Published in final edited form as:

Hum Mutat. 2015 October ; 36(10): 985–988. doi:10.1002/humu.22840.

Data sharing in the Undiagnosed Diseases Network

Catherine A. Brownstein[1,*,#], **Ingrid A. Holm**[1,#], **Rachel Ramoni**[2], **David B. Goldstein**[3], and **Members of the UDN**[Ψ]

[1]Division of Genetics and Genomics and the Manton Center for Orphan Disease Research, Boston Children's Hospital, Boston, MA

[2]Center for Biomedical Informatics, Harvard Medical School, Boston, MA

[3]Columbia Institute for Genomic Medicine, Genetics and Development, New York, New York, United States

Abstract

The Undiagnosed Diseases Network (UDN), builds on the successes of the Undiagnosed Diseases Program at the National Institutes of Health (NIH UDP). Through support from the NIH Common Fund, a coordinating center, six additional clinical sites, and two sequencing cores comprise the UDN. The objectives of the UDN are to: (1) improve the level of diagnosis and care for patients with undiagnosed diseases through the development of common protocols designed by an enlarged community of investigators across the Network; (2) facilitate research into the etiology of undiagnosed diseases, by collecting and sharing standardized, high-quality clinical and laboratory data including genotyping, phenotyping, and environmental exposure data; and (3) create an integrated and collaborative research community across multiple clinical sites, and among laboratory and clinical investigators, to investigate the pathophysiology of these rare diseases and to identify options for patient management. Broad-based data sharing is at the core of achieving these objectives, and the UDN is establishing the policies and governance structure to support broad data sharing.

Keywords

Big Data; Genetics; Genomics; Precision Medicine; Personalized Medicine; Matchmaker Exchange

Introduction

The UDN is structured to make advances in multiple disciplines related to the fields of genomics and undiagnosed diseases. The activities of the seven clinical sites, coordinating center, and two sequencing centers that constitute the UDN Network are shown in Table 1.

[*]Correspondence to: Catherine Brownstein, Boston Children's Hospital, Division of Genetics and Genomics, 3 Blackfan Circle, CLSB 16, Boston, MA 02115, catherine.brownstein@childrens.harvard.edu.

[#]the authors contributed equally to this work

[Ψ]Members listed in the Appendix

Conflict of interest disclosure

DBG is a consultant with LabCorp in their diagnostic exome service. The other authors have no conflict of interest.

In the following sections, we describe the history and evolution of the network in the context of the importance of data sharing and breaking down barriers to discovery.

History and Aims

A request for applications from the NIH Common Fund for the Coordinating Center (CC) to the UDN was released in November 2012 (http://grants.nih.gov/grants/guide/rfa-files/RFA-RM-12-020.html), and the request for sites was released on January 29, 2013 (http://grants.nih.gov/grants/guide/rfa-files/RFA-RM-13-004.html. The UDN CC was selected several months prior to the selection of the new Clinical Sites, which allowed for careful study and codification of the processes used and refined by the NIH UDP and selection for those suitable for Network-wide adoption and refinement. The request for a sequencing core was announced on August 7, 2013 (http://grants.nih.gov/grants/guide/rfa-files/RFA-RM-13-018.html), for which two centers were selected (Medical College of Wisconsin and Baylor).

The UDN is composed of a CC, 7 UDN Clinical Sites (6 new Sites plus the foundational NIH UDP), and 2 Sequencing Centers (Table 1).

The UDN has three major aims:

1. Improve the level of diagnosis and care for patients with undiagnosed diseases through the development of common protocols designed by an enlarged community of investigators.
2. Facilitate research into the etiology of undiagnosed diseases, by collecting and sharing standardized, high-quality clinical and laboratory data including genotyping, phenotyping, and documentation of environmental exposures.
3. Create an integrated and collaborative research community across multiple clinical sites and among laboratory and clinical investigators prepared to investigate the pathophysiology of new and rare diseases and share this understanding to identify improved options for optimal patient management.

The program goals are to provide improved patient access to state-of-the-art diagnostic methods, by expanding the available expertise and facilities serving patients with these unusual disorders and to accelerate discovery and innovation in diagnosing and treating these patients.

