An Informatics Roadmap Toward a FAIR Understanding of Mitochondrial Biology and Rare Mitochondrial Disease

Garlid, Anders Olav

UNIVERSITY OF CALIFORNIA

Los Angeles

An Informatics Roadmap Toward a FAIR Understanding of Mitochondrial Biology and Rare Mitochondrial Disease

A thesis submitted in partial satisfaction
of the requirements for the degree Doctor of Philosophy
in Molecular, Cellular, & Integrative Physiology (MCIP)

by

Anders Olav Garlid

2019

ABSTRACT OF THE DISSERTATION

An Informatics Roadmap Toward a FAIR Understanding
of Mitochondrial Biology and Rare Mitochondrial Disease

by

Anders Olav Garlid
Doctor of Philosophy in Molecular, Cellular, & Integrative Physiology (MCIP)
University of California, Los Angeles, 2019
Professor Peipei Ping, Chair

Mitochondrial biology is integral to our fundamental understanding of human health and many diseases. They exist in every human cell type except for red blood cells and have critical functions in metabolism, oxidative phosphorylation, oxidation-reduction, and as signaling hubs responsible for mediating protective mechanisms. Rare mitochondrial diseases (RMDs) are devastating and complex, affect multiple organ systems, and disproportionately impact young children. Despite copious existing knowledge and increased public interest, the knowledge is fragmented and difficult to access. Clinical case reports (CCRs) on RMDs contain valuable clinical insights, but they are scarce and lack the metadata necessary to facilitate their discovery among the two million CCRs on PubMed. The unstructured text data of CCRs is also ill-suited to computational approaches, limiting our ability to derive the knowledge contained within.

To address these issues, I assembled all available informatics tools and resources with mitochondrial components and used them to contribute to Gene Wiki pages that enable easy access to mitochondrial knowledge for researchers, students, clinicians, and patients. Through these efforts, I made mitochondrial gene, protein, and disease knowledge widely accessible with contributions of over 4MB of content across 541 Gene Wiki pages. Concurrently, I used Gene Wiki as an educational platform to train over 50 students in the biosciences and pre-medical studies in mitochondrial biology and disease, as well as instilling effective research and writing methods in biomedicine.

To impose structure on CCRs and render them FAIR (Findable, Accessible, Interoperable, Reusable), I developed and applied a standardized metadata template to RMD CCRs and codified patient symptomology with the International Statistical Classification of Disease and Related Health Problems (ICD) system. I created the open-source, cloud-based MitoCases RMD Knowledge Platform (http://mitocases.org/) to house data on 384 RMD CCRs, including 4,561 instances of 952 unique ICD codes. Supplementing CCRs with structured metadata amplifies machine-readable information content and provides a distinct improvement in searching for CCRs as compared to indexing by title and abstract. Finally, I employed these resources to conduct a thorough review of Barth syndrome and characterized the diversity of presentations, range of genetic etiologies, and treatment paradigms.

The dissertation of Anders Olav Garlid is approved.

Alex Ahn-Tuan Bui

Mario C. Deng

Thomas M. Vondriska

Xia Yang

Peipei Ping, Committee Chair

University of California, Los Angeles

2019

DEDICATION PAGE

This thesis is dedicated to my Mom, Randi Brannan, my Dad, Keith Garlid, my brother, Torleif Garlid, Fiona Palmer, Alec Emmons, and in loving memory of Nicole Deland, all of whom have supported me, pushed me, and inspired me to keep striving.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1. INTRODUCTION

CHAPTER 3. A METADATA EXTRACTION APPROACH FOR CLINICAL CASE REPORTS TO ENABLE ADVANCED UNDERSTANDING OF BIOMEDICAL CONCEPTS

CHAPTER 4. A REFERENCE SET OF CURATED BIOMEDICAL DATA AND METADATA FROM CLINICAL CASE REPORTS

CHAPTER 5. A FAIR REPRESENTATION OF MITOCHONDRIAL BIOLOGY AND RARE MITOCHONDRIAL DISEASE VIA GENEWIKI AND MITOCASES

CHAPTER 6. *TAZ* ENCODES TAFAZZIN, A TRANSACYLASE ESSENTIAL FOR CARDIOLIPIN FORMATION AND CENTRAL TO THE ETIOLOGY OF BARTH SYNDROME

LIST OF TABLES

CHAPTER 2. EQUIPPING PHYSIOLOGISTS WITH AN INFORMATICS TOOL CHEST: TOWARD AN INTEGRATED MITOCHONDRIAL PHENOME

CHAPTER 3. A METADATA EXTRACTION APPROACH FOR CLINICAL CASE REPORTS TO ENABLE ADVANCED UNDERSTANDING OF BIOMEDICAL CONCEPTS

CHAPTER 4. A REFERENCE SET OF CURATED BIOMEDICAL DATA AND METADATA FROM CLINICAL CASE REPORTS

CHAPTER 5. A FAIR REPRESENTATION OF MITOCHONDRIAL BIOLOGY AND RARE MITOCHONDRIAL DISEASE VIA GENEWIKI AND MITOCASES

CHAPTER 6. *TAZ* ENCODES TAFAZZIN, A TRANSACYLASE ESSENTIAL FOR CARDIOLIPIN FORMATION AND CENTRAL TO THE ETIOLOGY OF BARTH SYNDROME

ACKNOWLEDGEMENTS

Chapter Two is a reprint of the published version of: Anders O. Garlid, Jennifer S. Polson, Keith D. Garlid, Henning Hermjakob, and Peipei Ping. “Equipping Physiologists with an Informatics Tool Chest: Toward an Integrated Mitochondrial Phenome.” *Handb Exp Pharmacol*. 2017; 240:377-401. doi: 10.1007/164_2016_93. Review. PMID: 27995389. Reprinted with permission from Springer Nature. AOG and JSP contributed equally to researching, writing, and editing the chapter. KDG and HH assisted with revisions. PP conceived of the study and contributed to manuscript revisions.

Chapter Three is a reprint of the published version of: J. Harry Caufield, David A. Liem, Anders O. Garlid, Yijiang Zhou, Karol E. Watson, Alex A.T. Bui, Wei Wang, and Peipei Ping. “A Metadata Extraction Approach for Clinical Case Reports to Enable Advanced Understanding of Biomedical Concepts.” *J Vis Exp*. 2018; (139), e58392. doi: 10.3791/58392. PMID: 30295669. Reprinted as permitted by the publisher’s Creative Commons Attribution 4.0 International License (CC BY). JHC, DAL, and AOG led the study and contributed to study design, writing, and editing of the manuscript. JHC and AOG also contributed the video component of the publication (available in Supplementary Materials and at https://jove.com/video/58392). YZ, KEW, AATB, and WW contributed to study design. PP conceived of the study and contributed to study design and manuscript revisions.

Chapter Four is a reprint of the published version of: J. Harry Caufield, Yijiang Zhou, Anders O. Garlid, Sean P. Setty, David A. Liem, Quan Cao, Jessica M. Lee, Sanjana Murali, Sarah Spendlove, Wei Wang, Li Zhang, Yizhou Sun, Alex A.T. Bui, Henning Hermjakob, Karol E. Watson, and Peipei Ping. “A reference set of curated biomedical data and metadata from clinical case reports.” *Sci Data*. 2018; 5:180258. doi: 10.1038/sdata.2018.258. PMID: 30457569. Reprinted as permitted by the publisher’s Creative Commons Attribution License (CRC 3.0, by, nc, nd). JHC, YZ, and AOG led the study and contributed to study design, generation of the data, data validation, writing, and editing of the manuscript. JHC also contributed to software development. DAL and QC contributed to study design and generation of the data. JML and SM contributed to generation of the data. SS contributed to generation of the data, software development, and data validation. WW, LZ, HH, and KEW contributed to study design. PP conceived of the study and contributed to study design and editing of the manuscript.

Chapter Five covers original and unpublished work by Anders O. Garlid on the Mitochondrial Gene Wiki Project and the MitoCases Rare Mitochondrial Disease Knowledge Platform (http://mitocases.org/). Vladimir Guevara-Gonzalez contributed to platform development. Calvin T. Schaffer, Jaewoo Kim, and Hirsh Bhatt contributed to metadata extraction and Gene Wiki efforts. Peipei Ping is the PI and conceived of the studies.

Chapter Six is the submitted version of: Anders O. Garlid, Calvin T. Schaffer, Jaewoo Kim, Hirsh Bhatt, Vladimir Guevara-Gonzalez, and Peipei Ping. "*TAZ* encodes tafazzin, a transacylase essential for cardiolipin formation and central to the etiology of Barth Syndrome." *Gene*. In review. AOG researched, wrote, and revised the manuscript with assistance from CTS, JK, and HB. VGG contributed data analysis and R scripts for figure generation. PP conceived of the study and contributed to manuscript revisions.

The projects included in this dissertation were supported by the National Institutes of Health through the NIGMS Big Data to Knowledge (BD2K) initiative U54-GM114833 (Ping P); the NHLBI R35-HL135772 (Ping P); the Theodore C. Laubisch endowment at UCLA (Ping P); the NIBIB Medical Imaging Informatics Training Grant T32-EB016640 (2016-2018; Bui A); and the NIGMS Training in Molecular, Cellular, & Integrative Physiology T32-GM065823 (2015-2016; Tidball J). Funding for tuition and graduate stipend was further provided through a TAship with the Life Sciences department at UCLA (2018).

VITA/BIOGRAPHICAL SKETCH

Education:

2014, M.Sc. Mitochondrial Physiology, Dept. of Biology, Portland State University, Portland, Oregon

2011, B.Sc. Chemistry: Biochemistry, Dept. of Chemistry, Portland State University, Portland, Oregon

Teaching Experience:

2018 Spring/Fall T.A., Physiology & Human Biology, Dept. of Life Sci., University of California, Los Angeles

2012-2013 T.A., Principles of Biology, Dept. of Biology, Portland State University, Portland, Oregon

Work Experience:

2015-present, Graduate Research Associate in the laboratory of Dr. Peipei Ping, University of California, Los Angeles, Department of Physiology, Medicine, and Bioinformatics

2014-2015, Assistant to the Director, Peipei Ping, of the HeartBD2K Center of Excellence at UCLA, University of California, Los Angeles, Department of Physiology, Medicine, and Bioinformatics

2013-2014, Research Consultant in the laboratory of Dr. William Stanley, University of Sydney, Australia, Department of Physiology, School of Medical Sciences

2011-2014, Graduate Research Assistant in the laboratory of Prof. Keith Garlid and supervised by Prof. Jason Podrabsky, Portland, Oregon, Portland State University, Department of Biology

2008-2011, Undergraduate Research Assistant in the laboratory of Prof. Keith Garlid, Portland, Oregon, Portland State University, Department of Biology

Professional Honors:

2019, Human Proteome Organization (HUPO) Student Travel Award, "Informatics Map for Understanding Rare Mitochondrial Disease Symptomology" (Poster). HUPO World Congress 2019, Adelaide, Australia

2018, Keystone Symposia Young Investigator Scholarship (National Heart, Lung, and Blood Institute, Grant #1-R13 HL140913-01), "Extracting Clinical Insights from Case Reports of Rare Mitochondrial Diseases" (Poster). Keystone Symposia on Heart Failure, Keystone, CO

2010, Young Investigator Award, "Reactive Oxygen Species (ROS) and Signaling: Which ROS, and how does it move about the cell?" (Poster). 7th Mitochondrial Physiology Conference, Obergurgl, Austria

Scholarships and Sponsorships:

2016-2018, NIH T32 Training Award, "Medical Imaging Informatics Training Grant," University of California, Los Angeles, Department of Radiological Sciences & Bioinformatics, supported by T32 Award EB016640 from the National Institutes of Health (Alex Bui, PI)

2015-2016, NIH T32 Training Award, "Training in Molecular, Cellular, & Integrative Physiology," University of California, Los Angeles, Department of Physiology, supported by T32 Award GM065823 from the National Institutes of Health (James Tidball, PI)

2011-2014, Graduate Research Asst., Portland State University, Dept. of Biology, Portland, OR, supported by grants HL35673 and HL067842 from the National Heart, Lung and Blood Institute (Amir Askari, PI)

Publications:

A.O. Garlid, C.T. Schaffer, J. Kim, H. Bhatt, V. Guevara-Gonzalez, P. Ping (2019). "*TAZ* encodes tafazzin, a transacylase essential for cardiolipin formation and central to the etiology of Barth syndrome." *Gene*. [in review].

J.H. Caufield, Y. Zhou, **A.O. Garlid**, D.A. Liem, Q. Cao, J. Lee, S. Murali, S. Spendlove, W. Wang, L. Zhang, Y. Sun, A.A.T. Bui, H. Hermjakob, and K. Watson (2018). "A reference set of curated biomedical data and metadata from clinical case reports." *Nature Scientific Data.* PMID: 30457569.

J.H. Caufield, D.A. Liem, **A.O. Garlid**, Y. Zhou, K. Watson, A.A.T. Bui, W. Wang, and P. Ping (2018). "A Metadata Extraction Approach for Clinical Case Reports to Enable Advanced Understanding of Biomedical Concepts." *JoVE*. PMID: 30295669.

J. Wang, H. Choi, N. Chung, Q. Cao, D. Ng, B. Mirza, S. Scruggs, D. Wang, **A.O. Garlid**, and P. Ping (2018). "Integrated Dissection of the Cysteine Oxidative Modification Proteome During Cardiac Hypertrophy." *Journal of Proteome Research.* PMID: 30141336.

A.O. Garlid, J.S. Polson, K.D. Garlid, H. Hermjakob, and P. Ping (2016). "Equipping Physiologists with an Informatics Tool Chest: Toward an Integrated Mitochondrial Phenome." *Handb Exp Pharmacol: Pharmacology of Mitochondria,* 240:377-401. 2016 Dec 20. PMID: 27995389.

A.O. Garlid, M. Jaburek, J.P. Jacobs, and K.D. Garlid (2013). "Mitochondrial reactive oxygen species: which ROS signals cardioprotection?" *Am J Phys: Heart & Circ Phys*, 305 (7):H960-H968.

D.B. Foster, A.S. Ho, J. Rucker, **A.O. Garlid**, L. Chen, A. Sidor, K.D. Garlid, and B. O'Rourke (2012). "The Mitochondrial ROMK Channel is a Molecular Component of MitoK$_{ATP}$." *Circ. Res.*, 111:446-454.

External Presentations:

2019 "An Informatics Map for Understanding Rare Mitochondrial Disease Symptomology" (Poster). 27th ISMB / 18th ECCB, Basel, Switzerland

2019 "Instilling FAIR Principles with a Learning Platform for Mitochondria Biology on GeneWiki" (Poster). 27th ISMB / 18th ECCB, Basel, Switzerland

2018 "An Informatics Map to Understanding Mitochondrial Biology and Rare Mitochondrial Disease" (Invited speaker). UCLA MCIP Recruitment Retreat, Calamigos Ranch, Malibu, CA

2018 "Text mining" (Invited course lecturer). Bioinformatics M202 Graduate Lecture, UCLA.

2018 "An Informatics Map for Understanding Rare Mitochondrial Disease Symptomology" (Poster). UCLA QCBio 4th Annual Retreat, Calamigos Ranch, Malibu, CA

2018 "Advancing FAIR Principles and Data Science Awareness via Gene Wiki Training and Education Efforts" (Invited speaker) Intelligent Systems for Molecular Biology, Chicago, IL

2018 "NIH BD2K & Data Science Empowering Biomedical Discovery: Aztec Demo" (NIH BD2K Product Demos) Intelligent Systems for Molecular Biology, Chicago, IL

2018 "NIH BD2K & Data Science Empowering Biomedical Discovery: CaseOLAP Demo" (NIH BD2K Product Demos) Intelligent Systems for Molecular Biology, Chicago, IL

2018 "OPTM Fingerprinting for Deep Phenotyping of Cardiac Mitochondrial Proteome" (Poster) Systems Biology of Human Diseases, QCBio UCLA, Los Angeles, CA

2018 "An Informatics Map for Understanding Rare Mitochondrial Disease Symptomology" (Poster). Systems Biology of Human Diseases, QCBio UCLA, Los Angeles, CA

2018 "Extracting Clinical Insights from Case Reports of Rare Mitochondrial Diseases" (Poster). Keystone Symposia on Heart Failure, Keystone, CO

2016 "Signalosomes orchestrate ouabain-induced intracellular inotropic signaling" (Poster). American Heart Association Scientific Sessions, New Orleans, LA

2016 "Signalosomes orchestrate ouabain-induced intracellular inotropic signaling" (Poster). UCLA Cardiovascular Symposium, Los Angeles, CA

2016 "Adenosine and ischemic signalosomes require activation by reactive oxygen species to mediate cardioprotection" (Poster). International Society for Heart Research, Buenos Aires, Argentina

2016 "Novel Software Tools for Crowdsourcing Mitochondrial Protein Knowledge in Gene Wiki" (Poster). International Society for Heart Research, Buenos Aires, Argentina

2015 "Wiki data 2 Gene Wiki" (Hackathon presentation). 1st BD2K Hackathon, San Diego, CA

2015 "Novel Software Tools for Crowdsourcing Mitochondrial Protein Knowledge in Gene Wiki" (Poster and Invited speaker). International Society for Biocuration, Beijing, China

2014 "The Characterization, Design, and Function of the Mitochondrial Proteome: From Organs to Organisms" (Co-speaker with Amanda Lin). Human Proteome Organization, Madrid, Spain

2014 "The Mitochondriac" (Invited speaker). CVRL members, University of California, Los Angeles, CA

2013 "Cardioprotection by cardiac glycosides is mediated by signalosomes acting on mitochondrial p38 MAP kinase to open mitoK$_{ATP}$" (Poster). Society for Heart and Vascular Metabolism, Cambridge, MD

2012 "Na$^+$,K$^+$-ATPase Signaling to Mitochondria" (Invited speaker). Bioblast: Conference on Mitochondrial Competence, Innsbruck, Austria

2012 "Mitochondrial Reactive Oxygen Species (ROS): Which ROS is responsible for cardioprotective signaling?" (Invited speaker). The Hatter Cardiovascular Institute, University College London, England

2011 "Reactive oxygen species (ROS) in cardioprotection: which ROS does the signaling?" (Poster). International Society of Heart Research, Philadelphia, PA

2010 "Reactive Oxygen Species (ROS) and signaling: Which ROS, and how does it move about the cell?" (Poster). Mitochondrial Physiology Conference, Mitochondrial Physiology Society, Obergurgl, Austria

Chapter I

Introduction:

An Informatics Roadmap Toward a FAIR Understanding of Mitochondrial Biology and Rare Mitochondrial Disease

Introduction

Overview

Mitochondrial biology is integral to our fundamental understanding of human health and many diseases, from exercise to aging, cardiovascular disease to neurological complications, and a wide variety of complex mitochondrial diseases. Despite a large volume of existing knowledge and increased public interest, its access and comprehension remain elusive to the scientific community and public at large. With the intensifying pursuit of omics investigations and recruitment of large-scale clinical cohorts, vast numbers of enormous datasets in a wide variety of data types and inconsistent structure are constantly being generated, all on top of the ever-increasing rate of publication. Clinical communications in the form of case reports are also on the rise, but the information contained within is largely inaccessible by any technological means due to the unstructured nature of text data and the lack of structured and meaningful metadata. These documents present a tremendous opportunity for deriving important clinical insights, but their utility is hampered by the necessity of manual curation and human processing.

Challenges presented by this deluge of biomedical information require significant technological advances in order to increase our capacity to process, analyze, and integrate across data types so that we may derive scientific insight and medical knowledge. Overcoming these challenges also demands the widespread adoption of the FAIR Principles of Findability, Accessibility, Interoperability, and Reusability [1] to render biomedical data and other digital objects machine-readable and increase their overall utility in informatics pipelines and machine learning applications. The projects discussed in the following chapters are driven by a commitment to firmly establish these principles in the study of mitochondrial biology and rare mitochondrial diseases, focusing on informatics tools and resources, imposing structure on unstructured clinical case reports, contributions to citizen science efforts, education, and a new mitochondrial disease knowledge resource. A roadmap to the contents of this thesis is depicted in **Figure 1-1**.

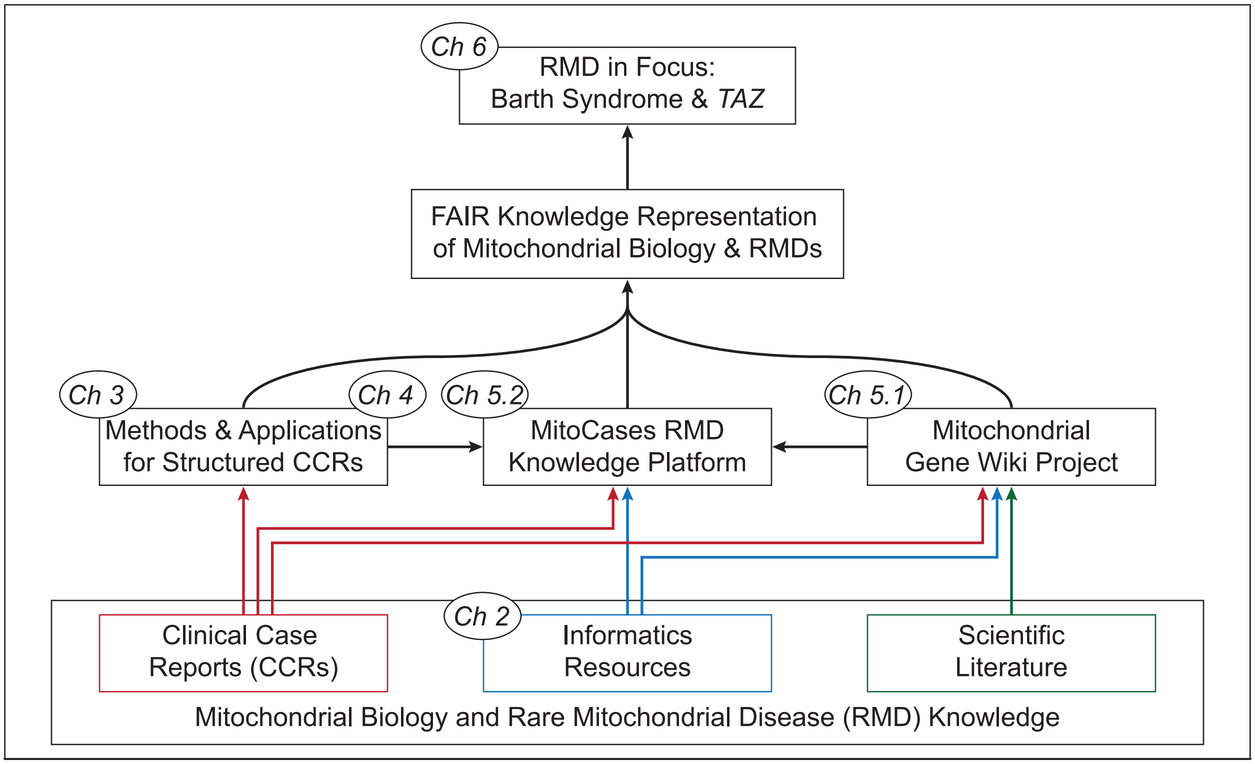


Figure 1-1: Schematic overview of chapters and their interplay. The focus of this thesis is grounded in mitochondrial biology and rare mitochondrial diseases (RMDs), knowledge of which is represented in three primary media: clinical case reports (CCRs), informatics resources, and scientific literature. Chapter 2 **(Ch 2)** reviews informatics tools and resources with mitochondrial components and their utility in modern research. The three sources of mitochondrial knowledge propagate upward to the second tier of entities, each of which form a component of the efforts undertaken in this project. "Methods & Applications for Structured CCRs" *(left)* is contributed to solely by CCRs, for which we developed a standardized metadata template to impose structure on the otherwise unstructured text data contained within those reports. Chapter 3 **(Ch 3)** provides details on the metadata template and instructions for its use by researchers and clinicians, while Chapter 4 **(Ch 4)** presents a collection of metadata extracted from 3,100 reports, characteristics of the dataset, as well as potential use cases for downstream analysis and text-mining applications. The "Mitochondrial Gene Wiki Project" *(right)*, discussed in Chapter 5.1 **(Ch 5.1)**, entails our contributions of over 4MB of content and 5,674 references across 541 Gene Wiki articles on mitochondrial genes and proteins by deriving information from all three sources of mitochondrial knowledge, including CCRs, Informatics Resources, and Scientific Literature. Chapter 5.2 **(Ch 5.2)** details the technical specifications and use cases for the cloud-based "MitoCases RMD Knowledge Platform" *(middle)*, which is contributed to by Informatics Resources and CCRs, as well as Structured CCRs from Chapters 3 & 4 and the Mitochondrial Gene Wiki Project contributions from Chapter 5.1. MitoCases houses structured metadata from RMD CCRs, detailed genetic information, and standardized symptomology records codified with the International Classification of Diseases and Related Symptoms, 10th and 11th revisions (ICD-10 and ICD-11), all searchable by a powerful query system that aids in case discovery and analysis. The efforts from Chapters 2-5 contribute synergistically to rendering the knowledge derived from the three sources of mitochondrial knowledge Findable, Accessible, Interoperable, and Reusable (FAIR). These efforts all feed upwards to contribute to a "FAIR Knowledge Representation of Mitochondrial Biology & RMDs". Chapter 6 **(Ch 6)** makes use of this FAIR mitochondrial knowledge to present an "RMD in Focus: Barth Syndrome & *TAZ*", a comprehensive review of a devastating disease caused by mutations in the *TAZ* gene and the resulting impairments to the tafazzin protein responsible for cardiolipin remodeling. Drawing content and data from efforts presented in previous chapters, Chapter 6 details the structure and function of tafazzin, the role of cardiolipin in mitochondrial structure and function, clinical presentations in Barth syndrome across organ systems, as well as current and proposed treatment paradigms.

Chapter 2. Omics and informatics resources for understanding mitochondrial biology.

Mitochondria play an integral role in all aspects of biology, across cell types, organ systems, and in both health and disease. Gaining a complete understanding of development, aging, signaling, and any other aspect across physiology and pathology requires that we consider mitochondria and the complex nature of its pervasive involvement throughout eukaryotic systems. Since their discovery in 1890 by Altmann [2], who originally described them as "bioblasts," researchers have made significant advances in our comprehension of these ubiquitous organelles and their diverse functions. With each advent of new technological breakthroughs, the interest in and activity around mitochondrial research sees a resurgence and advances our ability to comprehend them.

Part and parcel with each of these advances, the quantity and variety of data becomes ever greater and more complex, presenting significant challenges in managing and assimilating it all into a comprehensive understanding of the mitochondrial phenome. The widespread use of omics technologies and large-scale clinical cohort studies results in massive datasets that are driving the rise of Big Data and amplifying the inherent challenges of integrating high-dimensional and diverse data types. A staggering number of bioinformatics tools and resources are employed throughout research to address these needs, but it can be difficult to determine what tool or resource is appropriate for a particular field of study. It is becoming increasingly more important to have a roadmap to navigate the informatics landscape and find the appropriate tools for any given avenue of investigation.

This chapter presents an overview of mitochondria-specific resources as well as the mitochondria-related subsets contained within other larger tools, databases, and knowledgebases. We have compiled detailed descriptions and analysis of the various informatics resources that are helping researchers explore this vital organelle and gain insights into its form, function, and dynamics. We focus on integrated omics resources, including mitochondria-specific resources, mitochondrial components of general resources, available mitochondrial datasets, as well as analytical tools and computational methods for an informatics approach to understanding and amplifying the study of mitochondrial physiology.

Chapter 3. Methods for structuring text data in clinical case reports.

Clinical case reports (CCRs) are a fundamental mechanism for sharing observations and insights in medicine, facilitating communication between clinicians and providing a vital educational resource for medical students. Emerging diseases, disease subtypes lacking specific diagnostic consensus, and rare clinical presentations are frequently communicated in CCRs, along with their treatments and genetic etiologies [3-6]. The first treatment of rabies in humans by Louis Pasteur in 1885 [7, 8] and the first application of penicillin in patients [9] were reported in CCRs. Heart failure with preserved ejection fraction (HFpEF) may affect half of all heart failure patients [10], but it was only recognized as a distinct subtype of cardiovascular disease in the past few decades. Over two million CCRs are available on PubMed as of August 2019, over a quarter of which were published in the past ten years. These documents represent an invaluable resource for clinical insight across disease types and have great potential to aid in detecting disease trends, identifying at-risk subsets of the population, highlighting promising new therapeutic regimens, and preemptively stemming the damage done by unsafe drugs or treatment practices. The deluge of CCR publications begs for a computer-aided approach to process and analyze these reports as they are released; the sheer volume, complexity, and variety of the reports renders it all but impossible for humans to manage manually. Unfortunately, CCRs lack extensive, structured metadata and the information contained within is primarily unstructured text data, making them inaccessible to machine-based approaches and severely limiting our ability to derive clinical insight from them in an efficient manner.

To address this critical issue, we must envision a mechanism by which we can impose structure on the text data presented in CCRs, render them machine-readable, and thereby gain access to a treasure trove of medical insight as it becomes available. We must establish a standardized system for amplifying the utility of CCRs by supplementing them with extensive metadata and annotations that convey the clinical concepts they describe in a machine-readable format and provide detailed instructions for implementing such a system.

Accordingly, we developed a standardized CCR metadata template and associated methods for annotators to extract detailed clinical information from the reports and establish structure atop an otherwise unstructured text data resource [11, 12]. To ensure the accessibility and usability of this approach, we have published a comprehensive guide to the methodology behind this process to enable others to generate structured CCR metadata as well [12]. An accompanying video of the protocol clarifies the instructions and motivations behind this effort.

Chapter 4. Dissecting informatics blueprints of disease.

As informatics tools and pipelines become ever more prevalent and ubiquitous throughout research, it is imperative that the data and publications we produce are FAIR, structured, and machine-readable so that they can be utilized in these approaches. Natural language processing (NLP) and text-mining approaches require large collections of structured text data in order to successfully train machine learning models for named entity recognition (NER) and relationship extraction. The demand is particularly high in biomedical text-mining due to the complexity of the language and heterogeneity across domains. Medical Subject Headings (MeSH) and related tools [13, 14] facilitate indexing and enforce structure on biomedical documents, as do the curated resources of Informatics for Integrating Biology and the Bedside (i2b2) [15, 16] and the CRAFT [17], AnatEM [18], NCBI disease [19], and PubMed Phrases [20] corpora. Clinical narratives, however, like those presented in CCRs, present unique challenges with their wide variability in content, presentation of interrelated phenomena, and the temporal component of disease progression. Most existing resources for biomedical text-mining have not been designed to model these narratives and the indexing and structure for CCRs is limited. Furthermore, while clinical controlled vocabularies and coding systems including ICD-10 [21], LOINC [22], and SNOMED [23], are in widespread use in electronic medical records (EMRs), these systems are rarely utilized in published clinical reports. The lack of informative metadata to structure and index CCRs represents a significant challenge and hinders the utility of this important resource.

A standardized approach to enriching CCRs with metadata and imbuing them with structure is necessary to address these issues. Collaborative efforts between clinicians, researchers, and data scientists will be required to establish these standards and ensure their implementation across existing and future case report publications. Furthermore, the utility of this metadata must be demonstrated by highlighting an increase in information content and providing use cases that can be implemented by key stakeholders in biomedical research and clinical care. Additionally, the added metadata must be made FAIR so that it can be widely disseminated and integrated into informatics pipelines.

Recognizing the need to lead by example, we created a metadata template to impose structure on CCRs and aggregated metadata from a large collection of reports spanning 16 disease systems and covering over 100 rare diseases. We provide this extensive dataset of over 3,100 structured CCR metadata templates for others to employ in downstream analysis, developing text-mining tools, modeling medical language, and potentially automating the process for application to a more comprehensive set of reports [11]. Detailed use cases for the dataset guide those interested in incorporating it in their work, including researchers, physician investigators, clinicians, data scientists, IP officers, pharmaceutical companies for drug development, and those shaping government policies for clinical trials. The dataset also serves as an educational resource that highlights the variety of CCR writing styles and a range of completeness in describing clinical concepts. By offering a comparison between reports and the benefits of presenting a complete and detailed clinical narrative, we anticipate that clinicians and their co-authors will rise to the challenge and strive to produce better, more structured reports of their own accord using a standardized metadata template.

Chapter 5. Advancing FAIR mitochondrial biology via Gene Wiki and MitoCases.

Mitochondrial diseases are complex, rare, and fatal, frequently leading to disruption of mitochondrial proteomes and function. Unfortunately, decades after their discovery, our limited understanding of these diseases and their pathogenesis is still woefully inadequate, leading to delayed diagnoses and a dearth of treatment options. These issues are compounded by the fragmented nature of clinical case information, much of which is communicated entirely in the unstructured text data contained within clinical case reports. Similarly, much of the biomedical research community's accomplishments in advancing our understanding of mitochondrial biology are difficult to access, particularly for the general public. Wikipedia consistently ranks as one of the most popular websites across the world (currently 5th most viewed) and is widely regarded as a reliable source of scientific information; biomedical articles are both understandable to the average user and useful to those in scientific research for the comprehensive citations. However, many mitochondrial proteins and mitochondrial diseases have been inadequately represented and poorly annotated, posing a roadblock to knowledge discovery in mitochondria-related studies.

FAIR data resources that support knowledge discovery and amplify progress in biomedicine are necessary to advance our understanding of mitochondrial biology and disease, improve our ability to provide timely and accurate diagnoses, and develop effective treatment paradigms. The Gene Wiki Project [24], an effort to inspire citizen science within Wikipedia, embodies FAIR Principles by making complex biomedical knowledge widely available. We have taken a concerted effort to address deficiencies in mitochondrial representation on Wikipedia and, concurrently, realized an educational opportunity to train students on concepts of mitochondrial biology, research methods, and FAIR Principles. Additionally, we set out to decode clinical narratives by imposing structure on CCRs relating to rare mitochondrial diseases (RMDs) using the metadata template discussed in the previous chapters [11, 12] and codified patient symptoms using standardized ICD codes to construct a systematic understanding of symptomology among these diseases. Text data standardization and integration with existing protein resources and clinical ontologies renders metadata FAIR and enables the biomedical community to elevate disease knowledge and improve patient care.

With the participation of 35 high school summer interns, 12 undergraduate college students, and 2 graduate students, we contributed over 4MB of content across 541 articles and added nearly 5,700 references. We introduced critical research strategies and informatics tools to students and instilled FAIR Principles in the next generation of scientists, clinicians, and researchers. To house structured RMD CCR metadata, we built the MitoCases platform (http://mitocases.org/) for 384 reports on 8 RMDs, including deficiencies in complex I through V of the electron transport chain, carnitine deficiency, Barth syndrome, and megaconial-type congenital muscular dystrophy

(MDCMC). Across 384 RMD CCRs, 4,561 instances of 952 unique ICD-10 codes, along with detailed metadata, highlights shared and common symptoms as well as rare and unique characteristics, revealing pathogenesis and mechanistic insights underlying RMDs. MitoCases is a highly structured, FAIR data resource with a powerful search interface through which to discover CCRs of relevance, aiding in literature curation, case review, and diagnosis.

Chapter 6. Rare mitochondrial disease in focus: Barth syndrome and tafazzin.

Mitochondrial diseases have devastating effects on organ systems throughout the body, and they frequently present themselves during infancy or childhood. Their rarity makes them particularly difficult to diagnose and treat due to the lack of collective medical experience on their etiology and presentation. Barth syndrome is a rare mitochondrial disease that effects almost exclusively males and causes significant cardiovascular defects within the first years of life, among other serious metabolic, hematologic, and muscular phenotypes [25, 26]. No treatment exists to address the root cause of the disease. Mutations in the *TAZ* gene impair cardiolipin remodeling by its protein product, the transacylase tafazzin, and compromise mitochondrial structure and function. As a result, the only recourse is to attempt to manage symptoms as the patient progresses towards heart failure or is plagued by severe infections.

Leveraging the resources, tools, and approaches that were uncovered, developed, and applied throughout the preceding chapters, I endeavored to gain a thorough understanding of this particularly interesting and devastating rare mitochondrial disease. I conducted a deep investigation of tafazzin structure and function, explored the complex role of cardiolipin in mitochondrial form and function, and delved into the clinical narratives and genetic etiologies across the population of Barth syndrome patients described in the clinical literature and in the *TAZ* mutations database managed by the Barth Foundation.

The cumulative product of these efforts is the comprehensive review of tafazzin, cardiolipin, and Barth syndrome presented in this chapter, as well as the highly structured clinical and genetic data on these patients that is housed on the MitoCases platform. The review and associated metadata contain a wealth of information on a multi-factorial and inevitably lethal disease with no cure. We present a detailed view of the diverse genetic etiologies for Barth syndrome. Clinicians and researchers will find detailed discussions covering the spectrum of case presentations, including metabolic disorders and pathologies of the cardiovascular, musculoskeletal, neurological, and hematologic systems, as well as the range of treatments that have been attempted, disproven, or proposed for further study.

Closing statement

The chapters within detail efforts to establish “an informatics roadmap toward a FAIR understanding of mitochondrial biology and rare mitochondrial disease” by using and developing tools, resources, and protocols to make mitochondrial knowledge more Findable, Accessible, Interoperable, and Reusable through integration, standardization, and enforcing structure on otherwise unstructured text data.

References:

1. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. and Mons, B. (2016). "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific Data*. 3: 160018. doi:10.1038/sdata.2016.18.

2. Altmann, R. (1894). "Die Elementarorganismen und ihre Beziehungen zu den Zellen," Veit.

3. Ban, T.A. (2006). "The role of serendipity in drug discovery." *Dialogues in Clinical Neuroscience*. 8: 335-44.

4. Bayoumi, A.M. (2004). "The storied case report." *Canadian Medical Association Journal*. 171: 569-570. doi:10.1503/cmaj.1031503.

5. Cabán-Martinez, A.J. and García-Beltrán, W.F. (2012). "Advancing medicine one research note at a time: the educational value in clinical case reports." *BMC Research Notes*. 5: 293. doi:10.1186/1756-0500-5-293.

6. Vandenbroucke, J.P. (2001). "In defense of case reports and case series." *Ann Intern Med*. 134: 330-4. PMID:11182844. doi:10.7326/0003-4819-134-4-200102200-00017.

7. Pasteur, L. (1885). "Méthode pour prévenir la rage après morsure."

8. Pearce, J.M. (2002). "Louis Pasteur and rabies: a brief note." *J Neurol Neurosurg Psychiatry*. 73: 82. PMID:12082056. doi:10.1136/jnnp.73.1.82.

9. Keefer, C.S., Blake, F.G., Marshall, E.K., Lockwood, J.S. and Wood, W.B. (1943). "Penicillin in the treatment of infections: a report of 500 cases." *Journal of the American Medical Association*. 122: 1217-1224.

10. Andersen, M.J. and Borlaug, B.A. (2014). "Heart failure with preserved ejection fraction: current understandings and challenges." *Curr Cardiol Rep*. 16: 501. PMID:24893938. doi:10.1007/s11886-014-0501-8.

11. Caufield, J.H., Zhou, Y., Garlid, A.O., Setty, S.P., Liem, D.A., Cao, Q., Lee, J.M., Murali, S., Spendlove, S., Wang, W., Zhang, L., Sun, Y., Bui, A., Hermjakob, H., Watson, K.E. and Ping, P. (2018). "A reference set of curated biomedical data and metadata from clinical case reports." *Sci Data*. 5: 180258. PMID:30457569. doi:10.1038/sdata.2018.258.

12. Caufield, J.H., Liem, D.A., Garlid, A.O., Zhou, Y., Watson, K., Bui, A.A.T., Wang, W. and Ping, P. (2018). "A Metadata Extraction Approach for Clinical Case Reports to Enable Advanced Understanding of Biomedical Concepts." *J Vis Exp*. PMID:30295669. doi:10.3791/58392.

13. Liu, K., Peng, S., Wu, J., Zhai, C., Mamitsuka, H. and Zhu, S. (2015). "MeSHLabeler: improving the accuracy of large-scale MeSH indexing by integrating diverse evidence."

Bioinformatics. 31: i339-i347. doi:10.1093/bioinformatics/btv237.

14. Mork, J.G., Jimeno-Yepes, A. and Aronson, A.R. (2013). "The NLM Medical Text Indexer System for Indexing Biomedical Literature," BioASQ Workshop.

15. Stubbs, A., Kotfila, C., Xu, H. and Uzuner, Ö. (2015). "Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2." *Journal of biomedical informatics*. 58 Suppl: S67-77. doi:10.1016/j.jbi.2015.07.001.

16. Sun, W., Rumshisky, A. and Uzuner, O. (2013). "Evaluating temporal relations in clinical text: 2012 i2b2 Challenge." *Journal of the American Medical Informatics Association*. 20: 806-813. doi:10.1136/amiajnl-2013-001628.

17. Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D., Baumgartner, W.A., Cohen, K., Verspoor, K., Blake, J.A. and Hunter, L.E. (2012). "Concept annotation in the CRAFT corpus." *BMC Bioinformatics*. 13: 161. doi:10.1186/1471-2105-13-161.

18. Pyysalo, S. and Ananiadou, S. (2014). "Anatomical entity mention recognition at literature scale." *Bioinformatics*. 30: 868-875. doi:10.1093/bioinformatics/btt580.

19. Doğan, R.I., Leaman, R. and Lu, Z. (2014). "NCBI disease corpus: A resource for disease name recognition and concept normalization." *Journal of Biomedical Informatics*. 47: 1-10. doi:10.1016/j.jbi.2013.12.006.

20. Kim, S., Yeganova, L., Comeau, D.C., Wilbur, W.J. and Lu, Z. (2018). "PubMed Phrases, an open set of coherent phrases for searching biomedical literature." *Scientific Data*. 5: 180104. doi:10.1038/sdata.2018.104.

21. World Health Organization (1992). "International classification of diseases and related health problems, 10th revision." Geneva.

22. McDonald, C.J. (2003). "LOINC, a Universal Standard for Identifying Laboratory Observations: A 5-Year Update." *Clinical Chemistry*. 49: 624-633. doi:10.1373/49.4.624.

23. Millar, J. (2016). "The Need for a Global Language - SNOMED CT Introduction." *Stud Health Technol Inform*. 225: 683-5. PMID:27332304. 27332304

24. Huss, J.W., 3rd, Orozco, C., Goodale, J., Wu, C., Batalov, S., Vickers, T.J., Valafar, F. and Su, A.I. (2008). "A gene wiki for community annotation of gene function." *PLoS Biol*. 6: e175. PMID:18613750. doi:10.1371/journal.pbio.0060175.

25. Barth, P.G., Scholte, H.R., Berden, J.A., Van der Klei-Van Moorsel, J.M., Luyt-Houwen, I.E., Van 't Veer-Korthof, E.T., Van der Harten, J.J. and Sobotka-Plojhar, M.A. (1983). "An X-linked mitochondrial disease affecting cardiac muscle, skeletal muscle and neutrophil leucocytes." *Journal of the neurological sciences*. 62: 327-55.

26. Barth, P.G., Valianpour, F., Bowen, V.M., Lam, J., Duran, M., Vaz, F.d.r.M. and Wanders, R.J.A. (2004). "X-linked cardioskeletal myopathy and neutropenia (Barth syndrome): An update." *American Journal of Medical Genetics*. 126A: 349-354. doi:10.1002/ajmg.a.20660.

Chapter II

Equipping Physiologists with an Informatics Tool Chest: Toward an Integrated Mitochondrial Phenome

Equipping Physiologists with an Informatics Tool Chest: Toward an Integrated Mitochondrial Phenome

Anders Olav Garlid, Jennifer S. Polson, Keith D. Garlid, Henning Hermjakob, and Peipei Ping

Contents

Abstract

Understanding the complex involvement of mitochondrial biology in disease development often requires the acquisition, analysis, and integration of large-scale molecular and phenotypic data. An increasing number of bioinformatics tools are currently employed to aid in mitochondrial investigations, most notably in predicting or corroborating the spatial and temporal dynamics of mitochondrial molecules, in retrieving structural data of mitochondrial components, and

A.O. Garlid (✉) • J.S. Polson (✉) • K.D. Garlid
The NIH BD2K Center of Excellence in Biomedical Computing at UCLA, Department of Physiology, University of California, Los Angeles, CA 90095, USA
e-mail: aogarlid@gmail.com; jpolson@g.ucla.edu

H. Hermjakob
The NIH BD2K Center of Excellence in Biomedical Computing at UCLA, Department of Physiology, University of California, Los Angeles, CA 90095, USA
Molecular Systems Cluster, European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI), Cambridge, UK

P. Ping
The NIH BD2K Center of Excellence in Biomedical Computing at UCLA, Departments of Physiology, Medicine, and Bioinformatics, University of California, Los Angeles, CA 90095, USA

H. Singh, S.-S. Sheu (eds.), *Pharmacology of Mitochondria*,
Handbook of Experimental Pharmacology 240, DOI 10.1007/164_2016_93

in aggregating as well as transforming mitochondrial centric biomedical knowledge. With the increasing prevalence of complex Big Data from omics experiments and clinical cohorts, informatics tools have become indispensable in our quest to understand mitochondrial physiology and pathology. Here we present an overview of the various informatics resources that are helping researchers explore this vital organelle and gain insights into its form, function, and dynamics.

Keywords

Computation • Data science • FAIR data • Metadata • Omics • Open access • Visualization

1 Introduction

Mitochondria play a major role in a range of diseases and are promising targets for therapeutic approaches to neurodegenerative and metabolic disorders, ischemia-reperfusion injury and cardiomyopathies, and various forms of cancer (Lesnefsky et al. 2001; Vlasblom et al. 2014). In spite of this, very few mitochondrial drugs have completed clinical trials (Walters et al. 2012). Initially described as "bioblasts" in 1890 by Altmann, mitochondria are of vital importance in eukaryotic systems (Altmann 1894). In the 1940s, Claude developed methods for cell fractionation and differential centrifugation that allowed the separation of mitochondria from the cytosol and a closer study of these fascinating organelles in isolation (Bensley and Hoerr 1934; Claude and Fullam 1945; Claude 1946a, b). He established the localization of respiratory enzymes to the mitochondrion; Kennedy and Lehninger further identified this as the site of the citric acid cycle, oxidative phosphorylation, and fatty acid oxidation (Kennedy and Lehninger 1949). The distinctive dual-membrane characteristic of the mitochondrion and the organization of cristae within its matrix were uncovered in 1952 with the first high-resolution electron micrographs of isolated mitochondria by Palade (1952, 1953) (Daems and Wisse 1966). Imaging approaches also provided the means for the discovery of mitochondrial DNA (Nass 1963; Nass and Nass 1963; Schatz et al. 1964) which ultimately became the first component of the human genome to be completely sequenced (Anderson et al. 1981) more than 20 years before the massive Human Genome Project completed their goal of sequencing the entire human genome (Lander et al. 2001).

Technological advances and physiological discoveries have fomented renaissance periods in mitochondrial research. A notable example is the work of Peter Mitchell leading to the chemiosmotic theory (Mitchell 1961, 1966, 1968). Eventually accepted 20 years after its introduction, Mitchell's proposal broadened the scope from basic molecular biology and biochemistry to a more physiological approach, including explorations of the intricacies of the potassium cycle and matrix volume regulation through the study of channel and carrier membrane biology. These approaches have become progressively more advanced, leading to a greater understanding of mitochondrial physiology and its role in health and disease. The scientific community is now progressing from the classical, organelle-based study of mitochondria to the digital era of the informatics-based mitochondrial phenome. Informatics combines computer science, engineering, and statistics to develop tools and methods that empower researchers to navigate biological data, put it into context, and extract knowledge. Under the increasing influx of data, prowess in data science methods and awareness of available informatics resources are becoming essential components of the modern researcher's tool chest.

The most common high-throughput data come in the form of omics datasets, which assess the global conditions within the cell through one dimension of data, be it whole-genome sequencing for genomics, RNA-sequencing (RNA-seq) for transcriptomics, tandem mass spectrometry (MS/MS) for proteomics, or NMR spectroscopy for metabolomics studies (Field et al. 2009). Integrating levels of expression across many dimensions can elucidate physiological mechanisms underlying healthy and diseased phenotypes not revealed or biasedly portrayed by a single dimension. Developments in these technologies have manifested a dramatic increase in the sheer volume of publicly available scientific data; genomics data has been growing at a rate that exceeds Moore's law by a factor of 4 since 2008 (Gomez-Cabrero et al. 2014; O'Driscoll et al. 2013). This deluge of data has presented unique and unforeseen challenges. Omics investigations rely heavily on effective database management and annotation to contextualize molecular data and infer biological significance through statistical enrichment and class discovery techniques. The completeness and precision of existing annotations are therefore instrumental to harness omics techniques for disease phenotyping and mechanistic investigations. In light of these challenges, funding agencies, research organizations, and publishers around the world are adopting FAIR data principles, which maintain that data should be findable, accessible, interoperable, and reusable (FORCE11 2014). Building on these principles, omics datasets and analysis platforms must also be citable and scalable to allow the attribution of the work to the appropriate research groups and expansion to new and improved techniques with larger capacity. In this changing landscape, the concept of open-access data is gaining traction, asserting that data should be freely available for researchers to use, reuse, and disseminate (Molloy 2011).

This chapter provides an overview of the informatics tools and resources available to the modern researcher and how they may be used to inform a greater understanding of mitochondrial physiology. We focus on integrated omics resources, including mitochondria-specific resources, mitochondrial components of general resources, available mitochondrial datasets, as well as analytical tools and computational methods for an informatics approach to mitochondrial physiology.

2 Mitochondria-Specific Resources

The mitochondrial research community has undergone several paradigm shifts in conceptual focus and experimental design, progressing from hypothesis to data-driven approaches. The discovery of mitochondrial DNA (mtDNA) fostered a period of renewed interest in mitochondrial physiology and the development of new tools and techniques. Mitochondrial DNA was first discovered in 1963 by Nass et al., who described them as "intramitochondrial fibers with DNA characteristics" (Nass 1963; Nass and Nass 1963). The entirety of the human mitochondrial genome was sequenced in 1981 by Anderson et al. (1981), and mutations were first discovered less than a decade later, identifying the genetic basis of LHON, Kearns-Sayre, MELAS, and MERRF (Wallace et al. 1988a, b). mtDNA is 3,000 kb long and encodes only 13 proteins in the mitochondrial proteome, compared to over 1,500 from nuclear DNA. Aberrations in the mitochondrial genetic code result in diseases that are most often fatal, demonstrating their integral role. Maternal inheritance of mtDNA has provided a method for genotyping and tracing of genetic lineages. Computational approaches have been employed to identify genetic pressures for phylogenetic retention of mitochondrial genes as well as the debated mtDNA bottleneck mechanism, whereby cell–cell variability is utilized to avoid aggregation of deleterious mutations and loss of function of the uniparental mtDNA (Johnston and Williams 2016; Johnston et al. 2015).

Characterizing the mitochondrial genome and corresponding proteome requires a tailored approach due to its unique quality of having genetic contribution from both nuclear and mitochondrial DNA. The databases that contain these datasets must delineate the genetic sources, and the protein localization information can vary between datasets and detection methods. As such, the mitochondrial community has created a variety of tools to decipher the principles of mitochondrial physiology. In this section, we highlight mitochondria-specific resources geared exclusively toward the storage, maintenance, and manipulation of mitochondrial datasets, from "omics" repositories to community-curated resources of mitochondrial knowledge.

In genomics, analyses of an individual's genome is compared to a reference genome to determine individual genetic variations and aberrations; accordingly, the Cambridge Reference Sequence (CRS) was created in 1981 (Anderson et al. 1981), and updated in 1999 (Andrews et al. 1999). This reference sequence is stored within GenBank [NCBI Reference Sequence: NC_012920.1] and within MITOMAP, which provides information relating specifically to the human mitochondrial genome; this includes polymorphisms, mutations, and control regions, and allows users to upload and analyze sequences through the MITOMASTER web interface (Lott et al. 2013). Similarly, MitoCarta 2.0 provides a curated inventory of 1,158 human and mouse genes, as well as the proteins that localize to the mitochondrion. The inventory is generated using mass spectra of mitochondria isolated from 14 tissues and protein localization is determined via GFP tagging, microscopy, and machine learning. MS and microscopy results are integrated with six other genome-scale datasets of mitochondrial localization, lending greater accuracy to the determination of protein location (Pagliarini et al. 2008). Importantly, MitoCarta has integrated information from archived mitochondrial databases, such as MitoP2, in order to ensure that the knowledge contained within said databases remains accessible (Calvo et al. 2016). Datasets can be downloaded in Excel, MySQL, BED, and FASTA file formats and are publicly available. Datasets within MitoCarta have led to important insights such as the identification of targets for whole-exome sequencing disease analysis (Falk et al. 2012).

Characterizing the mitochondrial genome is particularly important in the context of human disease, where maternally inherited mutations can lead to deadly diseases such as MERAS and MERFF (Wallace et al. 1988a, b). Computational tools have been essential in characterizing these mutations (see Table 1). MitImpact is a repository of pathogenicity predictions as related to mitochondrial DNA mutations. Predictions are generated by assembling estimations as well as structural and evolutionary annotations for each missense mutation. The resource is comprehensive, and provides assessments of susceptibilities for previously characterized and unknown mutations resulting in amino acid sequence alteration (Castellana et al. 2015). Mutations currently characterized across populations are stored in the Human Mitochondrial DataBase (HmtDB), which focuses on mitochondrial diseases in population genetics (Rubino et al. 2012). The genome sequences within HmtDB are annotated based on population variability factors, using SiteVar software. Users can query and browse the database, analyze sequences for classification, and download the results with reference genomes. As of August 2016, HmtDB contains 28,196 complete normal genomes spanning multiple continents, and 3,539 complete patient genomes. Globally, the database contains data for close to 10,000 variant sites (Attimonelli et al. 2005). For more deleterious mutations, MitoBreak contains mitochondrial genome rearrangements comprising circular deletion, circular duplication, and linear breakpoints. Spanning seven species including human, each case lists the positions of the breakpoints, junction sequences, and clinical relevance found in publications. The resulting resource is crucial for studying structural alterations of mtDNA (Damas et al. 2014). MitoSeek is a software tool for obtaining various mitochondrial genome information from exome, whole genome, and RNA-seq data (Guo et al. 2013). The tool can be utilized for mitochondrial sequence extraction, assembly quality evaluation, relative copy number estimation, detection of mitochondrial heteroplasmy, somatic mutations, and structural mtDNA alterations (Jayaprakash et al. 2015). These mitochondria-specific tools have enabled greater efficiency and standardization in the analysis of genomics datasets.

Just a few short years after Marc Wilkins coined the term “proteome” in 1995 (Wasinger et al. 1995; Godovac-Zimmermann 2008), Rabilloud attempted the first characterization of all mitochondrial proteins using 2-D electrophoresis (Rabilloud et al. 1998). In the ensuing decades, tremendous progress has been made in defining the mitochondrial proteome and its subproteomes (Lotz et al. 2014), owing primarily to the remarkable developments in mass spectrometry technology (Yates 2013). To house the massive amount of data generated in these studies, MitoMiner is used as a data aggregator to store and analyze mitochondrial proteomics data obtained from MS and fluorescent protein tagging studies (Smith et al. 2012). It integrates with many other informatics resources, namely UniProt, Gene Homology, Online

Table 1 Mitochondrial resources and websites

Tool	URL	Description	Last updated
HmtDB	http://www.hmtdb.uniba.it:8080/hmdb/	Open resource hosting human mitochondrial genome sequences annotated with population and variability data, the latter being estimated through the application of SiteVar	09/2015
MitImpact	http://mitimpact.css-mendel.it/	Repository of pathogenicity predictions. These predictions are created through the assembly of precomputed and computed sets of estimations for all missense mutations; these mutations are then structurally and evolutionarily annotated	07/2016
MitoBreak	http://mitobreak.portugene.com	Database containing mitochondrial genome rearrangements through a list of circular deletion, circular duplication, and linear breakpoints	05/2014
MitoCarta 2.0	http://www.broadinstitute.org/node/7098/index.html	Provides a curated inventory of 1,158 human and mouse genes encoding proteins with strong scientific support of localization to the mitochondrion	06/2015
MitoFish/ MitoAnnotator	http://mitofish.aori.u-tokyo.ac.jp/	Contains the mitochondrial genomes of many model systems, including zebra fish. The database also contains phylogenetic information, as lineage can often be determined via mitochondrial DNA. MitoAnnotator automates the annotations of new sequences uploaded to the database, and has also reannotated the previously uploaded mitogenomes to gain new insights	08/2016
MITOMAP	http://www.mitomap.org/MITOMAP	Provides information relating specifically to the human mitochondrial genome, including polymorphisms, mutations, and control regions, and allows users to upload and analyze sequences through the MITOMASTER web interface	06/2016
MitoMiner	http://mitominer.mrc-mbu.cam.ac.uk	Data aggregator for the storage and analysis of mitochondrial proteomics data obtained from MS and fluorescent protein tagging studies	04/2016
MitoPedia	http://www.bioblast.at/index.php/MitoPedia	Encyclopedic resource and discussion platform specifically focused on mitochondrial knowledge relating to experimental design, methods, and terminology	05/2016

(continued)

Table 1 (continued)

Tool	URL	Description	Last updated
MitoSeek	https://github.com/riverlee/MitoSeek	Software tool for obtaining various types of mitochondrial genome information from exome, whole genome, and RNA-seq sequencing data	05/2015

Mendelian Inheritance in Man, HomoloGene, Integrated Mitochondrial Protein Index, KEGG, and PubMed. As such, MitoMiner provides an all-in-one platform for mitochondrial researchers interested in probing the mitochondrial proteome. MitoMiner currently encompasses 11 different species and integrates 46 large-scale proteomics studies in its database, providing output data in XML, JSON, GFF3, UCSC-BED, FASTA, and HTML formats, and programmatic access through REST APIs and platform-specific clients (Perl, Python, Ruby, and Java). Most importantly, MitoMiner is actively maintained and updated to accommodate changes to the integrated resources.

Other databases have been created for specific animal models, such as MitoFish. MitoFish contains the mitochondrial genomes of many fish species, including the common model system, zebra fish. The database also contains phylogenetic information, as lineage can often be determined via mitochondrial DNA. MitoAnnotator automates the annotations of new sequences uploaded to the database, and has also reannotated the previously uploaded mitogenomes to gain new insights (Iwasaki et al. 2013). MitoFish is particularly useful to mitochondrial researchers due to expert curation and automated annotation. Data contained within MitoFish has spurred efforts to determine the genetic basis for various adaptations in fish (Wang et al. 2016) as well as advancements in phylogeographic studies (Hirase et al. 2016). This includes the development of suffix tree-based marker detection methods for detecting short genetic sequences, resulting in improved approaches to annotating mitochondrial genomes or to detecting and correcting erroneous annotations (Moritz et al. 2014).

Collaboration among domain-specific communities is integral to creating studies for emerging physiological questions. Created in 2010 by Bioblast, MitoPedia was created as an encyclopedic resource and discussion platform specifically focused on mitochondrial knowledge relating to experimental design, methods, and terminology. Content is generated by contributions from domain scientists and mitochondrial physiologists with experience in cellular and mitochondrial isolation and experimentation. Experts in the field write, discuss, and contribute to articles relating to respirometry, fluorometry, spectrophotometry, mitochondrial swelling, membrane potential (Δψ), and ion flux experiments. Members of the Mitochondrial Physiology Society (MiPs) comprise the primary user base of the MitoPedia platform, which has been accessed over 40,000 times, with approximately 100–200 page views per month (Oroboros 2015). Many of the articles presented deal with respirometry experiments and the MiPs group actively endorses a move to standardized experimental protocols, drug concentrations, and terminology so as to have the most effective discussions among mitochondrial physiologists across the world.

Use Case for Investigating Mitochondrial DNA Mutations

Biomedical question: An investigator would like to study the role of mtDNA mutations in Leber's hereditary optic neuropathy (LHON), whether polymorphisms exist predominantly for a particular demographic, as well as information on current treatment efficacy and clinical trials.

Data science approach: Use HmtDB to find sequences, population genetics, and polymorphisms in human mitochondrial genomes. Searching for LHON returns 190 records in healthy and diseased patients. The same search on MITOMAP yields 61 selected references, and MitoMiner was recently used to identify mitochondrial proteins that are downregulated in LHON patients due to an mtDNA mutation (Tun et al. 2014). MitoPedia contains 13 entries of references and abstracts relating to LHON and mitochondrial function. Finally, clinicaltrials.gov (discussed in Sect. 3) lists a current Phase 2 randomized clinical trial on a small cohort (12 patients) "Investigating the Safety, Tolerability, and Efficacy of Elamipretide (MTP-131) Topical Ophthalmic Solution for the Treatment of Leber's Hereditary Optic Neuropathy" (https://clinicaltrials.gov/ct2/show/NCT02693119/). Using the patient records from HmtDB, one could determine differential prevalence across demographics, or predict other mutations that would result in a similar phenotype using the MitImpact tool.

3 General Bioinformatics Resources and Their Mitochondrial Components

Approaches to the study of mitochondrial physiology have undergone tremendous change with the advent of omics approaches and cloud computing, among other technologies and advancements. The resulting influx of information has the potential to generate vast amounts of knowledge, but only with the proper infrastructure in place to handle the load. Bioinformatics and cloud computing approaches allow more efficient and effective management of the wide variety of data sources that contribute to our generation of physiological knowledge and a greater understanding of mitochondria. Here we review the mitochondrial components of well-established, curated omics resources (see Table 2).

