Mitochondial Parts, Pathways, and Pathogenesis

by

Sarah E. Calvo

Submitted to the Harvard-MIT Division of Health Sciences and Technology
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN
BIOINFORMATICS AND INTEGRATIVE GENOMICS

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2009

ARCHIVES

MASSACHUSETTS INSTITUTE
OF TECHNOLOGY

SEP 17 2009

LIBRARIES

© Sarah E. Calvo, MMIX. All rights reserved.

The author hereby grants to MIT permission to reproduce and distribute publicly
paper and electronic copies of this thesis document in whole or in part.

Author ....
Harvard-MIT Division of Health Sciences and Technology
May 1, 2009

Certified by ....
Vamsi K. Mootha MD
Associate Professor
Department of Systems Biology, Harvard Medical School
Thesis Supervisor

Accepted by ....
Ram Sasisekharan PhD
Director, Harvard-MIT Division of Health Sciences and Technology
Chairman, Department Committee on Graduate Theses

Mitochondial Parts, Pathways, and Pathogenesis

by

Sarah E. Calvo

Submitted to the Harvard-MIT Division of Health Sciences and Technology
on May 1, 2009, in partial fulfillment of the
requirements for the degree of
DOCTOR OF PHILOSOPHY IN
BIOINFORMATICS AND INTEGRATIVE GENOMICS

Abstract

Mitochondria are cellular compartments that perform essential roles in energy metabolism, ion homeostasis, and apoptosis. Mitochondrial dysfunction causes disease in 1 in 5,000 live births and also has been associated with aging, neurodegeneration, cancer, and diabetes. To systematically explore the function of mitochondria in health and in disease, it is necessary to identify all of the proteins resident in this organelle and to understand how they integrate into pathways. However, traditional molecular and biochemistry methods have identified only half of the estimated 1200 mitochondrial proteins, including the 13 encoded by the tiny mitochondrial genome. Now, newly available genomic technologies make it possible to identify the remainder and explore their roles in cellular pathways and disease.

Toward this goal, we performed mass spectrometry, GFP tagging, and machine learning on multiple genomic datasets to create a mitochondrial compendium of 1098 genes and their protein expression across 14 mouse tissues. We linked poorly characterized proteins in this inventory to known mitochondrial pathways by virtue of shared evolutionary history. We additionally used our matched mRNA and protein measurements to demonstrate a widespread role of upstream open reading frames (uORFs) in blunting translation of mitochondrial and other cellular proteins.

Next we used the mitochondrial protein inventory to identify genes underlying inherited diseases of mitochondrial dysfunction. In collaboration with clinicians, we identified causal mutations in five genes underlying diseases including hepatocerebral mtDNA depletion syndrome, autosomal dominant mitochondrial myopathy, and several forms of inherited complex I deficiency. These discoveries have enabled the development of diagnostic tests now widely available. More broadly, the mitochondrial compendium provides a foundation for systematically exploring the organelle's contribution to both basic cellular biology and human disease.

Thesis Supervisor: Vamsi K. Mootha MD
Title: Associate Professor
Department of Systems Biology, Harvard Medical School

Acknowledgments

I would like to thank Vamsi Mootha for being an extraordinary mentor, whose enthusiasm, creativity, collaborative spirit and scientific rigor have helped shaped my own research approach. Without his guidance, my graduate career would not have been nearly as productive, enjoyable, or short. I would also like to thank David Altshuler and Eric Lander for their guidance on my thesis committee and for providing opportunities for me to share my research. I am particularly indebted to my friend and co-author Dave Pagliarini for sharing his knowledge and skills over the past three years. I thank my fellow Mootha lab members and BIG colleagues for their support and willingness to discuss ideas, and Rick Mitchell for his outstanding introduction to Pathology. I am grateful to the Broad Institute for its resources and collaborative mission, and to Steve Carr, Betty Chang, Shao-en Ong, Karl Clauser, Mike Zody, Xiaohui Xie, Manuel Garber, and Nick Patterson who made this collaborative atmosphere a daily reality.

Lastly, I thank my family and close friends who have been a constant source of encouragement. In particular, my parents, Rita and Joseph Calvo, and my grandparents, Joseph and Esther Aronson Rothenberg, have been a continued inspiration to me. I give special thanks to my sister Naomi and my partner Scott for making these graduate school years a joy.

Contents

Chapter 1

—

Introduction

Introduction

Analysis of the human genome sequence has revealed that our DNA contains roughly 20,000 protein-coding genes[1,2]. The function and cellular location of several thousand proteins have been established by detailed biochemical and molecular studies, which together have elucidated core pathways such as energy metabolism, cell signaling, and replication. Progress in characterizing the remaining proteins is being fueled by technological advances that allow simultaneous measurement of thousands of genes or proteins. However, methods to interpret these genomic and proteomic data are still in their infancy. Currently we know little beyond the sequence of roughly a third of all human genes.

In this dissertation, I focus on characterizing the protein components of one specific compartment of mammalian cells, the mitochondrion. Mitochondria generate the majority of the cell's supply of chemical energy, as well as performing critical roles in biosynthesis, intracellular signaling and apoptosis[3]. Mitochondrial dysfunction causes diseases ranging from neonatal fatalities to adult neurodegeneration, and is also associated with diabetes, cancer, and aging[4]. At the start of this project, approximately half of the estimated 1200 mitochondrial proteins had been identified through decades of biochemical and molecular studies. Identifying the rest will enable a systematic approach to understanding mitochondrial function and the molecular basis of disease. The goals of this dissertation work are to define a more comprehensive list of mitochondrial protein parts, to determine how some of these proteins function together in pathways, to identify regulatory elements within mitochondrial genes, and to identify causal mutations underlying mitochondrial disease.

In this chapter, I introduce the role of mitochondria in cellular metabolism, discuss how mitochondrial defects lead to disease, and review current progress in defining mitochondrial protein composition.

Mitochondrial form and function

Mitochondria are eukaryotic cellular structures separated from the cytoplasm by a double membrane. These organelles are best known for their role in generating adenosine triphosphate (ATP) by oxidative phosphorylation, which provides the chemical energy for most cellular reactions. In addition to ATP generation, mitochondria house machinery for the Krebs cycle, urea cycle, ion homeostasis, apoptosis, and the biosynthesis/metabolism of amino acids, fatty acids, steroids, lipids, cardiolipin, ubiquinol, iron-sulfur clusters, and heme[3].

Mitochondria are the only eukaryotic organelles, beside the nucleus, to contain DNA. Mitochondrial DNA (mtDNA) was first discovered in 1963[5], and human mtDNA was first sequenced in 1981[6] revealing a circular molecule containing 16,569 base pairs. Each mitochondrion contains approximately five copies of mtDNA[4]. There is strong evidence that mitochondria descended from an endosymbiotic bacterium early in eukaryotic evolution[7]. In the subsequent one and a half billion years of evolution, most of the endosymbiont's genome was lost or transferred to the host genome. Currently, mammalian mtDNA contains only 13 protein-coding genes, along with the ribosomal and tRNA genes needed for protein synthesis[7]. All remaining protein components are encoded by nuclear DNA (nDNA), translated in the cytoplasm, and actively transported into the mitochondrion.

Human mitochondria differ widely across tissues in number, structure, and function. Mitochondria number varies from hundreds to thousands per cell, depending on the tissue's metabolic demand, and can occupy up to 25% of cellular volume[7]. Interestingly, egg cells contain 100,000-300,000 copies of mtDNA, the largest number in any cell type[4]. Structurally, mitochondria exhibit morphological diversity across tissues. For example, cardiomyocyte mitochondria have tightly folded inner membranes (cristae) which increase OXPHOS capacity and ATP generation, whereas steroid-secreting cells typically have mitochondria with tubular cristae[3]. Functionally, mitochondria are highly specialized as well. For example, thermogenesis occurs mainly in brown adipose mitochondria, heme is synthesized in bone marrow mitochondria, and gluconeogenesis occurs mainly in liver and kidney mitochondria[7].

