14:1) at a final amount of 0.05 μg and 0.01 μg, respectively). Stored CSF TLEs were dried *in vacuo* and reconstituted in 50 μL of methanol containing the same post-extraction internal standards at a final amount of 0.02 μg each. The TLEs were analysed as outlined in Kyle *et al.*[15]. A Waters Acquity UPLC H class system interfaced with a Velos-ETD Orbitrap mass spectrometer was used for LC-ESI-MS/MS analyses. 10 μl of reconstituted sample was injected onto a Waters CSH column (3.0 mm × 150 mm x 1.7 μm particle size) and separated over a 34-minute gradient (mobile phase A: ACN/H2O (40:60) containing 10 mM ammonium acetate; mobile phase B: ACN/IPA (10:90) containing 10 mM ammonium acetate) at a flow rate of 250 μl/minute. Eluting lipids were introduced to the MS via electrospray ionization in both positive and negative modes, and lipids were fragmented using higher-energy collision dissociation (HCD) and collision-induced dissociation (CID).

Metabolite identification and data processing. Metabolite identifications and data processing were conducted as previously detailed[14]. GC-MS raw data files were processed using the Metabolite Detector software, version 2.0.6 beta[16]. Retention indices (RI) of detected metabolites were calculated based on the analysis of the FAMEs mixture, followed by their chromatographic alignment across all analyses after deconvolution. Metabolites were identified by matching experimental spectra to an augmented version of the Agilent Fiehn Metabolomics Retention Time Locked (RTL) Library[17], containing spectra and validated retention indices. All metabolite identifications were manually validated. The NIST 08 GC-MS library was also used to cross validate the spectral matching scores obtained using the Agilent library and to provide identifications for metabolites that were initially unidentified. The three most abundant fragment ions in the spectra of each identified metabolite were automatically determined by Metabolite Detector, and their summed abundances were integrated across the GC elution profile. A matrix of identified metabolites, unidentified metabolite features, and their corresponding abundances for each sample in the batch were exported for statistics.

Processing the data from the analyses of the reference population resulted in the identification of 81 plasma polar metabolites (across 16 super classes and 27 classes as categorized in the Human Metabolome Database[18,19], 116 urine metabolites (across 17 super classes and 28 classes), and 82 CSF metabolites (across 14 super classes and 26 classes) (Table 4)

Lipid identification and data processing. LC-MS/MS lipidomics data were analyzed using LIQUID (Lipid Informed Quantitation and Identification)[15]. Analysis parameters included an initial precursor mass error tolerance of 20 ppm (i.e. ±10 ppm), and fragment mass error tolerances of 20 ppm (±10 ppm) and 500 ppm (±250 ppm) for HCD and CID MS/MS events, respectively. Confident identifications were selected by manually evaluating the MS/MS spectra for diagnostic and corresponding acyl chain fragments of the identified lipid. In addition, the precursor isotopic profile, extracted ion chromatogram, and mass measurement error along with the elution time were evaluated. For certain lipids, multiple LC peaks having nearly identical MS/MS spectra were observed, suggesting the presence of lipid stereoisomers. In these cases, the stereochemistry of the lipid isomers could not be completely determined based on the LC-MS/MS data alone, and so these isomers are annotated with "_A", "_B" or "_C" at the end of the lipid name. Typically, the mass measurement error of confidently identified lipids was within ± 2.5 ppm. Given the time-consuming nature of manual validation of each identified lipid, a library of confident lipid identifications was generated from the reference dataset and select UDN participants (3 NIST QCs, 6 pooled plasma of reference population, 2 reference individuals, and 3 UDN participant). All LC-MS/MS data were aligned and gap-filled to this target database for feature identification using the identified lipid

