ORIGINAL RESEARCH ARTICLE | Genetics inMedicine

The National Institutes of Health Undiagnosed Diseases Program: insights into rare diseases

William A. Gahl, MD, PhD[1–3], Thomas C. Markello, MD, PhD[2], Camilo Toro, MD[1], Karin Fuentes Fajardo, BS[1], Murat Sincan, MD[3], Fred Gill, MD[4], Hannah Carlson-Donohoe, BA[3], Andrea Gropman, MD[2,5], Tyler Mark Pierson, MD, PhD[1,6], Gretchen Golas, MS, CRNP[2,7], Lynne Wolfe, MS, CRNP[1], Catherine Groden, MS, CRNP[1,2], Rena Godfrey, PA[1], Michele Nehrebecky, MS, CRNP[1], Colleen Wahl, MS, CRNP[1], Dennis M.D. Landis, MD[1], Sandra Yang, MS[1,2], Anne Madeo, MS[8], James C. Mullikin, PhD[9], for the NISC Comparative Sequencing Program, Cornelius F. Boerkoel, MD, PhD[1], Cynthia J. Tifft, MD, PhD[1,2] and David Adams, MD, PhD[1,3]

Purpose: This report describes the National Institutes of Health Undiagnosed Diseases Program, details the Program's application of genomic technology to establish diagnoses, and details the Program's success rate during its first 2 years.

Methods: Each accepted study participant was extensively phenotyped. A subset of participants and selected family members (29 patients and 78 unaffected family members) was subjected to an integrated set of genomic analyses including high-density single-nucleotide polymorphism arrays and whole exome or genome analysis.

Results: Of 1,191 medical records reviewed, 326 patients were accepted and 160 were admitted directly to the National Institutes of Health Clinical Center on the Undiagnosed Diseases Program service. Of those, 47% were children, 55% were females, and 53% had neurologic disorders. Diagnoses were reached on 39 participants (24%) on clinical, biochemical, pathologic, or molecular grounds; 21 diagnoses involved rare or ultra-rare diseases. Three disorders were diagnosed based on single-nucleotide polymorphism array analysis and three others using whole exome sequencing and filtering of variants. Two new disorders were discovered. Analysis of the single-nucleotide polymorphism array study cohort revealed that large stretches of homozygosity were more common in affected participants relative to controls.

Conclusion: The National Institutes of Health Undiagnosed Diseases Program addresses an unmet need, i.e., the diagnosis of patients with complex, multisystem disorders. It may serve as a model for the clinical application of emerging genomic technologies and is providing insights into the characteristics of diseases that remain undiagnosed after extensive clinical workup.

Genet Med 2012:14(1):51–59

Key Words: neurological disorders; rare disease; SNP arrays; undiagnosed disease; whole exome sequencing

INTRODUCTION

In the past decade, both the Office of Rare Diseases, now the Office of Rare Diseases Research (ORDR), and the Genetic and Rare Diseases Information Center recognized how difficult it was for patients with rare diseases to reach an accurate diagnosis.[1] Hence, in early 2008, the ORDR designated $280,000 to foster an initiative to investigate individuals with undiagnosed disorders. National Institutes of Health (NIH) and center directors approved of this concept, namely the creation of a "mystery disease" clinic reminiscent of the early years of the NIH intramural program. Consequently, the NIH Undiagnosed Diseases Program (UDP) was announced to approximately 90 patient advocacy groups and 25 news agencies on May 19, 2008. The explicit goals of the UDP were to achieve a diagnosis for patients who remained undiagnosed after an exhaustive workup and to discover new disorders that would provide insight into mechanisms of disease. An administrative assistant and two nurse practitioners were hired to coordinate admissions and clinical evaluations. The Program's popularity and publicity grew rapidly, prompting an increase in funding to $1.9M in fiscal year 2009 and $3.5M annually for fiscal year 2010–2012.

The emergence and growth of the UDP as a clinical outreach effort has coincided with the rapid expansion and application of genomic methods such as high-density single-nucleotide polymorphism (SNP) arrays and affordable whole exome sequencing (WES) and whole genome sequencing. The UDP provides an opportunity to combine data from extensive inpatient phenotyping with genomic analyses of participants and genetically informative immediate family members. This report describes the findings of the UDP during its first 2 years.

[1]NIH Undiagnosed Diseases Program, NIH, Bethesda, Maryland, USA; [2]Office of the Clinical Director, National Human Genome Research Institute, NIH, Bethesda, Maryland, USA; [3]Medical Genetics Branch, National Human Genome Research Institute, NIH, Bethesda, Maryland, USA; [4]NIH Clinical Center, NIH, Bethesda, Maryland, USA; [5]Department of Neurology, Children's National Medical Center, Washington, DC, USA; [6]Neurogenetics Branch, National Institute of Neurological Disorders and Stroke, NIH, Bethesda, Maryland, USA; [7]Office of Rare Diseases Research, Office of the Director, NIH, Bethesda, Maryland, USA; [8]Social and Behavioral Research Branch, Office of Rare Diseases Research, Office of the Director, NIH, Bethesda, Maryland, USA; [9]NIH Intramural Sequencing Center, National Human Genome Research Institute, NIH, Bethesda, Maryland, USA.
Correspondence: David Adams (david.adams@nih.gov)

Submitted 13 June 2011; accepted 12 August 2011; advance online publication 26 September 2011. doi:10.1038/gim.0b013e318232a005

MATERIALS AND METHODS

Patients

Individuals admitted to the NIH Clinical Center (NIH-CC) were enrolled in protocol 76-HG-0238, "Diagnosis and Treatment of Patients with Inborn Errors of Metabolism or Other Genetic Disorders," approved by the National Human Genome Research Institute (NHGRI) Institutional Review Board. Patients and/or their families gave written, informed consent.