Many resources span multiple dimensions, providing information on two or more omics data types. Xfam is a collection of databases including Rfam for RNA families (Nawrocki et al. 2015), Pfam for protein families (Finn et al. 2014a), and iPfam for protein family interactions (Finn et al. 2014b). Each database provides annotations that are crowdsourced through Wikipedia and links to other databases for more information

Table 2 Mitochondrial entries in existing big resources

Tool	URL	Data type(s)	Mitochondrial relevant entries
Flybase	http://flybase.org/	Fly genes, mutations, and stocks	377 genes, 11,057 stocks
HMDB	http://www.hmdb.ca/	Metabolites	17,682 metabolites
IMSR	http://www.findmice.org/	Mouse strains	4,011 strains
KEGG	http://www.genome.jp/kegg/	Pathway maps	55 relevant pathway maps
PDB	http://www.rcsb.org/pdb/	Protein structures	2,107 protein structures
Reactome	http://www.reactome.org/	Reactions and pathways	153 human pathways
UniProt	http://www.uniprot.org/	Proteins	4,889 reviewed proteins
Xfam	http://xfam.org/		
Rfam	http://rfam.xfam.org/	RNA families	RNA families
Pfam	http://pfam.xfam.org/	Protein families	1,460 proteins
iPfam	http://ipfam.org/	Protein family interactions	325 protein families, 44 ligands

on the protein sequence, protein structure, or RNA sequence of interest. Rfam contains information about noncoding RNAs (ncRNAs), structured cis-regulatory elements, and self-splicing RNAs. Each entry is represented by multiple sequence alignments, consensus secondary structures, and covariance models (CMs), which allow simultaneous modeling of RNA structure and sequence (Nawrocki et al. 2015). Currently, there are 861 mitochondrial RNA families within this database. Pfam utilizes hidden Markov models (HMMs) to generate multiple protein sequence alignments, allowing users to search sequence databases for homologous proteins with a specialized computational package. Sequence information is organized into higher level groupings of related families called clans, based on collections of Pfam-A entries related by sequence similarity, structure, or profile HMM (Finn et al. 2014a). Pfam has over 16,000 manually curated entries to date, 1,460 of which are annotated as mitochondrial. iPfam provides protein interaction information, based on structural information from all known structures contained in the Protein Data Bank (PDB) (Berman et al. 2000). Protein crystal structure is analyzed to identify protein domains, bonds, and small chemical ligands in each structure and bond length is estimated based on geometric and chemical properties of the sites (Finn et al. 2014b). There are 325 protein families and 44 ligands within iPfam that are mitochondrially related.

The Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) is the premier resource for protein structure data. The PDB contains over 120,000 crystallographic structures of proteins, nucleic acids, and protein/

nucleic acid complexes, along with complementary analysis and visualization tools (Berman et al. 2000). The PDB also accepts data depositions from the community and provides tools for data preparation and validation. Most data provided by the PDB is available as PDBML or PDB files, and is accessible over two REST web services: the SEARCH service for querying the database and the FETCH service for retrieving the structure data. Over 2,000 of the structures within PDB are mitochondrially associated. Structural information about these proteins yields valuable insight into their conformational arrangements and sites of posttranslational modification, and can help identify active sites for drug development.

The PDB assists in these developments by providing a stable, organized resource for accessing and sharing structural data. To supplement the structural data within PDB, users rely on UniProt, a publicly accessible, comprehensive resource of protein sequence, annotation, and localization data created and maintained by the European Bioinformatics Institute. UniProt contains both highly curated and machine-generated protein annotations, and has become the de facto source for protein information. As of September 2016, this resource contains 550,000 manually reviewed proteins for human (SwissProt), of which 4,889 are related to the mitochondria.

Perturbation in protein expression is a crucial component of omics research; the manifestations of such changes can be studied through metabolomics, which assess the entire pool of small molecules within the cell or organelle. As proteins act on metabolites, their fluctuations provide another, highly dynamic dimension of phenotypic data. One database, the Human Metabolome Database (HMDB), contains or links chemical data, clinical data, and molecular biology/biochemistry data (Wishart et al. 2007). The database has detailed, curated information for 41,993 water- and lipid-soluble metabolites. Within each entry, there are 110 data fields, with a large emphasis on chemical and clinical data; significantly, quantification information is available on each MetaboCard entry, and there exists protein sequence information for 5,701 entries. The HMDB supports extensive text, sequence, chemical structure, and relational query searches (Wishart et al. 2009). The database interfaces with many others, including some of those aforementioned, as well as PubChem, MetaCyc, ChEBI, and GenBank (Wishart et al. 2013). Of the over 40,000 metabolites within the database, mitochondrially related metabolites comprise over 17,000 MetaboCard entries.

Metabolite fluctuations are best understood through visualization of their respective pathways, as it is becoming increasingly apparent that their complex interplay is a burgeoning area of investigation. Two databases provide pathway visualization tools: KEGG PATHWAY and Reactome. The Kyoto Encyclopedia of Genes and Genomes (KEGG) currently contains 17 databases, each with a different focus in the domains of systems biology, genomics, chemistry, and medicine. KEGG is publicly accessible and supports tab-separated, plain text or KEGG database entry data formats, while the KEGG Application Programming Interface (API) allows customization of KEGG analyses (Kanehisa et al. 2014). These tools provide an in-depth look at signaling pathways and interactions among proteins, allowing a more complete view of pathways of interest, providing insight into related processes and species that might deserve study. Specifically, within KEGG

PATHWAY, there are 55 mitochondrial relevant pathway maps, many of which are associated with the progression of common human diseases. Another resource, Reactome, is an open-source platform for biological pathway visualization that is manually curated and peer reviewed by experts, with a focus on human reaction pathways (Matthews et al. 2009). The database contains deeply annotated pathway information from 19 distinct species and includes 1,786 pathways for human as of September 2016. Of these, there are over 150 mitochondrially associated pathways, reactions, and complexes identified and annotated within the platform. Reactome bundles many pathway-related visualization, interpretation, and analysis tools in one resource (Stein 2004). The pathway data can be viewed and analyzed directly from the Pathway Browser, accessed programmatically through a REST API, or downloaded in BioPAX, PSI-MITAB, SBML, and SBGN formats (Jupe et al. 2015). Reactome is widely used for different analyses, including the identification of biomarkers in a neurological model of PTSD (Jia et al. 2012; Bai et al. 2007).

Use Case for Investigating Mitochondrial Localization or Involvement

Biomedical question: 12-Lipoxygenase has been implicated in ischemic preconditioning pathways, but has it been identified as having mitochondrial localization? What resource can an investigator use to answer this question? Are lipoxygenases mitochondrially targeted? What metabolic changes result from perturbations in lipoxygenase expression or function? What interacting partners exist or is there any relevance as a drug target?

Data science approach: UniProt reveals a GO annotation of positive regulation of mitochondrial depolarization. Querying HMDB for 12-lipoxygenase returns five metabolites that are associated with the enzyme. These results also connect the user to a multitude of resources, including primary research papers that have been used for annotation and other sites or resources with relevant information. KEGG and Reactome can also be used to investigate reaction pathways related to the lipoxygenases; a KEGG query for 12-lipoxygenase returns eight pathways and Reactome lists seven reactions and their corresponding pathways.

4 Public Mitochondrial Datasets and Data Sharing

The digitization of scientific literature through resources like PubMed has made scientific publications easy to find for a much wider audience. However, due to the incremental nature of science, where new studies are based on the conclusions reached by past endeavors, knowledge remains distributed across many publications. This creates a gap in our ability to reuse or repurpose existing knowledge and renders high-throughput computational analyses of the literature immensely difficult. The growing demand for access to the data behind scientific publications assigns data repositories an increasingly important role as the backbone of modern scientific research. Sponsored by the National Institutes of Health, the National Center for Biotechnology Information (NCBI) created the Database of Genotypes and Phenotypes, or dbGaP, which contains hierarchical data of structured types. The inputted data are organized into investigations that explore the interface of genotype and phenotype and are linked to an accession number. Within the studies are phenotypic datasets, whereby certain variables are measured. Investigators can upload the raw datasets, provide important information, and give metadata to encourage reusability. If investigators performed analyses, those may also be uploaded. Upon study completion, investigators may upload supporting documents, which may contain further information such as study instructions, protocols/forms for data procurement, and other information necessary to use these datasets (Tryka et al. 2014). Conducting a query for "mitochondria" returns 21 mitochondrial related studies spanning 216 variables. Within this, there are ten supporting documents and 37 raw datasets. These are also documented in clinicaltrials.gov. Currently, the database is openly accessible to institutions, and there is an avenue by which individuals can gain controlled access. These resources are aggregated in Table 3.

As the creation of new data exponentially grew, the scientific community realized the need for consolidated, open omics repositories. One unique challenge was to take raw, unprocessed datasets and provide information and resources to enable

Table 3 Publicly accessible mitochondrial datasets

Dataset	URL	Description	Mitochondrial datasets and components
dbGaP	http://www.ncbi.nlm.nih.gov/gap	The database of genotypes and phenotypes contains hierarchical data of structured types	21 studies: 216 variables 1 analysis 10 documents 37 datasets
ProteomeXchange	http://www.proteomexchange.org/	Publicly accessible, centralized repository for proteomics datasets	32 datasets
OmicsDI	http://www.omicsdi.org/	Provides searchable metadata such as accession, description, sample/data protocols, and biological corroborations	2,819 total datasets: 326 proteomics 24 metabolomics 63 transcriptomics 122 genomics 208 multi-omics

access and collaborative analysis by other users. Founded by the creators of the prominent primary proteomics resource PRIDE (PRoteomics IDEntifications), the ProteomeXchange Consortium sought to create a repository for proteomics mass spectrometry datasets. The consortium consists of investigators who created many primary and secondary proteomics resources, and now has a publicly accessible, centralized repository (Vizcaino et al. 2014). Users can proceed through a full or partial submission workflow, using PRIDE as the initial resource. The PX Submission Tool facilitates the upload of datasets, as well as supporting documentation and metadata (Ternent et al. 2014). Of the over 2,500 datasets compiled on the ProteomeXchange resource, there are 32 proteomics datasets that are related to mitochondria. While many repositories exist that are specific to one omics domain, they are fragmented, operate independently, and highlight the need for an integrated repository, whereby omics datasets can be organized and disseminated in a systematic fashion. Within the Big Data to Knowledge (BD2K) Initiative, there exists OmicsDI, a repository for multiple datasets. One common challenge within omics datasets is that data is generated through differential protocols; as such, this hinders the analyses that can be performed. OmicsDI combats this by providing searchable metadata such as accession, description, sample/data protocols, and biological corroborations (Perez-Riverol et al. 2016). Querying OmicsDI for mitochondrial terms yields 2,800 datasets: of those, 122 are genomic, 63 are transcriptomic, 326 are proteomic, 24 are metabolomic, and 208 are multi-omic datasets.

The development of informatics resources for storage, maintenance, and analysis of scientific data addresses each component of the FAIR data doctrine. Standardization of experimental protocols in the “omics” sciences is necessary for more reliable data, while consistency in data formats and annotation facilitates more efficient data sharing. Curated resources and data repositories organize the information, rendering “omics” data findable and accessible, and lending added value to the data that might otherwise be undiscovered in an obscure format. Furthermore, analytic tools and platforms have been developed by the scientific community to facilitate the goals of interoperability and reusability. Analytic platforms make data mining and the generation of metadata much easier and more widely accessible, furthering data discovery and allowing others to try a novel approach on any of the available datasets. Scientists benefit from making their datasets publicly available and in a common format because their work becomes more visible, useable, and, ultimately, citable.

In addition to the technical challenges of data sharing, there exist legal and cultural considerations, especially with respect to clinical studies. One readily identifiable challenge is sharing clinical datasets in accordance with the Health Insurance Portability and Accountability Act (HIPAA), which mandates the protection of privacy with regard to medical information. Research efforts must establish protocols that de-identify information contained within electronic health/medical records (EHR/EMR), and package them into usable, queryable data for research purposes (Russo et al. 2016). More important, however, is the need to secure explicit consent from study participants such that their personal data can be shared across groups with different research foci. Currently, protocols are being developed to provide standardized consent forms for larger scale purposes with wide applicability, rather than for individual studies. New studies benefit from incorporating these stipulations into consent documents during study design, whereas researchers wishing to disseminate older datasets must review their consent documents to determine if sharing is appropriate or legal. Alternatively, a follow-up may be necessary to gain further consent from study participants (Kaye and Hawkins 2014). Already, research groups and agencies have created repositories with these issues in mind.

Because these data repositories are general and allow the storage of datasets originating from a broad range of biological fields, researchers may question the utility of those that were not originally constructed with a mitochondrial focus. Previously, datasets were considered disposable material, only useful for supporting the conclusions of the particular study. In the data science landscape, new methods and protocols have been developed for managing and cataloging data, providing researchers the ability to recycle and reuse datasets. There lies untapped potential in these datasets for mitochondrial researchers to explore and gain new insights, all while avoiding expensive and laborious data generation until it is necessary. This is especially applicable to large-scale omics datasets, which provide the opportunity for data-driven, untargeted approaches to streamline experimental design and generate targeted approaches for further hypothesis-driven, physiological studies.

5 Pipelines and Tools for Data Processing and Analysis

While access to datasets is crucial, it is only one step in being able to truly harness the knowledge contained within them. The first and often most laborious aspect is to process the data and wrangle it into a format that will be usable by intended tools and pipelines. Investigative teams often perform their own processing protocols, and do not leave adequate metadata and/or annotation for others to be able to recreate the study. To mitigate this, some omics investigators have sought to create best practice pipelines. The effects of a unified system and set of tools can be most concretely seen in the area of genomics, specifically with the Genome Analysis ToolKit (GATK), which focuses on sequence analysis to discover relevant variants (McKenna et al. 2010). Having established best practices for genomics analyses, GATK has also been expanded to have copy number and structural variation analysis capacities (Tennessen et al. 2012). Currently, GATK hosts a wide variety of tools to assist investigators along all steps of analysis, including sequence data processing tools, such as sequence aligners and readers; variant discovery, evaluation, and manipulation tools, such as a Bayesian genotyper and a variant filter; and annotation modules. The toolkit was originally intended for human exomes and genomes sequenced from Illumina technologies, but has been expanded to other experimental protocols and model systems, regardless of ploidy. GATK also hosts third-party tools that can be integrated with the previously established pipelines (DePristo et al. 2011; Van der Auwera et al. 2013). GATK and other analysis tools and pipelines are highlighted in Table 4.

Table 4 Analysis pipelines and platforms

Tool	URL	Description
COPaKB	https://amino.heartproteome.org/	Centralized platform featuring high-quality cardiac proteomics data and relevant cardiovascular phenotype information. The organellar modules constitute the mass spectral library and are utilized by COPaKB's unique high-performance search engine to identify and annotate proteins in the mass spectra files that are submitted by the user in mzML or DTA formats
Cytoscape	http://www.cytoscape.org/	Open-source software platform for the visualization of complex biological systems, such as molecular interaction networks and biological pathways. The platform enables enhancement of the network data through integration of various formats of metadata into the network structure
Galaxy	https://usegalaxy.org/	Informatics workflow management system and data integration platform that aims to make computational biology accessible to researchers with limited experience in computer programming. Provides a graphical user interface, customizable plug-ins, access to public datasets, and other users' workflows
GATK	https://software.broadinstitute.org/gatk/	Toolkit that focuses on sequence analysis to discover relevant variants; originally intended for human exomes and genomes sequenced from Illumina technologies, but has been expanded to other experimental protocols and model systems, regardless of ploidy
MetaboAnalyst	http://www.metaboanalyst.ca/	Analysis pipeline spanning a wide variety of data types; has the capacity for biomarker identification, as well as a host of other bioinformatics tools for the best standard metabolomics analyses using an extensive spectral library for enhanced metabolite identification
MetaCore	https://portal.genego.com/	Standalone program and Web application comprising multiple different analysis methods for varying types of high-throughput molecular data, including sequencing and gene expression, proteomic data, and metabolomic data. Also contains a manually curated database
MetazSecKB	http://bioinformatics.ysu.edu/secretomes/animal/index.php	Database presenting subcellular protein location based on manual curation of the scientific literature combined with UniProt sequence data and annotations. Employs an algorithm that utilizes multiple prediction tools, combining the predictions of publicly

(continued)

Table 4 (continued)

Tool	URL	Description
		available tools including TargetP, SignalP, Phobius, ScanProsite, WolfPSORT, FragAnchor, and TMHMM
OmicsPipe	http://sulab.org/tools/omics-pipe/	Integrates multiple best practice pipelines to provide a platform for processing raw data; aims to reduce the overhead for processing large datasets, and provides visualizations produced; currently, the platform has automated workflows for processing RNA-seq, whole-exome sequencing (WES), whole-genome sequencing (WGS), and ChIP-seq datasets
ProTurn	http://heartproteome.org/proturn	Scalable analysis platform that assesses the turnover rate of proteins. Uses a deuterium oxide (D_2O) labeling protocol and determines the kinetics by integrating MS peaks, determining isotope abundances, and using multivariate optimization. Available for use on a wide variety of datasets
TargetP	http://www.cbs.dtu.dk/services/TargetP/	Web-based tool for predicting the subcellular location of eukaryotic proteins and identity of secretory signal or transit peptides using N-terminal sequence information and a combination of machine learning methods. Commonly used by mitochondrial physiologists analyzing proteomics datasets

Many other emerging areas of omics are developing best practice guidelines; however, creating these guidelines requires significant effort for their implementation. To help investigators with this, OmicsPipe integrates multiple best practice pipelines to provide a platform for processing raw data. OmicsPipe aims to reduce the overhead for processing large datasets, and provides visualizations produced; currently, the platform has automated workflows for processing RNA-seq, whole-exome sequencing (WES), whole-genome sequencing (WGS), and ChIP-seq datasets (Fisch et al. 2015).

Upon processing the data, investigators are faced with myriad tools to perform analyses; without distinct computational knowledge, navigating these platforms can be daunting. Operating on a unified system can substantially streamline omics analyses; that being said, there are many stand-alone tools specific to one branch of omics that can elucidate valuable knowledge. Within the realm of proteomics, identifying the subcellular location of a protein is important for determination of function, genome annotation, drug development, and disease identification. Proteins are designed to play their role in specific locations, defining their function based on their environment. This becomes particularly important when discussing mitochondrial proteins. Generally speaking, there are two parallel approaches to determine whether a protein is localized to the mitochondrion: the approach of the biologist or biochemist, and that of the informatician. The former will devise a plan for isolation, fractionation, assays, and copurification with other known mitochondrial markers to look for biochemical evidence of localization to one organelle or another. The latter would investigate features of the gene or protein sequence utilizing computational approaches that would identify it as localized to mitochondria.

Proteins bound for localization in mitochondrial membranes contain an amino acid sequence signal at the N-terminus or an internal targeting sequence to direct them to their appropriate position, both of which are managed by different species of the translocase inner- and outer-membrane (TIM and TOM) complexes. An amphipathic helix in a protein's presequence is cleaved upon delivery (von Heijne 1986; Schatz and Gottfried 1993), while proteins lacking a presequence region remain in the cytosol (Fox 2012).

TargetP is a Web-based tool for predicting the subcellular location of eukaryotic proteins and identity of secretory signal or transit peptides using N-terminal sequence information and a combination of machine learning methods. This tool is commonly used by mitochondrial physiologists analyzing proteomics datasets, as evidenced by over 2,000 citations of the tool's two papers in the scientific literature. The user submits either an amino acid sequence or a FASTA file as the input, and retrieves a plain text file outlining the predictions (Emanuelsson et al. 2007). The predictions were found to be 90% accurate for non-plant proteins and 85% accurate for plant proteins (Klee and Ellis 2005; Emanuelsson et al. 2000). The tool is publicly accessible as a Web service or downloadable for local computation. It is one of the prediction tools used by UniProt to annotate mitochondrial peptides, along with Predotar, TMHMM, and Phobius, all of which make predictions based on N-terminal targeting sequences (Small et al. 2004; Käll et al. 2004; Krogh et al. 2001). In addition to localization prediction, TMHMM predicts transmembrane helices in protein sequences with 97% accuracy (Krogh et al. 2001) while Phobius was shown to predict the secondary structure of proteins with fewer instances of false classification and identify signaling proteins with fewer false positives than TargetP and TMHMM (Käll et al. 2004).

MetazSecKB combines results from each of these tools to increase prediction accuracy for secretome and subcellular proteome localization. The database presents subcellular protein location based on manual curation of the scientific literature combined with UniProt sequence data and annotations. When this annotation is lacking, MetazSecKB employs an algorithm that utilizes multiple prediction tools, combining the predictions of publicly available tools including TargetP, SignalP, Phobius, ScanProsite, WolfPSORT, FragAnchor, and TMHMM (Meinken et al. 2015). The accuracy of localization predictions increased significantly when these tools were utilized in concert, rather than individually. Accordingly, the algorithm combs data from each of the tools simultaneously, and then applies statistical and data mining techniques to acquire the most accurate localization predictions for eukaryotic secreted proteins (Min 2010). Over 135,000 proteins in *Homo sapiens* are represented in the database, approximately 21,000 of which localize to mitochondria; 3,737 of these are associated with the mitochondrial membrane and 17,623 are non-membrane proteins.

The Cardiac Organellar Protein Atlas Knowledgebase (COPaKB) is a centralized platform featuring high-quality cardiac proteomics data and relevant cardiovascular phenotype information (Zong et al. 2013). As of September 2016, COPaKB features 11 organellar modules, comprising 4,467 LC-MS/MS experiments from human, mouse, drosophila, and *Caenorhabditis elegans*. There are four mitochondrial specific modules for each species with over 1,000 proteins represented in each species. The organellar modules constitute the mass spectral library and are utilized by COPaKB's unique high-performance search engine to identify and annotate proteins in the mass spectra files that are submitted by the user in mzML or DTA formats. Data in COPaKB can be viewed within the browser, accessed via the REST API or downloaded in Excel XLS, XML, and JSON formats.

Protein expression data does not take into account the rate of synthesis and degradation of a certain protein, termed protein turnover. As such, measuring expression alone is not sufficient to understand the dynamics of protein levels within the mitochondria. Accordingly, tools have been developed that align with dynamics protocols. One such tool is ProTurn, which uses a deuterium oxide (D_2O) labeling protocol and determines the kinetics by integrating MS peaks, determining isotope abundances, and using multivariate optimization. Most importantly, this tool is scalable, which enables users to perform analysis on a wide range of datasets (Wang et al. 2014).

To elucidate the small-molecule perturbations that may occur in varying mitochondrial physiological states, two types of tools exist: one assesses the quantitative levels of metabolites, while the other synthesizes metabolite lists into known networks, so that these can be visualized by the investigator. Originally developed in 2009 (Xia et al. 2009), MetaboAnalyst has gone through many iterations, with updates in 2012 (Xia et al. 2012) and 2015 (Xia et al. 2015) bringing vital improvements. The current version accepts a wide variety of data types, including NMR spectra, MS spectra, and compound/concentration data. The user interface guides investigators through the analysis pipeline, beginning with dataset quality control by the user. Once quality control standards have been met, the platform will analyze the data, using an extensive spectral library for enhanced metabolite identification. The most current version, MetaboAnalyst 3.0, has the capacity for biomarker identification, as well as a host of other informatics tools for the best standard metabolomics analyses. The graphical output allows users to view the analysis results in a user-friendly format. This platform has been used extensively in mitochondrial studies to understand mitochondrial physiological function in the spinal cord in an ALS model, identify therapeutic targets of cardiomyopathy, and uncover the role of mitochondrial protein quality control in the context of physiological stress across many systems (Quintana et al. 2016; Cacabelos et al. 2016; Picard et al. 2015).

Cytoscape is an open-source software platform for the visualization of complex biological systems, such as molecular interaction networks and biological pathways. The platform enables enhancement of the network data through integration of various formats of metadata into the network structure. One powerful aspect of Cytoscape is its extendibility; third-party developers can access its API and develop applications on top of Cytoscape that readily implement the desired functionality. As of September 2016, the Cytoscape App Store contains 228 Apps, with 305,000 downloads in total (Ono 2015). Cytoscape is an excellent fit for visualization of networks, such as microRNA networks in the brain stem (DeCicco et al. 2015).

Stand-alone tools created by members of the omics community have proven integral to furthering omics research. However, these tools exist as fragments, which makes dissemination to the broader community a significant challenge. To combat this, Galaxy was created as an informatics workflow management system and data integration platform that aims to make computational biology accessible to researchers with limited experience in computer programming (Goecks et al. 2010). Galaxy provides a graphical user interface, customizable plug-ins, and access to public datasets and other users' workflows. This offers a robust peer-review mechanism in which the analyses conducted previously can be reproduced with little effort (Sandve et al. 2013). Because the workflows are hosted on the cloud and Galaxy servers perform the computational work, this greatly reduces the requirement for setting up expensive infrastructure to achieve research goals. One branch of Galaxy, Galaxy-P, contains workflows specifically designed to analyze proteomics datasets and integrate them with other forms of omics data, such as transcriptomics (Sheynkman et al. 2014). In addition to tools developed in the academic community, commercial tools have been developed with similar infrastructure. MetaCore™, developed by Thomson Reuters, exists as a stand-alone program as well as a Web application. The tool contains multiple different analysis methods for varying types of high-throughput molecular data, including sequencing and gene expression, proteomic data, and metabolomic data. In addition to an internal, manually curated database, MetaCore™ contains genomic analysis tools, a data mining toolkit, a pathway editor, and data parsers to adapt the wide range of omics data that can be uploaded (Cambiaghi et al. 2016).

Using these tools requires significant computational power; the previously established paradigm for in-house platforms is becoming extremely costly. The hardware usually becomes dated in about 3 years, and must be updated in order to maintain relevance. Even with extensive collaboration and shared infrastructure, the servers are used for only a fraction of the time they are available, which creates a highly inefficient cost per analysis. These problems have illustrated the need for a unified resource in which researchers can take advantage of ever-improving capacities and features, with significantly less up-front expense. Recent advancements point to cloud-based computing and storage systems. Operations on the cloud provide the same computational power but represent only a tiny fraction of the hardware and operational costs for an in-house server-based computational platform. The National Institutes of Health has proposed the cloud-based Commons environment, a cost-sharing model that will provide access to scalable storage and computational resources for the entire biomedical community. Commons Credits will serve as the currency for computational efforts and require minimal effort to apply so as to reduce the administrative overhead of establishing computational infrastructure, be it in-house or on business cloud servers (https://www.commons-credit-portal.org).

6 Conclusion

The current deluge of data brings significant challenges to which researchers must respond by continually improving methods and technologies for data management and dissemination. In adhering to the FAIR doctrine, data shall be accessible in all respects, and their analyses will also be presented in an intuitive, structured manner. As such, the resulting dataset and knowledge will be open to reuse, repurpose, and reanalysis so as to investigate different targets of interest. In this chapter, we have outlined a collection of tools and resources that serve to aid a deeper understanding of mitochondrial physiology and its role in health and disease. These tools range from community-generated encyclopedic resources, expert-curated databases, and repositories for data management and access to tools for analysis and visualization of biological processes. They allow greater access and reuse of data through annotation, metadata, and analysis platforms. The development of these tools and resources, as well as the openness of scientific data, has had a dramatic impact on the breadth, depth, and structure of data, as well as the reproducibility of experiments and analysis pipelines.

The physiology discipline stands at a unique cross section, and bridges data and clinical applications. It is through these tools that investigators are able to unlock mechanistic insight and access the potential for clinical translation. The advancement of multi-omic approaches, bioinformatics analyses, and open-access data has improved our basic understanding of physiology and pathology while spurring the development of personalized medicine and discovery of biomarkers for disease (Almeida 2010; de Graaf 2013). EHR is becoming more detailed, accessible, and multidimensional, and natural language processing is making it easier to conduct meta-analyses of disease treatments from de-identified patient records. With the concept of precision medicine, the paradigm of health and disease classification is shifting from broad generalization to distinct and individualized medical profiling (Hayes et al. 2014). In this landscape, researchers can significantly benefit from using computational and informatics tools to enable better scientific investigations.

Informatics science is transforming the scope of biomedical research, providing ample tools and methods by which to address the requirements of Big Data, personalized medicine, and next-generation scientific questions. New and improved infrastructure in the research and health sectors have resulted in a burgeoning expansion of data that requires research scientists and clinicians alike to investigate novel approaches in data science and informatics. It is at the interface of domain knowledge and computational bandwidth that mitochondrial research can synergistically propel forward, at a velocity not seen in isolated studies. As this data is shifting from disposable to indispensable, integrated approaches are rapidly demonstrating themselves as invaluable components of a biomedical researcher's tool chest.

Acknowledgements This work was supported by the NIH Big Data to Knowledge (BD2K) Center of Excellence Award (U54GM114883).

References

Almeida JS (2010) Computational ecosystems for data-driven medical genomics. Genome Med 2(9):67

Altmann R (1894) Die Elementarorganismen und ihre Beziehungen zu den Zellen. Verlag von Veit & Comp, Leipzig

Anderson S et al (1981) Sequence and organization of the human mitochondrial genome. Nature 290(5806):457–465

Andrews RM et al (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 23(2):147

Attimonelli M et al (2005) HmtDB, a human mitochondrial genomic resource based on variability studies supporting population genetics and biomedical research. BMC Bioinformatics 6(Suppl 4): S4

Bai X et al (2007) Third-generation human mitochondria-focused cDNA microarray and its bioinformatic tools for analysis of gene expression. Biotechniques 42(3):365–375

Bensley RR, Hoerr NL (1934) Studies on cell structure by the freezing-drying method VI. The preparation and properties of mitochondria. Anat Rec 60(4):449–455

Berman HM et al (2000) The Protein Data Bank. Nucleic Acids Res 28:235–242 Oxford University Press

Cacabelos D et al (2016) Early and gender-specific differences in spinal cord mitochondrial function and oxidative stress markers in a mouse model of ALS. Acta Neuropathol Commun 4:3

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

Cambiaghi A, Ferrario M, Masseroli M (2016) Analysis of metabolomic data: tools, current strategies and future challenges for omics data integration. Brief Bioinform. doi:10.1093/bib/bbw031 (Epub ahead of print)

Castellana S, Ronai J, Mazza T (2015) MitImpact: an exhaustive collection of pre-computed pathogenicity predictions of human mitochondrial non-synonymous variants. Hum Mutat 36(2): E2413–E2422

Claude A (1946a) Fractionation of mammalian liver cells by differential centrifugation: II. Experimental procedures and results. J Exp Med 84(1):61–89

Claude A (1946b) Fractionation of mammalian liver cells by differential centrifugation: I. Problems, methods, and preparation of extract. J Exp Med 84(1):51–59

Claude A, Fullam EF (1945) An electron microscope study of isolated mitochondria: method and preliminary results. J Exp Med 81(1):51–62

Daems WT, Wisse E (1966) Shape and attachment of the cristae mitochondriales in mouse hepatic cell mitochondria. J Ultrastruct Res 16(1):123–140

Damas J et al (2014) MitoBreak: the mitochondrial DNA breakpoints database. Nucleic Acids Res 42(Database issue):D1261–D1268

de Graaf D (2013) Multi-omic biomarkers unlock the potential of diagnostic testing. MLO Med Lab Obs 45(8):40, 42

DeCicco D et al (2015) MicroRNA network changes in the brain stem underlie the development of hypertension. Physiol Genomics 47(9):388–399

DePristo MA et al (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43(5):491–498

Emanuelsson O et al (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 300(4):1005–1016

Emanuelsson O et al (2007) Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2(4):953–971

Falk MJ et al (2012) Mitochondrial disease genetic diagnostics: optimized whole-exome analysis for all MitoCarta nuclear genes and the mitochondrial genome. Discov Med 14(79):389–399

Field D et al (2009) Omics data sharing. Science 326(5950):234–236

Finn RD et al (2014a) Pfam: the protein families database. Nucleic Acids Res 42(Database issue): D222–D230

Finn RD et al (2014b) iPfam: a database of protein family and domain interactions found in the Protein Data Bank. Nucleic Acids Res 42(Database issue):D364–D373

Fisch KM et al (2015) Omics Pipe: a community-based framework for reproducible multi-omics data analysis. Bioinformatics 31(11):1724–1728

FORCE11 (2014) Guiding principles for findable, accessible, interoperable and re-usable data publishing version b1.0. https://www.force11.org/node/6062

Fox TD (2012) Mitochondrial protein synthesis, import, and assembly. Genetics 192(4):1203–1234

Godovac-Zimmermann J (2008) 8th Siena meeting. From genome to proteome: integration and proteome completion. Expert Rev Proteomics 5(6):769–773

Goecks J et al (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 11(8):R86

Gomez-Cabrero D et al (2014) Data integration in the era of omics: current and future challenges. BMC Syst Biol 8(Suppl 2):I1

Guo Y et al (2013) MitoSeek: extracting mitochondria information and performing high-throughput mitochondria sequencing analysis. Bioinformatics 29(9):1210–1211

Hayes DF et al (2014) Personalized medicine: risk prediction, targeted therapies and mobile health technology. BMC Med 12:37

Hirase S et al (2016) Parallel mitogenome sequencing alleviates random rooting effect in phylogeography. Genome Biol Evol 8(4):1267–1278

Iwasaki W et al (2013) MitoFish and MitoAnnotator: a mitochondrial genome database of fish with an accurate and automatic annotation pipeline. Mol Biol Evol 30(11):2531–2540

Jayaprakash AD et al (2015) Mito-seek enables deep analysis of mitochondrial DNA, revealing ubiquitous, stable heteroplasmy maintained by intercellular exchange. Nucleic Acids Res 43(4):2177–2187

Jia M et al (2012) Biomarkers in an animal model for revealing neural, hematologic, and behavioral correlates of PTSD. J Vis Exp (68)

Johnston IG, Williams BP (2016) Evolutionary inference across eukaryotes identifies specific pressures favoring mitochondrial gene retention. Cell Syst 2(2):101–111

Johnston IG et al (2015) Stochastic modelling, Bayesian inference, and new in vivo measurements elucidate the debated mtDNA bottleneck mechanism. Elife 4:e07464

Jupe S, Fabregat A, Hermjakob H (2015) Expression data analysis with reactome. Curr Protoc Bioinformatics 49:8.20.1–8.20.9

Käll L, Krogh A, Sonnhammer ELL (2004) A combined transmembrane topology and signal peptide prediction method. J Mol Biol 338(5):1027–1036

Kanehisa M et al (2014) Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res 42(Database issue):D199–D205

Kaye J, Hawkins N (2014) Data sharing policy design for consortia: challenges for sustainability. Genome Med 6(1):4

Kennedy EP, Lehninger AL (1949) Oxidation of fatty acids and tricarboxylic acid cycle intermediates by isolated rat liver mitochondria. J Biol Chem 179(2):957–972

Klee EW, Ellis LB (2005) Evaluating eukaryotic secreted protein prediction. BMC Bioinformatics 6:256

Krogh A et al (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305(3):567–580

Lander ES et al (2001) Initial sequencing and analysis of the human genome. Nature 409 (6822):860–921

Lesnefsky EJ et al (2001) Mitochondrial dysfunction in cardiac disease: ischemia–reperfusion, aging, and heart failure. J Mol Cell Cardiol 33(6):1065–1089

Lott MT et al (2013) mtDNA variation and analysis using MITOMAP and MITOMASTER. Curr Protoc Bioinformatics 1(123):1.23.1–1.23.26

Lotz C et al (2014) Characterization, design, and function of the mitochondrial proteome: from organs to organisms. J Proteome Res 13(2):433–446

Matthews L et al (2009) Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res 37(Database issue):D619–D622

McKenna A et al (2010) The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20(9):1297–1303

Meinken J et al (2015) MetazSecKB: the human and animal secretome and subcellular proteome knowledgebase. Database 2015

Min XJ (2010) Evaluation of computational methods for secreted protein prediction in different eukaryotes. J Proteomics Bioinform 3:143–147

Mitchell P (1961) Coupling of phosphorylation to electron and hydrogen transfer by a chemiosmotic type of mechanism. Nature 191:144–148

Mitchell P (1966) Chemiosmotic coupling in oxidative and photosynthetic phosphorylation. Biol Rev Camb Philos Soc 41(3):445–502

Mitchell P (1968) Chemiosmotic coupling and energy transduction. Glynn Research, Bodmin, Cornwall

Molloy JC (2011) The Open Knowledge Foundation: open data means better science. PLoS Biol 9(12):e1001195

Moritz RL, Bernt M, Middendorf M (2014) Local similarity search to find gene indicators in mitochondrial genomes. Biology (Basel) 3(1):220–242

Nass MM, Nass S (1963a) Intramitochondrial fibers with DNA characteristics. I. Fixation and electron staining reactions. J Cell Biol 19:593–611

Nass S, Nass MM (1963b) Intramitochondrial fibers with DNA characteristics. II. Enzymatic and other hydrolytic treatments. J Cell Biol 19:613–629

Nawrocki EP et al (2015) Rfam 12.0: updates to the RNA families database. Nucleic Acids Res 43(Database issue):D130–D137

O’Driscoll A, Daugelaite J, Sleator RD (2013) ‘Big data’, Hadoop and cloud computing in genomics. J Biomed Inform 46(5):774–781

Ono, K (2015) Cytoscape: an open source platform for complex network analysis and visualization. http://www.cytoscape.org/

Oroboros (2015) MitoPedia – bioblast. http://www.bioblast.at/index.php/MitoPedia

Pagliarini DJ et al (2008) A mitochondrial protein compendium elucidates complex I disease biology. Cell 134(1):112–123

Palade GE (1952) The fine structure of mitochondria. Anat Rec 114(3):427–451

Palade GE (1953) An electron microscope study of the mitochondrial structure. J Histochem Cytochem 1(4):188–211

Perez-Riverol Y et al (2016) Omics discovery index – discovering and linking public omics datasets. bioRxiv:049205

Picard M et al (2015) Mitochondrial functions modulate neuroendocrine, metabolic, inflammatory, and transcriptional responses to acute psychological stress. Proc Natl Acad Sci U S A 112(48): E6614–E6623

Quintana MT et al (2016) Cardiomyocyte-specific human Bcl2-associated anthanogene 3 P209L expression induces mitochondrial fragmentation, Bcl2-associated anthanogene 3 haploin sufficiency, and activates p38 signaling. Am J Pathol 186(8):1989–2007

Rabilloud T et al (1998) Two-dimensional electrophoresis of human placental mitochondria and protein identification by mass spectrometry: toward a human mitochondrial proteome. Electrophoresis 19(6):1006–1014

Rubino F et al (2012) HmtDB, a genomic resource for mitochondrion-based human variability studies. Nucleic Acids Res 40(Database issue):D1150–D1159

Russo E et al (2016) Challenges in patient safety improvement research in the era of electronic health records. Healthcare (Amsterdam, Netherlands) pii: S2213-0764(15)30090-7. doi:10.1016/j.hjdsi.2016.06.005 (Epub ahead of print)

Sandve GK et al (2013) Ten simple rules for reproducible computational research. PLoS Comput Biol 9(10):e1003285

Schatz G, Gottfried S (1993) The protein import machinery of mitochondria. Protein Sci 2(2):141–146

Schatz G, Haslbrunner E, Tuppy H (1964) Deoxyribonucleic acid associated with yeast mitochondria. Biochem Biophys Res Commun 15(2):127–132

Sheynkman GM et al (2014) Using Galaxy-P to leverage RNA-Seq for the discovery of novel protein variations. BMC Genomics 15:703

Small I et al (2004) Predotar: a tool for rapidly screening proteomes for N-terminal targeting sequences. Proteomics 4(6):1581–1590

Smith AC, Blackshaw JA, Robinson AJ (2012) MitoMiner: a data warehouse for mitochondrial proteomics data. Nucleic Acids Res 40(Database issue):D1160–D1167

Stein LD (2004) Using the reactome database. Curr Protoc Bioinformatics Chapter 8:Unit8.7

Tennessen JA et al (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science 337(6090):64–69

Ternent T et al (2014) How to submit MS proteomics data to ProteomeXchange via the PRIDE database. Proteomics 14(20):2233–2241

Tryka KA et al (2014) NCBI's database of genotypes and phenotypes: dbGaP. Nucleic Acids Res 42(Database issue):D975–D979

Tun AW et al (2014) Profiling the mitochondrial proteome of Leber's Hereditary Optic Neuropathy (LHON) in Thailand: down-regulation of bioenergetics and mitochondrial protein quality control pathways in fibroblasts with the 11778G>A mutation. PLoS One 9(9):e106779

Van der Auwera GA et al (2013) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics 43:11.10.1–11.1033

Vizcaino JA et al (2014) ProteomeXchange provides globally coordinated proteomics data submission and dissemination. Nat Biotechnol 32(3):223–226

Vlasblom J et al (2014) Exploring mitochondrial system properties of neurodegenerative diseases through interactome mapping. J Proteomics 100:8–24

von Heijne G (1986) Mitochondrial targeting sequences may form amphiphilic helices. EMBO J 5(6):1335–1342

Wallace DC et al (1988a) Mitochondrial DNA mutation associated with Leber's hereditary optic neuropathy. Science 242(4884):1427–1430

Wallace DC et al (1988b) Familial mitochondrial encephalomyopathy (MERRF): genetic, pathophysiological, and biochemical characterization of a mitochondrial DNA disease. Cell 55(4): 601–610

Walters AM, Porter GA Jr, Brookes PS (2012) Mitochondria as a drug target in ischemic heart disease and cardiomyopathy. Circ Res 111(9):1222–1236

Wang D et al (2014) Characterization of human plasma proteome dynamics using deuterium oxide. Proteomics Clin Appl 8(7–8):610–619

Wang Y et al (2016) Mitogenomic perspectives on the origin of Tibetan loaches and their adaptation to high altitude. Sci Rep 6:29690

Wasinger VC et al (1995) Progress with gene-product mapping of the Mollicutes: Mycoplasma genitalium. Electrophoresis 16(7):1090–1094

Wishart DS et al (2007) HMDB: the Human Metabolome Database. Nucleic Acids Res 35(Database issue):D521–D526

Wishart DS et al (2009) HMDB: a knowledgebase for the human metabolome. Nucleic Acids Res 37(Database issue):D603–D610

Wishart DS et al (2013) HMDB 3.0 – The Human Metabolome Database in 2013. Nucleic Acids Res 41(Database issue):D801–D807

Xia J et al (2009) MetaboAnalyst: a web server for metabolomic data analysis and interpretation. Nucleic Acids Res 37(Web Server issue):W652–W660

Xia J et al (2012) MetaboAnalyst 2.0 – a comprehensive server for metabolomic data analysis. Nucleic Acids Res 40(Web Server issue):W127–W133

Xia J et al (2015) MetaboAnalyst 3.0 – making metabolomics more meaningful. Nucleic Acids Res 43(W1):W251–W257

Yates JR 3rd (2013) The revolution and evolution of shotgun proteomics for large-scale proteome analysis. J Am Chem Soc 135(5):1629–1640

Zong NC et al (2013) Integration of cardiac proteome biology and medicine by a specialized knowledgebase. Circ Res 113(9):1043–1053

Chapter III

A Metadata Extraction Approach for Clinical Case Reports to Enable Advanced Understanding of Biomedical Concepts

Video Article

A Metadata Extraction Approach for Clinical Case Reports to Enable Advanced Understanding of Biomedical Concepts

John Harry Caufield[1,2], David A. Liem[1,2,3], Anders O. Garlid[1,2], Yijiang Zhou[4], Karol Watson[1,3], Alex A. T. Bui[1,5,6,7], Wei Wang[1,7,8,9], Peipei Ping[1,2,3,7,8]

[1] The NIH BD2K Center of Excellence in Biomedical Computing, University of California, Los Angeles
[2] Department of Physiology, University of California, Los Angeles
[3] Department of Medicine/Cardiology, University of California, Los Angeles
[4] Department of Cardiology, First Affiliated Hospital, Zhejiang University School of Medicine
[5] Department of Radiological Sciences, University of California, Los Angeles
[6] Department of Bioengineering, University of California, Los Angeles
[7] Scalable Analytics Institute (ScAi), University of California, Los Angeles
[8] Department of Bioinformatics, University of California, Los Angeles
[9] Department of Computer Science, University of California, Los Angeles

Correspondence to: John Harry Caufield at jcaufield@mednet.ucla.edu

URL: https://www.jove.com/video/58392
DOI: doi:10.3791/58392

Keywords: Medicine, Issue 139, data science, medical informatics, text mining, annotation, curation, clinical case reports

Date Published: 9/20/2018

Citation: Caufield, J.H., Liem, D.A., Garlid, A.O., Zhou, Y., Watson, K., Bui, A.A., Wang, W., Ping, P. A Metadata Extraction Approach for Clinical Case Reports to Enable Advanced Understanding of Biomedical Concepts. *J. Vis. Exp.* (139), e58392, doi:10.3791/58392 (2018).

Abstract

Clinical case reports (CCRs) are a valuable means of sharing observations and insights in medicine. The form of these documents varies, and their content includes descriptions of numerous, novel disease presentations and treatments. Thus far, the text data within CCRs is largely unstructured, requiring significant human and computational effort to render these data useful for in-depth analysis. In this protocol, we describe methods for identifying metadata corresponding to specific biomedical concepts frequently observed within CCRs. We provide a metadata template as a guide for document annotation, recognizing that imposing structure on CCRs may be pursued by combinations of manual and automated effort. The approach presented here is appropriate for organization of concept-related text from a large literature corpus (*e.g.,* thousands of CCRs) but may be easily adapted to facilitate more focused tasks or small sets of reports. The resulting structured text data includes sufficient semantic context to support a variety of subsequent text analysis workflows: meta-analyses to determine how to maximize CCR detail, epidemiological studies of rare diseases, and the development of models of medical language may all be made more realizable and manageable through the use of structured text data.

Video Link

The video component of this article can be found at https://www.jove.com/video/58392/

Introduction

Clinical case reports (CCRs) are a fundamental means of sharing observations and insights in medicine. These serve as a basic mechanism of communication and education for clinicians and medical students. Historically, CCRs have also provided accounts of emerging diseases, their treatments, and their genetic backgrounds[1,2,3,4]. For example, the first treatment of human rabies by Louis Pasteur in 1885[5,6] and the first application of penicillin in patients[7] were both reported through CCRs. More than 1.87 million CCRs have been published as of April 2018, with over half a million within the last decade; journals are continuing to provide new venues for these reports[8]. Though unique in form and content, CCRs contain text data that are largely unstructured, contain a vast vocabulary, and concern interrelated phenomena, limiting their use as a structured resource. Significant effort is required to extract detailed metadata (*i.e.,* "data about data", or in this case, descriptions of document contents) from CCRs and establish them as a findable, accessible, interoperable, and reusable (FAIR)[9] data resource.

Here, we describe a process for extracting text and numerical values to standardize the description of specific biomedical concepts within published CCRs. This methodology includes a metadata template to guide annotation; see **Figure 1** for an overview of this process. Application of the annotation process to a large collection of reports (*e.g.,* several thousand of a specific type of disease presentation) permits assembly of a manageable and structured set of annotated clinical texts, achieving machine-readable documentation and biomedical phenomena embedded within each clinical presentation. Though data formats such as those provided by HL7 (*e.g.*, Version 3 of the Messaging Standard[10] or the Fast Healthcare Interoperability Resources [FHIR][11]), LOINC[12], and revision 10 of the International Statistical Classification of Diseases and Related Health Problems (ICD-10)[13] provide standards for describing and exchanging clinical observations, they do not capture the text surrounding these data, nor are they intended to. The results of our methodology are best used to enforce structure on CCRs and facilitate subsequent analysis, normalization through controlled vocabularies and coding systems (*e.g.*, ICD-10), and/or conversion to the clinical data formats listed above.

Mining CCRs is an active area of work within biomedical and clinical informatics. Though previous proposals to standardize the structure of case reports (*e.g.*, using HL7 v2.5[14] or standardized phenotype terminology[15]) are commendable, it is likely that CCRs will continue to follow a variety of different natural-language forms and document layouts, as they have for much of the past century. Under ideal conditions, authors of new case reports follow CARE guidelines[16] to ensure they are comprehensive. Approaches sensitive to both natural language and its relation to medical concepts may therefore be most effective in working with new and archived reports. Resources such as CRAFT[17] and those produced by Informatics for Integrating Biology and the Bedside (i2b2)[18] curation support natural language processing (NLP) approaches yet do not specifically focus on CCRs or clinical narratives. Similarly, medical NLP tools such as cTAKES[19] and CLAMP[20] have been developed but generally identify specific words or phrases (*i.e.*, entities) within documents rather than the general concepts commonly described in CCRs.

We have designed a standardized metadata template for features commonly included within CCRs. This template defines features to impose structure on CCRs—an essential precursor for in-depth comparisons of document contents-yet allows for sufficient flexibility to retain semantic context. Though we have designed the format associated with this template to be appropriate for both manual annotation and computationally-assisted text mining, we have ensured it is particularly easy to use for manual annotators. Our approach noticeably differs from more intricate (and, therefore, less immediately understandable to untrained researchers) frameworks such as FHIR[21]. The following protocol describes how to isolate document features corresponding to each template data type, with a single set of values corresponding to those in a single CCR.

The data types within the template are those most descriptive for CCRs and patient-focused medical documents in general. Annotation of these features promotes findability, accessibility, interoperability, and reusability of CCR text, primarily by giving it structure. The data types are in four general categories: document and annotation identification, case report identification (*i.e.,* document-level properties), medical content concepts (primarily concept-level properties), and acknowledgements (*i.e.,* features providing evidence of funding). In this annotation process, each document includes the full text of a CCR, omitting any document contents material independent to the case (*e.g.,* experimental protocols). CCRs are generally less than 1,000 words each; a single corpus should ideally be indexed by the same bibliographic database and be in the same written language.

The product of the approach described here, when applied to a CCR corpus, is a structured set of annotated clinical text. While this methodology can be performed fully manually and has been designed to be performed by domain experts without any informatics experience, it complements the natural language processing approaches specified above and provides data appropriate for computational analysis. Such analyses may be of interest to audiences of researchers beyond those who frequently read CCRs, including:

- those concerned with disease presentations, their key symptomology, usual diagnostic approaches, and treatments
- those who wish to compare the results of clinical trials with events described within the clinical literature, potentially providing additional observations and greater statistical power.
- bioinformatics, biomedical informatics, and computer science researchers who require structured medical language data sets or high-level understandings of medical narratives
- Government policy researchers focusing on how clinical trials may best reflect how diagnosis and treatment as it occurs in reality

Enforcing structure on CCRs can support numerous subsequent efforts to better understand both medical language and biomedical phenomena.

Protocol

1. Document and Annotation Identification

Note: Values in this category support the annotation process.

1. Using the annotation template, provide an identifier specific to this metadata set, *e.g.*, ***Case123***. The identifier format should be consistent throughout the project (*e.g., **Case001*** through ***Case500***).
2. Specify the date on which a document was read and annotated. Use a format resembling “Jan 10 2018” for consistency and readability.

2. Case Report Identification

Note: Values in this category provide document-level features and contribute to a document’s findability.

1. Be consistent with the format of each field across all annotations, *e.g.,* individual values should be separated by semicolons without following spaces in all entries. Use identical formats to those used in the original document *or* those used in a bibliographic database such as MEDLINE.
2. Provide the title of the document.
3. Provide the names of all authors of the document in the provided order. Normalize the format of all names, such that all names take the form of a single last name followed by any number of initials, *e.g.* Jane B. Park becomes ***Park JB***. Do not include titles. Separate multiple authors with a semicolon without additional punctuation, such that John A. Smith, Jane B. Park takes a form of **Smith JA;Park JB.**
4. Provide the year of publication of the document.
5. Provide the full title of the journal in which the document was published. A list of controlled journal names is provided by the NLM Catalog (https://www.ncbi.nlm.nih.gov/nlmcatalog)**.**
6. Provide the address of the home institution of the authors of the document, as specified in the document. This may include departments, geographic locations, and postal address details.

    1. If multiple locations are provided (*e.g.,* if affiliations differ between authors), specify only details for the corresponding author. If a corresponding author cannot be identified, use that of the first author, or do not specify an institution. If a corresponding author has multiple affiliations, specify both and separate with a semicolon.

7. Provide the corresponding author of the document, as specified within the document heading using the same format as that used in the Authors data type.
8. Provide a document identifier (*e.g.,* a PMID).
9. Provide a Digital Object Identifier, where possible and available, resolvable to the document URL (through https://www.doi.org/), not a PubMed Central page.
10. Provide a stable URL to the full text of the document, if available**.** To maximize accessibility, this may refer to the PubMed Central version.
11. Provide the document language. For documents available in multiple languages, provide both, separated with a semicolon.

3. Medical Content

Note: Values in this category identify document-level, concept-level, and text-level features. They serve to enhance a document's accessibility, interoperability, and reusability. These features provide ways to observe conceptual and semantic similarities between document content, with a focus on biomedical topics and events. Most categories in this section can include multiple text statements and each should be separated using a semicolon.

1. Include contextual detail in each field (*e.g.,* "mother had breast cancer at age 50") rather than providing only terms from a controlled vocabulary (*e.g.,* not "breast cancer" alone). Do not include extensive detail beyond each observation.
2. Omit commonly repeated words and phrases (*e.g.,* pronouns, the word "patient", and the phrases "complained of" or "presented with"). Though subjectivity across multiple annotators is likely, it may be reduced by having multiple annotators for each document and through automated normalization after data collection. Computational post-processing approaches will vary by subsequent analysis needs and are not discussed here in detail.
3. Provide the following information in the annotation template.
    1. Provide specific terms identified within a document, usually in its header, as key terms. Separate with a semicolon as terms may include other punctuation.
    2. Provide demographic values, specifically any text statements describing a patient's background, including sex and/or gender, age, ethnicity, or nationality.
    3. Provide geographic locations mentioned within the clinical narrative, other than specific institution addresses. This should not include anatomical locations/parts, but may include any geographic locale where the patient resides or travels.
    4. Provide life style values, including any text statements describing frequent patient activities or behaviors relevant to their general health. In practice, this frequently involves smoking or alcohol consumption habits, but may also include sun exposure, diet, or frequency of specific types of physical activity.
    5. Provide medical history values referring to family history. Include any text statements describing clinical observations of and events experienced by siblings, parents, and other family members. This includes genetic conditions and negative observations (*i.e., **family history was negative for*** a disease)**.**
    6. Provide values referring to Social History, including any text statements describing patient background not covered in Demography or Life Style. There may be overlaps in content between these categories. The statements may include occupational history and social habits.
    7. Provide values referring to the patient's medical and surgical history. Include any text statements describing any medical observations, treatments, or other events taking place prior to the beginning of the clinical presentation. This includes obstetric history and periods of good health, where noted.
    8. Specify one or more of the following 16 disease system categories. Note that these values are categorical rather than free-text. Categories are not comprehensive but should indicate most systems impacted by the events described in the clinical presentation and diagnosed disease.
        1. Follow a specific set of categories, based on the categories used in the International Statistical Classification of Diseases and Related Health Problems, revision 10 (ICD-10) code system. See **Table 1** for the list of disease system categories along with corresponding ICD-10 code ranges.

    9. Provide details of all signs and symptoms**.** Include any text statements describing any medical observations of signs or symptoms beginning at initial presentation, including their onset, duration, severity, and resolution, if provided. Do not include symptoms described in the outcome. These values may overlap with other types if symptoms continue from history to initial presentation.
    10. Provide details of any comorbidities. Include any terms or phrases describing distinct diseases present at the time of initial clinical presentation. There is likely overlap between these values and those in clinical history, though Comorbidity should not include terms identical to those in the Diagnosis.
    11. Provide details of all diagnostic techniques and procedures. Include the names of medical procedures done for diagnostic purposes, including examinations, tests, and imaging, as well as the conditions under which these tests were performed and relevant anatomical locations (*e.g.,* "upper extremity venous ultrasound"). Exclude test results.
    12. Provide details of diagnosis. Include any text statements describing diagnoses of disease, even if the final diagnosis is ambiguous.
    13. Provide all laboratory values and test results. Include names of diagnostic tests, their values, and conditions under which they were performed. This will involve overlap with terms used in the Diagnostic Techniques and Procedures data type. Both numerical and qualitative values (*e.g., **complete blood count was within normal limits***) are acceptable**.** If the names of diagnostic tests are not provided, use terms describing the results (*e.g., **leukopenia***), though they should also be included in the Signs and Symptoms.
    14. Provide details of pathology. Include any text statements describing results of pathology and histology studies, including gross pathology, immunology, and microscopy studies. Terms may overlap with those used in Diagnostic Techniques and Procedures (step 3.11), *e.g.,* with the procedures performed to obtain samples such as biopsy.

15. Provide all pharmacological therapies. Include any text statements describing drug therapies used in the course of treatment, including general terms such as ***antibiotics*** or specific drug names**.** Also, include descriptions of when and how drug therapies were stopped.
16. Provide all interventional procedures. Include any text statements describing therapeutic procedures used in the course of treatment, including invasive procedures, implantation of medical devices, and procedures done to facilitate other therapies. Also, include descriptions of when and how ongoing therapeutic procedures were stopped, if necessary.
17. Provide the patient outcome. Include any text statements describing health of the patient as of the end of the clinical presentation described in the report, including any follow-up tests.
18. Provide counts of all diagnostic images, figures, videos/animations, and tables. Include all counts of visual media included in the report, in the following format: Count of images; Count of figures; Count of videos or animations; Count of tables.
    1. Distinguish between images and figures in this way: images include any products of clinical diagnostics, including photographs, micrographs, electrocardiogram rhythm images, and other products of diagnostic imaging, while figures are all other images, generally including data plots and illustrations.

19. Provide evidence of relationships to other CCRs. This field may include identifiers (*e.g.,* PMIDs) of other reports in the data set cited by or referencing this report.
20. Provide evidence of relationships to clinical trials. This field may include identifiers of clinical trials citing this CCR. Identify trials by their ClinicalTrials.gov identifiers, preceded by NCT, or other stable identifier.
21. Include database crosslinks corresponding to this document, including identifiers, preferably as database names and stable URLs.

4. Acknowledgements

Notes: Values in this category identify document-level features yet have little consistent structure across publications. They provide details regarding the organizations providing support for a CCR and related work. This category also includes a field for the total count of references cited by an article: this is intended to provide a rough metric of the degree to which a document has conceptual relationships with other biomedical documents of any type. Within the four data types in this section, provide the following.

1. Specify all funding sources supporting the work and corresponding PI as well as relevant award numbers. The first value, Funding Source, should include the names of all organizations providing financial support for the work.
    1. Separate organizations with semicolons and spaces, *e.g.*, ***National Institutes of Health/National Cancer Institute***; ***DOE***; ***Smith-Park Foundation***.
    2. For the following value, Award Number, specify any award numbers or specific designations provided along with the recipients of the awards, where appropriate, as initials of the recipients in parentheses, *e.g.*, ***R01HL123123 (to JP)***, ***NS12312 (to JP, JS)***, ***research training fellowship (to JS)***. Authors may explicitly state that no corresponding information is available (*e.g.,* "no funding was received"); in these cases, use the text provided by the authors as the Funding Source value. Otherwise, the value should be NA.

2. Specify disclosures/conflicts of interest as specified by the authors, *e.g.*, ***JP is a consultant for DrugCo***. Authors may explicitly state that no corresponding information is available (*e.g.,* "no conflict of interest is declared"); in these cases, use the text provided by the authors as the Disclosures/Conflict of Interest value. Otherwise, as above, the value should be NA.
3. Specify a numerical count of all references cited by the document, not including those provided in any supplementary material. No reference text should be included in this field.

Representative Results

An example of the annotation process is shown in **Figure 2**. This case[22] describes a presentation of infection by the bacterial pathogen *Burkholderia thailandensis*. For reference, the relevant portion of this CCR is provided in plain text format in **Supplementary File 1**; some research findings are also presented in this report and are included for comparison. In practice, converting reports provided in HTML or PDF format to plain text may improve the efficiency and ease of metadata extraction.

Examples of two sets of completed CCR metadata annotations are provided in **Table 2**. The first of these examples is mock data to illustrate the ideal format of each value, while the second example contains values extracted from a published CCR on a rare condition, acrodermatitis enteropathica[23].

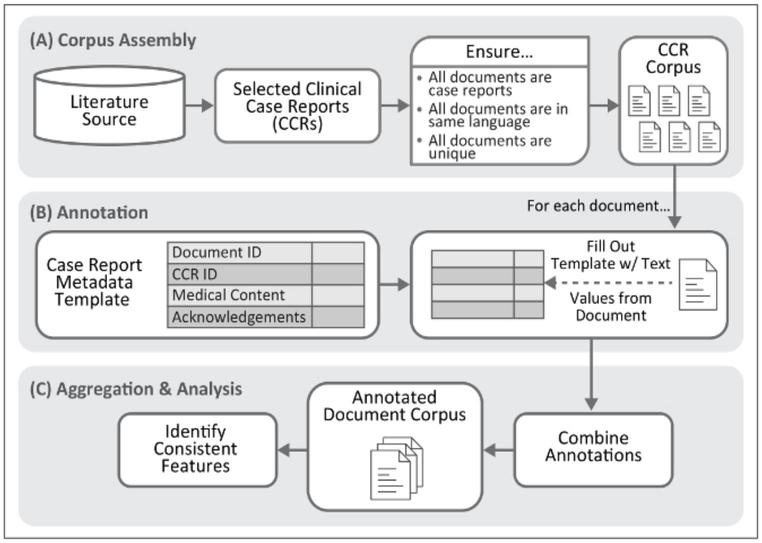
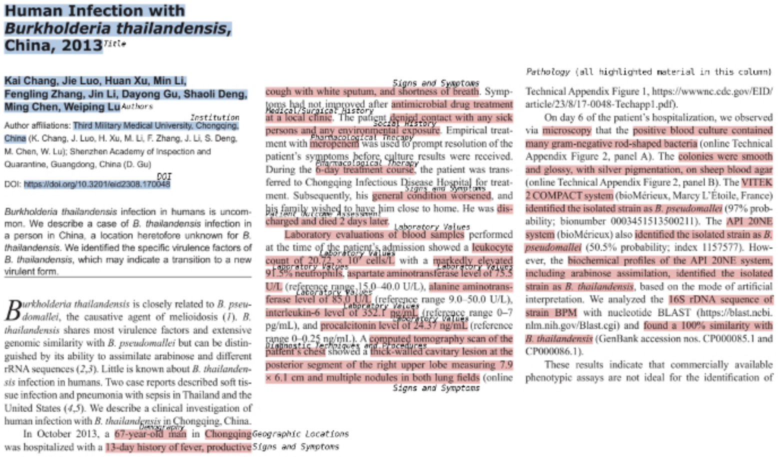

(A) Corpus Assembly
Literature Source
Selected Clinical Case Reports (CCRs)
Ensure...
• All documents are case reports
• All documents are in same language
• All documents are unique
CCR Corpus
(B) Annotation
For each document...
Case Report Metadata Template
Document ID
CCR ID
Medical Content
Acknowledgements
Fill Out Template w/ Text
Values from Document
(C) Aggregation & Analysis
Identify Consistent Features
Annotated Document Corpus
Combine Annotations

Figure 1. Workflow for Case Report Annotation. The protocol described here provides a method for identification of textual features frequently present within clinical case reports. This process requires assembly of a document corpus. The product of the annotation process, once aggregated into a single file, permits identification of text features associated with medical concepts and their descriptions within case reports. Please click here to view a larger version of this figure.


