The undiagnosed diseases program approach to diagnosis

Ellen F. Macnamaraa,*, and Precilla D’Souzaa, Undiagnosed Diseases Network, and Cynthia J. Tiffta,b

aNational Institutes of Health, Undiagnosed Diseases Program, Common Fund, Office of the Director, Bethesda, MD, USA
bOffice of the Clinical Director, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, USA

Abstract. Undiagnosed and rare conditions are collectively common and affect millions of people worldwide. The NIH Undiagnosed Diseases Program (UDP) strives to achieve both a comprehensive diagnosis and a better understanding of the mechanisms of disease for many of these individuals. Through the careful review of records, a well-orchestrated inpatient evaluation, genomic sequencing and testing, and with the use of emerging strategies such as matchmaking programs, the UDP succeeds nearly 30 percent of the time for these highly selective cases. Although the UDP process is built on a unique set of resources, review of the process, with case examples, demonstrates steps genetic professionals can take in both clinical and research settings, including increased HPO term usage to enhance identification of phenotypes, sequencing when clear objective findings are present, and matchmaking services to connect with others, to arrive at a diagnosis for their most challenging cases.

1. Introduction

The “diagnostic odyssey” is well known in medical genetics. Undiagnosed conditions affect three million Americans and include those with rare, difficult to identify conditions, atypical presentations of known conditions, and diseases yet to be discovered [1]. Patients and their families can spend years on this diagnostic odyssey and never arrive at a diagnosis. The average length of a diagnostic odyssey is eight years [1].

The Undiagnosed Diseases Program (UDP) at the National Institutes of Health (NIH) was established in May 2008. The goals of the UDP were two-fold: 1) to achieve a comprehensive diagnosis for patients who have already undergone an exhaustive evaluation and remain on the odyssey; and 2) to identify new biochemical, physiological, and cell biological pathways to better understand the mechanisms of disease. The many meanings of “diagnosis” include a simple histological description, a specific collection of clinical symptoms, or a genetic mutation. The UDP, however, believes that the “most satisfying definition of diagnosis includes an understanding of the disease pathogenesis, linking genetic and clinical findings and informing prognosis and therapy” [2].

To reach these goals, UDP team members review all available past medical records of applicants. Accepted individuals receive careful and extensive clinical evaluations at the NIH Clinical Center.
The patients and their families then undergo integrated genomic analyses. Following identification of candidate genes, collaborations are established to identify additional patients through matchmaking services, prove causality through in vitro cell studies, and perform functional assays in model systems.

In 2014, the UDP expanded to the Undiagnosed Diseases Network (UDN), and included six additional clinical sites, a coordinating center, two sequencing cores, a metabolomics core, a model organisms core, two sequencing centers, and a central biorepository. The UDN, modeled after the UDP, shares the mission of providing diagnoses and discovering new diseases through wider access to cross-disciplinary expertise and a collaborative network of experts [3]. This paper outlines the strategies employed by the UDP and UDN to achieve diagnoses, the lessons learned, and the application of these methods to clinical genetics.

2. Application and acceptance

Applicants to the UDN apply through an online portal with a referral letter from a clinician [4]. Applicants share medical and surgical records (for pediatric patients, birth records and growth curves are also requested), genetic testing results, imaging studies, and biopsy slides. These records are reviewed by experts in the disease category appropriate for the patient (e.g., neurology, rheumatology, immunology, etc.) and the UDN Board accepts cases with “objective findings, novel phenotypic manifestations, and a high likelihood of obtaining a diagnosis” [5]. Historically, the most informative notes for pediatric applicants have been consultant notes that speak to the patients’ symptomatology, developmental notes that address the patients’ trajectory, newborn records to assist in acquired versus congenital onset, and growth curves. Occasionally, diagnoses are made by careful review of these records; this is most commonly accomplished through receipt of original genetic testing reports that may have been previously misinterpreted.

