Published in final edited form as:

Methods. 2008 December ; 46(4): 248–255. doi:10.1016/j.ymeth.2008.10.002.

Identification of mitochondrial disease genes through integrative analysis of multiple datasets

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

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

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.

Keywords

functional genomics; mitochondria; genetic disorder; genome-wide dataset; discrimination analysis; positional cloning; functional gene network; machine learning; Bayesian data integration; support vector machine

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,

*To whom correspondence should be addressed. *Telephone:* +49 6221 387-389; *Fax:* +49 6221 387-518; *larsms@embl.de* (Lars M. Steinmetz).

misdiagnosis, epistasis, 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].

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 as well as 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.

Methods

A typical data integration procedure aimed at prioritizing candidate mitochondrial disease genes can be divided into three major steps (Figure 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.

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].

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 addition, discrimination analysis methods use a negative reference set, which in this case contains non-mitochondrial proteins. This set can be built from proteins that localize to other cellular compartments [19] or have annotation that does not include mitochondrial function. However, potential drawbacks of these options are dual localization and bias of the integration against well-annotated proteins. Another option is to use the implicit negative reference set—that is, proteins not in the positive reference set. This set is enriched for non-mitochondrial proteins (at least 90% expected) and is large enough to prevent bias. The drawback in this case is the underestimation of specificity.

Orthology—The human mitochondrion is a complex organelle with an estimated 1500 proteins that are highly conserved in sequence and localization. Due to experimental advantages, mitochondria of many model organisms are more thoroughly characterized [11]. Sequence similarity to known mitochondrial proteins in other organisms is therefore an indicator of mitochondrial function in human. Indeed, relatively high ranges of sensitivity (40–48%) and specificity (35–47%) were obtained for reciprocal best BLAST hits with yeast and *Neurospora* mitochondrial proteins in the construction of the human MitoP2 database [16]. This approach only detects human mitochondrial proteins conserved in other organisms; for example, the yeast mitochondrial proteome is about half the size of the human one [21]. In addition to inferring mitochondrial function through sequence similarity, orthology permits the transfer of genome-wide datasets relevant to mitochondrial function from model organisms to human.

Gene Expression—Transcription profiles in relevant conditions can indicate mitochondrial function. For example, yeast can grow by fermentation or by respiration, the latter requiring active mitochondria for ATP production. Genes are potentially mitochondrial if they show differential expression between the two conditions, in response to the diauxic shift from fermentation to respiration [29], or in response to overexpression of Hap4, a transcription factor of nuclear-encoded mitochondrial genes [30]. In mouse and human, genes upregulated in response to overexpression of *PGC-1 α* , one of several activators of mitochondrial proliferation [31], may play a role in mitochondrial biogenesis [19]. Mitochondrial genes can also be predicted based on coexpression (similarity of expression profiles to those of known mitochondrial genes) [8], which suggests coregulation. The specificity of transcription profiling is limited by the potential coregulation of mitochondrial and unrelated processes (15–43% in MitoP2-Yeast); its sensitivity is limited due to posttranscriptional and posttranslational regulation (19–43%). Gene expression profiles for conditions relevant to mitochondria can be accessed via the Gene Expression Omnibus (www.ncbi.nlm.nih.gov/geo), ArrayExpress (www.ebi.ac.uk/arrayexpress), and the Stanford Microarray Database (genome-www5.stanford.edu).

Subcellular Localization—Subcellular localization in yeast has been globally assessed using strain collections expressing GFP fusion proteins [32] and epitope tagged-proteins [33]. In mammalian cells, a new tagging technique has been developed whereby full-length GFP is synthesized only if the encoded protein is imported into mitochondria, eliminating the need to examine mitochondria themselves for fluorescence [34]. These approaches will only accurately classify proteins that physically reside in the organelle and whose localization is unaffected by the tag (specificity 26%, sensitivity 80%).

Knockout/Knockdown Phenotype—Mitochondrial function can be inferred by phenotypes in the absence of the gene. The yeast deletion collection has been used to screen for mitochondrial genes by identifying mutants with fitness defects in respiratory growth conditions [24,35]. Data generated with the yeast deletion collection is available from the Yeast Deletion Project and Proteomics of Mitochondria Database (www-deletion.stanford.edu/YDPM). In higher organisms, gene deletion is more difficult, and a common alternative is gene silencing using RNA interference (RNAi). Mitochondrial proteins have been identified by screening for knockdown phenotypes such as altered citrate synthase activity [36], mitochondrial morphology, oxygen consumption, or metabolism [37]. Genes lacking a knockout or knockdown phenotype because they are redundant or essential (the latter being the case for the mitochondrial outer membrane transporters, for example) will not be detected by these analyses (specificity 50%, sensitivity 37–45% in MitoP2-Yeast).