Data Sharing

The success of the UDN will depend on the collection and subsequent sharing of well-curated clinical and research data both within and outside of the network. These activities are grounded in a common Data Sharing and Use Agreement (DSUA). The DSUA is a legal agreement between institutions, in this case, signed by the designated institutional officials from each of the members of the UDN (including affiliated institutions), which establishes what data may be shared, the ways in which the information in the data set may be used, and how the data will be protected. The UDN DSUA was carefully reviewed in order to ensure consistency with the IRB protocols and consent forms, as well as relevant sections of the

UDN Manual of Operations, e.g., UDN Publications Policies. Coming to agreement about the content of the DSUA required several rounds of interaction with technology transfer offices of the member institutions (including affiliates), an activity that was led by the CC.

The UDN DSUA has features that are not typical of such agreements. For instance, in the context of the UDN, data sharing is integral to our ability to see participants, rather than simply facilitating secondary data analysis. To illustrate this point, consider that initial applications are received centrally by the CC through the online application, the UDN Gateway: the transfer of these data to the clinical sites at which these applications will be evaluated depends upon the existence of the DSUA. Another feature that distinguishes the UDN DSUA is that it explicitly enables the sharing of personally identifiable information, which is essential due to the nature of the research being conducted within the UDN. Furthermore, many DSUAs are constructed as pairwise agreements in which Site A agrees to share data with Site B: the UDN DSUA covers multi-directional data sharing among all components of the network. Finally, the UDN DSUA explicitly *requires* sharing of the researcher-facing data elements contained within the UDN Gateway, which have been agreed-upon by the UDN Steering Committee. Examples of these data elements include medical history information, human phenotype ontology-encoded phenotypes, and sequencing data.

UDN participants consent to have their data shared, in accordance with the UDN informed consenting process. Given that the UDN has adopted a central IRB model (with the central IRB being located at NHGRI), there is a great deal of consistency in the informed consent documents. The consent documents contain identical information, save for necessarily site-specific language (e.g., HIPAA authorization language). Each participant in the database will be associated with a UUID (Universal Unique Identifier), which will be used as the primary identifier for all data associated with that participant. Role-based access and physical security controls that are aligned with the sensitivity of the data at each point of use and access will be employed.

The UDN data sharing philosophy is consistent with the goals of the NIH Data Sharing Policy (NIH Data Sharing Policy. http://grants.nih.gov/grants/policy/data_sharing/). The NIH states “Data should be made as widely and freely available as possible while safeguarding the privacy of participants, and protecting confidential and proprietary data.”. Thus, the UDN DSUA also addresses external data sharing. In particular, the DSUA allows for the coordinating center to provision de-identified study data (HHS - Office for Civil Rights - HIPAA. http://www.hhs.gov/ocr/hipaa/) in the database of Genotypes and Phenotypes (dbGaP) or other controlled-access data repositories.

As has been covered in numerous forums (Knowledge Exchange, http://repository.jisc.ac.uk/5662/1/KE_report-incentives-for-sharing-researchdata.pdf), there are few incentives for an individual investigator to share data: it is an additional step in the research process and may be perceived as reducing the career benefits of having gathered the data. The UDN encourages data sharing via multiple approaches, including the provision of infrastructure, policy, and data standards to support data sharing; ensuring that data sharing would be an automatic consequence of Network operations; and by fostering a culture of collaboration, openness, and mutual trust. The previously-mentioned governance is key to maintaining this culture, as Network decision-making is transparent and collaborative.

Collaborations with external groups

UDN data may be shared with investigators who are not currently part of the UDN, if the initiatives are complementary and consistent with UDN goals. This is true on an individual participant and aggregate UDN population basis. For example, if there are useful experts in a patient's phenotype outside the UDN, the data would be shared with them on an as needed basis. There will also be data sharing initiatives on a larger scale. Once a candidate gene for a condition has been identified, finding the second case to support the hypothesis has been traditionally word of mouth and involves serendipity. That said, there have been several "matches" made this way, leading to a second case identified. The UDN is interested in collaborating with, and contributing to, external organizations to facilitate gene discovery and the identification of additional cases. To this end, in addition to collaborations within and between the UDN sites and workgroups, the Network has also been linking to other groups focused on similar goals and activities.