Mitochondrial disease

Since mitochondria are essential for so many cellular functions, it is not surprising that their defects lead to a variety of human diseases. Mitochondrial dysfunction causes well over 50 diseases ranging from neonatal fatalities to adult onset neurodegeneration (Table 1), and is a likely contributor to cancer and type II diabetes[8-10]. Additionally, acquired mitochondrial defects have been associated with aging[3,11]. Primary mitochondrial disease has an estimated prevalence of 1 in 5000 live births and is one of the most common inborn errors of metabolism[4].

Mitochondrial dysfunction causes a wide range of clinical presentations, usually involving highly metabolic tissues. Clinical features may include myopathy, encephalopathy, lactic acidosis, neurodegeneration, deafness, blindness, GI dysmotility, diabetes, and liver disease[4]. These disorders exhibit incredible heterogeneity and tissue-specificity[4]. For example, different mutations in gene $tRNA^{Ile}$ can cause cardiomyopathy or progressive external ophthalmoplegia[4]. In some cases family members with the same molecular mutation have different affected organ systems[4]. For some maternally inherited disorders, disease tissue-specificity is caused by the percent of mtDNA genomes carrying the mutation in a particular tissue (skewed heteroplasmy). For other cases, the molecular basis of tissue-specificity is not understood.

Hereditary mitochondrial diseases can show maternal, autosomal recessive, autosomal dominant, or X-linked inheritance. Maternal inheritance occurs for mutations in mtDNA, since almost all zygote mtDNA are derived from the egg cell, and paternal mtDNA is specifically degraded[4]. The one known case of paternal mtDNA inheritance[12] is likely the exception that proves the rule[13,14]. Mendelian recessive, dominant, and X-linked patterns of inheritance occur from single nDNA gene mutations. The complex patterns of expressivity and tissue-specificity of mitochondrial diseases may well result from the presence of multiple interacting genetic variants.

For patients with mitochondrial disease, diagnosis is extremely difficult and few effective therapies exist. Currently diagnosis of a "definite mitochondrial disease" requires complex clinical algorithms that utilize clinical, biochemical, imaging, and molecular features. Clinical criteria include the number of affected organ systems and specific forms of muscle and CNS abnormalities[15,16]. Biochemical indicators include abnormal lactate, pyruvate and alanine levels (measured from serum or cerebrospinal fluid) and enzymatic activity of respiratory chain complexes measured from invasive muscle biopsy samples[15,16]. Imaging includes MRI analysis of brain abnormalities[15,16]. Molecular features include identification of known pathogenic mutations in mtDNA or in several nDNA genes[15,16]. With the possible exception of molecular mutation, none of the features is specific enough for accurate diagnosis. Thus for clinical use, several complicated scoring algorithms have been developed to combine the diverse indicators, including the Walker[17], Bernier[15], and Morava criteria[16]. The diagnosis process is long, expensive, and often inconclusive. Additionally, few treatments are available apart from simply managing symptoms. Patients are often provided a nutritional supplement therapy including mixtures of coenzyme Q_{10}, L-carnitine, folic acid, creatine, lipoic acid, B_1, B_2, and/or B_{12}, although none of these supplements have shown sustained efficacy[4].

Over the past twenty years, research has elucidated the molecular basis of over fifty mitochondrial diseases[18]. Approximately 15% of patients have mutations in their mtDNA, which either arise spontaneously or are inherited maternally[4,19]. Over 300 different disease-causing mutations have been identified within human mtDNA[20]. The remaining 85% of patients likely have mutations in nDNA. To date, disease-causing mutations in 86 nDNA genes have been discovered (Table 1). Most of these encode proteins targeted to the mitochondrion, while a handful encode cytoplasmic or nuclear proteins regulating mitochondrial function (e.g. TAZ, PUS1, RRM2B).

Most known nDNA disease genes have been discovered by genetic analysis of large consanguineous families. Approaches such as linkage analysis, homozygosity mapping, or chromosomal transfer can narrow the search to a small chromosomal region[21-23], typically containing over 100 genes. Next, candidate genes are sequenced to find mutations that segregate with the disease. Pathogenic mutations can be experimentally validated by rescue of phenotype in patient cells or cellular models. These approaches are limited by the availability of large consanguineous families, and cannot be used to investigate sporadic cases.

Discovering additional disease-related nDNA genes will aid in diagnosis and treatment of mitochondrial disease. Identifying a causal gene defect can facilitate understanding of the pathogenesis of the disease, and can be used to create a cost effective diagnostic test. In addition, discovery of disease-related genes enables a molecular classification of mitochondrial disorders. A more fine-grained disease classification enables the better prediction of the disease progression, which is of substantial benefit to families, and the ability to assess treatment of specific disease forms. Thus identifying the gene defects underlying mitochondrial diseases enables development of patient diagnostics and possibly even therapies.

In this work, we implement a systematic approach to discovering nDNA genes underlying mitochondrial dysfunction. We first compile a catalog of mitochondrial protein parts, then infer the pathway function and regulation of a subset of genes, and finally apply the catalog to pinpoint candidate genes for mitochondrial disease based on inferred function or location within linkage regions. In the next section, I review recent approaches to define the mitochondrial proteome.

Defining the mitochondrial proteome

A comprehensive parts list of the human mitochondrion is an important resource for understanding the function of its essential pathways and for systematically elucidating disease processes.

Before cataloging the mitochondrial proteome, we first must specify the definition of a mitochondrial protein. This is complicated by several factors:

(i) proteins may localize to multiple subcellular compartments

(ii) proteins may localize to the mitochondrion only under certain conditions (e.g. apoptosis factors)

(iii) genes may have only a subset of splice forms that code for mitochondrial proteins

(iv) proteins may localize transiently to the outer mitochondrial membrane (e.g. transport, fusion and fission proteins)

In this thesis, I define the term mitochondrial proteome as the set of all proteins that reside within the mitochondria (or outer membrane) in at least one tissue or condition. Mitochondrial genes refer to those that encode at least one mitochondrial protein.

It is not known how many different proteins comprise the mammalian mitochondrion, although there are several methods that provide estimates. The most quoted figure of 1500 proteins was estimated in rat liver by Lopez et al. who isolated mitochondria by centrifugation and sucrose gradients, separated proteins by size and acidity on a 2-D gel, silver-stained for protein content, and counted the number of distinct silver spots[24]. However, that estimate mistakenly includes contaminant proteins and misses low-abundance proteins or those having similar isoelectric point and molecular weight. Another method uses homology to the genes determined to be mitochondrial in model organisms such as yeast. As yeast has over 800 mitochondrial proteins[25], it is assumed that the more complex mammalian organisms will have at least this many proteins.

For any given protein, there are several "gold-standard" methods to determine whether it resides in the mitochondrion. One method is to create a genetically tagged construct with an epitope or green fluorescent protein (GFP) tag, and to use microscopy to confirm that the reporter construct co-localizes with mitochondria. This method was used to identify 332 mitochondrial-localized proteins in yeast[26]. While this method is highly specific, its sensitivity is limited by interference of the tag in protein import, over-expression artifacts, and presence of necessary chaperones, modifiers and/or conditions in the tested cell type. A second method is to show that the protein is protected from proteinase K degradation in intact mitochondria, but not in mitochondria lacking a membrane potential[27,28]. While both of these labor-intensive methods can provide gold-standard proof of mitochondrial localization, not all proteins are amenable to these techniques. At the start of this project, only about 600 mammalian proteins had solid evidence of mitochondrial location[29].

In addition to focused methods, several groups have pioneered high-throughput methods to define the mitochondrial proteome, including mass spectrometry based proteomics, genetics, computational targeting sequence prediction, and gene expression. These methods are described in more detail below.

