nature genetics

Systematic identification of human mitochondrial disease genes through integrative genomics

Sarah Calvo[1–3], Mohit Jain[1–3], Xiaohui Xie[1], Sunil A Sheth[1–3], Betty Chang[1], Olga A Goldberger[1–3], Antonella Spinazzola[4], Massimo Zeviani[4], Steven A Carr[1] & Vamsi K Mootha[1–3]

The majority of inherited mitochondrial disorders are due to mutations not in the mitochondrial genome (mtDNA) but rather in the nuclear genes encoding proteins targeted to this organelle. Elucidation of the molecular basis for these disorders is limited because only half[1,2] of the estimated 1,500 mitochondrial proteins[3] have been identified. To systematically expand this catalog, we experimentally and computationally generated eight genome-scale data sets, each designed to provide clues as to mitochondrial localization: targeting sequence prediction, protein domain enrichment, presence of *cis*-regulatory motifs, yeast homology, ancestry, tandem-mass spectrometry, coexpression and transcriptional induction during mitochondrial biogenesis. Through an integrated analysis we expand the collection to 1,080 genes, which includes 368 novel predictions with a 10% estimated false prediction rate. By combining this expanded inventory with genetic intervals linked to disease, we have identified candidate genes for eight mitochondrial disorders, leading to the discovery of mutations in *MPV17* that result in hepatic mtDNA depletion syndrome[4]. The integrative approach promises to better define the role of mitochondria in both rare and common human diseases.

A comprehensive catalog of mitochondrial proteins is essential for a systematic approach to discovering related disease genes. However, the best experimental and computational techniques fall far short of accurately identifying the estimated 1,500 human genes encoding mitochondrial proteins, of which only 13 are within the mtDNA. Computational tools have long been available for detecting N-terminal signal sequences that direct proteins to this organelle[5]. However, not all mitochondrial proteins are imported by such mechanisms, and moreover, computational detection of these signals is imprecise. As a consequence, methods such as TargetP[5] achieve only 91% specificity and 60% sensitivity, which gives rise to a 69% false positive prediction rate when the method is applied genome-wide, because the prior probability of a protein localizing to the mitochondrion is only 7% (see Methods). More recently, experimental approaches using tandem mass spectrometry (MS/MS) have added to the current inventory of known mitochondrial proteins, but owing to the bias toward abundant proteins, these methods have identified only an additional ~150 mitochondrial proteins[6,7]. Hence, when used alone, existing approaches have limited sensitivity and specificity. Recent studies have illustrated how these limitations can be overcome by combining different genomic approaches, but because such methods require high-quality

Table 1 Eight genome-scale data sets used to predict mitochondrial localization

Method	Genome-scale data set	Proteins predicted	False discovery rate (%)
Targeting signal	TargetP on human/mouse orthologs	4,532	69
Protein domain	Pfam domain found only in eukaryotic mitochondrial proteins (SwissProt)	1,097	12
Cis motif	Errα motif in human/mouse promoters	597	78
Yeast homology	*S. cerevisiae* mitochondrial ortholog	763	34
Ancestry	*R. prowazekii* ortholog	2,075	66
Coexpression	Coexpression with known mitochondrial genes in human/mouse tissue atlases	867	40
MS/MS	Mouse mitochondria (brain, heart, liver, kidney)	697	38
Induction	Difference in gene expression during mitochondrial biogenesis induced by PGC-1α	2,361	68
Maestro		**1,451**	**10**

Eight individual methods and an integrated approach (named Maestro) were used to predict mitochondrial localization of all 33,860 Ensembl human proteins. The genome-wide false discovery rate was estimated from large gold standard training data.

[1]Broad Institute of MIT and Harvard, Cambridge, Massachusetts 02142, USA. [2]Center for Human Genetic Research, Massachusetts General Hospital, Boston, Massachusetts 02114, USA. [3]Department of Systems Biology, Harvard Medical School, Boston, Massachusetts 02446, USA. [4]Unit of Molecular Neurogenetics, National Neurological Institute 'C. Besta', 20126 Milan, Italy. Correspondence should be addressed to V.K.M. (vamsi@hms.harvard.edu).

Received 10 November 2005; accepted 9 March 2006; published online 2 April 2006; doi:10.1038/ng1776

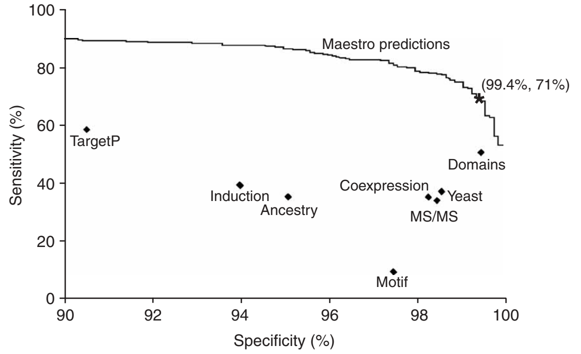

Figure 1 Sensitivity and specificity of mitochondrial prediction methods. Using training data of 654 known mitochondrial proteins (T_{mito}) and 2,847 nonmitochondrial proteins ($T_{\sim\mathrm{mito}}$), we estimate the sensitivity (percentage of T_{mito} correctly predicted) and specificity (percentage of $T_{\sim\mathrm{mito}}$ correctly predicted) of each prediction method. The accuracies of the eight individual data sets are shown at specific thresholds (see Methods), whereas the accuracy of Maestro is shown at a range of thresholds (black curve), with the chosen threshold marked by an asterisk.

genome-scale data sets and training data, they have been limited so far to studies in model organisms[8,9].