HMDB Metabolite Class	Plasma	Urine	CSF
Alcohols and Polyols	2	1	2
Alkylamines	1	—	—
Amines	—	1	1
Amino Acids and Derivatives	19	25	15
Anhydrohexose	1	—	1
Benzenoids	—	9	—
Benzoic Acid and Derivatives	1	—	—
Carboxylic Acids and Derivatives	6	10	4
Cyclic Alcohols and Derivatives	—	1	—
Diazines	1	—	—
Disaccharides	2	6	1
Fatty Acids and Conjugates	4	1	5
Furans	—	1	—
Glycerolipids	2	—	—
Glycerophospholipids	—	1	1
Glycosyl Compounds	—	1	1
Hydroxy Acids and Derivatives	6	15	11
Imidazolidines	—	1	—
Imidazopyrimidines	2	3	—
Indoles and derivatives	1	—	—
Keto acids and derivatives	5	1	1
Keto-Acids and Derivatives	1	1	2
Lactams	1	1	1
Lactones	—	—	2
Monosaccharides	6	12	11
Non-metal Oxoanionic Compounds	1	1	1
Organic carbonic acids and derivatives	—	—	1
Organic nitrogen compound	1	1	1
Organic Phosphoric Acids and Derivatives	1	—	—
Organooxygen compound	—	—	1
Phenols	1	—	—
Phenylpropanoids and polyketides	—	1	—
Piperidines	—	—	1
Purine Nucleosides and Analogues	—	3	—
Purine Nucleotides	1	—	—
Pyridines and derivatives	2	3	1
Pyrimidines and pyrimidine derivatives	—	1	1
Steroids and Steroid Derivatives	1	—	—
Sugar Acids and Derivatives	3	6	4
Sugar alcohols	4	7	8
Tetrapyrroles and Derivatives	1	1	1
Unclassified	4	1	3
Total	*81*	*116*	*82*

Table 4. Number of metabolites identified in the reference population biofluids by metabolite class.

name, observed *m/z*, and the retention time using MZmine 2[20] (see ***figshare***[8] **('Parameters used for MZmine2 processing of lipidomics data')**). Data from each ionization type were aligned and gap-filled separately. Aligned features were manually verified and peak apex intensity values were exported for statistical analysis. All subsequent batches were aligned to this library of confident lipid identifications.

To correct for batch retention time (RT) shifts for alignment to the reference library, an in-house tool to correct for linear RT shifts was used. The instrument files were converted into.mzXML files using MSConvert[21]. Each file was associated with a target list containing the name, RT, and *m/z* of the internal standards within a batch and was imported into MZmine. As the internal standards alone did not elute across the entire gradient, two lipids that were present in all samples in positive mode (carnitine(10:1) and CE(18:1)) and one in negative mode (HexCer(d18:1/24:0) were included in the target list as they eluted near the start and end of the gradient. The peak alignment of each target was manually validated, and corrected if needed, and the RT of each target lipid was exported. These targets acted as anchor points for the RT correction. Using the RT anchors for each target, all instrument files within a batch were shifted and aligned to the reference and new.mzXML were generated for

LipidMaps Lipid Subclass	Plasma	CSF
Carnitine Esters	4	—
NAE	2	1
CE	6	4
Ubiquinone	1	—
Cer	15	1
SM	35	26
HexCer	5	5
Hex2Cer	1	—
LPC	23	16
PC	88	52
LPCO	2	—
PCO	18	11
PCP	13	9
LPE	7	3
PE	26	13
PEO	6	4
PEP	40	9
LPI	2	—
PI	24	8
PS	1	—
DG	6	7
TG	137	39
Total	462	208

Table 5. Number of lipids identified in the reference population biofluids, by lipid subclass.
NAE = N-acylethanalamine; Cer = ceramide; SM = sphingomyelin; HexCer = Hexosylceramide; LacCer = Dihexosylceramide; LPC = monoacylglycerophosphocholine; PC = diacylglycerophosphocholine; LPCO = Monoalkylglycerophosphocholine; PCO = alkyl, acylglycerophosphocholine; PCP = 1Z-alkenyl acylglycerophosphocholine LPE = monoacylglycerophosphoethanolmine; PE = diacylglycerophosphoethanolmine; PEO = alkyl acylglycerophospho-ethanolamine; PEP = plasmalogen PE; LPI = PI = monoacylglycerophosphoinositol; PI = diacylglycerophosphoinositol; PS = diacylglycerophosphoserine; DG = diacylglyceride; TG = triacylglyceride.

subsequent alignment in MZmine. For all batches aligned to the reference list, lipid identifications were randomly selected (approximately 30 lipids) and verified using LIQUID to ensure that identification in the reference and sample batches matched.

Processing the data from the analyses of the reference population resulted in the identification of 462 plasma lipids across 6 lipid categories and 23 lipid subclasses (as categorized by LipidMaps)[22–24], and 208 CSF lipids across 6 lipid categories and 17 lipid subclasses (Table 5).