Proteomics

Mass spectrometry-based proteomics allows the identification of hundreds of proteins within a complex mixture. This method has been used to identify protein components of mitochondria purified from cells using centrifugation and density gradients. Over the past six years, studies of enriched mitochondria have identified over 800 proteins in yeast[30,31]; 615 from human heart[32]; 591 from mouse brain, heart, kidney, and liver[33]; 689 from rat muscle, heart and liver[34]; 297 from mouse liver[35]; 2533 from mouse brain, heart, kidney, liver, lung, and placenta[36]; 1130 from adipocyte 3T3-L1 cells[37]; and 1,162 from rat brain, liver, heart, and kidney[38]. While these studies represent substantial progress in defining the mitochondrial proteome across tissues, the mass spectrometry approach suffers from two major flaws. First, it has limited sensitivity for low abundance proteins. Second, it is extremely limited by the purity of the mitochondrial enrichment process, as the approach will identify co-purifying contaminants. As mitochondria are physically tethered to the endoplasmic reticulum and other organelles it is not possible to completely isolate these organelles and thus all proteomic data must be interpreted cautiously and analyzed by additional means to exclude contaminants.

Genetics

Mitochondrial components can be identified by impaired mitochondrial phenotypes in organisms with defective or missing proteins. In yeast, existing deletion strains for over 5000 genes enable a screen for mutants showing decreased growth on non-fermentable substrates compared to fermentable sugars. Steinmetz and colleagues used this approach to identify 466 putative mitochondrial genes whose deletion impaired respiration[39]. Dimmer and colleagues similarly identified 341 yeast proteins with respiration defects and an additional 15 with mitochondrial morphology abnormalities[40]. While this approach works well in yeast, it is less amenable to high throughput implementation in mammalian systems. Additionally, the genetics approach will miss proteins that are either redundant or essential[41].

Targeting sequence prediction

Experiments have shown that a short N-terminal protein sequence is sufficient to direct protein import into the mitochondria[42]. The identified signal contains an alpha-helix with positively charged residues on one side, and uncharged and hydrophobic residues on the other[42]. A plethora of computational algorithms to identify this three dimensional targeting signal have been developed, including TargetP[43], pTARGET[44], PSORT[45], iPSORT[46], Predotar[47], ngLoc[48], MitPred[49], MitoPred[50], and MitoProt[51]. However, this three-dimensional signal is not present in many *bona fide* mitochondrial proteins, and thus these methods are not sensitive. Additionally, these tools generate a high percent of false positives.

Gene expression

Identifying genes that are coordinately expressed during mitochondrial biogenesis is another high-throughput method to discover mitochondrial components. This method was pioneered in 1997 by DeRisi, Iyer and Brown who created microarrays to assay yeast mRNA transcripts activated during the metabolic shift from fermentation to respiration[52]. Mootha and colleagues used microarrays to profile mouse transcripts upregulated during mitochondrial biogenesis induced by overexpression of PGC-1alpha[53]. While these methods, and other coexpression studies across diverse tissues, can highlight co-regulated mitochondrial genes, they fail to detect mitochondrial components that are low-abundance, not tightly co-regulated, or activated only in specific cellular conditions.

While each of these high-throughput techniques provides clues to mitochondrial localization, none is individually accurate enough to confidently define the mitochondrial proteome. By integrating complementary data sets, we can hope to better define the mitochondrial proteome and use this information to aid disease gene discovery.

Integrative approaches to defining protein function and subcellular location

Several groups have integrated genomic data from multiple sources in order to define protein function[54,55] and subcellular localization[41,56-58]. These approaches are largely successful because the underlying data sources have complementary strengths and weaknesses. In general, the approaches use supervised learning, based on a training set of known positive and negative controls, in order to classify unknown data points. Common classifiers include Bayesian networks, decision trees, support vector machines and ensemble approaches such as boosting and bagging (see Appendix D). Classification accuracy is typically assessed by the prediction of known elements excluded from the training set, termed cross-validation.

Bayesian methods have been particularly successful at combining heterogeneous, noisy, and incomplete genomic data sets. Briefly, these methods define a prior probability that a protein has a given subcellular location (or function), and then update the probability using Bayes rule and available data in order to assign a posterior probability of subcellular location (or function). In its simplest form, naïve Bayes, the data sources (features) are assumed to be conditionally independent. Drawid and Gerstein developed a naïve Bayesian system to predict the subcellular location of ~6000 yeast proteins based on 30 features, such as sequence motifs and mRNA expression[56]. They achieved 75% accuracy, using training sets of ~1300 yeast proteins. Jansen, Gerstein and colleagues used a similar approach to predict yeast protein-protein interactions using nine genomic datasets including yeast two-hybrid, *in vivo* pull-down, and mRNA expression[54]. Because some of the datasets were highly dependent, Jansen et al. used both a fully connected Bayes subnetwork and a naïve Bayes sub-network to assign interaction probabilities. Accuracy in this setting is difficult to assess, since there are few negative controls for non-interacting proteins, but experimental validation of predicted interactions showed encouraging results[54]. Independently, Lee, Marcotte and colleagues developed a Bayesian-based interaction network in yeast designed to reconstruct functional, rather than physical, gene interactions[55]. They integrated mRNA coexpression, gene fusions, phylogenetic profiles, literature co-citation, and protein interaction experiments. These network reconstructions in yeast were possible only because of the high quality genomic datasets available for this model organism.

A variety of groups have used integrative methods specifically to predict the mitochondrial proteome. For example, Prokisch, Steinmetz and colleagues used an *ad hoc* integration of 22 genomic datasets in yeast[58]. For each gene they computed the MitoP2 score as the maximum *R* score (1- false discovery rate) from any individual feature or any combination of features[58]. Later, Prokisch and colleagues applied other integrative approaches including a support vector machine[57], linear classifier[41] and neural network[41] to refine the MitoP2 scores and apply these methods to yeast, human, and other model organisms[57,59,60]. While the *ad hoc* and support vector machine approaches can produce integrated scores, these scores are not readily interpretable, unlike the Bayesian approaches.

The above methods all integrate different types of genomic data using machine learning. A different integrative approach, based solely on sequence data, predicts gene function based on shared evolutionary history. In an elegant paper, Li and colleagues identified the molecular components of flagella and basal bodies using comparative genomics[61]. In their simple phylogenetic profiling approach, they identified 688 genes present in ciliated organisms (human and *Chlamydomonas*) that had no homologs in an un-ciliated organism (*Arabidopsis*) and showed that six of them disrupted flagellar activities when knocked down by RNAi. They used their collection to identify candidates within a linkage region for Bardet-Biedl syndrome, which is caused by basal body dysfunction[62], and discovered causal gene mutations[61]. Other, more sophisticated, phylogenetic approaches have also been developed to annotate gene function based on shared evolutionary history[63-71], but have not been applied toward disease gene discovery.

A comprehensive parts list enables systematic discovery of human mitochondrial disease genes

Here, we apply Bayesian integration methods, pioneered in model organisms[54,56], to identify the protein composition of human mitochondria (Chapter 2). We combine these predictions with large-scale mass spectrometry based proteomics and GFP-tagging/microscopy to define a high-quality "MitoCarta" inventory of 1098 mitochondrial genes and their protein expression across 14 mammalian tissues (Chapter 3). We estimate that this inventory contains ~100 false positives and misses ~100 *bona fide* mitochondrial proteins – thus we are approaching our goal of a comprehensive list of mitochondrial protein parts.

We use this inventory to elucidate the tissue-specificity, pathway function and regulation of a subset of mitochondrial proteins. First, we compare the differences in protein composition across mammalian tissues (Chapter 3). Second, we use shared evolutionary history to identify 19 proteins likely involved in the function of complex I of the electron transport chain (Chapter 3). Third, we take advantage of our mass spectrometry measurements to investigate the widespread role of upstream open reading frames in blunting translation of mitochondrial and other genes (Chapter 4). These insights illustrate how our mitochondrial protein inventory provides a foundation for exploring mitochondrial biology.