Sequence-Based Prediction—Mitochondrial proteins encoded by nuclear genes must be imported into the organelle. This process is believed to be triggered by N-terminal signal peptides that are recognized by the import machinery and cleaved once the protein reaches its destination. Several algorithms have thus been developed to predict signal peptides based on the patterns observed in the N-terminus of mitochondrial proteins [38]. They have achieved modest success, in part because the signature amphipathic helix is primarily a matrix-localization signal and targeting sequences for other mitochondrial compartments are not well-defined [39]. In addition, signal sequences do not follow straight-forward patterns and likely involve more of the protein than the N-terminus, especially in higher eukaryotes [40,41]. In some cases, proteins are localized by binding to another protein (a ‘hitchhiking’ mechanism) or based on signals in their mRNA sequences [39]. Some examples of these algorithms include TargetP [42] (www.cbs.dtu.dk/services/TargetP/), SubLoc [43] (www.bioinfo.tsinghua.edu.cn/SubLoc/), PSORT [44,45] (psort.nibb.ac.jp), Predotar [46] (urgi.versailles.inra.fr/predotar), and pTARGET [40] (<http://bioapps.rit.albany.edu/pTARGET/>) (specificity 5–15%, sensitivity 44–61%).

Mitochondrion-bound polysomes—Many mRNAs encoding mitochondrial proteins have been shown to associate with mitochondrion-bound polysomes in yeast and human [47]. Once transported by these polysomes, these transcripts are cotranslationally imported into mitochondria. By isolating mitopolysomes and hybridizing the associated sequences to a microarray, the identity and abundance of the genes involved has been investigated in yeast [48] (specificity 23%, sensitivity 13% in MitoP2-Yeast).

Proteomics—Protein identification by mass spectrometry is a direct technique to define the mitochondrial proteome. This is typically done by purifying mitochondria and identifying protein components using mass spectrometry [49–51]. Mass spectrometry-based methods have been used for comparative proteomics studies under several conditions in yeast [8] and in different mouse tissues [52]. These techniques, however, exhibit bias against proteins of low abundance (due to limitations of the dynamic range of detection) and membrane proteins [51,53]. Moreover, the mitochondrial purification method chosen can have a considerable effect on the proteins identified or result in contamination with cytoskeletal proteins [53] (specificity 37–83%, sensitivity 10–38%).

Protein-Protein Interactions—Mitochondrial function and localization is often inferred from interaction with known mitochondrial proteins. Evidence for protein-protein interactions can be derived from large-scale experiments, literature, functional associations, and orthology; recent reviews, however, indicate that much of this data may be spurious or biologically irrelevant [54–57]. Interaction data and resources are discussed further in the Functional Gene Networks section (specificity 26–62%, sensitivity 22–38%).

α -Proteobacterial Homology—Since the mitochondrion is derived from an ancient α -proteobacterium [12], several of its core components are homologous to α -proteobacterial proteins [21]. Sequence similarity to proteins in the α -proteobacterium *Rickettsia prowazekii*, the closest living relative of the mitochondrial ancestor, suggests mitochondrial function. Past studies have classified a protein as mitochondrial if it has a high-scoring reciprocal best BLAST hit with an *R. prowazekii* protein (specificity 14%, sensitivity 30%).

Text Mining—Scientific literature is rich in single-gene studies that offer evidence for mitochondrial function. The sheer volume of the literature makes manual extraction of this information unfeasible, necessitating the use of computational text mining algorithms. The majority of English sentences, however, cannot yet be systematically parsed and understood by these algorithms [58]. Nevertheless, text mining has been used to infer interactions [55,

56], functional annotations [59], and subcellular localization (SherLoc, www-bs.informatik.uni-tuebingen.de/Services/SherLoc/ [60]).

Methods for Data Integration

The goal of data integration here is to assess each gene's involvement in mitochondrial function. The discrimination analysis method considers each gene as a collection of its values in each dataset. The algorithm 'learns' how to use these values by fitting a mathematical model of the classification using reference set genes [9]. It then calculates a final score indicating how to classify each gene in the genome. To avoid overfitting—that is, to ensure the method will perform well not only on the reference set but on unseen data—one typically trains the method on a subset of the reference set (training set) and evaluates its performance on the remainder of the reference set (test set) [9]. Discrimination analysis algorithms vary by the type of model used and the criteria for assessing the quality of the generated model. We describe here three types of algorithms that are standard for data integration problems and have successfully been applied to the prediction of mitochondrial function.

Naïve Bayes—Naïve Bayes is one of the most straightforward discrimination analysis algorithms [9]. An odds ratio is calculated for each gene in each dataset as the enrichment for positive versus negative reference genes with the same values as that of the gene (quantitative data must be discretized); the product of a gene's odds ratios across datasets is its score, which is used to classify it as mitochondrial or not. This score equals the odds ratio of the gene being mitochondrial, assuming independence of the datasets given the gene classification. Although the independence assumption is not always fulfilled (for example, even among mitochondrial proteins, protein abundance is at least partially dependent on transcription levels), naïve Bayes can still yield good results in practice. For example, it was successfully employed to identify *MPV17* as a mitochondrial disease gene using eight datasets [19].