Statistical analysis. We have previously described in detail the statistical approach used for processing the data[6], and briefly summarize this below. To facilitate the identification of potentially disease-associated analyte profiles of UDN participants, a reference population of individuals with no known metabolic diseases was established as described above. Batches of samples from UDN participants were analysed and compared to this reference population as outlined in Webb-Robertson *et al.*[6]. Briefly, quality control (QC) processing of the reference dataset includes log2 transformation and the removal of any identified or unidentified features not present in at least 10% of the samples. Samples with missing or low abundance values and an uncorrelated pattern of expression by Pearson correlation and rMd-PAV[25] were assessed to determine whether the seemingly poor behaviour was most likely due to biological or to technical/sample preparation issues. If biological issues appeared to be the cause, the sample was retained in the current batch for further analysis; if technical issues appeared to be the cause, the sample was omitted from further analyses. QC processing for the participant samples included the same steps as for the reference samples; however, participant samples required stronger evidence before removal than reference samples.

Normalization of the reference data and the participant data was performed in two steps[6]. First, QC-RLSC accounted for batch effects, and was performed on a per-batch basis[11]. This required identical QC samples to be run in every batch of samples (for reference samples and UDN samples alike), as described above. Quality control–based robust LOESS signal correction (QC-RLSC) was implemented using the parameter values described previously[11]. Namely, a missingness threshold requiring the observation of a molecule in at least half of the QC samples, filtering of molecules with RSD above 30 percent, and possible polynomial degrees of first and second order. To account for differences in the amount of sample material analysed by GC-MS or LC-MS, QC-RLSC was followed by global median centering of each sample, where each log2 biomolecule value within a sample was normalized via subtraction of the corresponding sample median (also on the log2 scale)..

To identify unique features in the analyte profiles of participants, results were compared to those from the reference dataset[6]. A univariate approach was applied that compared the feature values of the participants to the mean and standard deviation of the feature values in the reference dataset using z-scores[26]. An absolute value z-score threshold was used to obtain a list of metabolites and/or lipids with outlying z-scores that may have potential diagnostic significance. Additionally, for a given participant and biomolecule, log2 fold changes relative to the reference data were computed as the difference between the participant's log2 value and the median log2 value of the reference population.

Data Records

The raw LC-MS and GC-MS data files in .raw and.D format, and converted files in .mzML format, and.CDF format, respectively were deposited and are publicly available at the MassIVE repository (MSV000084717[27], MSV000085506[28], MSV000085508[29]). The normalized values for all identified lipids and metabolites for the UDN individuals and reference population are also available in MassIVE. The evidence supporting the molecular identifications (e.g. fragment ion *m/z*, retention times) are provided (**figshare**[8] **('The evidence supporting the molecular identifications (e.g. fragment ion m/z, retention times)')**). The deposited data also contains the post-processed data including the log2 fold change, Z-score, and p-value for each lipid and metabolite per UDN individual in .csv format. For the lipid results, the identifications made in positive and negative ionization mode were consolidated into a single file. Family member data is included in the associated proband files. In addition, for the UDN probands that have been diagnosed, the diagnosis name and relevant gene information are provided (**figshare**[8] **('UDN probands with available diagnoses')**).

The data deposited to MassIVE contains up to three directories: peak, quant, and raw. Each biofluid data repository also contains automatically generated subdirectories prefixed with "ccms". Users of the data should obtain data from the peak, quant, and raw directories listed above and detailed below:

peak/ → Peak_List_Files/

1. Reference_pop_(biofluid)_lipid/ = LC-MS/MS instrument files in .mzXML format for the reference population
2. Reference_pop_(biofluid)_metab/ = GC-MS instrument files in .cdf format for the reference population
3. UDN_(biofluid)_lipid/ = LC-MS/MS instrument files in .mzXML format for the UDN participant
4. UDN_(biofluid)_metab/ = GC-MS instrument files in .cdf format for the UDN participant

quant/ → Quantification_Results/

1. UDN_(biofluid)_lipid_normalized_data/ = normalized data files for the associated biofluid per lipidomics batch analysis for lipids identified in both positive (POS) and negative (NEG) ionization mode
2. UDN_(biofluid)_lipid_results/ = The lipidomics result file containing the log2 fold change, Zscore, and p-value for each identified lipid per UND participant (and relatives, if applicable)
3. UDN_(biofluid)_metab_normalized_data/ = normalized data files for the associated the biofluid per metabolomics batch analysis for the identified metabolites
4. UDN_(biofluid)_metab_results/ = The metabolomics result file containing the log2 fold change, Zscore, and p-value for each identified metabolite per UND participant (and relatives, if applicable)