Finally, our protein inventory enables systematic discovery of molecular defects underlying human mitochondrial diseases. By intersecting our inventory with linkage intervals for inherited disorders, we can quickly pinpoint causal gene defects for inherited diseases – as we have shown for hepatocerebral mtDNA depletion[29,72] and complex I deficiency[28,73] (Chapters 2 and 3). We are currently tackling sporadic cases of complex I deficiency using new sequencing technologies (Chapter 5). Next, planned initiatives will use MitoCarta to identify variants in a broad spectrum of primary mitochondrial diseases. These discoveries can be immediately applied to improving diagnosis of primary mitochondrial disorders and to providing candidate genes for complex diseases that involve mitochondrial dysfunction.

The dissertation work described here was performed in an extremely collaborative environment. I designed and performed all computational aspects of the research. I helped to design all experiments in collaboration with colleagues expert in biochemistry, molecular biology, mass-spectrometry and mitochondrial disease. All experimental and clinical work was performed solely by my collaborators.

Together, our work toward the identification of mitochondrial parts, pathways and pathogenesis has helped elucidate the function of this organelle within the cell and provides a foundation for future systematic investigations of mitochondrial function in human health and disease.

Mitochondrial Diseases with nDNA Mutations	**nDNA genes**
Alpha methylacetoacetic aciduria	ACAT1
Anemia, sideroblastic, and spinocerebellar ataxia (ASAT)	ABCB7
Barth syndrome (BTHS)	TAZ
Cardioencephalomyopathy due to COX deficiency	SCO2
Charcot-Marie-Tooth disease, axonal, type 2A2 (CMT2A2)	MFN2
Coenzyme Q10 deficiency	PDSS1, PDSS2, COQ2, APTX, CABC1, ETFDH
Combined oxidative phosphorylation deficiency	EFG1, MRPS16, TSFM, TUFM, MRPS22
Complex 1 deficiency	NDUFS1, NDUFS2, NDUFS4, NDUFS6, NDUFS7, NDUFV1, NDUFV2, NDUFA1, NDUFAF1, NDUFAF2, C6orf66, **C20orf7**, ***FOXRED1***
Complex IV deficiency	COX10, COX6B1, SCO1
Complex V deficiency	ATPAF2, **TMEM70**
Encephalopathy, ethymalonic	ETHE1
Encephalopathy, hepatomegaly	HMGCL
Friedreich ataxia	FXN
Glycine encephalopathy	GLDC, GCST, GCSH
Gracile syndrome	BCS1L
HMG-CoA synthase deficiency	HMGCS2
Homozygous 2p16 deletion syndrome	PPM1B, PREPL, SLC3A1
Leigh syndrome (LS)	COX15, DLD, NDUFS3, NDUFS8, SDHA, SURF1, **C8orf38**
Leigh syndrome, French-Canadian type (LSFC)	LRPPRC
Leiomyoma, hereditary multiple, of skin	FH
Leukoencephalopathy with brainstem and spinal cord involvement and lactate elevation	DARS2
Microcephaly, amish type (MCPHA)	SLC25A19
Mitochondrial complex III deficiency	UQCRB, UQCRQ
Mitochondrial myopathy and sideroblastic anemia (MLASA)	PUS1
Mitochondrial neugastrointestinal encephalopathy syndrome	ECGF1
Mitochondrial phosphate carrier deficiency	SLC25A3
Mitochondrial trifuctional protein deficiency	HADHA, HADHB
Mohr-Tranebjaerg sydrome (MTS)	TIMM8A
mtDNA depletion syndrome, encephalomyopathic form with renal tubulopathy	RRM2B
mtDNA depletion syndrome, encephalomyopathic form	SUCLG1
mtDNA depletion syndrome, hepatocerebral form	DGUOK, **MPV17**
mtDNA depletion syndrome, myopathic form	TK2, SUCLA2
Myopathy, autosomal dominant	**CHCHD10**
Optic atrophy 1 (OPA1)	OPA1
Paragangliomas	SDHB, SDHC, SDHD
Parkinson disease 6, autosomal recessive early-onset	PINK1
Pontocerebellar hypoplasia	RARS2
Progressive external ophthalmoplegia, mtDNA deletions	POLG, SLC25A4, C10orf2, POLG2
Pyruvate carboxylase deficiency	PC
Pyruvate decarboxylase deficiency	PDHA1
Spastic paraplegia	HSPD1, REEP1, SPG7
Wilson disease	ATP7B
Wolfram Syndrome, mitochondrial form	WFS1

Table 1. Mitochondrial disorders caused by nuclear DNA mutations. Bold indicates genes discovered with aid of Maestro or MitoCarta. Compiled by Mohit Jain.

References

1. Lander, E.S. et al. Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
2. Clamp, M. et al. Distinguishing protein-coding and noncoding genes in the human genome. *Proc Natl Acad Sci U S A* **104**, 19428-33 (2007).
3. Scheffler, I.E. *Mitochondria*, (Wiley, 2008).
4. DiMauro, S., Hirano, M. & Schon, E.A. *Mitochondrial Medicine*, 348 (Informa Healthcare, 2006).
5. Nass, M.M. & Nass, S. Intramitochondrial Fibers with DNA Characteristics. I. Fixation and Electron Staining Reactions. *J Cell Biol* **19**, 593-611 (1963).
6. Anderson, S. et al. Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457-65 (1981).
7. Lodish, H. et al. *Molecular Cell Biology*, (W.H. Freeman and Company, 2004).
8. DiMauro, S. & Schon, E.A. Mitochondrial respiratory-chain diseases. *N Engl J Med* **348**, 2656-68 (2003).
9. Lowell, B.B. & Shulman, G.I. Mitochondrial dysfunction and type 2 diabetes. *Science* **307**, 384-7 (2005).
10. Wallace, D.C. A mitochondrial paradigm of metabolic and degenerative diseases, aging, and cancer: a dawn for evolutionary medicine. *Annu Rev Genet* **39**, 359-407 (2005).
11. Friguet, B., Bulteau, A.L. & Petropoulos, I. Mitochondrial protein quality control: implications in ageing. *Biotechnol J* **3**, 757-64 (2008).
12. Schwartz, M. & Vissing, J. Paternal inheritance of mitochondrial DNA. *N Engl J Med* **347**, 576-80 (2002).
13. Schwartz, M. & Vissing, J. No evidence for paternal inheritance of mtDNA in patients with sporadic mtDNA mutations. *J Neurol Sci* **218**, 99-101 (2004).
14. Taylor, R.W. et al. Genotypes from patients indicate no paternal mitochondrial DNA contribution. *Ann Neurol* **54**, 521-4 (2003).
15. Bernier, F.P. et al. Diagnostic criteria for respiratory chain disorders in adults and children. *Neurology* **59**, 1406-11 (2002).
16. Morava, E. et al. Mitochondrial disease criteria: diagnostic applications in children. *Neurology* **67**, 1823-6 (2006).
17. Walker, U.A., Collins, S. & Byrne, E. Respiratory chain encephalomyopathies: a diagnostic classification. *Eur Neurol* **36**, 260-7 (1996).
18. Online Mendelian Inheritance in Man, OMIM. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD) and National Center for Biotechnology Information, National Library of Medicine (Bethesda, MD).
19. Dimauro, S. & Davidzon, G. Mitochondrial DNA and disease. *Ann Med* **37**, 222-32 (2005).