Linear predictors—Linear predictor methods, such as linear discriminant analysis or logistic regression [9] can be viewed as a generalization of naïve Bayes. These methods assign a weight to each dataset, and the score of a gene depends on the weighted sum of its dataset values. The advantage over naïve Bayes is that weights of linear predictors are fitted to optimal values using the training set. An advantage it shares with naïve Bayes is that the dataset weights indicate the relative contribution of the datasets to the prediction. A disadvantage is that both methods can suffer from overfitting on high-dimensional data compared to SVMs (discussed next). A recent implementation of linear prediction expanded the catalog of yeast mitochondrial proteins and their interactions, many of which were experimentally verified [21].

Support Vector Machines—A more recently developed discrimination analysis algorithm is the support vector machine (SVM) [61]. SVMs place each gene in multidimensional space according to its values in each dataset and compute the plane that maximally separates positive from negative genes. Thus when fitting the mathematical model in SVMs, the boundary datapoints (points that are closest to each other but belong to different groups) are most important, whereas the other methods consider all datapoints. Both from a theoretical point of view and in practice, this approach performs very well on high-dimensional datasets compared to traditional methods. SVMs are versatile in terms of choosing the mathematical functions used to generate the model [61]. However, SVM fitting requires some parameter tuning, which is a manual and time-consuming process. Moreover, the predictor is not as easily interpretable as naïve Bayes or linear predictors. The MitoP2 database (Table 1), which is currently the most comprehensive resource available for mitochondrial proteins, provides SVM-based prediction and has facilitated the identification of mitochondrial disease genes [16].

Applying results of data integration—Performance of the discrimination analysis method can be evaluated by selecting a threshold and estimating the sensitivity (the fraction of reference proteins scoring above the threshold) and specificity (the fraction of negative reference proteins scoring below the threshold) of the results. Figure 2 depicts the performance in terms of sensitivity and specificity (for all possible thresholds) of multiple yeast mitochondrial gene predictions, including a heuristic method (described in Prokisch *et al.* [8]), a linear predictor [21], and an SVM-based method [16], relative to 24 individual datasets. At a given sensitivity, a higher specificity is preferable and vice versa. Hence, methods with curves further toward the right and the top are the most effective. This illustrates that while output varies among the integration methods, all of them clearly outperform the individual datasets. Data integration approaches to identify mitochondrial parts-lists usually select a threshold score (based on sensitivity and specificity) above which genes are deemed mitochondrial. For the purpose of prioritizing candidate disease genes, however, genes can simply be ranked for mutation screening according to their score.

Functional Gene Networks

Genome-wide datasets have provided insight into the molecular networks that govern biological processes, such as protein-protein interaction, metabolic, and regulatory networks. Functional gene networks are abstractions of these networks, where gene nodes are connected if they share a functional relationship. They are used to infer gene function based on the guilt-by-association principle: if two genes are network neighbours, they presumably contribute to the same biological process [55,62–65]. By placing genes in a functional context, these networks offer an additional means of candidate disease gene prioritization.

Functional gene networks are constructed by integrating a collection of datasets on gene-gene interactions using techniques described in the previous section. Evidence for interactions between genes can be derived from many sources: (a) their protein products physically interact [57,66]; (b) they exhibit genetic interactions indicated by, for example, a synthetic lethal phenotype [67]; (c) their loss-of-function phenotypes are similar; or (d) they share another functional relationship [55,64]. Many of the examples in the Datasets section are therefore also applicable to network construction. Several public network databases are detailed in Table 2. One example designed for disease gene identification is the tool Prioritizer, based on a human functional gene network generated by Bayesian integration [68].

Applications of Functional Gene Networks to Disease Gene Identification

Candidate disease genes can be prioritized according to the network proximity of disease genes associated with similar symptoms (Figure 3A). When grouping genes into modules according to their function, it has been observed that mutations in genes of a particular module often result in the same phenotype [63, 73]. Genes sharing functional modules with the approximately 150 mitochondrial disease genes identified to date [16] are therefore valid candidates for diseases with mitochondrial symptoms. Even in a yeast mitochondrial interaction network, orthologs of related disease genes were often found in the same functional module [21]. In addition, a recent analysis of human mitochondrial metabolism has identified several ‘co-sets’ consisting of groups of genes whose mutations (in particular, SNPs) lead to similar metabolic consequences [73].

When studying multigenic diseases, networks can help to prioritize groups of candidate genes. In the event that several disease loci result from linkage analysis, the most probable combinations of genes from the loci can be predicted based on their relative positions in the network (Figure 3B). This is based on the observation that genes causing the same disease are separated by a significantly shorter-than-average network distance [68], which is the foundation of Prioritizer (Table 2). Therefore, if one gene has already been implicated in a

multigenic disorder, additional candidates can be selected based on their network proximity to that gene. Without the advantages afforded by networks, testing all the combinations that result even from a few disease loci would be unfeasible.