Data resulting from UDN efforts will be deposited in dbGaP, maintained by the National Center for Biotechnology Information at the NIH. Data may also be shared with other controlled-access databases, registries, and repositories, such as PhenomeCentral (a collaborative effort of a number of rare disease research communities), the NIH Global Rare Diseases Registry, and potentially with other condition-specific registries, such as those listed in Orphanet.

The UDN is a member of Matchmaker Exchange's collaborative effort to address the common challenge of exome and genome sequencing in both the research and clinical settings. The UDN steering committee has voted to contribute data to the Matchmaker Exchange network via PhenomeCentral. PhenomeCentral is the ideal vehicle since it is a repository for secure data sharing working in the rare disorder community, and is also part of Matchmaker Exchange. It is hoped that the UDN and its participants will benefit from PhenomeCentral's remote matching API for finding genotypically and phenotypically similar patients within the Matchmaker Exchange network.

Data Sharing for Innovation in Genomic Analysis

While one critical goal of the UDN is to identify the causes of disease in the genomes of the individuals that are studied, another complementary goal is to contribute to the science of genome interpretation. In the rush of enthusiasm surrounding the remarkable advances in human genetics it is easy to forget that interpreting individual genomes is very hard, and remains error prone. This reality is especially important to keep in mind given that one explicit aim of genetic diagnostics is to inform treatment options. In this context, mistakes in genome interpretation can result in real clinical harm. For these reasons, the UDN is committed to evaluating and refining best practices in the evolving science of interpreting patient genomes. Beyond broad sharing, there are other cross-institution initiatives, such as sharing approaches to analysis, which will move the field forward via developing the building blocks necessary to support the science of genome interpretation.

To better visualize the data the UDN will work with, it is useful to keep in mind the usual outcomes of whole genome or whole exome diagnostic sequencing. When a single genome of a patient with a presumed serious genetic disease is interpreted, the outcomes that are now normally encountered can be classified into four groups:

1. A clear likely pathogenic genotype is identified,
2. An interesting candidate is identified, for example on the basis of a suggestive bioinformatic “signature”,
3. Multiple suggestive candidates are identified,
4. No good candidate is identified.

Although some mistakes are made in groups 1 and 4, these outcomes are generally straightforward. The main point to make about group 4 is the critical need to re-analyze the genomes regularly, given the rapid pace of new gene discovery. It is also a priority for the field to develop a view of whether there are any predominant explanations for some negative results, such as pathogenic variants that are refractory to identification by next generation sequencing, regulatory pathogenic variants that are difficult to identify as causal, or oligogenic models. The UDN hopes to address these questions at least in part through comparisons of whole exome and whole genome sequencing.

Groups 2 and 3 are where considerable methodological developments are needed to permit the most appropriate conclusions to be drawn in these more challenging settings. This begins by sharing phenotype and sequencing data among the UDN sites and throughout the Matchmaker Exchange as a way to identify unrelated individuals with the same phenotype and variants in the same genes. Data sharing at this level also facilities the sites’ doing their own large-scale analyses that would otherwise not be possible. The transparency of analysis protocols allows for the identification of best methods and practices (Brownstein, et al., 2014).

A percentage of exomes/genomes carry multiple candidate genes for their phenotypic condition. Resolution in this case depends on finding additional individuals and functional assessment of the multiple candidates, as well as detailed phenotyping to assess whether expected phenotype correlates of each of the candidates are observed. The UDN affords us the opportunity for appropriate genomes to receive this full interpretation.