20. Ruiz-Pesini, E. et al. An enhanced MITOMAP with a global mtDNA mutational phylogeny. *Nucleic Acids Res* **35**, D823-8 (2007).
21. Rotig, A. & Munnich, A. Genetic features of mitochondrial respiratory chain disorders. *J Am Soc Nephrol* **14**, 2995-3007 (2003).
22. Scaglia, F. et al. Clinical spectrum, morbidity, and mortality in 113 pediatric patients with mitochondrial disease. *Pediatrics* **114**, 925-31 (2004).
23. Shoubridge, E.A. Nuclear gene defects in respiratory chain disorders. *Semin Neurol* **21**, 261-7 (2001).
24. Lopez, M.F. et al. High-throughput profiling of the mitochondrial proteome using affinity fractionation and automation. *Electrophoresis* **21**, 3427-40 (2000).
25. Issel-Tarver, L. et al. Saccharomyces Genome Database. *Methods Enzymol* **350**, 329-46 (2002).
26. Kumar, A. et al. Subcellular localization of the yeast proteome. *Genes Dev* **16**, 707-19 (2002).
27. Pon, L.A. & Schon, E.A. *Mitochondria*, (Academic Press, 2007).
28. Sugiana, C. et al. Mutation of C20orf7 disrupts complex I assembly and causes lethal neonatal mitochondrial disease. *Am J Hum Genet* **83**, 468-78 (2008).
29. Calvo, S. et al. Systematic identification of human mitochondrial disease genes through integrative genomics. *Nat Genet* **38**, 576-82 (2006).
30. Reinders, J., Zahedi, R.P., Pfanner, N., Meisinger, C. & Sickmann, A. Toward the complete yeast mitochondrial proteome: multidimensional separation techniques for mitochondrial proteomics. *J Proteome Res* **5**, 1543-54 (2006).
31. Sickmann, A. et al. The proteome of Saccharomyces cerevisiae mitochondria. *Proc Natl Acad Sci U S A* **100**, 13207-12 (2003).
32. Taylor, S.W. et al. Characterization of the human heart mitochondrial proteome. *Nat Biotechnol* **21**, 281-6 (2003).
33. Mootha, V.K. et al. Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. *Cell* **115**, 629-40 (2003).
34. Forner, F., Foster, L.J., Campanaro, S., Valle, G. & Mann, M. Quantitative proteomic comparison of rat mitochondria from muscle, heart, and liver. *Mol Cell Proteomics* **5**, 608-19 (2006).
35. Foster, L.J. et al. A mammalian organelle map by protein correlation profiling. *Cell* **125**, 187-99 (2006).
36. Kislinger, T. et al. Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling. *Cell* **125**, 173-86 (2006).
37. Adachi, J., Kumar, C., Zhang, Y. & Mann, M. In-depth analysis of the adipocyte proteome by mass spectrometry and bioinformatics. *Mol Cell Proteomics* **6**, 1257-73 (2007).

38. Johnson, D.T. et al. Tissue heterogeneity of the mammalian mitochondrial proteome. *Am J Physiol Cell Physiol* **292**, C689-97 (2007).

39. Steinmetz, L.M. et al. Systematic screen for human disease genes in yeast. *Nat Genet* **31**, 400-4 (2002).

40. Dimmer, K.S. et al. Genetic basis of mitochondrial function and morphology in Saccharomyces cerevisiae. *Mol Biol Cell* **13**, 847-53 (2002).

41. Perocchi, F., Mancera, E. & Steinmetz, L.M. Systematic screens for human disease genes, from yeast to human and back. *Mol Biosyst* **4**, 18-29 (2008).

42. Pfanner, N. Protein sorting: recognizing mitochondrial presequences. *Curr Biol* **10**, R412-5 (2000).

43. 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**, 1005-16 (2000).

44. Guda, C. pTARGET: a web server for predicting protein subcellular localization. *Nucleic Acids Res* **34**, W210-3 (2006).

45. Nakai, K. & Horton, P. PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem Sci* **24**, 34-6 (1999).

46. Bannai, H., Tamada, Y., Maruyama, O., Nakai, K. & Miyano, S. Extensive feature detection of N-terminal protein sorting signals. *Bioinformatics* **18**, 298-305 (2002).

47. Small, I., Peeters, N., Legeai, F. & Lurin, C. Predotar: A tool for rapidly screening proteomes for N-terminal targeting sequences. *Proteomics* **4**, 1581-90 (2004).

48. King, B.R. & Guda, C. ngLOC: an n-gram-based Bayesian method for estimating the subcellular proteomes of eukaryotes. *Genome Biol* **8**, R68 (2007).

49. Kumar, M., Verma, R. & Raghava, G.P. Prediction of mitochondrial proteins using support vector machine and hidden Markov model. *J Biol Chem* **281**, 5357-63 (2006).

50. Guda, C., Fahy, E. & Subramaniam, S. MITOPRED: a genome-scale method for prediction of nucleus-encoded mitochondrial proteins. *Bioinformatics* **20**, 1785-94 (2004).

51. Claros, M.G. & Vincens, P. Computational method to predict mitochondrially imported proteins and their targeting sequences. *Eur J Biochem* **241**, 779-86 (1996).

52. DeRisi, J.L., Iyer, V.R. & Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680-6 (1997).

53. 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 U S A* **101**, 6570-5 (2004).

54. Jansen, R. et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302**, 449-53 (2003).

55. Lee, I., Date, S.V., Adai, A.T. & Marcotte, E.M. A probabilistic functional network of yeast genes. *Science* **306**, 1555-8 (2004).
56. Drawid, A. & Gerstein, M. A Bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *J Mol Biol* **301**, 1059-75 (2000).
57. Prokisch, H. et al. MitoP2: the mitochondrial proteome database--now including mouse data. *Nucleic Acids Res* **34**, D705-11 (2006).
58. Prokisch, H. et al. Integrative analysis of the mitochondrial proteome in yeast. *PLoS Biol* **2**, e160 (2004).
59. Elstner, M. et al. MitoP2: an integrative tool for the analysis of the mitochondrial proteome. *Mol Biotechnol* **40**, 306-15 (2008).
60. Prokisch, H. & Ahting, U. MitoP2, an integrated database for mitochondrial proteins. *Methods Mol Biol* **372**, 573-86 (2007).
61. Li, J.B. et al. Comparative genomics identifies a flagellar and basal body proteome that includes the BBS5 human disease gene. *Cell* **117**, 541-52 (2004).
62. Blacque, O.E. & Leroux, M.R. Bardet-Biedl syndrome: an emerging pathomechanism of intracellular transport. *Cell Mol Life Sci* **63**, 2145-61 (2006).
63. Date, S.V. & Marcotte, E.M. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat Biotechnol* **21**, 1055-62 (2003).
64. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. & Yeates, T.O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A* **96**, 4285-8 (1999).
65. Pagel, P., Wong, P. & Frishman, D. A domain interaction map based on phylogenetic profiling. *J Mol Biol* **344**, 1331-46 (2004).
66. Marcotte, E.M., Xenarios, I., van Der Bliek, A.M. & Eisenberg, D. Localizing proteins in the cell from their phylogenetic profiles. *Proc Natl Acad Sci U S A* **97**, 12115-20 (2000).
67. Wu, J., Hu, Z. & DeLisi, C. Gene annotation and network inference by phylogenetic profiling. *BMC Bioinformatics* **7**, 80 (2006).
68. Cokus, S., Mizutani, S. & Pellegrini, M. An improved method for identifying functionally linked proteins using phylogenetic profiles. *BMC Bioinformatics* **8 Suppl 4**, S7 (2007).
69. Ranea, J.A., Yeats, C., Grant, A. & Orengo, C.A. Predicting protein function with hierarchical phylogenetic profiles: the Gene3D Phylo-Tuner method applied to eukaryotic genomes. *PLoS Comput Biol* **3**, e237 (2007).
70. Barker, D. & Pagel, M. Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. *PLoS Comput Biol* **1**, e3 (2005).
71. Zhou, Y., Wang, R., Li, L., Xia, X. & Sun, Z. Inferring functional linkages between proteins from evolutionary scenarios. *J Mol Biol* **359**, 1150-9 (2006).

72. Spinazzola, A. et al. MPV17 encodes an inner mitochondrial membrane protein and is mutated in infantile hepatic mitochondrial DNA depletion. *Nat Genet* **38**, 570-5 (2006).

73. Pagliarini, D.J. et al. A mitochondrial protein compendium elucidates complex I disease biology. *Cell* **134**, 112-23 (2008).

Chapter 2

—

Systematic identification of human mitochondrial disease genes through integrative genomics

Sarah Calvo, Mohit Jain, Xiaohui Xie, Sunil A. Sheth, Betty Chang, Olga A. Goldberger, Antonella Spinazzola, Massimo Zeviani, Steven A. Carr, and Vamsi K. Mootha

This chapter is presented in the context of its contemporary science and originally appeared in *Nature Genetics* 38:576-582 (2006).