Discussion

Data integration techniques enable informed prioritization of candidate disease genes. Their power lies in combining the strengths of various high-throughput datasets and compensating for their individual limitations. Functional gene networks extend this potential by consolidating gene-gene relationships in the context of biological processes. Candidate disease genes can thus be prioritized by network proximity to predicted or known disease genes associated with the same or clinically related diseases.

As existing technologies for both high-throughput and more focused studies progress, and new ones emerge, the effectiveness of data integration methods and network models will only improve. For example, differential proteomics generates thousands of precise measurements of protein levels and thus complements transcription as a measure of gene activity and regulation under multiple conditions; tools for analyzing this data are rapidly becoming more accurate [51,74]. The quantification of small-molecule metabolites (also known as metabolite profiling) is contributing to the construction of network models and a comprehensive description of mitochondrial [75] and cellular physiology [76–78]. Additional mitochondria-specific phenotyping assays, including quantitative morphology [79] and membrane potential measurements [80], enable more refined characterization of mitochondrial function. In lab-on-a-chip technologies, multiple types of measurements can be generated by a single experiment in cell-like environments, enabled by micro- and nanofluidics [81].

By clarifying the dynamic function of genes, proteins, and metabolites, the application of more quantitative technologies is instrumental in refining models of biological networks. Quantitative models can already predict the behaviour of mitochondrial metabolism under various conditions [22]. Such models allow simulation of the effects of perturbations (e.g., inactivating a gene) on biological systems; thus, analyzing experimental data from a diseased system and inferring the causative defect will eventually be possible *in silico*. For example, a model of mitochondrial metabolism [22] has been used to identify the deficient respiratory chain complex in Leigh syndrome fibroblasts via metabolite profiling [82]. Another study has pinpointed subsets of gene networks showing variation in gene expression associated with quantitative trait loci (QTLs) for obesity; as a result, three novel obesity genes have been identified [83].

In addition to identifying disease genes, network models can be used to study disease mechanisms by elucidating molecular roles in maintaining health. In particular, alterations caused by disease in portions of the underlying networks can be investigated. Therapeutic strategies can then be targeted at causative defects, their most consequential downstream effects, or affected subnetworks [83,84]. As models are developed to encapsulate healthy and diseased states, they must be extended to capture the effects of genetic variation, including the thousands of polymorphisms yielding small effects or contributing to individual disease risk [85]. High-throughput genotyping and genome-wide association studies are making progress towards understanding these relationships [7], and as these advances are mirrored on the clinical side with more comprehensive diagnostic tools, the gap between genotype and phenotype will be reduced [17]. Eventually, personalized therapeutic strategies will be developed that cater to individual genetic makeup, environment, diet, and lifestyle [86]. It is certain that these developments will not be realized through any one technology, and thus data integration makes them attainable by drawing on all the strengths molecular biology has to offer.

Acknowledgments

The authors would like to acknowledge Amoolya Singh, Michael DiBernardo, Christine Panagiotidis, Stéphanie Blandin, Eugenio Mancera, Fabiana Perocchi, Himanshu Sinha, and Sebastian Kühner for their help in preparing this manuscript.