Operations

The application process for the UDP is generally patient initiated and requires a summary letter from the referring healthcare provider, who agrees to assume continuity of care for the patient after a potential NIH-CC admission. In addition, complete medical records are requested, including pertinent imaging and histological slides of biopsy material. A website (http://rarediseases.info.nih.gov/Resources/Rare_Diseases_Information.aspx) was established to explain these requirements. The UDP sends an acknowledgment of receipt of materials to the referring physician and the patient and creates an entry in a medical database.

The two UDP directors, specializing in either adult or pediatric cases, triage records to 1–5 consultants. These specialists may identify patients as suitable for their own clinics and protocols, in which case the patient receives pertinent contact information. Otherwise, the reviews are collated and the director reaches a disposition and sends an acceptance/rejection letter to the physician and patient. Suggestions for management or further pursuit of a diagnosis may be included in a rejection letter. Selection of patients for acceptance into the program is based on whether the clinical or laboratory findings of the patient would likely provide a path forward for research. Generally, cases are deferred or rejected when there has been insufficient workup or when a highly likely and testable diagnosis can be pursued locally. Cases where genetic causation seems likely are prioritized.

Patients accepted to the UDP service are scheduled for a week-long inpatient NIH-CC admission. Adult and pediatric UDP teams meet approximately weekly to triage records, follow-up on information requests, update the patient database, and sign disposition letters. A monthly UDP board meeting, complete with photographs, videos, radiographic images, and pathology slides, provides an NIH-wide forum for the discussion of several cases with more than 60 healthcare professionals participating.

Over the first 2 years of its existence, the UDP grew to employ 12 healthcare professionals and support staff. The UDP director and pediatric UDP director are the Clinical Director and Deputy Clinical Director of NHGRI, respectively. Salaried UDP physicians included a full-time adult neurologist, a part-time pediatrician-geneticist, and a part-time pediatric neurologist. Additionally, four full-time nurse practitioners (including two certified in pediatrics and two certified in family practice), a half-time nurse practitioner, a physician's assistant, and a research nurse comprised clinical UDP staff. Support staff included schedulers and administrative assistants. Several other professionals, including an NIH-CC internist, contract neurologists, and technicians received partial salary support from the UDP. More importantly, many NIH-CC specialists donated their time for the review of medical records. These consultants covered the fields of dermatology, pathology, ophthalmology, radiology, virology, hematology, cardiology, pulmonology, rheumatology, laboratory medicine, immunology, infectious diseases, endocrinology, hepatology, nephrology, dentistry, pharmacology, psychiatry, neurology, pain and palliative care, internal medicine, pediatrics, genetics, physiatry, sarcoidosis, metabolic bone disorders, and otolaryngology.

The UDP is supported by the ORDR in the NIH Office of the Director, the NIH-CC, and the NHGRI, which provides all administrative support. Governance, initiated in early 2010, includes an Internal Advisory Board and an External Advisory Board. The internal committee comprised intramural experts and officials, who met quarterly to provide operational advice, and reports to the NIH Director of Intramural Research. The external board consisted of two members of the NHGRI Board of Scientific Counselors and other national experts in medicine and genetics, who met biannually to provide evaluation and oversight, and reports to the NHGRI Director of Intramural Research.

DNA samples

Genomic DNA was extracted from peripheral whole blood using the Gentra Puregene Blood kit (Qiagen, Valencia, CA) per the manufacturer's instructions. For WES, an additional chloroform/phenol extraction step was performed to neutralize infectious agents.

SNP analysis

The NHGRI Genomics Core laboratory performed SNP determinations using the Illumina Bead Array Platform (1MDuo and OmniQuad1M arrays, Illumina, San Diego, CA), which identified specific polymorphisms every 3,000 bases on average. Genotype call rates were more than 99.7% for all samples analyzed. Genome-wide fluorescent intensities and genotype calls, as well as the genotype specific fluorescent intensities, were analyzed using Bead Studio and Genome Studio (Illumina).

Analyses of copy number variations were generated using PennCNV software[2] and then manually validated by visual inspection using Genome Studio.[3] Signal intensities revealed copy number variants (CNVs), i.e., single or double copy deletions or (triallelic) duplications.

Chromosomal mosaicism can be detected by SNP chip data analysis.[4] We developed a method for the accurate and precise quantitation of mosaicism levels. The method was adapted from a continuous distribution function deconvolution technique described by Wampler.[5] Uniparental disomy analysis made use of Genome Studio, wherein Boolean filters were applied to SNP genotypes from multiple individuals within the proband's immediate family.

Multiple small regions of contiguous homozygosity (RCH) are an expected feature of any human genome. Their length and number depend on population history, local recombination rates, and consanguinity.[6] Large RCHs, in contrast, are potential

candidate regions for disease-causing genes and are the basis for homozygosity mapping. We developed a method for determining the upper limit of expected RCH length for any given region of the genome as follows. RCHs were measured across the entire genome. The genome was divided into 1 Mb bins to allow the estimation of local, rather than genome wide, RCH characteristics. The distribution of RCH lengths was determined within each bin. As expected, the frequency of an RCH was inversely proportional to its length. The length distribution within each bin varied across the genome in a manner that correlated with local recombination rates. An empiric upper threshold was determined for RCH lengths within each bin using a control population; the thresholds were then applied to the patient cohort. The threshold for each bin used a set length (2 Mb), or number of SNPs (1,000), normalized to a scalar metric derived from the distribution of RCH lengths within the given bin. The metric was defined as the negative slope of a line fitted to the log-log plot of RCH frequency versus RCH length within a bin. The line was fitted by linear regression. A large slope value, for example, would be expected for a bin containing (or within) a hotspot for recombination. The threshold for identifying an RCH as anomalously large, therefore, would be expected to be low in a bin with a high recombination rate and vice versa. Our working definition of an anomalous RCH was set at 1, 2, or 3 times the calculated threshold for the bin.