Corresponding Supplementary Material can be found in Appendix A.

Systematic identification of human mitochondrial disease genes through integrative genomics

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. A limitation to elucidating the molecular basis for these disorders is that we currently know only about half[1,2] of the estimated 1,500 mitochondrial proteins[3]. To systematically expand this catalog, we experimentally and computationally generated eight genome-scale datasets, each designed to provide clues of 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 amino terminus signal sequences that direct proteins to this organelle[5]. However, not all mitochondrial proteins are imported via 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 69% false positive predictions when applied genome-wide since 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 due to the bias toward abundant proteins, these methods have identified only an additional ~150 mitochondrial proteins[6,7]. Hence, when used alone, existing approaches exhibit 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 genome-scale datasets 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 mitochondrial-localized proteins by generating and integrating datasets 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 datasets (Fig. 1a and Methods):

Targeting sequence (s_1): presence of an amino-terminus mitochondrial targeting sequence that directs protein import into the mitochondrion, identified by TargetP[5].

Protein domain (s_2): presence of domains found exclusively in eukaryotic sequences with known mitochondrial localization (based on SwissProt annotation).

***Cis*-motifs (s_3)**: presence of evolutionarily conserved transcriptional regulatory elements that we previously discovered to be enriched upstream of mitochondrial genes[10].

Yeast homology (s_4): sequence similarity to *S. cerevisiae* proteins with experimental evidence of mitochondrial localization (Saccharomyces Genome Database annotation).

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

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

MS/MS (s_7): peptide support from mitochondria extracted from multiple mouse tissues in a previous proteomic survey[6].

Induction (s_8): up-regulation of mRNA transcripts in a cellular model of mitochondrial biogenesis. We induced mitochondrial proliferation in a muscle cell line by overexpressing the transcriptional co-activator 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 non-mitochondrial proteins ($T_{\sim\mathrm{mito}}$) annotated to localize to other cellular compartments (see Methods and Supplementary Table 1). As can be seen in Figure 1b, the limited sensitivity and the relatively low specificity of each individual approach can generate a large proportion of false positives when applied genome-wide (Fig 1a).

To improve prediction accuracy, we integrated the eight approaches using a naïve Bayes classifier[8] that we implemented with a computer program called Maestro (see Methods). We trained Maestro on the gold standard positive and negative datasets and applied it to the Ensembl set of 33,860 human proteins. For each of the eight features, a likelihood of mitochondrial localization was calculated by comparing performance on T_{mito} to $T_{\sim\mathrm{mito}}$ at a range of scores (Fig. 2a). A composite Maestro score was computed by summing the log-likelihoods of eight individual features (Fig. 2b) in a naïve Bayes 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 corresponding to 10% false discovery rate and 99.4% specificity, 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, while 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). While our method does not appear to be biased with respect to protein function, molecular weight, charge, or abundance (data not shown), it appears to have lower sensitivity (14/38) for proteins localizing to the outer mitochondrial membrane[2], which may represent evolutionarily recent mitochondrial acquisitions given the fewer homologues 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 10-fold cross-validation (in rotation, training on 9/10 of the data and reserving 1/10 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, 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 10 novel predictions, 10 negative controls ($T_{\sim mito}$ proteins) and 10 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).

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 (score -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, Figure 3 shows that neither of the negative controls localized to the mitochondrion, whereas 8/9 Maestro predictions showed mitochondrial localization (HIBCH, GTPBP5, LOC91689, MPV17, TMEM70, H17, C6ORF210, SITPEC). The COROB2 protein showed clear non-mitochondrial 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 while 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 can be effective, it was applied 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 literature[16,17,21] (Supplementary Table 3). We then re-trained 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 mitochondrial-localized. 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 ~3 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 1). 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 which 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 1).

For one of the eight diseases, hepatic mtDNA depletion syndrome, we went one step further to re-sequence candidate genes in patients and controls. In a companion paper[4], we report the sequencing of these predictions in three unrelated families that has led to the discovery of segregating mutations in the prioritized candidate gene *MPV17*. Despite prior literature suggesting peroxisomal localization of MPV17[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 additionally used clues about their ancestry and gene regulation to improve coverage and specificity. While 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. While 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. Additionally, with increasing availability of genome-scale datasets, the integrative approach applied here to the mitochondrion can be readily extended to other cellular pathways in order to tackle a broader range of human diseases.

Methods

Human and mouse datasets. All genomic methods were applied to a common set of 33,860 human proteins from the Ensembl database (www.ensembl.org, 1/10/05). For the experiments performed on mouse models (MS/MS, induction, GNF 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 2/1/05). Since 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). As one human protein can have multiple mouse protein orthologs, a human protein is assigned the maximum ortholog score (separately for each dataset).

Training sets. T_{mito} was obtained from MitoP2 (ihg.gsf.de/mitop2, 1/10/05) and mapped to Ensembl proteins using SwissProt/Trembl identifiers (707 unique SwissProt/Trembl identifiers mapped to 654 Ensembl proteins). $T_{\sim\text{mito}}$ 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 alpha helix that directs import into the organelle. TargetP v1.1 predicts the subcellular location (mitochondria, secretory pathway, or other) based on the N-terminal 130aa protein sequence. Because of the high false discovery rate, we increased specificity by additionally considering targeting signals in orthologous mouse proteins. Human proteins were assigned scores 0-2, indicating mitochondrial targeting signals present within 0, 1, or 2 of the ortholog pairs.

Protein domain (s_2). Following MitoPred's methodology[14] for identifying mitochondrial domains, we utilized the ~60,000 SwissProt eukaryotic proteins containing annotations for 'subcellular location' (release 48.8, 1/23/06). 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 $S_{\sim\text{mito}}$ containing 15,322 proteins localized to other compartments (see Supplementary Methods). Pfam domains were determined for each protein based on the Sanger Center's precomputed analysis (ftp.sanger.ac.uk/pub/databases/Pfam/current_release/swisspfam, 1/23/06). We assigned each Pfam domain a categorical score (M+,M-,M±, N/A) based on whether the SwissProt proteins containing the domain were exclusively from S_{mito}, exclusively from $S_{\sim\text{mito}}$, found in both S_{mito} and $S_{\sim\text{mito}}$, 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 2kb 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 *Saccharomyces cerevisiae* has been extensively studied by experimental approaches. Using the Saccharomyces Genome Database, which currently lists 749 mitochondrial yeast genes (ftp.yeastgenome.org/yeast, 1/18/05), we identify potential mammalian homologs based on a simple all-vs-all protein comparison between species. A human protein was assigned a categorical score of 1 if the best yeast homolog (BLASTP expect < 1e-3, coverage > 50% of longer gene) was annotated as mitochondrial in yeast, and 0 otherwise.

Ancestry (s_5). Since 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, *Rickettsia prowazekii*[11] (genbank accession AJ235269). Since homology is difficult to determine at this distance, we assign each human protein a similarity score (BLASTP expect) to the best *Rickettsia* 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 of gene expression across 61 mouse tissues (GNF1M)[12] and 79 human tissues (Affymetrix HG-U133A and GNF1B)[12] (GEO accession GSE1133). Since 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 IDs were mapped to Ensembl protein IDs via data in EnsMart (www.ensembl.org) for the HG-U133A chip. Probe sets were assigned to all matching Ensembl proteins (e.g. 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 (Megablast -p 100 -W 20 -q -50 -D 3 -f), 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, variance=1. Rows with no 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. Note that for cross-validation studies, the N50 metric was recalculated for each set of training data.

Mass spectrometry (s_7). We re-analyzed the data from a previous survey[6] of mitochondrial proteins from 4 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, London). 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 1,2,3) 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 GEO (accession GSE4330). 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 induction measured in fold-change; 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 significant difference between case and control (p<0.05, measured by 1-tailed heteroscedastic student t-test) were considered (5,927 probe-sets).

Integration of genome-scale datasets. We explored a variety of computational methods for combining features provided by the eight different genome-scale datasets, including naïve Bayes, decision trees, and boosting (See Supplementary Methods). Of the methods we tested, a simple naïve Bayes integration, as outlined by Jansen et al[8] yielded the most accurate predictions.