We sought to construct high-quality predictions of human proteins localized to the mitochondrion by generating and integrating data sets that provide complementary clues about mitochondrial localization. Unlike existing computational methods that rely purely on sequence features within the protein, we also take advantage of recent insights into the ancestry and transcriptional regulation of the organelle. Specifically, for each human gene product *p*, we assign a score $s_i(p)$, using each of the following eight genome-scale data sets (**Table 1** and Methods):

The targeting signal score (s_1) indicates the presence or absence of an N-terminal mitochondrial targeting sequence that directs protein import into the mitochondrion, identified by a computational tool called TargetP[5].

The protein domain score (s_2) records the presence of protein domains found to be exclusively mitochondrial, exclusively nonmitochondrial or shared, based on the SwissProt annotation of all eukaryotic sequences.

The *cis*-motif score (s_3) indicates the presence or absence of evolutionarily conserved transcriptional regulatory elements that we previously discovered to be enriched upstream of mitochondrial genes[10].

The yeast homology score (s_4) indicates the presence or absence of an *S. cerevisiae* ortholog with experimental evidence of mitochondrial localization (Saccharomyces Genome Database annotation).

The ancestry score (s_5) measures the sequence similarity to proteins from *Rickettsia prowazekii*, the closest living bacterial relative of human mitochondria[11].

The coexpression score (s_6) measures transcriptional coexpression with known mitochondrial genes, using genome-scale atlases of RNA expression across diverse tissues[12]. We use a neighborhood metric[6] to score each gene's coexpression with known mitochondrial genes.

The MS/MS score (s_7) indicates the number of tissues in which the protein was detected in a previous proteomic survey of mitochondria isolated from four mouse tissues[6].

The induction score (s_8) measures the upregulation of mRNA transcripts in a cellular model of mitochondrial biogenesis. We induced mitochondrial proliferation in a muscle cell line by overexpressing the transcriptional coactivator PGC-1α[13] and assayed genome-wide RNA abundance with microarray profiling (see Methods).

Each of the above scores (s_1–s_8) can be used individually as a weak genome-wide predictor of mitochondrial localization. We assessed each method's performance using large 'gold standard' curated training sets: 654 mitochondrial proteins (T_{mito}) curated by the MitoP2 database[1] and 2,847 nonmitochondrial proteins ($T_{\sim\mathrm{mito}}$) annotated to localize to other cellular compartments (see Methods and **Supplementary Table 1** online). As can be seen in **Figure 1**, the limited sensitivity and the relatively low specificity of each individual approach can generate a large proportion of false positives when applied genome-wide (**Table 1**).

To improve prediction accuracy, we integrated the eight approaches using a naive Bayes classifier[8] that we implemented with a computer program called Maestro (see Methods). We trained Maestro on the gold standard positive and negative data sets and applied it to the Ensembl set of 33,860 human proteins. For each of the eight features, we calculated a likelihood of mitochondrial localization by comparing performance on T_{mito} to performance on $T_{\sim\mathrm{mito}}$ at a range of scores (**Fig. 2a**). We computed a composite Maestro score by summing the log-likelihoods of eight individual features (**Fig. 2b**) in a naive Bayesian integration (see Methods). We selected a score threshold, dependent on the application, and classified as mitochondrial all proteins scoring above the threshold. Using a conservative threshold of 5.65, corresponding to a false discovery rate

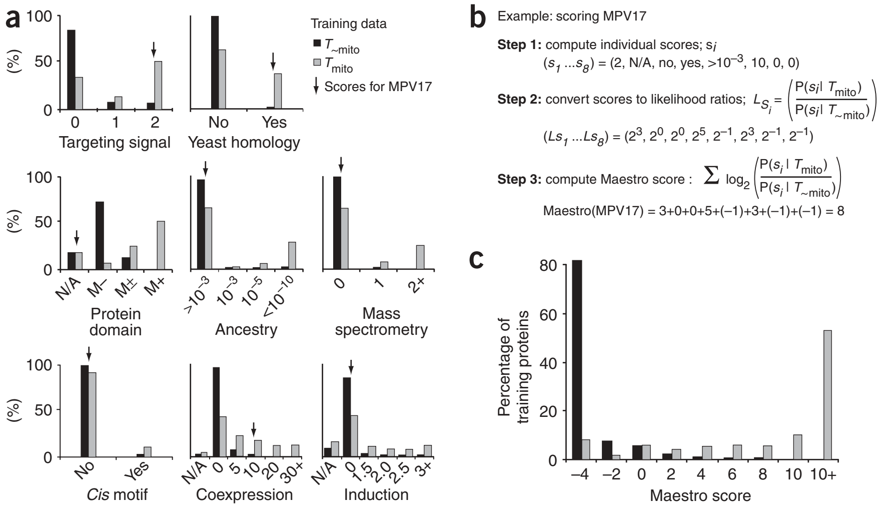

Figure 2 Integration of eight genome-scale approaches. (**a**) For each feature, the distribution of scores is plotted for the known mitochondrial proteins versus the known nonmitochondrial proteins. See Methods for complete details. (**b**) An example of the computation of the Maestro score for a query protein, MPV17. The arrows in **a** indicate the eight scores for MPV17, which are each converted to a likelihood ratio based on the training data distributions in **a** (probability of score given T_{mito} / probability of score given $T_{\sim\mathrm{mito}}$). The eight log-likelihood ratios are summed to compute the final Maestro score in a naive Bayesian integration. (**c**) The distribution of Maestro scores is plotted for training data, computed using cross-validation.