Title
Authors
Institution
DOI
Demography
Geographic Locations
Signs and Symptoms
Signs and Symptoms
Medical/Surgical History
Social History
Pharmacological Therapy
Pharmacological Therapy
Signs and Symptoms
Patient Outcome Assessment
Laboratory Values
Laboratory Values
Laboratory Values
Laboratory Values
Laboratory Values
Laboratory Values
Laboratory Values
Diagnostic Techniques and Procedures
Signs and Symptoms
Pathology (all highlighted material in this column)

Human Infection with *Burkholderia thailandensis*, China, 2013

Kai Chang, Jie Luo, Huan Xu, Min Li, Fengling Zhang, Jin Li, Dayong Gu, Shaoli Deng, Ming Chen, Weiping Lu

Author affiliations: Third Military Medical University, Chongqing, China (K. Chang, J. Luo, H. Xu, M. Li, F. Zhang, J. Li, S. Deng, M. Chen, W. Lu); Shenzhen Academy of Inspection and Quarantine, Guangdong, China (D. Gu)

DOI: https://doi.org/10.3201/eid2308.170048

Burkholderia thailandensis infection in humans is uncommon. We describe a case of *B. thailandensis* infection in a person in China, a location heretofore unknown for *B. thailandensis*. We identified the specific virulence factors of *B. thailandensis*, which may indicate a transition to a new virulent form.

Burkholderia thailandensis is closely related to *B. pseudomallei*, the causative agent of melioidosis (*1*). *B. thailandensis* shares most virulence factors and extensive genomic similarity with *B. pseudomallei* but can be distinguished by its ability to assimilate arabinose and different rRNA sequences (*2*,*3*). Little is known about *B. thailandensis* infection in humans. Two case reports described soft tissue infection and pneumonia with sepsis in Thailand and the United States (*4*,*5*). We describe a clinical investigation of human infection with *B. thailandensis* in Chongqing, China.

In October 2013, a 67-year-old man in Chongqing was hospitalized with a 13-day history of fever, productive cough with white sputum, and shortness of breath. Symptoms had not improved after antimicrobial drug treatment at a local clinic. The patient denied contact with any sick persons and any environmental exposure. Empirical treatment with meropenem was used to prompt resolution of the patient's symptoms before culture results were received. During the 6-day treatment course, the patient was transferred to Chongqing Infectious Disease Hospital for treatment. Subsequently, his general condition worsened, and his family wished to have him close to home. He was discharged and died 2 days later.

Laboratory evaluations of blood samples performed at the time of the patient's admission showed a leukocyte count of 20.72 × 10^9 cells/L with a markedly elevated 91.5% neutrophils, aspartate aminotransferase level of 75.5 U/L (reference range 15.0–40.0 U/L), alanine aminotransferase level of 85.0 U/L (reference range 9.0–50.0 U/L), interleukin-6 level of 352.1 pg/mL (reference range 0–7 pg/mL), and procalcitonin level of 24.37 ng/mL (reference range 0–0.25 ng/mL). A computed tomography scan of the patient's chest showed a thick-walled cavitary lesion at the posterior segment of the right upper lobe measuring 7.9 × 6.1 cm and multiple nodules in both lung fields (online Technical Appendix Figure 1, https://wwwnc.cdc.gov/EID/article/23/8/17-0048-Techapp1.pdf).

On day 6 of the patient's hospitalization, we observed via microscopy that the positive blood culture contained many gram-negative rod-shaped bacteria (online Technical Appendix Figure 2, panel A). The colonies were smooth and glossy, with silver pigmentation, on sheep blood agar (online Technical Appendix Figure 2, panel B). The VITEK 2 COMPACT system (bioMérieux, Marcy L'Étoile, France) identified the isolated strain as *B. pseudomallei* (97% probability; bionumber 0003451513500211). The API 20NE system (bioMérieux) also identified the isolated strain as *B. pseudomallei* (50.5% probability; index 1157577). However, the biochemical profiles of the API 20NE system, including arabinose assimilation, identified the isolated strain as *B. thailandensis*, based on the mode of artificial interpretation. We analyzed the 16S rDNA sequence of strain BPM with nucleotide BLAST (https://blast.ncbi.nlm.nih.gov/Blast.cgi) and found a 100% similarity with *B. thailandensis* (GenBank accession nos. CP000085.1 and CP000086.1).

These results indicate that commercially available phenotypic assays are not ideal for the identification of

Figure 2. Identification of Concept-Specific Text in a Clinical Case Report. Beginning with the text of a case report, a manual annotator may progress through the document, identifying segments of text corresponding to each component of the metadata template. Identification features are highlighted in blue. Text corresponding to medical concepts are in red and labeled with their type; all highlighted text in the third column refers to the Pathology type. Please click here to view a larger version of this figure.

Category	Description	ICD-10 Chapter	ICD-10 Code Range
cancer	Any type of cancer or malignant neoplasm.	II	C00-D49
nervous	Any disease of the brain, spine, or nerves.	VI	G00-G99
cardiovascular	Any disease of the heart or vascular system. Does not include hematological diseases.	IX	I00-I99
musculoskeletal and rheumatic	Any disease of the muscles, skeletal system, joints, and connective tissues.	XIII	M00-M99
digestive	Any disease of the gastrointestinal tract and digestive organs, including the liver and pancreas.	XI	K00-K95
obstetrical and gynecological	Any disease relating to pregnancy, childbirth, the female reproductive system, or the breasts.	XIV; XV	O00-O9A; N60-N98
infectious	Any disease causes by infectious microorganisms.	I	A00-B99
respiratory	Any disease of the lungs and respiratory tract.	X	J00-J99
hematologic	Any disease of the blood, bone marrow, lymph nodes, or spleen.	III	D50-D89
kidney and urologic	Any disease of the kidneys or bladder, including the ureters, as well as the male reproductive organs, including the prostate.	XIV	N00-N53; N99
endocrine	Any disease of the endocrine glands, as well as metabolic disorders.	IV	E00-E89
oral and maxillofacial	Any condition involving the mouth, jaws, head, face, or neck.	XI; XIII	K00-K14; M26-M27
eye	Any condition involving the eyes, including blindness.	VII	H00-H59
otorhinolaryngologic	Any condition of the ear, nose, and/or throat.	VIII	H60-H95; J30-J39
skin	Any disease of the skin.	XII	L00-L99
rare	A special category reserved for reports of rare diseases, defined as those impacting fewer than 200,000 individuals in the United States (see https://rarediseases.info.nih.gov/diseases)	*NA*	*NA*

Table 1. Disease Categories for Document Annotation. The categories listed here are those to be used for the Disease System data type in the document metadata template. As each disease presentation may involve multiple organ systems or etiologies, a single clinical case report may correspond to multiple categories. These categories largely follow those used to differentiate sections of the International Statistical Classification of Diseases and Related Health Problems, revision 10 (ICD-10) code system: corresponding ICD-10 chapters and code ranges are provided. Some categories, such as that for *oral and maxillofacial* disease, correspond to multiple sections of the ICD-10 system.

Data Type	Example #1	Example #2 (Cameron and McClain 1986)
Document and Annotation Identification		
Internal ID	CCR005	CCR2000
Annotation Date	Mar 2 2018	Mar 1 2018
Case Report Identification		
Title	A case of endocarditis.	Ocular histopathology of acrodermatitis enteropathica.
Authors	Grant AB;Chang CD	Cameron JD;McClain CJ
Year	2017	1986
Journal	World Journal of Medicine and Case Reports	British Journal of Ophthalmology
Institution	Department of Medicine, Division of Cardiology, First General Hospital, Boston, Massachusetts, USA	Department of Ophthalmology, University of Minnesota Medical School, Minneapolis, Minnesota 55455
Corresponding Author	Grant AB	Cameron JD
PMID	25555555	3756122
DOI	10.1011/wjmcr.2017.11.001	NA
Link	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9555555/	https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1040795/
Language	English	English
Medical Content		
Key Words	brucellosis; endocarditis; mitral valve	NA
Demography	37-year-old male	male child
Geographic Locations	Florida; Rio de Janeiro, Brazil	NA
Life Style	smoker; drinks alcohol occasionally	NA
Family History	third of five children of consanguineous parents; younger brother has chronic eczema	NA
Social History	construction worker	NA
Medical/Surgical History	history of fatigue	8 pound 9 ounce (3884 g) product of an uncomplicated, full term pregnancy; in good health until age 1 month when he developed a blistering skin rash on his cheeks; rash spread to involve the skin around the eyes, nose, and mouth; skin lesions were also noted on the abdomen and extremities; diarrhoea and failure to thrive; skin biopsy at that time showed parakeratosis typical of acrodermatitis enteropathica; treated over the next six years with intermittent courses of broad spectrum antibiotics, breast milk, and diodoquin; partially responded; developed total alopecia, intermittent acrodermatitis, and intermittent diarrhoea with suboptimal weight gain; spasticity attributed to central nervous system involvement by the ae had developed by 8 months of age; several episodes of cardiopulmonary arrest at 11 months; lack of co-ordination of his vocal cords; tracheostomy; by age 18 months the child developed searching nystagmus associated with bilateral optic atrophy and slight attenuation of retinal vessels as well as signs of psychomotor retardation; bilateral keratoconjunctivitis; skin rash; second skin biopsy performed at age 3 again showed parakeratosis typical for ae; severe skin rash and diarrhoea; bilateral gross anterior corneal opacities were seen which had completely resolved by the time he was reexamined at age five; frequent infections

		including otitis media, urinary tract infections, and skin infections
Disease System	cardiovascular; infectious	digestive; skin; eye; rare
Signs and Symptoms	palpitations and dyspnea in the previous week; presented with lethargy, headache, and chills	severe blepharoconjunctivitis and bilateral anterior corneal vascularisation; severe skin rash and diarrhoea; gram-negative bacterial sepsis; skin lesions typical of acrodermatitis enteropathica, absence of thymic tissue, marked degeneration of the optic nerves, chiasm, and optic tracts and extensive cerebellar degeneration
Comorbidity	hypertension; hyperlipidemia	NA
Diagnostic Techniques and Procedures	Physical examination; electrocardiography; blood cultures	ocular examination; necropsy
Diagnosis	Brucella endocarditis	acrodermatitis enteropathica
Laboratory Values	increase in c-reactive protein (9 mg/dl); alkaline phosphatase (250 u/l)	NA
Pathology	Brucella melitensis was cultured from blood samples	right and left eyes were similar in appearance; corneal epithelium was reduced in thickness to one to three cell layers of flattened squamous epithelial cells over the entire surface of the cornea; all polarity of the epithelium was lost. bowman's membrane could be identified only in the periphery of the right cornea. no bowman's membrane could be identified in the left cornea. neither degenerative nor inflammatory pannus could be identified in either eye; extensive atrophy of the circular and oblique muscles of the ciliary body; some posterior migration of lens capsular epithelium and early cortical degenerative changes; extensive degeneration of the retinal pigment epithelium throughout the posterior pole; retina was attached and showed mild autolytic changes throughout; some preservation of rod and cone outer segments in the posterior pole, however, these structures were completely lost anterior to the equator; extensive loss of the ganglion cell and nerve fibre layers of both eyes; nearly complete atrophy of the disc and adjacent optic nerve
Pharmacological Therapy	gentamycin 240 mg/iv/daily	NA
Inverventional Therapy	prosthetic valve replacement	NA
Patient Outcome Assessment	recovery was uneventful; discharged home	died in 1971 (age 7)
Diagnostic Imaging/Videotape Recording	2;1;0;1	7;0;0;0
Relationship to Other Case Reports	5555555	23430849
Relationship with Clinial Trial	NCT05555123	NA
Crosslink with Database	MedlinePlus Health Information: https://medlineplus.gov/ency/article/000597.htm	HighWire - PDF: http://bjo.bmj.com/cgi/pmidlookup?view=long&pmid=3756122; Europe PubMed Central: http://europepmc.org/abstract/MED/3756122; Genetic Alliance: http://www.diseaseinfosearch.org/result/143
Acknowledgements		
Funding Source	National Institutes of Health/National Heart, Lung, and Blood Institute	The Minnesota Lions Club; Research to Prevent Blindness; Veterans Administration; Office of Alcohol and Other Drug Abuse Programming of the State of Minnesota
Award Number	R01HL123123 (to AG)	NA
Disclosures/Conflict of Interest	Dr. Grant is a paid spokesperson for DrugCo.	NA

References	4	27

Table 2. Standardized Metadata Template for Clinical Case Reports, with Example Annotations. A set of features common to clinical case reports and facilitating their concept-level annotations is shown here. This template is arranged into three primary sections: Identification, Medical Content, and Acknowledgments, denoting the purpose and additional value afforded by each type of case report feature. This table contains two sets of example annotations, one of a fictionalized case report, and another set derived from a report on the condition acrodermatitis enteropathica[23].

Supplementary File 1. Text of a clinical case report (Chang *et al.* 2017). Please click here to download this file.

Discussion

Implementation of a standardized metadata template for CCRs can make their content more FAIR, expand their audience, and extend their applications. Following the traditional use of CCRs as educational tools in medical communications, healthcare trainees (*e.g.,* medical students, interns, and fellows), and biomedical researchers may find that summarized case report contents enable more rapid comprehension. The greatest strength of metadata standardization with CCRs, however, is that indexing these data transforms otherwise isolated observations into interpretable patterns. The protocol provided here can serve as the first step in a workflow for working with CCRs, whether this workflow consists of epidemiological analysis, post-marketing drug or treatment surveillance, or broader surveys of pathogenesis or therapeutic efficacy. Structured features identified within CCRs can provide a useful resource for researchers focusing on disease presentations and treatments, particularly for rare conditions. Clinical researchers may find data on past treatment regimens to analyze recorded symptoms or side effects and degree of improvement under previous standards of care. The data may also drive broader analyses of a new treatments based on efficacy, lack of adverse effects or toxicity, or on drug targeting differences in gender, age group, or genetic background.

The benefits provided by structured metadata are similarly applicable to computational workflows designed to parse or model medical language. Structured CCR features may also provide evidence of areas where report authors may provide more easily machine-readable (and in some cases, human-readable) content. Variance among CCRs can result from a lack of explicitly provided observations: *e.g.,* a patient's exact age may not be specified. Similarly, clinicians may not mention tests if the diagnostics or their results were considered trivial. By providing examples of gaps necessary for in-depth analysis, enforcing structure on CCRs highlights potential improvements. In a broader perspective, a greater availability of structured text data from medical documents supports natural language processing (NLP) efforts to learn from big data in healthcare[24,25].

Disclosures

The authors have nothing to disclose.

Acknowledgements

This work was supported in part by National Heart, Lung, and Blood Institute: R35 HL135772 (to P. Ping); National Institute of General Medical Sciences: U54 GM114833 (to P. Ping, K. Watson, and W. Wang); National Institute of Biomedical Imaging and Bioengineering: T32 EB016640 (to A. Bui); a gift from the Hoag Foundation and Dr. S. Setty; and the T.C. Laubisch endowment at UCLA (to P. Ping).

References

1. Ban, T.A. The role of serendipity in drug discovery. *Dialogues in Clinical Neuroscience.* **8** (3), 335-44, at <http://www.ncbi.nlm.nih.gov/pubmed/17117615> (2006).
2. Cabán-Martinez, A.J., García-Beltrán, W.F. Advancing medicine one research note at a time: the educational value in clinical case reports. *BMC Research Notes.* **5** (1), 293 (2012).
3. Vandenbroucke, J.P. In Defense of Case Reports and Case Series. *Annals of Internal Medicine.* **134** (4), 330 (2001).
4. Bayoumi, A.M. The storied case report. *Canadian Medical Association Journal.* **171** (6), 569-570 (2004).
5. Pasteur, L. Méthode pour prévenir la rage après morsure. *Comptes rendus de l'Académie des Sciences.* **101**, 765-774 (1885).
6. Pearce, J. Louis Pasteur and Rabies: a brief note. *Journal of Neurology, Neurosurgery & Psychiatry.* **73** (1), 82-82 (2002).
7. Keefer, C.S., Blake, F.G., Marshall, E.K.J., Lockwood, J.S., Wood, W.B.J. PENICILLIN IN THE TREATMENT OF INFECTIONS. *Journal of the American Medical Association.* **122** (18), 1217 (1943).
8. Akers, K.G. New journals for publishing medical case reports. *Journal of the Medical Library Association : JMLA.* **104** (2), 146-149 (2016).
9. Wilkinson, M.D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data.* **3**, 160018 (2016).
10. Beeler, G.W. HL7 Version 3-An object-oriented methodology for collaborative standards development. *International Journal of Medical Informatics.* **48** (1-3), 151-161 (1998).
11. *HL7 FHIR Release 3 (STU; v3.0.1-11917).* at <http://hl7.org/implement/standards/fhir/>. (2018).
12. McDonald, C.J. LOINC, a Universal Standard for Identifying Laboratory Observations: A 5-Year Update. *Clinical Chemistry.* **49** (4), 624-633 (2003).
13. *CDC/National Center for Health Statistics ICD-10-CM Official Guidelines for Coding and Reporting.* at <https://www.cdc.gov/nchs/data/icd/10cmguidelines_fy2018_final.pdf> (2017).
14. Rajeev, D. *et al.* Development of an electronic public health case report using HL7 v2.5 to meet public health needs. *Journal of the American Medical Informatics Association.* **17** (1), 34-41 (2010).
15. Biesecker, L. Mapping phenotypes to language: a proposal to organize and standardize the clinical descriptions of malformations. *Clinical Genetics.* **68** (4), 320-326 (2005).

16. Riley, D.S. *et al.* CARE guidelines for case reports: explanation and elaboration document. *Journal of Clinical Epidemiology.* **89**, 218-235 (2017).
17. Cohen, K.B. *et al.* Coreference annotation and resolution in the Colorado Richly Annotated Full Text (CRAFT) corpus of biomedical journal articles. *BMC Bioinformatics.* **18** (1), 372 (2017).
18. Sun, W., Rumshisky, A., Uzuner, O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *Journal of the American Medical Informatics Association.* **20** (5), 806-813 (2013).
19. Savova, G.K. *et al.* Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association.* **17** (5), 507-513 (2010).
20. Soysal, E. *et al.* CLAMP - a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association.* **25** (3), 331-336 (2018).
21. Bender, D., Sartipi, K. HL7 FHIR: An Agile and RESTful approach to healthcare information exchange. *Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems.* 326-331 (2013).
22. Chang, K. *et al.* Human Infection with Burkholderia thailandensis, China, 2013. *Emerging Infectious Diseases.* **23** (8), 1416-1418 (2017).
23. Cameron, J.D., McClain, C.J. Ocular histopathology of acrodermatitis enteropathica. *British Journal of Ophthalmology.* **70** (9), 662-667 (1986).
24. Maddox, T.M., Matheny, M.A. Natural Language Processing and the Promise of Big Data. *Circulation: Cardiovascular Quality and Outcomes.* **8** (5), 463-465 (2015).
25. Kreimeyer, K. *et al.* Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review. *Journal of Biomedical Informatics.* **73**, 14-29 (2017).

SCIENTIFIC DATA

OPEN

Data Descriptor: A reference set of curated biomedical data and metadata from clinical case reports

J. Harry Caufield[1,2,*], Yijiang Zhou[1,2,3,*], Anders O. Garlid[1,2,*], Shaun P. Setty[4], David A. Liem[1,2,5], Quan Cao[1,2], Jessica M. Lee[1,2], Sanjana Murali[1,2], Sarah Spendlove[1,2], Wei Wang[1,6,7,8], Li Zhang[3], Yizhou Sun[1,7], Alex Bui[1,6,9], Henning Hermjakob[1,10], Karol E. Watson[1,5] & Peipei Ping[1,2,5,6,8]

Received: 18 June 2018

Accepted: 27 September 2018

Published: 20 November 2018

Clinical case reports (CCRs) provide an important means of sharing clinical experiences about atypical disease phenotypes and new therapies. However, published case reports contain largely unstructured and heterogeneous clinical data, posing a challenge to mining relevant information. Current indexing approaches generally concern document-level features and have not been specifically designed for CCRs. To address this disparity, we developed a standardized metadata template and identified text corresponding to medical concepts within 3,100 curated CCRs spanning 15 disease groups and more than 750 reports of rare diseases. We also prepared a subset of metadata on reports on selected mitochondrial diseases and assigned ICD-10 diagnostic codes to each. The resulting resource, Metadata Acquired from Clinical Case Reports (MACCRs), contains text associated with high-level clinical concepts, including demographics, disease presentation, treatments, and outcomes for each report. Our template and MACCR set render CCRs more findable, accessible, interoperable, and reusable (FAIR) while serving as valuable resources for key user groups, including researchers, physician investigators, clinicians, data scientists, and those shaping government policies for clinical trials.

Design Type(s)	data integration objective • metadata search and retrieval objective •
Measurement Type(s)	case report textual entity
Technology Type(s)	digital curation
Factor Type(s)	disease (OBI) • Study Publication Date • geographic location
Sample Characteristic(s)	Homo sapiens

[1] The NIH BD2K Center of Excellence in Biomedical Computing, University of California at Los Angeles, Los Angeles, CA 90095, USA. [2] Department of Physiology, University of California at Los Angeles, Los Angeles, CA 90095, USA. [3] Department of Cardiology, First Affiliated Hospital, Zhejiang University School of Medicine, 310003, Hangzhou, Zhejiang, P.R. China. [4] Department of Pediatric and Adult Congenital Cardiac Surgery, Miller Children's and Women's Hospital and Long Beach Memorial Hospital, Long Beach, CA 90806, USA. [5] Department of Medicine/Cardiology, University of California at Los Angeles, Los Angeles, CA 90095, USA. [6] Department of Bioinformatics, University of California at Los Angeles, Los Angeles, CA 90095, USA. [7] Department of Computer Science, University of California at Los Angeles, Los Angeles, CA 90095, USA. [8] Scalable Analytics Institute (ScAi), University of California at Los Angeles, Los Angeles, CA 90095, USA. [9] Department of Radiological Sciences, University of California at Los Angeles, Los Angeles, CA 90095, USA. [10] Molecular Systems Cluster, European Molecular Biology Laboratory-European Bioinformatics Institute, Wellcome Genome Campus, Cambridge, UK. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to P.P. (email: ppingucla@gmail.com)

Background & Summary

Clinical case reports (CCRs) are a fundamental means of sharing observations and insights in medicine. A wealth of knowledge exists within this venerated and actively growing area of medical publishing[1]. Unfortunately, many of these largely unstructured text data lack adequate metadata denoting specific clinical events. As a result, our ability to curate a comprehensive set of reports relevant to a disease of interest is inadequate, and extraction of the clinical insights contained within is limited. Our Metadata Acquired from CCRs (MACCRs), prepared by domain experts, enriches CCRs with detailed metadata, establishing a findable, accessible, interoperable, and reusable (FAIR)[2] data resource and empowering researchers to form *in silico* cohorts of disease, amplify small sample size studies, and leverage the cumulative power of many reports for statistical analyses.

More than 1.89 million CCRs have been published as of August 2018, with over half a million in the last decade. CCRs serve as teaching tools, elucidating the reasoning behind diagnoses and management conclusions. Throughout history, CCRs have provided accounts of emerging diseases, their treatments, and their genetic backgrounds[3–5]. The first treatment of human rabies by Louis Pasteur in 1885[6,7], the first application of penicillin in patients[8], and an early study of the retroviral human T-cell lymphoma virus (HTLV)[9] were all reported through CCRs[3]. CCRs are a first line of evidence, serving as both a source of individual pathologies and the basis of study for population-level trends that may otherwise go unnoticed[5,10]. These reports remain the only formally published mechanism for exchanging clinical observations and are not subject to the extensive privacy concerns of electronic health records (EHRs).

The text data in these CCRs are largely unstructured, vary widely in content, and contain interrelated phenomena, limiting their use as a structured resource. Existing resources for indexing and enforcing structure on biomedical documents, including Medical Subject Headings (MeSH) and their associated tools[11,12], the ongoing Informatics for Integrating Biology and the Bedside (i2b2) curated resources[13,14], and the CRAFT[15], AnatEM[16], NCBI disease[17], and newly-available PubMed Phrases[18] corpora are useful for focused natural language processing (NLP) approaches. Annotations of free-text resources, such as adverse reactions on drug labels[19], are additional sources of expert-guided structure for unstructured text. However, these resources have limited applications with CCRs, as most of them have not been designed to model clinical narratives. As a result, there exists a significant gap between the clinical data we generate and our ability to convert it to knowledge. Clinical controlled vocabularies and coding systems (e.g., ICD-10[20] or LOINC[21]) can help to bridge this gap but have rarely been used with published clinical reports. Furthermore, despite the existence of several resources (e.g., immuneXpresso[22]) offering utility on biomolecular text mining in the context of disease, information is fragmented with limited clinical perspective. Therefore, we set out to expand the value of CCRs as a vital biomedical knowledge resource through extensive metadata creation.

This MACCR dataset contains free-text selections from 3,100 CCRs, with over 32,000 manually annotated medical features across a variety of clinical presentations structured into 15 different types of medical concepts (Table 1; Figs 1 and 2). We curated the CCRs to ensure they would provide a general model of clinical language, then performed expert-guided annotation to extract a comprehensive and nuanced array of textual features. As CCRs often describe infrequently observed symptoms and diseases, our MACCR set includes over 700 reports of rare disease presentations, with additional focus on 7 selected mitochondrial diseases. Our goal in developing this resource is to provide a manageable, structured set of metadata on clinical events and case descriptions. In so doing, we render case reports more findable and the information contained within more accessible. One effect of making CCRs more FAIR is to facilitate the discovery and study of these invaluable sources of clinical insight, enhancing the potential for researchers and clinicians to gain a better understanding of diseases and their treatment. To guide those interested in employing the MACCR dataset in their research, we present 5 feasible primary study objectives (detailed in Usage Notes) that may be pursued through downstream analysis by researchers, physician investigators, clinicians, data scientists, IP officers, pharmaceutical companies, and those shaping government policies for clinical trials.

Methods

To assemble our set of Metadata Acquired from Clinical Case Reports (MACCRs), we performed three primary stages of tasks. First, we designed a structured data template including the primary features of most case reports and curated a corpus of case report documents along with their associated metadata records. Next, we extracted metadata from each document in the corpus using manual and automated methods. Finally, all metadata records were aggregated into a single set of documents and verified. An overview of our approach is provided in Fig. 1.

Metadata template and curation

Our data template (Table 1) includes metadata on three aspects of information collected for each CCR: identification, medical content, and acknowledgements. The identification features are primarily used to distinguish documents by bibliographic features, including the title, authors, journal, publication date, and unique identifiers such as DOI and PMID. The majority of the metadata collected in the medical content is concept-level. This includes text segments corresponding to 15 different clinical concepts, detailed in Table 1, where a segment may vary in length from a single word to multiple sentences. A

Field	Data Type	Example
Case Report Identification (Findable)		
Title	Text	Case report: a case of cardiogenic shock and hyperparathyroidism.
Authors	Text	Neeley AB, Mossman ET
Year	Text	2017
Journal	Text	Midwest Journal of Medicine
Institution	Text	Department of Cardiology, Mt Vernon Hospital, Mt Vernon, Wisconsin, USA
Corresponding Author*	Text	Neeley AB
PMID	Identifier	29999555
DOI	Identifier	10.1011/mwjmed.2017.10.001
Link	Identifier	http://www.mwjmed.org/doi/full/10.1011/mwjmed.2017.10.001
Language*	Text	English
Medical Content (Accessible, Interoperable, Reusable)		
Key Words	Text	Shock, cardiogenic; hyperparathyroidism; fatigue; headache
Demography**	Text	Male; 40 years of age
Geographic Locations***	Text	Mt Vernon, Wisconsin, USA
Life Style	Text	Smoker
Family History	Text	no family history of heart disease
Social History	Text	worked as a truck driver
Medical/Surgical History	Text	history of fatigue; splenectomy performed six years previously
Disease System	Text	Cardiovascular diseases
Signs and Symptoms	Text	presented with lethargy, headache, diaphoresis, and twitching in all four limbs; cardiac enzyme levels were elevated, ventricular tachycardia
Comorbidity	Text	alopecia
Diagnostic Techniques and Procedures	Text	Electrocardiogram; dual energy X-ray absorptiometry (DXA)
Diagnosis	Text	hyperparathyroidism
Laboratory Values	Text	serum calcium concentration was 3.0 mmol per liter; complete blood cell counts normal
Pathology	Text	endomyocardial biopsy did not reveal a myocardial pathology
Pharmaceutical Therapy	Text	bisphophonates
Interventional Therapy	Text	ventilated on the 2nd day post-surgery due to respiratory distress
Patient Outcome Assessment	Text	Patient developed refractory shock; died of persistent ventricular tachycardia
Diagnostic Imaging/Videotape Recording****	Numerical	3;0;0;0
Relationship to Other Case Reports*	Text / Identifier	PMID: 5555555
Relationship with Clinical Trial*	Text / Identifier	PMID: 5551111
Crosslink with Database*	Text / Identifier	MedlinePlus Health Information : https://medlineplus.gov/parathyroiddisorders.html
Acknowledgements		
Funding Source	Text	National Institutes of Health/National Heart, Lung, and Blood Institute
Award Number	Identifier	R01HL123123 (to AN)
Disclosures/Conflict of Interest	Text	Dr. Neeley is a paid consultant for Medicaltech Inc.
References	Numerical	12

Table 1. Standardized metadata template for clinical case reports. *The template and associated processing workflow support use of these fields, though their values are not provided in the MACCR set. **Demography details are converted to consistent values prior to inclusion in the MACCR set. ***If not provided within document text, geographic location is inferred from the associated institution. ****Purely a numerical count of the total number of clinical images, figures, videos, and tables, respectively, published along with the main text of the report.A set of features common to clinical case reports and facilitating their concept-level metadata extraction. This template is arranged into three primary sections: Identification, Medical Content, and Acknowledgments, denoting the purpose and additional value afforded by each type of case report feature. Here, we have also included relevancy of the first two categories to promoting FAIR standards. A single document contains the majority of these features; metadata records include the span of these features (i.e., the value referring to a single concept). Examples shown here are simulated but representative of dataset contents. Data Type refers to the type of source data, rather than the dataset contents; this may be "Text" if free-text (this may contain numerical components, though these are identified in subsequent processing steps), "Identifier" if a unique database identifier or other structured value specific to the document, or "Numerical". Please note that the Acknowledgements section provided here, including the Disclosures/Conflict of Interest statement, is an example only and not intended to claim any funding provided to or competing interest by the authors.

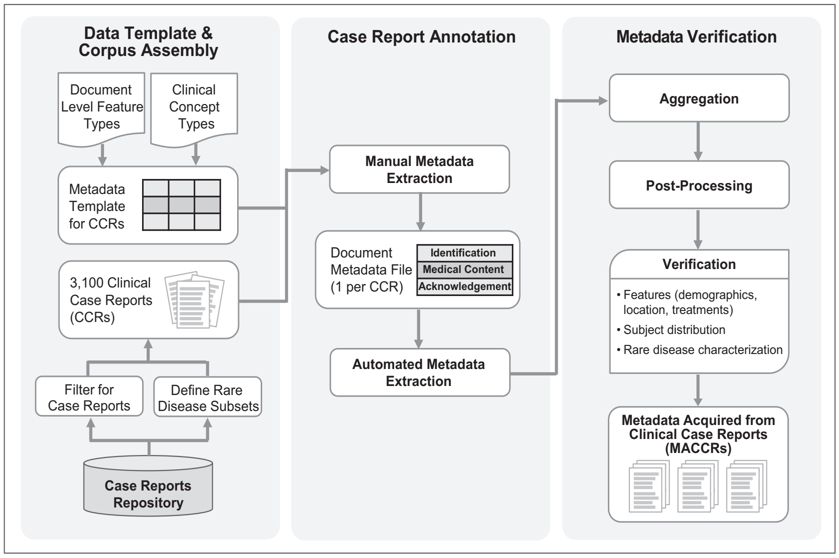

Data Template & Corpus Assembly
Document Level Feature Types
Clinical Concept Types
Metadata Template for CCRs
3,100 Clinical Case Reports (CCRs)
Filter for Case Reports
Define Rare Disease Subsets
Case Reports Repository
Case Report Annotation
Manual Metadata Extraction
Document Metadata File (1 per CCR)
Identification
Medical Content
Acknowledgement
Automated Metadata Extraction
Metadata Verification
Aggregation
Post-Processing
Verification
• Features (demographics, location, treatments)
• Subject distribution
• Rare disease characterization
Metadata Acquired from Clinical Case Reports (MACCRs)

Figure 1. Data creation workflow. To assemble the set of Metadata Acquired from Clinical Case Reports (MACCRs) we first assembled a corpus of 3,100 published case reports. Using a document metadata template including document-level identification and acknowledgement features (i.e., citation data such as title; Medical Subject Headings [MeSH terms]) and concept-level medical content features (e.g., descriptions of patient demography, clinical signs and symptoms, or outcomes), a team of medical experts manually identified text from each document corresponding to each feature. More specific terms were identified through automated approaches. To finalize this dataset, we aggregated all document metadata records into a single file. We normalized categorical features, verified, and cleaned these data, which are available as the MACCR set.

single concept may also correspond to multiple text segments within a given document. The acknowledgements metadata provides details for disclosures and sponsorship.

The metadata in our dataset are sourced from CCRs indexed by MEDLINE. We first defined the CCR corpus from which our metadata are extracted using our *heartCases* software (https://github.com/UCLA-BD2K/heartCases); *heartCases* aggregates metadata provided with each MEDLINE entry to determine features common to a set of documents (e.g., their top publication years, journals, and MeSH descriptors). We then assigned each CCR to at least one of 16 disease categories based on their MeSH descriptors or presence of a MeSH descriptor in a title (Table 2). The set of terms for each disease category was defined using corresponding segments of the MeSH Tree (https://meshb.nlm.nih.gov/treeView); for example, all CCRs in the “digestive” category match primary and entry terms at or below the following points on the MeSH Tree: A03.159, A03.556, A03.620, A03.734, C06.130, C06.198, C06.267, C06.301, C06.405, C06.552, C06.689, C06.844, and G10.261, or a total of 3,356 different terms. A match may include a matched MeSH term used to index the document or presence of the term in the document title. With the goal of assembling a generally representative set of CCRs, we selected documents from the larger corpus such that the assignment to each disease category resembled that seen across all CCRs. The most popular topics across CCRs in general are therefore popular topics in our source CCR set as well. In an effort to ensure representation of a variety of disease presentations – one of the inherent strengths of case reports – we also selected reports of rare diseases (i.e., those affecting fewer than 200,000 individuals in the United States, as defined by the NIH NCATS Genetic and Rare Diseases Information Center [https://rarediseases.info.nih.gov/diseases/]) such that the set included 5 to 10 reports each for over 100 rare diseases. The resulting reference metadata set is sourced from 3,100 CCRs spanning 15 major disease categories, as well as a subset of rare diseases (Table 2).

Manual annotation was performed by 12 annotators familiar with medical terminology and, for subsets of rare diseases, with the clinical features underlying these diseases (e.g., for rare mitochondrial diseases, annotators possessed an understanding of the underlying mitochondrial physiology and corresponding mutations). Our roster of curators comes from clinical fellows (4), post-doctoral fellows (6), and senior graduate students (2). Full instructions followed by annotators are detailed in our *Metadata Extraction Guide*, included with the data files (Metadata Extraction Guide, Data Citation 1). For each CCR in the corpus, one annotator read the full text of the document and extracted text phrases

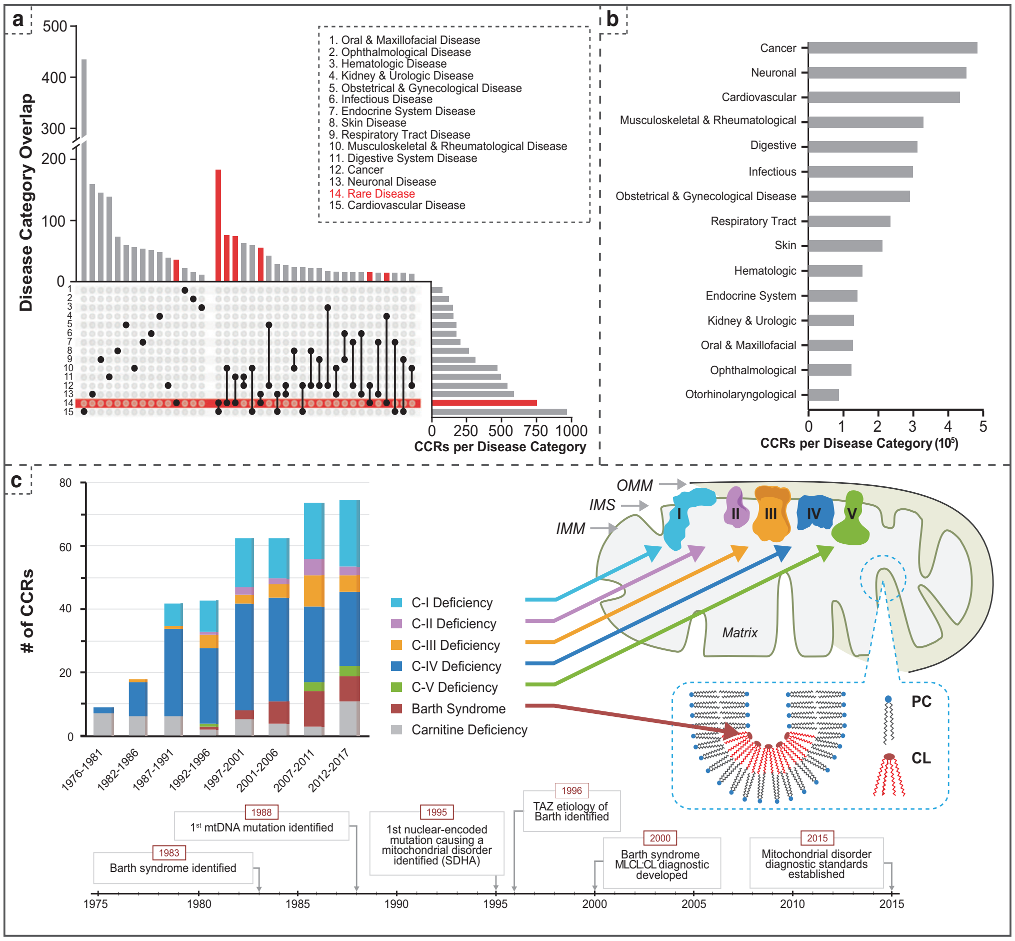


Figure 2. Contents of the MACCR dataset. (**a**) Concept overlap among CCRs in the MACCR set. We assigned each report to one or more disease categories based on involvement of particular organ systems. Reports describing presentations of rare diseases, including mitochondrial rare diseases, were assigned to the Rare Disease category as well. Here, we show the total count of reports labeled with single or multiple disease categories (Disease Category Overlap, top) as an UpSet plot[59]. Counts of reports involving rare diseases (n = 751) are highlighted in red. Total counts irrespective of overlap with other categories are also shown at right. For example, 435 CCRs involve cardiovascular disease (CVD) without specific involvement of other organ systems, yet 967 CCRs involve CVD alone *or* along with other disease categories. Otorhinolaryngologic reports constitute the smallest category in the MACCR set (n = 58); their counts are omitted here. (**b**) Distribution of disease categories across all published case reports. Here, we determined disease category assignment across all 1.89 million published CCRs (as of August 2018) using sets of MeSH terms corresponding to each category. As in Part A, a report may belong to multiple categories. More than a quarter of all reports in this broad set involve cancer, differing from the report distribution in the MACCR set, though in both sets, cancer, cardiovascular disease, and neuronal disease are the most common three disease categories. (**c**) Contribution of mitochondrial disease CCRs. 246 CCRs cover a sample of rare mitochondrial diseases, including Barth syndrome, carnitine deficiency, and deficiencies of the respiratory chain complexes. The distribution of CCR publication year is displayed on the left for each disease, and the affected components of the mitochondrion are represented in the diagram to the right. Complex I, II, III, IV, and V deficiencies each cause impairment in their respective

component of the respiratory chain, resulting in a range of cardiovascular, neurological, muscular, and metabolic phenotypes. Barth syndrome is caused by a mutation in the tafazzin protein that renders it incapable of creating properly formed cardiolipin (CL) for the inner mitochondrial membrane (IMM). Phosphatidylcholine (PC) is unable to form the tight bends of the cristae, severely limiting energy generation and leading to cardiovascular complications. The timeline below depicts key advancements and discoveries relating to rare mitochondrial disease diagnosis. OMM: outer mitochondrial membrane. IMS: intermembrane space.

Category	Description
cancer	cancer or neoplasms
nervous	brain, spine, or nerve involvement
cardiovascular	heart or cardiovascular involvement, not including conditions specific to the blood
musculoskeletal and rheumatic	muscle, bone, joints, or connective tissue involvement
digestive	gastrointestinal involvement, including liver, pancreas, or gallbladder
obstetrical and gynecological	pregnancy, childbirth, the female reproductive system, or the breasts
infectious	infection by microorganisms
respiratory	respiratory tract involvement
hematologic	blood, bone marrow, lymph nodes, or spleen involvement
kidney and urologic	kidney or bladder involvement; any involvement of the male reproductive organs
endocrine	endocrine gland involvement and metabolic disorders
oral and maxillofacial	mouth, jaws, head, face, or neck, including dental issues
eye	eye involvement and visual issues
otorhinolaryngologic	ear, nose, or throat involvement, including hearing issues
skin	skin involvement
rare	rare diseases; see above

Table 2. Disease system categories. A set of categories for grouping case reports on the basis of related symptoms, co-morbidities, and etiologies. A report may belong to more than one category, particularly in clinical presentations involving multiple disease systems and/or cancer of one or more systems. The special category of 'rare' is specific to reports of rare diseases affecting no more than 200,000 individuals in the United States at a time.

corresponding to each component of the data template, avoiding extended discussion sections or clinical studies except when these sections were the only source of target metadata. They populated the template with these phrases, delimiting each distinct phrase with a semicolon, then completed the annotation set by adding bibliographic metadata as presented in the document. Annotators assigned each CCR one or more of 16 different disease categories based on both document content and MeSH descriptors, each corresponding to an organ system or general classification of disease presentations (e.g., cardiovascular, endocrine, infectious, or rare disease, etc.; Table 2). These categories were assigned based on presence of related symptoms, co-morbidities, and etiologies rather than primary diagnosis in order to determine conceptual overlap between cases. We did not include a category for congenital or genetic disorders and focused instead on the clinical signs of these presentations. Finally, annotators determined the count of additional data elements in each CCR, counting clinical images (i.e., any photograph, micrograph, or direct result of a diagnostic procedure such as an electrocardiogram), figures (i.e., any assembled image or data visualization), tables, and videos, not counting supplemental materials.

ICD codes and interoperability

In order to facilitate interoperability with existing ontologies and support efforts at gaining a systematic understanding of disease, we assigned codes from a standard set of diagnostic identifiers to reports from a subset of the MACCR set. We sought to reveal shared and common symptoms as well as rare and unique characteristics underlying mitochondrial disease, constructing digital maps of disease symptomology for documents in the rare mitochondrial disease (RMD) subset using ICD-10-CM codes (International Statistical Classification of Diseases and Related Health Problems, tenth revision, clinical modification, 2018 release). The document corpus is assembled from 246 CCRs, each describing an individual presentation of one of six selected mitochondrial diseases (Barth syndrome, primary carnitine deficiency, or a deficiency of mitochondrial complex II, III, IV or V). For each document, two separate annotators familiar with mitochondrial disease pathologies identified specific concepts within the full text of the case report corresponding to ICD-10-CM codes, including those for symptoms. Codes are included if their concepts are part of a given patient's clinical presentation, but not if they are only discussed or proposed. In final data files, observations are treated as binary values (1 if identified, 0 if not identified) and split into two different sets: in the first, each code is provided separately, while in the second, codes are aggregated into categories on the basis of their ICD-10 code blocks (https://www.cdc.gov/nchs/data/icd/10cmguidelines_fy2018_final.pdf). For example, one or more codes between C00 and D49 assigned for a given document will yield a score of 1 for the "Neoplasms" category corresponding to this code block. By incorporating ICD-10 codes and MeSH descriptors, we enhance interoperability in this subset of the MACCR set.

Data quality control and validation

Quality control was implemented throughout the manual curation process with regular milestone meetings and a closely collaborative research environment designed to align and standardize methods for metadata extraction and foster consistency among all curators. Manual metadata extraction from individual CCRs was followed by aggregation and additional quality control through post-processing. All metadata records were combined into a single file using Python and R scripts (*Extract.py* and *QualityControl.R*; see Code Availability). This workflow uses basic natural language processing functions to perform the following: retrieval and verification of document details (e.g., title, database identifiers, and publication details), assignment of each document to one of our major disease categories, and conversion of patient age to a numerical value. For patient age identification in particular, all values are treated as integers, rounded down (e.g., a patient aged 5.5 years is assigned a value of 5; those aged <1 year are assigned a value of zero), or are estimated when not explicitly provided (e.g., a report of a patient in their "sixties" is assigned an age value of 65). Features with insufficient detail in the text are assigned a value of "NA". All text features are checked to ensure formatting consistency, and to verify database cross-links: titles and author names are compared to MEDLINE records, DOIs and links are confirmed and added where missing, and journal names are normalized to include their full names as presented in the NLM Catalog (https://www.ncbi.nlm.nih.gov/nlmcatalog/journals), but with the preceding "the" omitted. All text field delimiters (i.e., those denoting separate text segments within a single field) are confirmed to be semicolons. Additionally, most text fields are converted to lowercase characters to enable easier comparison and named entity recognition (NER[23,24]). Final validation of the dataset is performed through observation of distributions of features within the MACCR set comparable to those seen in larger collections of clinical documents (see Technical Validation for additional details).

Geographic distribution analysis is performed for validation and visualization of MACCR set features. We developed an R Shiny app designed specifically for performing this analysis with clinical case reports (see Code availability). Briefly, this app first identifies the occurrence of all case reports indexed by MEDLINE on the basis of their AD, or Affiliation, field. This field is largely unstructured and has changed in format over time, so in order to provide additional detail and consistency, the text is processed to identify specific names of countries and US states. Then, the institutional affiliation field of each MACCR record is parsed in an identical manner. Using a 2-proportion Z-test, counts of locations are compared between the set of all available CCRs and those in the subset contributing metadata to our MACCR set to identify locations with a statistically significant (i.e., higher or lower) difference in proportion. The differences are visualized on a world map with US states treated independently from each other.

Three additional files contain citation details, corresponding MeSH headings, and named entities contained within the MACCR set. The citations for documents corresponding to each metadata record are loaded into a Mendeley citation database and converted to BibTeX format. The list of all unique MeSH headings is prepared by isolating unique MeSH descriptors, without modifiers, from each MEDLINE-format citation of those corresponding to each metadata record in the MACCR set. To determine the extent to which the metadata text segments correspond to a controlled vocabulary of biomedical terms (i.e., with NER), we identify all named entities up to three words long present in each medical content field across all 3,100 MACCRs, based on entities within MeSH and SNOMED-CT as per the UMLS Metathesaurus[25].

Code availability

The code used to process and verify the MACCR dataset, along with documentation, is available at https://github.com/UCLA-BD2K/MACCRs. This repository includes all utilities used in the assembly and verification of the MACCR metadata set, with the exception of the following two pieces of software. Analysis of the complete corpus of CCRs, as part of the verification of this dataset, was done using *heartCases*, available at https://github.com/UCLA-BD2K/heartCases. The R Shiny App used for analysis and visualization of case report geographic distributions is available at https://github.com/UCLA-BD2K/Significant-Mapping.

Data Records

Starting from the manually curated set of CCRs as defined above, we obtain the full text records of each report from PubMed/MEDLINE corresponding to each respective PubMed entry identifier (PMID). Text corresponds to the PubMed Central document version where possible; all other text is curated from publicly-available document PDFs provided by journal publishers. As CCRs vary in structure and format (e.g., section headings vary, and a case description may be just one component of a document), curators identify a single, primary case presentation section within each published record, then identify text corresponding to the concepts within the CCR metadata extraction template (Table 1). The contents of the MACCR set are, therefore, metadata with respect to each CCR.

The primary data file (MACCRs.tsv, Data Citation 1) is provided in UTF-8, tab-delimited format, such that the metadata for each CCR is a single line in the file. Each column corresponds to a distinct metadata type. Within text columns, distinct text segments are delimited using semicolons and most are in lowercase only to facilitate easier searching. This file contains 1 metadata record for each of 3,100 unique documents. The corresponding reports have publication dates spanning from 1956 to 2018 and were originally published in 1,020 different journals. Across the 15 different concept-level free-text features identified within each report, the set contains 2,980 unique descriptions of diagnostic techniques/procedures and 3,026 unique clinical diagnoses, among other descriptions in context. Descriptions of all data fields are provided in our *MACCR File Guide* (MACCR File Guide, Data Citation 1). Full citations for all CCRs in the MACCR set are provided in BibTeX format within the citation file (MACCR_citations.bib, Data Citation 1).

We have provided our metadata extraction template (TEMPLATE.xlsx, Data Citation 1) to facilitate adaptation to new studies through the creation of similar CCR-based datasets. The template is provided as an Excel spreadsheet to ensure universality and ease of use. Each data type is identified in each row of the first column and corresponding values extracted from the CCR text are provided in the fourth column. The fifth column is used to identify values provided through PubMed records and indexing alone; for medical content, these values are MeSH descriptors (if they exist) and any provided modifiers. The second and third columns contain counts to indicate the presence of content in the fourth and fifth columns, such that a value of '1' corresponds to any value other than a blank cell. To enable this comparison, cells without any relevant information for a particular CCR are left blank. Please see the *Metadata Extraction Guide* (Metadata Extraction Guide for the MACCR set, Data Citation 1) for further details of the process.

We additionally provide the set of all unique MeSH headings applicable to the source CCRs for the MACCR set (MACCR_mesh.tsv, Data Citation 1). This file provides a list of headings in the first column, with one unique heading per line and without further modifications. The second column provides the MeSH root category (as per the 2018 release; https://www.nlm.nih.gov/mesh/) for each heading as a single letter, e.g., the heading "Spleen" is in the Anatomy category, or category A. For headings with multiple potential codes within the MeSH hierarchy, the category of the first listed in the MeSH index file is used (e.g., "Outcome Assessment (Health Care)" has codes in categories H and N and is considered to be in category H). The headings "Male" and "Female" have no location in the MeSH tree and no category. Here, all reports contributing metadata to the MACCR set have at least one associated MeSH descriptor through MEDLINE.

Named entity recognition (NER)[23,24] results (MACCR_entities.tsv, Data Citation 1) are provided as an additional means of illustrating concepts within the MACCR set. Each line in this file contains the PMID of the report corresponding to each metadata set, followed by a list of named entities identified within selected metadata fields, as indicated in the heading. Named entities are MeSH descriptors and SNOMED-CT terms, as available through UMLS resources[25].

Mitochondrial disease reports, covering six different diseases (Barth syndrome, primary carnitine deficiency, or a deficiency of mitochondrial complex II, III, IV or V), contribute 246 reports to this dataset. As the majority of these reports describe rare mitochondrial diseases, we refer to this subset as the RMD subset. Metadata in the RMD subset include codes from ICD-10-CM such that symptoms and diagnoses mentioned within each CCR are each identified using the most closely matching and specific ICD-10-CM code, including symptom codes (codes R00.0 through R99). 500 unique codes were identified across all RMD CCRs, with a total of 2,119 codes assigned. The presence or absence of each of these 500 ICD-10-CM codes for each CCR is provided in its own file (MACCR_RMD_ICD10.tsv, Data Citation 1), with one CCR per row, identified by its PMID in the first column and the RMD category (barth [for Barth syndrome], carnitine [for primary carnitine deficiency], or complex_I through complex_V [for mitochondrial complex deficiencies]) in the second column. Each of the 500 unique ICD-10-CM codes identified in the RMD set is represented in the following 500 columns. A value of "1" indicates a given code matches clinical events described in the CCR, while a value of "0" indicates matching material is not observed, though its omission from CCR text may not conclusively indicate its absence. Additionally, we provide a compressed version of these data indicating presence or absence of categories of codes in lieu of individual codes (MACCR_RMD_ICD10_Categories.tsv, Data Citation 1). This file contains the PMID for one CCR and an RMD category in its first and second columns, respectively, as in the file described above. The following columns correspond to one of 20 chapter titles from ICD-10-CM, each represented by a block of codes. A value of "1" indicates at least one ICD-10-CM code in the specified chapter's code block matches clinical events described in the CCR, while a value of "0" indicates matching material was not observed. Inclusion of category-based observation only in this file reduces the total observations to 1,073 across 20 blocks of codes.

This dataset itself meets the FAIR Data Principles. All files are provided through both Figshare (Data Citation 1) and through Dryad (Data Citation 2). Metadata are assigned a globally unique and persistent identifier and registered through both of these searchable resources to make them *F*indable; the metadata

FAIRshake metric	Score	Support for metric
1. A standardized ID or accession number is used to identify the dataset.	Yes; 1	Dataset provided with unique DOIs by Figshare and Dryad.
2. The dataset is described with metadata using a formal, broadly applicable vocabulary.	Yes; 1	Dataset is described using formal biomedical terminology, including diagnostic techniques and procedures, signs and symptoms, etc., as well as MeSH terms and ICD-10 codes.
3. Information is provided on the experimental methods used to generate the data.	Yes; 1	All methods for ACCR metadata template generation provided along with raw data files.
4. The dataset is hosted in an established data repository, if a relevant repository exists.	Yes; 1	Dataset is hosted on the Figshare and Dryad repositories.
5. The dataset can be downloaded for free from the repository.	Yes; 1	Dataset can be downloaded in .tsv format for free from the Figshare and Dryad repositories.
6. Version information is provided for the dataset.	Yes; 1	Versioning begins with v1, and updated versions of the set will be v2, v3 etc.
7. Contact information is provided for the creator(s) of the dataset.	Yes; 1	Contact information for the lead investigator is provided (Dr. Peipei Ping).
8. Information is provided describing how to cite the dataset.	Yes; 1	Citation information is provided along with this publication.
9. Licensing information is provided on the dataset's landing page.	Yes; 1	Licensing information is provided through Figshare, Dryad, and *Scientific Data*.

Table 3. **Support for FAIRShake metrics.** Here, we provide scores for each of the metrics in the FAIRshake (https://fairshake.cloud) rating system. The MACCR dataset fully meets each metric and therefore has a score of "1" for each.

are retrievable via this identifier using an open, standardized, and free communications protocol to make them *A*ccessible; the metadata set uses a formal, broadly applicable vocabulary and domain-recognized ontologies (MeSH and ICD-10) to make them *I*nteroperable; and the metadata files contains detailed provenance, licensing, and versioning information to make them *R*eusable. Out of the 9 metrics used by FAIRShake (https://fairshake.cloud/), our dataset provides all 9 of the necessary values (Table 3).

Technical Validation

Distribution of disease categories *vs.* published case reports

We intend the MACCR set to be representative of the semantic and lexical variation present in reports from a wide variety of medical presentations and specialties. Rather than focusing on reports describing a single type of disease, the metadata in the MACCR set is sourced from clinical presentations spanning 15 disease groups, along with an additional category for rare disease presentations (Fig. 2a, and Table 2). CCRs often describe rare disease presentations and clinically-relevant relationships not accessible through any other public source. As one example, CCRs provided evidence for the theory that gadolinium-based contrast agents are a trigger for the rare disease nephrogenic systemic fibrosis[26,27]. By including a wide variety of medical concepts in our corpus, we highlight the value of CCRs, particularly in their descriptions of novel and often complex diagnostic processes. Further, we provide validation of MACCR content relative to that seen across all CCRs outside the corpus.

Both the MACCR set and the set of all published CCRs are predominantly composed of reports of cancer, cardiovascular disease, and neuronal disease (Fig. 2a-b). The variety of topics covered within the MACCR set is demonstrated by the number of MeSH descriptors among the contributing reports: the source reports for the MACCRs are indexed with 5,326 unique terms (MACCR_mesh.tsv, Data Citation 1), or 12,980 unique terms with their modifiers. For comparison, the full 2018 MeSH ontology includes 28,239 descriptors (https://www.nlm.nih.gov/mesh/) and the set of all 1.89 million CCRs published as of August 2018 is indexed with 24,842 unique terms, or 786,256 unique terms with modifiers. Of the 5,326 MeSH terms in the MACCR set, 2,042 describe Diseases (category C), 1,208 describe Chemicals and Drugs (category D), and 793 describe Analytical, Diagnostic and Therapeutic Techniques, and Equipment (category E).

Rare disease subgroup membership

Our curation process ensures that the MACCR set includes descriptions of rare diseases across a variety of disease subtypes. Rare diseases are defined as those with an incidence in the U.S. of fewer than 1 in 20,000 individuals. These diseases are infrequently reported on in CCRs, primarily due to their inherent rarity. The contribution of these reports to the MACCR set offers a novel advantage by providing a more comprehensive representation of the language used in presentations of these diseases as compared to more frequently described clinical presentations. Figure 2a highlights the contribution of rare diseases to the MACCR set.

The MACCR set includes a focused subset of 246 rare mitochondrial disease (RMD) reports that describe presentations of one of six rare mitochondrial diseases. The distribution of these reports in the

MACCR set across their years of publication reflects time points of discovery and increased diagnosis (Fig. 2c, left); the timeline presented in Fig. 2c highlights significant advances in the identification of mitochondrial diseases and their etiology, as well as the development of diagnostic tools and standards. These mitochondrial diseases in particular can be distinguished on the basis of deficiencies in crucial molecular components: complex I, II, III, IV, and V deficiencies impair individual components of the respiratory chain, while Barth syndrome is characterized by malformed cardiolipin, disturbing mitochondrial membranes and reducing respiratory chain function (Fig. 2c, right). Though the resulting sets of symptoms overlap noticeably, the overall range of symptoms covers cardiovascular, neurological, muscular, and metabolic phenotypes, as evidenced by the range of ICD-10-CM codes assigned to this subset of CCRs. Our subset of mitochondrial diseases therefore demonstrates how a collective understanding of the presentation of a specific condition may be best gained through comparisons with those of biologically similar diseases.

To validate the ICD-10-CM codes we assigned to each RMD report, we compare our code assignments to those expected for specific diseases. The list of diagnostic criteria identified for our set of six diseases generally follows consensus statements on mitochondrial disease diagnosis and treatment provided by Parikh *et al.*[28,29] and the Mitochondrial Medicine Society (http://mitosoc.org/news/). Because these recommendations involve conditions that are frequently more specific than those described by ICD-10, we use this list primarily as a source of high-level types of symptoms. For example, acidosis (ICD-10-CM E87.2) is routinely described in each of the selected diseases, with the exception of Complex III deficiency. This symptom appears in 69 out of 246 reports, often as a specific form; some forms are represented by a unique ICD-10 code, such as 3-methylglutaconic aciduria (E71.111), while others, including lactic acidosis, rely on the general code for identification. Among the 5 mitochondrial respiratory chain complex deficiencies, none have a specific diagnostic code within ICD-10; we collectively annotate them with two codes for mitochondrial metabolism disorders (E88.40 and E88.49). Complex II and Complex V deficiencies are described in 12 and 7 reports in this subset, respectively, and are associated with novel hypertrophic cardiomyopathies (I42.2; seen in 7 reports across both disorders) and a lack of development in childhood (R62.50; seen in 7 reports across both). Complex III deficiency is associated with hypoglycemia (E16.2) in 17 out of 29 cases but is otherwise notable for its lack of distinguishing characteristics in this set. Generalized muscle weakness (M62.81), while common across all RMDs, is especially frequently described within reports on primary carnitine deficiency (18 out of 45, compared to 32 out of all 201 other reports). We also compare our code assignments to Barth syndrome diagnosis criteria; specifically, that the disease presentation frequently involves neutropenia and 3-methylglutaconic aciduria[30]. Out of 30 Barth syndrome reports, 26 describe neutropenia (D70.9), 24 describe dilated cardiomyopathy (I42.0), 16 describe cardiomegaly (I51.7), 11 describe acidosis (E87.2), and 8 describe 3-methylglutaconic aciduria (E71.111).