Briefly, we use the training sets T_{mito} and $T_{\text{-mito}}$ to convert each of the eight individual genome-scale scores ($s_1 \dots s_8$) into a likelihood ratio, defined as $L(s_1 \dots s_8) = P(s_1 \dots s_8|$

T_{mito})/$P(s_1 \ldots s_8|\ T_{\sim\text{mito}})$, which is then simplified to $L(s_{1..}s_8)=\prod_{i=1}^{8}\frac{P(s_i|T_{mito})}{P(s_i|T_{\sim mito})}$ assuming that the features are independent. We define the Maestro score for a gene product as log *L* (see Fig. 2b), which we assign to every gene product in the human genome. An underlying assumption of the naïve Bayes procedure is that the individual datasets are independent of each other, though 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). 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 datasets, for Fig. 1 display only, 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: 1e-3; coexpression: 10; mass spec: 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 postives, 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 fdr. We scale by the training set sizes by computing the genome-wide false discovery rate fdr = (1-SP)/(1-SP + SN* O_{prior}) where specificity SP =TN/(TN+FP), sensitivity SN=TP/(TP+FN), TN=true negatives, FN=false negatives, and O_{prior} = 1,500/21,000.

Validation by tandem mass spectrometry. 30 proteins were selected from within the set of mouse proteins not previously identified in MS/MS studies[6] and which showed intermediate mRNA expression in liver tissue[12] (10th-90th percentile, equivalent to expression values 80-1300). Within this set, we selected 10 high-scoring novel Maestro predictions, 10 randomly selected $T_{\sim\text{mito}}$ proteins, and 10 randomly selected T_{mito} proteins. The 10 novel predictions selected were: NP_848710, BC051227, Mterfd3, Lace1, NP_061376, NP_776146, NP_080687, Q9DCB8, D5ertd33e, 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 tested for purity using immunoblot analysis. Duplicate lanes of purified mitochondrial proteins were size separated by a 10-20% gradient SDS-PAGE. 20 slices from each gel lane were excised, reduced, alkylated, and then subjected 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, San Jose, CA). 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 10 selected predictions (HIBCH [TC115062], GTPBP5 [TC100454], LOC91689 [BC024237], MPV17 [TC118652], TMEM70 [BC002748], H17 [BC013902], C6ORF210 [BC039906], SLC35C1 [BC001427], SITPEC [BC008279], CORO2B [BC026335]) and two negative controls were amplified by PCR (TAQ polymerase from Qiagen) with sequence-specific primers that contained restriction enzymes sites. In addition, forward primers included a Kozak sequence (CCACC), while reverse primers were designed to eliminate stop codons and designed to be in-frame with the C-terminal GFP. The PCR products were cloned into the pacGFP1-N2 vector (Clontech) and then sequence verified on the 5' ends.

Approximately 1 x 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% carbon dioxide atmosphere. 2 l of Lipofectamine 2000 (Invitrogen) was added to 48 l of Opti-MEM I Reduced Serum Media (Invitrogen) and incubated at room temperature for 5 minutes. 2.5 g of DNA was added to a final volume of 50 l Opti-MEM I media and was combined with the transfection mixture and added to the cells. These transfected cells were incubated for 24 hours and then transferred to 8-well coverglass plates. Cells were stained with 50 nM MitoTracker Red CMXRos and 1:10000 diluted Hoechst 33258 (Molecular Probes) for 30 minutes at 37°C and washed twice with PBS. Cells were subsequently fixed with 3.7% formaldehyde in PBS for 15 minutes at room temperature. Cells were washed twice with PBS and mounted in SlowFade Gold anti-fade media. Fluorescence microscopy was performed at 63x oil objective using a Zeiss widefield microscope. Multiple images were captured for the constructs and reviewed for co-localization of GFP and MitoTracker red signals.

Data access. In addition to predicting the human mitochondrial proteome, we performed the analogous Bayes integration on all mouse proteins. Data for the eight datasets and Maestro predictions are provided for the 33,860 human proteins (Supplementary Table 4) and 31,037 mouse proteins (Supplementary Table 5). Microarray induction data is available from GEO (accession GSE4330).

Acknowledgements

We thank C. Guda for performing MitoPred analysis, J. Bunkenborg for performing Mascot searches using previously published mass spectrometry data, J. Evans of MIT 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 VKM.

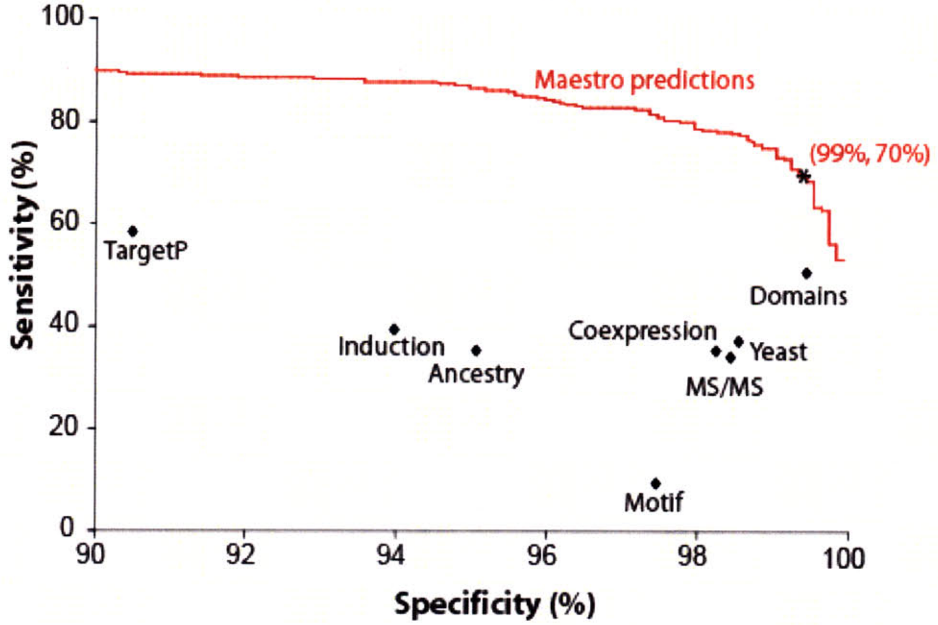
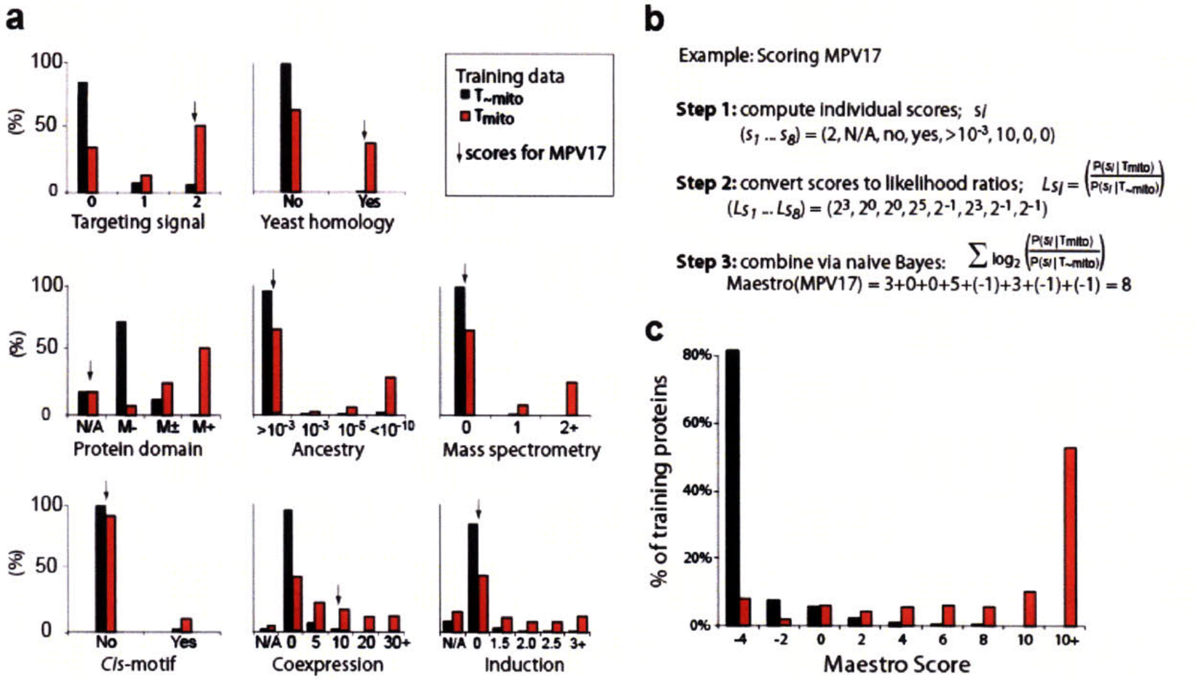