Additionally, when considering the outcome of many interpreted genomes, we are often left with interesting candidates on the basis of a bioinformatic signature, but not enough is known about them to draw conclusions. For example, in an analysis of 103 trios, it was reported that 28.2% of the 103 patients unresolved by a recessive disorder have a “hot zone” (defined in (Petrovski, et al., 2013)) *de novo* variant, versus 6.0% of the 728 control trios ($p=3.0\times10^{-10}$). 19.4% of the 103 patients unresolved by a recessive disorder have a hot zone *de novo* variant in an essential gene, versus 2.1% of the 728 control trios ($p=8.8\times10^{-11}$). This translates to a 90% excess (Zhu, et al., 2015). These results emphasize that even when there are no clearly causal pathogenic variants, bioinformatics analyses demonstrate unequivocally the presence of real risk factors. Optimizing these bioinformatic predictors of pathogenic variants is therefore a priority for the field, as is learning the contexts when suggestive genomic findings should be shared with patients and their care providers.

The novel approach here is focused on what is necessary to resolve these candidate variants. By aggregation of candidates such as hot zone variants and sharing phenotypic, genotypic, and functional work across institutions, discoveries will be made that otherwise would not be possible.

Patient Participant – driven data sharing

In addition to data sharing by UDN researchers with those outside of the network, UDN patients and families may lead their own data sharing efforts, as illustrated by the work of Matthew Might, who is the UDN Coordinating Center's patient and family advisor. His son was the first suspected case of a disorder caused by a mutation in *NGLY1*(Enns, et al., 2014); Might chronicled his son's journey in a blog post, which led to the identification of a second case (Might, 2012). To date, there are over 35 confirmed cases of *NGLY1* deficiency (Might, 2015).

Conclusions

Following its first year, the UDN is poised to make great strides in the fields of genomics and informatics. The network is developing tools and best practices that are being shared and utilized by the genomics and informatics communities and beyond. The UDN is pioneering techniques and approaches that will advance the fields of genetics and genomics. Most of the phenotypes enrolled in the UDN are incredibly rare and unique. Sharing phenotype and sequencing data may allow for identification of other patients with the same phenotype, smoothing the path to solving the molecular basis of such phenotypes. The commitment to data sharing is pioneering and an example of where the field is moving. It is hoped that this will result in new diagnoses for undiagnosed patients, and overall improvements in health care, and enhanced understanding of the biology of disease. Exciting partnerships with Matchmaker exchange and PhenomeCentral will also facilitate discovery.

Acknowledgments

Funding:

This work was supported by NIH Common Fund, through the Office of Strategic Coordination/Office of the NIH Director and managed by NHGRI U01 HG007530, HG007690, and HG007672.

The authors wish to thank the patients and families for participating in the UDN.

References

Brownstein CA, Beggs AH, Homer N, Merriman B, Yu TW, Flannery KC, DeChene ET, Towne MC, Savage SK, Price EN, Holm IA, et al. An international effort towards developing standards for best practices in analysis, interpretation and reporting of clinical genome sequencing results in the CLARITY Challenge. Genome Biol. 2014; 15(3):R53. [PubMed: 24667040]

Enns GM, Shashi V, Bainbridge M, Gambello MJ, Zahir FR, Bast T, Crimian R, Schoch K, Platt J, Cox R, et al. Mutations in NGLY1 cause an inherited disorder of the endoplasmic reticulum-associated degradation pathway. Genet Med. 2014; 16(10):751–8. [PubMed: 24651605]

Might M. Hunting down my son’s killer. 2012

Might, M. Cases of NGLY1. Brownstein, C., editor. 2015. talk on how many cases of NGLY1 are now confirmed. ed

Petrovski S, Wang Q, Heinzen EL, Allen AS, Goldstein DB. Genic intolerance to functional variation and the interpretation of personal genomes. PLoS Genet. 2013; 9(8):e1003709. [PubMed: 23990802]

Zhu X, Petrovski S, Xie P, Ruzzo EK, Lu YF, McSweeney KM, Ben-Zeev B, Nissenkorn A, Anikster Y, Oz-Levi D, et al. Whole-exome sequencing in undiagnosed genetic diseases: interpreting 119 trios. Genet Med. 2015