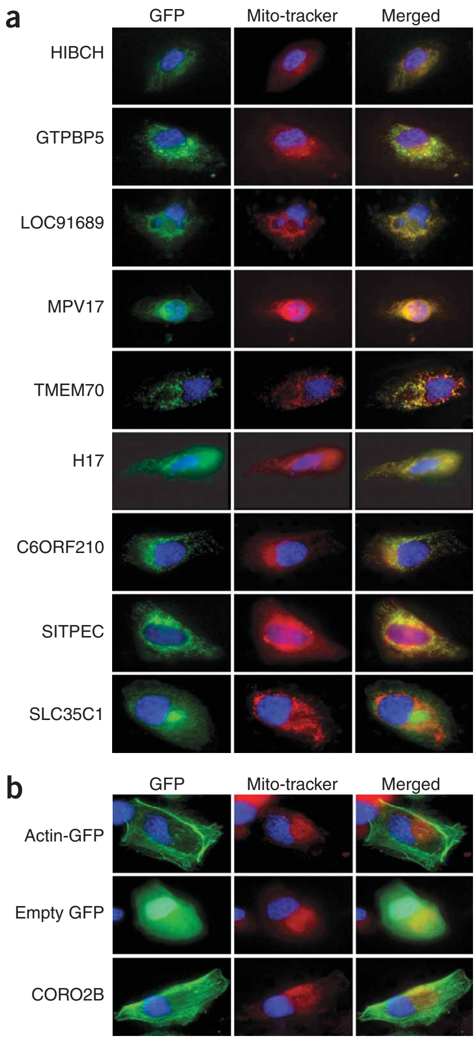

Figure 3 Experimental validation of novel mitochondrial predictions. GFP fusion constructs of selected mitochondrial predictions or controls were expressed in HeLa cells, stained with markers for mitochondria (MitoTracker Red) and nuclei (Hoechst, blue) and were then analyzed by fluorescence microscopy. (**a**) Nine novel Maestro predictions were analyzed, and all but SLC35C1 showed mitochondrial localization. (**b**) Negative controls actin, GFP and CORO2B (predicted to be mitochondrial by MitoPred and TargetP but not by Maestro) were analyzed and showed nonmitochondrial localization.

of 10% and specificity of 99.4%, Maestro properly predicted 71% of the known mitochondrial proteins (**Fig. 2c**) as well as an additional 797 proteins (encoded by 592 genes) not in the training data. Nearly half of these proteins or their mammalian orthologs are annotated with gene ontology or keyword terms associated with mitochondria, and the remaining 490 (encoded by 368 genes) have no apparent link to this organelle and thus are completely novel predictions. Our novel predictions show considerable overlap with MitoPred[14], the best existing computational prediction algorithm, but with greater sensitivity and specificity on our training data (**Supplementary Fig. 1** online). Although our method does not seem to be biased with respect to protein function, molecular weight, charge or abundance (data not shown), it seems to have lower sensitivity (14/38) for proteins localizing to the outer mitochondrial membrane[2], which may represent evolutionarily recent mitochondrial acquisitions, given the lower number of homologs in fungi and bacteria (data not shown). The 490 novel predictions include a large number of previously uncharacterized proteins as well as characterized proteins, such as the Toll signaling pathway protein SITPEC[15] (**Fig. 3a**), which we now link to the mitochondrion.

To assess the accuracy of the 490 novel protein predictions, we used a computational approach as well as two experimental techniques. First, using tenfold cross-validation (in rotation, training on nine-tenths of the data and reserving one-tenth for testing), we correctly predicted 70% of T_{mito} (sensitivity) and 99.5% of $T_{\sim\mathrm{mito}}$ (specificity) at a genome-wide false discovery rate of 10% (comparable to the 71% sensitivity and 99.4% specificity achieved without cross-validation).

Second, we used a targeted proteomics approach (using a technique known as dynamic inclusion) to test 30 selected proteins to determine if they were detected in highly purified liver mitochondria. We specifically analyzed MS/MS spectra of peptide fragments with molecular weights matching an ‘inclusion list’ of target peptides, chosen to contain ten novel predictions, ten negative controls ($T_{\sim\mathrm{mito}}$ proteins) and ten positive controls (T_{mito} proteins not previously identified using MS/MS). The purified mitochondrial extract from mouse liver contained peptide spectra matching 100% of novel predictions, 0% of negative controls and 70% of positive controls (see Methods and **Supplementary Table 2** online).

Third, we used epitope tagging and fluorescence microscopy to validate selected candidates spanning a wide range of scores. We chose nine novel predictions at a range of Maestro scores (6–36), two negative controls (actin and GFP) and one protein (CORO2B) predicted to be mitochondrial by other computational tools[5,14] but not by Maestro (a score of –3). We tested mitochondrial localization of these 12 proteins using a combination of GFP tagging and fluorescence microscopy (see Methods). When expressed in HeLa cells, neither of the negative controls localized to the mitochondrion (**Fig. 3**), whereas 8/9 Maestro predictions showed mitochondrial localization (HIBCH, GTPBP5, LOC91689, MPV17, TMEM70, H17, C6ORF210, SITPEC). The CORO2B protein showed nonmitochondrial localization, consistent with its low Maestro score. Together, these three approaches confirm mitochondrial localization for 18/19 novel predictions and support the robustness of the Maestro predictions.

