From case studies to community knowledge base: MSeqDR provides a platform for the curation and genomic analysis of mitochondrial diseases

Marni J. Falk,[1] Lishuang Shen,[2] and Xiaowu Gai[2]

[1]Division of Human Genetics, Department of Pediatrics, The Children's Hospital of Philadelphia and University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania 19104, USA; [2]Center for Personalized Medicine, Children's Hospital Los Angeles, Los Angeles, California 90027, USA

Mitochondrial disease is now recognized to represent a highly heterogeneous group of genetic disorders that impair energy metabolism and can potentially involve every organ system, with more than 250 causative genes already confirmed across both the nuclear and mitochondrial genomes (Koopman et al. 2012). Novel variants in a recently recognized nuclear disease gene are reported in this issue to cause an autosomal-recessive multisystemic mitochondrial disease affecting two cousins from a Lebanese family (Joshi et al. 2016). As has now essentially become commonplace in the highly heterogeneous classes of neurologic and mitochondrial diseases, massively parallel genomic sequencing in a single family with a complex phenotype proved to be an effective means to identify potentially pathogenic novel mutations in a nuclear gene encoding a mitochondrial protein that was recently linked to cerebellar ataxia (Choquet et al. 2015; Jobling et al. 2015). Protein modeling, functional analyses in patient fibroblasts, and wild-type cDNA rescue were then used to conclusively demonstrate the causality of the novel *PMPCA* mutations in their patients' more expansive phenotype that extended beyond isolated cerebellar atrophy to also include multisystemic mitochondrial disease. In this iterative fashion that directly relies on continual publication of case reports and data sharing within community resources, community knowledge steadily grows of new gene disorders, new mutations within known genes, and their expanded phenotypes, further improving community recognition of the functional significance of the up to 1500 proteins that localize within mitochondria (Calvo et al. 2016). However, given the large number of mitochondrial disease genes having low prevalence, their wide spectrum of clinical presentations, and often overlapping clinical phenotypes (Koopman et al. 2012), it is increasingly challenging for any one clinician or researcher to recognize or stay current on all potential gene causes for a patient's disease presentation. Large public databases such as ClinVar serve as centralized warehouses for annotation assertions on potential gene causes for diverse disease presentations but remain dependent on user contributions (Landrum et al. 2016).

To aid the community of mitochondrial disease researchers in curating and accessing the rapidly expanding knowledge base of mitochondrial disease genetics, a Web-based resource has been built, the Mitochondrial Disease Sequence Data Resource (MSeqDR; https://mseqdr.org). This multidimensional data resource aims to support both clinicians and researchers in the mitochondrial disease community as conceived and built by the MSeqDR Consortium (Falk et al. 2015), a grass-roots effort begun in 2012 by mitochondrial disease researchers, clinicians, and bioinformaticians with the support of the United