Chronological features

The reports corresponding to metadata in the MACCR set were originally published between 1956 and 2018, with 455 CCRs published from 2017 to 2018. Out of all 3,100 reports contributing metadata to the MACCR set, 2,112 (68.1%) were published since 2007. Among 1.89 million CCRs indexed through PubMed/MEDLINE with publication dates from 1936 to August 2018, more than 595,000 (31.5%) were published since 2007. This skew toward more recent years in the MACCR set, relative to the distribution of all published CCRs, is partially a conscious choice made during curation to focus on cases with well-defined diagnoses and general similarity in terminology. This difference also reflects the greater accessibility of clinical case reports within the last decade.[1] Two prominent sources are *British Medical Journal Case Reports* (*BMJ Case Rep*) and *Journal of Medical Case Reports* (*J Med Case Rep*), which began publishing in 2008 and 2007, respectively. These two journals alone have published more than 20,000 CCRs and contribute 166 reports (5.3%) to the MACCR set. Publications with longer histories, such as the *New England Journal of Medicine* (*NEJM*), continue to publish CCRs as well. Though their editorial stance on CCRs has changed over time[31], *NEJM* has published nearly 10,000 case reports since 1949 and 2,224 since 2007, of which 101 and 75 are included in the MACCR set, respectively.

Demographic and geographic distribution of MACCRs

The distribution of demographic features among patients described in our MACCR dataset shows it is not excessively skewed toward one subset of the patient population. Patient age presents a generally consistent distribution: 890 (28.7%) reports describe clinical presentations with pediatric patients (i.e., less than 21 years old), 1,104 (35.6%) reports describe patients of at least 21 and no more than 50 years old, and 1,051 (33.9%) reports describe patients of 51 years old or older. Just 55 (1.7%) reports did not provide enough information to know or infer patient age. Some reports may span years or decades of symptoms or treatment, so age remains a rough demographic value (in this case, age refers to patient age at the beginning of a given clinical narrative). Even so, we believe this is evidence that the MACCR set is not excessively biased in favor of a single age group.

Patient sex provides additional evidence of demographic and conceptual variety among CCRs contributing to the MACCR set. Over the 3,100 CCRs, 1,415 (45.6%) reports concern female patients, 1,536 (49.5%) concern male patients, and the remainder (149; 4.8%) did not specify a patient's sex or were unclear. One potential source of imbalance between male and female patients is the presence of

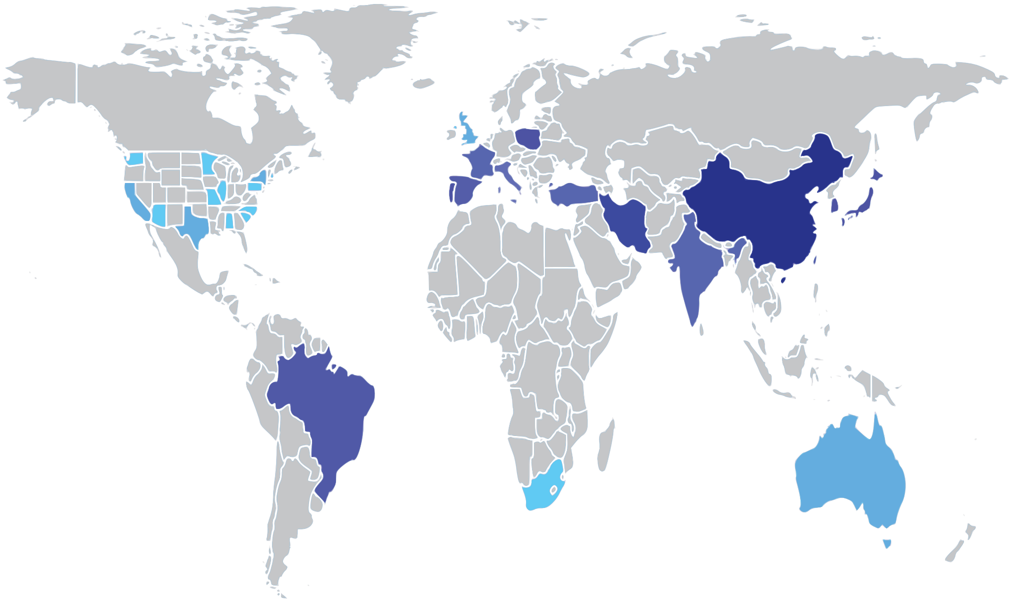


Figure 3. Geographic regions with difference in publication frequency. Here, we visualize the distribution of locations of reports in the MACCR set across the world and within US states. Darker shades correspond to a greater significant difference between the count of CCRs in the MACCR set for a particular region (determined by institutional affiliation of the corresponding author) and the count of all CCRs for that region (as specified by MEDLINE index).

numerous cardiovascular disease (CVD) CCRs in our set: male patients with CVD symptoms are more likely to receive corresponding diagnostics and treatment *vs.* female patients[32,33]. Similarly, we find the assumption that all obstetrical and gynecological cases should involve female patients is generally true: out of 176 cases in this category, 160 identified female patients specifically; 14 reports did not explicitly state the patient's gender, though it was implicitly inferred as female; and of the remaining 2 cases involving male patients, both describe conditions impacting male fetuses. Across all female patients, 360 are pediatric (<21), 551 were at least 21 and less than 51 years old, and 480 were 51 or older. Of the male patients, the same counts are 470, 513, and 530, respectively.

This demographic variety extends to the geographic origin of the MACCRs: Fig. 3 provides a global view of differences between MACCR geographic origins and those of all CCRs. Across all CCRs published as of August 2018, more than 700,000 originate from the United States. Additional visualizations of the geographic distribution, including its breakdown by disease category, are shown in a supplementary animation (MACCR Supplementary File 1, Data Citation 1). Their distribution in this set is very similar to that seen across all MACCRs, even when accounting for individual states, though we note that location indexing is rarely comprehensive and provides only a rough estimate. Most reports in the MACCR set have corresponding authors from the US, Japan, and China, likely due to curation of reports from more recent years, and hence more publications from nations with higher citation rates in recent years[34]. As a result, China and Japan have a slightly greater representation in the MACCR set than in CCRs as a whole. It is important to note that these geographic identities relate to the corresponding author, which does not necessarily equate to the treatment location or origin of the patient. Patient geo-location data is not available from CCRs, though these data would be highly informative. Still, knowledge of the lead investigator's geo-location provides crucial insight as to where important clinical work is occurring.

Quality and value of case report metadata

The value of the MACCR dataset is derived from features created through metadata extraction. We recognize this is an *added* value, as the CCRs have existing metadata and subject headings provided by MEDLINE and MeSH (e.g., the content of each title, author list, and other bibliographic information). Unlike MeSH assignments available through PubMed, we have examined the full clinical narrative communicated within each case report and therefore furnish more detail than that permitted by the short list of MeSH descriptors associated with each document. Our inclusion of descriptions of clinical events and activities as they are described in the case reports themselves covers terminology not present in

MeSH, particularly for the names of drugs and diagnostic or therapeutic procedures. Our metadata extraction approach also differs from a code-centric approach and from NER-driven text mining approaches in that it relies upon assignment of document text to general clinical concepts rather than to stringently-defined concepts or entities. Neither MeSH nor ICD-10-CM have been designed with clinical narratives in mind: the former is intended to index biomedical documents while the latter translates clinical diagnoses into coded endpoints. Our metadata extraction approach fills the niche these methods were not intended to address.

As part of our metadata extraction process, we used a scoring system to quantify the total number of features manually identified across all MACCRs. A CCR with a full set of values corresponding to clinical concepts (i.e., each of the concepts has at least one associated text value) is assigned a medical content score of 18, the highest value. One point each is provided for presence of key words and for enough detail to determine a disease category, while the remaining points reflect types of medical content. The minimum medical content score is 1. The average medical content score across all metadata sets is 10.9 with a standard deviation of +/−3.52, indicating each set of metadata provides more than 10 new details on average, relative to available metadata, or more than 33,000 new details across all metadata records in the set. More than 7,400 details are from CCRs discussing rare disease presentations. It is relevant to note that the presence of text values corresponding to each medical concept is not consistent across all concepts and across the full MACCR set; we believe this indicates material was simply not included in the CCRs rather than omitted during the metadata extraction process. For example, 2,980 (96%) of the MACCRs contain material describing diagnostic techniques and procedures, 2,349 (75.7%) contain material describing clinical outcomes, and just 186 (6%) provide descriptions of patient social history.

We may also consider the metadata value in terms of information entropy. Determining entropy on a per-character basis (i.e., Shannon entropy) allows us to calculate an average entropy per concept, with all "NA" values treated as a value with an entropy of 0 bits (as these values provide no additional information). We use this metric in lieu of estimates of readability (e.g., Flesch-Kincaid[35] or SMOG[36] scores) as these metrics are heavily biased by the frequency of complex vocabulary common to medical language. We intend the entropy values to serve as estimates of differences in information content between fields in the main MACCR dataset. Table 4 contains these average entropy values, along with character, word, and segment (i.e., each semicolon-delimited phrase) counts for each medical concept. This approach essentially combines two metrics (i.e., average entropy for a measurement of overall information content and treatment of "NA" values to adjust for missing values) such that the resulting entropy values denote estimates of each concept's overall semantic complexity relative to others. Concepts with average entropy close to 4, such as Diagnostic Techniques and Procedures, not only contain more values than other concepts but each value is more complex, primarily as a function of length.

Usage Notes

We anticipate the MACCR set will aid clinicians and clinical researchers in gaining a better understanding of disease presentations, including their key symptomology, diagnostic approaches, and treatment. Researchers in bioinformatics, clinical informatics, and information extraction will find the MACCRs useful as a set of medical language labeled at multiple levels. Individual researchers faced with small sample sizes may use the MACCRs to enhance their statistical power through incorporation of additional observations, or as a starting point for the assembly of *in silico* patient cohorts. We envision a researcher could generate MACCRs of their own using our metadata standard template to assemble these cohorts and leverage the cumulative power for statistical analysis. The resulting metadata enables deep text mining alongside MeSH, ICD-10, and other clinical ontologies. To establish the utility of the dataset and guide those interested in employing it in their research, we present use cases that may be pursued by researchers, physician investigators, clinicians, data scientists, IP officers, pharmaceutical companies for drug development, and those shaping government policies for clinical trials. This set of metadata in medical language yields a rich resource for providing insight into the events and biological phenomena within each clinical presentation.

Analysis of case report features

The immediately applicable uses of the MACCR set are those involving aggregation and analysis of features particular to each CCR. As the reports contributing metadata to this set reflect a substantial variety of disease presentations and features, our structured data provide a multitude of options for subsequent analysis. In the simplest case, extraction of sets of terms associated with a particular disease or disease category establishes a set of starter terms for further study. For example, extraction of all diagnostic procedures used in respiratory disease cases in the MACCR set (e.g., chest x-ray, lung function test, or CT scan) allows researchers to better direct future literature searches by including a set of commonly used treatments. These terms, while not comprehensive for any particular topic, form a representative set of term *vs.* concept associations and include nonspecific terms as well (e.g., CT scans are not specific to diagnostics for respiratory tract disease).

The rich term vocabulary available within the MACCR set permits more in-depth analysis and application to additional documents. Researchers may find this metadata particularly helpful in studying differences in treatment approaches across disease type or on the basis of the impacted organ system. We

Concept	Average Entropy (bits, +/−standard deviation)	Character Count	Word Count	Segment Count
Keywords	2.17 +/− 2.04	127,932	8,326	6,636
Geographic Locations	0.35 +/− 1.01	6,085	901	358
Life Style	0.55 +/− 1.35	29,244	4,862	521
Family History	1.15 +/− 1.83	138,162	21,342	1,717
Social History	0.23 +/− 0.90	12,310	2,022	249
Medical/Surgical History	3.02 +/− 1.84	804,975	119,816	8,783
Signs and Symptoms	3.96 +/− 0.94	1,460,450	218,276	16,467
Comorbidities	0.96 +/− 1.63	33,978	3,918	1,329
Diagnostic Techniques and Procedures	3.98 + / − 0.87	1,369,668	195,000	15,936
Diagnosis	3.85 +/− 0.66	206,418	24,432	4,718
Laboratory Values	2.80 +/− 2.12	990,769	146,240	5,238
Pathology	2.32 +/− 2.11	853,084	121,009	2,865
Pharmacological Therapy	2.74 +/− 1.99	422,402	60,270	3,863
Interventional Therapy	2.60 +/− 1.94	399,831	57,967	4,909
Patient Outcome Assessment	3.07 +/− 1.77	440,602	66,786	4,526

Table 4. **Text properties and entropy of medical concept metadata records.** For each medical concept used in the metadata extraction process, we determined its average character-level entropy (Shannon entropy) across all text values in the concept, along with its standard deviation. As length of text can contribute to estimates of its complexity, we also include counts of characters (not including delimiters or spaces), words, and segments (i.e., phrases between delimiters) for each concept across the MACCR set. Values of "NA" are considered to have an entropy of zero and do not contribute to character, word, or segment counts.

suggest that an initial analysis of this set be managed through extraction of a set of terms in a comprehensive dictionary, such as RxNorm[37]. Starting with observations listed in the Drug Therapy column in each metadata record, for example, rule-based and NER methods can identify compound names of interest. Enrichment of any name or group of names among a given subset of the metadata will reveal broader phenomena, e.g., antibiotics may correlate with infectious disease cases.

Because the MACCR set includes English medical language from a variety of locales, it is a representative set of medical text not specific to a single region, organization, or population. The dataset includes structured demographic features (including age, sex, and geo-location) to serve as easily-parsed features for correlative analyses. A more in-depth search enables geography-based analyses: by combining both document metadata and free-text from metadata categories (specifically, patient demographics or other features describing the patient), these fields can be mined for names of major cities, states, and countries. They may then be mapped and quantified to visualize the case report distribution (as demonstrated in Fig. 3). Methodologies developed using this dataset are appropriate for multi-level geographic term identification (i.e., from specific to general location) with a larger set of clinical reports, especially as researchers may find that certain features of CCRs are more informative for geographic location than others. Geographic trends revealed using the MACCR set may reveal broader phenomena to be followed up in new studies. They may also provide evidence of an unexpected focus in regional publication for specific diseases. We may expect CCRs involving a regional epidemic of a specific infectious disease to be predominantly written by clinicians in that area. Alternatively, these cases may also be popular among clinicians describing the spread of infectious disease to new locations.

Incorporating CCR analysis into broader studies permits exploration of undefined or poorly defined diagnoses. Some rare diseases may only exist, conceptually, as subsets of more common disease presentations. Heart failure, for example, is such a common condition that it may be responsible for more hospitalizations than any other condition, yet half of heart failure patients may suffer from a particular subtype of the syndrome, heart failure with preserved ejection fraction[38], or HFpEF. Despite its current prevalence, HFpEF was only recognized as a distinct condition within the last several decades. The first observations of disease presentations lacking specific diagnostic consensus may therefore be contained only in CCRs and are unlikely to appear in formal epidemiological studies.

Support for mitochondrial and rare disease characterization

The MACCR dataset documented here contains a rare disease subset, including a set of rare mitochondrial diseases. There are over 7,000 rare diseases affecting over 300 million individuals worldwide, 58 of which are RMDs (https://rarediseases.info.nih.gov/diseases). We have generated metadata acquired from CCRs describing 7 mitochondrial diseases for this dataset, as well as an additional set of ICD-10-CM diagnostic and symptom codes. The RMDs curated for this dataset have a limited number of publications in the medical literature, so our metadata collectively represent a substantial portion of the published clinical observations of these diseases. Each of the RMDs shares similar features in that they all involve mitochondrial abnormalities, yet each produces markedly different phenotypes and clinical signs. A number of analysis routes are available with the current dataset. We suggest that researchers use this set as a model to identify additional cases in the literature not explicitly identified as mitochondrial diseases. These implicit cases may be more common than anticipated but may be predicted based on presentations sharing numerous signs and symptoms with known cases. The resulting predicted cases would comprise a model for future studies of rare and/or idiopathic disease. For RMDs in particular, the MACCR set establishes a basis for both expected and frequently correlated symptomology, facilitated by the ICD-10-CM on RMD CCR symptomology.

Additional data on the genetic and molecular basis of the selected RMDs provides a more specific diagnostic picture and allows researchers to interface with knowledgebases to conduct proteome and pathway analysis. This information can offer mechanistic insight for the potential pathogenesis of disease. The etiology of many RMDs lies in mutations to mitochondrial DNA (mtDNA), but the majority are caused by mutations to nuclear-encoded genes that play various critical roles in mitochondrial biosynthesis and function (Fig. 2c). Barth syndrome, for instance, is caused by mutations to the *TAZ* (*G4.5*) gene at Xq28, resulting in malformed mitochondrial membranes due to a nonfunctional tafazzin enzyme that is responsible for adding linoleic acid to cardiolipin (CL) through its acyltransferase activity. Because of its prominence in the mitochondrial inner membrane and its intimate association with the electron transport chain complexes, improperly formed CL severely limits mitochondrial energy production and results in a variety of complications, including dilated cardiomyopathy, hypertrophic dilated cardiomyopathy, left-ventricular non-compaction, and endocardial fibroelastosis. Barth syndrome was first described in 1983[30,39] and over 100 distinct mutations to the *TAZ* gene have been identified since the genetic basis was discovered in 1996[40]. Recognizing the common feature of diminished CL and elevated concentrations of monolysocardiolipin (MLCL) in Barth syndrome, an HPLC-MS/MS bloodspot assay was developed for diagnosis by an elevated MLCL:CL ratio[41]. The MACCR set can serve as a valuable resource for uncovering common proteins, pathways, and metabolites of interest in RMDs, which may lead to the identification of potential biomarkers. The depth of information can be further amplified by integrating this dataset with other publicly available resources and knowledgebases such as *UniProt*[42] (http://www.uniprot.org), *Reactome*[43,44] (https://reactome.org), and the Human Metabolome Database (HMDB)[45] (http://www.hmdb.ca).

This dataset will aid users in gaining a better understanding of rare diseases and their treatments through downstream analysis of the structured text data. The CCR metadata derived from reports on Barth syndrome, for example, are directly relevant to developing and evaluating treatments for this and related conditions, as well as diagnostic and treatment planning in the clinic. By reviewing existing CCR metadata, a researcher could investigate past treatment regimens to analyze recorded symptoms or side effects and the degree of improvement under prevailing standards of care. Similarly, clinicians could employ metadata extracted from CCRs to support differential diagnosis and treatment planning in a patient with a suspected mitochondrial disorder. Furthermore, we envision that researchers and clinicians might utilize the metadata standard template and protocols to generate additional MACCRs and construct *in silico* patient cohorts of their own, enhancing their ability to compare existing treatments and evaluate newly developed therapies. For example, Elamipretide (Stealth Biotherapeutics, Newton, Massachusetts) is currently under stage II and III clinical trials for treatment of Barth syndrome and other mitochondrial myopathies by protecting properly formed cardiolipin from damage[46]. As new case reports on patients treated with Elamipretide become available, additional MACCRs might be leveraged to compare treatments, perhaps identifying side effects or varying efficacy of the new drug in different patient groups.

Support for congenital heart disease characterization

Congenital heart defects (CHD) and disease are unfortunately common clinical issues occurring in an estimated 1% of births in the United States[47]. Similarly, they are frequently represented among the reports in the MACCR dataset, 243 of which involve CHD. Though general descriptions of CHD diagnoses (e.g., “ventricular septal defect” or “hypoplastic left heart syndrome”) lend themselves well to indexing by MeSH, more nuanced narratives surrounding CHD presentations require more detailed examination. As a brief example, among the CHD reports within the MACCR set, the top MeSH descriptors include “Infant, Newborn”, “Echocardiography”, and “Abnormalities, Multiple”. A comparison of the metadata for these reports reveals some variation in age: of the 240 CCRs for which age is specified or can be inferred, mean age is about 22.7 years, with 76 cases involving patients less than a year old and 137 involving pediatric patients (any under 21 years of age). In terms of additional detail, however, just 49 (20.2%) of these CCRs specifically mention heart failure, but nearly as many (47) mention cyanosis or cyanotic conditions. Though a larger collection of CHD reports may be required, metadata extraction and investigation of these reports may provide solid evidence for new biomarker candidates in an area where few currently exist.

Models of medical language

The MACCR set contains rich, contextual descriptions of medical events. Individual words and phrases in the set are not explicitly assigned to a given ontology or vocabulary but are included within our medical concept categories. For example, instead of indicating a document describes a "myocardial infarction" and/or identifying this phrase in each document, if a document mentions events such as a heart attack, we assign the event to the appropriate medical concept (and assign a disease category; in this case, cardiovascular disease). The corresponding text segment in each case includes phrases such as "family history was positive for myocardial infarction in a sibling at age 54 years"[48]. As compared with an approach of processing unstructured case report text with NLP tools, our resource supplies an intermediate level of structure sufficient to retain the context of the segment. This semantic context contains information denoting relations between concepts and events with enough detail to assign additional diagnostic categories and codes. In the example of a myocardial infarction, the metadata often includes the details necessary to determine subtypes (e.g., ST-elevation *vs.* non-ST elevation, or ICD-10-CM codes of I21.3 *vs.* I21.4). Employing NER alone on the source text may yield only phrases such as "myocardial infarction", and MeSH descriptors generally do not index documents to this level of detail. The MACCR set labels text segments with medical concepts, allowing collections of phrases, rather than named entities, to be associated with higher-level concepts. Our resource thereby enables an additional degree of interoperability between CCRs, controlled vocabularies (e.g., MeSH), and diagnostic coding systems (e.g., ICD-10), while supplying a rich collection of contextual and concept-labeled clinical text features.

The contents of the MACCR set provide the structured training data necessary for developing computational models of higher-level features in clinical text. Computational linguists as well as researchers in clinical informatics and medical NLP may use the MACCR records to develop concept-level models of medical language, allowing for context-based machine learning and alternatives to dictionary-driven NLP approaches[49,50]. The MACCR approach supports generation of term frequency sets, word vectors, and basic entity-level analysis (e.g., with UMLS resources[25]) while retaining clinical concepts such as medical history or demographics. NER systems such as cTAKES[49] (http://ctakes.apache.org/) or CLAMP[50] (http://clamp.uth.edu/) support identification of procedures and signs/symptoms using the features within MACCR records. Additionally, NER and rule-based phrase matching approaches draw connections between MACCR content and biomedical knowledgebases (i.e., *UniProt*[42] [http://www.uniprot.org], *Reactome*[43,44] [https://reactome.org], or the Disease Ontology[51] [http://disease-ontology.org/]). The combination of extensive knowledgebases with advanced computational models supports transformation of clinical observations into biomedical insight.

Our dataset facilitates expansion of ongoing developments in NLP principles and methods to medical documents, especially alongside the substantial extant resources for contextualization and distant supervision of computational approaches to understanding medical language. The broad demand from the community and far-reaching significance of NLP approaches has been a springboard for novel approaches in biomedical research and beyond, evidenced by rapid development of tools and resources in a variety of research fields[52–55]. Beyond clinical informatics research, tools developed to enforce structure on otherwise unstructured biomedical text – including that in electronic health records – offer a major source of untapped biomedical knowledge[56,57]. These tools will require significant software development and engineering efforts, yet once the resulting knowledge becomes structured and searchable, it will be of greater interest and utility to clinicians, data scientists, and physician investigators, as well as intellectual property specialists and officials determining policies for clinical trials.

Manual curation is currently the most suitable option for capturing comprehensive details associated with high-level concepts in biomedical literature. Though it may someday yield similar results, automated, machine learning-driven medical language analysis presents distinct limitations in precision and recall, producing numerous false positive and false negative results as compared to human annotators[58]. With this dataset, we present a resource appropriate for training new machine learning models on the types of language common to clinical case reports: vocabulary, common phrases, and association with high-level medical concepts. The resulting models may then support further human curation and metadata extraction, assembly of more fine-grained knowledge structures (e.g., knowledge graphs), and transfer learning to train more complex medical language models. Our dataset is therefore entirely complementary to biomedical text resources such CRAFT[15] and those available through i2b2[13,14]. Though these datasets provide records of specific concepts and features, the MACCR set furnishes rich metadata of clinical concepts across a wide variety of disease types. In instances where machine learning methods may require considerably a larger amount of text for training, we suggest using the MACCR set as an initial training step and in combination with other text resources.

An educational resource for writing better case reports

The MACCR set contains metadata for reports spanning disease types and medical specialties, highlighting a wide variety of CCR writing styles and a range of completeness in describing relevant clinical concepts. In some cases, variance among MACCRs is the result of a lack of explicitly stated observations: e.g. a patient's exact age or family history may be omitted. Similarly, clinicians may not mention tests if the diagnostics or their results were considered trivial. The richness of our dataset offers a basis for comparison among cases. Clinical investigators may observe the extent to which expected clinical concepts are or are not discussed in case reports. This analysis may be particularly informative if otherwise similar cases ares found to differ in diagnosis. Other features may be more useful for subsequent analyses if provided in a more specific, quantifiable manner; a CCR with a patient described as "55 years old", for instance, will be more informative than one with the description "middle-aged." Further analysis of the specific features within CCRs will provide clear examples of how clinicians may write more informative, citable, and computationally readable CCRs.

While, at present, case reports are primarily read by academic physicians for educational purposes, implementation of the standardized metadata template to enrich these documents can expand the audience and application of case reports. For example, case report user groups may include medical students, interns and fellows, epidemiologists, and statisticians. These audiences would not only be able to more easily identify relevant CCRs, but also derive valuable information from improved indexing and categorization. In turn, these improvements will lead to better understanding of clinical phenotypes and relationships of an individual case to a larger representative patient population. As another example, healthcare organizations and policymakers (e.g., FDA) can retrieve CCR metadata as an additional source for tracking unusual disease occurrences, epidemiological trends, and post-marketing drug surveillance. Moreover, pharmaceutical industries can design a survey on case reports of drugs with unexpected indications or unrecognized side-effects to assist in modifying usage instructions and direct future development.

To address the key clinical items commonly missing in case reports, we envision a solution that integrates what PubMed has already accomplished with MeSH terms using both metadata extraction and coding with ICD-10-CM. This strategy would support further classification with systems such as the International Classification of Health Interventions (ICHI) (http://www.who.int/classifications/ichi/en/) to compensate for missing items. The resulting curated, indexed, and structured CCR metadata could ultimately interface with preclinical -omics research, clinical cohort studies, and clinical trials to advance understanding of disease progression, management, and clinical outcomes. To surmount the ever-growing amount of free-text information with limited metadata, indexing, and accessibility, computational platforms and in-depth search algorithms will enable better recognition of CCR contents and relevant clinical trials to elevate text data analysis, advance medical science, and improve patient care.

As a time-honored tradition in medical publication and a treasured source of clinical data, clinical case reports augment our understanding of disease etiology, pathogenesis, miscellaneous diagnosis, and therapeutic efficacy. These reports provide valuable clinical narratives relevant to clinicians and biomedical researchers. The growing volume of case reports published each year stands testament to their popularity and usefulness to their targeted clinical readership, but this size, coupled with the isolated, unstructured, and heterogeneous nature of case reports' contents, also presents a challenge to index, annotate, and query case report data. In this report, we created a standardized metadata template and metrics, as well as a test dataset consisting of 3,100 CCRs spanning 16 disease categories. In the course of assembling our dataset, we evaluated the caliber of the existing metadata employed for case reports in PubMed and confirmed a discrepancy between the medical content and the metadata meant to describe it. Our MACCR set addresses this discrepancy by adding rich metadata and serves as a valuable resource for biomedical researchers developing novel approaches to advance medical science and improve patient care[20].

References

1. Akers, K. G. New journals for publishing medical case reports. *J. Med. Libr. Assoc.* **104,** 146–149 (2016).
2. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3,** 160018 (2016).
3. Cabán-Martinez, A. J. & García-Beltrán, W. F. Advancing medicine one research note at a time: the educational value in clinical case reports. *BMC Res. Notes* **5,** 293 (2012).
4. Nissen, T. & Wynn, R. The recent history of the clinical case report: a narrative review. *JRSM Short Rep* **3,** 1–5 (2012).
5. Vandenbroucke, J. P. In defense of case reports and case series. *Ann. Intern. Med.* **134,** 330 (2001).
6. Pasteur, L. Méthode pour prévenir la rage après morsure. *Comptes rendus l'Académie des Sci* **101,** 765–774 (1885).
7. Pearce, J. Louis Pasteur and Rabies: a brief note. *J. Neurol. Neurosurg. Psychiatry* **73,** 82–82 (2002).
8. Keefer, C. S., Blake, F. G., Marshall, E. K. J., Lockwood, J. S. & Wood, W. B. J. Penicillin in the treatment of infections. *J. Am. Med. Assoc* **122,** 1217 (1943).
9. Poiesz, B. J. *et al.* Detection and isolation of type C retrovirus particles from fresh and cultured lymphocytes of a patient with cutaneous T-cell lymphoma. *Proc. Natl. Acad. Sci. U. S. A* **77,** 7415–9 (1980).
10. Nissen, T. & Wynn, R. The clinical case report: a review of its merits and limitations. *BMC Res. Notes* **7,** 264 (2014).
11. Mork, J. G., Jimeno-Yepes, A. & Aronson, A. R. The NLM Medical Text Indexer System for indexing biomedical literature. in *BioASQ Workshop* (2013).
12. Liu, K. *et al.* MeSHLabeler: improving the accuracy of large-scale MeSH indexing by integrating diverse evidence. *Bioinformatics* **31,** i339–i347 (2015).
13. Sun, W., Rumshisky, A. & Uzuner, O. Evaluating temporal relations in clinical text: 2012 i2b2 Challenge. *J. Am. Med. Informatics Assoc* **20,** 806–813 (2013).
14. Stubbs, A., Kotfila, C., Xu, H. & Uzuner, Ö. Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2. *J. Biomed. Inform.* **58**(Suppl,): S67–77 (2015).
15. Bada, M. *et al.* Concept annotation in the CRAFT corpus. *BMC Bioinformatics* **13,** 161 (2012).
16. Pyysalo, S. & Ananiadou, S. Anatomical entity mention recognition at literature scale. *Bioinformatics* **30,** 868–875 (2014).
17. Doğan, R. I., Leaman, R. & Lu, Z. NCBI disease corpus: A resource for disease name recognition and concept normalization. *J. Biomed. Inform.* **47,** 1–10 (2014).
18. Kim, S., Yeganova, L., Comeau, D. C., Wilbur, W. J. & Lu, Z. PubMed Phrases, an open set of coherent phrases for searching biomedical literature. *Sci. Data* **5,** 180104 (2018).

19. Demner-Fushman, D. *et al.* A dataset of 200 structured product labels annotated for adverse drug reactions. *Sci. Data* **5,** 180001 (2018).
20. World Health Organization. International classification of diseases and related health problems, 10th revision. (1992).
21. McDonald, C. J. LOINC, a universal standard for identifying laboratory observations: a 5-year update. *Clin. Chem.* **49,** 624–633 (2003).
22. Kveler, K. *et al.* Immune-centric network of cytokines and cells in disease context identified by computational mining of PubMed. *Nat. Biotechnol.* **36,** 651–659 (2018).
23. Grishman, R. & Sundheim, B. Message Understanding Conference-6. In *Proceedings of the 16th conference on Computational linguistics* **1,** 466, Association for Computational Linguistics (1996).
24. Tjong Kim Sang, E. F. & De Meulder, F. Introduction to the CoNLL-2003 shared task. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003* **4,** 142–147, Association for Computational Linguistics (2003).
25. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* **32,** D267–70 (2004).
26. Cowper, S. E. *et al.* Scleromyxoedema-like cutaneous diseases in renal-dialysis patients. *Lancet (London, England)* **356,** 1000–1 (2000).
27. Grobner, T. Gadolinium--a specific trigger for the development of nephrogenic fibrosing dermopathy and nephrogenic systemic fibrosis? *Nephrol. Dial. Transplant* **21,** 1104–8 (2006).
28. Parikh, S. *et al.* Diagnosis and management of mitochondrial disease: a consensus statement from the Mitochondrial Medicine Society. *Genet. Med.* **17,** 689–701 (2015).
29. Parikh, S. *et al.* Patient care standards for primary mitochondrial disease: a consensus statement from the Mitochondrial Medicine Society. *Genet. Med.* **19,** 1380 (2017).
30. Barth, P. G. *et al.* X-linked cardioskeletal myopathy and neutropenia (Barth syndrome): An update. *Am. J. Med. Genet.* **126A,** 349–354 (2004).
31. The Case Report. *N. Engl. J. Med.* **277,** 827–827 (1967).
32. Mosca, L., Barrett-Connor, E. & Kass Wenger, N. Sex/gender differences in cardiovascular disease prevention: what a difference a decade makes. *Circulation* **124,** 2145–2154 (2011).
33. Wenger, N. K. Gender disparity in cardiovascular disease: bias or biology? *Expert Rev. Cardiovasc. Ther.* **10,** 1401–11 (2012).
34. Smith, M. J., Weinberger, C., Bruna, E. M. & Allesina, S. The scientific impact of nations: journal placement and citation performance. *PLoS One* **9,** e109195 (2014).
35. Kincaid, J. P., Braby, R. & Mears, J. E. Electronic authoring and delivery of technical information. *J. Instr. Dev* **11,** 8–13 (1988).
36. McLaughlin, G. H. SMOG grading-a new readability formula. *J. Read* **12,** 639–646 (1969).
37. Bennett, C. C. Utilizing RxNorm to support practical computing applications: capturing medication history in live electronic health records. *J. Biomed. Inform.* **45,** 634–41 (2012).
38. Andersen, M. J. & Borlaug, B. A. Heart failure with preserved ejection fraction: current understandings and challenges. *Curr. Cardiol. Rep.* **16,** 501 (2014).
39. Barth, P. G. *et al.* An X-linked mitochondrial disease affecting cardiac muscle, skeletal muscle and neutrophil leucocytes. *J. Neurol. Sci.* **62,** 327–55 (1983).
40. Bione, S. *et al.* A novel X-linked gene, G4.5. is responsible for Barth syndrome. *Nat. Genet.* **12,** 385–9 (1996).
41. Kulik, W. *et al.* Bloodspot assay using HPLC-tandem mass spectrometry for detection of Barth syndrome. *Clin. Chem.* **54,** 371–8 (2008).
42. The UniProt Consortium. UniProt: the universal protein knowledgebase. *Nucleic Acids Res* **45,** D158–D169 (2017).
43. Fabregat, A. *et al.* The Reactome Pathway Knowledgebase. *Nucleic Acids Res* **46,** D649–D655 (2018).
44. Milacic, M *et al.* Annotating cancer variants and anti-cancer therapeutics in reactome. *Cancers (Basel)* **4,** 1180–211 (2012).
45. Wishart, D. S. *et al.* HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res* **46,** D608–D617 (2018).
46. Birk, A. V *et al.* The mitochondrial-targeted compound SS-31 re-energizes ischemic mitochondria by interacting with cardiolipin. *J. Am. Soc. Nephrol.* **24,** 1250–61 (2013).
47. Hoffman, J. I. . & Kaplan, S. The incidence of congenital heart disease. *J. Am. Coll. Cardiol.* **39,** 1890–1900 (2002).
48. Kazmi, A. S. & Wall, B. M. Reversible congestive heart failure related to profound hypocalcemia secondary to hypoparathyroidism. *Am. J. Med. Sci.* **333,** 226–229 (2007).
49. Savova, G. K. *et al.* Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J. Am. Med. Informatics Assoc.* **17,** 507–513 (2010).
50. Soysal, E. *et al.* CLAMP – a toolkit for efficiently building customized clinical natural language processing pipelines. *J. Am. Med. Informatics Assoc* **25,** 331–336 (2018).
51. Kibbe, W. A. *et al.* Disease Ontology 2015 update: an expanded and updated database of human diseases for linking biomedical knowledge through disease data. *Nucleic Acids Res* **43,** D1071–8 (2015).
52. Fernandes, A. C. *et al.* Identifying suicide ideation and suicidal attempts in a psychiatric clinical research database using natural language processing. *Sci. Rep* **8,** 7426 (2018).
53. Volanakis, A. & Krawczyk, K. SciRide Finder: a citation-based paradigm in biomedical literature search. *Sci. Rep* **8,** 6193 (2018).
54. Court, C. J. & Cole, J. M. Auto-generated materials database of Curie and Néel temperatures via semi-supervised relationship extraction. *Sci. Data* **5,** 180111 (2018).
55. Mandloi, S. & Chakrabarti, S. PALM-IST: pathway assembly from literature mining - an information search tool. *Sci. Rep* **5,** 10021 (2015).
56. Maddox, T. M. & Matheny, M. A. Natural language processing and the promise of Big Data. *Circ. Cardiovasc. Qual. Outcomes* **8,** 463–465 (2015).
57. Pivovarov, R. & Elhadad, N. Automated methods for the summarization of electronic health records. *J. Am. Med. Informatics Assoc* **22,** 938–947 (2015).
58. Ching, T. *et al.* Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* **15,** 20170387 (2018).
59. Lex, A., Gehlenborg, N., Strobelt, H., Vuillemot, R. & Pfister, H. UpSet: visualization of intersecting sets. *IEEE Trans. Vis. Comput. Graph.* **20,** 1983–92 (2014).

Data Citations

1. Caufield, J.H. *et al. figshare* https://doi.org/10.6084/m9.figshare.c.4220324 (2018).
2. Caufield, J.H. *et al. Dryad Digital Repository* https://doi.org/10.5061/dryad.r36cn90 (2018).

Acknowledgements

This work was supported in part by National Heart, Lung, and Blood Institute: R35 HL135772 (to P. Ping); National Institute of General Medical Sciences: U54 GM114833 (to P. Ping, K. Watson, W. Wang and A. Bui), the T.C. Laubisch endowment at UCLA (to P. Ping); and the China Scholarship Council

201606325015 (to Y. Zhou). The authors would like to thank K. Komya, K. Rastogi, A. Tsai, and all our CCR metadata extraction volunteers for their assistance in the assembly of the MACCR set and this manuscript.

Author Contributions

J.H.C., Y.Z., and A.O.G. led the study and contributed to: study design, generation of the data, data validation, writing, and editing of the manuscript. J.H.C. also contributed to software development. D.A.L. and Q.C. contributed to study design and generation of the data. J.M.L. and S.M. contributed to generation of the data. S.S. contributed to generation of the data, software development, and data validation. W.W., L.Z., H.H., and K.E.W. contributed to study design. P.P. conceived of the study and contributed to study design and editing of the manuscript.

Additional Information

Competing interests: The authors declare no competing interests.

How to cite this article: Caufield, J. H. *et al.* A reference set of curated biomedical data and metadata from clinical case reports. *Sci. Data.* 5:180258 doi: 10.1038/sdata.2018.258 (2018).

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

Chapter 5. A FAIR Representation of Mitochondrial Biology and Rare Mitochondrial Disease via GeneWiki and MitoCases

Introduction

Mitochondria are responsible for a wide array of cellular processes including fatty acid and glycolytic metabolism, oxygen handling through oxidative phosphorylation, energy production through the electron transport chain, as well as inter- and intra-cellular signaling governing protective, apoptotic, and proliferative pathways. With its central role in these essential cellular functions, mitochondrial biology is integral to our fundamental understanding of human health, aging, and disease. Mitochondrial dysfunction is implicated in cardiovascular, neurological, and muscular disorders; rare mitochondrial diseases (RMDs) caused by mutations to mitochondrial genes in both nuclear and mitochondrial DNA are devastating and largely untreatable. Despite a large volume of existing knowledge and increased public interest in mitochondrial biology and related diseases, its access and comprehension remain elusive to the scientific community and largely inaccessible to the general public. Our limited understanding of RMDs and their pathogenesis, compounded by the fragmented nature of clinical case information and unstructured text data, leads to delayed diagnoses and leaves us with a dearth of treatment options. We must make the knowledge we do possess more readily findable, accessible, and integrated across data types, enabling researchers and clinicians to amplify their ability to advance our understanding of mitochondrial biology and related diseases.

The FAIR Data Principles [1] promote the philosophy that scientific knowledge presented in any form should be made Findable, Accessible, Interoperable, and Reusable to ensure that it has the greatest impact in advancing our understanding of biology and disease. Briefly, publications, datasets, and metadata should be assigned a globally unique and persistent identifier such as a digital object identifier (DOI) to make them Findable; the digital objects should be retrievable via their identifier using an open, standardized, and free communications protocol to make them Accessible; data and metadata should use a formal, broadly applicable vocabulary and domain-recognized ontologies to make them Interoperable; and metadata on the digital objects should contain detailed provenance, licensing, and version information to make them Reusable. The FAIR

Principles have been well-received and adopted internationally by government funding agencies [2]. The principles are relatively easy to implement and evaluate, with the potential to increase the reach and impact of any given research project or dataset. Many biomedical resources and knowledge, however, remain decidedly unFAIR, particularly the information that is conveyed only by unstructured text data.

We have established the Mitochondrial Gene Wiki Project and the MitoCases Mitochondrial Disease Knowledge Platform to improve findability, accessibility, interoperability, and reusability of mitochondrial biology and mitochondrial disease related knowledge. The Mitochondrial Gene Wiki Project, discussed in **Section 1**, has made vast amounts of mitochondrial knowledge available through contributions to Wikipedia articles on mitochondria-related genes and proteins. This is an ideal starting point for researchers, students, and patients exploring mitochondrial biology and related diseases. The project has also been an effective educational tool for introducing biomedical literature curation, research methods, knowledgebase access, and scientific writing to students contemplating academic research and pre-medical studies. The MitoCases Rare Mitochondrial Disease Knowledge Platform, presented in **Section 2**, houses detailed metadata on RMD clinical case reports (CCRs), provides a powerful search mechanism for case discovery, and makes all data available for download or via API access. Researchers can use the platform to identify case reports on a disease of interest or involving a particular gene, protein, or pathway; clinicians can search for a set of symptoms to find similar cases and provide better patient care; and data scientists can access the dataset to develop tools for natural language processing (NLP) and named entity recognition (NER).

Section 1. Gene Wiki

1.1 Background and Approach

1.1.1 Gene Wiki Project and citizen science.

A wide range of scientific and clinical concepts, biomolecules, and diseases are represented on Wikipedia, providing a reliable source of detailed information and references for the general public and biomedical community. The Gene Wiki Project [3], an effort within Wikipedia, was established to make biomedical knowledge readily available on the platform and to encourage the participation of citizen scientists in contributing to the resource. They kickstarted the effort to annotate all human genes by automatically generating 9,000 “stub” pages [3] and populating them with basic content from Entrez Gene [4]. The Gene Wiki Project develops and maintains informatics tools to enable greater community contribution, including a simple “Gene Wiki Generator” web service that allows users to automatically create a similar stub page on their gene of interest [5]. The Gene Wiki Project has further elevated the utility of Wikipedia by automatically instantiating and populating “infoboxes” on each page with structured gene and protein data from Entrez Gene, UniProt, and the PDB. The Gene Wiki has since imported all human and mouse genes and proteins into Wikidata as structured entities, including nearly 60,000 human genes and 30,000 human proteins [6]. They have also implemented Semantic Wiki Links on the platform that facilitate machine-readable triplets that can convey the relationship between entities [7]. The resulting product is a broad collection of structured pages that may be more effectively digested by the public and contributed to by citizen scientists.

1.1.2 Underrepresentation of mitochondrial knowledge on Wikipedia.

Mitochondrial gene, protein, and pathway articles are also available on Wikipedia, but many mitochondria-related entities remained inadequately represented and poorly annotated. We identified a large collection of pages related to mitochondrial genes and proteins in need of significant updates. Many of these were stub pages that had yet to see any contributions, while many others were altogether nonexistent. Researchers have a tendency to prefer studying and functionally annotating only a small collection of genes for which copious information exists [8]. Information is indeed available on these mitochondrial entities, but their pages had not seemed to garner the attention of the community.

1.1.3 Project focus.

We have undertaken a concerted effort to identify and address these deficiencies on Wikipedia. The goal of the Mitochondrial Gene Wiki effort is to expand the breadth of the Gene Wiki Project, with a particular focus on the cardiac mitochondrial proteome. We have endeavored to establish well-organized examples of Wiki pages for the proteins within the conserved cardiac mitochondrial proteome, those with cardiovascular significance in coronary artery disease (CAD), as well as genes and proteins implicated in rare mitochondrial diseases. We wrote detailed protocols for analyzing current Gene Wiki pages for both their organization and content, curating knowledge from PubMed and key knowledgebases, and writing well-referenced articles suited for the Wikipedia framework and its audiences.

We selected five sub-proteomes from the human cardiac mitochondrial proteome for analysis and improvement, including metabolism (312 genes), oxidative phosphorylation (101), proteolysis (58), apoptosis (54) and redox (31) proteins (**Fig. 5-1**). This list was initially derived from proteomics datasets covering the cardiac mitochondrial proteome [9, 10]. Additionally, we targeted a collection of membrane proteins (17) and tRNA (18) that are implicated in rare mitochondrial disease, contractile proteins (56) critical to cardiac function, as well as proteins identified as risk factors in coronary artery disease (CAD; 25). The entire collection of proteins was compiled in the master article record spreadsheet, with links to the article pages, assignments to specific team members, initial article analysis, improvement metrics for each section, and the total article content in bytes before and after editing. Trainees and team members chose assignments within one or more of these sub-proteomes, based on their expertise, training, and research interests, then analyzed each of the 672 protein pages within Gene Wiki and identified those in need of improvement and update.

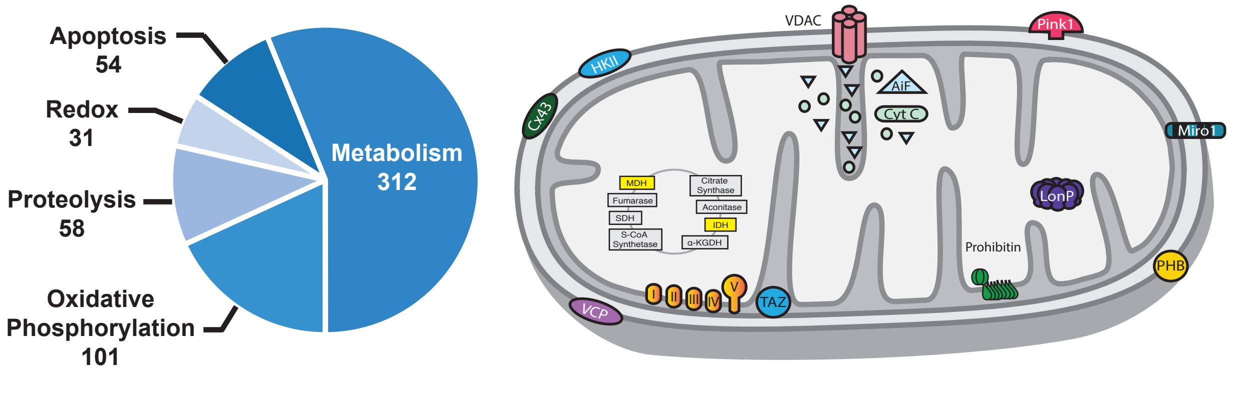


Figure 5-1. Targeting cardiac mitochondrial subproteomes for article improvement on Gene Wiki. Five subproteomes from the human cardiac mitochondrial proteome were selected for analysis and improvement. These include metabolism (312 genes), oxidative phosphorylation (101), proteolysis (58), apoptosis (54) and redox (31). Additional clusters of mitochondrial membrane proteins, contractile proteins, proteins implicated as risk-factors for coronary artery disease (CAD), tRNA-related proteins, as well as proteins and genes implicated in rare mitochondrial diseases were chosen for article improvement. Trainees and team members chose to improve articles within one or more of these subproteomes, based on their expertise, training, and research interests.

1.1.4 Gene Wiki as a training platform.

Wikipedia has been used successfully as an educational tool in several domains [11-13], and it provides a unique opportunity in the biomedical realm as well. Many students progressing through biology programs and pre-medical studies with the goal of entering into academic research or clinical professions lack essential research and writing skills necessary to communicate effectively in the scientific community. These skills are difficult to acquire and are rarely taught explicitly during the course of undergraduate studies. Without joining a research lab and actively participating in journal clubs and lab meetings, it is entirely possible to progress into medical school or graduate studies without ever learning how to find research papers, read them critically, or write an appropriate scientific discussion. Furthermore, the increasing prevalence of informatics tools and resources presents another challenge for aspiring biomedical professionals who will undoubtedly encounter them and need to navigate them throughout their continued studies and into their career. The large gap in coverage of mitochondrial genes and proteins provides a valuable space in which to introduce students to fundamental biomedical knowledge about mitochondria and impart these key skills for scientific research. Therefore, we used the Mitochondrial Gene Wiki effort to teach the basics of cardiovascular physiology, mitochondrial biology, and the role of mitochondria in health and disease. We imparted research skills in biomedical literature curation and critically assessing scientific publications, knowledgebase access, the effective use of informatics tools, and how to synthesize all of this information into impactful scientific communication. In training students from all educational levels, we made significant contributions to Gene Wiki pages and instilled the FAIR Principles in the next generation of biomedical researchers and clinicians.

1.2 Methods

The high-level workflow followed throughout this project, depicted in **Fig. 5-2**, entails (i) Gene Wiki article analysis, (ii) manual curation of detailed protein and gene information from a variety of sources, including PubMed, WikiData, UniProt, COPaKB, and others, and (iii) knowledge synthesis into an updated and improved Gene Wiki article. Ultimately, these contributions of scientific knowledge and insight on key biomolecules in health and disease are (v) made available to the community in a FAIR and open-source manner in the form of Gene Wiki articles and (iv) publications. Each step in this process was supported by detailed standard operating procedures

(SOPs), as well as regular meetings and inter-operator evaluation of completed tasks. We designed a master record for article analysis before and after improvement by our team, providing a mechanism for detailed statistics on individual and group-wide contribution.

1.2.1 User onboarding, training, and accounts.

Initial training and onboarding for new team members included an overview of the goals of the Mitochondrial Gene Wiki effort and an introduction to mitochondrial biology and cardiovascular physiology. Additionally, each team member read about the essentials of Wikipedia contributions detailed in the Wikipedia Tutorial article (https://en.wikipedia.org/wiki/Wikipedia:Tutorial), the Wikipedia Manual of Style (https://en.wikipedia.org/wiki/Wikipedia:Manual_of_Style), and the Wikipedia Etiquette article (https://en.wikipedia.org/wiki/Wikipedia:Etiquette). These provide all of the necessary information about basic editing instructions, how and when to create new pages, how to use Semantic Wiki links appropriately, citations, formatting style guides, standard practices regarding edit reversals or significant changes, indicating changes at the time of submission, as well as a general guide to etiquette within the Wikipedia community. These articles are particularly

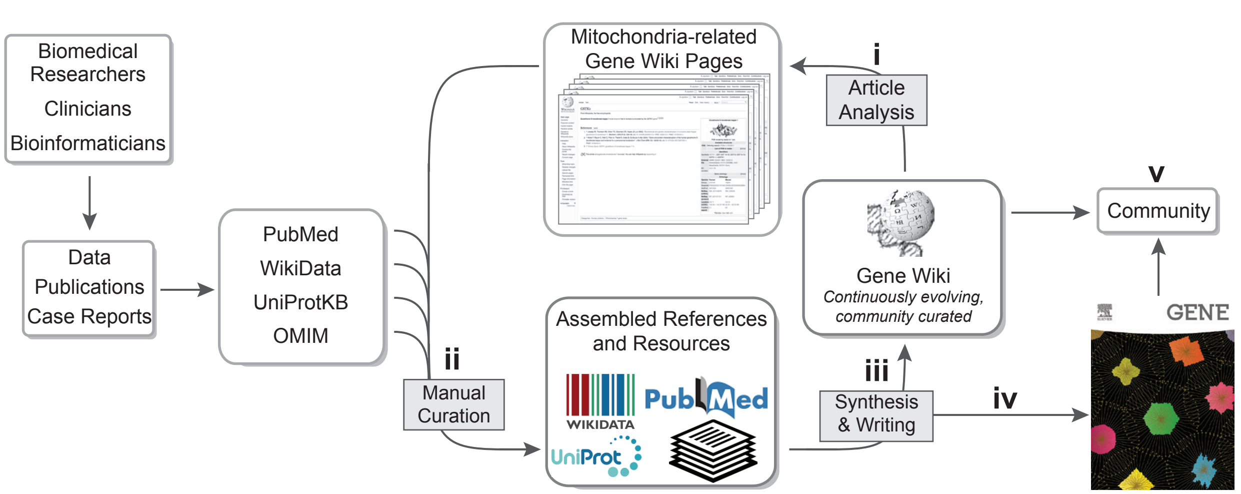


Figure 5-2. Mitochondrial Gene Wiki article improvement workflow. (i) Articles of interest on Gene Wiki are first analyzed for completeness and quality of key sections (Introduction, Structure, Function, and Clinical Significance), the inclusion of ample and recent references, and the presence of extensive semantic linkages throughout the article. **(ii)** Upon identifying articles for improvement, comprehensive gene, protein, and clinical knowledge is curated from PubMed, UniProt, Reactome, WikiData, OMIM, and other knowledgebases and data sources to assemble a complete view of available information on the gene of interest. **(iii)** The assembled information is then synthesized and enhanced through integration into a complete Gene Wiki article and, **(iv)** for select genes of interest, submitted for publication to the Gene journal through their Gene Wiki Reviews mechanism. **(v)** The resulting articles are made available to the community, amplifying access to critical knowledge on mitochondrial biology.

useful when unsure of how to accomplish a certain type of formatting, how to give appropriate attribution, and which citations require special tags for their Creative Commons licensing. Each team member within the Mitochondrial Gene Wiki effort created an individual user name so as to track their contributions. Institution and group user accounts are forbidden by Wikipedia terms of use, so each member submitted their user name for inclusion in our master article record for later analysis. Team members were asked to include a tag in their signature line to better track contributions outside of the master article record maintained in-house.

1.2.2 Gene Wiki article analysis.

Team members first searched for each of their assigned proteins and genes on Wikipedia for existing pages, checking for synonyms and alternate naming conventions before identifying missing pages as such within the master article record. The existing Gene Wiki articles were analyzed for completeness, organization, the quality and length of key sections (Introduction, Structure, Function, and Clinical Significance), the presence of the Gene Wiki infobox, the inclusion of ample, relevant, and recent references, and the presence of extensive Semantic Wiki links throughout the article connecting it to other biological concepts in articles across Wikipedia. Within the master article record, space is provided for scoring the key article sections on a scale of 1-3, indicating whether the section was nonexistent (1), incomplete (2), or complete (3). To be considered complete, the Structure, Function, and Clinical Significance sections are each expected to contain 250-500 words or more, accompanied by appropriate references for each statement and Semantic Wiki links to other articles, and be prefaced with an appropriate section heading. The Introduction is expected to be concise. Excessive information about a particular topic earns an “incomplete” rating for the section and is ultimately moved to a more appropriate section, such as Structure, Function, or Clinical Significance. The presence of the Gene Wiki infobox is also indicated and marked for addition or update if it is absent or incomplete. Finally, team members record the size of the article in bytes prior to any editing.

1.2.3 Knowledge curation.

Comprehensive gene, protein, and clinical information is curated for the protein of interest using a wealth of resources, including NCBI Gene, PubMed, UniProt, the Cardiac Organellar Protein

Atlas Knowledgebase (COPaKB), the Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB), Reactome, IntAct, and Online Mendelian Inheritance in Man (OMIM). Under the SOP for research activities, team members are advised to begin their search on NCBI Gene (https://ncbi.nlm.nih.gov/gene) with the gene ID and the "[Homo sapiens(human)]" tag. The summary typically provides a reasonable overview of known characteristics of the gene, and basic information such as the location of the gene in the human genome is contained under the "Genomic context" section. The "Bibliography" and "GeneRIFs" (Genetic References Into Functions) sections supply a large collection of related articles that are directly relevant to the gene and its functions. Once these references have been exhausted, we query PubMed for the gene or protein name and search for recent review articles. These articles aid in becoming familiarized with the gene and/or protein. Their reference lists provide many avenues for deeper information. UniProt [14] (https://uniprot.org/) is a preferred source of information on the protein and its functions, from molecular weight, variants, post-translational modifications and interactions, as well as key functional and structural domains. COPaKB [15] is a cardiac proteome-focused knowledgebase containing mass spectrometry datasets and RCSB PDB (http://rcsb.org/) houses structural models of many proteins. UniProt, COPaKB, and RCSB PDB are recommended sources for fleshing out the Structure and Function sections of the Gene Wiki articles. Reactome [16] (https://reactome.org/) and IntAct [17] (https://ebi.ac.uk/intact/) furnish detailed pathway and interaction information to construct a comprehensive Interactions section. Finally, the Clinical Significance section is best addressed using OMIM (https://ncbi.nlm.nih.gov/omim) resource, an extensive collection of genetic phenotypes, allelic variants, and the associated references.

1.2.4 Gene Wiki article improvement.

The assembled information is then synthesized and integrated into a Gene Wiki article, depending on the completeness of the existing page. Our team members are advised to write a draft of the document first, to avoid any loss of work due to a bad connection or accidentally closing the Wikipedia editing tab of their browser. Missing pages are instantiated according to the instructions provided (https://en.wikipedia.org/wiki/Wikipedia:How_to_create_a_page), and headings for the Structure, Function, Interaction, and Clinical Significance sections are created. Next, the relevant information from the curation step is incorporated into each section with references to support each claim. The Introduction section is pared down to the essential information to provide a concise summary of the contents of the article and critical information relating to protein function and disease relevance. Longer explanations are moved to the relevant sections. Semantic Wiki links are added throughout the article and, in keeping with the Wikipedia Tutorial, only at the first appearance of a concept to avoid overlinking. If it is missing, the Gene Wiki infobox is populated on the page through the BioGPS site (http://biogps.org/) or by adding the '{{infobox gene}}' string to the markup code editor at the top of the page. Upon submitting the edited article, infobox content is imported in the background by the ProteinBoxBot, managed by the Gene Wiki Project team [6]. Upon submitting the final version of the article, team members recorded key measures of their contribution, including the number of semantic links and references they added, which sections they improved or created, as well as the initial and final size of the article in bytes.

1.2.5 Converting Gene Wiki articles to peer-reviewed publications.

Select genes or proteins of interest, particularly those without recent or comprehensive review articles available, are submitted for publication to the *Gene* journal through their Gene Wiki Reviews mechanism [18]. This initiative provides a venue through which the contributions of scientists to the Gene Wiki platform can be recognized and provide tangible professional benefit in the form of a peer-reviewed publication in a well-known and circulated journal. Through the efforts of the Mitochondrial Gene Wiki team, two publications have been secured through this mechanism, including a comprehensive review of the Hsp70 family chaperone BiP (binding immunoglobulin protein), encoded by the *HSPA5* gene [19], as well as a thorough review of the tafazzin-encoding *TAZ* gene implicated in Barth syndrome, a complex, rare mitochondrial disease affecting cardiolipin remodeling [20]. Scientific contributions to Wikipedia are largely underappreciated and garner no professional recognition. Gene Wiki Reviews provide an effective way to amplify the visibility of these efforts and to publish reviews on important genes and proteins from devoted researchers.

1.2.6 Article standardization, evaluation, and quality control.

Each article followed a standardized format for ease of public consumption. Furthermore, articles were evaluated using the master article record and the metrics recorded by each contributor. The number of new pages created, sections improved, semantic links and references added, and overall size of the articles in bytes were logged and used for subsequent analysis of the overall contributions by the team and by each individual team member.

The efforts undertaken by the Mitochondrial Gene Wiki team were accomplished with significant contribution from younger students, including dedicated high school age students in advanced summer internship programs, undergraduates of all levels looking for research experience, as well as Master's and Doctoral level graduate students working on projects related to some of the selected genes and proteins of interest. Each trainee was individually guided through the research, curation, writing, and editing process to ensure that the submitted articles represented the highest caliber and sufficiently covered the key pieces of knowledge on each protein. For their first two articles, trainees worked individually with a senior lab member – either a doctoral student, a postdoctoral scholar, a clinician, or a senior scientist with strong writing abilities and knowledge of the proteins under study. Many trainees worked in pairs or groups of three for another set of 2-3 articles. For the remainder of their training they were assigned a small collection of related proteins to work on individually. All trainee articles were reviewed in depth with the assistance of a senior lab member upon submission and revised with their supervision when necessary.

1.3 Results and Discussion

1.3.1 Initial article analysis.

Our article scoring protocols provided a mechanism by which to evaluate Gene Wiki pages and track the progress of the Mitochondrial Gene Wiki effort over time. At the outset, only 1% of the 672 mitochondrial protein and gene articles slated for improvement achieved a score that would allow us to consider them "complete." Another 8% already had a significant amount of content and required some restructuring, organization, and the addition of more recent references. About a quarter of the pages (24%) were rated as "somewhat complete," and were typically lacking cohesive structure, organization, or were poorly cited. In total, roughly one third of our targeted proteins had at least a start at a reasonably well constructed article available.

The majority of the pages, however, were either missing entirely or severely lacking in content. Nearly a quarter of the genes and proteins we identified (22%) lacked any representation on

Wikipedia whatsoever. These pages required construction from the ground up, including enlisting the ProteinBoxBot for initial instantiation of the Gene Wiki infobox and writing each of the key sections from scratch. Another half of the pages (46%) were rated “mostly incomplete,” which typically meant that they were little more than a placeholder with a title and perhaps an infobox with basic gene and protein information. These pages are referred to on Wikipedia as a “stub” page and are usually accompanied by a stub tag, which is used as an invitation to contribute to that page. Upon adding to these pages and writing a more complete article, the editor is expected to remove the tag.

1.3.2 Article improvement progression and metrics.

The Mitochondrial Gene Wiki team worked to improve these articles over the course of several years (**Fig. 5-3**). The initial effort in the first year saw an increase of “complete” articles from 1% to 41%, primarily by tackling “mostly incomplete” articles and reducing their share from 46% to 19%. Missing articles are frequently nonexistent due to the dearth of available information on these genes and proteins. They are frequently small subunits or chaperones that are not well characterized in the literature and require significantly more effort to unearth enough content for a full article. In the first year, these decreased from 22% to 13%. “Somewhat complete” articles also decreased from 24% to 17%, while “Mostly complete” pages increased from 8% to 10%.