The ENT software program and/or the Genome Studio Boolean search facilities were used to determine sites of recombination within single-family pedigrees. Segregation analysis was used to identify regions of the genome that had been inherited in a manner consistent with genetic models being considered for each family.

Whole genome and whole exome analyses

Whole genome sequencing or WES was performed on probands and genetically informative family members by the NIH Intramural Sequencing Center.

Paired end whole genome sample preparation (short insert), flow cell preparation, and massively parallel reversible terminator-based sequencing were performed using the Illumina (San Diego, CA) HiSeq 2000 and GaIIx instruments per the manufacturer's recommendations.[7] For each sequencing run, 100-bp paired end read lengths were used. Approximately 16 lanes of GAIIx (50Gb configuration) or eight lanes of HiSeq 2000 sequence data were enough to create a high-quality alignment of the genomic sequence.

For WES, solution hybridization exome capture was carried out using the Sureselect Human All Exon System (Agilent Technologies, Santa Clara, CA). This technique uses biotinylated RNA baits to hybridize to sequences that correspond to exons.[8] The manufacturer's protocol version 1.0, compatible with Illumina paired end sequencing, was used, with the exception that DNA fragment size and quality were measured using a 2% agarose gel stained with Sybr Gold instead of using an Agilent Bioanalyser. The capture regions total approximately 38 Mb or approximately 50 Mb. This kit covers the 1.22% of the human genome corresponding to the Consensus Conserved Domain Sequences database and >1,000 noncoding RNAs.

Flow cell preparation and 76-bp paired end read sequencing were carried out as per protocol for the GAIIx sequencer (Illumina). Approximately two lanes on a GAIIx flow cell were used per exome sample to generate sufficient reads to produce the coverage needed for high-quality aligned sequence. The quality of coverage must meet a threshold, whereby the percentage of Consensus Coding Sequence exomic bases with most probable genotype quality scores >10 exceeds 85.[9] A most probable genotype score of 10 co-occurs with average coverage in the 10×–20× range but more importantly is roughly equivalent to a Phred Sanger sequencing quality score of 28 or 1 expected error per 500 base calls.[9]

Variation filtering

Data manipulation (filtering) software. Data manipulation and filtering steps were carried out using the VarSifter software developed by Jamie Teer (unpublished data, software available at http://research.nhgri.nih.gov/software/VarSifter/).

Public databases of previously described variants. Common reported variants were identified using dbSNP 130, dbSNP 131, and the 1,000 genomes project data as they became available.[10,11] Once available, all data were analyzed by a filter comprising all 1,000 genome SNPs with an average heterozygosity >1%.

Mendelian consistency filtering. The acquisition of sequences from multiple family members allowed for the filtering of variants that did not segregate in a manner consistent with a Mendelian and/or genetic model.

Pathogenicity. Pathogenicity of individual variants was assessed using CDPred.[12] CDPred assigns a numeric score to each variation that can be aligned to a residue in the National Center for Biotechnology Information Conserved Domain database.[13] Scores for unaligned bases were assigned using the BLOSSUM62 scoring matrix.

Empiric data from within cohort. As data accumulated, a set of highly polymorphic genes was identified. Examples included olfactory receptor and human leukocyte antigen genes. Additional variations occurred in multiple families with dissimilar phenotypes. Both types of variants were used to filter candidate variant lists.

RESULTS

Demographics

Over the initial 2-year period of its existence, the UDP received 2,990 inquiries, and medical records were submitted for 1,191. The age of applicants to the UDP ranged from the first year of life to >70 years; 60% were females. Of the 1,191 applicants, 699 (59%) were rejected; 44 (6%) of the rejection letters held open the possibility of reconsideration, and 183 (26%) offered some advice for further evaluation or referral. The Program accepted

326 individuals; of these, 242 were admitted to the UDP service, and the remaining 84 were seen by other NIH services. Of the UDP admissions, 160 completed their NIH-CC admission, and 66 were scheduled to be admitted later. Although 16 patients were unable to travel to the NIH-CC for clinical phenotyping, their DNA was received and analyzed. DNA was also collected on 293 family members of accepted patients; each was enrolled in the protocol, consented, and assigned a UDP identification number.

Of the 310 (84 + 160 + 66) UDP patients admitted to the NIH-CC, 122 (39%) were children, i.e., <18 years of age. The acceptance rate for children was 122 of 257 or 47%, compared with an acceptance rate of 188 of 934 or 20% for adults. Of the accepted patients, 55% were females.

Deaths

After the initial 32 months of the Program, 23 of the 2,990 UDP patients had died. Five died before their complete medical records were received, 13 died while their records were being evaluated or while awaiting admission, and five died after evaluation at the NIH-CC.

Phenotypic categories

By far the most common phenotypic category of both pediatric and adult cases was neurologic disorders, encompassing 43% of applicants and 53% of accepted patients; the numbers of males and females were equivalent (**Table 1**). Other prominent categories included gastrointestinal disease, fibromyalgia and chronic fatigue syndrome, immunology, rheumatology (including connective tissue disorders), psychiatry (including 33 cases fitting the definition of Morgellon disease), pain, dermatology, and cardiovascular disease. In assigning categories, connective tissue disorders were considered rheumatologic. In the applicant group, six cases of mal de debarquement syndrome and eight cases of severe sound sensitivity were included in the neurology category.