raw/ → Raw_Spectrum_Files/

1. Reference_pop_(biofluid)_lipid/ = LC-MS/MS Thermo instrument files in .raw format for the reference population
2. Reference_pop_(biofluid)_metab/ = GC-MS Agilent instrument files in .D format for the reference population
3. UDN_(biofluid)_lipid/ = LC-MS/MS Thermo instrument files in .raw format for the UDN participant
4. UDN_(biofluid)_metab/ = GC-MS Agilent instrument files in .D format for the UDN participant

updates/ → 2020-03-10_alchemistmatt_b439c281 → quant → Quantification_Results → UDN_urine_metab_normalized_data/ = normalized data file for the reference population for urine metabolites.

Technical Validation

To ensure unbiased data production, randomization orders were created and followed for sample extraction, GC-MS derivatization, and MS run orders. Family units, meaning probands and their relatives, were analysed within the same batch. Batch sizes were limited to the number of samples that could be analysed by GC-MS in approximately one day due to the stability of the chemically derivatized metabolites. Approximately 33 samples composed a batch. Samples were randomized based on sex, age, ethnicity (if provided), family association, and clinical site (if samples from more than 1 clinical site were available at the time of batching). Randomization orders were created as sufficient samples accumulated to make up a new batch over the 2.5 years of the study. To monitor data quality, the QC samples (across all molecules) were evaluated with prior batches collected to verify removal of batch effects via normalization. In addition, on a batch-by-batch basis, data quality was monitored by visual inspection of the log2 internal standard values across all samples within a batch.

To evaluate the consistency of the data collection process the coefficient of variation (CV) was utilized. The CV is defined as the standard deviation divided by the mean and lower values signify lower variability. Using data from the reference population, QC samples per platform and biofluid, the median CV along with the first and

Platform & Biofluid	Min.	1st Qu.	Median	Mean	3rd Qu.	Max
Metabolite Urine	12.61	16.4	18.22	23.52	22.38	195.9
Metabolite Plasma	16.93	24.61	30.38	41.23	44.17	189.6
Metabolite CSF	12.68	23.53	30.27	41.89	49.80	201.4
Lipid POS Plasma	9.73	13.32	14.91	17.56	16.91	179.3
Lipid NEG Plasma	20.64	24.20	27.68	30.84	32.36	96.25
Lipid POS CSF	13.62	24.32	26.16	33.05	32.00	103.30
Lipid NEG CSF	25.16	32.08	46.34	45.16	51.22	87.42

Table 6. The coefficient of variation (CV) per platform and biofluid for the reference population QC samples calculated from raw values. Min = minimum CV, 1st Qu. = first quartile, 3rd Qu. = third quartile, Max. = maximum CV, POS = positive mode ionization, NEG = negative mode ionization.

