Contents lists available at ScienceDirect

Methods

journal homepage: www.elsevier.com/locate/ymeth

Identification of mitochondrial disease genes through integrative analysis of multiple datasets

Raeka S. Aiyar, Julien Gagneur, Lars M. Steinmetz *

European Molecular Biology Laboratory, Meyerhofstraße 1, 69117 Heidelberg, Germany

ARTICLE INFO

Article history:
Accepted 8 October 2008
Available online 16 October 2008

Keywords:
Functional genomics
Mitochondria
Genetic disorder
Genome-wide dataset
Discrimination analysis
Positional cloning
Functional gene network
Machine learning
Bayesian data integration
Support vector machine

ABSTRACT

Determining the genetic factors in a disease is crucial to elucidating its molecular basis. This task is challenging due to a lack of information on gene function. The integration of large-scale functional genomics data has proven to be an effective strategy to prioritize candidate disease genes. Mitochondrial disorders are a prevalent and heterogeneous class of diseases that are particularly amenable to this approach. Here we explain the application of integrative approaches to the identification of mitochondrial disease genes. We first examine various datasets that can be used to evaluate the involvement of each gene in mitochondrial function. The data integration methodology is then described, accompanied by examples of common implementations. Finally, we discuss how gene networks are constructed using integrative techniques and applied to candidate gene prioritization. Relevant public data resources are indicated. This report highlights the success and potential of data integration as well as its applicability to the search for mitochondrial disease genes.

1. Introduction

A central task in elucidating the molecular basis of a genetic disease is identification of the causative defect—that is, the gene or genes whose mutations result in the disease. This knowledge is also crucial for developing effective therapeutic strategies that target the key molecular players rather than simply alleviating the symptoms. For the majority of Mendelian or suspected Mendelian diseases, however, the genetic basis remains undetermined [1]; complex diseases driven by multiple genetic and environmental factors have proven even more challenging [2].

The most common approach used for disease gene identification is positional cloning to pinpoint the disease locus. In this approach, linkage analysis using polymorphic genetic markers isolates a region of the genome that segregates with the disease phenotype [3]. Candidate genes are then selected from this region and screened for mutations in a patient population. This approach has helped to identify the causative defect of approximately 2400 genetic diseases (Online Mendelian Inheritance in Man (OMIM) [1,4]). However, its success is limited by the genetic complexity of most diseases, which exacerbates issues such as inadequate sample sizes, a lack of informative meiotic crossover events, genetic heterogeneity, misdiagnosis, epistatis, and incomplete penetrance [2,5–7]. Consequently, positional cloning frequently fails to identify a disease locus. When a locus is identified, it often contains more candidate genes than one laboratory can feasibly screen [2]. Theoretically, these candidates can be filtered according to their biological annotation. In practice, however, annotation is usually insufficient, and therefore informed prioritization requires a more complete understanding of gene function.

High-throughput technologies generate extensive data on gene function: these technologies include genome sequencing, gene expression arrays, protein–protein interaction screens, RNA interference, mass spectrometry, and metabolite profiling. An advantage of these technologies is that all measurements in a dataset are made under uniform conditions, enabling quantitative comparison. Furthermore, data is generated on uncharacterized genes, providing indications of their function. Orthology allows this data to be transferred across species, and text mining consolidates information from decades of single-gene studies. Each genome-scale approach is biased towards different functional subsets of genes and prone to certain errors; consequently, overlap among different datatypes is often limited [8]. In order to effectively predict gene function, a combined analysis of these datasets is required to capitalize on their strengths and compensate for their limitations. Data integration techniques are efficient at extracting information from multiple datasets [9], and are thus an invaluable tool for candidate gene prioritization [8,10,11].

* Corresponding author. Fax: +49 6221 387 518.
E-mail address: larsms@embl.de (L.M. Steinmetz).

1046-2023/$ - see front matter © 2008 Elsevier Inc. All rights reserved.
doi:10.1016/j.ymeth.2008.10.002

Integrative approaches have been particularly useful in the functional characterization of mitochondria [11]. In addition to producing the majority of cellular ATP through respiration, this organelle plays a central role in metabolism, ion storage, oxidative stress management, signal transduction, antiviral response, and apoptosis [12,13]. Due to the diverse and fundamental nature of these processes, mitochondrial dysfunction can impair multiple organ systems [14]. Mitochondrial diseases primarily affect tissues with high energy requirements (e.g., central nervous system, muscle, liver) [13,15], and include myopathies, dystrophies, and neurodegenerative disorders. (Mitochondrial disease descriptions can be found on the United Mitochondrial Disease Foundation website www.umdf.org and in the OMIM database [4].) The frequency of mitochondrial diseases is significantly higher than expected based on the estimated number of protein components [15]. Mitochondrial diseases follow either maternal, Mendelian, or complex inheritance since some proteins (13 in human) are encoded by the mitochondrial genome, while the vast majority are nuclear-encoded [14,16]. For all of these reasons, it is believed that many diseases with an unknown molecular basis are mitochondrial [12,17].

