Equipping Physiologists with an Informatics Tool Chest: Toward an Integerated 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 ($\Delta\psi$), 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