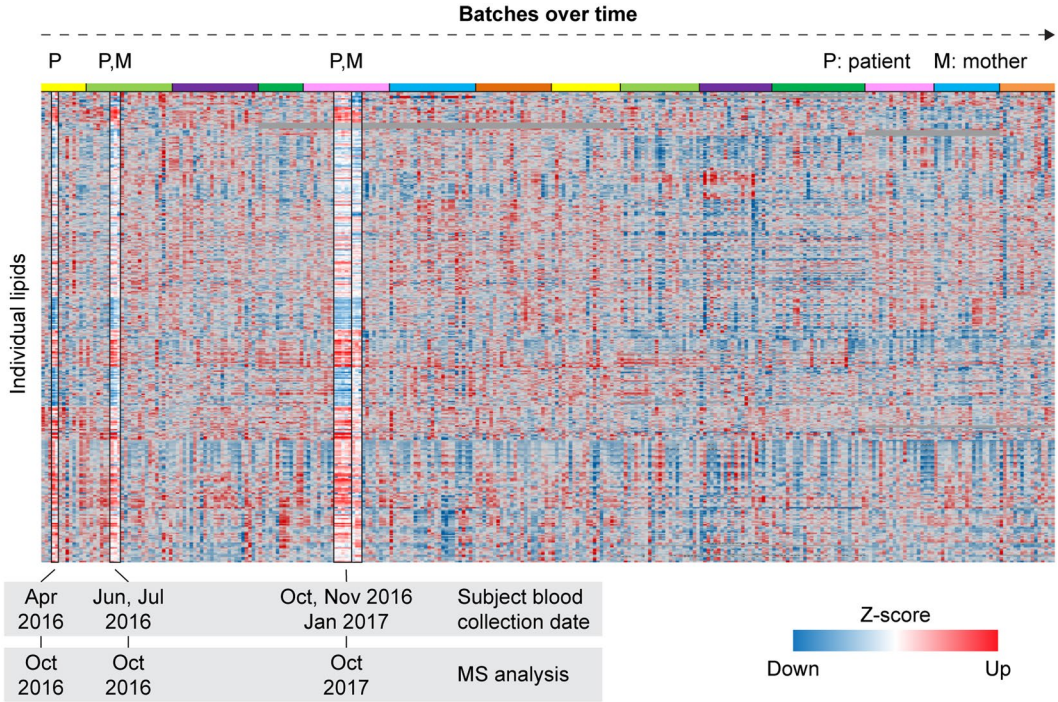

Fig. 2 Z-score map of plasma lipidomics data for all UDN individuals (n = 294). One proband (P) and her mother (M) had samples analysed in multiple batches over the course of one year. The proband's samples were collected at different times for each batch and analysed on the MS in October 2016 (2 batches, 2 samples per batch) and October 2017 (1 batch, 5 samples). The proband's lipid profile remained very similar between each analysis batch. Top coloured bar indicates the different batches over time.

third quartile are shown in Table 6. The median CV of the lipid negative mode CSF is the greatest, possibility due to the lower number of samples available and lipids identified.

To evaluate the reproducibility of the results, we assessed the lipidomics results from one UDN proband from whom we had 9 samples collected over a period of 9 months. The samples were analysed in 3 different batches, separated by up to 1 year (Fig. 2). The proband's mother had 3 samples analysed in two batches at time intervals coinciding with the proband's samples. As shown in Fig. 2, the Z-score pattern of both the proband's and the mother's samples remain consistent between batches across the one-year timespan between the collection and analysis of the first and last set of samples (October 2016 to October 2017).

Code availability

Statistical processing and analyses were performed in R version 3.4.0. Quality control and median normalization were performed using the R package *pmartR* version 0.9.0, freely available on GitHub (https://github.com/pmartR/pmartR)[30]. Default parameter values for *pmartR* function calls were used. QC-RLSC and the calculation of log2 fold changes and z-scores were carried out using in-house R functions and are available on Github (https://github.com/pmartR/qcrlsc).

Received: 6 January 2020; Accepted: 9 March 2021;
Published online: 21 April 2021