Mitochondrial Disease Foundation (UMDF) and North American Mitochondrial Disease Consortium (NAMDC). Thus, the focus of MSeqDR is to provide a centralized community resource in which to deposit, curate, annotate, analyze, and share accurate genomic knowledge on the many different mitochondrial diseases, genes, and variants that are continually being recognized to cause mitochondrial disease (McCormick et al. 2013). MSeqDR includes distinct domains in an interactive Web interface designed to intuitively support the diverse genomic data needs of clinicians, researchers, and diagnostic laboratories working in the field of mitochondrial disease. Major domains include (1) MSeqDR Gbrowse, a genome browser with custom tracks specific to diverse aspects of mitochondrial biology and disease; (2) MSeqDR-LSDB, a locus-specific database that provides curated data on more than 1360 nuclear and mitochondrial genes linked to mitochondrial biology and disease, as well as their transcripts and variants (MSeqDR currently synchronizes with ClinVar by periodically extracting variant annotations relevant to mitochondrial diseases from ClinVar [Landrum et al. 2016]); (3) a suite of public and custom bioinformatics tools to support direct genomic analyses by MSeqDR users on their own data sets of either nuclear or mitochondrial genomes; (4) phenome-based data and analysis tools to catalog features of mitochondrial diseases using defined terms from the Human Phenotype Ontology (HPO) (Groza et al. 2015); (5) cloud- and Web-based exome and/or genome analysis in Genesis 2.0 (formerly GEM.app) (Gonzalez et al. 2013), which enables discovery of additional patient cases for rare genetic disorders via the Matchmaker Exchange (Gonzalez et al. 2015), and (6) submission tools to facilitate genomic data set and individual variant deposition into MSeqDR. Additionally, variant annotation tools have been implemented within MSeqDR to simplify the work required to formulate variants in ClinVar submission format. Collaboration with the Clinical Genomic Resource (ClinGen) project team and the National Center for Biotechnology Information (NCBI) is underway to enable coordinated and automated ClinVar submission via MSeqDR. In this way, there will be a synergistic relationship between the community-specific MSeqDR resource and the more general ClinVar resource. Specifically, it is our goal that the mitochondrial disease community use MSeqDR as their primary system for deposition and analysis of genomic data, with subsequent sharing of interpreted variants into ClinVar in a way that will be directly managed by MSeqDR. This might potentially include reference in the ClinVar variant entries to connect a submitted interpretation with the supporting observations housed in MSeqDR, accessible to all users that have access authentication. Detailed explanation of each MSeqDR component has been recently published (Shen et al. 2016). In addition, step-by-step user tutorials and practice exercises are provided on the MSeqDR website to familiarize users with the diverse functionality of MSeqDR. Conveniently, a central search portal is accessible from every page within MSeqDR, allowing users to efficiently survey all relevant data within the resource on any given disease, gene, or variant of interest. All sites within MSeqDR are internally linked and directly link out to many external databases. Although accounts are provided free to all academic users, login is required to access all of the data and different domain capabilities within MSeqDR. Public access is only provided to public data sets by default. Through the account system, users gain access to data sets shared by other MSeqDR consortium members. However, MSeqDR users retain the option to keep their own data sets private, share them with other MSeqDR users, or share them with the general public.

As was done for the *PMPCA* mutations identified by Joshi et al. (2016), novel disease genes and variants can be readily added to MSeqDR-LSDB. A custom MSeqDR accession number is assigned to all annotated pathogenic variants that are either submitted by users or batch extracted from other databases such as ClinVar or Ensembl, each containing a unique "MSCV" (MSeqDR clinically related variant) identifier with a seven-digit code. Users can readily submit variants, in either VCF (variant call format) or HGVS (Human Genome Variation Society) format, using the custom MSeqDR "pathogenic variant submission tool"

(https://mseqdr.org/submission.php). Authenticated users can add variants, as well as logged comments on any variant annotation, with assistance by the semiautomated variant annotation tool that is available within MSeqDR. Expert review panels are now being organized within the MSeqDR community to review the levels of evidence for pathogenicity of all reported variants within all MSeqDR-LSDB curated genes, following the guidelines set by the ClinGen Consortium (Rehm et al. 2015). Active development is also underway to incorporate phenotype entry and analysis tools in MSeqDR, to provide deeper knowledge for individual variants of the often complex, multisystem phenotypes seen in affected individuals (Shen et al. 2016). Specifically, MSeqDR is actively working to capture standardized phenotype data in controlled vocabularies, including HPO and MeSH (Medical Subject Headings). PhenoTips is being used for phenotype data input (Girdea et al. 2013) and combined with the associated phenotype-driven exploration and analysis tools in MSeqDR that support HPO, OMIM (Online Mendelian Inheritance in Man), and MeSH terminology.

In summary, MSeqDR provides a centralized Web portal in which to deposit, organize, curate, analyze, discover, educate, and share genomic knowledge necessary for the global community to better diagnose and understand the complex and ever-growing cadre of mitochondrial diseases. It is our hope that MSeqDR will prove to be a widely used and useful resource for data deposition and analysis by the diverse community of genomic researchers, mitochondrial disease clinicians, and genetic diagnostic laboratories who continue to discover and evaluate new mitochondrial disease patients, genes, and variants.

ADDITIONAL INFORMATION

Tarnopolsky, Isabelle Thiffault, David Thorburn, Johan Van Hove, Mannis van Oven, Lynne Wolfe, Lee-Jun Wong, Philip Yeske, Douglas C. Wallace, Zhe Zhang, and Stephan Zuchner.