References

1. Online Mendelian Inheritance in Man, OMIM (TM). McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information. Bethesda, MD: National Library of Medicine; [October 2, 2008]. World Wide Web URL www.ncbi.nlm.nih.gov/omim
2. Botstein D, Risch N. *Nat Genet* 2003;33:228–237. [PubMed: 12610532]
3. Lander ES, Botstein D. *Genetics* 1989;121:185–199. [PubMed: 2563713]
4. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. *Nucleic Acids Res* 2005;33(Database Issue)
5. Dudbridge F, Gusnanto A, Koeleman BP. *Hum Genomics* 2006;2:310–317. [PubMed: 16595075]
6. Hoh J, Ott J. *Nat Rev Genet* 2003;4:701–709. [PubMed: 12951571]
7. Kruglyak L. *Nature Reviews Genetics*. 2008
8. Prokisch H, Scharfe C, Camp DG, Xiao W, David L, Andreoli C, Monroe ME, Moore RJ, Gritsenko MA, Kozany C, Hixson KK, Mottaz HM, Zischka H, Ueffing M, Herman ZS, Davis RW, Meitinger T, Oefner PJ, Smith RD, Steinmetz LM. *PLoS Biol* 2004;2
9. Hastie, T.; Tibshirani, R.; Friedman, JH. *The Elements of Statistical Learning*. Springer; 2001.
10. Giallourakis C, Henson C, Reich M, Xie X, Mootha VK. *Annu Rev Genomics Hum Genet* 2005;6:381–406. [PubMed: 16124867]
11. Perocchi F, Mancera E, Steinmetz LM. *Mol Biosyst* 2008;4:18–29. [PubMed: 18075670]
12. Wallace DC. *Annu Rev Genet* 2005;39:359–407. [PubMed: 16285865]
13. Chan DC. *Cell* 2006;125:1241–1252. [PubMed: 16814712]
14. Wallace DC. *Science* 1999;283:1482–1488. [PubMed: 10066162]
15. Schaefer AM, Taylor RW, Turnbull DM, Chinnery PF. *Biochimica et biophysica acta* 2004;1659:115–120. [PubMed: 15576042]
16. Prokisch H, Andreoli C, Ahting U, Heiss K, Ruepp A, Scharfe C, Meitinger T. *Nucleic Acids Res* 2006;34
17. Haas RH, Parikh S, Falk MJ, Saneto RP, Wolf NI, Darin N, Cohen BH. *Pediatrics* 2007;120:1326–1333. [PubMed: 18055683]
18. Westermann B, Neupert W. *Nat Biotech* 2003 Mar;21(3):239–240.
19. Calvo S, Jain M, Xie X, Sheth SA, Chang B, Goldberger OA, Spinazzola A, Zeviani M, Carr SA, Mootha VK. *Nature Genetics* 2006;38:576–582. [PubMed: 16582907]
20. Pagliarini DJ, Calvo SE, Chang B, Sheth SA, Vafai SB, Ong S-E, Walford GA, Sugiana C, Boneh A, Chen WK, Hill DE, Vidal M, Evans JG, Thorburn DR, Carr SA, Mootha VK. *Cell* 2008;134:112–113. [PubMed: 18614015]
21. Perocchi F, Jensen LJ, Gagneur J, Ahting U, von Mering C, Bork P, Prokisch H, Steinmetz LM. *PLoS Genetics* 2006;2:e170+. [PubMed: 17054397]
22. Vo TD, Greenberg HJ, Palsson BO. *J Biol Chem* 2004;279:39532–39540. [PubMed: 15205464]
23. Cotter D, Guda P, Fahy E, Subramaniam S. *Nucleic Acids Res* 2004;32
24. Steinmetz LM, Scharfe C, Deutschbauer AM, Mokranjac D, Herman ZS, Jones T, Chu AM, Giaever G, Prokisch H, Oefner PJ, Davis RW. *Nat Genet* 2002;31:400–404. [PubMed: 12134146]
25. D’Elia D, Catalano D, Licciulli F, Turi A, Tripoli G, Porcelli D, Saccone C, Caggese C. *Mitochondrion* 2006;6:252–257. [PubMed: 16982216]
26. Heazlewood JL, Millar AH. *Nucleic Acids Res* 2005;33
27. Mootha VK, Lepage P, Miller K, Bunkenborg J, Reich M, Hjerrild M, Delmonte T, Villeneuve A, Sladek R, Xu F, Mitchell GA, Morin C, Mann M, Hudson TJ, Robinson B, Rioux JD, Lander ES. *Proc Natl Acad Sci U S A* 2003;100:605–610. [PubMed: 12529507]