References

1. Nicholson, J. K. *et al.* Metabolic phenotyping in clinical and surgical environments. *Nature* **491**, 384–392, https://doi.org/10.1038/nature11708 (2012).
2. Garcia-Cazorla, A., Mochel, F., Lamari, F. & Saudubray, J. M. The clinical spectrum of inherited diseases involved in the synthesis and remodeling of complex lipids. A tentative overview. *Journal of inherited metabolic disease* **38**, 19–40, https://doi.org/10.1007/s10545-014-9776-6 (2015).
3. Lamari, F., Mochel, F. & Saudubray, J. M. An overview of inborn errors of complex lipid biosynthesis and remodelling. *Journal of inherited metabolic disease* **38**, 3–18, https://doi.org/10.1007/s10545-014-9764-x (2015).
4. Ramoni, R. B. *et al.* The Undiagnosed Diseases Network: Accelerating Discovery about Health and Disease. *American journal of human genetics* **100**, 185–192, https://doi.org/10.1016/j.ajhg.2017.01.006 (2017).
5. Network, T. U. D. Undiagnosed Diseases Network. (2017).
6. Webb-Robertson, B. M. *et al.* Statistically-driven Metabolite and Lipid Profiling of Patients from the Undiagnosed Diseases Network. Analytical chemistry, https://doi.org/10.1021/acs.analchem.9b03522 (2019).
7. Gahl, W. A. *et al.* The National Institutes of Health Undiagnosed Diseases Program: insights into rare diseases. *Genetics in medicine: official journal of the American College of Medical Genetics* **14**, 51–59, https://doi.org/10.1038/gim.0b013e318232a005 (2012).
8. Kyle, J. E. *et al.* A resource of lipidomics and metabolomics data from individuals with undiagnosed diseases. *figshare* https://doi.org/10.6084/m9.figshare.12440342 (2021).
9. Bowden, J. A. *et al.* Harmonizing lipidomics: NIST interlaboratory comparison exercise for lipidomics using SRM 1950-Metabolites in Frozen Human Plasma. *Journal of lipid research* **58**, 2275–2288, https://doi.org/10.1194/jlr.M079012 (2017).
10. Simon-Manso, Y. *et al.* 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. *Analytical chemistry* **85**, 11725–11731, https://doi.org/10.1021/ac402503m (2013).
11. Dunn, W. B. *et al.* Procedures for large-scale metabolic profiling of serum and plasma using gas chromatography and liquid chromatography coupled to mass spectrometry. *Nature protocols* **6**, 1060–1083, https://doi.org/10.1038/nprot.2011.335 (2011).
12. Folch, J., Lees, M. & Sloane Stanley, G. H. A simple method for the isolation and purification of total lipides from animal tissues. *The Journal of biological chemistry* **226**, 497–509 (1957).
13. Nakayasu, E. S. *et al.* MPLEx: a Robust and Universal Protocol for Single-Sample Integrative Proteomic, Metabolomic, and Lipidomic Analyses. mSystems 1, https://doi.org/10.1128/mSystems.00043-16 (2016).
14. Webb-Robertson, B. J. *et al.* A Statistical Analysis of the Effects of Urease Pre-treatment on the Measurement of the Urinary Metabolome by Gas Chromatography-Mass Spectrometry. *Metabolomics: Official journal of the Metabolomic Society* **10**, 897–908, https://doi.org/10.1007/s11306-014-0642-1 (2014).
15. Kyle, J. E. *et al.* LIQUID: an-open source software for identifying lipids in LC-MS/MS-based lipidomics data. *Bioinformatics (Oxford, England)* **33**, 1744–1746, https://doi.org/10.1093/bioinformatics/btx046 (2017).
16. Hiller, K. *et al.* MetaboliteDetector: comprehensive analysis tool for targeted and nontargeted GC/MS based metabolome analysis. *Analytical chemistry* **81**, 3429–3439, https://doi.org/10.1021/ac802689c (2009).
17. Kind, T. *et al.* FiehnLib: mass spectral and retention index libraries for metabolomics based on quadrupole and time-of-flight gas chromatography/mass spectrometry. *Analytical chemistry* **81**, 10038–10048, https://doi.org/10.1021/ac9019522 (2009).
18. Djoumbou Feunang, Y. *et al.* ClassyFire: automated chemical classification with a comprehensive, computable taxonomy. *Journal of cheminformatics* **8**, 61, https://doi.org/10.1186/s13321-016-0174-y (2016).
19. Wishart, D. S. *et al.* HMDB 3.0–The Human Metabolome Database in 2013. *Nucleic acids research* **41**, D801–807, https://doi.org/10.1093/nar/gks1065 (2013).
20. Pluskal, T., Castillo, S., Villar-Briones, A. & Oresic, M. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC bioinformatics* **11**, 395, https://doi.org/10.1186/1471-2105-11-395 (2010).
21. Adusumilli, R. & Mallick, P. Data Conversion with ProteoWizard msConvert. *Methods in molecular biology (Clifton, N.J.)* **1550**, 339–368, https://doi.org/10.1007/978-1-4939-6747-6_23 (2017).
22. Fahy, E. *et al.* A comprehensive classification system for lipids. *Journal of lipid research* **46**, 839–861, https://doi.org/10.1194/jlr.E400004-JLR200 (2005).
23. Fahy, E. *et al.* Update of the LIPID MAPS comprehensive classification system for lipids. *Journal of lipid research* **50**(Suppl), S9–14, https://doi.org/10.1194/jlr.R800095-JLR200 (2009).
24. Liebisch, G. *et al.* Shorthand notation for lipid structures derived from mass spectrometry. *Journal of lipid research* **54**, 1523–1530, https://doi.org/10.1194/jlr.M033506 (2013).
25. Matzke, M. M. *et al.* Improved quality control processing of peptide-centric LC-MS proteomics data. *Bioinformatics (Oxford, England)* **27**, 2866–2872, https://doi.org/10.1093/bioinformatics/btr479 (2011).
26. Miller, M. J. *et al.* Untargeted metabolomic analysis for the clinical screening of inborn errors of metabolism. *Journal of inherited metabolic disease* **38**, 1029–1039, https://doi.org/10.1007/s10545-015-9843-7 (2015).
27. Kyle, J. E. *et al.* Undiagnosed Disease Network Study Results, Urine lipidomics and metabolomics. *MassIVE* https://doi.org/10.25345/C5H10H (2020).
28. Kyle, J. E. *et al.* Undiagnosed Disease Network Study Results, CSF lipidomics and metabolomics. *MassIVE* https://doi.org/10.25345/C5XH7Q (2020).
29. Kyle, J. E. *et al.* Undiagnosed Disease Network Study Results, Plasma lipidomics and metabolomics. *MassIVE* https://doi.org/10.25345/C5P11F (2020).
30. Stratton, K. G. & Webb-Robertson, B. M. *pmartR: Quality Control and Statistics for Mass Spectrometry-Based Biological Data.* **18**, 1418–1425, https://doi.org/10.1021/acs.jproteome.8b00760 (2019).