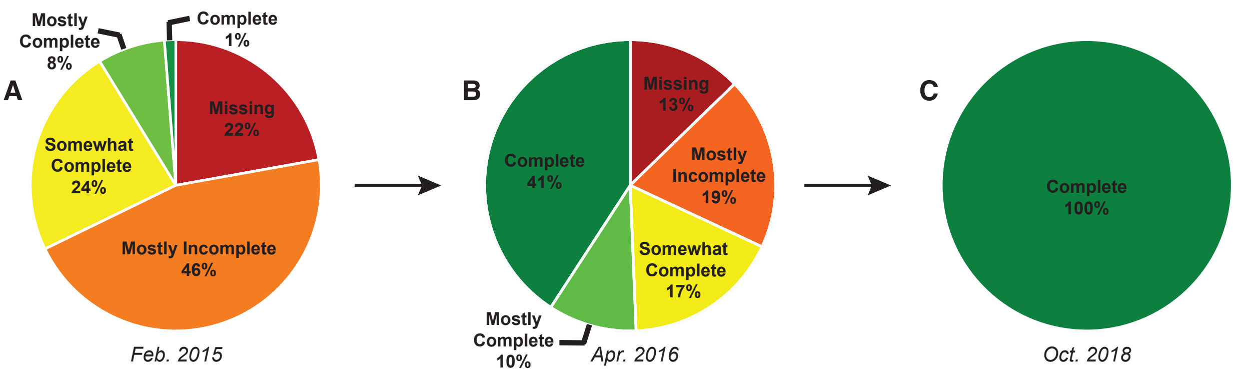


Figure 5-3. Mitochondrial Gene Wiki article improvement progression. (A) Initial scoring of the selected Gene Wiki articles revealed that only 1% of the articles could be deemed “complete.” Over the course of a year, the Mitochondrial Gene Wiki team worked to improve these articles. **(B)** Through this initial effort, we raised the number of “complete” articles from 1% to 41%, a result of the reduction of “mostly incomplete” articles from 46% to 19% and of “missing” articles from 22% to 13%. **(C)** All sub-proteomes were completed in the following two years along with other mitochondria-related protein clusters, including membrane, contractile, coronary artery disease (CAD), and tRNA-related proteins.

In the following two years, Gene Wiki articles on all of the selected proteins were completed from the five key sub-proteomes, along with the membrane, contractile, coronary artery disease (CAD), rare mitochondrial disease, and tRNA-related proteins. The master article record allowed us to measure the size of our contribution by a collection of metrics, including pages created, improved, and completed; sections improved; and the number of references, Semantic Wiki links, and kilobytes (kB) added (**Table 5-1**). Given that all of the content added to these pages is free text, a total contribution of over 4MB is really quite significant. Over 5,600 references were added across 541 pages, making mitochondrial research more findable and accessible, and potentially more reusable through newfound citations and awareness of the work. Wikipedia is an excellent starting point for the beginning of a research or writing effort, particularly when ample references are made available and cited appropriately.

1.3.3 Mitochondrial Gene Wiki as an educational platform.

Beyond improving articles and access to scientific knowledge, the Mitochondrial Gene Wiki project was also implemented as an effective educational mechanism for imbuing students pursuing careers in science and medicine with effective research and writing techniques. Trainees from all educational levels were guided through each step of the process to develop critical skills in curation, both from the literature and a variety of informatics sources; article and data source evaluation; knowledge synthesis; and scientific writing. In all, we trained 35 high school summer interns, 12 undergraduate college students, and 2 graduate students over the course of three years. Throughout the process, we used the Mitochondrial Gene Wiki article improvement

Cluster	*Pages Created*	*Pages Improved*	*Pages Completed*	*Sections Improved*	*References Added*	*Links Added*	*kB Added*
Metabolism	21	139	160	650	1,153	4,224	963.57
OXPHOS	36	80	116	547	863	4,607	761.89
Proteolysis	6	66	72	287	997	1,966	674.43
Apoptosis	8	54	62	374	498	1,980	442.75
Redox	0	15	15	74	162	282	64.93
Membrane	0	17	17	77	148	842	110.84
Contractile	0	56	56	262	1,287	3,220	617.10
CAD	4	21	25	99	451	809	315.54
tRNA	0	18	18	62	115	456	79.39
Total	**75**	**466**	**541**	**2,432**	**5,674**	**18,386**	**4,030.42**

Table 5-1. Mitochondrial Gene Wiki article improvement metrics. Overall article improvements are shown by protein/gene cluster, including pages created, improved, and completed; sections improved; as well as references, Semantic Wiki links, and kilobytes (kB) added.

efforts as a vehicle to introduce them to cardiovascular physiology, mitochondrial biology, clinical concepts, and FAIR Principles in biomedicine. We view this as a critical aspect of training for the next generation of researchers and clinicians.

Section 2. MitoCases

2.1 Background and Approach

2.1.1 Clinical reports on rare mitochondrial disease.

Rare mitochondrial diseases (RMDs) are complex and difficult to diagnose, with limited treatment options and no available cures. Rare diseases are defined in the U.S. by the Rare Diseases Act of 2002 as those affecting fewer than 200,000 individuals [21], or about 1 in 1,500 individuals. The low incidence of rare mitochondrial diseases makes it particularly difficult to develop effective treatments, conduct clinical trials, or ensure awareness among health care professionals. Clinical case reports (CCRs) provide clinicians and researchers with a wealth of information about a wide range of diseases, their defining characteristics, and successful as well as failed attempts at treatment. As of August 2019, PubMed hosts over two million CCRs, and the rate of publication of this type of report is continually increasing. A sample of 30 RMDs are covered by fewer than 3,000 of these reports, demonstrating their rarity as well as the importance of being able to easily find and access the clinical insights contained within. However, much of this information remains relatively hidden as a result of ineffective indexing and a lack of comprehensive metadata on the reports. Unstructured text data is not accessible by computational means and requires extensive manual effort to curate, read, and evaluate cases before synthesizing the information into actionable clinical knowledge.

2.1.2 CCR curation challenges.

The wide range of clinical signs and symptoms in mitochondrial diseases makes them particularly difficult to diagnose and to differentiate between without in-depth genetic profiling, which is sometimes not readily or quickly available. Furthermore, the rarity of these diseases means that comprehensive clinical profiles of patients and best practices in treatment and therapeutics are sometimes unavailable or difficult to source. Clinical case reports represent a key mechanism for communication between clinicians, particularly regarding rare and uncommon presentations.

However, this critical resource is poorly indexed, making it difficult to quickly and accurately identify relevant documents. PubMed is the primary source used for accessing CCRs, but the lack of extensive, case-specific metadata on these reports makes curating relevant reports an inordinately time-consuming task. Pertinent documents are easily missed while many irrelevant documents are acquired, slowing the sourcing of critical case information with the necessary investigation of inaccuracies and false leads.

PubMed does have a feature for searching CCRs, but the search mechanism is limited to the title, abstract, and MeSH (Medical Subject Headings) terms [22, 23]. MeSH terms provide some metadata on the content of the reports, but the system is not ideally suited for complex queries probing the symptoms, genetics, or demographics of patients discussed within. MeSH terms related to patient age are aligned with common clinical classifications ("Infant", including "Infant, Newborn"; "Child", including "Child, Preschool"; "Adolescent"; and "Adult", including "Young Adult", "Middle Aged", and "Aged"), but case reports are frequently mislabeled or tagged with multiple labels not related to the patient discussed in the report, making it difficult to filter cases using these tags. Similarly, gender tags exist but are inconsistently utilized. Frequently, CCRs are labeled with both "Male" and "Female" terms, compromising the utility of the tag as a search mechanism. MeSH terms exist for many disease classifications as well, but they are also used inconsistently; in a collection of 30 reports on Barth syndrome, only 10 were labeled with the "Barth syndrome" MeSH term, and 17 with the more general "syndrome" term. There are MeSH terms for a variety of symptoms, biomolecules, and other related concepts, but these, again, are used inconsistently and do not provide comprehensive or reliable metadata related to the reports. Searching for a set of symptoms typically returns very few results.

In fairness, MeSH was not designed to provide highly detailed metadata on case reports. Still, the lack of such a system is hampering the ability of clinicians to access reports relevant to their cases and makes it more difficult for them to make accurate, timely diagnosis and provide adequate treatment and therapeutics to their patients. Researchers are similarly limited in their ability to study translational aspects of pathways and drugs related to particular diseases. Furthermore, data scientists and text-mining researchers lack reliable and accessible collections of text data or comprehensive ontologies of clinical language, presenting a significant barrier to development of advanced NLP approaches necessary for powerful clinical decision support systems.

2.1.3 Rendering RMD CCRs FAIR.

We set out to make the knowledge represented in CCRs on RMDs more FAIR by generating detailed, highly structured metadata on the reports and creating a platform from which it can be effectively accessed, searched, and analyzed. We manually curated 384 reports on 8 RMDs, including deficiencies in complex I through V of the electron transport chain, carnitine deficiency, megaconial type congenital muscular dystrophy (MDCMC), and Barth syndrome. Leveraging the collective expertise of our group regarding mitochondrial physiology and clinical paradigms, we imposed structure on the reports using the standardized Metadata Acquired from CCR (MACCR) template [24, 25] discussed in **Chapters 3 and 4**, including detailed bibliographic, demographic, and medical content from the CCR. We also constructed a digital map of RMD symptomology by codifying patient symptoms described in each CCR using clinical controlled vocabularies ICD-10 and ICD-11 (International Statistical Classification of Diseases and Related Health Problems, 10th and 11th revisions) [26, 27].

In alignment with the FAIR Data Principles, we built a cloud-based MediaWiki platform called MitoCases (http://mitocases.org/) to house all metadata and ICD codes, providing biomedical, clinical, and text-mining researchers with a large dataset of structured text data for downstream analysis. All metadata is downloadable in whole, by disease, or in subsets based on results via our Elasticsearch functionality, and the MitoCases API allows direct programmatic access to platform functionalities. We provide Jupyter notebooks and present potential use cases that may be pursued through downstream analysis by researchers, physician investigators, clinicians, data scientists, text-mining researchers, pharmaceutical companies, and clinical trial coordinators. The entirety of the underlying codebase and analysis tools are available to the public on GitHub (https://github.com/aogarlid/mitocases). The platform empowers biomedical researchers to create *in silico* RMD cohorts, amplify small sample size studies by integrating with larger cohorts, and leverage the cumulative power and value of many patient reports for statistical analysis. MitoCases provides a FAIR Data resource of standardized and structured CCRs on RMDs and integrates existing ICD ontologies to enable the biomedical community to elevate understanding of mitochondrial disease and improve patient care.

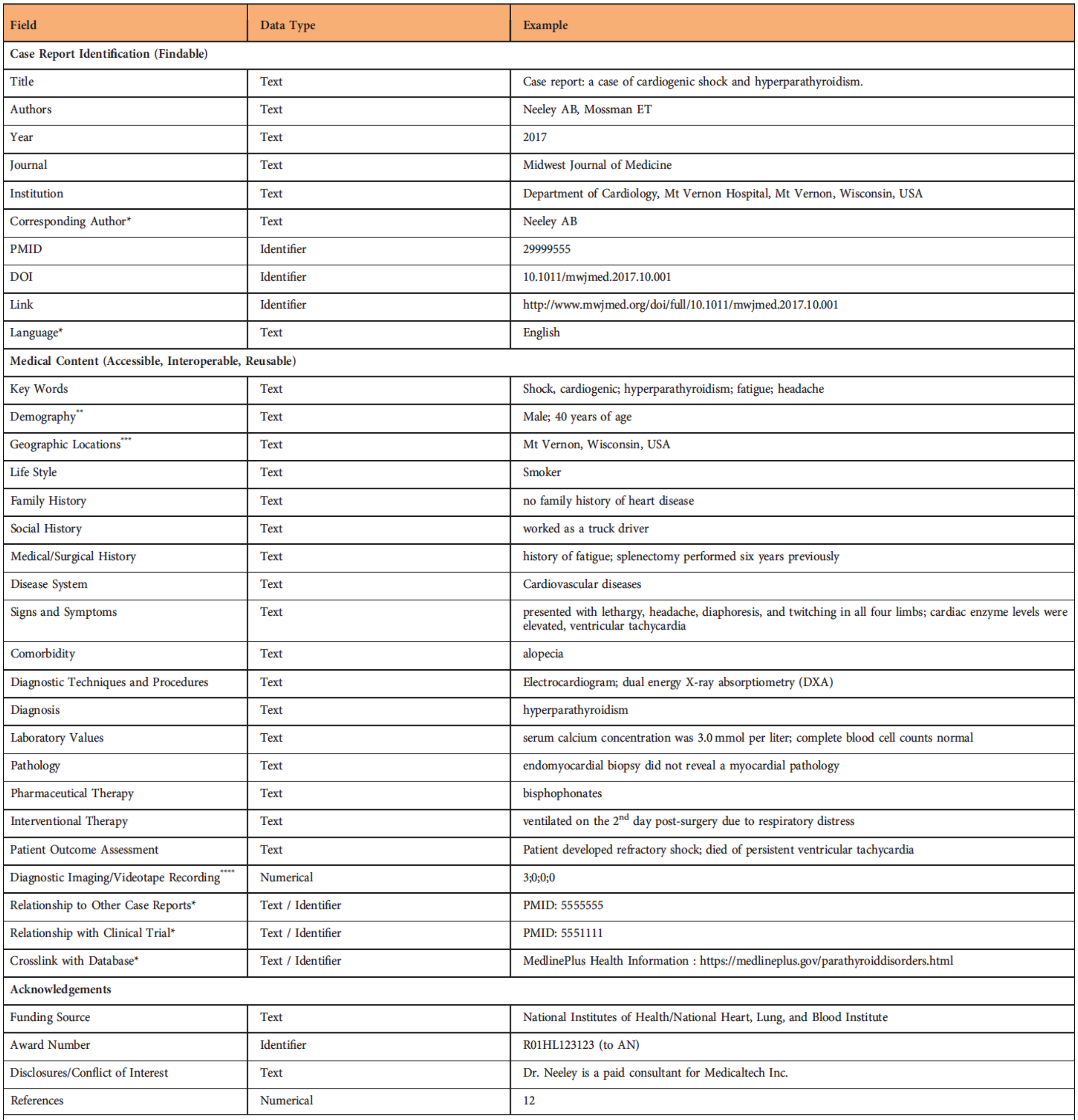
2.2 Methods

2.2.1 Manual case report curation.

There are currently more than two million clinical case reports available on PubMed, over 3,000 of which describe patients with a sample of 30 different RMDs. We selected 8 of these diseases for article curation, including deficiencies in complex I-V of the electron transport chain, systemic carnitine deficiency, Barth syndrome, and megaconial-type congenital muscular dystrophy. Our manual curation approach filtered the 788 CCRs discovered across these diseases for articles describing a single patient, written in English, with the full text available for download. The resulting set of articles slated for metadata extraction comprised 384 CCRs: Complex I deficiency (84 reports), Complex II deficiency (12 reports), Complex III deficiency (29 reports), Complex IV deficiency (170 reports), Complex V deficiency (7 reports), systemic carnitine deficiency (45 reports), Barth syndrome (30 reports), and congenital megaconial type muscular dystrophy (MDCMC; 7 reports).

2.2.2 Metadata extraction with MACCR template.

To impose structure on clinical information relating to RMDs, we created a metadata template for CCRs (the MACCR template) with bibliographic, demographic, and medical content sections, as well as genetic information (**Table 5-2**) [24, 25]. Each section and the corresponding fields were developed in close collaboration with clinicians and researchers to ensure full representation of the relevant information contained within these reports. We manually extracted metadata from each CCR using the standardized MACCR template. Full instructions are provided in the *Metadata Extraction Guide* [28], originally published as a Data Citation in the *Sci. Data* manuscript [25] presented in **Chapter 4**. Annotators begin by filling the case report identification fields of the template with bibliographic information about the article, including the title, authors, year of publication, journal, institution, corresponding author, PMID, DOI, and the language in which the report is written. The fields in the Medical Content section identify document-level, concept-level, and text-level features of the report. For this section, annotators read the CCR closely and



Field	Data Type	Example
Case Report Identification (Findable)		
Title	Text	Case report: a case of cardiogenic shock and hyperparathyroidism.
Authors	Text	Neeley AB, Mossman ET
Year	Text	2017
Journal	Text	Midwest Journal of Medicine
Institution	Text	Department of Cardiology, Mt Vernon Hospital, Mt Vernon, Wisconsin, USA
Corresponding Author*	Text	Neeley AB
PMID	Identifier	29999555
DOI	Identifier	10.1011/mwjmed.2017.10.001
Link	Identifier	http://www.mwjmed.org/doi/full/10.1011/mwjmed.2017.10.001
Language*	Text	English
Medical Content (Accessible, Interoperable, Reusable)		
Key Words	Text	Shock, cardiogenic; hyperparathyroidism; fatigue; headache
Demography**	Text	Male; 40 years of age
Geographic Locations***	Text	Mt Vernon, Wisconsin, USA
Life Style	Text	Smoker
Family History	Text	no family history of heart disease
Social History	Text	worked as a truck driver
Medical/Surgical History	Text	history of fatigue; splenectomy performed six years previously
Disease System	Text	Cardiovascular diseases
Signs and Symptoms	Text	presented with lethargy, headache, diaphoresis, and twitching in all four limbs; cardiac enzyme levels were elevated, ventricular tachycardia
Comorbidity	Text	alopecia
Diagnostic Techniques and Procedures	Text	Electrocardiogram; dual energy X-ray absorptiometry (DXA)
Diagnosis	Text	hyperparathyroidism
Laboratory Values	Text	serum calcium concentration was 3.0 mmol per liter; complete blood cell counts normal
Pathology	Text	endomyocardial biopsy did not reveal a myocardial pathology
Pharmaceutical Therapy	Text	bisphophonates
Interventional Therapy	Text	ventilated on the 2nd day post-surgery due to respiratory distress
Patient Outcome Assessment	Text	Patient developed refractory shock; died of persistent ventricular tachycardia
Diagnostic Imaging/Videotape Recording****	Numerical	3;0;0;0
Relationship to Other Case Reports*	Text / Identifier	PMID: 5555555
Relationship with Clinical Trial*	Text / Identifier	PMID: 5551111
Crosslink with Database*	Text / Identifier	MedlinePlus Health Information : https://medlineplus.gov/parathyroiddisorders.html
Acknowledgements		
Funding Source	Text	National Institutes of Health/National Heart, Lung, and Blood Institute
Award Number	Identifier	R01HL123123 (to AN)
Disclosures/Conflict of Interest	Text	Dr. Neeley is a paid consultant for Medicaltech Inc.
References	Numerical	12

Table 5-2. Standardized metadata template for clinical case reports. The metadata template provides a set of features common to clinical case reports and facilitates concept-level metadata extraction. This template is arranged into three primary sections: Identification, Medical Content, and Acknowledgments. The first two categories are geared towards promoting aspects of FAIR standards, as indicated. A single document contains the majority of these features and the metadata template was designed in consultation with clinicians to ensure all key concepts were addressed. Data Type refers to the type of source data, including free-text ("Text"), unique database identifier or other structured value specific to the document ("Identifier"), or "Numerical" values, such as number of references. Examples shown here are simulated but representative of dataset contents. It should be noted that the Disclosures/Conflict of Interest listed here are fabricated and do not apply to the author. **Demography details are converted to consistent values to facilitate Case Report Search. ***If not provided within document text, geographic location is inferred from the associated institution. ****Numerical count of the total number of clinical images, figures, videos, and tables, respectively, published along with the main text of the report.

extract the exact text content from the report to fill the relevant fields: key words, demographics, geographic locations, life style, family history, social history, medical/surgical history, disease system, signs and symptoms, comorbidities, diagnostic techniques and procedures, diagnosis, laboratory values, pathology, pharmacological therapies, interventional therapies, patient outcome assessment, and counts of visual media included in the report (images, figures, videos or animations, and tables). Finally, the Acknowledgements section provides additional details about the report, including funding source, award number of any support mentioned, disclosures or conflicts of interest, and a numerical value of the number of references cited by the report.

2.2.3 Digital symptomology map with ICD codes.

For added interoperability with existing standardized clinical vocabularies, we identified symptoms discussed in each report and represented them using ICD-10 and ICD-11 codes. The symptomology dataset includes the ICD codes and their descriptions, as well as the text content from the CCR describing the symptom. Each code is contained on a new row in an Excel sheet and identified by the article PMID. Annotators read the CCR line by line to identify all mentions of symptoms pertaining directly to the patient, but avoiding those symptoms mentioned in the background description of a typical case presentation, as well as those mentioned in the discussion of similar or related cases. For each identified symptom, annotators copy the text describing it into the appropriate column, then search for the symptom codes in the ICD-10 and ICD-11 browsers (ICD-10: https://icd.who.int/browse10/2016/en and the alternative https://icd10data.com/; ICD-11: https://icd.who.int/browse11/l-m/en). Symptoms are frequently presented in the case reports with descriptions that are quite different than that of the ICD code, which requires that annotators consult available resources such as Wikipedia and PubMed to find synonymous phrases that will help identify the appropriate code. When there are seemingly multiple valid coding options, annotators are instructed to identify the best fit.

Some ICD codes have an “Other specified” or “Unspecified” option. For example, the “Heart failure, unspecified” ICD-11 code (BD1Z) would be employed for a case report that describes the patient as having “died from cardiac failure” without specifying a particular form of cardiac failure. Similarly, the “Other specified abnormalities of breathing” ICD-11 code (MD11.Y) is used for a case report describing a patient with “tachypnea”, as there is no unique code for this type of breathing issue. Neither ICD-10 nor ICD-11 contain a specific code for “lactic acidosis” so the general code is used (ICD-10 E87.2 “acidosis” and ICD-11 5C73.Y “other specified acidosis”). Finally, it is important to note that while ICD-10 does contain some codes for specific diseases, the recent release of ICD-11 appears to focus on symptoms to the exclusion of diseases (e.g., Barth syndrome is included in ICD-10 but not in ICD-11). When there is a discrepancy in representation of a symptom between the two systems, the symptom will still be recorded with the available code and the column for the other coding system is filled with “NA”. As the dataset grows, we are logging these discrepancies and will submit important codes missing from ICD-11 for inclusion in an upcoming revision.

2.2.4 Inter-annotator agreement.

To ensure comprehensive metadata extraction across all annotators and consistency between them, we monitored inter-annotator agreement by engineering a small degree of overlap between their assigned tasks [29]. Roughly 10% of the total collection of reports were assigned to all annotators and the resulting MACCRs and ICD codes were evaluated for consistency between team members as well as their relative comprehensiveness [30]. The identities of the repeat assignments were not revealed to the annotators so as to prevent any alteration in their typical methodology. Annotation tasks were assigned in batches, with each batch containing repeat assignments. This provided a mechanism for intra-annotator evaluation, or the consistency of a single annotator’s performance and comprehensiveness over time [29]. New annotators joining the team were assigned the past repeat tasks to evaluate their performance against past and existing team members, and to provide suggestions for improvement where necessary. Additionally, all new members were coached extensively and guided individually through their first set of assignments by an experienced annotator.

2.2.5 Creating the MitoCases platform.

With extensive metadata on RMD CCRs in hand, we developed the cloud-based MitoCases Rare Mitochondrial Disease Knowledge Platform (http://mitocases.org/) to house the metadata and provide a user-friendly interface with a powerful search function, download capabilities, an upload pipeline to contribute additional RMD CCR metadata, envisioned use case scenarios, and sample code for researchers to use in their own downstream analysis. MitoCases is hosted on an AWS instance and built upon a MediaWiki framework using MySQL for data storage and additional applications for API access (Flask) and search capabilities (Elasticsearch). The following section, **2.3 Technical Specifications**, details the front- and back-end components and architecture of the MitoCases platform.

2.3 Technical Specifications

The cloud-based MitoCases Rare Mitochondrial Disease Knowledge Platform is housed on an Amazon Web Services (AWS) Elastic Compute Cloud (EC2) t3.large instance with two virtual CPUs (vCPU) and 8 GB memory, providing ample computing power and storage space for the largely text-based datasets. MitoCases is built upon a MediaWiki framework with a MySQL database, Elasticsearch functionality, and a Flask-based API to allow for a modular, decoupled, and extensible platform (**Fig. 5-4**). All open source technologies, API access, detailed documentation, and use case scenarios for key user groups ensure that the platform and the data contained within is highly usable and FAIR, regardless of the technical capabilities of the user.

2.3.1 MediaWiki framework.

MediaWiki is built upon a free open source software (FOSS) bundle known as the LAMP

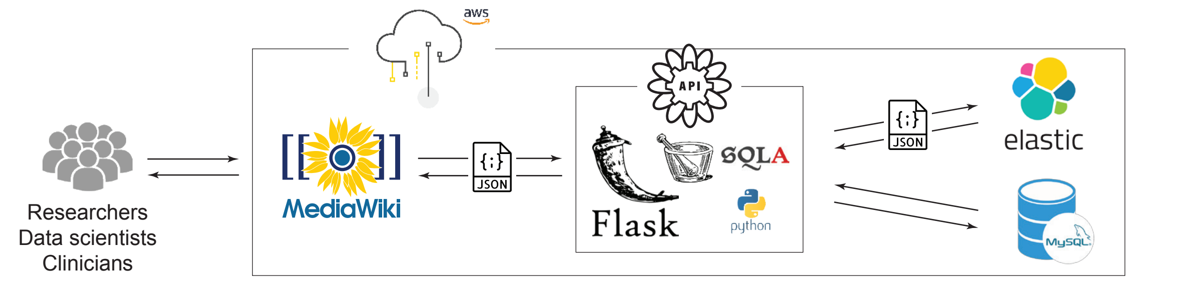


Figure 5-4: Cloud-based MitoCases platform architecture. The MitoCases platform is accessible to researchers, clinicians, and data scientists through a custom-configured MediaWiki user interface built on a LAMP stack (Linux, Apache, MySQL, and PHP). The Data API (application programming interface) is driven by the Python-based Flask web application framework and mediates all back-end functionality and data handling throughout MitoCases. The SQLalchemy SQL toolkit facilitates interactions with the MySQL database where MitoCases data is stored, and Elasticsearch provides a full-text search engine configured to query the MySQL database. JSON (JavaScript Object Notation) files are used for communication between the various components to deliver queries, download requests, and uploads. The MitoCases Platform is housed on an AWS (Amazon Web Services) Elastic Compute Cloud (EC2) T3.large instance.

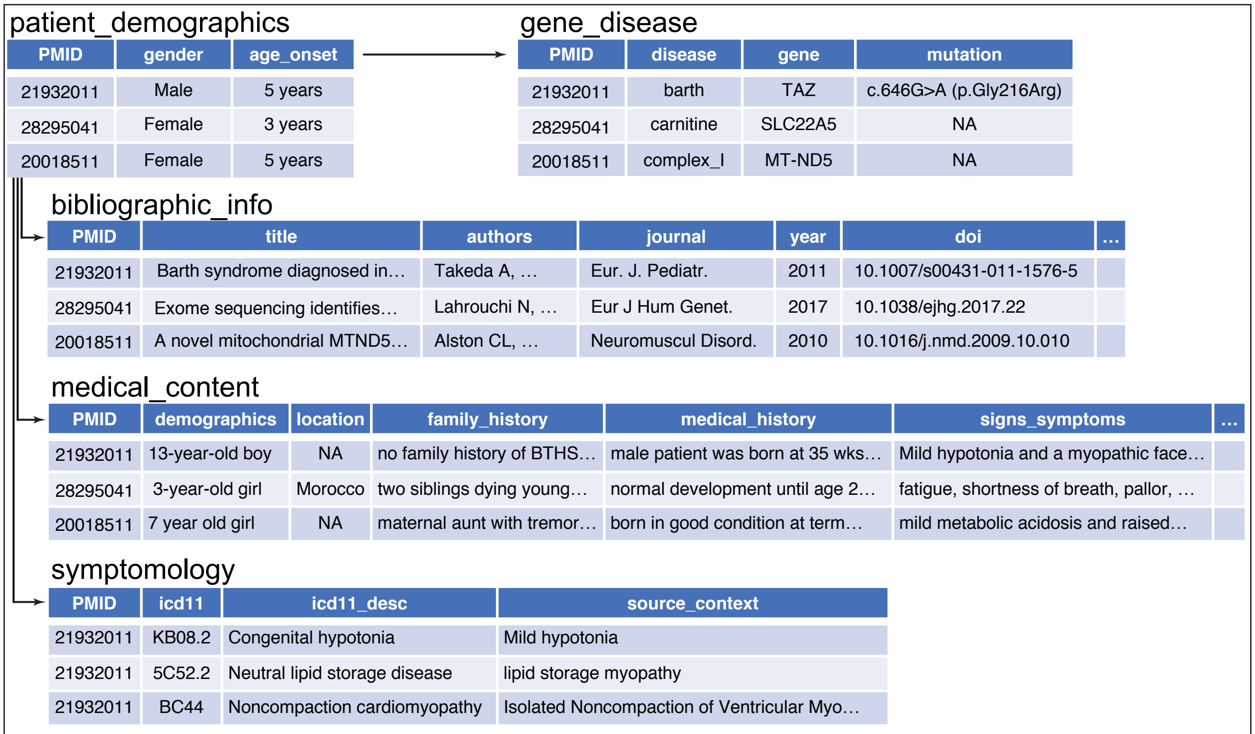
stack, combining the Linux operating system, Apache web server, MySQL Relational Database Management System (RDBMS), and the PHP programming language. For the purposes of the MitoCases platform, we used the proxy module so Apache can act as proxy and reverse proxy, allowing requests to pass through the website URL and be redirected to specified endpoints, such as the Flask API (application programming interface) used throughout the back end development for uploads, downloads, and data management. The WSGI (Web Server Gateway Interface) module allows Apache to receive requests from the API using Python code, which is used throughout this implementation for data handling, analysis, and visualization. MediaWiki furnishes clean, structured data entries in an easily recognizable and navigable format. The MediaWiki API enables programmatic page building with additional markup and easily incorporates new features.

2.3.2 MySQL database.

The underlying metadata and ICD codes extracted from CCRs on MitoCases utilizes the MySQL RDBMS for data warehousing and extract, transform, load (ETL) functions, chosen specifically for its inherent integration and optimization within the MediaWiki infrastructure. MySQL facilitates simple queries to link data by their common relationships, typically referred to as a primary key. To accommodate reports discussing multiple patients and to prevent reuse of the primary key, MitoCases employs a dual-column primary key: the article PMID (PubMed identification number) and the patient number (**Fig. 5-5**). The primary table ("demographics") contains the article PMID, patient number, and individual patient demographics, including the age of onset and gender. The disease classifications and affected genes and mutations (where identified) are contained in the "gene_disease" table, which uses the PMID and patient number as a foreign key to accommodate cases with multiple diagnoses. The "symptomology" table contains structured metadata on patient symptoms encoded using ICD-10 and ICD-11 and also uses the PMID and patient number as a foreign to accommodate multiple symptom entries per patient. The "bibliographic_info" table contains detailed provenance information and the "medical_content" table houses the MACCR metadata templates assembled for each patient. The bibliographic and medical content tables employ the article PMID and patient number as the primary key to prevent any double entries and maintain data integrity and fidelity. This allows data retrieval across all tables with simple left-join queries to link diseases, genes, and symptoms back to the patient in question and their demographic details. The design is easily extensible through instantiation of new tables sharing the PMID-patient key, allowing later incorporation of additional related data, such as ICD procedure codes or inclusion of SNOMED codes in the symptomology table. This data structure, with tight association across tables through a unique identifier, is better suited for an RDBMS such as MySQL as compared to a NoSQL solution, which is meant for data with no tabular relationships.

2.3.3 Flask API.

Flask and additional Python libraries form the MitoCases API, which handles all user requests for search, uploads, and downloads, effectively decoupling the front-end user interface from the back end to allow for better extensibility and modularization. Flask is a microframework for web



patient_demographics

PMID	gender	age_onset
21932011	Male	5 years
28295041	Female	3 years
20018511	Female	5 years

gene_disease

PMID	disease	gene	mutation
21932011	barth	TAZ	c.646G>A (p.Gly216Arg)
28295041	carnitine	SLC22A5	NA
20018511	complex_I	MT-ND5	NA

bibliographic_info

PMID	title	authors	journal	year	doi	…
21932011	Barth syndrome diagnosed in…	Takeda A, …	Eur. J. Pediatr.	2011	10.1007/s00431-011-1576-5	
28295041	Exome sequencing identifies…	Lahrouchi N, …	Eur J Hum Genet.	2017	10.1038/ejhg.2017.22	
20018511	A novel mitochondrial MTND5…	Alston CL, …	Neuromuscul Disord.	2010	10.1016/j.nmd.2009.10.010	

medical_content

PMID	demographics	location	family_history	medical_history	signs_symptoms	…
21932011	13-year-old boy	NA	no family history of BTHS…	male patient was born at 35 wks…	Mild hypotonia and a myopathic face…	
28295041	3-year-old girl	Morocco	two siblings dying young…	normal development until age 2…	fatigue, shortness of breath, pallor, …	
20018511	7 year old girl	NA	maternal aunt with tremor…	born in good condition at term…	mild metabolic acidosis and raised…	

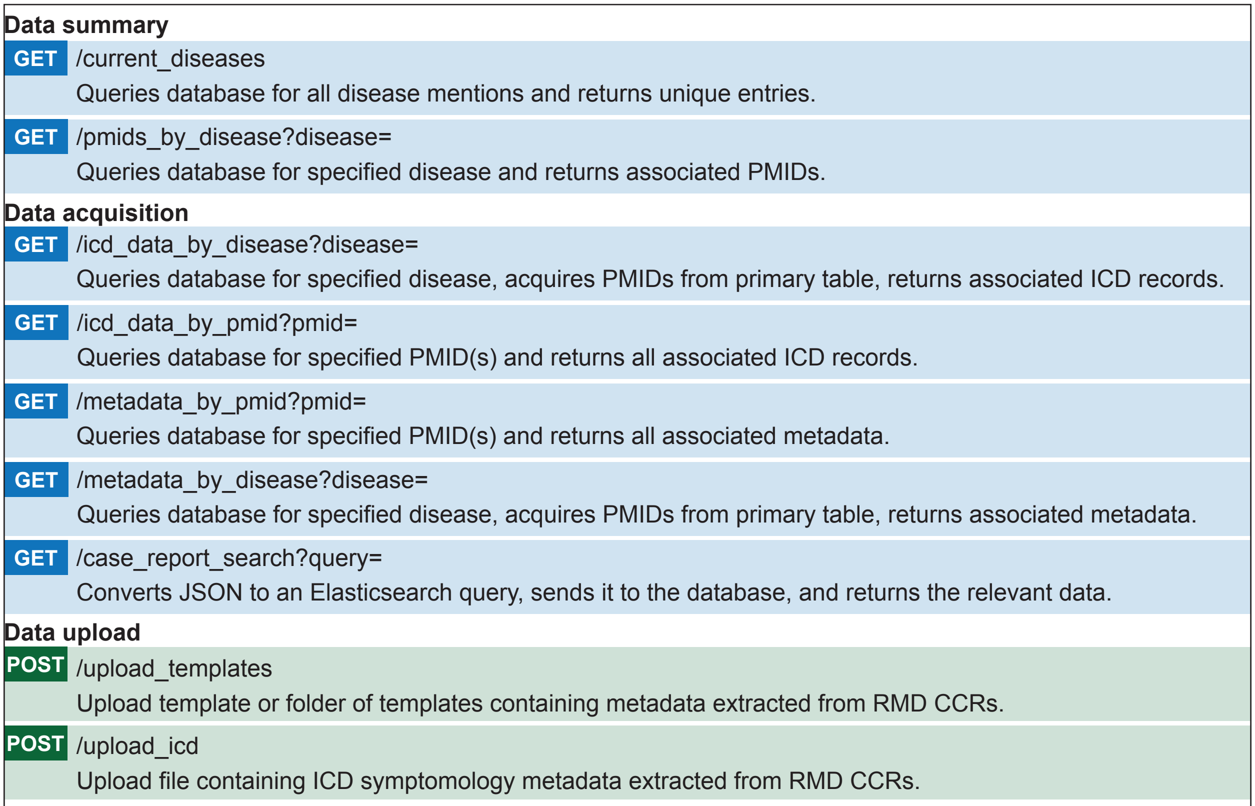
symptomology

PMID	icd11	icd11_desc	source_context
21932011	KB08.2	Congenital hypotonia	Mild hypotonia
21932011	5C52.2	Neutral lipid storage disease	lipid storage myopathy
21932011	BC44	Noncompaction cardiomyopathy	Isolated Noncompaction of Ventricular Myo…

Figure 5-5: MitoCases MySQL database schema. The data on MitoCases is contained across five tables: “patient_demographics”, “gene_disease”, “bibliographic_info”, “medical_content”, and “symptomology”. The PMID serves as the primary key in the patient_demographics, bibliographic_info, and medical_content tables to ensure that only one entry per PMID is ever accepted. For the gene_disease and symptomology tables, PMID is used instead as a foreign key such that multiple diseases and genetic associations can be entered per patient, as well as a comprehensive record of symptoms. The bibliographic_info and medical_content tables are shown here with only a sample of the total column headings contained within each; the complete tables contain all fields represented in the MACCR metadata template depicted in **Table 2**. Additional tables (not shown) contain a master database of ICD-10 and ICD-11 codes and their descriptions.

application development in Python and ships with limited functionality beyond its ability to serve and receive HTTP requests. Apache’s proxy module redirects all URLs with the base URL “mitocases.org/mito_api/” from the MediaWiki framework to Flask to handle all API endpoints (**Fig. 5-6**). The API provides language-agnostic programmatic access to all available data such that technically proficient users can employ any programming language that supports web requests to analyze the data in a controlled and easily accessible manner.

MitoCases depends on several different applications and software, which are effectively compartmentalized and incorporated using Flask. The Apache web server grants access to the Flask server, facilitated by additional modules including mod_wsgi, reverse_proxy, and the settings listed in the mito_api.conf file. External Python libraries including Elasticsearch, Elasticsearch_dsl, and requests allow interaction with search functionalities, and internal modules tie their functionality together for use as single functions. All database access is accomplished through the SQLalchemy Python library, allowing Flask to construct SQL queries in Python. Flask largely



Data summary

GET /current_diseases
Queries database for all disease mentions and returns unique entries.

GET /pmids_by_disease?disease=
Queries database for specified disease and returns associated PMIDs.

Data acquisition

GET /icd_data_by_disease?disease=
Queries database for specified disease, acquires PMIDs from primary table, returns associated ICD records.

GET /icd_data_by_pmid?pmid=
Queries database for specified PMID(s) and returns all associated ICD records.

GET /metadata_by_pmid?pmid=
Queries database for specified PMID(s) and returns all associated metadata.

GET /metadata_by_disease?disease=
Queries database for specified disease, acquires PMIDs from primary table, returns associated metadata.

GET /case_report_search?query=
Converts JSON to an Elasticsearch query, sends it to the database, and returns the relevant data.

Data upload

POST /upload_templates
Upload template or folder of templates containing metadata extracted from RMD CCRs.

POST /upload_icd
Upload file containing ICD symptomology metadata extracted from RMD CCRs.

Figure 5-6: Data API endpoints. Data summaries, downloads, uploads, and queries to the MySQL database are all handled through the Data API. MitoCases allows direct programmatic access via the Data API for GET and POST requests from advanced users.

serves as a connector between other services and software designed in-house. Using Flask as an intermediary furnishes a common mechanism to incorporate other useful software in a language that has low onboarding time, with features that are easy to use and require minimal knowledge of web services for extensibility and modularization.

Flask serves as the back-end API for MitoCases, handling downloads, uploads, and case report search by sending and receiving JSON files between the Flask server and other web applications, including MediaWiki, Elasticsearch, and MySQL. API endpoints are easily generated using the Flask decorator function, which imbues simple Python functions with additional functionality. For instance, the app.route decorator associates a MitoCases URL with a Python function, sending requests to that URL directly to the Python code. The process takes one line of code and a new response object accessible by the input function. The response object contains all the data passed when making the request, what type of request was sent, and the status of the request. Our expected user community of data science and text-mining researchers is more likely to possess some Python experience than web development experience, providing an easy workflow, rapid development time, and ease of extension for even beginners in Python.

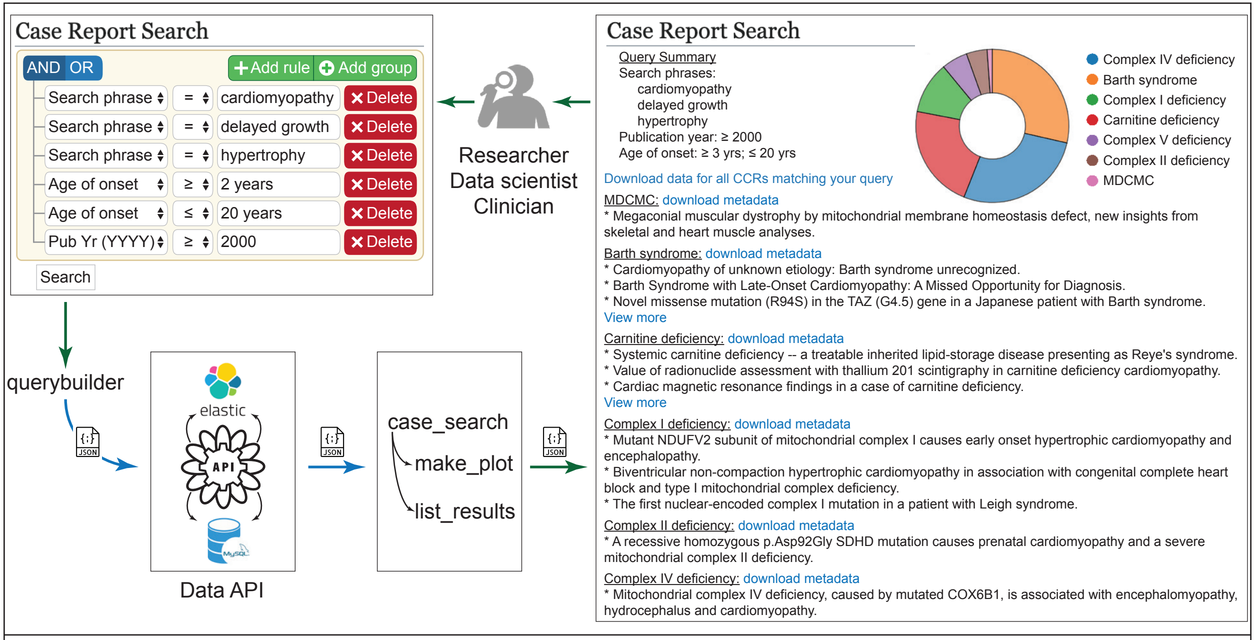
The Flask framework offers a simple and lightweight solution for uniting diverse applications in comparison to other options such as Django and web.py; most other options supply features and packages (“batteries included”) that are unnecessary for the use case, or implement their features in such a way that could present a barrier to entry for new developers. Django, for instance, is a large framework with native support for an admin page, form submission, and its own ORM (object relationship mapper), but these features are implemented in a complex manner that may not be clear to a novice Python user. Similarly, web.py has a range of features that are unnecessary for this implementation. URL routing is much more complex, requiring all URLs to be listed in a separate object, a separate class for each endpoint, and, within each class, a method for every type of HTTP request supported by the endpoint. Additionally, request information is implicitly incorporated into the function defined in each class, rather than being provided in a corresponding request object. In contrast, Flask provides a simple one-page tutorial and requires only a rudimentary understanding of HTTP requests for a novice developer to begin accessing the

API or extending MitoCases functionalities for any desired purpose. Furthermore, Flask benefits from the built-in app.route decorator for fast API endpoint creation through a single line of code.

2.3.4 Elasticsearch.

Elasticsearch facilitates MitoCases user queries across all available metadata housed in the database, providing users with the ability to retrieve case report metadata by disease, keyword, demographics, or any combination using Boolean logic in a powerful search engine. Elasticsearch employs a RESTful (REpresentational State Transfer) software architectural style which ensures reliable interoperability between computer systems across the internet by using a stateless protocol and standard operations. The inherent extensibility of the Elasticsearch RESTful API and full Python support through SQLalchemy allows easy integration with Flask for communication with other web services, as well as seamless integration with the underlying code on MitoCases. Search and storage services are decoupled by assigning the roles to Elasticsearch and MySQL, respectively, avoiding a full system failure in the event that the search function fails.

The MitoCases Case Report Search functionality allows users to build complex queries governed by Boolean logic to include or exclude all documents with a particular search phrase, or based on a disease, age range, gender, or publication date (**Fig. 5-7**). The search mechanism is composed of a querybuilder function, a plotting function for the results summary, and functions that build the section titles, results lists, and download links. The querybuilder Queries are verified for proper formatting by the querybuilder package and the user is notified to address errors or empty fields before submitting. Once it has been constructed and submitted by the user, querybuilder converts it to a JSON and sends it to Flask to process into an Elasticsearch query. The Elasticsearch query is sent through Flask to SQLalchemy, which constructs SQL queries in Python to send to the MySQL database. Flask retrieves the search results and generates usable information for front-end display functions using the case_report_search API endpoint. A summary of the results is displayed as a donut chart, categorized by disease, using the create_donut_chart function and d3.js. The list of search results, categorized by disease, are generated by the create_title_section function. Download links call jszip to load .zip files with metadata from the results, either in entirety or by disease, and filesaver.js enacts the download to the local machine.



Case Report Search

AND OR · +Add rule · ⊕ Add group

Search phrase	=	cardiomyopathy	✕ Delete
Search phrase	=	delayed growth	✕ Delete
Search phrase	=	hypertrophy	✕ Delete
Age of onset	≥	2 years	✕ Delete
Age of onset	≤	20 years	✕ Delete
Pub Yr (YYYY)	≥	2000	✕ Delete

Search

Case Report Search

Query Summary
Search phrases:
cardiomyopathy
delayed growth
hypertrophy
Publication year: ≥ 2000
Age of onset: ≥ 3 yrs; ≤ 20 yrs

Download data for all CCRs matching your query

MDCMC: download metadata
* Megaconial muscular dystrophy by mitochondrial membrane homeostasis defect, new insights from skeletal and heart muscle analyses.

Barth syndrome: download metadata
* Cardiomyopathy of unknown etiology: Barth syndrome unrecognized.
* Barth Syndrome with Late-Onset Cardiomyopathy: A Missed Opportunity for Diagnosis.
* Novel missense mutation (R94S) in the TAZ (G4.5) gene in a Japanese patient with Barth syndrome.
View more

Carnitine deficiency: download metadata
* Systemic carnitine deficiency -- a treatable inherited lipid-storage disease presenting as Reye's syndrome.
* Value of radionuclide assessment with thallium 201 scintigraphy in carnitine deficiency cardiomyopathy.
* Cardiac magnetic resonance findings in a case of carnitine deficiency.
View more

Complex I deficiency: download metadata
* Mutant NDUFV2 subunit of mitochondrial complex I causes early onset hypertrophic cardiomyopathy and encephalopathy.
* Biventricular non-compaction hypertrophic cardiomyopathy in association with congenital complete heart block and type I mitochondrial complex deficiency.
* The first nuclear-encoded complex I mutation in a patient with Leigh syndrome.

Complex II deficiency: download metadata
* A recessive homozygous p.Asp92Gly SDHD mutation causes prenatal cardiomyopathy and a severe mitochondrial complex II deficiency.

Complex IV deficiency: download metadata
* Mitochondrial complex IV deficiency, caused by mutated COX6B1, is associated with encephalomyopathy, hydrocephalus and cardiomyopathy.

Figure 5-7: MitoCases Case Report Search. MitoCases provides a powerful Case Report Search mechanism that allows users to design complex queries for cases of interest using keywords and phrases as well as filtering by disease, patient gender, age, and publication year. The user can define multiple groupings of search terms and filters, with AND/OR logical rules within and between the groups. The querybuilder module converts a valid query into a JSON object that is passed to Elasticsearch and MySQL through the Data API and on to the case_search module that displays the results. The user is presented with a summary of their search terms and filters and a donut plot summarizing the results, categorized by an interactive donut plot. The case reports matching the search terms are listed below, categorized by disease, with links to PubMed on each case report, links to download all available metadata for each disease, as well as a download link to acquire all available metadata across all diseases.

2.3.5 Downloads.

Downloads are furnished by the metadata_by_disease and icd_data_by_disease API endpoints to serve files. Users can acquire all available MitoCases data through the MediaWiki UI, which provides simple links to download all forms of metadata in .zip files. They may also use the search feature for a more limited set of reports related to their current study or patient. More advanced users may interact with the API directly to download all available data, retrieve specific subsets of the data, or construct queries of their own.

2.3.6 Uploads.

MitoCases provides the ability to contribute metadata on case reports relating to diseases already included as well as those not yet represented on the platform (**Fig. 5-8**). A simple user interface facilitates uploading a single metadata file or multiple files contained within a folder on the user's

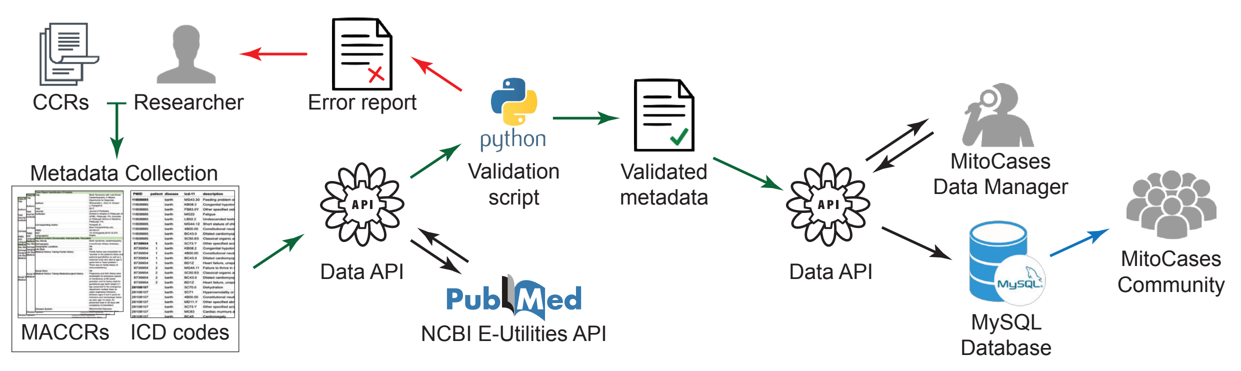


Figure 5-8: Metadata upload functionality for user contributions. Researchers can contribute their self-assembled metadata templates and symptomology tables on a disease of interest via the MitoCases upload functionality. Blank and example templates are provided along with detailed instructions to guide metadata creation, which can be submitted individually or in bulk uploads by folder. The MitoCases Data API queries both the NCBI PubMed E-Utilities API and DOI.org for the PMIDs and DOIs identified in the upload files, then employs a Python-based validation script to confirm that both identifiers match the title supplied and that there are no missing fields or invalid characters. When necessary, an error report is generated and sent back to the user with specific issues and links to the relevant instructions needed to remedy them. Upon successful validation, the templates are delivered through the Data API to the MitoCases server for approval by a data manager prior to its ultimate upload to the MySQL database, where it will be made accessible to the MitoCases community and other researchers and clinicians.

local machine. Detailed instructions, blank templates, and examples are provided for collecting metadata from individual case reports and generating the MACCRs templates as well as ICD-11 symptom metadata. Upon uploading, the Flask API delivers the templates to a Python validation script that will check that all fields are filled, no errant or unsupported characters exist, and that proper line separators are used. The script also communicates with the NCBI Entrez Programming Utilities (E-Utils) API [31] and the DOI.org resolution system (http://doi.org/) [32] to confirm that the submitted URLs, DOIs, PMIDs, and titles are correctly associated with one another. In the event of any errors, a report is delivered to the user to alert them to the specific fields that require attention with references to the relevant instructions that should be revisited and adhered to for a successful upload and inclusion in the dataset. Upon successfully passing the validation step, the user is notified that their submission has been delivered for manager review, and their e-mail address is requested so that they may be notified upon its ultimate approval. User accounts with fast-tracked approval will be provided for those who consistently submit valid and useful metadata. Upon data validation, the metadata is delivered to the cloud platform and stored pending approval by the MitoCases data managers. Once the data has been approved, it is routed through the API to the MySQL database and distributed across the relevant tables.

2.4 Results and Discussion

2.4.1 Overview and summary of MitoCases metadata.

MitoCases houses a large collection of demographics, medical, and genetic metadata, all of which is available on the platform to browse or download. The metadata is also searchable using the Case Report Search feature, providing users the ability to assemble complex queries to identify cases of interest. Across all 384 curated CCRs, 4,561 instances of 952 unique ICD-10 codes were identified for the individual patients. Each of these codes, their descriptions, and the text content in the CCRs from which they were extracted is indexed by the Case Report Search functionality, providing a straightforward method for identifying case reports with a particular set of symptoms. Key characteristics of all metadata are summarized on the home page as well as on each disease page. This includes distributions reflecting the patients' age of onset, gender, genetic information, as well as a geographic distribution of publication localities across the world. A summary of the metadata contents is provided in **Table 5-3**.

		Age (in yrs, unless noted)				**Gender (%)**		**ICD-10 codes**		
Disease	# CCRs	Median	Mean	Min	Max	Male	Female	Total	Unique	Top gene
Barth	30	7.5mo	5	< 1 mo	30	97	3	340	95	*TAZ*
Carnitine	45	4	11.4	< 1 mo	68	48	52	366	153	*SLC22A5*
Complex I	84	1	7.2	< 1 mo	64	54	46	1463	482	*MT-TL1*
Complex II	12	7 mo	6.4	< 1 mo	33	62	38	141	69	*SDHA*
Complex III	29	10 mo	8.7	< 1 mo	57	57	43	347	141	*MT-CYB*
Complex IV	170	13 mo	7.7	< 1 mo	62	48	52	1715	503	*SURF1*
Complex V	7	< 1 mo	3.8	< 1 mo	22	43	57	73	39	*TMEM70*
MDCMC	7	2.5	6.1	< 1 mo	16	71	29	116	64	*CHKB*
Overall	**384**	**1.1 yr**	**7.6 yr**	**< 1 mo**	**68 yr**	**55**	**45**	**4,561**	**952**	***TAZ***

Table 5-3. Summary statistics on MitoCases metadata contents. The MitoCases platform contains detailed metadata on 384 CCRs relating to 8 different RMDs, including Barth syndrome, carnitine deficiency, deficiencies in complexes I-V of the electron transport chain, and congenital megaconial-type muscular dystrophy (MDCMC). Patient ages range from newborns under 1 month of age up to patients in their 60's, though the diseases disproportionately affect infants and young children. The gender distribution is quite balanced, aside from Barth syndrome, which occurs almost exclusively in males. 4,561 instances of nearly 952 unique ICD-10 codes were used to codify the symptomology of these patients. The genetic etiology of Barth syndrome is entirely attributed to mutations in the tafazzin-encoding *TAZ* gene and only *CHKB* mutations were reported in the MDCMC cases. The genetic basis is more varied for the rest of the diseases investigated here, shown in **Figure 5-10**.

2.4.2 Comparing representation of patient gender.

The gender reported for each patient is included in a top-level demographics table within the database, along with the genetic etiology of their disease. Frequently, gender is missing entirely or both male and female MeSH terms will be included for a given CCR on PubMed. In many cases, this appears to be due to a mention of other patients with a similar presentation that are discussed within the report. However, this presents difficulties when attempting to limit search results to one gender or the other. In the MitoCases dataset, metadata is derived only for the primary patient of the report, removing any ambiguity and facilitating a better search mechanism. The gender distribution across all diseases is presented on the main page of the MitoCases platform, and for the individual diseases on each disease page.

2.4.3 Comparing age categories and their utility.

The age categories available in the MeSH hierarchy are appropriate and aligned with clusters employed in the clinical community, but it is difficult to make use of these terms when searching for case reports. Inputting numerical values does not call up the relevant MeSH terms. Similar issues of ambiguity as in the gender tags also occur, with multiple, contradictory age groups listed for a single report. These issues further limit users' ability to search for cases relevant to specific age ranges. In contrast, the MitoCases platform contains specific ages for each patient, with newborns classified as < 1 month, 1 month resolution for patients below 2 years of age, and 1 year resolution beyond. The ages are presented in common written form on the site for ease of understanding by our users (i.e., "Under 1 month", "13 months", "25 years", etc.), while the back-end data handling treats months as fractions of years to ensure the appropriate search results are retrieved. The age distribution across all cases is included on the main page and also by disease on each individual page. Similar to the MeSH system, we chose to employ 10 age bins in our distribution presentations, including Newborn (<2 months), Infant (<2 years), Preschool child (3–5 years), Child (6–13 years), Adolescent (14–18 years), Young adult (19–24 years), Adult (25–48 years), Middle aged (49–64 years), Aged (65–78 years), and Elderly (79–98 years).

2.4.4 Comparing representation of patient symptomology.

We analyzed PubMed MeSH terms in comparison to ICD codes in Barth syndrome CCRs to evaluate the degree of additional information contained on MitoCases (**Fig. 5-9**). Excluding demographic terms (age, sex), the top MeSH terms identified in the 30 Barth Syndrome CCRs were “syndrome” (n=17), “transcription factors” (n=15), “cardiomyopathy, dilated” (n=13), “Barth syndrome” (n=10), and “neutropenia” (n=10). “Cardiomyopathy, dilated”, “neutropenia”, and “cardiomyopathies” are the only three symptoms identified in the top 10 MeSH terms. While they are indeed key features of the disease, the number of cases tagged with MeSH terms for those symptoms does not cover all of the cases in which they are discussed. Interestingly, only one third of the these CCRs are tagged with the proper MeSH term for “Barth syndrome”. Among 30 CCRs on individual patients with Barth syndrome, we extracted 340 instances of 95 unique ICD-10 codes and 332 instances of 123 unique ICD-11 codes, in addition to the rest of the detailed medical and demographic content contained in the MACCRs. The top ICD codes identified in the 30 Barth syndrome CCRs are “constitutional neutropenia” (n=25), “cardiomegaly” (n=23), “dilated cardiomyopathy” (n=18), “congenital hypotonia” (n=17), and “classical organic aciduria”

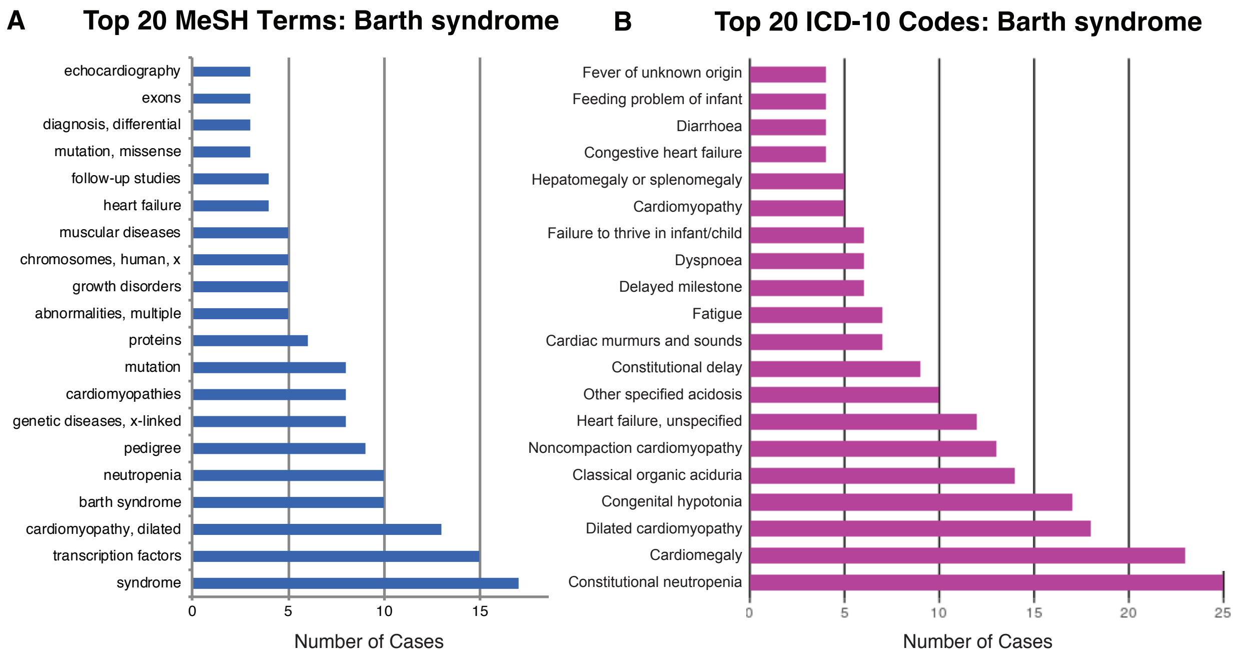


Figure 5-9. Comparing MeSH terms and ICD-11 codes for Barth syndrome CCRs. (A) Excluding demographic terms (age, gender), the top MeSH terms identified in the 30 Barth Syndrome CCRs were “syndrome”, “transcription factors”, and “cardiomyopathy, dilated”. Interestingly, only one third of the these CCRs have the proper MeSH term for “Barth syndrome”. **(B)** The top ICD-10 codes identified for the same reports were “constitutional neutropenia”, “cardiomegaly”, “dilated cardiomyopathy”, “congenital hypotonia”, and “classical organic aciduria”, followed by other forms of cardiomyopathy, heart failure, acidosis, and developmental issues. ICD annotation accurately represents the primary features of this disease and, with an effective query mechanism, improves case report indexing and findability.