Genetic analysis

Of the 160 patients admitted to the UDP service at the NIH-CC, 139 had DNA analysis involving million SNP arrays, using either the 1M Duo or the Omni-Quad chips. Genetic material was also collected from 141 unaffected family members of the UDP patients and subjected to SNP analysis, providing a control group for comparison with the patient group. Regardless of the type of SNP chip used, there were no differences in CNVs (i.e., double deletions, single deletions, or duplications) between the two groups, specifically in number or size (**Table 2**). However, the patient group exhibited many more runs of anomalous contiguous homozygosity, regardless of whether the criterion for such a "run" was one, two, or three thresholds above the norm for that million base pair segment of the human genome (**Table 3**).

The lack of a difference in the average sizes or numbers of SNPs between probands and closely related unaffected individuals is a direct reflection of the fact that many small CNVs can be found in all individuals. Unlike copy number variations visible in karyotypes, or even those seen in clinical level CGH arrays, the majority of small variants do not impinge on coding sequence. Large and/or known clinically significant CNV are likely underrepresented in the UDP cohort as they are often cited as explanatory diagnoses in practice. We observed CNVs ranging in size from three to many hundreds of SNPs in a row,

Table 1 Primary phenotypes of UDP applicants and accepted patients

Primary phenotype	Records received			Accepted patients		
	M	F	Total	M	F	Total
Allergy	2	10	12	0	1	1
Cardiovascular	18	22	40	12	10	22
Dermatology	20	27	47	5	3	8
Endocrine	12	21	33	8	5	13
ENT	1	1	2	0	0	0
Fibromyalgia/CFS	21	58	79	1	2	3
Gastrointestinal	36	70	106	2	11	13
Gynecology	0	8	8	0	3	3
Hematology	5	18	23	1	6	7
Infectious disease	10	8	18	2	1	3
Immunology	31	32	63	6	9	15
Metabolic	5	4	9	2	1	3
Neurology	236	276	512	78	86	164
Oncology	3	9	12	0	2	2
Ophthalmology	8	3	11	3	1	4
Orthopedics	3	6	9	1	1	2
Pain	14	37	51	0	0	0
Psychiatry	19	34	53	0	4	4
Pulmonary	12	16	28	7	7	14
Renal	8	11	19	2	8	10
Rheumatology	16	40	56	7	12	19
Total	480	711	1,191	137	173	310

UDP, Undiagnosed Diseases Program.

Table 2 SNP analyses for UDP patients and controls: copy number variants

	Double deletions			Single deletions			Duplications		
	#	SNPs	kb	#	SNPs	kb	#	SNPs	kb
1M Duo[a]									
Controls[b] (*N* = 52)	3	12	26	23	17	52	15	21	70
Probands (*N* = 57)	3	10	21	28	17	39	18	26	71
OmniQuad[a]									
Controls (*N* = 89)	55	13	4	130	15	19	40	16	32
Probands (*N* = 82)	59	13	4	150	13	23	39	17	30

The number sign ("#") indicates the average number of regions per patient; SNPs, average number of SNPs within a region; kb, average size of each.

SNP, single-nucleotide polymorphism; UDP, Undiagnosed Diseases Program.

[a]Two different SNP chips were used. The 1M Duo provides ~50 PennCNV calls, whereas the OmniQuad chip provides ~240 PennCNV calls. [b]Controls were unaffected UDP family members; three controls were excluded from the 1M Duo chip analysis because of poor quality chip result, i.e., a call rate <99%.

Table 3 SNP analyses for UDP patients and controls: regions of anomalous continuous homozygosity (no. per 100 individuals)

	Controls (*N* = 145)	Probands (*N* = 128)
SNP thresholds[a]		
>1	56	101
>2	12	28
>3	4	12
kb thresholds		
>1	414	691
>2	61	89
>3	21	36

Controls were unaffected family members of UDP probands. Of 145 controls, 60 were analyzed by the 1M Duo chips and 85 using the OmniQuad chips; of 128 UDP patients, 53 were analyzed by the 1M Duo chips and 75 using the OmniQuad chips.

SNP, single-nucleotide polymorphism; UDP, Undiagnosed Diseases Program.

[a]Thresholds were determined separately for number of continuously homozygous SNPs and for number of continuously homozygous base pairs, as described in "Methods." Numbers of regions exceeding 1, 2, and 3 normal thresholds were normalized to 100 individuals for comparison of controls and UDP patients.

and from 10 kb to several megabases for all CNV categories, and for both probands and controls.

Thirty-two affected individuals from 26 UDP families had whole genome sequencing (3) or WES (29); DNA from 78 unaffected family members was also submitted for WES. An average family (4.2 members) had a mean of 134,634 variants identified.

Diagnoses

Most UDP diagnoses were known conditions. The reasons that the diagnoses had not been made previously were diverse; for example, rare and ultra-rare conditions may not have been considered, prior laboratory data may have been misleading, and newer tests and refined disease definitions may have emerged.