28. Spinazzola A, Viscomi C, Fernandez-Vizarra E, Carrara F, D'Adamo P, Calvo S, Marsano RM, Donnini C, Weiher H, Strisciuglio P, Parini R, Sarzi E, Chan A, Dimauro S, Rötig A, Gasparini P, Ferrero I, Mootha VK, Tiranti V, Zeviani M. *Nature Genetics* 2006;38:570–575. [PubMed: 16582910]
29. DeRisi JL, Iyer VR, Brown PO. *Science* 1997;278:680–686. [PubMed: 9381177]
30. Lascaris R, Bussemaker HJ, Boorsma A, Piper M, van der Spek H, Grivell L, Blom J. *Genome Biol* 2003;4
31. D'Souza AD, Parikh N, Kaech SM, Shadel GS. *Mitochondrion* 2007;7:374–385. [PubMed: 17890163]
32. Huh WK, Falvo JV, Gerke LC, Carroll AS, Howson RW, Weissman JS, O'Shea EK. *Nature* 2003;425:686–691. [PubMed: 14562095]
33. Kumar A, Agarwal S, Heyman JA, Matson S, Heidtman M, Piccirillo S, Umansky L, Drawid A, Jansen R, Liu Y, Cheung K-H, Miller P, Gerstein M, Roeder SG, Snyder M. *Genes Dev* 2002;16:707–719. [PubMed: 11914276]
34. Ozawa T, Umezawa Y. *Methods Mol Biol* 2007;390:119–130. [PubMed: 17951684]
35. Dimmer KS, Fritz S, Fuchs F, Messerschmitt M, Weinbach N, Neupert W, Westermann B. *Mol Biol Cell* 2002;13:847–853. [PubMed: 11907266]
36. Chen J, Shi X, Padmanabhan R, Wang Q, Wu Z, Stevenson SC, Hild M, Garza D, Li H. *Genome research* 2008;18:123–136. [PubMed: 18042644]
37. Ichishita R, Tanaka K, Sugiura Y, Sayano T, Mihara K, Oka T. *J Biochem.* 2008
38. Emanuelsson O, Brunak S, von Heijne G, Nielsen H. *Nat Protoc* 2007;2:953–971. [PubMed: 17446895]
39. Nakai K, Horton P. *Methods Mol Biol* 2007;390:429–466. [PubMed: 17951705]
40. Guda C, Subramaniam S. *Bioinformatics* 2005;21:3963–3969. [PubMed: 16144808]
41. Andrade MA, O'Donoghue SI, Rost B. *Journal of molecular biology* 1998;276:517–525. [PubMed: 9512720]
42. Emanuelsson O, Nielsen H, Brunak S, von Heijne G. *J Mol Biol* 2000;300:1005–1016. [PubMed: 10891285]
43. Hua S, Sun Z. *Bioinformatics* 2001;17:721–728. [PubMed: 11524373]
44. Nakai K, Horton P. *Trends Biochem Sci* 1999;24:34–36. [PubMed: 10087920]
45. Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K. *Nucleic Acids Res* 2007;35
46. Small I, Peeters N, Legeai F, Lurin C. *PROTEOMICS* 2004;4:1581–1590. [PubMed: 15174128]
47. Sylvestre J, Margeot A, Jacq C, Dujardin G, Corral-Debrinski M. *Molecular Biology of the Cell* 2003;14:3848–3856. [PubMed: 12972568]
48. Marc P, Margeot A, Devaux F, Blugeon C, Corral-Debrinski M, Jacq C. *EMBO Rep* 2002;3:159–164. [PubMed: 11818335]
49. Delahunty C, Yates. *Methods* 2005;35:248–255. [PubMed: 15722221]
50. Sickmann A, Reinders J, Wagner Y, Joppich C, Zahedi R, Meyer HE, Schönfisch B, Perschil I, Chacinska A, Guiard B, Rehling P, Pfanner N, Meisinger C. *Proc Natl Acad Sci U S A* 2003;100:13207–13212. [PubMed: 14576278]
51. Domon B, Aebersold R. *Science* 2006;312:212–217. [PubMed: 16614208]
52. Mootha VK, Bunkenborg J, Olsen JV, Hjerrild M, Wisniewski JR, Stahl E, Bolouri MS, Ray HN, Sihag S, Kamal M, Patterson N, Lander ES, Mann M. *Cell* 2003;115:629–640. [PubMed: 14651853]
53. Da Cruz S, Parone PA, Martinou JC. *Expert Rev Proteomics* 2005;2:541–551. [PubMed: 16097887]
54. von Mering C, Krause R, Snel B, Cornell M, Oliver SG, Fields S, Bork P. *Nature* 2002;417:399–403. [PubMed: 12000970]
55. von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, Krüger B, Snel B, Bork P. *Nucleic Acids Res* 2007;35
56. Ramani AK, Bunesco RC, Mooney RJ, Marcotte EM. *Genome Biol* 2005;6
57. Bader JS, Chaudhuri A, Rothberg JM, Chant J. *Nat Biotechnol* 2004;22:78–85. [PubMed: 14704708]
58. Rebholz-Schuhmann D, Kirsch H, Couto F. *PLoS Biol* 2005;3