The expanded collection of 1,451 human mitochondrial proteins (1,080 genes) represents the most complete set to date and is useful for identifying genes underlying human diseases characterized by mitochondrial pathology. These disorders are clinically characterized by neurological disease (seizures, strokes, ataxia), skeletal and cardiac muscle myopathy, blindness, deafness, diabetes or lactic acidosis[16,17]. The molecular basis for the majority of cases presenting with these symptoms remains unknown, and although several hundred genes may be involved, only a few dozen have been successfully identified using strategies such as linkage analysis, homozygosity mapping, candidate gene sequencing or chromosomal transfer[18–20]. These methods typically implicate large chromosomal intervals containing many genes that, in principle, can be prioritized by our list of mitochondrial predictions.

In order to assess whether this approach could be effective, we applied it to all mitochondrial disorders with previously identified underlying nuclear genes. We compiled a list of 56 nuclear genes underlying clinical mitochondrial disorders by carefully reviewing the literature[16,17,21] (**Supplementary Table 3** online). We then retrained Maestro by conservatively removing all 2,004 genes related to any disease phenotype according to the Online Mendelian Inheritance in Man (OMIM) database. Of the 56 known mitochondrial disease genes, Maestro correctly identified 86% as localized to the

Table 2 Novel candidates for mitochondrial diseases

Disease (OMIM)	Clinical symptoms	Linkage region	Size (Mb)	Gene loci	Mitochondrial candidates
Hepatic mtDNA depletion	Encephalomyopathy, liver failure, hepatocerebral mtDNA depletion	D2S2373– D2S2259 (ref. 4)	21.9	151	HADHB, HADHA, ASXL2, MRPL33, PRO1853, COX7A2L, MPV17, CAD, TP53I3, SLC30A6, EIF2B4, RBJ
MEHMO (300148)	Mental retardation, epileptic seizures, hypogonadism and hypogenitalism, microcephaly and obesity	CYBB–DXS365 (ref. 24)	18.0	70	MGC4825, ENSG00000182432, PDK3, GK, ACOT9, PRDX4
Friedreich ataxia 2 (601992)	Autosomal recessive ataxia	D9S285–D9S1874 (ref. 25)	21.1	147	HINT2, STOML2, NDUFB6, DNAJA1, ACO1
Paragangliomas 2 (601650)	Tumors of the head and neck including the carotid body	D11S956–PYGM (ref. 26)	6.1	158	PRDX5, GLYAT, GLYATL2, GLYATL1, FLJ20487, COX8A, MRPL16, BAD, LRP16, TRPT1
Multiple mitochondrial dysfunctions syndrome (605711)	Feeding difficulty, weakness, lethargy, decreasing responsiveness after birth	A053XF9–D2S441 (ref. 27)	8.6	44	ENSG00000119838, MDH1, CCT4, RAB1A
Striatonigral degeneration, infantile (271930)	Choreoathetosis, abnormal eye movements, seizures, mental retardation	D19S596–D19S867 (ref. 28)	1.3	65	BCAT2, BAX
Optic atrophy 4 (605293)	Autosomal dominant optic atrophy	D18S34–D18S479 (ref. 29)	8.8	39	ATP5A1, ACAA2
Wolfram Syndrome, mitochondrial form (604928)	Insulin-dependent diabetes mellitus and optic atrophy	D4S1591–D4S3240 (ref. 30)	7.6	35	HADHSC, PPA2
Total			**93.4**	**709**	**43**

For each mitochondrial disease, (column 1) we narrow the search of gene candidates within the linkage interval (column 3) from all gene loci (column 5) down to a small number of mitochondrial candidates (column 6, ordered by decreasing score, with novel Maestro predictions underlined).

mitochondrion. For the subset of the 29 human disease genes identified through linkage analysis, Maestro typically reduced the number of candidates from ∼100 genes in the linkage interval to about three mitochondrial candidates and, in 86% of the cases, correctly predicted the causal gene as encoding a mitochondrial protein.

We next applied our predictions to eight human mitochondrial disorders that have been mapped to genomic intervals but for which no causal gene has yet been identified (**Table 2**). For each disease, we reduced the large number of linked genes to a manageable number of candidates, relying on a threshold corresponding to 15% false discovery rate. We identified mitochondrial candidates for all eight disorders and provided novel candidates for five of them. Many of the novel candidates represent genes of unknown function that otherwise would not have warranted further investigation. The eight diseases include a novel form of hepatic mtDNA depletion, an X-linked lethal pediatric syndrome termed MEHMO, and multiple mitochondrial dysfunction syndrome (**Table 2**).

For one of the eight diseases, hepatic mtDNA depletion syndrome, we went one step further and resequenced candidate genes in patients and controls. In a companion paper[4], we report the sequencing of these predictions in three unrelated families, which led to the discovery of segregating mutations in the prioritized candidate gene *MPV17*. Despite prior literature suggesting peroxisomal localization of MPV17 (ref. 22), our analysis indicated a high Maestro score for mitochondrial localization, as confirmed through fluorescence microscopy (**Fig. 3**) and detailed subcellular localization studies[4].