Since the inception of the UDN, approximately 39% of applicants have been accepted network wide. Applicants have a wide-range of objective and subjective symptoms. Walley, et. al., reviewed 151 “Not Accepted” patients and 50 “Accepted” patients in the UDN to determine what distinguished these groups from one another [6]. On average, accepted patients were younger, had a longer percentage of time with illness and an earlier onset of disease, and were referred by specialists. Neurology was the primary category for symptoms in both the “Accepted” and “Not Accepted” patients, followed by allergy & immunology and musculoskeletal disease.

Making a distinction between undiagnosed individuals with and without objective findings increases the likelihood of successful evaluation in the UDN. Those with objective findings often have a genetic etiology and therefore are more likely to benefit from the resources of the UDN [6].

3. Admission and evaluation

The UDN functions under a single protocol with a common Institutional Review Board and enrolls all patients and consenting family members [7]. Patients accepted to the NIH UDP, one of the UDN clinical sites, are consented and admitted to the NIH Clinical Center, generally for a five-day comprehensive inpatient evaluation. The Clinical Center is supported by intramural NIH funding and patients are enrolled regardless of their ability to pay. This assures that patients are accepted into the UDP on the uniqueness of their clinical case. Alternative approaches across other NIH-funded UDN sites include both inpatient and outpatient models. The cost is borne by a combination of third party payors for clinically-indicated testing and extramural NIH grant funding.

Patients at NIH are seen by a wide range of specialists engaged in a highly collaborative diagnostic approach. The standard pediatric UDP consultations include: genetics, nutrition, neurology,
audiology, ophthalmology, neuropsychiatry, rehabilitative medicine, and physical, speech, and occupational therapy. More than half of the pediatric patients, i.e., 56% between 2008 and 2015 (unpublished observations), received a single 3-4-hour long sedation allowing for neuroimaging and multiple evaluations that were potentially painful or difficult to accomplish in an awake child. Patients had an average of three procedures following imaging; they include lumbar puncture, skin biopsy, and dilated eye exam. Additional procedures often involved auditory evoked response testing, dental evaluation and cleaning, electromyogram and nerve conduction studies, dysmorphology evaluation in an uncooperative patient, catheterization for urine collection, and large blood draws. More than 300 children, many with multisystem disease American Society of Anesthesiology (ASA) Score III and above, have been sedated without complications. The NIH diagnostic evaluations have led to clinical diagnoses, management recommendations, targeted testing, and occasionally treatment.

At the conclusion of the week-long evaluation the proband’s phenotype is recorded in a secure database using standardized terms taken from the Human Phenotype Ontology (HPO) [8]. PhenoTips® software, embedded within the uniquely designed UDPICS database, facilitates entry of HPO terms [9–11]. Since accurately identifying family members as “affected” or “unaffected” with respect to the proband’s disorder is critical for genetic analysis, family members are often evaluated for subtle manifestations of the proband’s phenotype; the use of HPO terms allows for harmonization of phenotypic information across family members.

HPO terms are publicly available through https://hpo.jax.org/app/ and the goal is to enable the integration of phenotypic information across scientific fields and databases and even species [12]. Since its main application involves rare disorders, HPO terms have been informed by OMIM, Orphanet, and Decipher entries. There are many benefits to using HPO terms, including specificity and creation of an annotation profile that allows for identification of candidate genes, candidate pathways, and potential model organisms. Additionally, HPO terms allow for comparisons across groups and research endeavors by creating standardized descriptions of symptoms. Greater use of HPO terms across the genetics community, for example when submitting clinical sequencing requests, would facilitate case sharing and recognition of similarly affected individuals.

4. Genomic sequencing and testing

Since many rare disorders are genetically-based, the most effective tool at the disposal of the UDP has been sequencing (rare disease statistics, 2015, globalgenes.org). Over the last 5 years the majority of children and a smaller number of adult applicants have already had non-diagnostic exome sequencing prior to their acceptance into the program. Hence, additional sequencing (including exome, genome, and/or RNA) performed on the proband and family members and analyzed through research pipelines have been important for diagnosis.