Acknowledgements

We are grateful for the participation of patients, family members, and their referring clinicians. We would also like to thank the UDN investigators and the following individuals, institutions and funding sources for the donation of reference 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 the Vanderbilt University Medical Center; and Dr. Devin Oglesbee at Mayo Clinic (with support by Mayo Clinic's Department of Laboratory Medicine and Pathology). This work was funded by the Undiagnosed Diseases Network (1U01TR001395) supported by the National Institutes of Health Common Fund Program. 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). PNNL is operated by Battelle Memorial Institute for the DOE under contract DEAC05-76RLO1830.

Author contributions

D.M.K. and T.O.M. conceived the project. J.E.K. provided logistical management of the project and performed lipidomics analysis. E.M.Z. performed sample preparation and GC-MS data acquisition. Y.M.K. performed metabolomics analysis. K.J.B. performed LC-MS data acquisition. K.G.S., B.J.W.R. and K.M.W. designed normalization and statistical analysis. K.G.S. performed statistical analysis. J.E.K., D.M.K. and T.O.M. wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to D.M.K. or T.O.M.

Reprints and permissions information is available at www.nature.com/reprints.

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

Undiagnosed Diseases Network

Carlos A. Bacino[4], Neil A. Hanchard[4], Richard A. Lewis[4], Jill A. Rosenfeld[4], Daryl A. Scott[4], Alyssa A. Tran[4], Patricia A. Ward[4], Lindsay C. Burrage[4], Gary D. Clark[4], Mercedes E. Alejandro[4], Jennifer E. Posey[4], Michael F. Wangler[4], Brendan H. Lee[4], William J. Craigen[4], Hugo J. Bellen[4], Sarah K. Nicholas[4], Bret L. Bostwick[4], Susan L. Samson[4], Alica M. Goldman[4], Paolo M. Moretti[4], Christine M. Eng[4], Donna M. Muzny[4], James P. Orengo[4], Tiphanie P. Vogel[4], Seema R. Lalani[4], David R. Murdock[4], Mahshid S. Azamian[4], Jordan S. Orange[4], Lisa T. Emrick[4], Shweta U. Dhar[4], Ashok Balasubramanyam[4], Lorraine Potocki[4], Shinya Yamamoto[4], Yaping Yang[4], Shan Chen[4], Fariha Jamal[4], Lefkothea Karaviti[4], Ronit Marom[4], Sharyn A. Lincoln[5], Chris A. Walsh[5], Alan H. Beggs[5], Lance H. Rodan[5], Joan M. Stoler[5], Gerard T. Berry[5], Laurel A. Cobban[6], Calum A. MacRae[6], Joel B. Krier[6], Edwin K. Silverman[6], Elizabeth L. Fieg[6], Richard L. Maas[6], Joseph Loscalzo[6], Aaron Aday[6], Susan Korrick[6], David B. Goldstein[7], Nicholas Stong[7], Jennifer A. Sullivan[8], Rebecca C. Spillmann[8], Loren D. M. Pena[8], Queenie K.-G. Tan[8], Nicole M. Walley[8], Yong-hui Jiang[8], Allyn McConkie-Rosell[8], Kelly Schoch[8], Vandana Shashi[8], Heidi Cope[8], Ingrid A. Holm[9], Isaac S. Kohane[9], Alexa T. McCray[9], Cecilia Esteves[9], Kimberly LeBlanc[9], Matthew Might[9], Emily Kelley[9], Elizabeth A. Worthey[10], Daniel C. Dorset[10], Braden E. Boone[10], Shawn E. Levy[10], Camille L. Birch[10], Angela L. Jones[10], Donna M. Brown[10], David P. Bick[10], J. Scott Newberry[10], Jozef Lazar[10], Thomas May[10], David A. Sweetser[11], Lauren C. Briere[11], J. Carl Pallais[11], Cynthia M. Cooper[11], Frances High[11], Melissa Walker[11], Heather A. Colley[12], Laura A. Mamounas[12], Teri A. Manolio[12], Elizabeth A. Burke[12], Rena A. Godfrey[12], Catherine A. Groden[12], William A. Gahl[12], Lynne A. Wolfe[12], Thomas C. Markello[12], C. Christopher Lau[12], David D. Draper[12], Sarah E. Gould[12], Michele E. Nehrebecky[12], Colleen E. Wahl[12], Gabriel F. Batzli[12], Ellen F. Macnamara[12], Jyoti G. Dayal[12], David J. Eckstein[12], John J. Mulvihill[12], Cynthia J. Tifft[12], Tiina K. Urv[12], Anastasia L. Wise[12], Jennifer L. Murphy[12], Andrea L. Gropman[12], Ellen M. Howerton[12], Donna M. Krasnewich[12], Jean M. Johnston[12], Barbara N. Pusey[12], David R. Adams[12], Valerie V. Maduro[12], May Christine V. Malicdan[12], Mariska Davids[12], Tyra Estwick[12], Donna Novacic[12], Prashant Sharma[12], Camilo Toro[12], Guoyun Yu[12], Babak Behnam[12], Precilla D'Souza[12], Carlos Ferreira[12], Marie Morimoto[12], Eva H. Baker[12], John Yang[12], Jean-Philippe F. Gourdine[13], Matthew Brush[13], Melissa Haendel[13], Euan A. Ashley[14], Jonathan A. Bernstein[14], Jacinda B. Sampson[14], Diane B. Zastrow[14], Noah D. Friedman[14], Jason D. Merker[14], Colleen E. McCormack[14], Paul G.