In summary, we have integrated eight complementary genomic approaches to expand the catalog of human mitochondrial proteins. Whereas previous methods to compile this catalog have relied on sequence properties of the proteins[5,14], we have used additional clues about their ancestry and gene regulation to improve coverage and specificity. Although the augmented catalog represents a significant step forward, we believe there are still another ∼500 genes yet to be identified. With advances in high-throughput experimental methods to detect localization, refined methods to identify targeting signals, and more extensive training data, the goal of a comprehensive mitochondrial proteome will become achievable. Although the expanded inventory of mitochondrial proteins has proven valuable in discovering the molecular basis of monogenic diseases, in the future such a catalog may enable us to chart the role of the mitochondrion in common human disorders such as type 2 diabetes, cardiomyopathy and neurodegenerative diseases. Finally, with increasing availability of genome-scale data sets, the integrative approach applied here to the mitochondrion can be extended readily to other cellular pathways in order to tackle a broader range of human diseases.

METHODS

Human and mouse data sets. All genomic methods were applied to a common set of 33,860 human proteins from the Ensembl database. For the experiments performed on mouse models (MS/MS, induction, mouse tissue coexpression), mouse proteins were mapped to human counterparts based on an Ensembl orthology mapping that relies on synteny and gene sequence similarity (EnsMart). As the Ensembl orthology mapping is performed at the gene level (using the longest transcript for each gene locus), we computed a protein-level orthology mapping with each protein inheriting all orthologs from its gene locus (**Supplementary Fig. 2** online). As one human protein can have multiple mouse protein orthologs, a human protein is assigned the maximum ortholog score (separately for each data set).

Training sets. T_{mito} was obtained from MitoP2 and mapped to Ensembl proteins using SwissProt/Trembl identifiers (707 unique SwissProt/Trembl identifiers mapped to 654 Ensembl proteins). $T_{\sim\mathrm{mito}}$ was created from the set of all Ensembl human and mouse orthologs with GO annotations to specific compartments outside of the mitochondrion (**Supplementary Table 1**).

Targeting sequence (s_1). A subset of the known nuclear-encoded mitochondrial proteins contain an N-terminal amphiphilic α helix that directs import into the organelle. TargetP v1.1 predicts the subcellular location (mitochondrion, secretory pathway or other) on the basis of the N-terminal 130-residue protein sequence. Because of the high false discovery rate, we increased specificity by considering targeting signals in orthologous mouse proteins. Human proteins were assigned scores of 0–2, indicating mitochondrial targeting signals present within zero, one or two of the ortholog pairs.

Protein domain (s_2). Following MitoPred's methodology[14] for identifying mitochondrial domains, we used the ~60,000 SwissProt eukaryotic proteins containing annotations for 'subcellular location' (release 48.8). We filtered out low-confidence annotations (excluding 'by similarity', 'potential', 'probable' and 'possible' entries) and partitioned the rest into two sets: S_{mito}, containing 3,459 mitochondrial proteins, and $S_{\sim\mathrm{mito}}$, containing 15,322 proteins localized to other compartments (**Supplementary Methods** online). Pfam domains were determined for each protein based on the Sanger Center's precomputed analysis. We assigned each Pfam domain a categorical score (M+, M–, M± or N/A) on the basis of whether the SwissProt proteins containing the domain were exclusively from S_{mito}, exclusively from $S_{\sim\mathrm{mito}}$, found in both S_{mito} and $S_{\sim\mathrm{mito}}$, or not present in either set. Note that for cross-validation studies, all human proteins were removed from S_{mito} to avoid overestimating sensitivity.

***Cis*-regulatory motifs (s_3).** Binding sites of three transcription factors have been shown to lie upstream of mitochondrial genes: Errα (TGACCTTG), Gapba (GGAARY) and NRF1 (GCGCNYGCGC)[10]. For each motif, we identified all genes with a binding site occurring within the 2-kb window surrounding the annotated transcription start site of orthologous genes in both the human and mouse genomes. Of the three motifs, only Errα was specific enough to be informative (likelihood $L = 4$), and genes containing this motif were assigned a categorical score of 1 or 0 depending on the presence of a motif in the vicinity of the annotated transcription start site in both the human and mouse orthologs.

Yeast homology (s_4). The mitochondrial proteome of the yeast *S. cerevisiae* has been extensively studied by experimental approaches. Using the Saccharomyces genome database, which currently lists 749 mitochondrial yeast genes, we identify potential mammalian homologs based on a simple all-versus-all protein comparison between species. A human protein was assigned a categorical score of 1 if the best yeast homolog (BLASTP expect value $< 1 \times 10^{-3}$, coverage $> 50\%$ of longer gene) was annotated as mitochondrial in yeast and was assigned a score of 0 otherwise.

Ancestry (s_5). Because the mitochondrion is theorized to have evolved from a bacterial endosymbiont, we searched for ancestral bacterial homology by comparing all human proteins to the closest bacterial progenitor of mitochondria, *R. prowazekii*[11] (GenBank AJ235269). As homology is difficult to determine at this distance, we assign each human protein a similarity score (BLASTP expect) to the best *R. prowazekii* homolog.