59. Jaeger S, Gaudan S, Leser U, Rebholz-Schuhmann D. BMC Bioinformatics 2008;9
60. Shatkay H, Höglund A, Brady S, Blum T, Dönnies P, Kohlbacher O. Bioinformatics 2007;23:1410–1417. [PubMed: 17392328]
61. Schölkopf, B.; Smola, AJ. Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning). The MIT Press; 2001.
62. Aravind L. Genome Res 2000;10:1074–1077. [PubMed: 10958625]
63. Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabasi A-L. PNAS 2007;104:8685–8690. [PubMed: 17502601]
64. McGary K, Lee I, Marcotte E. Genome Biology 2007;8
65. Tian W, Zhang LV, Taşan M, Gibbons FD, King OD, Park J, Wunderlich Z, Cherry JM, Roth FP. Genome biology 2008;9
66. Krogan NJ, Cagney G, Yu H, Zhong G, Guo X, Ignatchenko A, Li J, Pu S, Datta N, Tikuisis AP, Punna T, Peregrín-Alvarez JM, Shales M, Zhang X, Davey M, Robinson MD, Paccanaro A, Bray JE, Sheung A, Beattie B, Richards DP, Canadian V, Lalev A, Mena F, Wong P, Starostine A, Canete MM, Vlasblom J, Wu S, Orsi C, Collins SR, Chandran S, Haw R, Rilstone JJ, Gandi K, Thompson NJ, Musso G, St. Onge RP, Ghanny S, Lam MHY, Butland G, Altaf-Ul AM, Kanaya S, Shilatifard A, O'Shea E, Weissman JS, Ingles JC, Hughes TR, Parkinson J, Gerstein M, Wodak SJ, Emili A, Greenblatt JF. Nature. 2006
67. Tong AHY, Lesage G, Bader GD, Ding H, Xu H, Xin X, Young J, Berriz GF, Brost RL, Chang M, Chen Y, Cheng X, Chua G, Friesen H, Goldberg DS, Haynes J, Humphries C, He G, Hussein S, Ke L, Krogan N, Li Z, Levinson JN, Lu H, Menard P, Munyana C, Parsons AB, Ryan O, Tonikian R, Roberts T, Sdicu A-M, Shapiro J, Sheikh B, Suter B, Wong SL, Zhang LV, Zhu H, Burd CG, Munro S, Sander C, Rine J, Greenblatt J, Peter M, Bretscher A, Bell G, Roth FP, Brown GW, Andrews B, Bussey H, Boone C. Science 2004;303:808–813. [PubMed: 14764870]
68. Franke L, Bakel H, Fokkens L, de Jong ED, Egmont-Petersen M, Wijmenga C. Am J Hum Genet 2006;78:1011–1025. [PubMed: 16685651]
69. Breitkreutz BJ, Stark C, Regulj T, Boucher L, Breitkreutz A, Livstone M, Oughtred R, Lackner DH, Bähler J, Wood V, Dolinski K, Tyers M. Nucleic Acids Res 2008;36
70. Güldener U, Münsterkötter M, Oesterheld M, Pagel P, Ruepp A, Mewes HW, Stümpflen V. Nucleic Acids Res 2006;34
71. Vastrik I, D'Eustachio P, Schmidt E, Joshi-Tope G, Gopinath G, Croft D, de Bono B, Gillespie M, Jassal B, Lewis S, Matthews L, Wu G, Birney E, Stein L. Genome Biology 2007;8:R39+. [PubMed: 17367534]
72. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y. Nucleic Acids Res 2008;36:D480–D484. [PubMed: 18077471]
73. Jamshidi N, Palsson BØ. Mol Syst Biol 2006;2:38. [PubMed: 16820779]Epub 2006 Jul 4. Review.
74. Mueller LN, Brusniak MY, Mani DR, Aebersold R. J Proteome Res 2008;7:51–61. [PubMed: 18173218]
75. Amara C, Marcinek D, Shankland E, Schenkman K, Arakaki L, Conley K. Methods 2008;46:312–318. [PubMed: 18930151]
76. Kell DB. Drug Discovery Today 2006;11:1085–1092. [PubMed: 17129827]
77. Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BØ. Proc Natl Acad Sci U S A 2007 Feb 6;104(6):1777–1782. [PubMed: 17267599]Epub 2007 Jan 31.
78. Bundy JG, Papp B, Harmston R, Browne RA, Clayson EM, Burton N, Reece RJ, Oliver SG, Brindle KM. Genome Res 2007;17:510–519. [PubMed: 17339370]
79. Song W, Bossy B, Martin OJ, Hicks A, Lubitz S, Knott AB, Bossy-Wetzel E. Methods 2008;46:295–303. [PubMed: 18952177]
80. Koopman W, Distelmaier F, Esseling J, Smeitink J, Willems P. Methods 2008;46:304–311. [PubMed: 18929665]
81. Teh SY, Lin R, Hung LH, Lee AP. Lab on a chip 2008;8:198–220. [PubMed: 18231657]
82. Vo TD, Lee Paul WN, Palsson BO. Mol Genet Metab 2007 May;91(1):15–22. [PubMed: 17336115] Epub 2007 Mar 1.

83. Chen Y, Zhu J, Lum PY, Yang X, Pinto S, Macneil DJ, Zhang C, Lamb J, Edwards S, Sieberts SK, Leonardson A, Castellini LW, Wang S, Champy M-F, Zhang B, Emilsson V, Doss S, Ghazalpour A, Horvath S, Drake TA, Lusis AJ, Schadt EE. *Nature*. 2008
84. Hood L, Heath JR, Phelps ME, Lin B. *Science* 2004;306:640–643. [PubMed: 15499008]
85. The Wellcome Trust Case Control Consortium. *Nature* 2007;447:661–678. [PubMed: 17554300]
86. Nicholson JK. *Mol Syst Biol* 2006;2
87. Ohlmeier S, Kastaniotis AJ, Hiltunen JK, Bergmann U. *J Biol Chem* 2004;279:3956–3979. [PubMed: 14597615]
88. Drawid A, Gerstein M. *Journal of Molecular Biology* 2000;301:1059–1075. [PubMed: 10966805]
89. Pflieger D, Le Caer JP, Lemaire C, Bernard BA, Dujardin G, Rossier J. *Anal Chem* 2002;74:2400–2406. [PubMed: 12038767]
90. Claros MG. *Comput Appl Biosci* 1995;11:441–447. [PubMed: 8521054]

Figure 1.