Of the 160 patients (65 males and 95 females) evaluated at the NIH-CC on the UDP service, 39 (15 males and 24 females), including seven children, received a diagnosis (**Table 4**). For 12 the diagnosis was based solely on clinical findings, and for 19 the diagnosis was based on molecular findings. The diagnoses included two undescribed disorders (in separate patients), 23 rare or ultrarare diseases (in 28 patients), and nine common

Table 4 UDP diagnoses

Frequency	UDP no.	Age (years)	Sex	Diagnosis	Basis	Comment
New diseases	797	54	F	ACDC[a]	C, P, B, M^S	*NT5E* mutations
	1,103	51	M	ACDC	C, M^S	*NT5E* mutations
	1,112	49	F	ACDC	C, P, B, M^S	*NT5E* mutations
	1,433	44	F	ACDC	C, M^S	*NT5E* mutations
	1,889	53	M	ACDC	C, M^S	*NT5E* mutations
	2,457	44	F	ACDC	C, B, P, M^T	*NT5E* mutations
	1,706	44	F	Familial distal myopathy	C, P, M^E	*HINT3* mutation
<60 cases reported	283	48	F	Leukodystrophy with axonal spheroids	C, P	
	338	14	M	Spinocerebellar ataxia and hereditary spastic paraplegia	C, P, B, M^E	Only case of biallelic *AFG3L2* mutations
	499	7	F	Pitt-Hopkins syndrome	C, M^T	*TCF4* mutation
	563	3	M	Hereditary benign intraepithelial dyskeratosis	C, P, M^S	4q35.2 duplication
	887	5	F	Congenital disorder of glycosylation IIb	C, B, M^T	Glucosidase I deficiency; sib of 1,248
	1,173	48	M	Autosomal dominant cerebellar ataxia	C, P, M^S	*LMNB1* duplication
	1,248	10	M	Congenital disorder of glycosylation IIb	C, B, M^T	Glucosidase I deficiency
	2,226	38	F	Aceruloplasminemia	C, B, M^T	*Cp* mutations; neurological involvement
<1/100,000	333	45	F	Facial dysautonomia	C	
	357	36	M	Hereditary spastic paraplegia	C, M^T	*SPG4* mutations
	608	4	F	Smith-Magenis syndrome	C, M^T	*RAI1* mutation
	679	31	F	CSF tetrahydro-biopterin deficiency	C, B	Incidental to devastating cerebral deficits
	752	41	F	Immune-mediated cerebellar degeneration	C, P	Responded to rituximab

If a diagnosis was already suggested by the referring center, confirming that diagnosis was not sufficient to claim success.

B, biochemical; C, clinical; CSF, cerebrospinal fluid; M^E, molecular whole exome; M^S, molecular SNP; M^T, molecular-targeted sequencing; P, pathological; SNP, single-nucleotide polymorphism; UDP, Undiagnosed Diseases Program.

[a]ACDC, arterial calcifications due to CD73 deficiency; cases 1,112, 797, 1,433, 1,103, and 1,889 are sibs.

Table 4 Continued on next page

Table 4 Continued.

Frequency	UDP no.	Age (years)	Sex	Diagnosis	Basis	Comment
	984	17	F	GM1 gangliosidosis	C, B, M[E]	Normal initial enzyme activity
	997	53	F	Amyloid myopathy	C, B, P	Multiple myeloma
	1,074	45	M	Amyotrophic lateral sclerosis	C, M[T]	*SOD1* mutation found
	1,262	46	M	Progressive spastic paraparesis	C, M[T]	*SPG7* mutations
	1,857	24	F	CSF tetrahydro-biopterin deficiency	C, B	Responded to treatment
	1,924	36	F	Call-Fleming syndrome	C, P	
1–10/100,000	714	40	M	Primary progressive multiple sclerosis	C, P	
	1,155	29	M	Neuromyelitis optica	C	
	2,019	69	M	Progressive supranuclear palsy and corticobasal ganglia degeneration	C	
	2,566	57	M	Corticobasal ganglionic degeneration	C	
Common	368	35	F	Fibromyalgia	C	
	800	36	F	Fibromyalgia	C	
	855	56	F	Psychogenic tic cough	C	
	932	44	F	Somatization	C	
	936	47	M	Morgellon disease	C	
	1,137	54	F	Multiple myeloma	C, P	
	1,628	54	F	Functional gait disorder	C	
	1,868	19	F	Psychogenic movement disorder	C	
	1,913	36	M	Fibromyalgia	C	

If a diagnosis was already suggested by the referring center, confirming that diagnosis was not sufficient to claim success.

B, biochemical; C, clinical; CSF, cerebrospinal fluid; M[E], molecular whole exome; M[S], molecular SNP; M[T], molecular-targeted sequencing; P, pathological; SNP, single-nucleotide polymorphism; UDP, Undiagnosed Diseases Program.

[a]ACDC, arterial calcifications due to CD73 deficiency; cases 1,112, 797, 1,433, 1,103, and 1,889 are sibs.

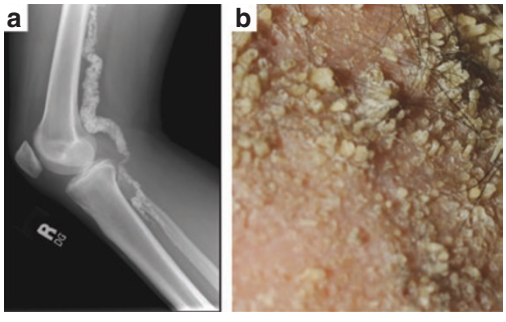

Figure 1 Descriptive information on unusual UDP cases. (**a**) Plain radiograph of femoral and popliteal arteries of UDP797, with arterial calcification due to deficiency of CD73. (**b**) Keratin spicules protruding from the scalp hair follicles of UDP441, a 50-year-old woman.

disorders (five psychogenic diagnoses, three fibromyalgia, and one multiple myeloma).