Fisher[14], Jean M. Davidson[14], Annika M. Dries[14], Gregory M. Enns[14], Marta M. Majcherska[14], Chloe M. Reuter[14], Daryl M. Waggott[14], Jennefer N. Kohler[14], Terra R. Coakley[14], Kevin S. Smith[14], Matthew T. Wheeler[14], Devon Bonner[14], Liliana Fernandez[14], Jason Hom[14], Yong Huang[14], Shruti Marwaha[14], Chunli Zhao[14], Julian A. Martínez-Agosto[15], Esteban C. Dell'Angelica[15], Jeanette C. Papp[15], Emilie D. Douine[15], Stan F. Nelson[15], Martin G. Martin[15], Christina GS. Palmer[15], Neil H. Parker[15], Manish J. Butte[15], Amanda J. Yoon[15], Sandra K. Loo[15], Brent L. Fogel[15], Katrina M. Dipple[15], Janet S. Sinsheimer[15], Patrick Allard[15], Hayk Barseghyan[15], Naghmeh Dorrani[15], Hane Lee[15], Eric Vilain[15], Ascia Eskin[15], Genecee Renteria[15], Rebecca Signer[15], Jijun Wan[15], Allison Zheng[15], Monte Westerfield[16], John A. Phillips III[17], Joy D. Cogan[17], John H. Newman[17], Amy K. Robertson[17], Rizwan Hamid[17], Anna Bican[17], Elly Brokamp[17], Laura Duncan[17], Mary Kozuira[17], Lynette Rives[17] & Lisa Shakachite[17]

[4]Baylor College of Medicine, Houston, Texas, USA. [5]Boston Children's Hospital, Boston, Massachusetts, USA. [6]Brigham and Women's Hospital, Boston, Massachusetts, USA. [7]Columbia University, New York City, New York, USA. [8]Duke University Medical Center, Durham, North Carolina, USA. [9]Harvard Medical School, Boston, Massachusetts, USA. [10]HudsonAlpha Institute for Biotechnology, Huntsville, Alabama, USA. [11]Massachusetts General Hospital, Boston, Massachusetts, USA. [12]National Institutes of Health, Bethesda, Maryland, USA. [13]Oregon Health and Science University, Portland, Oregon, USA. [14]Stanford University, Stanford, California, USA. [15]University of California Los Angeles, Los Angeles, California, USA. [16]University of Oregon, Eugene, Oregon, USA. [17]Vanderbilt University Medical Center, Nashville, Tennessee, USA.