The identification of mitochondrial disease genes is limited by the insufficient characterization of the mitochondrial proteome. To date, only half the 1500 proteins expected to localize within human mitochondria have been identified [16,18]. Since mitochondrial and cellular functions are tightly integrated, a full characterization of mitochondria requires complementing this set with extraorganellar proteins involved in, for example, mitochondrial transcriptional regulation, biogenesis, metabolic branches, and signalling. Systematic approaches have been directed at identifying these components (i.e., predicting "parts lists") [18–20], generating interaction networks [21], and developing mathematical models [22]. Several catalogs of mitochondrial genes have been predicted (Table 1): the most comprehensive of these is the MitoP2 database, generated using data integration techniques [16].

By predicting genes involved in mitochondrial function, data integration techniques have proven successful at prioritizing candidate mitochondrial disease genes. For example, a screen of multiple parameters in the *Saccharomyces cerevisiae* deletion collection identified 466 genes whose deletion impairs respiration. The high conservation of yeast and human mitochondria allowed these genes to be mapped to their human counterparts and prioritized as mitochondrial disease candidates [24]. Another study on Leigh syndrome (a cytochrome c oxidase deficiency) integrated RNA and protein expression data to select *LRPPRC* as the top candidate in the disease locus. Mutations discovered in this gene confirmed its causative role in the disease [27]. A later integration of eight datasets established that mutations in *MPV17* cause an infantile mitochondrial DNA depletion disorder, despite the gene's previous peroxisomal annotation [19,28]. Such success stories illustrate the power of integrative genomics.

In this report, we explain how data integration approaches are used to prioritize candidate genes according to mitochondrial function. We describe various datasets that can be used to determine mitochondrial function, compare data integration strategies, and present applications of computational network models.

2. Methods

A typical data integration procedure aimed at prioritizing candidate mitochondrial disease genes can be divided into three major steps (Fig. 1). The first step consists of collecting multiple datasets on mitochondrial function (previous efforts have used up to 25 [16,19,21]), including a reference set containing known mitochondrial genes. In the second step, a discrimination analysis method [9] is trained on the reference set to classify genes as mitochondrial or not; it is then used to optimally integrate the input datasets into a score reflecting the probability that each gene in the genome is mitochondrial. This score is used in the third step to prioritize candidate genes from a disease locus.

2.1. Datasets on mitochondrial function

Here, we highlight several types of data that can be used in an integrative approach to predict proteins physically residing in mitochondria and genes functionally related to the organelle. Both classes of genes must be considered for the study of mitochondrial disease, because of the interdependence of mitochondrial and cellular processes. Selecting complementary datatypes will maximize the information captured by the integration. While the accuracy of input datasets may vary, the discrimination analysis algorithm will compensate for these variations if high-quality positive and negative reference sets are supplied. Each dataset can be evaluated by estimating its sensitivity and specificity; this is usually done by selecting a threshold and calculating the sensitivity as the fraction of reference proteins captured, and the specificity as the fraction of negative reference proteins excluded. Ranges of sensitivity and specificity calculated in the construction of the human MitoP2 database are indicated for applicable datasets (otherwise, MitoP2-Yeast calculations are shown) [16].

2.1.1. Reference set

For the purpose of mitochondrial gene prediction, it is advisable to build the reference set from genes with definitive mitochondrial function based on single-gene studies. A good example is the manually-curated reference set of 870 mitochondria-localized human proteins used to construct the MitoP2 database (Table 1) [16]. In

Table 1
Mitochondrial parts-list databases

Database name	Organism	Contents	URL
MitoP2 [16]	Yeast, mouse, *A. thaliana*, *N. crassa*, human	Known and predicted proteins with mitochondrial localization and/or function; diseases	www.mitop.de
MitoCarta [20]	Human, mouse	Mitochondria-localized proteins, tissue-specific	www.broad.mit.edu/pubs/MitoCarta/
MitoProteome [23]	Human	Mitochondrial protein sequences from experimental and public databases	www.mitoproteome.org
HMPDb	Human	Proteins involved in mitochondrial biogenesis, function; diseases	bioinfo.nist.gov/hmpd
YMPD	Yeast	Proteins with mitochondrial localization and/or function	bmerc-www.bu.edu/projects/mito
YDPM [24]	Yeast	Mitochondria-specific yeast deletion collection phenotypes, proteomics and gene expression	deletion.stanford.edu/YDPM
MitoDrome [25]	Fruit fly	Nuclear-encoded mitochondrial proteins	www2.ba.itb.cnr.it/MitoDrome
AMPDB [26]	*Arabidopsis*	Predicted and verified mitochondria-localized proteins	www.plantenergy.uwa.edu.au/ampdb/