Appendix: Members of the UDN

First name	Last name	UDN site	email
Carlos	Bacino	Baylor	cbacino@bcm.edu
Brendan	Lee	Baylor	blee@bcm.edu
Christine	Eng	BMC SEQ	ceng@bcm.edu
Narayanan	Veeraghavan	BMC SEQ	narayanv@bcm.edu
David	Bernick	CC	David_Bernick@hms.harvard.edu
Catherine	Brownstein	CC	catherine.brownstein@childrens.harvard.edu
Ingrid	Holm	CC	ingrid.holm@childrens.harvard.edu
Issac	Kohane	CC	isaac_kohane@hms.harvard.edu
Alexa	McCray	CC	alexa_mccray@hms.harvard.edu
Rachel	Ramoni	CC	rachel_ramoni@hms.harvard.edu
Kimberly	Splinter	CC	kimberly_splinter@hms.harvard.edu
Vandana	Shashi	Duke	vandana.shashi@duke.edu
David	Goldstein	Duke/Colum bia	dg2875@columbia.edu
Alan	Beggs	Harvard	beggs@enders.tch.harvard.edu
Joseph	Loscalzo	Harvard	jloscalzo@partners.org
Calum	MacRae	Harvard	camacrae@bics.bwh.harvard.edu
Edwin	Silverman	Harvard	ed.silverman@channing.harvard.edu
Joan	Stoler	Harvard	joan.stoler@childrens.harvard.edu
David	Sweetser	Harvard	dsweetser@partners.org
Tina	Hambuch	Illumina	thambuch@illumina.com
Howard	Jacob	MCW SEQ	jacob@mcw.edu
Kim	Strong	MCW SEQ	kstrong@mcw.edu
Elizabeth	Worthey	MCW SEQ	eworthey@mcw.edu
Teri	Manolio	NIH	manolio@nih.gov
John	Mulvihill	NIH	John.Mulvihill@nih.gov
Anastasia	Wise	NIH	anastasia.wise@nih.gov
Euan	Ashley	Stanford	Euan@stanford.edu
Jonathan	Bernstein	Stanford	Jon.Bernstein@stanford.edu

First name	Last name	UDN site	email
Paul	Fisher	Stanford	pfisher@stanford.edu
Matt	Wheeler	Stanford	wheelerm@stanford.edu
Katrina	Dipple	UCLA	Kdipple@mednet.ucla.edu
Stan	Nelson	UCLA	snelson@ucla.edu
Christina	Palmer	UCLA	cpalmer@mednet.ucla.edu
Eric	Vilain	UCLA	evilain@ucla.edu
Camilo	Toro	UDP	toroc@mail.nih.gov
David	Adams	UDP	david.adams@nih.gov
Bill	Gahl	UDP	gahlw@mail.nih.gov
Cynthia	Tifft	UDP	ctifft@nih.gov
Rizwan	Hamid	Vanderbilt	rizwan.hamid@vanderbilt.edu
John	Newman	Vanderbilt	john.newman@vanderbilt.edu
John	Phillip	Vanderbilt	John.a.phillips@vanderbilt.edu

Table 1

Activities of each site in the UDN.

UDN Site	Role	Affiliated Institutions
Baylor College of Medicine (Houston TX)	UDN Clinical Site	N/A
Baylor College of Medicine (Houston TX)	Sequencing Core	N/A
Columbia University (New York NY)	UDN Clinical Site	Duke University (Durham NC)
Harvard Medical School	Coordinating Center	Boston Children’s Hospital (Boston MA)
		Harvard School of Public Health (Boston MA)
		Clinical Assistance Programs (Framingham MA)
Brigham and Women’s Hospital (Boston MA)	UDN Clinical Site	Boston Children’s Hospital (Boston MA)
		Massachusetts General Hospital (Boston MA)
Medical College of Wisconsin (Milwaukee WI)	Sequencing Core	Illumina, Inc. (San Diego CA)
National Human Genome Research Institute (Bethesda MD)	UDN Clinical Site	N/A
Stanford Medicine (Stanford CA)	UDN Clinical Site	N/A
University of California Los Angeles (Los Angeles CA)	UDN Clinical Site	N/A
Vanderbilt University Medical Center (Nashville TN)	UDN Clinical Site	N/A