(n=14), followed by other forms of cardiomyopathy, heart failure, acidosis, and developmental issues. These are the top symptoms seen in a classical presentation of Barth syndrome and they are representative of the symptoms discussed in each case. Codifying symptomology with ICD codes, therefore, provides more granular detail of each case and facilitates a more effective mechanism for physicians and biomedical researchers to identify cases of interest.

2.4.5 Comparing representation of genetic information.

We evaluated the top genes identified in the metadata from 384 CCRs on 8 selected RMDs to explore the range of genetic etiologies across diseases (**Fig. 5-10**). The most frequently identified genetic mutations across all 8 diseases were in *TAZ* (n=25), which encodes the tafazzin protein altered in Barth Syndrome, large mt-DNA deletions (n=13), found primarily in complex IV deficiency, and mutations in *SLC22A5* (n=13), which encodes the OCTN2 protein responsible for carnitine transport and is implicated in carnitine deficiency. This provides detail not otherwise available through a PubMed search. In the collection of Barth syndrome CCRs, for instance, there are several MeSH terms that allude to genetic information, including "pedigree", "genetic

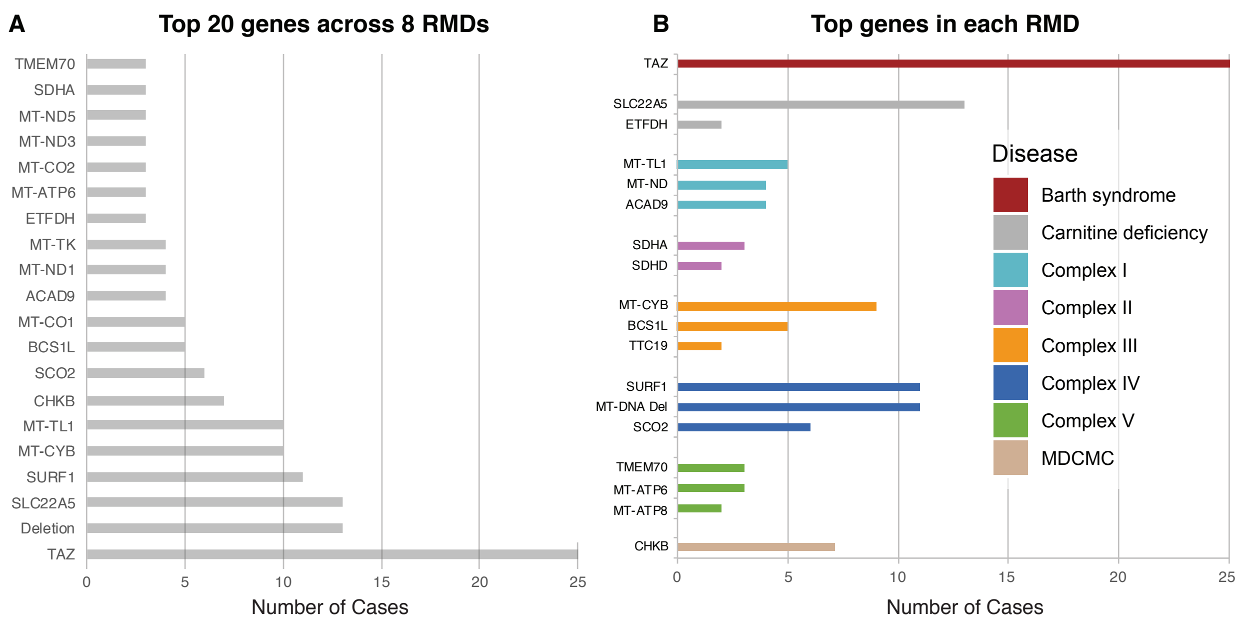
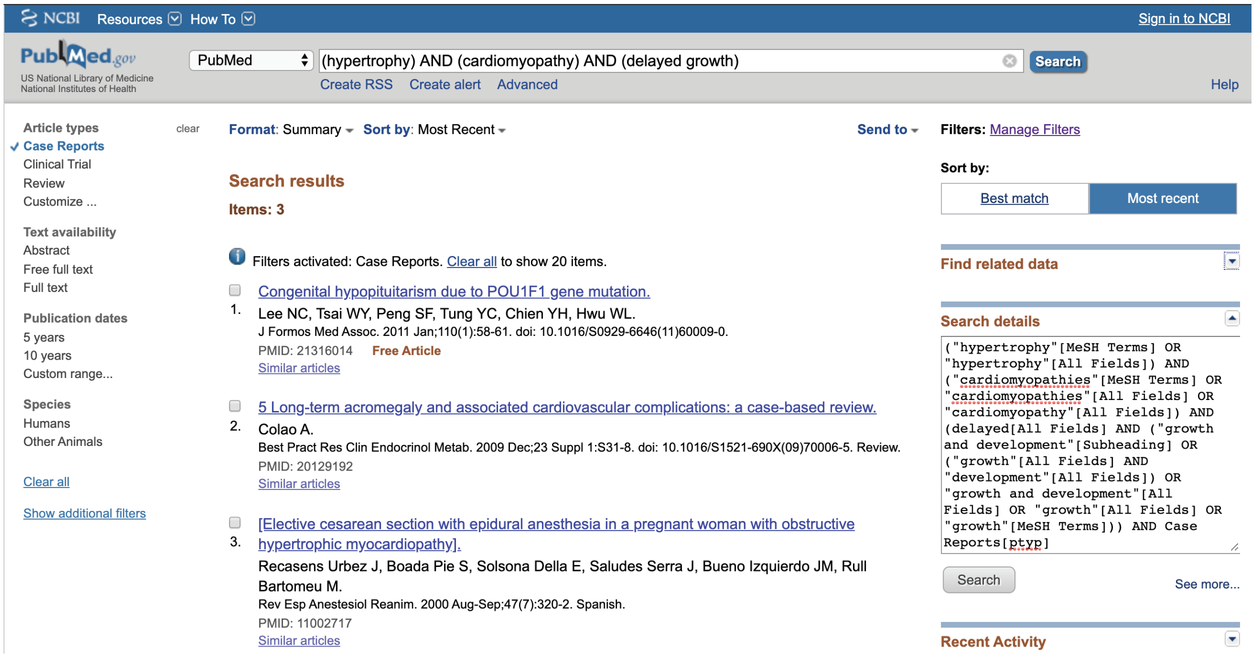


Figure 5-10. Genetic etiologies in selected mitochondrial diseases. The MitoCases dataset contains genetic information for each patient case report, where available. **(A)** The most frequently identified genetic mutations across 7 rare mitochondrial diseases were in *TAZ* (N=25), which encodes the tafazzin protein altered in Barth syndrome, large mt-DNA deletions (N=13) found primarily in complex IV deficiency, and *SLC22A5* (N=13), which encodes the OCTN2 protein responsible for carnitine transport and is implicated in carnitine deficiency. The top genes identified in each disease are displayed in panel **(B).**

diseases, x-linked", "mutation", "chromosomes, human, x", "mutation, missense", and "exons". All of these are certainly related to the disease and the causative *TAZ* mutations, but they do not provide much detail or utility in terms of indexing a case of Barth syndrome. Five of the Barth syndrome CCRs do not contain genetic information or identify a mutation in the *TAZ* gene; each of these were published prior to the identification of the genetic etiology of the disease or before genetic sequencing was in common usage for disease diagnosis.

2.4.6 Comparing search functionalities.

MitoCases Elasticsearch system handles complex queries constructed by the user and accommodates Boolean logic between and within groups of search terms. Users can filter CCRs by demographics and publication year, and they can construct search terms that probe the underlying database containing genetics, symptoms, and other medical and demographic details. This provides the ability to create a complex query with granular detail and elevates the current paradigm of searching by MeSH terms, abstracts, and titles. We compared the Case Report Search function on MitoCases to that of PubMed to evaluate whether the detailed metadata provides additional utility relative to searching across titles, abstracts, and MeSH terms. It appears that the added metadata enables MitoCases to significantly outstrip the capabilities of PubMed, which typically returns limited results for a search of more than two symptoms. For example, PubMed returns only three results when searching across the two million available case reports with the keywords "hypertrophy", "cardiomyopathy", and "delayed growth" (interpreted by PubMed as: ("hypertrophy"[MeSH Terms] OR "hypertrophy"[All Fields]) AND ("cardiomyopathies"[MeSH Terms] OR "cardiomyopathies"[All Fields] OR "cardiomyopathy"[All Fields]) AND (delayed[All Fields] AND ("growth and development"[Subheading] OR ("growth"[All Fields] AND "development"[All Fields]) OR "growth and development"[All Fields] OR "growth"[All Fields] OR "growth"[MeSH Terms])) AND Case Reports[ptyp]). The PubMed search is shown in **Fig. 5-11**. We ran the same query on MitoCases with further constraints for a specified age of onset between 3 and 20 years old with a publication date after the year 2000, as depicted in **Fig. 5-7**. From the limited dataset of only 384 CCRs available on MitoCases, the Case Report Search feature returns 31 results across 7 different diseases. Results are available for download *en masse* or for each individual disease, and PubMed links are provided for direct access to the original reports.



NCBI Resources How To Sign in to NCBI

PubMed.gov US National Library of Medicine National Institutes of Health

PubMed (hypertrophy) AND (cardiomyopathy) AND (delayed growth) Search

Create RSS Create alert Advanced Help

Article types clear
✓ Case Reports
Clinical Trial
Review
Customize ...

Text availability
Abstract
Free full text
Full text

Publication dates
5 years
10 years
Custom range...

Species
Humans
Other Animals

Clear all

Show additional filters

Format: Summary Sort by: Most Recent Send to

Search results

Items: 3

Filters activated: Case Reports. Clear all to show 20 items.

□ 1. Congenital hypopituitarism due to POU1F1 gene mutation.
Lee NC, Tsai WY, Peng SF, Tung YC, Chien YH, Hwu WL.
J Formos Med Assoc. 2011 Jan;110(1):58-61. doi: 10.1016/S0929-6646(11)60009-0.
PMID: 21316014 Free Article
Similar articles

□ 2. 5 Long-term acromegaly and associated cardiovascular complications: a case-based review.
Colao A.
Best Pract Res Clin Endocrinol Metab. 2009 Dec;23 Suppl 1:S31-8. doi: 10.1016/S1521-690X(09)70006-5. Review.
PMID: 20129192
Similar articles

□ 3. [Elective cesarean section with epidural anesthesia in a pregnant woman with obstructive hypertrophic myocardiopathy].
Recasens Urbez J, Boada Pie S, Solsona Della E, Saludes Serra J, Bueno Izquierdo JM, Rull Bartomeu M.
Rev Esp Anestesiol Reanim. 2000 Aug-Sep;47(7):320-2. Spanish.
PMID: 11002717
Similar articles

Filters: Manage Filters

Sort by:
Best match Most recent

Find related data

Search details

```
("hypertrophy"[MeSH Terms] OR
"hypertrophy"[All Fields]) AND
("cardiomyopathies"[MeSH Terms] OR
"cardiomyopathies"[All Fields] OR
"cardiomyopathy"[All Fields]) AND
(delayed[All Fields] AND ("growth
and development"[Subheading] OR
("growth"[All Fields] AND
"development"[All Fields]) OR
"growth and development"[All
Fields] OR "growth"[All Fields] OR
"growth"[MeSH Terms])) AND Case
Reports[ptyp]
```

Search See more...

Recent Activity

Figure 5-11: PubMed Case Report Search Results. PubMed returns only three results when searching across the over two million available case reports with the keywords “hypertrophy”, “cardiomyopathy”, and “delayed growth”. For comparison, the same search was conducted using the MitoCases Case Report feature, with the added constraints for cases with an age of onset between 3 and 20 years old and a publication date after the year 2000, as depicted in **Fig. 5-7**. MitoCases returns 31 results across 7 different diseases from the limited dataset of metadata on only 384 CCRs, representing a demonstrable improvement over the current paradigm of indexing by title, abstract, and MeSH terms.

2.5 Key Stakeholders and Use Cases

The MitoCases platform is designed for use by several key stakeholder groups, including clinicians, clinical researchers and other research scientists, as well as data scientist and text-mining researchers. Here, we describe these key user groups and our currently envisioned use cases for the platform.

2.5.1 Clinicians.

MitoCases is of particular utility for clinicians seeing patients with mitochondrial disease or suspected mitochondrial disease, particularly those cases with complex presentations and sparsely available clinical literature. The platform was created initially to address the lack of an effective system for sourcing case reports on RMDs. The metadata provided on the current collection of 384 reports from 8 RMDs contains detailed case information relating specifically to the primary patient discussed in the report, including gender, age of onset, genetic information (where available), and a wide range of medical information from family history, medical history, signs and symptoms, laboratory results, among others (**Table 5-2**). The Case Report Search function provides a mechanism for clinicians to search for reports within any age range, filter by gender, or enter search terms relating to genetics, and medical information contained within the reports. Patient symptoms have also been manually identified and codified by ICD-10 and ICD-11, and the search function probes these codes, their descriptions, as well as the contextual content from the report itself.

This extensive database of carefully extracted metadata from a large collection of reports on a range of RMDs, paired with the powerful Elasticsearch mechanism, provides clinicians the ability to retrieve reports that are actually relevant to their patient's specific symptoms, drug treatments, age, gender, genetic mutations, and other key medical concepts. Where PubMed searches give zero results, MitoCases consistently returns a collection of relevant reports, categorized by disease, with links to the article on PubMed for more in-depth reading, and download links to acquire the relevant metadata on those cases. MitoCases furnishes a sorely needed and capable search tool for clinicians to find relevant case reports with greatly improved accuracy and reliable results, reducing the time spent on literature curation so that they can develop a plan of action for differential diagnosis and effective treatment.

2.5.2 Research scientists and clinical researchers.

Much like clinicians, research scientists and clinical researchers are severely limited in their ability to curate a comprehensive set of case reports related to a disease of interest. In investigating the potential translational relevance of a signaling pathway, gene, or drug, researchers may turn to the clinical literature to determine whether there have been similar applications, and what kind of benefits and risks have already been discovered. Without an effective mechanism to conduct a more comprehensive search within the literature, they will spend a disproportionate amount of time curating reports and weeding out irrelevant results and false positives.

MitoCases allows clinicians and clinical researchers to create *in silico* cohorts of disease. Rare diseases present a particularly challenging area in conducting effective clinical trials because of the scarcity of patients with those diseases. By leveraging the cumulative power of large collections of reports, better statistical models can be generated, and more reliable inferences can be made. For instance, when a new drug designed to treat RMDs is released, there is a limited patient population available to determine its effectiveness and safety. As these patients are treated and their clinical support teams monitor their progress, case reports are eventually published and provide insight into whether this might be a good treatment regimen for future patients with similar presentations. As more publications are released, the magnitude of benefits and risks can be weighed by creating *in silico* cohorts from the patients described in these reports. These efforts do not require patient consent, nor do they deal with any HIPAA (Health Insurance Portability and Accountability Act of 1996) considerations, presenting an opportunity for researchers to conduct studies across the patient population throughout history.

Additionally, researchers can contribute to MitoCases by uploading metadata on a disease of interest, adding to the overall utility of the platform and presenting opportunities for collaboration and publications resulting from their submissions. The upload page has a simple user interface with detailed instructions for curation, filling the metadata template, and identifying symptoms and their corresponding ICD-11 codes. User submissions can include single or multiple metadata entries and are immediately subjected to a validation script that will alert them to any errors and direct them to the relevant information in the instructions to improve their dataset prior to resubmission (technical specifications detailed in *2.3.6 Uploads*). The data will be subject to verification by the MitoCases team, and the user will be notified upon its ultimate approval and integration into the database. Regular users who consistently submit valid metadata will be offered fast-tracked approval status and will see their submissions immediately included in the MitoCases database. Data contribution will be credited to the uploader and any future publications utilizing the data will attribute credit accordingly. By submitting additional data to the MitoCases platform, researchers will be able to utilize search features to find other similar cases among an expanding collection of diseases, and further downstream analysis methods and features that will be continually added to the platform.

2.5.3 Data scientists and text-mining researchers.

Clinical informatics is a rapidly growing field that is amplifying the abilities of clinicians and researchers to access and analyze information across a broad range of scientific disciplines. Text-mining has the potential to revolutionize the way we conduct curation efforts and handle electronic health records (EHRs). However, many of the current text-mining approaches attempting to accurately model clinical language or consistently identify and categorize relevant clinical terms are limited by the unstructured nature of text data. Without access to large datasets of structured text data, it is difficult to build these models and for clinical text-mining to reach its full potential. The MitoCases dataset provides access to carefully curated and manually generated metadata from domain experts on a range of RMDs. This is a valuable resource for modeling clinical language and gives researchers freely accessible, well indexed, structured text data.

Text-mining researchers can download the entirety of the MitoCases dataset for use in modeling clinical language and communication. The metadata template employed here was developed with the input of several clinicians who carefully considered each included field, establishing a structure for case report communications that does not otherwise exist. The variety of communication patterns across all included case reports provides a good representation of the multitude of ways that a clinician might write about lab results, signs and symptoms, family history, and therapeutic approaches. With this structure imposed upon the reports, data scientists can begin to model clinical language, develop improved named entity recognition (NER) models [33, 34], and conduct downstream text-mining analyses. As the collection expands and a wider range of diseases are represented on the platform, the dataset will become ever more useful to data scientists who require very large bodies of structured text data on which to train machine learning algorithms. Currently, an automated system for identifying symptoms via NER and mapping them to ICD codes is in development. Once it is implemented and integrated on the MitoCases platform, the amount of data available will increase exponentially and provide a much greater representation of RMD symptomology as a whole.

The MitoCases “Use cases” page houses potential analysis approaches for researchers of all types. To continue expanding functionality, submissions from text-mining researchers are welcomed for incorporation on the platform, providing access for clinician and basic researchers to conduct more advanced studies of their data. These analysis methods will be cited and attributed to the author, expanding their ability to reach domain scientists and amplify their usage statistics. Furthermore, testing new models and tools on text data from CCRs provides a solid method for rigor-testing against real-world challenges in the biomedical field.

Conclusions

We identified the need for a more FAIR and structured representation of mitochondrial biology and mitochondrial disease and set out to improve these deficiencies by contributing to mitochondrial Gene Wiki pages, enforcing structure on RMD CCRs with a standardized metadata template, and creating the MitoCases RMD Knowledge Platform to house the structured data. In addition, we used these efforts as an educational platform for training students pursuing biomedical degrees and pre-medical studies in mitochondrial physiology in health and disease, as well as important research techniques for literature curation, review, knowledge extraction, and scientific communication.

Our extensive contributions to Gene Wiki pages on mitochondrial genes and proteins have elevated the available content with significantly more references, Semantic Wiki Links to key relations, as well as linkages to key knowledgebases and informatics resources. This serves the biomedical community and the public at large by communicating complex concepts and current research in an easily interpreted manner and on a heavily utilized platform. Researchers and students studying a particular gene or protein will frequently initiate their search on Wikipedia, where one can find a wealth of information neatly summarized and heavily cited. This provides a useful starting point and also increases the exposure of the references used on the pages to a much wider audience. Vast quantities of structured information are also stored on Wikidata, displayed in the Gene Wiki infobox on individual gene and protein pages, and is represented through Semantic Wiki Links. This structured data provides a mechanism by which data scientists can begin studying relationships between proteins, genes, diseases, and drugs that are represented on Wikipedia. Because of the vibrant community, the information is reliably accurate and always well cited, increasing its utility tremendously. Patients with rare mitochondrial diseases or their family members, many of whom have no scientific background, also use Wikipedia to begin learning about their diagnosis and to find information about relevant societies and patient groups. This is a critical resource for making knowledge available to many communities, and we view the mitochondrial representation as an important subset of the overall resource.

Clinical case reports are widely used among clinicians and medical students, but their inherent lack of structure and limited mechanisms for effective indexing and categorization hinders the discovery of relevant cases, particularly for rare diseases. These deficiencies also limit the success of text-mining and information extraction applications. We recognized the need for structured, machine-readable metadata in RMD CCRs and generated a large dataset using our MACCRs standardized metadata template [24, 25]. We also codified the symptomology of each patient in each CCR with the ICD-10 [26] and ICD-11 [27] clinical coding systems employed in electronic health records (EHRs) by clinicians and health care professionals around the world [35]. These data are housed on the MitoCases platform (http://mitocases.org/), providing clinicians with a powerful tool to more effectively discover case reports of patients with similar presentations to one of their own. Similarly, researchers and clinician investigators can use the tool to construct *in silico* cohorts of disease, study drug interactions and treatment paradigms, and design studies with translational value in mind. The MitoCases platform also accepts metadata uploads from researchers and physician investigators interested in contributing their expertise.

Ultimately, we aim to achieve a complete representation of all 58 rare mitochondrial diseases in the dataset, in addition to as-yet undiscovered mitochondrial diseases. Data scientists interested in text-mining and NLP can use the structured data to develop named entity extraction tools and train machine learning algorithms for modeling clinical language. Future directions entail development of NLP and NER pipelines to automate metadata extraction of patient demographics, symptoms, drug treatments, and genetic etiology on the full corpus of RMD CCRs. The MitoCases platform may also provide a useful example on which to model additional resources dedicated to the structured representation of clinical case information across a much wider range of diseases. As a fully open-source and FAIR resource, all underlying source code is available and free to use, in part or in its entirety.

References:

1. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J.G., Groth, P., Goble, C., Grethe, J.S., Heringa, J., 't Hoen, P.A.C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.-A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. and Mons, B. (2016). "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific Data*. 3: 160018. doi:10.1038/sdata.2016.18.

2. Mons, B., Neylon, C., Velterop, J., Dumontier, M., da Silva Santos, L.O.B. and Wilkinson, M.D. (2017). "Cloudy, increasingly FAIR; revisiting the FAIR Data guiding principles for the European Open Science Cloud." *Information Services & Use*. 37: 49-56.

3. Huss, J.W., 3rd, Orozco, C., Goodale, J., Wu, C., Batalov, S., Vickers, T.J., Valafar, F. and Su, A.I. (2008). "A gene wiki for community annotation of gene function." *PLoS Biol*. 6: e175. PMID:18613750. doi:10.1371/journal.pbio.0060175.

4. Maglott, D., Ostell, J., Pruitt, K.D. and Tatusova, T. (2007). "Entrez Gene: gene-centered information at NCBI." *Nucleic Acids Res*. 35: D26-31. PMID:17148475. doi:10.1093/nar/gkl993.

5. Huss, J.W., 3rd, Lindenbaum, P., Martone, M., Roberts, D., Pizarro, A., Valafar, F., Hogenesch, J.B. and Su, A.I. (2010). "The Gene Wiki: community intelligence applied to human gene annotation." *Nucleic Acids Res*. 38: D633-9. PMID:19755503. doi:10.1093/nar/gkp760.

6. Burgstaller-Muehlbacher, S., Waagmeester, A., Mitraka, E., Turner, J., Putman, T., Leong, J., Naik, C., Pavlidis, P., Schriml, L., Good, B.M. and Su, A.I. (2016). "Wikidata as a semantic framework for the Gene Wiki initiative." *Database : the journal of biological databases and curation*. 2016: baw015. PMID:26989148. doi:10.1093/database/baw015.

7. Good, B.M., Clarke, E.L., Loguercio, S. and Su, A.I. (2012). "Building a biomedical semantic network in Wikipedia with Semantic Wiki Links." *Database : the journal of biological databases and curation*. 2012: bar060-bar060. PMID:22434829. doi:10.1093/database/bar060.

8. Su, A.I. and Hogenesch, J.B. (2007). "Power-law-like distributions in biomedical publications and research funding." *Genome Biol*. 8: 404. PMID:17472739. doi:10.1186/gb-2007-8-4-404.

9. Lau, E., Cao, Q., Ng, D.C., Bleakley, B.J., Dincer, T.U., Bot, B.M., Wang, D., Liem, D.A., Lam, M.P., Ge, J. and Ping, P. (2016). "A large dataset of protein dynamics in the mammalian heart proteome." *Sci Data*. 3: 160015. PMID:26977904. doi:10.1038/sdata.2016.15.

10. Lotz, C., Lin, A.J., Black, C.M., Zhang, J., Lau, E., Deng, N., Wang, Y., Zong, N.C., Choi, J.H., Xu, T., Liem, D.A., Korge, P., Weiss, J.N., Hermjakob, H., Yates, J.R., 3rd, Apweiler, R. and Ping, P. (2014). "Characterization, design, and function of the mitochondrial proteome: from organs to organisms." *J Proteome Res*. 13: 433-46. PMID:24070373. doi:10.1021/pr400539j.

11. Raitman, R., Augar, N. and Zhou, W. (2005). "Employing wikis for online collaboration in the e-learning environment: Case study," Third International Conference on Information Technology and Applications (ICITA'05). IEEE, pp. 142-146.

12. Konieczny, P. (2012). "Wikis and Wikipedia as a teaching tool: Five years later."

13. Parker, K. and Chao, J. (2007). "Wiki as a teaching tool." *Interdisciplinary Journal of e-learning and Learning Objects*. 3: 57-72.

14. The UniProt Consortium (2017). “UniProt: the universal protein knowledgebase.” *Nucleic Acids Research*. 45: D158-D169. doi:10.1093/nar/gkw1099.

15. Zong, N.C., Li, H., Li, H., Lam, M.P., Jimenez, R.C., Kim, C.S., Deng, N., Kim, A.K., Choi, J.H., Zelaya, I., Liem, D., Meyer, D., Odeberg, J., Fang, C., Lu, H.J., Xu, T., Weiss, J., Duan, H., Uhlen, M., Yates, J.R., 3rd, Apweiler, R., Ge, J., Hermjakob, H. and Ping, P. (2013). “Integration of cardiac proteome biology and medicine by a specialized knowledgebase.” *Circ Res*. 113: 1043-53. PMID:23965338. doi:10.1161/CIRCRESAHA.113.301151.

16. Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., Milacic, M., Roca, C.D., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Viteri, G., Weiser, J., Wu, G., Stein, L., Hermjakob, H. and D’Eustachio, P. (2018). “The Reactome Pathway Knowledgebase.” *Nucleic acids research*. 46: D649-D655. doi:10.1093/nar/gkx1132.

17. Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D. and Apweiler, R. (2004). “IntAct: an open source molecular interaction database.” *Nucleic Acids Res*. 32: D452-5. PMID:14681455. doi:10.1093/nar/gkh052.

18. Tsueng, G., Good, B.M., Ping, P., Golemis, E., Hanukoglu, I., van Wijnen, A.J. and Su, A.I. (2016). “Gene Wiki Reviews-Raising the quality and accessibility of information about the human genome.” *Gene*. 592: 235-8. PMID:27150585. doi:10.1016/j.gene.2016.04.053.

19. Wang, J., Lee, J., Liem, D. and Ping, P. (2017). “HSPA5 Gene encoding Hsp70 chaperone BiP in the endoplasmic reticulum.” *Gene*. 618: 14-23. PMID:28286085. doi:10.1016/j.gene.2017.03.005.

20. Garlid, A.O., Schaffer, C.T., Kim, J., Bhatt, H., Guevara-Gonzalez, V. and Ping, P. (2019). “*TAZ* encodes tafazzin, a transacylase essential for cardiolipin formation and central to the etiology of Barth syndrome.” *Gene*. doi:(in press).

21. 107th United States Congress (2002). “H.R 4013: Rare Diseases Act of 2002.” Public Law 107–280.

22. Liu, K., Peng, S., Wu, J., Zhai, C., Mamitsuka, H. and Zhu, S. (2015). “MeSHLabeler: improving the accuracy of large-scale MeSH indexing by integrating diverse evidence.” *Bioinformatics*. 31: i339-i347. doi:10.1093/bioinformatics/btv237.

23. Mork, J.G., Jimeno-Yepes, A. and Aronson, A.R. (2013). “The NLM Medical Text Indexer System for Indexing Biomedical Literature,” BioASQ Workshop.

24. Caufield, J.H., Liem, D.A., Garlid, A.O., Zhou, Y., Watson, K., Bui, A.A.T., Wang, W. and Ping, P. (2018). “A Metadata Extraction Approach for Clinical Case Reports to Enable Advanced Understanding of Biomedical Concepts.” *J Vis Exp*. PMID:30295669. doi:10.3791/58392.

25. Caufield, J.H., Zhou, Y., Garlid, A.O., Setty, S.P., Liem, D.A., Cao, Q., Lee, J.M., Murali, S., Spendlove, S., Wang, W., Zhang, L., Sun, Y., Bui, A., Hermjakob, H., Watson, K.E. and Ping, P. (2018). “A reference set of curated biomedical data and metadata from clinical case reports.” *Sci Data*. 5: 180258. PMID:30457569. doi:10.1038/sdata.2018.258.

26. World Health Organization (1992). “International classification of diseases and related health problems, 10th revision.” Geneva.

27. World Health Organization (2018). “International classification of diseases and related health problems, 11th revision.” Geneva.

28. Caufield, J.H., Zhou, Y., Garlid, A.O., Setty, S.P., Liem, D.A., Cao, Q., Lee, J.M., Murali, S.,

Spendlove, S., Wang, W., Zhang, L., Sun, Y., Bui, A., Hermjakob, H., Watson, K.E. and Ping, P. (2018). “Data from: A reference set of curated biomedical data and metadata from clinical case reports.” Dryad Digital Repository. doi:doi:10.5061/dryad.r36cn90.

29. Hovy, E. and Lavid, J. (2010). “Towards a ‘science’of corpus annotation: a new methodological challenge for corpus linguistics.” *International journal of translation*. 22: 13-36.

30. Bird, S., Klein, E. and Loper, E. (2009). “Natural language processing with Python: analyzing text with the natural language toolkit.” O’Reilly Media, Inc.

31. National Center for Biotechnology Information (U.S.) “Entrez programming utilities help,” NCBI help manual. National Center for Biotechnology Information, Bethesda, MD. 1553894

32. International DOI Foundation (2019). “DOI Resolution Documentation.” DOI.org. Updated: 2018-08-22. Accessed: 2019-08-16. https://doi.org/factsheets/DOIProxy.html

33. Grishman, R. and Sundheim, B. (1996). “Message Understanding Conference-6,” Proceedings of the 16th conference on Computational linguistics -. Morristown, NJ, USA, pp. 466. doi:10.3115/992628.992709.

34. Tjong Kim Sang, E.F. and De Meulder, F. (2003). “Introduction to the CoNLL-2003 shared task,” Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 -. Morristown, NJ, USA, pp. 142-147. doi:10.3115/1119176.1119195.

35. Sweet, L.E. and Moulaison, H.L. (2013). “Electronic Health Records Data and Metadata: Challenges for Big Data in the United States.” *Big Data*. 1: 245-51. PMID:27447257. doi:10.1089/big.2013.0023.

Chapter VI

TAZ encodes tafazzin, a transacylase essential for cardiolipin formation and central to the etiology of Barth syndrome

TAZ encodes tafazzin, a transacylase essential for cardiolipin formation and central to the etiology of Barth syndrome

Anders O. Garlid[1,2], Calvin T. Schaffer[1,2], Jaewoo Kim[1,2],
Hirsh Bhatt[1,2], Vladimir Guevara-Gonzalez[1,5], Peipei Ping[1,2,3,4,6 †]

[1]Cardiovascular Data Science Training Program at UCLA, Departments of [2]Physiology, [3]Medicine/Cardiology, [4]Bioinformatics, [5]Mathematics, and [6]Scalable Analytics Institute (ScAi), University of California at Los Angeles, CA 90095, USA

[†]To Whom Correspondence Should be Addressed: Peipei Ping, Ph.D., FAHA, FISHR, UCLA David Geffen School of Medicine, Departments of Physiology, Medicine (Cardiology) and Bioinformatics, Suite 1-609 MRL Building, 675 Charles E. Young Dr. South, Los Angeles, CA 90095-1760, Phone: 310-267-5624, ppingucla@gmail.com.

Abstract

Tafazzin, which is encoded by the *TAZ* gene, catalyzes transacylation to form mature cardiolipin and shows preference for the transfer of a linoleic acid (LA) group from phosphatidylcholine (PC) to monolysocardiolipin (MLCL) with influence from mitochondrial membrane curvature. The protein contains domains and motifs involved in targeting, anchoring, and an active site for transacylase activity. Tafazzin activity affects many aspects of mitochondrial structure and function, including that of the electron transport chain, fission-fusion, as well as apoptotic signaling. *TAZ* mutations are implicated in Barth syndrome, an underdiagnosed and devastating disease that primarily affects male pediatric patients with a broad spectrum of disease pathologies that impact the cardiovascular, neuromuscular, metabolic, and hematologic systems, and its overexpression has been linked to various cancers.

Keywords

mitochondrial biology; rare mitochondrial diseases; apoptosis; clinical case reports

1. Introduction

TAZ, also known as *G4.5*, is a 10kb gene located at position 28 on the q arm of chromosome X within a gene-rich, 450kb cluster of 13-16 small genes with CpG islands initially identified by Bione *et al.* as potential candidates for involvement in neuromuscular and cardiovascular disorders [1, 2]. *TAZ* encodes the transacylase protein tafazzin (**Table 1**), so named by Bione *et al.* based on a masochistic comic character named Tafazzi from an Italian sports show, apparently due to the difficulty they encountered in the original identification and characterization of this protein [2]. Tafazzin, located in the inner and outer mitochondrial membranes (IMM and OMM), acts as an acyl-specific transacylase that is essential to lipid metabolism through cardiolipin (CL) remodeling. CL remodeling, in turn, is essential for mitochondrial respiratory chain homeostasis, and disruptions to this process as a result of *TAZ* mutations have been shown to be a major cause of the complex, multi-system Barth syndrome. Mutations in the *TAZ* gene are associated with severe cardiovascular defects observed in Barth syndrome (BTHS), including endocardial fibroelastosis (EFE), X-linked dilated cardiomyopathy 3A (CMD3A, and left ventricular noncompaction (LVNC). As such, the gene is known by several aliases, including BTHS, EFE and EFE2, CMD3A, and LVNCX (**Table 1**). Due to similar naming conventions and associations with various cancers for both proteins, the *TAZ* gene and its tafazzin protein product have been confused in the literature with the TAZ *protein*, which is encoded by the *WWTR1* gene. To date, there is no cure for Barth syndrome, and treatments for tafazzin deficiencies have focused on symptom-based management. Deeper investigation of *TAZ*, tafazzin, and cardiolipin is necessary to increase our collective understanding of mitochondrial biology and may help find treatments for Barth syndrome and mitochondrial myopathies.

2. Structure

Tafazzin is part of a superfamily of acyltransferases based on conserved regions and motifs identified through sequence alignment with other acyltransferase proteins involved in phospholipid biosynthesis [3, 4]. The crystal structure of tafazzin has not been determined by x-ray crystallography; the closest homologous protein is plant glycerol-3-phosphate acyltransferase (G3PAT) from Cucurbita moscata and Spinacea oleracea (PDB ID: 1IUQ) [5], which has just 18.1% sequence identity with human tafazzin [6]. The half-life of tafazzin in mammalian cells is much shorter than that of many other mitochondrial proteins at only 3-6 hours [7]; the median half-life for mouse mitochondrial proteins is 17.2 days in the heart and 4.26 days in the liver [8]. This rapid turnover rate has likely contributed to the difficulty in elucidating the structure of tafazzin or acquiring a detailed understanding of its post-translational modifications (PTMs) from mass spectrometry data.

2.1. Tafazzin active site.

The putative phospholipid-binding site of human tafazzin is a 57 amino acid cleft with two open ends and a stretch of conserved, positively charged residues, based on homology modeling using ALAdeGAP for improved amino acid sequence alignment [9, 10]. Tafazzin and related hydrolases contain a conserved histidine residue required for their enzymatic action [3]. In human tafazzin, His-69 (His-77 in yeast) and Asp-74 form part of a conserved HX_4D motif seen in acyltransferases composed of a histidine (His) and aspartic acid (Asp) separated by any 4 amino acids (X_4) [11-13]. The HX_4D motif facilitates the Asp-His dyad mechanism seen commonly in serine proteases, whereby Asp raises the pK_a of His and aids the deprotonation of a hydroxyl group [14].

2.2. Mitochondrial localization and membrane anchoring.

Mitochondrial localization and membrane anchoring domains in tafazzin are of critical importance to its role in cardiolipin remodeling. In the H9c2 rat cardiomyocyte cell line, *TAZ* encodes two peptides external to the active site of the tafazzin protein that act independently to direct it to mitochondria [15]. The first of these sequences, encoded in exon 3 and spanning residues 84-95 on the tafazzin protein, targets exclusively to mitochondria, while the second, encoded in exon 7/8 and forming residues 185-220, also targets other cytosolic compartments [15]. The yeast *TAZ1* orthologue is homologous to human *TAZ* and has been used extensively to study the structure of tafazzin, its function, and in modeling Barth syndrome [16-19]. In yeast, tafazzin has been shown to localize to membrane leaflets facing the intermembrane space (IMS) between the IMM and OMM, where it associates peripherally due to its lack of a transmembrane domain [12, 20]. A hydrophobic sequence from residues 215-232 in yeast tafazzin confers its characteristic interfacial anchoring behavior in both the IMM and OMM [21]. Together, the translocase of the outer membrane (TOM) and the translocase of the inner membrane (TIM) facilitate tafazzin's movement across and insertion into the outer membrane, as well as its anchoring to inner membrane regions of intermediate density [21].

2.3. *TAZ* mutations and effects on tafazzin structure.

In characterizing the mutations of a family of patients with Barth syndrome, many unique forms of tafazzin were identified based on differential splicing events, ranging in length from 129 to 292 amino acids and affecting regions throughout the protein [2]. Many of the shorter forms of the protein lack a 30-residue hydrophobic N-terminus thought to contain a localization signal sequence, as well as modifications to the hydrophilic center of the protein in a 71 amino acid domain profuse with glycine and glutamic acid [2]. Mutations within the localization region result in mistargeting that directs the protein into inner membrane leaflets facing the matrix, rather than facing the IMS [21]. Whited *et al.* categorized *TAZ* mutations into 7 functional classes based on the pathogenic loss-of-function mechanisms of each mutation [22]. The largest class of mutations, Class 1, contains frameshift and splice-site mutations along the length of the gene. Class 2 and 3 mutations are both found in the membrane anchor domain: Class 2 mutations, including V224R, V223D, and I226P variants, represent pleiotropic biochemical defects and often result in mitochondrial mistargeting, while Class 3 mutations (G230R) affect tafazzin macromolecular assembly. Class 4 mutations are composed primarily of missense mutations resulting in catalytically inactive tafazzin and Class 5 mutations, including L90P and N109V, encode hypomorphic alleles which retain transacylase activity. Class 6 mutations, including A88R and L148H, have folding and assembly defects, and Class 7 mutations result in temperature sensitive proteins that undergo activity loss before degradation. There is limited knowledge regarding a link between the different classes of *TAZ* mutations and disease severity. Whether tafazzin is rendered catalytically inactive, mistargeted, or incapable of membrane anchoring, there does not appear to be a clear distinction between phenotypic presentations of patients with different mutations. The diverse nature of *TAZ* mutations is clearly demonstrated in the expansive database maintained and regularly updated by the Barth Syndrome Foundation, which actively collects new data from healthcare professionals on both pathogenic and benign variants (https://barthsyndrome.org/research/tazdatabase.html). Mutations along the length of the *TAZ* gene, their frequency, and their pathogenicity (benign, pathogenic, or unknown effect) are depicted in **Figure 1**. Exonic, pathogenic variants along the length of the tafazzin protein, their frequency, and the type of genetic mutation from which they arise (deletion, frame shift, point mutation, or stop codon) are represented in **Figure 2**, along with the primary protein domains extracted from the literature.

3. Function

Tafazzin plays a critical role in cardiolipin remodeling, limits the structural diversity of CL molecular species, and restricts CL composition to two fatty acids, typically linoleic and oleic acids [23]. Tafazzin displays a preference for the transfer of linoleic acid (LA) from phosphatidylcholine (PC) to monolysocardiolipin (MLCL) and may be affected by and contribute to the negative curvature of the IMM and OMM [24]. Through its effects on CL, tafazzin impacts many aspects of mitochondrial structure-function, including inner membrane curvature, oxidative phosphorylation (OXPHOS), supercomplex formation, oxidative stress repair, apoptosis, and fission and fusion [25-29].

3.1. Transacylase activity.

Tafazzin is an acyl-specific transferase that catalyzes reversible acyl transfer reactions between phospholipids and lysophospholipids in a CoA-independent manner, playing a critical role in the deacylation-reacylation cycle of cardiolipin [13, 24, 30]. Generally, transacylases exhibit phospholipase activity and catalyze acylation and deacylation through a free enzyme acyl intermediate. Tafazzin, on the other hand, does not exhibit phospholipase activity, nor does it utilize the free enzyme acyl intermediate mechanism; it acylates and re-acylates, but deacylation occurs independently of tafazzin [13]. After *de novo* synthesis of CL from phosphatidylglycerol by CL synthase (Crd1 in yeast, hCLS1 in humans) [31, 32], the remodeling process is initiated with cardiolipin deacylation to form MLCL by the cardiolipin specific phospholipase Cld1 in yeast [33] or the calcium-independent phospholipase A2 (iPLA2) in humans [34-38]. In order for MLCL produced by Cld1 to be exposed to tafazzin in the IMS, it must be transported through a different and unknown remodeling step [39]. In mammals, tafazzin functions along with other enzymes to achieve CL remodeling, including MLCL acyltransferase (MLCLAT), acyl-CoA:lysocardiolipin acyltransferase (ALCAT), and phospholipase [31]. Remodeling by tafazzin adds an acyl residue to immature CL, most frequently in the form of a linoleoyl residue in humans [40, 41]. Tafazzin reacylates MLCL in a single-step acyl group transfer reaction (**Figure 3**) from a variety of phospholipids (PL), including CL, PC, phosphatidylethanolamine (PE), and phosphatidic acid (PA). Thus, tafazzin effectively acts as a shuttle for specific acyl groups between different phospholipids [13, 42].

3.2. Acyl specificity and sensing curvature.

Tafazzin shows a clear preference for the transfer of an LA group from PC to MLCL to form mature CL [24]. This remodeling process converts cardiolipin into a mature composition that contains a predominance of tetralinoleoyl moieties. This results in an enrichment of tetralinoleoyl-cardiolipin (CL_4) in the IMM [41]. Indeed, Xu *et al.* report that, in *Drosophila melanogaster*, tafazzin can catalyze acyl transfer using multiple substrates, yet has a preference for the transfer of linoleoyl groups from PC to MLCL at a rate 10 times greater than that of oleoyl groups and twenty times greater than that of arachidonoyl groups, indicating a clear predilection for CL and PC substrates [13]. Conflicting explanations for this preference have been proposed, namely that tafazzin has an inherent enzymatic preference for specific acyl residues, or that it acts on the basis of energy minimization and is influenced by the surrounding mitochondrial microenvironment. Abe *et al.* propose that tafazzin exhibits acyl specificity for the PC to MLCL reaction, and that its function is predominately centered on the transacylation of unsaturated acyl PC to MLCL under any conditions [12]. Schlame *et al.*, on the other hand, propose the 'thermodynamic remodeling' hypothesis, whereby a perturbation of the lipid bilayer state and the physical properties of the lipid membrane determines tafazzin's preference for specific acyl groups [30, 43]. Schlame *et al.* propose that alternative phospholipases and acyltransferases (MLCAT and ALCAT), as well as the thermodynamic properties of lipids, provide the acyl specificity in CL remodeling and that tafazzin itself has no kinetic properties that suggest any sort of acyl-specificity. This mechanism proposes that since CL formation by tafazzin is reversible and has a minimal overall free energy, tafazzin's role is to non-specifically transfer acyl groups among phospholipids to achieve optimal lipid composition and reduce the impact of membrane constraints [24, 43].

According to the model proposed by Schlame *et al.* tafazzin specificity ultimately depends on the physical characteristics and packing properties of the lipid domain, including structural order and state. In *D. melanogaster*, tafazzin requires phospholipids that have a propensity to form non-bilayer phases such as HII phase, which is characterized by its negative curvature, disorganized acyl chains, and low packing order [24, 43]. Stable lipid bilayers were found to have the lowest rate of reaction by tafazzin, while lipids in the hexagonal or micellar phases, which were characterized by packing order changes due to positive or negative curvature, had significantly higher reaction rates. In addition to the rate of reaction, curvature was also shown to determine the specificity of acyl transfer [30]. The cristae of the IMM have negatively curved lipid monolayers and a predominance of phospholipids with small polar head groups, such as CL, and asymmetric, unsaturated hydrocarbon chains, such as linoleic acid. CL and linoleic acid specificity may thus be driven by curvature segregation of phospholipids based on physical properties of the lipid domains, which causes tafazzin to transacylate phospholipids that are located in negatively curved monolayers [20, 43]. In *Saccharomyces cerevisiae*, however, Abe *et al.* determined that tafazzin can efficiently catalyze a transacylation reaction even in a highly ordered lipid bilayer domain. Further, they posit that tafazzin has a unique acyl chain specificity for the PC to MLCL reaction in which tafazzin acts selectively to transfer PC's *sn*-2 acyl chain to MLCL's *sn*-1 and *sn*-2 positions. They determined that these reactions can occur in any environment, regardless of packing order and thermodynamic considerations [12]. These studies illustrate the propensity of tafazzin to transfer a linoleoyl group from PC to MLCL; however, more research into the specific mechanisms is required to fully understand the process and specificity of tafazzin's actions. Further experiments with NMR analysis by groups such as Epand *et al.* may aid in elucidating these mechanisms due to its ability to probe curvature properties of lipid assemblies and observe structures with minor isotropic resonance [30].

3.3. Tafazzin and cardiolipin in mitochondrial structure and function.

Cardiolipin, modified by tafazzin, constitutes 13 - 20% of the total phospholipid mass and exhibits a cone-shaped structure that facilitates its distribution into mitochondrial cristae [27, 44]. In the tafazzin-impaired fibroblasts of Barth syndrome patients, a greater proportion of saturated acyl chain substitution compromises this cone-shaped structure, and CL is heavily depleted with a concurrent accumulation of MLCL species [45]. CL assists in various aspects of OXPHOS, supporting the stability and function of the mitochondrial respiratory chain complexes through linkages between acyl chains [41]. CL binds selectively to the c-rings of ATP synthase, which is required for the function and assembly of the ATP synthase. It also smooths the rotation [46] and facilitates dimerization for efficient ATP synthesis [47]. The structural properties of CL and its pK above 8 facilitate trapping proteins in the IMS. This is thought to achieve proton localization for ATP synthase function and minimization of pH fluctuations [26]. CL also interacts with other proteins such as the ATP/ADP translocase, pyruvate carrier, and carnitine carrier, assisted by glycerol bridges which enable flexibility for interaction with diverse surface shapes.

The lipid-to-protein mass ratio of OXPHOS complexes located in the mitochondrial cristae is 22:78 [48], meaning that each complex is surrounded by just 40-400 lipid molecules [49]. The molecular packing of lipids in a bilayer with such a high protein density causes elastic stress on the curvature of the membrane [50]. Tafazzin remodeling is triggered by OXPHOS complex assembly so as to mitigate this stress and stabilize the membrane by generating CL species with reduced free energy [49]. The OXPHOS complexes I, III, and IV also form supercomplexes within the mitochondrial cristae such as the $I_1III_2IV_{n=1-4}$ "respirasome" [51-54]. CL is directly involved in the formation and maintenance of supercomplexes, providing structural support for trimer- and tetramerization [55, 56]. Furthermore, supercomplexes are disrupted and destabilized in Barth syndrome patients due to the loss of mature CL from the IMM [57].

Cardiolipin's intimate association with the electron transport chain brings it into close proximity with reactive oxygen species (ROS) generated by OXPHOS complexes and which have been shown to target CL. The proximity and its enrichment in long-chain polyunsaturated fatty acid (PUFA) chains make CL susceptible to the attack. In the process of lipid peroxidation, highly reactive oxygen free radicals oxidize the fatty acid chains of CL to form lipid peroxides [58]. Oxidative damage leads to a loss of functional CL, a basis for mitochondrial dysfunction [59, 60]. CL remodeling removes and replaces acyl chains damaged by oxidative stress and is thought to play a key role in oxidative stress repair mechanisms [25] and the recovery of the normal oxidative functions of mitochondria [39]. Aside from its damaging effects, ROS is also critical in mitochondrial and intracellular signaling, particularly in the context of cardioprotection from ischemia-reperfusion injury [61]. Phospholipids in the bilayer, such as cardiolipin, can be oxidized to form hydroperoxy fatty acids, which act as secondary messengers from mitochondria [62].

CL forms membrane domains localized to negatively curved regions and induced by mitochondrial creatine kinase (mtCK) and cytochrome *c* that play critical roles in energy transfer, apoptosis, and functional recovery from ischemic insult [63-65]. The microdomains occur at contact sites where the IMS narrows such that the IMM and OMM are positioned in close proximity to one another [63, 64, 66]. The IMS at these contact sites is replete with mtCK, which induces their formation, recruits CL, and provides stabilization [63, 67]. Mature tetralinoleoyl-CL species generated by tafazzin remodeling are required for the formation of these domains and CL_4 depletion disrupts their formation, which may explain the mitochondrial impairment observed in Barth syndrome and cardiac ischemia-reperfusion injury [66, 68, 69]. Further, mtCK is functionally coupled to adenine nucleotide translocase (ANT) in the IMM to facilitate efficient energy transfer by shuttling high-energy phosphates from the mitochondria to the cytosol through the voltage-dependent anion channel (VDAC) of the outer membrane [65, 70]. During ischemia, for example, this IMS structure-function is disrupted, reducing the functional coupling of mtCK and ANT and increasing the permeability of the OMM to ADP, thereby limiting energy transfer processes and exacerbating damage from an ischemic event such as a heart attack [65]. The cardioprotective ischemic preconditioning (IPC) protocol opens the mitochondrial ATP-sensitive K^+-channel (mitoK_{ATP}), which causes matrix swelling and results in preservation of IMS volume, contact sites, and tight coupling between mtCK and ANT [61, 65, 71]. CL clustering at these contact sites is dependent on the octameric structure of mtCK, which readily binds to anionic phospholipids and may mediate intermembrane contact by binding to VDAC on the OMM [63].

Cardiolipin mediates apoptosis through its interactions with members of the Bcl-2 family, caspases, Bid, Bax, and Bak, with a direct impact on the apoptotic signaling cascade [28]. The total level of CL as well as the oxidative state of its acyl side chains directly impacts apoptosis by regulating cytochrome *c* mobilization; decreased CL content or oxidation of the normally unsaturated acyl side chains releases cytochrome *c* from the membrane [28]. This can be prevented with antioxidants [72, 73] and by the presence of mitochondrial redox proteins [74, 75]. Therefore, CL remodeling by tafazzin restores cytochrome *c* affinity for CL and its localization in the membrane by replacing oxidized fatty acids with non-oxidized acyl groups [31]. Mobilized cytochrome *c* released from the mitochondrial membrane activates caspases 8 and 9, which cleave Bid to produce its truncated and active form, tBid [76]. Once recruited by mitochondrial carrier homologue 2 (MTCH2), tBid activates apoptosis via Bax and Bak activation [77]. Caspases 8 and 9, activate caspase 3, which drives apoptosis and inhibits ROS production [78]. The apoptotic pathway is further amplified by this ROS inhibition, as it is a ROS signal that is responsible for protective mitoK$_{ATP}$ opening that blocks the mitochondrial permeability transition (MPT) and prevents apoptosis [62, 79].

Additionally, through interactions with both inner and outer mitochondrial membranes and proteins such as the GTPase Opa1, CL plays an important role in mitochondrial fission, fusion, and mitophagy [29, 80]. *TAZ* deficiency reduces the generation of mitophagosomes and prevents initiation of mitophagy, further exacerbating the already reduced function of mitochondrial populations in *TAZ*-deficient organisms [80]. Under mitochondria-stress conditions, CL promotes mitochondrial fusion and membrane tethering with L-Opa1 and *trans*-Opa1, respectively, further illustrating its multifunctional importance in mitochondrial form and function [81].

4. Clinical Significance

Considering its critical role in the construction and maintenance of the IMM, it is of little surprise that *TAZ* has been implicated in a broad spectrum of disease pathologies that impact the cardiovascular, neuromuscular, metabolic, and hematologic systems. *TAZ* mutations are specifically associated with the multi-faceted Barth syndrome, and its expression levels have been studied in several varieties of cancer, including thyroid [82], rectal [83], prostate [84], and cervical [85] cancers.

4.1. Barth syndrome.

TAZ mutations cause Barth syndrome, also known as 3-Methylglutaconic Aciduria Type II (3MGA2), an X-linked autosomal recessive disorder that encompasses a complex phenotype, with cardiovascular, musculoskeletal, neurological, metabolic, and hematologic consequences [86-88]. Characterized initially by Barth *et al.* in 1983 [88] as a uniformly lethal disease that affects only males, it has now been found that the age distribution ranges between 0 to 49 years, and symptoms peak around puberty [89]. At least one female patient with BTHS has been identified [90]. The Barth Syndrome Foundation reports that 151 living Barth patients have been identified up to 2012 and 10 new patients are diagnosed each year in the United States with no apparent racial or ethnic predilections. BTHS is estimated to appear in 1 out of every 300,000 to 400,000 live births, although predictions have suggested that the prevalence is actually closer to 1 out of every 140,000 live births as a result of the generally accepted notion that the disease is underdiagnosed [91, 92]. In an effort to impose structure on otherwise unstructured clinical language contained in clinical case reports on BTHS (among other diseases), Caufield *et al.* extracted metadata from the reports and characterized patient symptomology using codes from the International Statistical Classification of Diseases and Related Health Problems (ICD-10) [93, 94]. The MitoCases platform (http://mitocases.org/) houses data on mitochondrial diseases, including Barth syndrome. **Figure 4** displays the distribution of 997 instances of 206 unique ICD-10 codes represented across 54 clinical case reports covering 133 patients with BTHS, including top symptoms overall as well as top cardiovascular symptoms. Classifying Barth syndrome symptomology using ICD-10 codes has the potential to facilitate a greater understanding of the disease, its phenotypes, as well as to aid in diagnosis and treatment.

BTHS diagnosis and treatment is complicated and frequently delayed due to the complexity and variation of disease presentation. Early cardiomyopathy and hypertrophy combined with neutropenia (a low neutrophil count in the blood) is a hallmark of the disease, but confirmation of the diagnosis typically relies on genetic analysis of *TAZ*. Over 160 mutations or errors in the *TAZ* gene have been identified in BTHS patients, with a wide variety of onset, progression, and severity [87]. 3-methylglutaconic acid (3-MGA) and CL content levels are often used to identify BTHS, but they are not always a reliable indicator, which led some to propose using an HPLC-tandem mass spectrometry blood spot assay to measure the ratio of MLCL to CL_4. Although indirect, this highly specific biochemical measure of tafazzin function has the potential to provide a clinically valid method for BTHS diagnosis [95-98]. Combining biochemical analyses with physical tests, such as the 6-minute walk test (6MWT), may allow clinicians to determine the extent of the musculoskeletal impact and cardiac function in patients who survive infancy and those with unknown mitochondrial deficiencies. A combination of these procedures may help to improve diagnostic abilities and shape patient-specific treatment plans [98].

4.1.1. Cardiovascular pathology.

Cardiomyopathy is a major characteristic of BTHS resulting from *TAZ* mutations. *TAZ* mutations lead to altered acyl chain composition and lipid peroxidation, and this can result in a failure of the sarcomeric action required to generate a sufficient power stroke. Disruption of the uniform contraction of sarcomeres can severely weaken the tissue, enlarge the left ventricular chamber, result in partial or incomplete contraction, and lead to decreased ejection volume. This results in the gradual thinning of the ventricular wall, stretching and dilation of cardiac chambers, and a cardiomyopathic phenotype of Barth syndrome, characterized by a weakened heart and diminished contractility [89, 99]. Among all patients, about 95% exhibited a history of cardiomyopathy, with 41.5% of all diagnosed cardiomyopathies occurring from birth to one month of age. Furthermore, statistical analysis revealed that cardiac function of patients declines over time [100]. Contrary to ATP depletion, Wang *et al.* suggest ROS as the main cause of cardiomyocyte dysfunction and cardiovascular impairments such as defective sarcomere assembly and contractile stress [101]. The typical approach to treating the cardiovascular symptoms of Barth syndrome is to follow the treatment paradigm for heart failure. This includes: 1) diuretics for fluid retention (e.g., spironolactone or furosemide), 2) angiotensin-converting enzyme (ACE) inhibitors for vasodilation to reduce afterload (e.g., captopril), 3) positive inotropes to increase contractility and as an antiarrhythmic (e.g., digoxin), and 4) beta blockers to reduce heart rate (e.g., propranolol, carvedilol). Regular echocardiography is used to monitor cardiovascular function and ejection fraction [102, 103]. Severe forms of cardiac symptoms in BTHS patients necessitate heart transplantation. Spencer *et al.* reported nine out of 73 (12%) patients referred to the BTHS Registry (https://barthsyndromeregistry.patientcrossroads.org) who have undergone cardiac transplantation are alive at the last update [86, 100]. Transplantation is generally successful [86]; among four BTHS patients described in Mangat *et al*., one developed a severe infection but they did not show an increased rate of rejection and rated their quality of life as good [104, 105].

Cardiomyopathy in BTHS includes dilated cardiomyopathy (DCM) and left ventricular noncompaction (LVNC) [89]. DCM is a specific type of cardiomyopathy characterized by an enlarged heart that is limited in function due to its inability to contract and pump blood efficiently [106 , 107]. A patient with BTHS resulting from a c.83T>A mutation in tafazzin exhibited DCM with an ejection fraction of 30%, providing a direct association between the gene and DCM [108]. LVNC is a condition that exhibits prominent trabeculations and deep intertrabecular recesses in the left ventricle that resemble a spongy structure on the ventricular wall [109]. One such case involves a family of 6 affected members that presented with LVNC with BTHS due to *TAZ* mutations [110]. Isolated noncompaction of the ventricular myocardium (INVM) has also been found to affect the right ventricle and the interventricular septum [89, 110]. Despite the general occurrence of cardiomyopathy, there have been instances of BTHS caused by *TAZ* mutations with mild or late-onset cardiac involvement, as seen in Woiewodski *et al.* and again in Rigaud *et al.* Each discuss a cohort of BTHS patients exhibiting varying levels of cardiomyopathy, including two infantile patients who did not present with cardiomyopathy at the time of diagnosis [111], another infantile patient whose autopsy revealed no cardiomyopathy, and one 12-year-old patient with no manifestation of cardiomyopathy [112]. There is no clear structural or functional reason for the relatively mild presentations of certain patients, nor a direct mechanistic link between different mutations and disease presentations, representing an intriguing area of research necessary to glean a greater understanding of tafazzin and its role in disease.

4.1.2. Musculoskeletal pathology.

Although skeletal myopathy is often a typical characteristic of patients with disease-causing *TAZ* mutations, it manifests itself in a wide range of symptoms from nonexistent to severe. One of the most common musculoskeletal symptoms in BTHS patients is general and localized weakness. This includes overt muscle weakness and increased exertional fatigue due to skeletal myopathy and exaggerated by the cardiovascular complications associated with Barth syndrome [113]. Hypotonia, fatigue, and weakness can present early in life, persist, and may result in delayed motor development; most patients can walk unassisted by 2 years of age. Common phenotypes include short stature and facial dysmorphia and can extend to rarer phenotypes such as clubfoot (bilateral talipes) [114]. Christodoulou *et al.* describes 6 cases of BTHS from four families with dysmorphic features, all of which exhibited persistent short stature. Four of the patients had also been found to exhibit similar myopathic facial appearances in conjunction with neuromuscular, cardiovascular and infectious symptoms [115]. A growth curve generated by examining 73 BTHS patients in Roberts *et al.* revealed a common down-shift in weight, length, and height relative to the normal population. Developmental delays are prevalent in BTHS patients with motor skills being the most affected, as indicated by a 65% prevalence of a delay in sitting up and a 71.6% delay in walking [100].

Developmental delay has been treated with some success using cornstarch supplementation. This alternate source of glucose production ameliorates muscle wasting due to overnight fasting [86]. Other treatments, including oral arginine and carnitine supplementation, have centered on treating metabolic deficiencies, which improves cardiac function and muscle weakness in some patients [27, 112, 116]. However, while carnitine supplementation was initially offered as a treatment paradigm for all cases of BTHS [117], its effect has subsequently been called into question, and no formal assessment of the utility of arginine supplementation has been published. Thus far, both carnitine and arginine have demonstrated efficacy only in patients with those specific deficiencies [27, 112, 116].

4.1.3. Neurological pathology.

Neurological complications tend to manifest as mild cognitive impairments in BTHS patients with *TAZ* mutations. While these patients were found to have a higher incidence of cognitive impairment [118] and mild learning and speech difficulties [100], many patients were found to have normal cognitive development, including a three-generation family with no cognitive impairment despite BTHS diagnosis [114]. The limited neurologic involvement of BTHS is interesting given that tafazzin has been shown to play an important role in brain mitochondrial respiration and normal cognitive function [119]. One postulate contends that the brain’s reliance on glucose, over tissues in the heart and liver that require high mitochondrial activity, allows the brain to have a more diverse and less tetralinoleoyl-dependent CL composition. Reducing the need for highly symmetric remodeled tetralinoleoyl-CL to achieve sufficient mitochondrial function may allow the brain to mitigate or avoid the detrimental effects of a tafazzin deficiency [41]. Indeed, CL in the brain has higher amounts of arachidonic (AA) and docosahexaenoic (DHA) acids, distinct from the preference for tetralinoleoyl-CL seen in other tissues [41]. Over 80% of CL in the liver and heart take the 18:2n-6 form, whereas the brain demonstrates less of a preference and has a higher concentration of saturated acyl chains, with only 48% polyunsaturated fatty acids and just 20%

of its CL in the 18:2n-6 form [120]. It has also been proposed that the higher ROS scavenging capability of the brain, which is about 100 times higher than the rate of ROS generation [121], may allow the brain to avoid the harmful effects more effectively than in other tissues even though it generates higher total levels of ROS [119].