REFERENCES

Calvo SE, Clauser KR, Mootha VK. 2016. MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins. *Nucl Acids Res* **44**(D1): D1251–D1257.

Choquet K, Zurita-Rendón O, La Piana R, Yang S, Dicaire MJ, Care4Rare Consortium, Boycott KM, Majewski J, Shoubridge EA, Brais B, et al. 2016. Autosomal recessive cerebellar ataxia caused by a homozygous mutation in *PMPCA*. *Brain* **139**(Pt 3): e19.

Falk MJ, Shen L, Gonzalez M, Leipzig J, Lott MT, Stassen AP, Diroma MA, Navarro-Gomez D, Yeske P, Bai R, et al. 2015. Mitochondrial Disease Sequence Data Resource (MSeqDR): a global grass-roots consortium to facilitate deposition, curation, annotation, and integrated analysis of genomic data for the mitochondrial disease clinical and research communities. *Mol Genet Metab* **114**: 388–396.

Girdea M, Dumitriu S, Fiume M, Bowdin S, Boycott KM, Chenier S, Chitayat D, Faghfoury H, Meyn MS, Ray PN, et al. 2013. PhenoTips: patient phenotyping software for clinical and research use. *Hum Mutat* **34**: 1057–1065.

Gonzalez MA, Lebrigio RF, Van Booven D, Ulloa RH, Powell E, Speziani F, Tekin M, Schüle R, Züchner S. 2013. GEnomes Management Application (GEM.app): a new software tool for large-scale collaborative genome analysis. *Hum Mutat* **34**: 842–846.

Gonzalez M, Falk MJ, Gai X, Postrel R, Schule R, Zuchner S. 2015. Innovative genomic collaboration using the GENESIS (GEM.app) platform. *Hum Mutat* **36**: 950–956.

Groza T, Kohler S, Moldenhauer D, Vasilevsky N, Baynam G, Zemojtel T, Schriml LM, Kibbe WA, Schofield PN, Beck T, et al. 2015. The human phenotype ontology: semantic unification of common and rare disease. *Am J Hum Genet* **97**: 111–124.

Jobling RK, Assoum M, Gakh O, Blaser S, Raiman JA, Mignot C, Roze E, Durr A, Brice A, Levy N, et al. 2015. *PMPCA* mutations cause abnormal mitochondrial protein processing in patients with non-progressive cerebellar ataxia. *Brain* **138**(Pt 6): 1505–1517.

Joshi M, Anselm I, Shi J, Bale TA, Towne M, Schmitz-Abe K, Crowley L, Giani FC, Kazerounian S, Markianos K, et al. 2016. Mutations in the substrate binding glycine-rich loop of the mitochondrial processing peptidase-α protein (PMPCA) cause a severe mitochondrial disease. *Cold Spring Harb Mol Case Stud* **3**: a000786.

Koopman WJ, Willems PH, Smeitink JA. 2012. Monogenic mitochondrial disorders. *N Engl J Med* **366**: 1132–1141.

Landrum MJ, Lee JM, Benson M, Brown G, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Hoover J, et al. 2016. ClinVar: public archive of interpretations of clinically relevant variants. *Nucleic Acids Res* **44**(D1): D862–D868.

McCormick E, Place E, Falk MJ. 2013. Molecular genetic testing for mitochondrial disease: from one generation to the next. *Neurotherapeutics* **10**: 251–261.

Rehm HL, Berg JS, Brooks LD, Bustamante CD, Evans JP, Landrum MJ, Ledbetter DH, Maglott DR, Martin CL, Nussbaum RL, et al. 2015. ClinGen—the Clinical Genome Resource. *N Engl J Med* **372**: 2235–2242.

Shen L, Diroma MA, Gonzalez M, Navarro-Gomez D, Leipzig J, Lott MT, van Oven M, Wallace DC, Muraresku CC, Zolkipli-Cunningham Z, et al. 2016. MSeqDR: a centralized knowledge repository and bioinformatics web resource to facilitate genomic investigations in mitochondrial disease. *Hum Mutat*. doi: 10.1002/humu.22974.