4.1. Sequencing analysis

The NIH UDP approach to exome and genome analysis is based on the implicit understandings that (1) there are limited pre-sequencing clues for diagnosis, (2) families have seemingly unique constellation of features, and (3) pedigrees are too small for conventional linkage analysis [13]. From 2008 to 2012, sequencing in the UDP was performed through the NIH Intramural Sequencing Center Comparative Sequencing Program. Since 2012, however, sequencing has been performed by one of two clinical laboratories funded by the UDN. An initial Clinical Laboratory Improvement Amendments (CLIA) report is generated for probands and family members; if no known disorder is identified, then research analysis begins.
The UDP begins with an annotated variant candidate list from the exome or, more commonly, the genome, sequencing data [14]. Each variant is given a genotype quality score based on a Bayesian statistic of the Most Probable Genotype (MPG) and a ratio of MPG to the coverage of any given variant [15]. Variants are then filtered for variant type, giving priority to coding sequence variants that result in missense or nonsense mutations, canonical slice site variants, or insertions/deletions. This list is further filtered by population frequency, assuming that disease-causing alleles in our population are likely rare, highly penetrant, and responsible for significant health problems [13].

The candidate list then requires manual curation. Variants within highly polymorphic genes (e.g., TTN) can be excluded; other individual base pairs are excluded due to consistent alignment problems, with the reference sequences having minor alleles [16]. This list is then filtered for segregation with disease among family members. SNP array data are also incorporated to reveal recombination mapping, mosaicism, regions of homozygosity, uniparental disomy, confirmation of parentage, and detection of copy number variants [17, 18]. The smaller candidate list is then reviewed for individual variants’ goodness of fit with clinical data, pathogenicity, and previously reported information. Pierson, et al, describes one of the first cases diagnosed with this process in the UDP [19].

Over time, the NIH UDP processes have evolved to include strategies to solve the completeness problem including exome capture, inclusion of intronic variants, and evaluation of medium-sized structural variants [20]. Automated programs have also been developed for ethnicity matched genotyping, salvage pathways for Mendelian inconsistencies, exon deletion filtration, and pedigree aware BAM file noise evaluation [21]. Internally referred to as the “forwards-backwards” analysis, this most recent toolset demonstrates the successful implementation of our analytic pipeline, since affected individuals had significantly more seemingly pathogenic, variants than their unaffected siblings. In other words, the pipeline is developed with enough constraint to keep the likely causal variant in the candidate list without creating an unmanageable number of variants.

4.2. Novel inheritance of known disease genes

The key to interpreting genetic analyses is determining which variants may be causing a proband’s disease. It is well known that different variants in the same gene can lead to different diseases, such as beta-galactosidase mutations causing either GM1 gangliosidosis or mucopolysaccharidosis type IVB, or BRCA2 variants causing either susceptibility to breast cancer or Fanconi anemia. Nevertheless, variants are frequently ruled out when a known disease is caused by variants in the gene but the phenotype of the proband does not match.

We have identified two novel human diseases by considering new inheritance patterns of known disease-causing genes. Monoallelic variants in GARS are known to cause Charcot-Marie-Tooth disease, type 2D (OMIM 601472). A seven-year-old female with extreme growth retardation and multiple organ involvement was found to have compound heterozygous variants in GARS [22]. Neither parent had symptoms of Charcot-Marie-Tooth disease and both had normal nerve conduction velocities. Functional studies demonstrated loss-of-function for both variants and, given the proband’s overlap with other ARS-related recessive conditions, the phenotypic and inheritance pattern of GARS-associated diseases was expanded.

Saul-Wilson syndrome (OMIM 618150), is a rare skeletal dysplasia described in 1982, but the genetic etiology was unknown until 2018 [23]. A UDN participant was evaluated and met clinical criteria for Saul-Wilson. Genomic sequencing analysis identified a de novo variant in COG4. Biallelic variants in COG4, however, are the known cause of type IIj congenital disorder of glycosylation (OMIM 613489). The UDN recruited twelve new patients with Saul-Wilson syndrome and confirmed that this disease is caused by the same recurrent monoallelic variant in COG4 (p.Gly516Arg), likely producing a gain of function.
4.3. Genome sequencing – structural variant calling

Since 2013, genome sequencing has been employed for UDP participants with prior non-diagnostic clinical exome sequencing and a compelling phenotype. Genome sequencing allows for better detection of structural variants as demonstrated by our recent discovery of a novel disease, Kilquist syndrome [24, 25]. The proband was previously known to have uniparental isodisomy for chromosome 5, so genome analysis targeted this region for candidate genes. Trio sequencing identified a 22kb homozygous deletion of \textit{SLC12A2}, that had not been identified by prior exome sequencing. We compared the HPO phenotype terms for the proband with the findings in a previously published murine model of SLC12A2 (NKCC1) and found compelling overlap, suggesting that this deletion was causative [26, 27].