4.1.4. Metabolic disorder.

3-methylglutaconic aciduria (3-MGA) is a major indicator of a variety of syndromes including BTHS, and is the result of mutations, including those in *TAZ*, that are linked to mitochondrial dysfunction [122]. 3-MGA refers to increased levels of the organic acids 3-methylglutaconic acid, 3-methylglutaric acid, and 2-ethyl-hydracrylic acid in urine [123]. BTHS patients typically demonstrate a large and consistent increase in the excretion of 3-MGA [124]. A diagnosis of 3-MGA type II is synonymous with BTHS. While most BTHS patients exhibit varying severities of 3-MGA, a case report by Schmidt *et al.* describes a 15 year-old-boy with typical BTHS symptoms, such as dilated cardiomyopathy, but normal levels of organic acids, amino acids, and mucopolysaccharides in urine. Thus, there was no diagnosis of 3-MGA, despite a *TAZ* missense mutation in nucleotide 877 at exon 8 [125]. Therapeutics such as riboflavin or coenzyme Q10 have been reported to show substantial improvement in some patients with 3-MGA [123]. Overall, however, metabolic treatments vary between patients and are largely designed to target symptomatic deficiencies rather than the underlying cause of the disease [86, 126].

4.1.5. Hematologic pathology.

Neutropenia is one of the most frequent characteristics of BTHS caused by *TAZ* mutations, characterized by a decline in total number of circulating neutrophils accompanied by an increase in monocytes and eosinophils with no fluctuations in lymphocyte numbers [86]. Makaryan *et al.* found that neutropenia in BTHS is caused by a disruption of mitochondrial membrane potential as well as caspase-3 activation resulting in an increased rate of apoptosis of myeloid progenitor cells [127, 128]. Neutropenia is a particularly variant symptom, and can present itself in many different forms, from severe to mild, cyclical to non-cyclical, and intermittent to chronic [129]. Severe chronic neutropenia (SCN), defined by an absolute neutrophil count of < 500/μL, is the most detrimental phenotype [130].

In a cohort study, Roberts *et al.* describe 73 patients with BTHS and indicate that 69.1% self-reported neutropenia with varying severity [100]. Ranging from a complete lack of neutrophils to a mild decline, neutropenia may be absent at presentation and change over the course of the disease in the same patient [86]. For instance, all seven members of a family with *TAZ* mutations exhibited no signs of neutropenia [131], while another case of siblings with severe BTHS both exhibited intermittent neutropenia [97]. Including persistent or intermittent forms of neutropenia, nearly 90% of BTHS patients exhibit the symptom to some degree [86], though it is mentioned in just over 65% of available clinical case reports on Barth syndrome (**Figure 4**). Neutropenia is an immune system deficiency that results in diminished response to invading organisms. Therefore, decreased defense mechanisms leads to serious bacterial infections including prolonged upper respiratory tract infections, mouth ulcers (chronic aphthous stomatitis) due to *Candida* infections, inflamed gums and perianal dermatitis, as well as sepsis and multi-organ failure, which are frequently treated by prophylactic antibiotics [86, 89, 130]. Among 73 BTHS patients in Roberts *et al.*, 65% of those with neutropenia exhibited mouth ulcers, relative to only 35% of patients without neutropenia, while 28% had a history of pneumonia and 10% had a history of blood infections [100]. Granulocyte colony-stimulating factor (G-CSF) has been identified as an effective and safe treatment for SCN [130], leading to improvement in many BTHS patients with symptoms including aphthous ulcers, bacterial infections, and lethargy [86].

4.1.6. Therapies in Barth syndrome.

Although several therapeutic strategies have proved successful in select clinical presentations, treatments are focused on treating the cardiovascular, musculoskeletal, and metabolic disorders, rather than the root cause of the disease. There is currently no cure for Barth syndrome [92]. Based on the observation that the fatty acid environment of cells impacts CL composition, ATP synthesis, membrane potential, and ROS production, dietary fatty acids have been suggested as a therapeutic strategy to target mitochondrial lipid metabolism and ameliorate effects on bioenergetics and cardiac function in mitochondrial diseases such as BTHS [132]. It is unclear whether these treatments have significant effects in clinical practice. Direct modulation of CL content by lipid replacement therapy using CL nanodisks has also been tested in cell and animal models of Barth syndrome. Apoptosis induced by shRNA-mediated knockdown of *TAZ* in cultured HL60 myeloid progenitor cells [127] is attenuated by incubation with CL nanodisks and confers a significant increase in cellular CL content [133]. However, translation to an *in vivo* setting was unsuccessful, with no alteration in the CL profile of either wildtype mice or a *TAZ* knockdown mouse model of Barth syndrome [134]. Another study aimed to investigate whether overexpression of an alternate CL remodeling enzyme could restore CL in *TAZ*-deficient cells. Lymphoblasts from Barth syndrome patients transfected with *MLCLAT1* saw increased CL levels, improved mitochondrial basal respiration and proton leak, and reduced superoxide production, but only partial compensation for respiratory function and no restoration of OXPHOS supercomplex formation [135, 136]. These results show some promise, but it remains to be seen whether they can be recapitulated in a live animal model.

Elevated ROS and oxidative stress have been proposed as significant culprits in Barth syndrome and the development of cardioskeletal myopathy in these patients [137-139]. *In vitro* studies of *TAZ*-deficient cardiomyocytes treated with the mitochondrially-targeted Mito-Tempo antioxidant demonstrated improved contractile function, cardiac hypertrophy, and cell death [101, 140]. Mice with *TAZ* deficiency (TAZKD) and mitochondria-specific overexpression of catalase, however, developed cardiomyopathy and muscle weakness similar to the Barth syndrome mouse, indicating that amelioration of oxidative stress is insufficient in the *in vivo* setting [141].

Peroxisome proliferator-activated receptors (PPARs) and the PPAR-gamma coactivator-1alpha (PGC-1α) are central to energy metabolism and bioenergetics in mitochondria, presenting opportunities for treatment in a variety of mitochondrial and metabolic disorders. Bezafibrate is a fibric acid derivative pan-agonist of PPAR signaling pathways that activates oxidative metabolism genes [142]. In patients with dyslipidemia or metabolic syndrome, bezafibrate reduces triglyceride levels and the incidence of myocardial infarction [143]. It also significantly decreases HbA1c in diabetic patients with dyslipidemia [144]. Because of its role in mitochondrial bioenergetics, it has been proposed as a potential treatment for Barth syndrome as well. In a TAZKD mouse model with isoproterenol (iso) treatment to induce more significant cardiac dysfunction, bezafibrate rescued iso-induced heart failure with marked increases in left ventricular fractional shortening and ejection fraction and prevention of the development of cardiomyopathy [142]. However, the treatment also caused a significant reduction in CL content and increase in MLCL/CL ratio in both wild type and *TAZ* knockdown mice, a common biomarker for Barth syndrome. Concurrently, mitochondrial biogenesis was amplified drastically, as indicated by a two-fold increase in mtDNA content and mitochondrial citrate synthase activity in bezafibrate-treated hearts. Additionally, the dose used in the mouse model was 60-80 times greater than is typically prescribed in dyslipidemic humans. The modification of CL content and dosage discrepancy presents significant hurdles to determine the mechanism of action, further evaluate the importance of MLCL and CL concentrations, and conduct toxicity studies before any enrollment in clinical trials [142].

Gene replacement therapy presents another avenue of exploration that has the potential to address underlying tafazzin deficiencies resulting from the *TAZ* mutations that cause Barth syndrome. Recombinant adeno-associated virus (rAAV) vectors provide stable and long-lasting gene transfer to the nucleus of an organism's cells using a non-pathogenic virus with minimal immune response [145]. The successful application of an AAV-delivered gene therapy in spinal muscular atrophy also establishes an important precedent for the safety and efficacy of this approach in a clinic setting [146]. AAV serotype 9 (AAV9) demonstrates high affinity for the heart and skeletal muscle, making it ideal for application to Barth syndrome [147]. In the TAZKD mouse model of Barth syndrome, an AAV9-*TAZ* vector with a desmin promotor resulted in significant *TAZ* gene and tafazzin protein expression levels in the heart and muscle and minimal levels in the liver [148]. Measures of muscular strength and fatigue as well as whole body activity (e.g., exercise and distance travelled) of the treated mice improved significantly. Increased fractional shortening and ejection fraction as well as decreased heart weight/body weight ratio indicate significant improvements in cardiac function. Mitochondrial structure and function defects were ameliorated, with improved cristae and sarcomeric organization, greater numbers and size, as well as improved mitochondrial respiration and OXPHOS complex activity [148]. In further studies, multiplex tandem mass tagging-based proteomics has provided a deeper mechanistic insight into the progression of Barth syndrome and its impact on critical proteins involved in cardiac development, heart failure, transcription, translation, and carnitine biosynthesis. [149]. The striking result of AAV9-*TAZ* gene therapy across a range of treatment ages in the mouse model of Barth syndrome paints an optimistic picture for its potential as a future clinical option for these patients.

Elamipretide (MTP-131, SS-31, Bendavia) is a novel mitochondria-targeted tetrapeptide designed to temporarily bind to CL and protect it from oxidative damage by blocking CL-mediated conversion of cytochrome *c* into a peroxidase, thereby preserving cristae structure, promoting OXPHOS, and maintaining mitochondrial integrity [150, 151]. In the canine model of intracoronary microembolization-induced chronic heart failure [152], long-term therapy with elamipretide was demonstrated to improve left ventricular ejection fraction, normalize key plasma biomarkers including tumor necrosis factor-alpha (TNF-α) and C-reactive protein (CRP), and reverse mitochondrial deficiencies in the heart [153] and skeletal muscle [154]. In explanted human cardiac ventricular tissue from patients with a wide demographic range, elamipretide improved impaired mitochondrial function in heart failure and had no effect on normal mitochondrial function in nonfailing hearts. Additionally, supercomplex function was improved, but no change was observed in the activities of OXPHOS complexes II or V [155]. In a clinical trial for patients with heart failure with reduced ejection fraction (HFrEF), elamipretide was safe, well-tolerated, and achieved significant decreases in left ventricular end-diastolic volume (LVEDV) and end-systolic volume (LVESV) in the highest dose cohort. Ejection fraction also improved in the treatment group as compared to those administered a placebo, though the measures did not reach statistical significance [156]. Elamipretide is currently in Phase 2 clinical trials (TAZPOWER) to treat Barth syndrome specifically, and recruitment is in progress for a Phase 3 clinical trial (MMPOWER-3) in patients with primary mitochondrial myopathies [157]. The previous animal studies and clinical trials in heart failure bode well for successful applications of Elamipretide in the Barth syndrome and mitochondrial myopathy patient populations.

4.2. Cancer.

TAZ overexpression, distinct from Barth syndrome-causing mutations, has been linked to various cancers, most commonly rectal cancer, thyroid neoplasm, and cervical cancer. In a cohort study analysis of 140 rectal cancer patients, *TAZ* overexpression was linked to increased expression of oncogenes FXYD-3 and Livin, cell anti-apoptosis response, and abnormal cell growth, as well as an indicator of the stage, type, and progression of rectal cancer [83]. In thyroid neoplasms, *TAZ* overexpression distinguishes follicular variants of papillary carcinomas from classic papillary carcinomas [82], which may present opportunities to provide patients with personalized approaches to treatment based on the difference in prognosis between variants [158]. In human cervical cancer cell lines, tafazzin levels were observed to increase from normal tissue, to squamous intraepithelial lesions, to squamous cervical carcinoma, leading to the proposal that *TAZ* induces cervical cancer progression by inhibiting apoptosis and promoting cancer cell growth, viability, and tumorigenesis [85]. *TAZ* is hypothesized to inhibit apoptosis in cervical cancer cells by limiting cleavage of caspase 3 and caspase 9, which play key roles in apoptotic pathways [85]. Cytochrome *c* released by mitochondria activates caspase 9, which cleaves and activates Bid into tBid. Caspase 3, once activated by caspase 9, inhibits ROS production, driving apoptosis [78]. Low concentration signaling ROS is responsible for protective mitoK_{ATP} opening, which blocks the mitochondrial permeability transition (MPT) and prevents apoptosis [79]. Tafazzin's role in establishing lipid content in CL as a consequence of remodeling has also been implicated in prostate cancer with the finding that palmitoleic acid content of CL was higher in prostate tumor tissue and that palmitic acid had the ability to stimulate prostate cancer cell proliferation and reduce the rate of apoptosis [84]. Extensive studies into the mechanistic role of tafazzin and CL remodeling in the context of cancer has yet to surface, but the effects of impaired mitochondrial function, injury to mitochondrial respiration [159], reduced apoptotic activity, and altered lipid environments [160] are common threads in explaining increased cancer cell proliferation and tumor growth.

5. Conclusions

Tafazzin is a transacylase responsible for remodeling cardiolipin in the mitochondrial membrane and plays an integral role in maintaining mitochondrial structure and function. The tight bends of the cristae in the inner membrane require a specific acyl profile, afforded by the activity of tafazzin. The protein has targeting and anchoring domains that direct it to the IMM and OMM and position it to face the IMS, build mature, tetra-linoleoyl cardiolipin species, and repair damaged membranes. Mutations in *TAZ* produce a dysfunctional or improperly localized protein that causes Barth syndrome, a multi-factorial and devastating disease that presents in infancy and results in heart failure, neutropenia, and musculoskeletal abnormalities. Current therapeutic paradigms are wide-ranging and attempt to treat the symptoms of individual systems. No cure exists for Barth syndrome, though there are a number of different treatments in development aimed at modulating metabolic processes, reducing oxidative stress, protecting CL from degradation, as well as conducting targeted gene-replacement of *TAZ*. Overexpression of *TAZ* may also lead to cancer by affecting cell proliferation and apoptosis. Additional studies are necessary to fully characterize and understand the unique and integral role of tafazzin in mitochondrial biology and in the manifestation of Barth syndrome and various cancers.

Acknowledgements

The authors would like to acknowledge the support of T32 EB16640 to Anders O. Garlid, R35 HL135772 to Peipei Ping, and the Theodore C. Laubisch endowment at UCLA to Peipei Ping. The authors are grateful to Keith Garlid and Hirsh Bhatt for their significant contributions and revisions to this manuscript.

This review and the corresponding Gene Wiki article are written as part of the Gene Wiki Review series – a series resulting from a collaboration between the journal GENE, the Gene Wiki Initiative, and the BD2K initiative. The Gene Wiki Initiative is supported by National Institutes of Health (R01 GM089820, U54 GM114833). Additional support for Gene Wiki Reviews is provided by Elsevier, the publisher of GENE. The corresponding Gene Wiki entry for this review can be found here: https://en.wikipedia.org/wiki/tafazzin/.

Abbreviations

3-MGA	3-methylglutaconic aciduria
AA	arachidonic acid
aa	amino acid
rAAV	recombinant adeno-associated virus
AAV9	adeno-associated virus serotype 9
ACE	angiotensin-converting enzyme
ALCAT	acyl-CoA:lysocardiolipin acyltransferase
ANT	adenine nucleotide translocase
BTHS	Barth syndrome
CL	cardiolipin

CL$_4$	tetralinoleoyl-cardiolipin
CMD3A	cardiomyopathy, dilated 3A (X-linked)
DCM	dilated cardiomyopathy
DHA	docosahexaenoic acid
EFE2	endocardial fibroelastosis 2
G3PAT	glycerol-3-phosphate acyltransferase
G-CSF	granulocyte colony-stimulating factor
HFrEF	heart failure with reduced ejection fraction
IMM	inner mitochondrial membrane
IMS	intermembrane space
INVM	isolated noncompaction of the ventricular myocardium
iPLA2	calcium-independent phospholipase A2
LA	linoleic acid
LVEDV	left ventricular end-diastolic volume
LVESV	left ventricular end-systolic volume
LVNC	left ventricular noncompaction
mitoK$_{ATP}$	mitochondrial ATP-sensitive K$^+$-channel
MLCL	monolysocardiolipin
MLCLAT	monolysocardiolipin acyltransferase
mtCK	mitochondrial creatine kinase
nt	nucleotide
OMM	outer mitochondrial membrane
OXPHOS	oxidative phosphorylation
PA	phosphatidic acid
PC	phosphatidylcholine
PE	phosphatidylethanolamine
PL	phospholipids
PPAR	peroxisome proliferator-activated receptor
PGC-1α	PPAR-gamma coactivator-1alpha
PUFA	polyunsaturated fatty acid

ROS	reactive oxygen species
SCN	severe chronic neutropenia
TIM	translocase of the inner membrane
TOM	translocase of the outer membrane
VDAC	voltage-dependent anion channel

Competing Interests

The authors have no competing interests to declare.

Figures and Tables

Species	Homo sapiens (human)	Mus musculus (mouse)
Gene	*TAZ*	*TAZ*
Synonyms	BTHS, EFE, EFE2, CMDA3, LVNCX, G4.5	5031411C02Rik, 9130012G04Rik, AW107266, AW552613, G4.5
NCBI Gene ID	6901	66826
Chromosomal location	chrXq28:154,411,524-154,421,726	chrX:74,282,697-74,290,151
Length (nt)	10,171	7,454
Exons/Introns	11/10	10/9
NCBI Gene ID	6901	66826
UniProt ID	Q16635	I7HJS2
Ensembl ID	ENSG00000102125	ENSMUST00000069722.12
Length (aa)	292	263
Molecular weight (Da)	33,459	30,433

Table 1. Essential properties and identifiers of *TAZ* and tafazzin. The table contains a summary of the essential properties, identifiers, and names of the *TAZ* gene and the encoded tafazzin protein in Homo sapiens (human) and Mus musculus (mouse) [161].

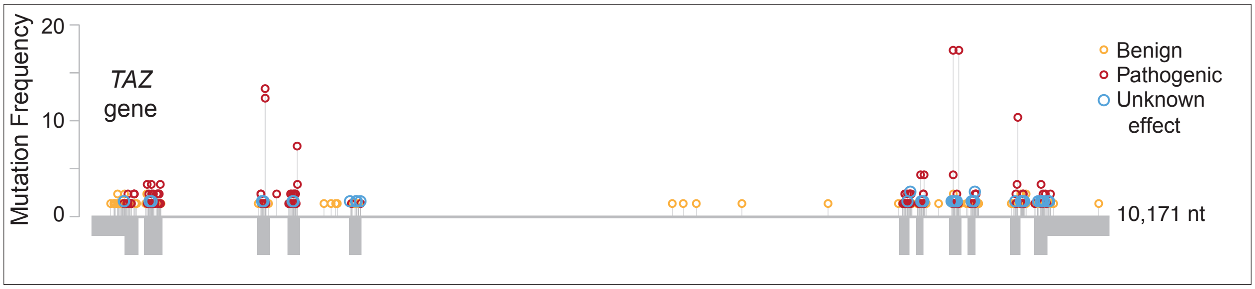


Figure 1. *TAZ* genetic mutations, frequency and pathogenicity. The frequency of intronic and exonic mutations across the length of the *TAZ* gene are shown here, categorized by pathogenicity and mapped to the nucleotide (nt) position on the gene, with exons represented by thick grey bars below the x-axis. Patient mutation data was acquired from the Barth Syndrome Foundation *TAZ* database (https://barthsyndrome.org/research/tazdatabase.html).

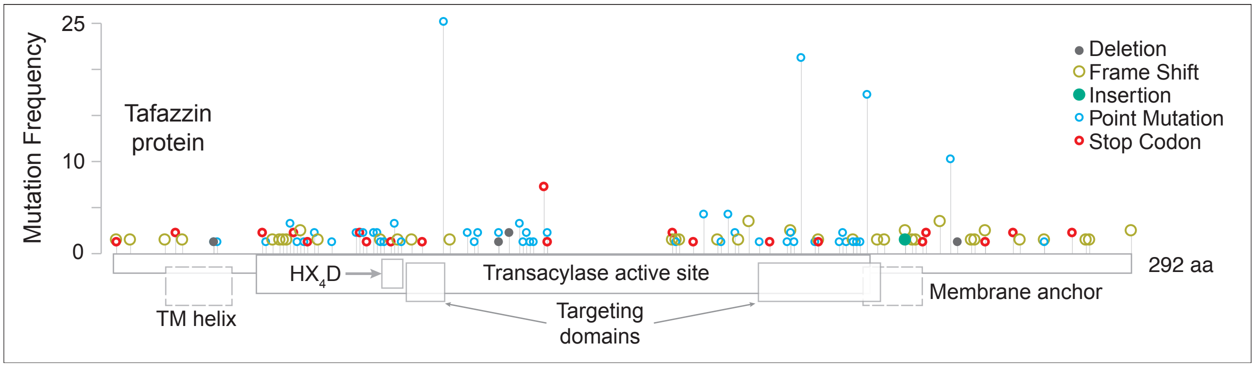


Figure 2. Tafazzin domains, mutation frequency, and mutation type. The frequency of pathogenic, exonic mutations across the length of the tafazzin protein are shown here, categorized by mutation type, with key functional domains of the tafazzin protein displayed below the x-axis. Tafazzin contains a transmembrane helix (TM helix) and a membrane anchor at positions 15-34 and 215-232, respectively. The acyltransferase active site spans 176 amino acids (aa) from residue 41-217, with His77 forming part of the His-Asp motif (HX_4D). Mitochondrial targeting domains are encoded in exon 3 as well as exon 7/8, spanning protein residues 84-95 and 185-220, respectively. Patient mutation data was acquired from the Barth Syndrome Foundation *TAZ* database (https://www.barthsyndrome.org/research/tazdatabase.html).

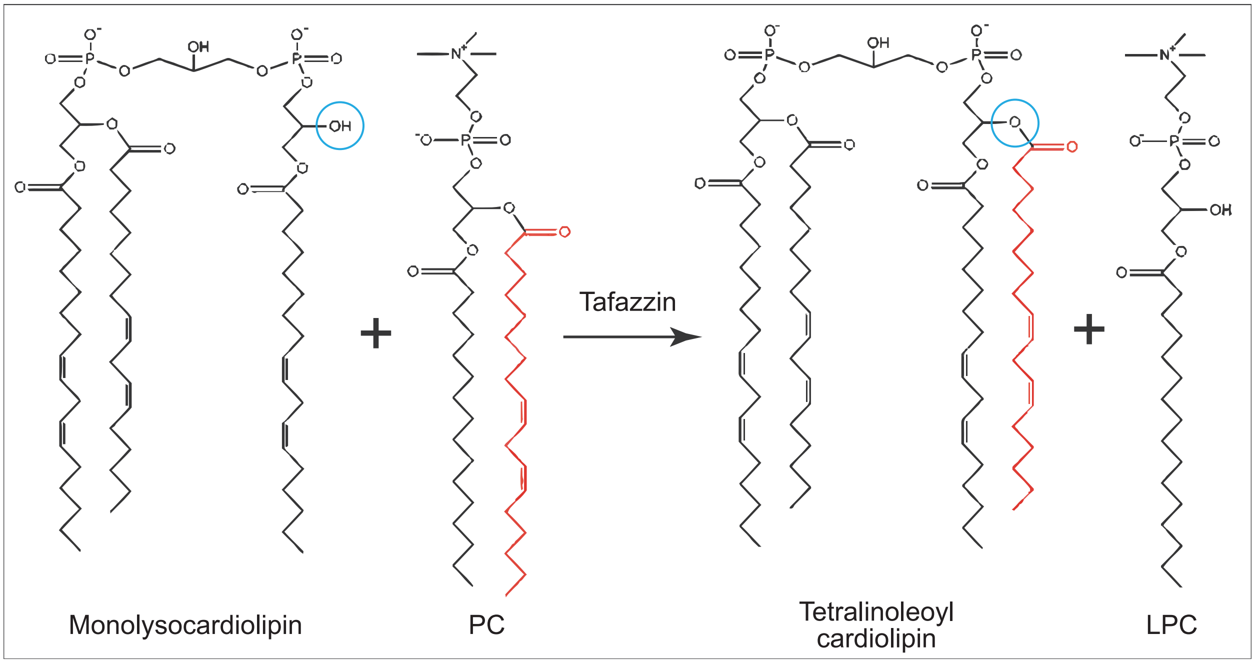


Figure 3. Mechanism of acyltransferase activity by tafazzin. Tafazzin acts as a shuttle for specific acyl groups between different phospholipids to generate mature cardiolipin, Tafazzin transfers an acyl side chain from a phospholipid such as phosphatidylcholine (PC) to reacylate monolysocardiolipin (MLCL) in a single-step acyl group transfer reaction, resulting in the formation of lysophosphatidylcholine (LPC) and the mature tetralinoleoyl form of cardiolipin. The red acyl side chains indicate the acyl group that is transferred by tafazzin, and the blue circles indicate the location on MLCL where the new acyl chain is added to form the mature tetralinoleoyl cardiolipin.

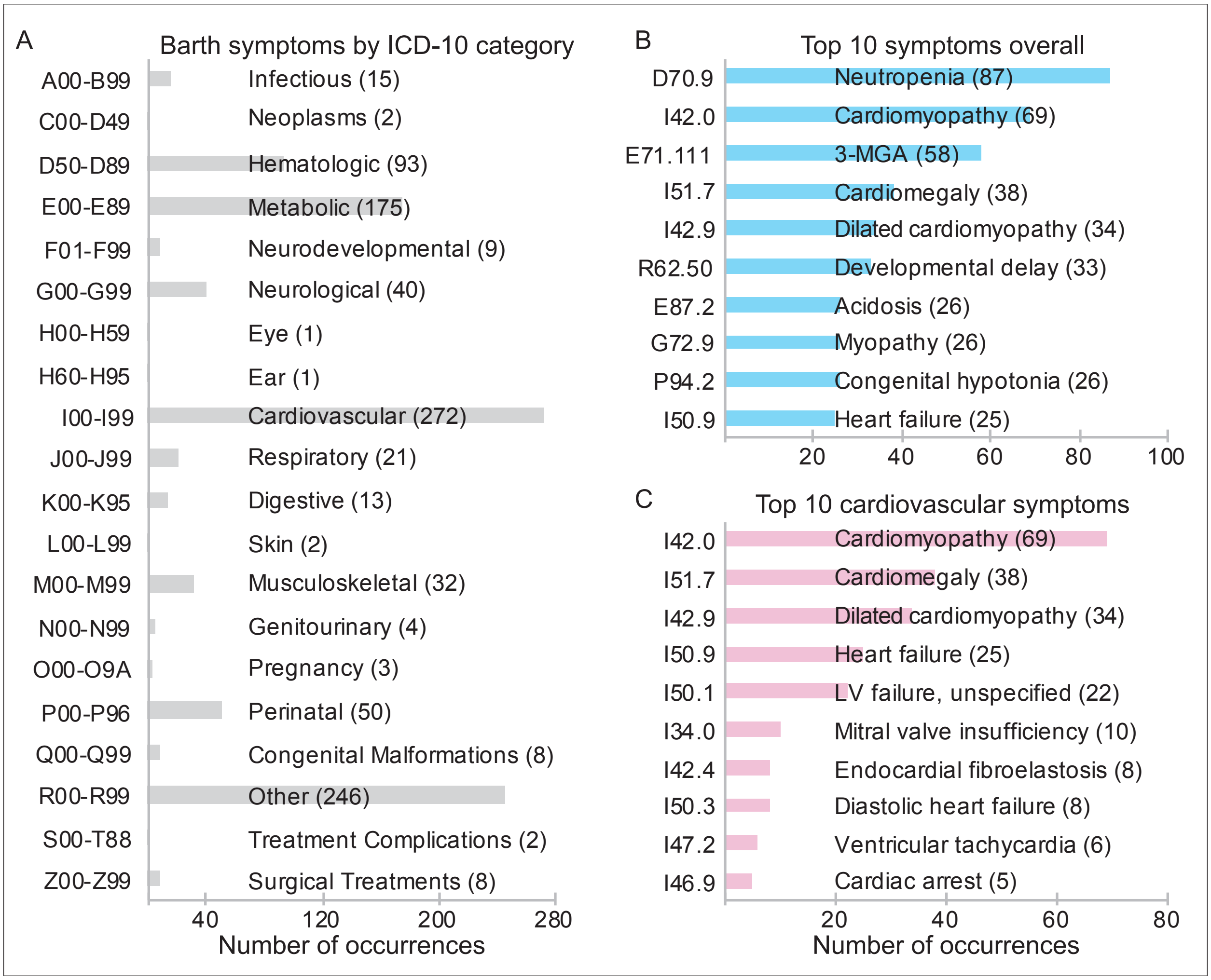


Figure 4. Complex symptomology of Barth syndrome codified by ICD-10. Symptom occurrence codified using the 10th edition of the International Statistical Classification of Diseases and Related Health Problems (ICD-10) from 133 Barth syndrome patients described in 54 clinical case reports (please see **References Cited: Clinical Case Reports**). Panel **(A)** presents the full collection of 997 instances of 206 unique ICD-10 codes across all patients in these reports, grouped by disease category [93, 94]. Cardiovascular diseases and symptoms are the most highly represented among all ICD-10 categories (n = 272). Panel **(B)** highlights the top 10 symptoms from all ICD-10 categories. Panel **(C)** depicts the top 10 cardiovascular symptoms. All data is housed on the MitoCases platform (http://mitocases.org/) along with detailed metadata on the medical information contained in the text of each CCR.

References Cited

1. Bione, S., Tamanini, F., Maestrini, E., Tribioli, C., Poustka, A., Torri, G., Rivella, S. and Toniolo, D. (1993). “Transcriptional organization of a 450-kb region of the human X chromosome in Xq28.” *Proc Natl Acad Sci U S A*. 90: 10977-81. PMID:8248200. 8248200

2. Bione, S., D’Adamo, P., Maestrini, E., Gedeon, A.K., Bolhuis, P.A. and Toniolo, D. (1996). “A novel X-linked gene, G4.5. is responsible for Barth syndrome.” *Nat Genet*. 12: 385-9. PMID:8630491. doi:10.1038/ng0496-385.

3. Neuwald, A.F. (1997). “Barth syndrome may be due to an acyltransferase deficiency.” *Curr Biol*. 7: R465-6. PMID:9259571. 9259571

4. Acehan, D., Xu, Y., Stokes, D.L. and Schlame, M. (2007). “Comparison of lymphoblast mitochondria from normal subjects and patients with Barth syndrome using electron microscopic tomography.” *Lab Invest*. 87: 40-8. PMID:17043667. doi:10.1038/labinvest.3700480.

5. Tamada, T., Feese, M.D., Ferri, S.R., Kato, Y., Yajima, R., Toguri, T. and Kuroki, R. (2004). “Substrate recognition and selectivity of plant glycerol-3-phosphate acyltransferases (GPATs) from Cucurbita moscata and Spinacea oleracea.” *Acta Crystallogr D Biol Crystallogr*. 60: 13-21. PMID:14684887. doi:10.1107/s0907444903020778.

6. Waterhouse, A., Bertoni, M., Bienert, S., Studer, G., Tauriello, G., Gumienny, R., Heer, F.T., de Beer, T.A.P., Rempfer, C., Bordoli, L., Lepore, R. and Schwede, T. (2018). “SWISS-MODEL: homology modelling of protein structures and complexes.” *Nucleic Acids Res*. 46: W296-W303. PMID:29788355. doi:10.1093/nar/gky427.

7. Xu, Y., Malhotra, A., Claypool, S.M., Ren, M. and Schlame, M. (2015). “Tafazzins from Drosophila and mammalian cells assemble in large protein complexes with a short half-life.” *Mitochondrion*. 21: 27-32. PMID:25598000. doi:10.1016/j.mito.2015.01.002.

8. Kim, T.Y., Wang, D., Kim, A.K., Lau, E., Lin, A.J., Liem, D.A., Zhang, J., Zong, N.C., Lam, M.P. and Ping, P. (2012). “Metabolic labeling reveals proteome dynamics of mouse mitochondria.” *Mol Cell Proteomics*. 11: 1586-94. PMID:22915825. doi:10.1074/mcp.M112.021162.

9. Hijikata, A., Yura, K., Noguti, T. and Go, M. (2011). “Revisiting gap locations in amino acid sequence alignments and a proposal for a method to improve them by introducing solvent accessibility.” *Proteins*. 79: 1868-77. PMID:21465562. doi:10.1002/prot.23011.

10. Hijikata, A., Yura, K., Ohara, O. and Go, M. (2015). “Structural and functional analyses of Barth syndrome-causing mutations and alternative splicing in the tafazzin acyltransferase domain.” *Meta Gene*. 4: 92-106. PMID:25941633. doi:10.1016/j.mgene.2015.04.001.

11. Heath, R.J. and Rock, C.O. (1998). “A conserved histidine is essential for glycerolipid acyltransferase catalysis.” *J Bacteriol*. 180: 1425-30. PMID:9515909. 9515909

12. Abe, M., Hasegawa, Y., Oku, M., Sawada, Y., Tanaka, E., Sakai, Y. and Miyoshi, H. (2016). “Mechanism for Remodeling of the Acyl Chain Composition of Cardiolipin Catalyzed by Saccharomyces cerevisiae Tafazzin.” *J Biol Chem*. 291: 15491-502. PMID:27268057. doi:10.1074/jbc.M116.718510.

13. Xu, Y., Malhotra, A., Ren, M. and Schlame, M. (2006). “The enzymatic function of tafazzin.” *J Biol Chem*. 281: 39217-24. PMID:17082194. doi:10.1074/jbc.M606100200.

14. Tang, Y., Xia, H. and Li, D. (2018). “Membrane Phospholipid Biosynthesis in Bacteria.” in: Cao, Y. (Ed.),

Advances in Membrane Proteins: Part I: Mass Processing and Transportation. Springer Singapore, Singapore, pp. 77-119. 10.1007/978-981-13-0532-0_4.

15. Dinca, A.A., Chien, W.M. and Chin, M.T. (2018). “Identification of novel mitochondrial localization signals in human Tafazzin, the cause of the inherited cardiomyopathic disorder Barth syndrome.” *J Mol Cell Cardiol*. 114: 83-92. PMID:29129703. doi:10.1016/j.yjmcc.2017.11.005.

16. Claypool, S.M., McCaffery, J.M. and Koehler, C.M. (2006). “Mitochondrial mislocalization and altered assembly of a cluster of Barth syndrome mutant tafazzins.” *J Cell Biol*. 174: 379-90. PMID:16880272. doi:10.1083/jcb.200605043.

17. Gu, Z., Valianpour, F., Chen, S., Vaz, F.M., Hakkaart, G.A., Wanders, R.J. and Greenberg, M.L. (2004). “Aberrant cardiolipin metabolism in the yeast taz1 mutant: a model for Barth syndrome.” *Mol Microbiol*. 51: 149-58. PMID:14651618. doi:10.1046/j.1365-2958.2003.03802.x.

18. Ma, L., Vaz, F.M., Gu, Z., Wanders, R.J. and Greenberg, M.L. (2004). “The human TAZ gene complements mitochondrial dysfunction in the yeast taz1Delta mutant. Implications for Barth syndrome.” *J Biol Chem*. 279: 44394-9. PMID:15304507. doi:10.1074/jbc.M405479200.

19. Vaz, F.M., Houtkooper, R.H., Valianpour, F., Barth, P.G. and Wanders, R.J. (2003). “Only one splice variant of the human TAZ gene encodes a functional protein with a role in cardiolipin metabolism.” *J Biol Chem*. 278: 43089-94. PMID:12930833. doi:10.1074/jbc.M305956200.

20. Gawrisch, K. (2012). “Tafazzin senses curvature.” *Nat Chem Biol*. 8: 811-2. PMID:22987008. doi:10.1038/nchembio.1068.

21. Herndon, J.D., Claypool, S.M. and Koehler, C.M. (2013). “The Taz1p transacylase is imported and sorted into the outer mitochondrial membrane via a membrane anchor domain.” *Eukaryot Cell*. 12: 1600-8. PMID:24078306. doi:10.1128/EC.00237-13.

22. Whited, K., Baile, M.G., Currier, P. and Claypool, S.M. (2013). “Seven functional classes of Barth syndrome mutation.” *Hum Mol Genet*. 22: 483-92. PMID:23100323. doi:10.1093/hmg/dds447.

23. Schlame, M. (2008). “Cardiolipin synthesis for the assembly of bacterial and mitochondrial membranes.” *J Lipid Res*. 49: 1607-20. PMID:18077827. doi:10.1194/jlr.R700018-JLR200.

24. Schlame, M., Xu, Y. and Ren, M. (2017). “The Basis for Acyl Specificity in the Tafazzin Reaction.” *J Biol Chem*. 292: 5499-5506. PMID:28202545. doi:10.1074/jbc.M116.769182.

25. Baile, M.G., Sathappa, M., Lu, Y.W., Pryce, E., Whited, K., McCaffery, J.M., Han, X., Alder, N.N. and Claypool, S.M. (2014). “Unremodeled and remodeled cardiolipin are functionally indistinguishable in yeast.” *J Biol Chem*. 289: 1768-78. PMID:24285538. doi:10.1074/jbc.M113.525733.

26. Haines, T.H. and Dencher, N.A. (2002). “Cardiolipin: a proton trap for oxidative phosphorylation.” *FEBS Lett*. 528: 35-9. PMID:12297275. 12297275

27. Vernon, H.J., Sandlers, Y., McClellan, R. and Kelley, R.I. (2014). “Clinical laboratory studies in Barth Syndrome.” *Mol Genet Metab*. 112: 143-7. PMID:24751896. doi:10.1016/j.ymgme.2014.03.007.

28. Schug, Z.T. and Gottlieb, E. (2009). “Cardiolipin acts as a mitochondrial signalling platform to launch apoptosis.” *Biochim Biophys Acta*. 1788: 2022-31. PMID:19450542. doi:10.1016/j.bbamem.2009.05.004.

29. Frohman, M.A. (2015). “Role of mitochondrial lipids in guiding fission and fusion.” *J Mol Med (Berl)*. 93: 263-9. PMID:25471483. doi:10.1007/s00109-014-1237-z.

30. Epand, R.M., D'Souza, K., Berno, B. and Schlame, M. (2015). "Membrane curvature modulation of protein activity determined by NMR." *Biochim Biophys Acta*. 1848: 220-8. PMID:24835017. doi:10.1016/j.bbamem.2014.05.004.

31. Ye, C., Shen, Z. and Greenberg, M.L. (2016). "Cardiolipin remodeling: a regulatory hub for modulating cardiolipin metabolism and function." *J Bioenerg Biomembr*. 48: 113-23. PMID:25432572. doi:10.1007/s10863-014-9591-7.

32. Chen, D., Zhang, X.Y. and Shi, Y. (2006). "Identification and functional characterization of hCLS1, a human cardiolipin synthase localized in mitochondria." *Biochem J*. 398: 169-76. PMID:16716149. doi:10.1042/BJ20060303.

33. Beranek, A., Rechberger, G., Knauer, H., Wolinski, H., Kohlwein, S.D. and Leber, R. (2009). "Identification of a cardiolipin-specific phospholipase encoded by the gene CLD1 (YGR110W) in yeast." *J Biol Chem*. 284: 11572-8. PMID:19244244. doi:10.1074/jbc.M805511200.

34. Hsu, Y.H., Dumlao, D.S., Cao, J. and Dennis, E.A. (2013). "Assessing phospholipase A2 activity toward cardiolipin by mass spectrometry." *PLoS One*. 8: e59267. PMID:23533611. doi:10.1371/journal.pone.0059267.

35. Yoda, E., Hachisu, K., Taketomi, Y., Yoshida, K., Nakamura, M., Ikeda, K., Taguchi, R., Nakatani, Y., Kuwata, H., Murakami, M., Kudo, I. and Hara, S. (2010). "Mitochondrial dysfunction and reduced prostaglandin synthesis in skeletal muscle of Group VIB Ca2+-independent phospholipase A2gamma-deficient mice." *J Lipid Res*. 51: 3003-15. PMID:20625036. doi:10.1194/jlr.M008060.

36. Malhotra, A., Edelman-Novemsky, I., Xu, Y., Plesken, H., Ma, J., Schlame, M. and Ren, M. (2009). "Role of calcium-independent phospholipase A2 in the pathogenesis of Barth syndrome." *Proc Natl Acad Sci U S A*. 106: 2337-41. PMID:19164547. doi:10.1073/pnas.0811224106.

37. Mancuso, D.J., Sims, H.F., Han, X., Jenkins, C.M., Guan, S.P., Yang, K., Moon, S.H., Pietka, T., Abumrad, N.A., Schlesinger, P.H. and Gross, R.W. (2007). "Genetic ablation of calcium-independent phospholipase A2gamma leads to alterations in mitochondrial lipid metabolism and function resulting in a deficient mitochondrial bioenergetic phenotype." *J Biol Chem*. 282: 34611-22. PMID:17923475. doi:10.1074/jbc.M707795200.

38. Mancuso, D.J., Han, X., Jenkins, C.M., Lehman, J.J., Sambandam, N., Sims, H.F., Yang, J., Yan, W., Yang, K., Green, K., Abendschein, D.R., Saffitz, J.E. and Gross, R.W. (2007). "Dramatic accumulation of triglycerides and precipitation of cardiac hemodynamic dysfunction during brief caloric restriction in transgenic myocardium expressing human calcium-independent phospholipase A2gamma." *J Biol Chem*. 282: 9216-27. PMID:17213206. doi:10.1074/jbc.M607307200.

39. Baile, M.G., Whited, K. and Claypool, S.M. (2013). "Deacylation on the matrix side of the mitochondrial inner membrane regulates cardiolipin remodeling." *Mol Biol Cell*. 24: 2008-20. PMID:23637464. doi:10.1091/mbc.E13-03-0121.

40. Minkler, P.E. and Hoppel, C.L. (2010). "Separation and characterization of cardiolipin molecular species by reverse-phase ion pair high-performance liquid chromatography-mass spectrometry." *J Lipid Res*. 51: 856-65. PMID:19965604. doi:10.1194/jlr.D002857.

41. Houtkooper, R.H., Turkenburg, M., Poll-The, B.T., Karall, D., Perez-Cerda, C., Morrone, A., Malvagia, S., Wanders, R.J., Kulik, W. and Vaz, F.M. (2009). "The enigmatic role of tafazzin in cardiolipin metabolism." *Biochim Biophys Acta*. 1788: 2003-14. PMID:19619503. doi:10.1016/j.bbamem.2009.07.009.

42. Schlame, M. (2013). "Cardiolipin remodeling and the function of tafazzin." *Biochim Biophys Acta*. 1831:

582-8. PMID:23200781. doi:10.1016/j.bbalip.2012.11.007.

43. Schlame, M., Acehan, D., Berno, B., Xu, Y., Valvo, S., Ren, M., Stokes, D.L. and Epand, R.M. (2012). “The physical state of lipid substrates provides transacylation specificity for tafazzin.” *Nat Chem Biol*. 8: 862-9. PMID:22941046. doi:10.1038/nchembio.1064.

44. Daum, G. and Vance, J.E. (1997). “Import of lipids into mitochondria.” *Prog Lipid Res*. 36: 103-30. PMID:9624424. 9624424

45. Oemer, G., Lackner, K., Muigg, K., Krumschnabel, G., Watschinger, K., Sailer, S., Lindner, H., Gnaiger, E., Wortmann, S.B., Werner, E.R., Zschocke, J. and Keller, M.A. (2018). “Molecular structural diversity of mitochondrial cardiolipins.” *Proc Natl Acad Sci U S A*. 115: 4158-4163. PMID:29618609. doi:10.1073/pnas.1719407115.

46. Mehdipour, A.R. and Hummer, G. (2016). “Cardiolipin puts the seal on ATP synthase.” *Proc Natl Acad Sci U S A*. 113: 8568-70. PMID:27439859. doi:10.1073/pnas.1609806113.

47. Acehan, D., Malhotra, A., Xu, Y., Ren, M., Stokes, D.L. and Schlame, M. (2011). “Cardiolipin affects the supramolecular organization of ATP synthase in mitochondria.” *Biophys J*. 100: 2184-92. PMID:21539786. doi:10.1016/j.bpj.2011.03.031.

48. Lotan, R. and Nicolson, G.L. (1981). “Plasma membranes of eukaryotes.” Van Nostrand-Reinhold, Princeton, NJ.

49. Xu, Y., Anjaneyulu, M., Donelian, A., Yu, W., Greenberg, M.L., Ren, M., Owusu-Ansah, E. and Schlame, M. (2019). “Assembly of the complexes of oxidative phosphorylation triggers the remodeling of cardiolipin.” *Proc Natl Acad Sci U S A*. 116: 11235-11240. PMID:31110016. doi:10.1073/pnas.1900890116.

50. Brown, M.F. (2017). “Soft Matter in Lipid-Protein Interactions.” *Annu Rev Biophys*. 46: 379-410. PMID:28532212. doi:10.1146/annurev-biophys-070816-033843.

51. Stuart, R.A. (2008). “Supercomplex organization of the oxidative phosphorylation enzymes in yeast mitochondria.” *J Bioenerg Biomembr*. 40: 411-7. PMID:18839289. doi:10.1007/s10863-008-9168-4.

52. Schagger, H. (2001). “Respiratory chain supercomplexes.” *IUBMB Life*. 52: 119-28. PMID:11798023. doi:10.1080/15216540152845911.

53. Schagger, H. and Pfeiffer, K. (2000). “Supercomplexes in the respiratory chains of yeast and mammalian mitochondria.” *EMBO J*. 19: 1777-83. PMID:10775262. doi:10.1093/emboj/19.8.1777.

54. Gu, J., Wu, M., Guo, R., Yan, K., Lei, J., Gao, N. and Yang, M. (2016). “The architecture of the mammalian respirasome.” *Nature*. 537: 639-43. PMID:27654917. doi:10.1038/nature19359.

55. Mileykovskaya, E. and Dowhan, W. (2014). “Cardiolipin-dependent formation of mitochondrial respiratory supercomplexes.” *Chem Phys Lipids*. 179: 42-8. PMID:24220496. doi:10.1016/j.chemphyslip.2013.10.012.

56. Zhang, M., Mileykovskaya, E. and Dowhan, W. (2002). “Gluing the respiratory chain together. Cardiolipin is required for supercomplex formation in the inner mitochondrial membrane.” *J Biol Chem*. 277: 43553-6. PMID:12364341. doi:10.1074/jbc.C200551200.

57. McKenzie, M., Lazarou, M., Thorburn, D.R. and Ryan, M.T. (2006). “Mitochondrial respiratory chain supercomplexes are destabilized in Barth Syndrome patients.” *J Mol Biol*. 361: 462-9. PMID:16857210. doi:10.1016/j.jmb.2006.06.057.

58. Paradies, G., Petrosillo, G., Pistolese, M. and Ruggiero, F.M. (2002). “Reactive oxygen species affect mitochondrial electron transport complex I activity through oxidative cardiolipin damage.” *Gene*. 286: 135-41. PMID:11943469. 11943469

59. Lesnefsky, E.J. and Hoppel, C.L. (2008). “Cardiolipin as an oxidative target in cardiac mitochondria in the aged rat.” *Biochim Biophys Acta*. 1777: 1020-7. PMID:18515061. doi:10.1016/j.bbabio.2008.05.444.

60. Shi, Y. (2010). “Emerging roles of cardiolipin remodeling in mitochondrial dysfunction associated with diabetes, obesity, and cardiovascular diseases.” *J Biomed Res*. 24: 6-15. PMID:23554606. doi:10.1016/S1674-8301(10)60003-6.

61. Garlid, K.D. (2000). “Opening mitochondrial K(ATP) in the heart--what happens, and what does not happen.” *Basic Res Cardiol*. 95: 275-9. PMID:11005581. 11005581

62. Garlid, A.O., Jaburek, M., Jacobs, J.P. and Garlid, K.D. (2013). “Mitochondrial reactive oxygen species: which ROS signals cardioprotection?” *Am J Physiol Heart Circ Physiol*. 305: H960-8. PMID:23913710. doi:10.1152/ajpheart.00858.2012.

63. Epand, R.F., Tokarska-Schlattner, M., Schlattner, U., Wallimann, T. and Epand, R.M. (2007). “Cardiolipin clusters and membrane domain formation induced by mitochondrial proteins.” *J Mol Biol*. 365: 968-80. PMID:17097675. doi:10.1016/j.jmb.2006.10.028.

64. Renner, L.D. and Weibel, D.B. (2011). “Cardiolipin microdomains localize to negatively curved regions of Escherichia coli membranes.” *Proc Natl Acad Sci U S A*. 108: 6264-9. PMID:21444798. doi:10.1073/pnas.1015757108.

65. Laclau, M.N., Boudina, S., Thambo, J.B., Tariosse, L., Gouverneur, G., Bonoron-Adele, S., Saks, V.A., Garlid, K.D. and Dos Santos, P. (2001). “Cardioprotection by ischemic preconditioning preserves mitochondrial function and functional coupling between adenine nucleotide translocase and creatine kinase.” *J Mol Cell Cardiol*. 33: 947-56. PMID:11343417. doi:10.1006/jmcc.2001.1357.

66. Pennington, E.R., Sullivan, E.M., Fix, A., Dadoo, S., Zeczycki, T.N., DeSantis, A., Schlattner, U., Coleman, R.A., Chicco, A.J., Brown, D.A. and Shaikh, S.R. (2018). “Proteolipid domains form in biomimetic and cardiac mitochondrial vesicles and are regulated by cardiolipin concentration but not monolyso-cardiolipin.” *J Biol Chem*. 293: 15933-15946. PMID:30158245. doi:10.1074/jbc.RA118.004948.

67. Speer, O., Back, N., Buerklen, T., Brdiczka, D., Koretsky, A., Wallimann, T. and Eriksson, O. (2005). “Octameric mitochondrial creatine kinase induces and stabilizes contact sites between the inner and outer membrane.” *Biochem J*. 385: 445-50. PMID:15294016. doi:10.1042/BJ20040386.

68. Sparagna, G.C., Chicco, A.J., Murphy, R.C., Bristow, M.R., Johnson, C.A., Rees, M.L., Maxey, M.L., McCune, S.A. and Moore, R.L. (2007). “Loss of cardiac tetralinoleoyl cardiolipin in human and experimental heart failure.” *J Lipid Res*. 48: 1559-70. PMID:17426348. doi:10.1194/jlr.M600551-JLR200.

69. Paradies, G., Paradies, V., Ruggiero, F.M. and Petrosillo, G. (2015). “Cardiolipin alterations and mitochondrial dysfunction in heart ischemia/reperfusion injury.” *Clinical Lipidology*. 10: 415-429.

70. Saks, V., Dzeja, P., Schlattner, U., Vendelin, M., Terzic, A. and Wallimann, T. (2006). “Cardiac system bioenergetics: metabolic basis of the Frank-Starling law.” *J Physiol*. 571: 253-73. PMID:16410283. doi:10.1113/jphysiol.2005.101444.

71. Costa, A.D. and Garlid, K.D. (2009). “MitoKATP activity in healthy and ischemic hearts.” *J Bioenerg Biomembr*. 41: 123-6. PMID:19353252. doi:10.1007/s10863-009-9213-y.

72. Petrosillo, G., Ruggiero, F.M. and Paradies, G. (2003). "Role of reactive oxygen species and cardiolipin in the release of cytochrome c from mitochondria." *FASEB J*. 17: 2202-8. PMID:14656982. doi:10.1096/fj.03-0012com.

73. Tyurina, Y.Y., Kini, V., Tyurin, V.A., Vlasova, II, Jiang, J., Kapralov, A.A., Belikova, N.A., Yalowich, J.C., Kurnikov, I.V. and Kagan, V.E. (2006). "Mechanisms of cardiolipin oxidation by cytochrome c: relevance to pro- and antiapoptotic functions of etoposide." *Mol Pharmacol*. 70: 706-17. PMID:16690782. doi:10.1124/mol.106.022731.

74. Enoksson, M., Fernandes, A.P., Prast, S., Lillig, C.H., Holmgren, A. and Orrenius, S. (2005). "Overexpression of glutaredoxin 2 attenuates apoptosis by preventing cytochrome c release." *Biochem Biophys Res Commun*. 327: 774-9. PMID:15649413. doi:10.1016/j.bbrc.2004.12.067.

75. Ran, Q., Liang, H., Gu, M., Qi, W., Walter, C.A., Roberts, L.J., 2nd, Herman, B., Richardson, A. and Van Remmen, H. (2004). "Transgenic mice overexpressing glutathione peroxidase 4 are protected against oxidative stress-induced apoptosis." *J Biol Chem*. 279: 55137-46. PMID:15496407. doi:10.1074/jbc.M410387200.

76. Kantari, C. and Walczak, H. (2011). "Caspase-8 and bid: caught in the act between death receptors and mitochondria." *Biochim Biophys Acta*. 1813: 558-63. PMID:21295084. doi:10.1016/j.bbamcr.2011.01.026.

77. Katz, C., Zaltsman-Amir, Y., Mostizky, Y., Kollet, N., Gross, A. and Friedler, A. (2012). "Molecular basis of the interaction between proapoptotic truncated BID (tBID) protein and mitochondrial carrier homologue 2 (MTCH2) protein: key players in mitochondrial death pathway." *J Biol Chem*. 287: 15016-23. PMID:22416135. doi:10.1074/jbc.M111.328377.

78. Brentnall, M., Rodriguez-Menocal, L., De Guevara, R.L., Cepero, E. and Boise, L.H. (2013). "Caspase-9, caspase-3 and caspase-7 have distinct roles during intrinsic apoptosis." *BMC Cell Biol*. 14: 32. PMID:23834359. doi:10.1186/1471-2121-14-32.

79. Costa, A.D. and Garlid, K.D. (2008). "Intramitochondrial signaling: interactions among mitoKATP, PKCepsilon, ROS, and MPT." *Am J Physiol Heart Circ Physiol*. 295: H874-82. PMID:18586884. doi:10.1152/ajpheart.01189.2007.

80. Hsu, P., Liu, X., Zhang, J., Wang, H.G., Ye, J.M. and Shi, Y. (2015). "Cardiolipin remodeling by TAZ/tafazzin is selectively required for the initiation of mitophagy." *Autophagy*. 11: 643-52. PMID:25919711. doi:10.1080/15548627.2015.1023984.

81. Ban, T., Ishihara, T., Kohno, H., Saita, S., Ichimura, A., Maenaka, K., Oka, T., Mihara, K. and Ishihara, N. (2017). "Molecular basis of selective mitochondrial fusion by heterotypic action between OPA1 and cardiolipin." *Nat Cell Biol*. 19: 856-863. PMID:28628083. doi:10.1038/ncb3560.

82. Chevillard, S., Ugolin, N., Vielh, P., Ory, K., Levalois, C., Elliott, D., Clayman, G.L. and El-Naggar, A.K. (2004). "Gene expression profiling of differentiated thyroid neoplasms: diagnostic and clinical implications." *Clin Cancer Res*. 10: 6586-97. PMID:15475448. doi:10.1158/1078-0432.CCR-04-0053.

83. Pathak, S., Meng, W.J., Zhang, H., Gnosa, S., Nandy, S.K., Adell, G., Holmlund, B. and Sun, X.F. (2014). "Tafazzin protein expression is associated with tumorigenesis and radiation response in rectal cancer: a study of Swedish clinical trial on preoperative radiotherapy." *PLoS One*. 9: e98317. PMID:24858921. doi:10.1371/journal.pone.0098317.

84. Sapandowski, A., Stope, M., Evert, K., Evert, M., Zimmermann, U., Peter, D., Page, I., Burchardt, M. and Schild, L. (2015). "Cardiolipin composition correlates with prostate cancer cell proliferation." *Mol*

Cell Biochem. 410: 175-85. PMID:26314254. doi:10.1007/s11010-015-2549-1.

85. Chen, M., Zhang, Y. and Zheng, P.S. (2017). “Tafazzin (TAZ) promotes the tumorigenicity of cervical cancer cells and inhibits apoptosis.” *PLoS One*. 12: e0177171. PMID:28489874. doi:10.1371/journal.pone.0177171.

86. Clarke, S.L., Bowron, A., Gonzalez, I.L., Groves, S.J., Newbury-Ecob, R., Clayton, N., Martin, R.P., Tsai-Goodman, B., Garratt, V., Ashworth, M., Bowen, V.M., McCurdy, K.R., Damin, M.K., Spencer, C.T., Toth, M.J., Kelley, R.I. and Steward, C.G. (2013). “Barth syndrome.” *Orphanet J Rare Dis*. 8: 23. PMID:23398819. doi:10.1186/1750-1172-8-23.

87. Aprikyan, A.A. and Khuchua, Z. (2013). “Advances in the understanding of Barth syndrome.” *Br J Haematol*. 161: 330-8. PMID:23432031. doi:10.1111/bjh.12271.

88. Barth, P.G., Scholte, H.R., Berden, J.A., Van der Klei-Van Moorsel, J.M., Luyt-Houwen, I.E., Van ‘t Veer-Korthof, E.T., Van der Harten, J.J. and Sobotka-Plojhar, M.A. (1983). “An X-linked mitochondrial disease affecting cardiac muscle, skeletal muscle and neutrophil leucocytes.” *J Neurol Sci*. 62: 327-55. PMID:6142097. 6142097

89. Barth, P.G., Valianpour, F., Bowen, V.M., Lam, J., Duran, M., Vaz, F.d.r.M. and Wanders, R.J.A. (2004). “X-linked cardioskeletal myopathy and neutropenia (Barth syndrome): An update.” *American Journal of Medical Genetics*. 126A: 349-354. doi:10.1002/ajmg.a.20660.

90. Cosson, L., Toutain, A., Simard, G., Kulik, W., Matyas, G., Guichet, A., Blasco, H., Maakaroun-Vermesse, Z., Vaillant, M.C., Le Caignec, C., Chantepie, A. and Labarthe, F. (2012). “Barth syndrome in a female patient.” *Mol Genet Metab*. 106: 115-20. PMID:22410210. doi:10.1016/j.ymgme.2012.01.015.

91. Cantlay, A.M., Shokrollahi, K., Allen, J.T., Lunt, P.W., Newbury-Ecob, R.A. and Steward, C.G. (1999). “Genetic analysis of the G4.5 gene in families with suspected Barth syndrome.” *J Pediatr*. 135: 311-5. PMID:10484795. 10484795

92. Barth Syndrome Foundation (2019). “What is Barth Syndrome?” Barth Syndrome Foundation. Updated: 3 March 2019. Accessed: 27 April, 2019. https://www.barthsyndrome.org/barthsyndrome/whatisbarthsyndrome.html

93. Caufield, J.H., Zhou, Y., Garlid, A.O., Setty, S.P., Liem, D.A., Cao, Q., Lee, J.M., Murali, S., Spendlove, S., Wang, W., Zhang, L., Sun, Y., Bui, A., Hermjakob, H., Watson, K.E. and Ping, P. (2018). “A reference set of curated biomedical data and metadata from clinical case reports.” *Sci Data*. 5: 180258. PMID:30457569. doi:10.1038/sdata.2018.258.

94. Caufield, J.H., Zhou, Y., Garlid, A.O., Setty, S.P., Liem, D.A., Cao, Q., Lee, J.M., Murali, S., Spendlove, S., Wang, W., Zhang, L., Sun, Y., Bui, A., Hermjakob, H., Watson, K.E. and Ping, P. (2018). “Data from: A reference set of curated biomedical data and metadata from clinical case reports.” Dryad Digital Repository. doi:10.5061/dryad.r36cn90.

95. Kulik, W., van Lenthe, H., Stet, F.S., Houtkooper, R.H., Kemp, H., Stone, J.E., Steward, C.G., Wanders, R.J. and Vaz, F.M. (2008). “Bloodspot assay using HPLC-tandem mass spectrometry for detection of Barth syndrome.” *Clin Chem*. 54: 371-8. PMID:18070816. doi:10.1373/clinchem.2007.095711.

96. Houtkooper, R.H., Rodenburg, R.J., Thiels, C., van Lenthe, H., Stet, F., Poll-The, B.T., Stone, J.E., Steward, C.G., Wanders, R.J., Smeitink, J., Kulik, W. and Vaz, F.M. (2009). “Cardiolipin and monolysocardiolipin analysis in fibroblasts, lymphocytes, and tissues using high-performance liquid chromatography-mass spectrometry as a diagnostic test for Barth syndrome.” *Anal Biochem*. 387: 230-7. PMID:19454236. doi:10.1016/j.ab.2009.01.032.

97. Bowron, A., Honeychurch, J., Williams, M., Tsai-Goodman, B., Clayton, N., Jones, L., Shortland, G.J., Qureshi, S.A., Heales, S.J. and Steward, C.G. (2015). "Barth syndrome without tetralinoleoyl cardiolipin deficiency: a possible ameliorated phenotype." *J Inherit Metab Dis*. 38: 279-86. PMID:25112388. doi:10.1007/s10545-014-9747-y.

98. Thompson, W.R., DeCroes, B., McClellan, R., Rubens, J., Vaz, F.M., Kristaponis, K., Avramopoulos, D. and Vernon, H.J. (2016). "New targets for monitoring and therapy in Barth syndrome." *Genet Med*. 18: 1001-10. PMID:26845103. doi:10.1038/gim.2015.204.

99. Ikon, N. and Ryan, R.O. (2017). "Barth Syndrome: Connecting Cardiolipin to Cardiomyopathy." *Lipids*. 52: 99-108. PMID:28070695. doi:10.1007/s11745-016-4229-7.

100. Roberts, A.E., Nixon, C., Steward, C.G., Gauvreau, K., Maisenbacher, M., Fletcher, M., Geva, J., Byrne, B.J. and Spencer, C.T. (2012). "The Barth Syndrome Registry: distinguishing disease characteristics and growth data from a longitudinal study." *Am J Med Genet A*. 158A: 2726-32. PMID:23045169. doi:10.1002/ajmg.a.35609.