One new disease discovered was arterial calcification due to deficiency of CD73 resulting from biallelic mutations in *NT5E* (**Figure 1a**; OMIM #211800).[14] CD73 is an enzyme present on vascular cell plasma membranes that produces adenosine and inorganic phosphate from AMP.[15] The second undescribed disease was a familial distal myopathy in which WES identified a single mutation in *HINT3*, an aprataxin-related gene (**Table 5**).[16] Ultrarare disorders, which we defined as reported in <60 individuals, were diagnosed in eight individuals. Two siblings had the congenital disorder of glycosylation type IIb (OMIM# 606056), which had only been reported in one other individual.[17] Two brothers had spinocerebellar ataxia and

Table 5 Whole exome sequence filtering data for three UDP families

Family	Operation	Cumulative residual variants[a]
UDP 0337	None	120,469
	1,000 genome 1%	59,482
	SNP chip-based family linkage file[b]	53,087
	Kill file[c]	50,467
	Homozygous recessive	11
	Homozygous recessive + not in db130[d]	3[e]
	Compound heterozygous[f]	121
	Compound heterozygous + not in db130	71
UDP 0984	None	135,292
	1,000 genome 1%	68,200
	SNP chip-based family linkage file[b]	52,426
	Kill file[c]	48,532
	Homozygous recessive	39
	Homozygous recessive + not in db130[b]	8
	Compound heterozygous[f]	286
	Compound heterozygous + not in db130	143
UDP1706	None	122,195
	1,000 genome 1%	52,683
	SNP chip-based family linkage file[b]	28,359
	Kill file[c]	28,359
	Dominant	1,675
	Dominant + not in db130	872
	Dominant + not in db130 + deleterious coding	64

These lists can be further ranked or sorted by several means, e.g., by predicted pathogenicity.

SNP, single-nucleotide polymorphism; UDP, Undiagnosed Diseases Program.

[a]Mutation types include frameshifting indels, nonsynonymous missense mutations, canonical splice site mutations, and stop mutations. [b]The SNP chip-based family linkage file can have a much larger impact than seen for these two families, depending on where recombinations occurred and how many participants with informative meioses are available. [c]The Kill file includes genes in the human leukocyte antigen region, very highly polymorphic genes such as the mucin genes, genes that code for the taste and olfactory receptors, and known pseudogenes with open coding regions. [d]Using the "not in db130 filter" needs to be done with caution because it contains pathogenic variants. [e]Includes *AFG3L2*. [f]For compound heterozygotes, the filter includes any combination of variants that follow a predicted Mendelian pattern. In practice, however, one of the two variants identified will be intronic, 3′ UTR, etc. Therefore, one can look at only the genes with two "good" mutations. For example, for UDP 0984, the list of 143 variants only represents 17 genes with two mutations of the type surveyed. One of those 17 was the responsible gene, *GLB1*.

hereditary spastic paraplegia due to homozygous mutations in *AFG3L2* (**Table 5**) (spinocerebellar ataxia 28; OMIM #610246); *AFG3L2* mutations had been associated with spinocerebellar ataxia[18] and hereditary spastic paraplegia[19] separately but never with both in the same individual. A 48-year-old man was diagnosed with adult-onset autosomal dominant leukodystrophy (OMIM #169500),[20] due to duplication of *LMNB1* on chromosome 5q; *LMNB1* encodes lamin B1, a protein of the inner nuclear membrane.

Rare conditions, which we defined as occurring in fewer than one in 10,000 individuals, were diagnosed in 15. These included a 53-year-old woman with massively increased muscle bulk from amyloid myopathy due to multiple myeloma and clinical diagnoses of neuromyelitis optica, Call-Fleming syndrome, progressive supranuclear palsy and corticobasal ganglion degeneration, leukodystrophy with axonal spheroids (OMIM #221790), facial dysautonomia, and corticobasal ganglion degeneration and molecular diagnoses of hereditary spastic paraplegia type 7 (OMIM #607259) and type 4 (OMIM #182601),[21,22] amyotrophic lateral sclerosis (OMIM #105400) due to an *SOD1* mutation, [23] Smith-Magenis syndrome (OMIM #182290) due to an *RAI1* mutation,[24] GM1 gangliosidosis (OMIM #230500), aceruloplasminemia (OMIM #604290),[25] Pitt-Hopkins syndrome (OMIM #610954),[26] and hereditary benign intraepithelial dyskeratosis (OMIM #127600).[27]

SNP arrays yielded three diagnoses, i.e., arterial calcification due to deficiency of CD73, autosomal dominant leukodystrophy, and hereditary benign intraepithelial dyskeratosis. WES was responsible for finding the mutations in the *AFG3L2*, *GLB1*, and *HINT3* genes. For *AFG3L2*, the use of a homozygous recessive model allowed for filtering the number of variants from a total of 120,469 to two homozygous recessive variants, one of which was *AFG3L2* (**Table 5**). To find *GLB1*, 135,292 variants were filtered to eight homozygous recessive and 143 compound heterozygous variants involving 17 genes (**Table 5**). For *HINT3*, 122,195 variants were reduced to 64.

In addition to patients admitted to the NIH-CC, one patient was investigated remotely by a UDP staff neurologist. This 20-year-old man was diagnosed with spinal muscular atrophy with respiratory distress-1 due to mutations in *IGHMBP2* (OMIM #604320).[28] He is the oldest known patient with the disorder and is described elsewhere.[29]

Intriguing but unsolved cases included (1) two women with increased circulating vascular endothelial growth factor and either thrombotic microangiopathy or hepatic and bone hemangiomas; (2) a man with renal stones and elevated vitamin D levels; (3) a young woman with fibro-inflammatory tumors of her lungs, liver, and pterygomaxillary region; (4) two women with decreased cerebrospinal fluid tetrahydrobiopterin and neurotransmitter levels who responded to sapropterin supplementation; (5) a child with developmental delays and copper storage in Zone 3 of the liver; (6) a child with idiopathic renal tubular Fanconi syndrome and hearing loss; (7) a woman with lung nodules and thick pulmonary mucus; (8) a woman with a possible pathogenic mutation in platelet-derived growth factor-α; (9) a woman with follicular keratosis producing painful spikes of keratin protruding from her skin and scalp (**Figure 1b**); and (10) a woman with autoimmune-mediated cerebellar degeneration.