Data integration procedure for prioritization of mitochondrial disease candidate genes. Input datasets informative about mitochondrial function (left) and reference sets (center, below) of known mitochondrial (green spots) and non-mitochondrial genes (black spots) are collected. The reference sets are used to train the data integration method to combine the input datasets to calculate a score for each gene reflecting the probability that it is involved in mitochondrial function. The ranked genes are then cross-referenced with positional candidates in a disease locus (right, highlighted in yellow), providing a basis for prioritization.

Figure 2.

Comparing integration methods by sensitivity and specificity. Performance of three integration methods in predicting yeast mitochondrial proteins are shown by continuous curves obtained by assigning a threshold to the score ranging from its most stringent value (bottom-right corner) to its least stringent value (top-left corner). At a given threshold, sensitivity is calculated as the fraction of the reference set covered by the predicted set, and specificity is calculated as the fraction of the predicted set confirmed by the reference set. The two machine-learning based methods (linear predictor [21] and MitoP2 SVM [16]) outperform the original heuristic method (MitoP2 2004 [8]). The SVM outperforms the linear predictor only in the range of 45–65% specificity. The 24 input datasets (black dots) are A: *Neurospora* ortholog with mitochondrial localization, B: Huh *et al.*, 2003 [32] (mitochondrial localization), C: Kumar *et al.*, 2002 [33] (mitochondrial localization), D: Sickmann *et al.*, 2003 [50] (proteomics), E: Steinmetz *et al.*,

2002 [24] (deletion phenotype), F: Prokisch et al, 2004 [8] (proteomics), G: Lascaris *et al.*, 2003 [30] (Hap4-induced genes), H: von Mering *et al.*, 2002 [54] (medium confidence interaction with mitochondrial protein), I: Dimmer *et al.*, 2002 [35] (petite phenotype), J: human ortholog with mitochondrial localization, K: *R. prowazekii* ortholog, L: Ohlmeier *et al.*, 2004 [87] (proteomics), M: Bayesian prediction [88], N: Predotar [46] (signal peptide, score >50), O: Pflieger *et al.*, 2002 [89] (proteomics), P: von Mering *et al.*, 2002 [54] (high confidence interaction with mitochondrial protein), Q: MitoProt [90] (import prediction, score >0.80), R: Prokisch *et al.* [16] (*Neurospora* proteomics), S: PSORT [44] (signal peptide), T: Prokisch *et al.*, 2004 [8] (>1.2-fold differential expression, glucose versus lactate), U: Marc *et al.*, 2002 [48] (mitochondrion-bound polysomes, MLR>80), V: von Mering *et al.*, 2002 [54] (low confidence interaction with mitochondrial protein), W: deRisi *et al.*, 1997 [29] (>2-fold increase in diauxic shift when OD600 = 7.3), X: *E. cuniculi* ortholog (negative predictor).

Figure 3a Symptom matching in gene networks to identify disease candidates. Modules containing genes in a hypothetical disease locus for a neuromuscular dystrophy with ataxia are shown: green nodes represent genes implicated in diseases causing ataxia, blue represents genes implicated in diseases with different symptoms, and gray represents genes not associated with disease. One of the positional candidate genes (outlined in red) shares a module with genes implicated in diseases causing ataxia. This candidate would therefore be prioritized relative to the others in the locus.

Figure 3b Predicting candidate gene combinations for multigenic diseases using networks. Three genes, one from each hypothetical disease locus (highlighted in yellow), are network neighbours and therefore functionally related: these comprise the combination most likely to be responsible for the disease. This approach reduces the number of gene combinations that must be screened for mutations, given the size of typical linkage intervals.

Table 1**Mitochondrial Parts-List Databases**

Database Name	Organism	Contents	URL
MitoP2 [16]	Yeast, Mouse, <i>A. thaliana</i> , <i>N. crassa</i> , 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	bmrc-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]	<i>Arabidopsis</i>	Predicted and verified mitochondria-localized proteins	www.plantenergy.uwa.edu.au/ampdb/

Table 2

Interaction Network Databases

Database Name	Organism	Contents	URL
BioGRID [69]	Various	Protein and genetic interactions	www.thebiogrid.org/
MPact (MIPS) [70]	Yeast	Protein-protein interactions	mips.gsf.de/genre/proj/impact
Reactome [71]	Various	Biological pathways and reactions	www.reactome.org
KEGG [72]	Various	Set of biological system databases, including pathways, interactions	www.genome.jp/kegg
STRING [55]	Various	Known and predicted protein-protein interactions (physical and functional)	string.embl.de
Funckenstein [65]	Yeast	Functional gene network (guilt-by-association and guilt-by-profiling)	func.med.harvard.edu/yeast/
Prioritizer [68]	Human	Disease gene prediction based on functionalprioritizer.nl gene network	functionalprioritizer.nl