5. Emerging strategies

5.1. Matchmaking - internally

Participation in a network, with multiple sites evaluating and sequencing patients, enables data-sharing and, therefore, matchmaking, across sites [28]. Clinical sequencing labs employ similar strategies, but the robust clinical information available through the UDN increases its ability to find similarly affected patients.

Exome sequencing identified seven individuals with significant phenotypic overlap and \textit{de novo} variants in \textit{TRAF7} (Tokita, 2018). When the UDP was alerted to this match, an internal database search was performed and an additional patient with a \textit{de novo} \textit{TRAF7} variant and substantial phenotypic overlap was identified. We frequently employ this searching strategy for potential matches as new disease genes are identified both within and outside of our cohort.

5.2. Matchmaking – Externally

5.2.1. PhenomeCentral

Matchmaking services, or “genetic dating sites”, have been developed to allow researchers and clinicians to find similarly affected individuals using phenotypic or genetic data. The UDN utilizes PhenomeCentral, a restricted access network for clinicians, researchers, and scientists, to share patient phenotype and genotype data [29]. De-identified patient information is submitted to the database and the user is provided a list of cases that appear most similar to the data submitted. Cases can be submitted as private (not shared in matchmaking), match-able (seen only when matched), or public (visible to all users).

PhenomeCentral participates in the Matchmaker Exchange (MME), a federated network of databases with genotypic and phenotypic information through an application programming interface (API) [30]. MME enables searches of multiple databases with a single query and users can choose where to deposit their data depending on the type of data they are submitting. Other sites within MME include GeneMatcher, DECIPHER, MyGene2, and \textit{matchbox} [31–34]. Increased use of MME by clinicians will improve the ability to identify new diseases and provide diagnoses for patients and their families across platforms.

5.2.2. Social media outreach

In addition to secure matchmaking sites, the UDN employs social media to share patient data from consented individuals. Participant Pages are created to summarize a proband’s medical history, significant findings, and candidate genes. The goal of these pages is two-fold: to find similarly affected individuals who may help clarify the cause of the condition; and to find external researchers who may
have expertise in the candidate gene. Participant Pages are modeled after the success of Matthew and Cristina Might in identifying additional cases of NGLY1 deficiency. By harnessing the power of social media and a well-written blog post, the Might family identified additional patients with variants in NGLY1 [35, 36]. Finding additional patients led to the identification of a new disease and a natural history study [37, 38].

The UDN has successfully utilized these pages, and new disease discoveries have benefited from the addition of probands found through Participant Pages. The newly described syndrome caused by variants in NACC1 utilized these webpages to identify two additional patients for their cohort [39]. A fifth patient was identified with variants in IRF2BPL utilizing a gene page, designed in the same format as participant pages, but focusing on the gene itself [40]. These pages are subsequently shared on Facebook and Twitter to reach participants directly. Rare disease patients are already using social media to find one another; these pages increase our reach and allow us to “meet” participants where they are [41].

5.3. RNA Sequencing

RNA sequencing (RNA-seq) allows for the direct probing of variation in RNA content and in RNA sequence. In combination with exome or genome sequencing, RNA-seq can help prioritize variants that would otherwise be uninterpretable or lost to filtration. RNA-seq is especially helpful in determining the effect of splice variants and can occasionally identify splicing defects not recognized in exome and genome sequence data.