DISCUSSION

In its initial 2 years, the NIH UDP processed 1,191 applications from individuals who were predominantly women; the typical patient was aged 30–70 years with neurologic issues. Diagnoses were established in approximately 24% of the 160 cases seen during its first 2 years; ongoing investigations will probably reveal additional diagnoses. Many cases likely represent new diseases. Most of the 39 diagnoses occurred in adults (32/94 vs. 7/66 for children); however, the children had more extensive genetic testing before their UDP evaluation. In addition to diagnosing several common disorders on clinical grounds, the UDP defined two new disorders, identified 21 rare diagnoses in 23 individuals on molecular or biochemical bases, and expanded the phenotype of numerous disorders. Two new diseases were discovered. Several sets of patients, defined by constellations of similar findings, were identified. In analyzing SNP array data compared with a control group, the UDP patients had more regions of homozygosity but the same number of CNVs.

The UDP also addressed the role that next generation sequencing could play in the diagnosis of complicated, challenging medical conditions. WES (performed on 32 patients) proved critical for the diagnosis of six different disorders. This success rate was for individuals who had already been thoroughly investigated at other academic centers and would have been greater for a less selected population of patients. Our approach of combining thorough phenotyping with next-generation sequencing highlighted several general principles:

1. Mendelian filtering (recombination mapping plus variant segregation analysis) is a powerful tool for filtering genome-wide sequencing data. It requires the inclusion of SNP and WES data from informative family members in addition to the proband. The additional family members often decreased the number of final candidate sequence variants by 1–2 orders of magnitude (unpublished data).
2. Mendelian filtering works best for certain minimum family sizes, which are in turn dependent on a proposed genetic model. The family sizes are smaller than those required for linkage analysis. For example, to set up a filter for an autosomal recessive family, a minimal pedigree would include the proband, one sib and two parents.
3. Filters from databases such as dbSNP should use the subset of data with frequency annotation to reduce the chance of excluding pathogenic variations.
4. In many cases, adequate filtering can reduce the number of candidate genes to 20 or fewer, rendering further analysis manageable.
5. WES can be more economical than wholesale ordering of CLIA-certified single gene sequencing if many gene candidates are being considered. Methodological knowledge and careful data analysis must be used to verify that genes of interest are adequately sequenced.
6. Accurate and meticulous phenotyping of affected family members by physical examination and diagnostic testing is essential to verify and expand details in the medical record of the proband. For example, it is common for our examination to be strikingly different from that portrayed in the medical records. In some cases, the discrepancy is explained by a long interval between the last major medical evaluation and the UDP workup.

To validate the sequencing findings and to enhance the understanding of these patients' disorders, basic research scientists and physician scientists are recruited from outside of the UDP. Contributing within their areas of expertise, they investigate disorders as examples of aberrations in new or poorly delineated metabolic pathways. To enhance this collaboration and further engage the community, the UDP plans to present deidentified unsolved cases on a web-based international portal, accessible to designated world experts.

Future aspirations of the UDP involve expanding research to provide a bridge between the clinical workup and basic scientists with specific expertise related to the patient's disease. This includes evaluation of selected fibroblast cultures for intracellular organelle defects involving morphological abnormalities or vesicular storage and of other fibroblast cultures for mitochondrial defects by measuring oxygen consumption and acid production.[30] Additionally, metabolites of selected body fluids, including cerebrospinal fluid, are being analyzed using nuclear magnetic resonance and mass spectrometry. Plans are also underway to create induced pluripotent stem cells from fibroblasts to investigate disorders of inaccessible tissues such as neurons.

The UDP experience has provided insights into other aspects of healthcare.[31] First, the remarkable percentage of applications with neurologic problems (43%) emphasizes the need for

advances in neurologic diagnostics and therapeutics. Second, regardless of geography, one medical subspecialist had often assumed primary responsibility for complicated patients, leading to a restricted focus on a subset of disease manifestations. Third, the NIH was often the first opportunity for many patients to access a coordinated multidisciplinary evaluation.

In summary, the UDP serves the needs of a very unique population of individuals, to the mutual benefit of the patients and the medical community. Its position at the interface of clinical care and basic science also provides a rare opportunity to integrate research tools into patient solutions.

ACKNOWLEDGMENTS

This work was supported by the Intramural Division of the National Human Genome Research Institute and the National Institute of Neurological Disorders and Stroke, the NIH Clinical Center, the NIH Office of the Director, the Office of Rare Diseases Research (Dr Stephen Groft), the NIH Clinical Center (Dr John Gallin), and the Intramural Research Programs of the National Human Genome Research Institute and the National Institute of Neurological Disorders and Stroke. The authors thank the NIH UDP patients, their families, and their physicians for making this enterprise a truly united effort; Roxanne Fischer and Richard Hess for their excellent technical assistance and the entire UDP staff for their dedicated service; and the NIH Intramural Sequencing Center for performing the whole exome and whole genome sequencing and analysis. The NHGRI Genomics Core provided superb SNP array results.

DISCLOSURE

The authors declare no conflict of interest.