101. Wang, G., McCain, M.L., Yang, L., He, A., Pasqualini, F.S., Agarwal, A., Yuan, H., Jiang, D., Zhang, D., Zangi, L., Geva, J., Roberts, A.E., Ma, Q., Ding, J., Chen, J., Wang, D.Z., Li, K., Wang, J., Wanders, R.J., Kulik, W., Vaz, F.M., Laflamme, M.A., Murry, C.E., Chien, K.R., Kelley, R.I., Church, G.M., Parker, K.K. and Pu, W.T. (2014). "Modeling the mitochondrial cardiomyopathy of Barth syndrome with induced pluripotent stem cell and heart-on-chip technologies." *Nat Med*. 20: 616-23. PMID:24813252. doi:10.1038/nm.3545.

102. McCanta, A.C., Chang, A.C. and Weiner, K. (2008). "Cardiomyopathy in a child with neutropenia and motor delay." *Curr Opin Pediatr*. 20: 605-7. PMID:18781126. doi:10.1097/MOP.0b013e32830a990a.

103. Masarone, D., Valente, F., Rubino, M., Vastarella, R., Gravino, R., Rea, A., Russo, M.G., Pacileo, G. and Limongelli, G. (2017). "Pediatric Heart Failure: A Practical Guide to Diagnosis and Management." *Pediatr Neonatol*. 58: 303-312. PMID:28279666. doi:10.1016/j.pedneo.2017.01.001.

104. Mangat, J., Lunnon-Wood, T., Rees, P., Elliott, M. and Burch, M. (2007). "Successful cardiac transplantation in Barth syndrome--single-centre experience of four patients." *Pediatr Transplant*. 11: 327-31. PMID:17430492. doi:10.1111/j.1399-3046.2006.00629.x.

105. Ronghe, M.D., Foot, A.B., Martin, R., Ashworth, M. and Steward, C.G. (2001). "Non-Epstein-Barr virus-associated T-cell lymphoma following cardiac transplantation for Barth syndrome." *Acta Paediatr*. 90: 584-6. PMID:11430723. 11430723

106. Soares, P., Rocha, G., Pissarra, S., Soares, H., Flor-de-Lima, F., Costa, S., Moura, C., Doria, S. and Guimaraes, H. (2017). "Neonatal dilated cardiomyopathy." *Rev Port Cardiol*. 36: 201-214. PMID:28256370. doi:10.1016/j.repc.2016.10.007.

107. Araco, M., Merlo, M., Carr-White, G. and Sinagra, G. (2017). "Genetic bases of dilated cardiomyopathy." *J Cardiovasc Med (Hagerstown)*. 18: 123-130. PMID:27661610. doi:10.2459/JCM.0000000000000432.

108. Zapala, B., Platek, T. and Wybranska, I. (2015). "A novel TAZ gene mutation and mosaicism in a Polish family with Barth syndrome." *Ann Hum Genet*. 79: 218-24. PMID:25776009. doi:10.1111/ahg.12108.

109. Shemisa, K., Li, J., Tam, M. and Barcena, J. (2013). "Left ventricular noncompaction cardiomyopathy." *Cardiovasc Diagn Ther*. 3: 170-5. PMID:24282766. doi:10.3978/j.issn.2223-3652.2013.05.04.

110. Bleyl, S.B., Mumford, B.R., Brown-Harrison, M.C., Pagotto, L.T., Carey, J.C., Pysher, T.J., Ward, K. and Chin, T.K. (1997). "Xq28-linked noncompaction of the left ventricular myocardium: prenatal diagnosis and pathologic analysis of affected individuals." *Am J Med Genet*. 72: 257-65. PMID:9332651. 9332651

111.Woiewodski, L., Ezon, D., Cooper, J. and Feingold, B. (2017). "Barth Syndrome with Late-Onset Cardiomyopathy: A Missed Opportunity for Diagnosis." *J Pediatr*. 183: 196-198. PMID:28108107. doi:10.1016/j.jpeds.2016.12.070.

112. Rigaud, C., Lebre, A.S., Touraine, R., Beaupain, B., Ottolenghi, C., Chabli, A., Ansquer, H., Ozsahin, H., Di Filippo, S., De Lonlay, P., Borm, B., Rivier, F., Vaillant, M.C., Mathieu-Dramard, M., Goldenberg, A., Viot, G., Charron, P., Rio, M., Bonnet, D. and Donadieu, J. (2013). "Natural history of Barth syndrome: a national cohort study of 22 patients." *Orphanet J Rare Dis*. 8: 70. PMID:23656970. doi:10.1186/1750-1172-8-70.

113. Spencer, C.T., Byrne, B.J., Bryant, R.M., Margossian, R., Maisenbacher, M., Breitenger, P., Benni, P.B., Redfearn, S., Marcus, E. and Cade, W.T. (2011). "Impaired cardiac reserve and severely diminished skeletal muscle O(2) utilization mediate exercise intolerance in Barth syndrome." *Am J Physiol Heart Circ Physiol*. 301: H2122-9. PMID:21873497. doi:10.1152/ajpheart.00479.2010.

114. Ades, L.C., Gedeon, A.K., Wilson, M.J., Latham, M., Partington, M.W., Mulley, J.C., Nelson, J., Lui, K. and Sillence, D.O. (1993). "Barth syndrome: clinical features and confirmation of gene localisation to distal Xq28." *Am J Med Genet*. 45: 327-34. PMID:8434619. doi:10.1002/ajmg.1320450309.

115. Christodoulou, J., McInnes, R.R., Jay, V., Wilson, G., Becker, L.E., Lehotay, D.C., Platt, B.A., Bridge, P.J., Robinson, B.H. and Clarke, J.T. (1994). "Barth syndrome: clinical observations and genetic linkage studies." *Am J Med Genet*. 50: 255-64. PMID:8042670. doi:10.1002/ajmg.1320500309.

116. Ferreira, C., Thompson, R. and Vernon, H. (1993). "Barth Syndrome." in: Adam, M.P., Ardinger, H.H., Pagon, R.A., Wallace, S.E., Bean, L.J.H., Stephens, K. and Amemiya, A. (Eds.), GeneReviews((R)). Seattle (WA). 25299040

117. Ino, T., Sherwood, W.G., Cutz, E., Benson, L.N., Rose, V. and Freedom, R.M. (1988). "Dilated cardiomyopathy with neutropenia, short stature, and abnormal carnitine metabolism." *J Pediatr*. 113: 511-4. PMID:3411399. 3411399

118. Mazzocco, M.M., Henry, A.E. and Kelly, R.I. (2007). "Barth syndrome is associated with a cognitive phenotype." *J Dev Behav Pediatr*. 28: 22-30. PMID:17353728. doi:10.1097/01.DBP.0000257519.79803.90.

119. Cole, L.K., Kim, J.H., Amoscato, A.A., Tyurina, Y.Y., Bay, R.H., Karimi, B., Siddiqui, T.J., Kagan, V.E., Hatch, G.M. and Kauppinen, T.M. (2018). "Aberrant cardiolipin metabolism is associated with cognitive deficiency and hippocampal alteration in tafazzin knockdown mice." *Biochim Biophys Acta Mol Basis Dis*. 1864: 3353-3367. PMID:30055293. doi:10.1016/j.bbadis.2018.07.022.

120. Corazzi, L. and Roberti, R. (2009). "Lipids of Brain Mitochondria." in: Lajtha, A., Tettamanti, G. and Goracci, G. (Eds.), Handbook of Neurochemistry and Molecular Neurobiology: Neural Lipids. Springer US, Boston, MA, pp. 199-221. 10.1007/978-0-387-30378-9_8.

121. Starkov, A.A., Andreyev, A.Y., Zhang, S.F., Starkova, N.N., Korneeva, M., Syromyatnikov, M. and Popov, V.N. (2014). "Scavenging of H2O2 by mouse brain mitochondria." *J Bioenerg Biomembr*. 46: 471-7. PMID:25248416. doi:10.1007/s10863-014-9581-9.

122. Su, B. and Ryan, R.O. (2014). "Metabolic biology of 3-methylglutaconic acid-uria: a new perspective."

J Inherit Metab Dis. 37: 359-68. PMID:24407466. doi:10.1007/s10545-013-9669-0.

123. Wortmann, S.B., Kluijtmans, L.A., Engelke, U.F., Wevers, R.A. and Morava, E. (2012). “The 3-methylglutaconic acidurias: what’s new?” *J Inherit Metab Dis*. 35: 13-22. PMID:20882351. doi:10.1007/s10545-010-9210-7.

124. Wortmann, S.B., Duran, M., Anikster, Y., Barth, P.G., Sperl, W., Zschocke, J., Morava, E. and Wevers, R.A. (2013). “Inborn errors of metabolism with 3-methylglutaconic aciduria as discriminative feature: proper classification and nomenclature.” *J Inherit Metab Dis*. 36: 923-8. PMID:23296368. doi:10.1007/s10545-012-9580-0.

125. Schmidt, M.R., Birkebaek, N., Gonzalez, I. and Sunde, L. (2004). “Barth syndrome without 3-methylglutaconic aciduria.” *Acta Paediatr*. 93: 419-21. PMID:15124852. 15124852

126. Reynolds, S. (2015). “Successful management of Barth syndrome: a systematic review highlighting the importance of a flexible and multidisciplinary approach.” *J Multidiscip Healthc*. 8: 345-58. PMID:26251611. doi:10.2147/JMDH.S54802.

127. Makaryan, V., Kulik, W., Vaz, F.M., Allen, C., Dror, Y., Dale, D.C. and Aprikyan, A.A. (2012). “The cellular and molecular mechanisms for neutropenia in Barth syndrome.” *Eur J Haematol*. 88: 195-209. PMID:22023389. doi:10.1111/j.1600-0609.2011.01725.x.

128. Makaryan, V., Dror, Y. and Aprikyan, A.A. (2009). “Loss of Tafazzin (TAZ) Function and Accelerated Apoptosis of Human Bone Marrow Stem and Myeloid Progenitors in Barth Syndrome.” *Blood*. 114: 549.

129. Steward, C.G., Groves, S.J., Taylor, C.T., Maisenbacher, M.K., Versluys, B., Newbury-Ecob, R.A., Ozsahin, H., Damin, M.K., Bowen, V.M., McCurdy, K.R., Mackey, M.C., Bolyard, A.A. and Dale, D.C (2019). “Neutropenia in Barth syndrome: characteristics, risks, and management.” *Curr Opin Hematol*. 26: 6-15. PMID:30451719. doi:10.1097/MOH.0000000000000472.

130. Dale, D.C., Bolyard, A.A., Schwinzer, B.G., Pracht, G., Bonilla, M.A., Boxer, L., Freedman, M.H., Donadieu, J., Kannourakis, G., Alter, B.P., Cham, B.P., Winkelstein, J., Kinsey, S.E., Zeidler, C. and Welte, K. (2006). “The Severe Chronic Neutropenia International Registry: 10-Year Follow-up Report.” *Support Cancer Ther*. 3: 220-31. PMID:18632498. doi:10.3816/SCT.2006.n.020.

131. Ferri, L., Donati, M.A., Funghini, S., Cavicchi, C., Pensato, V., Gellera, C., Natacci, F., Spaccini, L., Gasperini, S., Vaz, F.M., Cooper, D.N., Guerrini, R. and Morrone, A. (2015). “Intra-individual plasticity of the TAZ gene leading to different heritable mutations in siblings with Barth syndrome.” *Eur J Hum Genet*. 23: 1708-12. PMID:25782672. doi:10.1038/ejhg.2015.50.

132. Monteiro, J.P., Oliveira, P.J. and Jurado, A.S. (2013). “Mitochondrial membrane lipid remodeling in pathophysiology: a new target for diet and therapeutic interventions.” *Prog Lipid Res*. 52: 513-28. PMID:23827885. doi:10.1016/j.plipres.2013.06.002.

133. Ikon, N., Su, B., Hsu, F.F., Forte, T.M. and Ryan, R.O. (2015). “Exogenous cardiolipin localizes to mitochondria and prevents TAZ knockdown-induced apoptosis in myeloid progenitor cells.” *Biochem Biophys Res Commun*. 464: 580-5. PMID:26164234. doi:10.1016/j.bbrc.2015.07.012.

134. Ikon, N., Hsu, F.F., Shearer, J., Forte, T.M. and Ryan, R.O. (2018). “Evaluation of cardiolipin nanodisks as lipid replacement therapy for Barth syndrome.” *J Biomed Res*. 32: 107-112. PMID:29336355. doi:10.7555/JBR.32.20170094.

135. Taylor, W.A., Mejia, E.M., Mitchell, R.W., Choy, P.C., Sparagna, G.C. and Hatch, G.M. (2012). “Human trifunctional protein alpha links cardiolipin remodeling to beta-oxidation.” *PLoS One*. 7: e48628.

PMID:23152787. doi:10.1371/journal.pone.0048628.

136. Mejia, E.M., Zegallai, H., Bouchard, E.D., Banerji, V., Ravandi, A. and Hatch, G.M. (2018). “Expression of human monolysocardiolipin acyltransferase-1 improves mitochondrial function in Barth syndrome lymphoblasts.” *J Biol Chem*. 293: 7564-7577. PMID:29563154. doi:10.1074/jbc.RA117.001024.

137. Saric, A., Andreau, K., Armand, A.S., Moller, I.M. and Petit, P.X. (2015). “Barth Syndrome: From Mitochondrial Dysfunctions Associated with Aberrant Production of Reactive Oxygen Species to Pluripotent Stem Cell Studies.” *Front Genet*. 6: 359. PMID:26834781. doi:10.3389/fgene.2015.00359.

138. Dudek, J., Cheng, I.F., Balleininger, M., Vaz, F.M., Streckfuss-Bomeke, K., Hubscher, D., Vukotic, M., Wanders, R.J., Rehling, P. and Guan, K. (2013). “Cardiolipin deficiency affects respiratory chain function and organization in an induced pluripotent stem cell model of Barth syndrome.” *Stem Cell Res*. 11: 806-19. PMID:23792436. doi:10.1016/j.scr.2013.05.005.

139. Xu, Y., Sutachan, J.J., Plesken, H., Kelley, R.I. and Schlame, M. (2005). “Characterization of lymphoblast mitochondria from patients with Barth syndrome.” *Lab Invest*. 85: 823-30. PMID:15806137. doi:10.1038/labinvest.3700274.

140. He, Q., Harris, N., Ren, J. and Han, X. (2014). “Mitochondria-targeted antioxidant prevents cardiac dysfunction induced by tafazzin gene knockdown in cardiac myocytes.” *Oxid Med Cell Longev*. 2014: 654198. PMID:25247053. doi:10.1155/2014/654198.

141. Johnson, J.M., Ferrara, P.J., Verkerke, A.R.P., Coleman, C.B., Wentzler, E.J., Neufer, P.D., Kew, K.A., de Castro Bras, L.E. and Funai, K. (2018). “Targeted overexpression of catalase to mitochondria does not prevent cardioskeletal myopathy in Barth syndrome.” *J Mol Cell Cardiol*. 121: 94-102. PMID:30008435. doi:10.1016/j.yjmcc.2018.07.001.

142. Huang, Y., Powers, C., Moore, V., Schafer, C., Ren, M., Phoon, C.K., James, J.F., Glukhov, A.V., Javadov, S., Vaz, F.M., Jefferies, J.L., Strauss, A.W. and Khuchua, Z. (2017). “The PPAR pan-agonist bezafibrate ameliorates cardiomyopathy in a mouse model of Barth syndrome.” *Orphanet J Rare Dis*. 12: 49. PMID:28279226. doi:10.1186/s13023-017-0605-5.

143. Arbel, Y., Klempfner, R., Erez, A., Goldenberg, I., Benzekry, S., Shlomo, N., Fisman, E.Z., Tenenbaum, A. and Group, B.I.P.S. (2016). “Bezafibrate for the treatment of dyslipidemia in patients with coronary artery disease: 20-year mortality follow-up of the BIP randomized control trial.” *Cardiovasc Diabetol*. 15: 11. PMID:26794137. doi:10.1186/s12933-016-0332-6.

144. Teramoto, T., Shirai, K., Daida, H. and Yamada, N. (2012). “Effects of bezafibrate on lipid and glucose metabolism in dyslipidemic patients with diabetes: the J-BENEFIT study.” *Cardiovasc Diabetol*. 11: 29. PMID:22439599. doi:10.1186/1475-2840-11-29.

145. Schnepp, B.C., Clark, K.R., Klemanski, D.L., Pacak, C.A. and Johnson, P.R. (2003). “Genetic fate of recombinant adeno-associated virus vector genomes in muscle.” *J Virol*. 77: 3495-504. PMID:12610125. doi:10.1128/jvi.77.6.3495-3504.2003.

146. Mendell, J.R., Al-Zaidy, S., Shell, R., Arnold, W.D., Rodino-Klapac, L.R., Prior, T.W., Lowes, L., Alfano, L., Berry, K., Church, K., Kissel, J.T., Nagendran, S., L’Italien, J., Sproule, D.M., Wells, C., Cardenas, J.A., Heitzer, M.D., Kaspar, A., Corcoran, S., Braun, L., Likhite, S., Miranda, C., Meyer, K., Foust, K.D., Burghes, A.H.M. and Kaspar, B.K. (2017). “Single-Dose Gene-Replacement Therapy for Spinal Muscular Atrophy.” *N Engl J Med*. 377: 1713-1722. PMID:29091557. doi:10.1056/NEJMoa1706198.

147. Bish, L.T., Morine, K., Sleeper, M.M., Sanmiguel, J., Wu, D., Gao, G., Wilson, J.M. and Sweeney, H.L. (2008). “Adeno-associated virus (AAV) serotype 9 provides global cardiac gene transfer superior

to AAV1, AAV6, AAV7, and AAV8 in the mouse and rat." *Hum Gene Ther*. 19: 1359-68. PMID:18795839. doi:10.1089/hum.2008.123.

148. Suzuki-Hatano, S., Saha, M., Rizzo, S.A., Witko, R.L., Gosiker, B.J., Ramanathan, M., Soustek, M.S., Jones, M.D., Kang, P.B., Byrne, B.J., Cade, W.T. and Pacak, C.A. (2019). "AAV-Mediated TAZ Gene Replacement Restores Mitochondrial and Cardioskeletal Function in Barth Syndrome." *Hum Gene Ther*. 30: 139-154. PMID:30070157. doi:10.1089/hum.2018.020.

149. Suzuki-Hatano, S., Saha, M., Soustek, M.S., Kang, P.B., Byrne, B.J., Cade, W.T. and Pacak, C.A. (2019). "AAV9-TAZ Gene Replacement Ameliorates Cardiac TMT Proteomic Profiles in a Mouse Model of Barth Syndrome." *Mol Ther Methods Clin Dev*. 13: 167-179. PMID:30788385. doi:10.1016/j.omtm.2019.01.007.

150. Szeto, H.H. (2014). "First-in-class cardiolipin-protective compound as a therapeutic agent to restore mitochondrial bioenergetics." *Br J Pharmacol*. 171: 2029-50. PMID:24117165. doi:10.1111/bph.12461.

151. Birk, A.V., Chao, W.M., Bracken, C., Warren, J.D. and Szeto, H.H. (2014). "Targeting mitochondrial cardiolipin and the cytochrome c/cardiolipin complex to promote electron transport and optimize mitochondrial ATP synthesis." *Br J Pharmacol*. 171: 2017-28. PMID:24134698. doi:10.1111/bph.12468.

152. Sabbah, H.N., Stein, P.D., Kono, T., Gheorghiade, M., Levine, T.B., Jafri, S., Hawkins, E.T. and Goldstein, S. (1991). "A canine model of chronic heart failure produced by multiple sequential coronary microembolizations." *Am J Physiol*. 260: H1379-84. PMID:1826414. doi:10.1152/ajpheart.1991.260.4.H1379.

153. Sabbah, H.N., Gupta, R.C., Kohli, S., Wang, M., Hachem, S. and Zhang, K. (2016). "Chronic Therapy With Elamipretide (MTP-131), a Novel Mitochondria-Targeting Peptide, Improves Left Ventricular and Mitochondrial Function in Dogs With Advanced Heart Failure." *Circ Heart Fail*. 9: e002206. PMID:26839394. doi:10.1161/CIRCHEARTFAILURE.115.002206.

154. Sabbah, H.N., Gupta, R.C., Singh-Gupta, V. and Zhang, K. (2019). "Effects of elamipretide on skeletal muscle in dogs with experimentally induced heart failure." *ESC Heart Fail*. 6: 328-335. PMID:30688415. doi:10.1002/ehf2.12408.

155. Chatfield, K.C., Sparagna, G.C., Chau, S., Phillips, E.K., Ambardekar, A.V., Aftab, M., Mitchell, M.B., Sucharov, C.C., Miyamoto, S.D. and Stauffer, B.L. (2019). "Elamipretide Improves Mitochondrial Function in the Failing Human Heart." *JACC Basic Transl Sci*. 4: 147-157. PMID:31061916. doi:10.1016/j.jacbts.2018.12.005.

156. Daubert, M.A., Yow, E., Dunn, G., Marchev, S., Barnhart, H., Douglas, P.S., O'Connor, C., Goldstein, S., Udelson, J.E. and Sabbah, H.N. (2017). "Novel Mitochondria-Targeting Peptide in Heart Failure Treatment: A Randomized, Placebo-Controlled Trial of Elamipretide." *Circ Heart Fail*. 10. PMID:29217757. doi:10.1161/CIRCHEARTFAILURE.117.004389.

157. ClinicalTrials.gov [Internet] "Identifier NCT03098797: A Trial to Evaluate Safety, Tolerability and Efficacy of Elamipretide in Subjects With Barth Syndrome (TAZPOWER)." National Library of Medicine (US). Updated: 15 Feb, 2019. Accessed: 29 Apr. https://clinicaltrials.gov/ct2/show/NCT03098797

158. Shi, X., Liu, R., Basolo, F., Giannini, R., Shen, X., Teng, D., Guan, H., Shan, Z., Teng, W., Musholt, T.J., Al-Kuraya, K., Fugazzola, L., Colombo, C., Kebebew, E., Jarzab, B., Czarniecka, A., Bendlova, B., Sykorova, V., Sobrinho-Simoes, M., Soares, P., Shong, Y.K., Kim, T.Y., Cheng, S., Asa, S.L., Viola, D., Elisei, R., Yip, L., Mian, C., Vianello, F., Wang, Y., Zhao, S., Oler, G., Cerutti, J.M., Puxeddu, E., Qu, S., Wei, Q., Xu, H., O'Neill, C.J., Sywak, M.S., Clifton-Bligh, R., Lam, A.K., Riesco-Eizaguirre, G., Santisteban, P., Yu, H., Tallini, G., Holt, E.H., Vasko, V. and Xing, M. (2016). "Differential

Clinicopathological Risk and Prognosis of Major Papillary Thyroid Cancer Variants.” *J Clin Endocrinol Metab*. 101: 264-74. PMID:26529630. doi:10.1210/jc.2015-2917.

159. Warburg, O.H. (1969). “The prime cause and prevention of cancer.” K. Triltsch.

160. Kiebish, M.A., Han, X., Cheng, H., Chuang, J.H. and Seyfried, T.N. (2008). “Cardiolipin and electron transport chain abnormalities in mouse brain tumor mitochondria: lipidomic evidence supporting the Warburg theory of cancer.” *J Lipid Res*. 49: 2545-56. PMID:18703489. doi:10.1194/jlr.M800319-JLR200.

161. NCBI (2019). “TAZ tafazzin [Homo sapiens (human)] Gene ID: 6901.” National Center for Biotechnology Information.

References Cited: Clinical Case Reports

Adwani, S.S., Whitehead, B.F., Rees, P.G., Morris, A., Turnball, D.M., Elliott, M.J. and de Leval, M.R., 1997. Heart transplantation for Barth syndrome. Pediatr Cardiol 18, 143-5. PMID: 9049131. doi: 10.1007/s002469900135.

Aljishi, E. and Ali, F., 2010. Barth syndrome: an X-linked cardiomyopathy with a novel mutation. Indian J Pediatr 77, 1432-3. PMID: 20981509. doi: 10.1007/s12098-010-0222-y.

Alter, P. and Maisch, B., 2007. Non-compaction cardiomyopathy in an adult with hereditary spherocytosis. Eur J Heart Fail 9, 98-9. PMID: 16731037. doi: 10.1016/j.ejheart.2006.03.008.

Ances, B.M., Sullivan, J., Weigele, J.B., Hwang, V., Messe, S.R., Kasner, S.E. and Liebeskind, D.S., 2006. Stroke associated with Barth syndrome. J Child Neurol 21, 805-7. PMID: 16970891. doi: 10.1177/08830738060210090901.

Bachou, T., Giannakopoulos, A., Trapali, C., Vazeou, A. and Kattamis, A., 2009. A novel mutation in the G4.5 (TAZ) gene in a Greek patient with Barth syndrome. Blood Cells Mol Dis 42, 262-4. PMID: 19261493. doi: 10.1016/j.bcmd.2008.11.004.

Baksiene, M., Benusiene, E., Morkuniene, A., Ambrozaityte, L., Utkus, A. and Kucinskas, V., 2016. A novel intronic splice site tafazzin gene mutation detected prenatally in a family with Barth syndrome. Balkan J Med Genet 19, 95-100. PMID: 28289596. doi: 10.1515/bjmg-2016-0043.

Barth, P.G., Van den Bogert, C., Bolhuis, P.A., Scholte, H.R., van Gennip, A.H., Schutgens, R.B. and Ketel, A.G., 1996. X-linked cardioskeletal myopathy and neutropenia (Barth syndrome): respiratory-chain abnormalities in cultured fibroblasts. J Inherit Metab Dis 19, 157-60. PMID: 8739954.

Borna, N.N., Kishita, Y., Ishikawa, K., Nakada, K., Hayashi, J.I., Tokuzawa, Y., Kohda, M., Nyuzuki, H., Yamashita-Sugahara, Y., Nasu, T., Takeda, A., Murayama, K., Ohtake, A. and Okazaki, Y., 2017. A novel mutation in TAZ causes mitochondrial respiratory chain disorder without cardiomyopathy. J Hum Genet 62, 539-547. PMID: 28123175. doi: 10.1038/jhg.2016.165.

Bowron, A., Honeychurch, J., Williams, M., Tsai-Goodman, B., Clayton, N., Jones, L., Shortland, G.J., Qureshi, S.A., Heales, S.J. and Steward, C.G., 2015. Barth syndrome without tetralinoleoyl cardiolipin deficiency: a possible ameliorated phenotype. J Inherit Metab Dis 38, 279-86. PMID: 25112388. doi: 10.1007/s10545-014-9747-y.

Brady, A.N., Shehata, B.M. and Fernhoff, P.M., 2006. X-linked fetal cardiomyopathy caused by a novel mutation in the TAZ gene. Prenat Diagn 26, 462-5. PMID: 16548007. doi: 10.1002/pd.1438.

Brion, M., de Castro Lopez, M.J., Santori, M., Perez Munuzuri, A., Lopez Abel, B., Bana Souto, A.M., Martinez Soto, M.I. and Couce, M.L., 2016. Prospective and Retrospective Diagnosis of Barth Syndrome Aided by Next-Generation Sequencing. Am J Clin Pathol 145, 507-13. PMID: 27124939. doi: 10.1093/ajcp/aqw025.

Cardonick, E.H., Kuhlman, K., Ganz, E. and Pagotto, L.T., 1997. Prenatal clinical expression of 3-methylglutaconic aciduria: Barth syndrome. Prenat Diagn 17, 983-8. PMID: 9358581.

Christodoulou, J., McInnes, R.R., Jay, V., Wilson, G., Becker, L.E., Lehotay, D.C., Platt, B.A., Bridge, P.J., Robinson, B.H. and Clarke, J.T., 1994. Barth syndrome: clinical observations and genetic linkage studies. Am J Med Genet 50, 255-64. PMID: 8042670. doi: 10.1002/ajmg.1320500309.

Cosson, L., Toutain, A., Simard, G., Kulik, W., Matyas, G., Guichet, A., Blasco, H., Maakaroun-Vermesse, Z., Vaillant, M.C., Le Caignec, C., Chantepie, A. and Labarthe, F., 2012. Barth syndrome in a female patient. Mol Genet Metab 106, 115-20. PMID: 22410210. doi: 10.1016/j.ymgme.2012.01.015.

Dedieu, N., Giardini, A., Steward, C.G., Fenton, M., Karimova, A., Hsia, T.Y. and Burch, M., 2013. Successful mechanical circulatory support for 251 days in a child with intermittent severe neutropenia due to Barth syndrome. Pediatr Transplant 17, E46-9. PMID: 23190323. doi: 10.1111/petr.12027.

Fan, Y., Steller, J., Gonzalez, I.L., Kulik, W., Fox, M., Chang, R., Westerfield, B.A., Batra, A.S., Wang, R.Y., Gallant, N.M., Pena, L.S., Wang, H., Huang, T., Bhuta, S., Penny, D.J., McCabe, E.R. and Kimonis, V.E., 2013. A Novel Exonic Splicing Mutation in the TAZ (G4.5) Gene in a Case with Atypical Barth Syndrome. JIMD Rep 11, 99-106. PMID: 23606313. doi: 10.1007/8904_2013_228.

Ferri, L., Dionisi-Vici, C., Taurisano, R., Vaz, F.M., Guerrini, R. and Morrone, A., 2016. When silence is noise: infantile-onset Barth syndrome caused by a synonymous substitution affecting TAZ gene transcription. Clin Genet 90, 461-465. PMID: 26853223. doi: 10.1111/cge.12756.

Ferri, L., Donati, M.A., Funghini, S., Cavicchi, C., Pensato, V., Gellera, C., Natacci, F., Spaccini, L., Gasperini, S., Vaz, F.M., Cooper, D.N., Guerrini, R. and Morrone, A., 2015. Intra-individual plasticity of the TAZ gene leading to different heritable mutations in siblings with Barth syndrome. Eur J Hum Genet 23, 1708-12. PMID: 25782672. doi: 10.1038/ejhg.2015.50.

Ferri, L., Donati, M.A., Funghini, S., Malvagia, S., Catarzi, S., Lugli, L., Ragni, L., Bertini, E., Vaz, F.M., Cooper, D.N., Guerrini, R. and Morrone, A., 2013. New clinical and molecular insights on Barth syndrome. Orphanet J Rare Dis 8, 27. PMID: 23409742. doi: 10.1186/1750-1172-8-27.

Folsi, V., Miglietti, N., Lombardi, A., Boccacci, S., Utyatnikova, T., Donati, C., Squassabia, L., Gazzola, L., Bosio, I., Borghi, A., Grassi, V., Notarangelo, L.D. and Plebani, A., 2014. Cardiomyopathy in a male patient with neutropenia and growth delay. Ital J Pediatr 40, 45. PMID: 24887148. doi: 10.1186/1824-7288-40-45.

Hanke, S.P., Gardner, A.B., Lombardi, J.P., Manning, P.B., Nelson, D.P., Towbin, J.A., Jefferies, J.L. and Lorts, A., 2012. Left ventricular noncompaction cardiomyopathy in Barth syndrome: an example of an undulating cardiac phenotype necessitating mechanical circulatory support as a bridge to transplantation. Pediatr Cardiol 33, 1430-4. PMID: 22427193. doi: 10.1007/s00246-012-0258-z.

Huang, S.C., Wu, E.T., Chiu, S.N., Hwu, W.L., Wu, M.H. and Wang, S.S., 2008. Mitral annuloplasty in an infant with Barth syndrome and severe mitral insufficiency: first case report and determination of annular diameter. J Thorac Cardiovasc Surg 136, 1095-7. PMID: 18954662. doi: 10.1016/j.jtcvs.2008.01.031.

Huhta, J.C., Pomerance, H.H. and Barness, E.G., 2005. Clinicopathologic conference: Barth Syndrome. Fetal Pediatr Pathol 24, 239-54. PMID: 16396830. doi: 10.1080/15227950500405429.

Imai-Okazaki, A., Kishita, Y., Kohda, M., Yatsuka, Y., Hirata, T., Mizuno, Y., Harashima, H., Hirono, K., Ichida, F., Noguchi, A., Yoshida, M., Tokorodani, C., Nishiuchi, R., Takeda, A., Nakaya, A., Sakata, Y., Murayama, K., Ohtake, A. and Okazaki, Y., 2018. Barth Syndrome: Different Approaches to Diagnosis. J Pediatr 193, 256-260. PMID: 29249525. doi: 10.1016/j.jpeds.2017.09.075.

Karkucinska-Wieckowska, A., Trubicka, J., Werner, B., Kokoszynska, K., Pajdowska, M., Pronicki, M., Czarnowska, E., Lebiedzinska, M., Sykut-Cegielska, J., Ziolkowska, L., Jaron, W., Dobrzanska, A., Ciara, E., Wieckowski, M.R. and Pronicka, E., 2013. Left ventricular noncompaction (LVNC) and low mitochondrial membrane potential are specific for Barth syndrome. J Inherit Metab Dis 36, 929-37. PMID: 23361305. doi: 10.1007/s10545-013-9584-4.

Katsushima, Y., Fujiwara, I., Sakamoto, O., Ohura, T., Miyabayashi, S., Ohnuma, A., Yamaguchi, S. and Iinuma, K., 2002. Normal pituitary function in a Japanese patient with Barth syndrome. Eur J Pediatr 161, 67-8. PMID: 11808885.

Kelley, R.I., Cheatham, J.P., Clark, B.J., Nigro, M.A., Powell, B.R., Sherwood, G.W., Sladky, J.T. and Swisher, W.P., 1991. X-linked dilated cardiomyopathy with neutropenia, growth retardation, and 3-methylglutaconic aciduria. J Pediatr 119, 738-47. PMID: 1719174.

Kim, G.B., Kwon, B.S., Bae, E.J., Noh, C.I., Seong, M.W. and Park, S.S., 2013. A novel mutation of the TAZ gene in Barth syndrome: acute exacerbation after contrast-dye injection. J Korean Med Sci 28, 784-7. PMID: 23678274. doi: 10.3346/jkms.2013.28.5.784.

Kirwin, S.M., Vinette, K.M., Schwartz, S.B., Funanage, V.L. and Gonzalez, I.L., 2007. Multiple transmissions of Barth syndrome through an oocyte donor with a de novo TAZ mutation. Fertil Steril 87, 976.e5-7. PMID: 17241629. doi: 10.1016/j.fertnstert.2006.07.1543.

Lindenbaum, R.H., Andrews, P.S. and Khan, A.S., 1973. Two cases of endocardial fibroelastosis--possible x-linked determination. Br Heart J 35, 38-40. PMID: 4685904.

Man, E., Lafferty, K.A., Funke, B.H., Lun, K.S., Chan, S.Y., Chau, A.K. and Chung, B.H., 2013. NGS identifies TAZ mutation in a family with X-linked dilated cardiomyopathy. BMJ Case Rep 2013. PMID: 23345479. doi: 10.1136/bcr-2012-007529.

Marziliano, N., Mannarino, S., Nespoli, L., Diegoli, M., Pasotti, M., Malattia, C., Grasso, M., Pilotto, A., Porcu, E., Raisaro, A., Raineri, C., Dore, R., Maggio, P.P., Brega, A. and Arbustini, E., 2007. Barth syndrome associated with compound hemizygosity and heterozygosity of the TAZ and LDB3 genes. Am J Med Genet A 143a, 907-15. PMID: 17394203. doi: 10.1002/ajmg.a.31653.

Mazurova, S., Tesarova, M., Magner, M., Houstkova, H., Hansikova, H., Augustinova, J., Tomek, V., Vondrackova, A., Zeman, J. and Honzik, T., 2013. Novel mutations in the TAZ gene in patients with Barth syndrome. Prague Med Rep 114, 139-53. PMID: 24093814. doi: 10.14712/23362936.2014.16.

McCanta, A.C., Chang, A.C. and Weiner, K., 2008. Cardiomyopathy in a child with neutropenia and motor delay. Curr Opin Pediatr 20, 605-7. PMID: 18781126. doi: 10.1097/MOP.0b013e32830a990a.

Momoi, N., Chang, B., Takeda, I., Aoyagi, Y., Endo, K. and Ichida, F., 2012. Differing clinical courses and outcomes in two siblings with Barth syndrome and left ventricular noncompaction. Eur J Pediatr 171, 515-20. PMID: 21987083. doi: 10.1007/s00431-011-1597-0.

Rigaud, C., Lebre, A.S., Touraine, R., Beaupain, B., Ottolenghi, C., Chabli, A., Ansquer, H., Ozsahin, H., Di

Filippo, S., De Lonlay, P., Borm, B., Rivier, F., Vaillant, M.C., Mathieu-Dramard, M., Goldenberg, A., Viot, G., Charron, P., Rio, M., Bonnet, D. and Donadieu, J., 2013. Natural history of Barth syndrome: a national cohort study of 22 patients. Orphanet J Rare Dis 8, 70. PMID: 23656970. doi: 10.1186/1750-1172-8-70.

Ronghe, M.D., Foot, A.B., Martin, R., Ashworth, M. and Steward, C.G., 2001. Non-Epstein-Barr virus-associated T-cell lymphoma following cardiac transplantation for Barth syndrome. Acta Paediatr 90, 584-6. PMID: 11430723.

Ronvelia, D., Greenwood, J., Platt, J., Hakim, S. and Zaragoza, M.V., 2012. Intrafamilial variability for novel TAZ gene mutation: Barth syndrome with dilated cardiomyopathy and heart failure in an infant and left ventricular noncompaction in his great-uncle. Mol Genet Metab 107, 428-32. PMID: 23031367. doi: 10.1016/j.ymgme.2012.09.013.

Rugolotto, S., Prioli, M.D., Toniolo, D., Pellegrino, P., Catuogno, S. and Burlina, A.B., 2003. Long-term treatment of Barth syndrome with pantothenic acid: a retrospective study. Mol Genet Metab 80, 408-11. PMID: 14654353.

Sabater-Molina, M., Guillen-Navarro, E., Garcia-Molina, E., Ballesta-Martinez, M.J., Escudero, F. and Ruiz-Espejo, F., 2013. Barth syndrome in adulthood: a clinical case. Rev Esp Cardiol (Engl Ed) 66, 68-70. PMID: 22999963. doi: 10.1016/j.recesp.2012.05.015.

Sakamoto, O., Kitoh, T., Ohura, T., Ohya, N. and Iinuma, K., 2002. Novel missense mutation (R94S) in the TAZ (G4.5) gene in a Japanese patient with Barth syndrome. J Hum Genet 47, 229-31. PMID: 12032589. doi: 10.1007/s100380200030.

Schmidt, M.R., Birkebaek, N., Gonzalez, I. and Sunde, L., 2004. Barth syndrome without 3-methylglutaconic aciduria. Acta Paediatr 93, 419-21. PMID: 15124852.

Singh, H.R., Yang, Z., Siddiqui, S., Pena, L.S., Westerfield, B.H., Fan, Y., Towbin, J.A. and Vatta, M., 2009. A novel Alu-mediated Xq28 microdeletion ablates TAZ and partially deletes DNL1L in a patient with Barth syndrome. Am J Med Genet A 149a, 1082-5. PMID: 19396829. doi: 10.1002/ajmg.a.32822.

Steward, C.G., Newbury-Ecob, R.A., Hastings, R., Smithson, S.F., Tsai-Goodman, B., Quarrell, O.W., Kulik, W., Wanders, R., Pennock, M., Williams, M., Cresswell, J.L., Gonzalez, I.L. and Brennan, P., 2010. Barth syndrome: an X-linked cause of fetal cardiomyopathy and stillbirth. Prenat Diagn 30, 970-6. PMID: 20812380. doi: 10.1002/pd.2599.

Sweeney, R.T., Davis, G.J. and Noonan, J.A., 2008. Cardiomyopathy of unknown etiology: Barth syndrome unrecognized. Congenit Heart Dis 3, 443-8. PMID: 19037987. doi: 10.1111/j.1747-0803.2008.00226.x.

Takeda, A., Sudo, A., Yamada, M., Yamazawa, H., Izumi, G., Nishino, I. and Ariga, T., 2011. Barth syndrome diagnosed in the subclinical stage of heart failure based on the presence of lipid storage myopathy and isolated noncompaction of the ventricular myocardium. Eur J Pediatr 170, 1481-4. PMID: 21932011. doi: 10.1007/s00431-011-1576-5.

Thiels, C., Fleger, M., Huemer, M., Rodenburg, R.J., Vaz, F.M., Houtkooper, R.H., Haack, T.B., Prokisch, H., Feichtinger, R.G., Lucke, T., Mayr, J.A. and Wortmann, S.B., 2016. Atypical Clinical Presentations of TAZ Mutations: An Underdiagnosed Cause of Growth Retardation? JIMD Rep 29, 89-93. PMID: 26724946. doi: 10.1007/8904_2015_525.

Valianpour, F., Wanders, R.J., Overmars, H., Vreken, P., Van Gennip, A.H., Baas, F., Plecko, B., Santer, R., Becker, K. and Barth, P.G., 2002. Cardiolipin deficiency in X-linked cardioskeletal myopathy and neutropenia (Barth syndrome, MIM 302060): a study in cultured skin fibroblasts. J Pediatr 141, 729-

33. PMID: 12410207. doi: 10.1067/mpd.2002.129174.

Vesel, S., Stopar-Obreza, M., Trebusak-Podkrajsek, K., Jazbec, J., Podnar, T. and Battelino, T., 2003. A novel mutation in the G4.5 (TAZ) gene in a kindred with Barth syndrome. Eur J Hum Genet 11, 97-101. PMID: 12529714. doi: 10.1038/sj.ejhg.5200926.

Wang, J., Guo, Y., Huang, M., Zhang, Z., Zhu, J., Liu, T., Shi, L., Li, F., Huang, H. and Fu, L., 2017. Identification of TAZ mutations in pediatric patients with cardiomyopathy by targeted next-generation sequencing in a Chinese cohort. Orphanet J Rare Dis 12, 26. PMID: 28183324. doi: 10.1186/s13023-016-0562-4.

Woiewodski, L., Ezon, D., Cooper, J. and Feingold, B., 2017. Barth Syndrome with Late-Onset Cardiomyopathy: A Missed Opportunity for Diagnosis. J Pediatr 183, 196-198. PMID: 28108107. doi: 10.1016/j.jpeds.2016.12.070.

Yen, T.Y., Hwu, W.L., Chien, Y.H., Wu, M.H., Lin, M.T., Tsao, L.Y., Hsieh, W.S. and Lee, N.C., 2008. Acute metabolic decompensation and sudden death in Barth syndrome: report of a family and a literature review. Eur J Pediatr 167, 941-4. PMID: 17846786. doi: 10.1007/s00431-007-0592-y.

Yoo, T.Y., Kim, M.R., Son, J.S., Lee, R., Bae, S.H., Chung, S., Kim, K.S., Seong, M.W. and Park, S.S., 2016. Identification of a Novel De Novo Mutation of the TAZ Gene in a Korean Patient with Barth Syndrome. J Cardiovasc Ultrasound 24, 153-7. PMID: 27358708. doi: 10.4250/jcu.2016.24.2.153.

Zapala, B., Platek, T. and Wybranska, I., 2015. A novel TAZ gene mutation and mosaicism in a Polish family with Barth syndrome. Ann Hum Genet 79, 218-24. PMID: 25776009. doi: 10.1111/ahg.12108.

Chapter VII

Conclusions and Future Directions

Chapter 7. Conclusion

At the outset of this collection of projects, we recognized that mitochondrial knowledge is scattered, fragmented, and unstructured. This presents a barrier for researchers, clinicians, data scientists, and students attempting to find and access information to facilitate their contributions to science and patient care. Clinical case reports (CCRs) present a unique challenge due to the complexity of clinical presentations and clinical language, the inconsistency and lack of standardization among reports, and the inherent absence of structure in text data. These characteristics prevent investigators from extracting the rich clinical insights contained within, and it prevents integrating CCRs into computational pipelines and machine learning models of disease. We aimed to make knowledge pertaining to mitochondrial biology and related diseases more FAIR (Findable, Accessible, Interoperable, and Reusable) through standardization, integrating existing ontologies, knowledgebases, and coding systems, and imposing structure on otherwise unstructured text data. We focused on elevating the quantity, quality, and structure of FAIR knowledge on mitochondrial genes and proteins as well as rare mitochondrial diseases (RMDs).

We devised a multifaceted set of approaches to accomplish this overarching goal, including assembling a collection of informatics tools and resources pertaining to mitochondria, addressing the lack of mitochondrial coverage on Gene Wiki, designing a standardized approach for supplementing CCR corpora with highly structured metadata, and establishing a knowledge platform to house those data and integrate with existing ontologies and knowledgebases. We used a combination of all of these resources, approaches, and resulting data to conduct a thorough review of Barth syndrome, an RMD of particular interest because of its prominent cardiovascular phenotypes and deep underpinnings in mitochondrial form and function.

Many informatics resources are available that are either mitochondria-specific or contain mitochondrial subsets within the data. Most are active and updated regularly, but some are underutilized, unintegrated, and/or defunct or out of date. The review presented in **Chapter 2** provides an extensive index and analysis of a wide variety of tools and resources, including large, generalized omics databases with subsets of mitochondrial data, mitochondria-specific resources, and informatics tools and pipelines for data processing and analysis. These comprise the key components of a modern mitochondrial researcher's informatics tool chest.

Despite the availability of ample resources for the study of mitochondrial biology and copious biomedical knowledge contained in the literature, we found that mitochondrial genes and proteins were severely underrepresented on Gene Wiki [1, 2]. We addressed the gap in coverage of this vital organelle by evaluating 672 proteins identified in comprehensive mass spectrometry datasets of the cardiac mitochondria proteome [3-5]. The "Mitochondrial Gene Wiki Project" contributed over 4MB of content and nearly 6,000 references across 541 pages on mitochondrial genes and proteins to improve the findability and accessibility of essential mitochondrial knowledge, papers, and data resources for each entity. The Mitochondrial Gene Wiki Project became a key training effort for new members and volunteers in the lab as well, providing deep physiological knowledge and clinical insight for students from all educational levels, as well as training in critical aspects of research, from curating and evaluating literature to synthesizing knowledge and writing appropriate scientific communications.

Students, investigators, and patients alike commonly begin a research project on Wikipedia, in hopes of finding a well-organized, highly informative, and comprehensively cited article on their gene or protein of interest. Students and researchers can use this article as a launch point to find interacting partners through integration with IntAct [6], related pathways on Reactome [7], or pertinent articles on PubMed through the links in the references section. Clinicians and medical students also benefit from the biomedical content on Wikipedia as a quick a reliable source of information on pharmaceuticals, therapeutics, disease symptomology, and biological entities ranging from genes to proteins to metabolites. The structured data contained within Semantic Wiki Links [8-10] and the integrated Wikidata [11] entities also provides data scientists with an avenue for extracting detailed relationships between entities. Patients or their family members might be visiting to get a sense of what a particular diagnosis for a rare mitochondrial disease means and what kind of patient advocacy groups exist. Each of these key stakeholder groups benefits from the comprehensive representation of each mitochondrial gene, protein, channel, pathway, processes, and diseases.

The Gene Wiki Project also presents significant opportunities in training and education beyond the focused efforts undertaken here. In collaboration with the University of California, Los Angeles School of Life Sciences, we are developing a streamlined implementation of our Gene Wiki training methodologies for use in the Life Sciences 7C Physiology and Human Biology course. This course is the culmination of a year-long series that forms the core curriculum of a variety of undergraduate life science majors. The associated weekly discussion sessions give students the opportunity to work with a smaller group of classmates, guided by graduate student Teaching Assistants (TAs) from a wide range of research backgrounds. These sessions are a perfect setting for group work on Gene Wiki pages. Many students will gain their first experience in using PubMed, critically evaluating scientific publications, interacting with informatics resources, and conducting collaborative writing efforts, all of which are vital skills for future researchers and clinicians. Over 1,200 students take the LS Core Education Courses each year, providing a significant pool of potential contributors to this essential resource. Implementing these training efforts will ensure that the students will encounter the most prominent resources and become familiar with some of the most crucial activities for developing one’s scientific career. Furthermore, this presents the opportunity to introduce the FAIR Principles and data science approaches to a wide swath of students at one of the top universities in the country, amplifying the reach and consideration of the increasingly complex world of data and scientific knowledge.

Natural language generation (NLG) with GPT-2 [12, 13] presents an interesting avenue of research with potential application to Gene Wiki writing tasks. Given a prompt, the GPT-2 algorithm is capable of generating coherent paragraphs of text after training on vast stores of text data. This technology is still in its infancy, but it does show promise for processing large amounts of information about a gene or protein and generating a summary of those contents. NLG-derived gene summaries could identify and summarize a large quantity of relevant research articles, reducing the time devoted to manual curation efforts. The process would require human guidance and careful manual review to ensure the accuracy of the content, but it would greatly increase the comprehensiveness of Gene Wiki articles and the representation of scientific literature in the public domain. Gene Wiki also provides an ideal testing and training ground for the development of this nascent technology using a semi-supervised approach.

We amplified the unstructured text data in CCRs by designing a standardized metadata template with which to extract detailed and structured metadata characterizing the clinical presentations, treatment regimens, and diagnostic procedure described in the reports. We generated MACCRs (Metadata Annotated from CCRs) templates for over 3,100 CCRs across a wide range of diseases to establish a large dataset of structured data with which to study clinical language. For a subset of 384 CCRs on 8 RMDs, we further extracted and codified all patient symptoms from each report using the clinical codes from the 10th and 11th revisions of the International Statistical Classification of Diseases and Related Health Problems (ICD). The manually curated and annotated datasets are now housed on the newly formed MitoCases RMD Knowledge Platform (http://mitocases.org/).

MitoCases provides a more effective way to search for clinical reports on RMDs by symptoms, genetics, age rage, gender, or any other medical term contained within the reports. The extensive case information available on each report permits granular searches that are not possible when restricted to Medical Subject Heading (MeSH) terms and the content in titles and abstracts for indexing. This is of particular use to clinicians searching for cases similar to a patient they are seeing. Searching for CCRs with a set of symptoms as keywords on PubMed typically returns no results – MeSH was simply not designed to accommodate this level of detail, and the abstracts rarely provide extensive descriptions of symptoms. Physician investigators and researchers can also make use of this feature when attempting to construct an *in silico* cohort of a disease or trying to find cases in which a certain drug was used or a particular protein was measured or identified.

By demonstrating the utility of the MitoCases platform for CCRs on mitochondrial diseases and engaging those clinicians who undertake the effort to write these reports, we hope to inspire them to include symptomology metadata in their future publications and to lobby publishers to use the metadata for indexing purposes. One possibility is to simply include a patient’s coded ICD symptom list from their EHR, which presents a very low barrier to entry into this effort. Our initial outreach efforts have focused on the mitochondrial disease community at the United Mitochondrial Disease Foundation symposia and we aim to expand these efforts to other biomedical communities and clinical societies. It is important to note that the structured metadata for CCRs and EHRs should not replace the unstructured patient narratives but act instead as supplemental material to bolster the FAIRness of these documents. Unstructured narratives provide a nuanced and detailed view of patients' medical history, as well as the lines of reasoning undertaken by clinicians in their differential diagnosis efforts [14]. Combining these narratives with a structured component representing symptomology, disease diagnosis, genetic background, and treatment regimens serves to amplify the utility of CCRs and leverage their inherent value by making them more readily Findable, Accessible, Interoperable, and, ultimately Reusable.

Data scientists interested in studying clinical language may find particular utility in the large set of structured clinical narratives across a diverse set of presentations. This provides the opportunity to work with well-structured biomedical data and can serve as useful testing grounds for machine learning models, natural language processing (NLP) applications, named entity recognition (NER) [15, 16], and phrase mining [17]. These technologies might be applied to identify distinct entities and phrases within clinical text referring to patient symptoms, drug treatments, genes, proteins, and disease names. These approaches can be further enhanced via entity typing (ClusType) [18] to identify classes of entities detected through NER and attribute detection methods (MetaPAD) [19] to contextualize the conditions around observed entities and their interactions. With defined entity types, corpus-based set expansion (SetExpan) [20]g ontologies and knowledgebases may be used to discover additional instances of entities with matching features throughout the corpus of RMD CCRs. The projects presented here welcome the development of text-mining and automation tools to achieve a more comprehensive representation of the 58 known rare mitochondrial diseases and their available case presentations.

Text corpora standardization provides the basis for efficient document search and curation by exposing contents to search algorithms and enabling integration into knowledge graphs. The overarching aim of this project is to enable text data contained within CCRs to be machine-readable, facilitating downstream analysis and meta-analysis. There has been extensive development of metadata templates for structured data [21-24], but standardization of text data can only be accomplished and implemented with community consent and participation. The genomics and proteomics fields have seen significant successes in standardization by way of a community-driven process that identifies challenges, builds consensus, and leads to solutions. We envision clinicians and their co-authors as well as biomedical and data science researchers acting as an interdisciplinary community to establish and implement widely used standards for structuring text data. MitoCases is a model for wider implementation of metadata standard templates that will increase the utility of CCRs and our ability to derive knowledge from the clinical insights contained within. We recognize that a universal metadata template for all CCRs across every disease is impractical. Instead, we envision an opportunity to engage research and biomedical communities to form working groups that can leverage domain-specific expertise to create a compendium of templates capable of handling text data from a wide range of scientific and clinical documents. Ideally, the National Library of Medicine (NLM) could become involved and eventually incorporate more detailed metadata on PubMed for greater indexing capabilities of CCRs and other scientific literature.

These efforts depend upon dedication to the FAIR Principles throughout research such that all pieces of knowledge and data may be incorporated into the larger scheme of a holistic scientific knowledge representation. Ultimately, we envision the representation of all clinical and biomedical knowledge in a fully integrated knowledge graph that incorporates basic scientific concepts, experimental data, and clinical insights from all corners of research. With such a knowledge model, we could identify points of contention, opposing mechanistic hypotheses, or wholly missing pieces of critical knowledge. This would provide a roadmap to those explorations in need of further attention and resolution, driving research groups to test reproducibility, acquire more positive evidence for a claim, or to reject the underlying hypothesis. We will gain a deeper mechanistic understanding of genetics, protein biology, and disease presentations through the instantiation and development of such a resource, paving the way for improved patient care.

References:

1. Huss, J.W., 3rd, Lindenbaum, P., Martone, M., Roberts, D., Pizarro, A., Valafar, F., Hogenesch, J.B. and Su, A.I. (2010). “The Gene Wiki: community intelligence applied to human gene annotation.” *Nucleic Acids Res*. 38: D633-9. PMID:19755503. doi:10.1093/nar/gkp760.

2. Huss, J.W., 3rd, Orozco, C., Goodale, J., Wu, C., Batalov, S., Vickers, T.J., Valafar, F. and Su, A.I. (2008). “A gene wiki for community annotation of gene function.” *PLoS Biol*. 6: e175. PMID:18613750. doi:10.1371/journal.pbio.0060175.

3. Lau, E., Cao, Q., Ng, D.C., Bleakley, B.J., Dincer, T.U., Bot, B.M., Wang, D., Liem, D.A., Lam, M.P., Ge, J. and Ping, P. (2016). “A large dataset of protein dynamics in the mammalian heart proteome.” *Sci Data*. 3: 160015. PMID:26977904. doi:10.1038/sdata.2016.15.

4. Lotz, C., Lin, A.J., Black, C.M., Zhang, J., Lau, E., Deng, N., Wang, Y., Zong, N.C., Choi, J.H., Xu, T., Liem, D.A., Korge, P., Weiss, J.N., Hermjakob, H., Yates, J.R., 3rd, Apweiler, R. and Ping, P. (2014). “Characterization, design, and function of the mitochondrial proteome: from organs to organisms.” *J Proteome Res*. 13: 433-46. PMID:24070373. doi:10.1021/pr400539j.

5. Zong, N.C., Li, H., Li, H., Lam, M.P., Jimenez, R.C., Kim, C.S., Deng, N., Kim, A.K., Choi, J.H., Zelaya, I., Liem, D., Meyer, D., Odeberg, J., Fang, C., Lu, H.J., Xu, T., Weiss, J., Duan, H., Uhlen, M., Yates, J.R., 3rd, Apweiler, R., Ge, J., Hermjakob, H. and Ping, P. (2013). “Integration of cardiac proteome biology and medicine by a specialized knowledgebase.” *Circ Res*. 113: 1043-53. PMID:23965338. doi:10.1161/CIRCRESAHA.113.301151.

6. Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D. and Apweiler, R. (2004). “IntAct: an open source molecular interaction database.” *Nucleic Acids Res*. 32: D452-5. PMID:14681455. doi:10.1093/nar/gkh052.

7. Fabregat, A., Jupe, S., Matthews, L., Sidiropoulos, K., Gillespie, M., Garapati, P., Haw, R., Jassal, B., Korninger, F., May, B., Milacic, M., Roca, C.D., Rothfels, K., Sevilla, C., Shamovsky, V., Shorser, S., Varusai, T., Viteri, G., Weiser, J., Wu, G., Stein, L., Hermjakob, H. and D’Eustachio, P. (2018). “The Reactome Pathway Knowledgebase.” *Nucleic acids research*. 46: D649-D655. doi:10.1093/nar/gkx1132.

8. Good, B.M., Clarke, E.L., Loguercio, S. and Su, A.I. (2012). “Building a biomedical semantic network in Wikipedia with Semantic Wiki Links.” *Database : the journal of biological databases and curation*. 2012: bar060-bar060. PMID:22434829. doi:10.1093/database/bar060.

9. Krötzsch, M., Vrandečić, D. and Völkel, M. (2006). “Semantic mediawiki.” International semantic web conference. 935-942.

10. Völkel, M., Krötzsch, M., Vrandecic, D., Haller, H. and Studer, R. (2006). “Semantic wikipedia.” Proceedings of the 15th international conference on World Wide Web. 585-594.

11. Burgstaller-Muehlbacher, S., Waagmeester, A., Mitraka, E., Turner, J., Putman, T., Leong, J., Naik, C., Pavlidis, P., Schriml, L., Good, B.M. and Su, A.I. (2016). “Wikidata as a semantic framework for the Gene Wiki initiative.” *Database : the journal of biological databases and curation*. 2016: baw015. PMID:26989148. doi:10.1093/database/baw015.

12. Sutskever, A., Radford, J., Wu, D., Amodei, D., Amodei, J., Clark, M. and Brundage, I. (2019). “Better Language Models and Their Implications.” Updated: 2019-02-14. Accessed: https://openai.com/blog/better-language-models/

13. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D. and Sutskever, I. (2019). “Language models are unsupervised multitask learners.” *OpenAI Blog*. 1.

14. Barry, J. (2010). “Value of unstructured patient narratives.” *Health Manag Technol*. 31: 6-7.

15. Grishman, R. and Sundheim, B. (1996). “Message Understanding Conference-6.” Proceedings of the 16th conference on Computational linguistics -. 466. Morristown, NJ, USA. 1996.

16. Tjong Kim Sang, E.F. and De Meulder, F. (2003). “Introduction to the CoNLL-2003 shared task.” Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003 -. 142-147. Morristown, NJ, USA. 2003.

17. Liu, J., Shang, J., Wang, C., Ren, X. and Han, J. (2015). “Mining Quality Phrases from Massive Text Corpora.” *Proc ACM SIGMOD Int Conf Manag Data*. 2015: 1729-1744. PMID:26705375. doi:10.1145/2723372.2751523.

18. Ren, X., El-Kishky, A., Wang, C., Tao, F., Voss, C.R., Ji, H. and Han, J. (2015). “ClusType: Effective Entity Recognition and Typing by Relation Phrase-Based Clustering.” *KDD*. 2015: 995-1004. PMID:26705503. doi:10.1145/2783258.2783362.

19. Jiang, M., Shang, J., Cassidy, T., Ren, X., Kaplan, L., Hanratty, T. and Han, J. (2019). “MetaPAD: Meta pattern discovery from massive text corpora.” 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 2017. 877-886.

20. Shen, J., Wu, Z., Lei, D., Shang, J., Ren, X. and Han, J. (2017). “SetExpan: Corpus-Based Set Expansion via Context Feature Selection and Rank Ensemble.” Machine Learning and Knowledge Discovery in Databases. 288-304. Cham. 2017//.

21. Goncalves, R.S. and Musen, M.A. (2019). “The variable quality of metadata about biological samples used in biomedical experiments.” *Sci Data*. 6: 190021. PMID:30778255. doi:10.1038/sdata.2019.21.

22. Martinez-Romero, M., O’Connor, M.J., Shankar, R.D., Panahiazar, M., Willrett, D., Egyedi, A.L., Gevaert, O., Graybeal, J. and Musen, M.A. (2017). “Fast and Accurate Metadata Authoring Using Ontology-Based Recommendations.” *AMIA Annu Symp Proc*. 2017: 1272-1281. PMID:29854196. 29854196

23. Lou, Y., Tu, S.W., Nyulas, C., Tudorache, T., Chalmers, R.J.G. and Musen, M.A. (2017). “Use of ontology structure and Bayesian models to aid the crowdsourcing of ICD-11 sanctioning rules.” *J Biomed Inform*. 68: 20-34. PMID:28192233. doi:10.1016/j.jbi.2017.02.004.

24. Caufield, J.H., Zhou, Y., Garlid, A.O., Setty, S.P., Liem, D.A., Cao, Q., Lee, J.M., Murali, S., Spendlove, S., Wang, W., Zhang, L., Sun, Y., Bui, A., Hermjakob, H., Watson, K.E. and Ping, P. (2018). “A reference set of curated biomedical data and metadata from clinical case reports.” *Sci Data*. 5: 180258. PMID:30457569. doi:10.1038/sdata.2018.258.