A recent UDP case was solved using just this strategy. Agnostic analysis of RNA-seq data for splicing variants identified a novel splice junction in a patient with a non-diagnostic genome sequence (unpublished data). The splice junction was not observed in any control sample and review of the genome data showed a de novo, deep intronic (+1242bp) single nucleotide variant believed to be responsible for the aberrant splicing in PHIP (OMIM #617991). The proband’s phenotype was consistent with published cases of individuals with truncating variants in PHIP, thus identifying the diagnosis for this proband.

5.4. Model Organisms

The goal of the UDN Model Organism Screening Center (MOSC) is to provide compelling data from studies in worms, flies, or zebrafish that either support or refute disease causality of specific variants [42]. To do this, the MOSC utilizes a standard pipeline for review of candidate genes. Variants are prioritized for potential study in a model organism when they: 1) are in novel or candidate disease-causing genes; or 2) are novel variants in a known disease-causing gene in a patient with a novel phenotype.

Variants that pass an initial quality control review (i.e., rare, no existing disease association, and identification of potential second case) are then reviewed by each model organism group. Cases are further prioritized based on whether there is an ortholog or paralog in the model system, whether the amino acid is conserved in the species, and whether the gene or variant has been previously studied in a model system. If a case meets criteria, it is assigned a model organism and the group begins their research.

The MOSC has been instrumental in solving cases in the UDN. One such case was that of a 7-year-old male with global developmental delay, hypotonia, expressive speech delay, intellectual disability, and dysmorphic features. Trio exome sequencing identified a de novo variant in EBF3 [43]. Two additional probands were identified with de novo variants at the same amino acid and overlapping phenotypes. The phenotype was recapitulated in a fly model. Flies were selected for this project as the Knot gene
has a 62% identify and a 70% similarity with human EBF3. Multiple experiments demonstrated that the two EBF3 variants failed to rescue the lethality of the Knot knockout flies, while wildtype EBF3 was able to rescue Knot knockouts, thus demonstrating the loss of function of the de novo EBF3 variants. This functional validation of the EBF3 variants identified a new neurodevelopmental disorder and provided three families with a diagnosis.

5.5. Metabolomics

The UDN Metabolomics Core provides “untargeted and targeted quantitative metabolic approaches for bioinformatic, and clinical interpretation of specimens” [42]. The Core performs global, untargeted glycan, lipid, and mitochondrial metabolite profiling to identify priority targets for further study. The core continues to develop assays for identifying new metabolites and offers interpretations of identified metabolites.

Since inborn errors of metabolism have been associated with a wide range of symptoms that affect multiple organ systems, abnormal metabolites are a consideration for any UDN participant. Because metabolites in plasma or urine are subject to external factors, such as medications, diet, and supplements, metabolomic analysis in the UDN is performed on cultured fibroblasts grown and prepared under identical controlled conditions. Although metabolomics has not yet solved a case, we are hopeful that their agnostic and hypothesis driven analyses will improve our diagnostic capabilities, both in identifying and ruling out candidate genes.

6. Conclusions

The UDN evaluates a highly selected group of participants who have usually spent years on the diagnostic odyssey traveling to multiple academic centers in search of a diagnosis. Despite using many cutting-edge tools, the UDN fails more often than it succeeds; only 27.5% of all UDP participants (35% of pediatric participants) have received a diagnosis. However, for many of the 286 individuals who have participated in the UDP over the last 10 years, receiving a diagnosis has been life-changing. For some diagnoses, treatment is possible, for others a change in management or access to additional services has been facilitated. Diagnoses that have led to treatment include: diagnoses of KMT2B which have led to treatment with deep-brain stimulation and treatment of congenital serine biosynthesis defect due to biallelic PSAT1 variants with serine [44, 45]. For other families, diagnoses of ADNP, KAT6A, GLB1 mutations among many others have led to inclusion in ongoing natural history studies. For some families, determining an inheritance pattern has provided family planning options and healthy children have been born to parents who would have otherwise avoided pregnancy. Some families with the same diagnosis have joined together to start advocacy groups for mutual support and to raise funds for therapeutic research. Yet for most participants who receive a diagnosis the odyssey is over, and they can move on to the next phase of their meaningful lives [3]. What’s in a name? Everything!!

References


[42] About the Undiagnosed Diseases Network.