REFERENCES

1. National Technical Information Services. *Report of the National Commission on Orphan Diseases*. National Technical Information Services: Springfield, VA, 1989:17.
2. Wang K, Li M, Hadley D, et al. PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. Genome Res 2007;17:1665–1674.
3. Manoli I, Golas G, Westbroek W, et al. Chediak-Higashi syndrome with early developmental delay resulting from paternal heterodisomy of chromosome 1. Am J Med Genet A 2010;152A:1474–1483.
4. Conlin LK, Spinner N. Cytogenetics into cytogenomics: SNP arrays expand the screening capabilities of genetics laboratories. Illumina Corporation: San Diego, CA, 2008.
5. Wampler JE. Analysis of the probability distribution of small random samples by nonlinear fitting of integrated probabilities. Anal Biochem 1990;186:209–218.
6. Gusev A, Mandoiu II, Pasaniuc B. Highly scalable genotype phasing by entropy minimization. IEEE/ACM Trans Comput Biol Bioinform 2008;5:252–261.
7. Bentley DR, Balasubramanian S, Swerdlow HP, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 2008;456:53–59.
8. Gnirke A, Melnikov A, Maguire J, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol 2009;27:182–189.
9. Teer JK, Bonnycastle LL, Chines PS, et al. Systematic comparison of three genomic enrichment methods for massively parallel DNA sequencing. Genome Res 2010;20:1420–1431.
10. Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 2001;29:308–311.
11. Siva N. 1000 Genomes project. Nat Biotechnol 2008;26:256.
12. Johnston JJ, Teer JK, Cherukuri PF, et al. Massively parallel sequencing of exons on the X chromosome identifies RBM10 as the gene that causes a syndromic form of cleft palate. Am J Hum Genet 2010;86:743–748.
13. Marchler-Bauer A, Lu S, Anderson JB, et al. CDD: a Conserved Domain Database for the functional annotation of proteins. Nucleic Acids Res 2011;39(Database issue):D225–D229.
14. St Hilaire C, Ziegler SG, Markello TC, et al. NT5E mutations and arterial calcifications. N Engl J Med 2011;364:432–442.
15. Colgan SP, Eltzschig HK, Eckle T, Thompson LF. Physiological roles for ecto-5′-nucleotidase (CD73). Purinergic Signal 2006;2:351–360.
16. Brenner C. Hint, Fhit, and GalT: function, structure, evolution, and mechanism of three branches of the histidine triad superfamily of nucleotide hydrolases and transferases. Biochemistry 2002;41:9003–9014.
17. De Praeter CM, Gerwig GJ, Bause E, et al. A novel disorder caused by defective biosynthesis of N-linked oligosaccharides due to glucosidase I deficiency. Am J Hum Genet 2000;66:1744–1756.
18. Di Bella D, Lazzaro F, Brusco A, et al. Mutations in the mitochondrial protease gene AFG3L2 cause dominant hereditary ataxia SCA28. Nat Genet 2010;42:313–321.
19. Bonn F, Pantakani K, Shoukier M, Langer T, Mannan AU. Functional evaluation of paraplegin mutations by a yeast complementation assay. Hum Mutat 2010;31:617–621.
20. Padiath QS, Saigoh K, Schiffmann R, et al. Lamin B1 duplications cause autosomal dominant leukodystrophy. Nat Genet 2006;38:1114–1123.
21. Brugman F, Scheffer H, Wokke JH, et al. Paraplegin mutations in sporadic adult-onset upper motor neuron syndromes. Neurology 2008;71:1500–1505.
22. de Bot ST, van den Elzen RT, Mensenkamp AR, et al. Hereditary spastic paraplegia due to SPAST mutations in 151 Dutch patients: new clinical aspects and 27 novel mutations. J Neurol Neurosurg Psychiatr 2010;81:1073–1078.
23. Millecamps S, Salachas F, Cazeneuve C, et al. SOD1, ANG, VAPB, TARDBP, and FUS mutations in familial amyotrophic lateral sclerosis: genotype-phenotype correlations. J Med Genet 2010;47:554–560.
24. Bi W, Saifi GM, Shaw CJ, et al. Mutations of RAI1, a PHD-containing protein, in nondeletion patients with Smith-Magenis syndrome. Hum Genet 2004;115:515–524.
25. Hellman NE, Gitlin JD. Ceruloplasmin metabolism and function. Annu Rev Nutr 2002;22:439–458.
26. Zweier C, Peippo MM, Hoyer J, et al. Haploinsufficiency of TCF4 causes syndromal mental retardation with intermittent hyperventilation (Pitt-Hopkins syndrome). Am J Hum Genet 2007;80:994–1001.
27. Allingham RR, Seo B, Rampersaud E, et al. A duplication in chromosome 4q35 is associated with hereditary benign intraepithelial dyskeratosis. Am J Hum Genet 2001;68:491–494.
28. Grohmann K, Schuelke M, Diers A, et al. Mutations in the gene encoding immunoglobulin mu-binding protein 2 cause spinal muscular atrophy with respiratory distress type 1. Nat Genet 2001;29:75–77.
29. Pierson TM, Tart G, Adams D, et al. Infantile-onset spinal muscular atrophy with respiratory distress-1 diagnosed in a 20-year-old man. Neuromuscul Disord 2011;21:353–355.
30. Ferrick DA, Neilson A, Beeson C. Advances in measuring cellular bioenergetics using extracellular flux. Drug Discov Today 2008;13:268–274.
31. Gahl WA, Tifft CJ. The NIH Undiagnosed Diseases Program: lessons learned. JAMA 2011;305:1904–1905.