Gene coexpression (s_6). Because functionally related genes tend to share expression patterns, we score every gene for its expression similarity to the set of known mitochondrial genes (T_{mito}). We define a 'N50' metric as the number of T_{mito} genes within a gene's 50 closest neighbors (euclidean distance)[10]. We used two expression studies that have been shown to be the most informative for coexpression of mitochondrial genes: the GNF1 survey (GEO GSE1133) of gene expression across 61 mouse tissues (GNF1M)[12] and 79 human tissues (Affymetrix HG-U133A and GNF1B)[12]. Because not all human transcripts were represented on the chips for the human GNF survey, we increased sensitivity by combining data from human and mouse tissues: the N50 values were averaged for orthologs present in both the human and mouse GNF sets; otherwise, the value from either the human or mouse GNF data was used. Probe set identifiers were mapped to Ensembl protein identifiers via data in EnsMart for the HG-U133A chip. Probe sets were assigned to all matching Ensembl proteins (for example, alternate transcripts), and Ensembl proteins matching more than one probe set were assigned the highest N50 score. This mapping was not available for the GNF1 chips; thus, the mapping was computed by comparing the individual probe sequences for the GNF1 chips against the Ensembl cDNA transcript sequences (Mega BLAST with the following parameters: percent identity = 100%, word size = 20, nucleotide mismatch penalty = −50) and ensuring that at least 7 of the 11 probes per probe set all hit the same gene. To identify genes with informative expression patterns, microarray rows were clipped to smooth low-intensity values (any expression level < 20 was replaced with 20) and normalized to mean = 0 and variance = 1. Rows lacking a post-normalization value > 1.5 were excluded. A total of 29,806 human transcripts had probes meeting the filtering requirements in either the human or mouse GNF surveys and were assigned scores (0–50) based on the N50 metric. For cross-validation studies, the N50 metric was recalculated for each set of training data.

Mass spectrometry (s_7). We reanalyzed the data from a previous survey[6] of mitochondrial proteins from four mouse tissues (liver, kidney, heart, brain) by comparing the original spectra to the current Ensembl protein database, with tryptic constraints and initial mass tolerances < 0.13 Da in the search software Mascot (Matrix Sciences). We then scored each human protein with the total number of tissues (0–4) in which its mouse ortholog achieved a Mascot score > 20.

Transcriptional activation during mitochondrial proliferation (s_8). Cultured mouse myoblasts (C2C12 cells) were differentiated into myotubes and on day 3 were infected with an adenovirus expressing either green fluorescent protein (GFP) or PGC-1α[13,23]. Extending previous studies[23], gene expression was measured in triplicate at three time points (days 1, 2 and 3) by hybridizing targets to the Affymetrix MG-U74v2 set (A,B, and C chips containing 28,381 probe sets). Results from the 63 samples were deposited in the Gene Expression Omnibus database (GEO). Data from the three chips were concatenated, and then the microarray intensities were sample normalized via linear fit to the median scan. The score represents fold change in expression; dividing average intensity in PGC1α-treated cells (average of replicates on days 2, 3) by average intensity in GFP control cells. Only those probes showing a significant difference between case and control ($P < 0.05$; one-tailed heteroscedastic Student's *t*-test) were considered (5,927 probe sets).

Integration of genome-scale data sets. We explored a variety of computational methods for combining features provided by the eight different genome-scale data sets, including naive Bayes, decision trees and boosting (**Supplementary Methods**). Of the methods we tested, a simple naive Bayesian integration, as outlined previously[8], yielded the most accurate predictions.

Briefly, we use the training sets T_{mito} and $T_{\sim\mathrm{mito}}$ to convert each of the eight individual genome-scale scores ($s_1 \ldots s_8$) into a likelihood ratio, defined as $L(s_1 \ldots s_8) = P(s_1 \ldots s_8 | T_{\mathrm{mito}}) / P(s_1 \ldots s_8 | T_{\sim\mathrm{mito}})$, which is then simplified to

$$L(s_1 \ldots s_8) = \prod_{i=1}^{8} \frac{P(s_i \,|T_{\mathrm{mito}})}{P(s_i \,|T_{\sim\mathrm{mito}})}$$

assuming that the features are independent. We define the Maestro score for a gene product as log L (**Fig. 2b**), which we assign to every gene product in the human genome. An underlying assumption of the naive Bayes procedure is that the individual data sets are independent of each other, although in practice this assumption can rarely be strictly satisfied, which may lead to overly optimistic estimates of the likelihood for some genes. We tried to minimize this effect by using a relatively high threshold to maintain a high specificity for the prediction. Of note, we find that the Maestro score is linear with respect to the true likelihood over a range of scores, but at high scores it clearly overestimates the likelihood (**Supplementary Fig. 3** online). Therefore, the Maestro score is a proxy for the likelihood, but care should be taken in interpreting high scores.

In order to compare performance of data sets in **Table 1** and **Figure 1**, we chose the following thresholds based on the differential distribution of scores on training data (**Fig. 2a**): targeting signal, 1; domain, M+; *cis* motif, yes; yeast homology, yes; ancestry, 1×10^{-3}; coexpression, 10; mass spectrometry, 1; induction, 1.5.

False discovery rates. The false discovery rate (FDR) is the proportion of all predictions that are false; FDR = FP / (FP + TP), where FP and TP represent the false positives and true positives, respectively, estimated from gold-standard negative and positive training sets. If the sizes of the training sets do not accurately reflect the prior odds (O_{prior}) of the predictions, then the FP and TP must be first scaled to avoid underestimating the false positive rate. We scale by the training set sizes by the following computation: genome-wide FDR = $(1 - \text{SP})/(1 - \text{SP} + \text{SN} \times O_{\text{prior}})$, where specificity SP = TN/(TN + FP), sensitivity SN = TP/(TP + FN) and O_{prior} = 1,500/21,000 (TN, true negatives; FN, false negatives).

Validation by tandem mass spectrometry. We selected 30 proteins from within the set of mouse proteins not previously identified in MS/MS studies[6] that showed intermediate mRNA expression in liver tissue[12] (10th–90th percentile, equivalent to expression values 80–1,300). Within this set, we selected ten high-scoring novel Maestro predictions, ten randomly selected $T_{\sim\text{mito}}$ proteins and ten randomly selected T_{mito} proteins. The ten novel predictions selected were NP_848710, BC051227, Mterfd3, Lace1, NP_061376, NP_776146, NP_080687, Q9DCB8, D5ertd33e and NP_079619.

Mitochondria were prepared from livers of C57BL/6J mice by a combination of density centrifugation and Percoll purification, as previously described[6], and were tested for purity by immunoblot analysis. Duplicate lanes of purified mitochondrial proteins were separated by size on a 10–20% gradient SDS-PAGE. We excised 20 slices from each gel lane and then reduced, alkylated and subjected them to in-gel tryptic digestion. Peptides extracted from the gel slices were then analyzed by reverse-phase liquid chromatography tandem mass spectrometry using an LTQ-Orbitrap (Thermo). Mass spectra were acquired by targeted acquisition using inclusion lists derived from a set of 30 proteins, representing between 5 to 12 peptides per protein, with MS/MS fragmentation selection criteria of masses set within a very narrow mass window. MS/MS spectra were quality filtered and then searched against the Ensembl mouse protein database (see above) using the software tool Spectrum Mill MS Proteomics Workbench. See **Supplementary Methods** and **Supplementary Table 2** for additional details.

Cell culture, transfection, and microscopy. Full-length cDNAs (Invitrogen and Origene) corresponding to ten selected predictions (HIBCH, GTPBP5, LOC91689, MPV17, TMEM70, H17, C6ORF210, SLC35C1, SITPEC and CORO2B) and two negative controls were amplified by PCR (using Qiagen Taq polymerase) with sequence-specific primers that contained restriction enzymes sites. In addition, forward primers included a Kozak sequence (CCACC), and reverse primers were designed to eliminate stop codons and to be in-frame with the C-terminal GFP. The PCR products were cloned into the pacGFP1-N2 vector (Clontech), and the sequence was verified on the 5′ ends.

Approximately 1×10^5 HeLa cells were seeded in 24-well plates and incubated overnight in DMEM supplemented with 10% FBS at 37 °C in a humidified 5% CO_2 atmosphere. We added 2 µl of Lipofectamine 2000 (Invitrogen) to 48 µl of Opti-MEM I Reduced Serum Medium (Invitrogen) and incubated the mixture at 22 °C for 5 min. We added 2.5 µg of DNA to a final volume of 50 µl Opti-MEM I medium, combined this with the transfection mixture and then added it to the cells. These transfected cells were incubated for 24 h and then transferred to eight-well coverglass plates. Cells were stained with 50 nM MitoTracker Red CMXRos and 1:10,000 diluted Hoechst 33258 (Molecular Probes) for 30 min at 37 °C and were washed twice with PBS. Cells were subsequently fixed with 3.7% formaldehyde in PBS for 15 min at room temperature. Cells were washed twice with PBS and mounted in SlowFade Gold anti-fade media. Fluorescence microscopy was performed with a 63× oil-immersion objective on a Zeiss wide-field microscope. Multiple images were captured for the constructs and reviewed for colocalization of GFP and MitoTracker red signals.

Data access. In addition to predicting the human mitochondrial proteome, we performed the analogous Bayesian integration on all mouse proteins. Data for the eight data sets and Maestro predictions are provided for the 33,860 human proteins (**Supplementary Table 4** online) and the 31,037 mouse proteins (**Supplementary Table 5** online).

URLs. Emsembl and EnsMart: http://www.ensembl.org (10 January 2005 and 1 February 2005, respectively); MitoP2: http://ihg.gsf.de/mitop2 (10 January 2005); Pfam: ftp://ftp.sanger.ac.uk/pub/databases/Pfam/ (23 January 2006); Saccharomyces genome database: ftp://ftp.yeastgenome.org/yeast (18 January 2005).

Accession codes. Microarray data are available from GEO (GSE4330).

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS
We thank C. Guda for performing MitoPred analysis, J. Bunkenborg for performing Mascot searches using previously published mass spectrometry data, J. Evans of the Massachusetts Institute of Technology for assistance with microscopy and L. Gaffney for assistance with illustrations. We thank N. Patterson, L. Peshkin, B. Gewurz and E. Lander for valuable discussions and review of the manuscript. This work is funded by a grant from the United Mitochondrial Disease Foundation, a Burroughs Wellcome Fund Career Award in the Biomedical Sciences and a grant from the American Diabetes Association/ Smith Family Foundation awarded to V.K.M.

COMPETING INTERESTS STATEMENT
The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Andreoli, C. *et al.* MitoP2, an integrated database on mitochondrial proteins in yeast and man. *Nucleic Acids Res.* **32**, D459–D462 (2004).
2. Cotter, D., Guda, P., Fahy, E. & Subramaniam, S. MitoProteome: mitochondrial protein sequence database and annotation system. *Nucleic Acids Res.* **32**, D463–D467 (2004).
3. Lopez, M.F. *et al.* High-throughput profiling of the mitochondrial proteome using affinity fractionation and automation. *Electrophoresis* **21**, 3427–3440 (2000).
4. Spinazzola, A. *et al.* *MPV17* encodes an inner mitochondrial membrane protein and is mutated in infantile hepatic mitochondrial DNA depletion. *Nat. Genet.*, advance online publication 2 April 2006 (doi:10.1038/ng1765).
5. Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *J. Mol. Biol.* **300**, 1005–1016 (2000).
6. Mootha, V.K. *et al.* Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. *Cell* **115**, 629–640 (2003).
7. Taylor, S.W. *et al.* Characterization of the human heart mitochondrial proteome. *Nat. Biotechnol.* **21**, 281–286 (2003).
8. Jansen, R. *et al.* A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302**, 449–453 (2003).
9. Prokisch, H. *et al.* Integrative analysis of the mitochondrial proteome in yeast. *PLoS Biol.* **2**, e160 (2004).
10. Mootha, V.K. *et al.* Erralpha and Gabpa/b specify PGC-1alpha-dependent oxidative phosphorylation gene expression that is altered in diabetic muscle. *Proc. Natl. Acad. Sci. USA* **101**, 6570–6575 (2004).
11. Andersson, S.G. *et al.* The genome sequence of Rickettsia prowazekii and the origin of mitochondria. *Nature* **396**, 133–140 (1998).
12. Su, A.I. *et al.* A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc. Natl. Acad. Sci. USA* **101**, 6062–6067 (2004).
13. Lin, J. *et al.* Transcriptional co-activator PGC-1 alpha drives the formation of slow-twitch muscle fibres. *Nature* **418**, 797–801 (2002).
14. Guda, C., Fahy, E. & Subramaniam, S. MITOPRED: a genome-scale method for prediction of nucleus-encoded mitochondrial proteins. *Bioinformatics* **20**, 1785–1794 (2004).
15. Kopp, E. *et al.* ECSIT is an evolutionarily conserved intermediate in the Toll/IL-1 signal transduction pathway. *Genes Dev.* **13**, 2059–2071 (1999).
16. Finsterer, J. Mitochondriopathies. *Eur. J. Neurol.* **11**, 163–186 (2004).
17. Zeviani, M. Mitochondrial disorders. *Suppl. Clin. Neurophysiol.* **57**, 304–312 (2004).
18. Rotig, A. & Munnich, A. Genetic features of mitochondrial respiratory chain disorders. *J. Am. Soc. Nephrol.* **14**, 2995–3007 (2003).
19. Scaglia, F. *et al.* Clinical spectrum, morbidity, and mortality in 113 pediatric patients with mitochondrial disease. *Pediatrics* **114**, 925–931 (2004).

20. Shoubridge, E.A. Nuclear gene defects in respiratory chain disorders. *Semin. Neurol.* **21**, 261–267 (2001).
21. Thorburn, D.R. Mitochondrial disorders: prevalence, myths and advances. *J. Inherit. Metab. Dis.* **27**, 349–362 (2004).
22. Zwacka, R.M. *et al.* The glomerulosclerosis gene Mpv17 encodes a peroxisomal protein producing reactive oxygen species. *EMBO J.* **13**, 5129–5134 (1994).
23. Mootha, V.K. *et al.* PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267–273 (2003).
24. Steinmuller, R., Steinberger, D. & Muller, U. MEHMO (mental retardation, epileptic seizures, hypogonadism and -genitalism, microcephaly, obesity), a novel syndrome: assignment of disease locus to xp21.1-p22.13. *Eur. J. Hum. Genet.* **6**, 201–206 (1998).
25. Christodoulou, K. *et al.* Mapping of the second Friedreich's ataxia (FRDA2) locus to chromosome 9p23-p11: evidence for further locus heterogeneity. *Neurogenetics* **3**, 127–132 (2001).
26. Mariman, E.C., van Beersum, S.E., Cremers, C.W., Struycken, P.M. & Ropers, H.H. Fine mapping of a putatively imprinted gene for familial non-chromaffin paragangliomas to chromosome 11q13.1: evidence for genetic heterogeneity. *Hum. Genet.* **95**, 56–62 (1995).
27. Seyda, A. *et al.* A novel syndrome affecting multiple mitochondrial functions, located by microcell-mediated transfer to chromosome 2p14–2p13. *Am. J. Hum. Genet.* **68**, 386–396 (2001).
28. Basel-Vanagaite, L. *et al.* Infantile bilateral striatal necrosis maps to chromosome 19q. *Neurology* **62**, 87–90 (2004).
29. Kerrison, J.B. *et al.* Genetic heterogeneity of dominant optic atrophy, Kjer type: Identification of a second locus on chromosome 18q12.2–12.3. *Arch. Ophthalmol.* **117**, 805–810 (1999).
30. El-Shanti, H., Lidral, A.C., Jarrah, N., Druhan, L. & Ajlouni, K. Homozygosity mapping identifies an additional locus for Wolfram syndrome on chromosome 4q. *Am. J. Hum. Genet.* **66**, 1229–1236 (2000).