Figure 1: Accuracy of genome-wide mitochondrial prediction methods. Performance of eight individual predictors of mitochondrial localization. The rate of genome-wide false discovery (column 4) was estimated based on "gold standard" training data of 654 known mitochondrial proteins (T_{mito}) and 2,847 non-mitochondrial proteins ($T_{\sim\mathrm{mito}}$) at specific thresholds (see Methods). **b.** Tradeoff in accuracy between a dataset's sensitivity (%T_{mito} correctly predicted) and specificity (%$T_{\sim\mathrm{mito}}$ correctly predicted). The accuracy of the eight individual datasets is shown at specific thresholds (as in **a**), whereas the accuracy of Maestro is displayed at a range of thresholds (red curve), with the chosen threshold marked by an asterisk.



Example: Scoring MPV17

Step 1: compute individual scores; s_i
$(s_1 \ldots s_8) = (2, \text{N/A}, \text{no}, \text{yes}, >10^{-3}, 10, 0, 0)$

Step 2: convert scores to likelihood ratios; $Ls_i = \left(\frac{P(s_i \mid T_{mito})}{P(s_i \mid T_{\sim mito})}\right)$
$(Ls_1 \ldots Ls_8) = (2^3, 2^0, 2^0, 2^5, 2^{-1}, 2^3, 2^{-1}, 2^{-1})$

Step 3: combine via naive Bayes: $\sum \log_2 \left(\frac{P(s_i \mid T_{mito})}{P(s_i \mid T_{\sim mito})}\right)$
Maestro(MPV17) = 3+0+0+5+(-1)+3+(-1)+(-1) = 8

Figure 2. Integration of eight genome-scale approaches.

a. For each feature, the distribution of scores is plotted for the known mitochondrial proteins (red) compared to the known non-mitochondrial proteins (black). See Methods for complete details. **b.** The example shows how to compute 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 $T_{\sim mito}$). Based on a naïve Bayes integration, the eight log-likelihood ratios are summed to compute the final Maestro score. **c.** The distribution of Maestro scores is plotted for training data, computed using cross-validation.

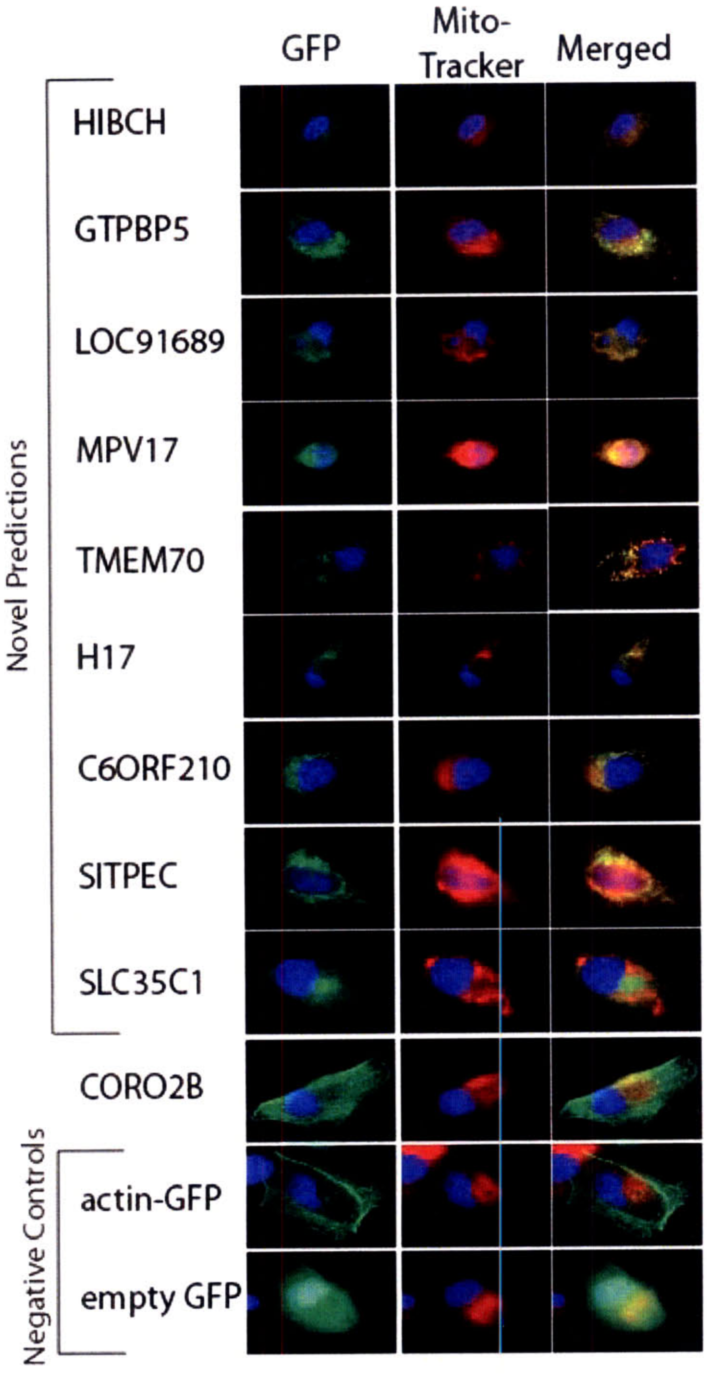
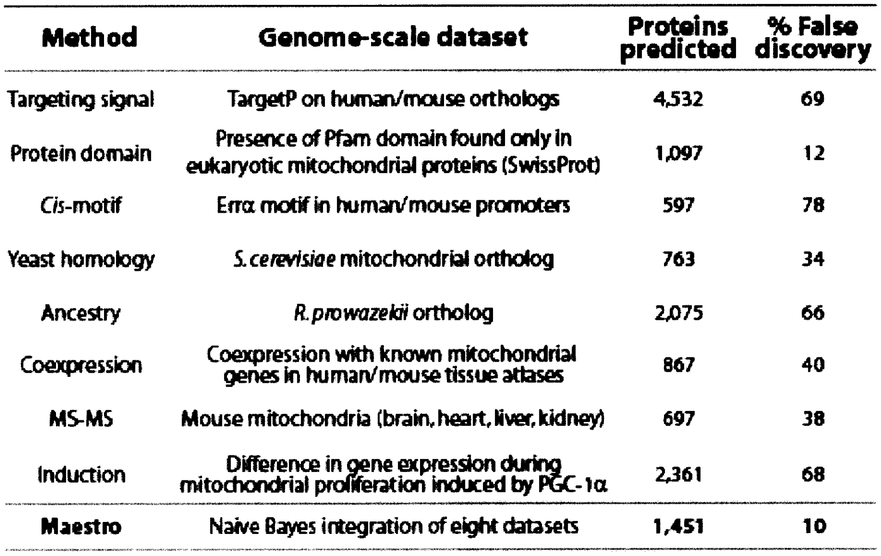


Figure 3. Experimental validation of novel mitochondrial predictions.

GFP fusion constructs of selected mitochondrial predictions or controls were expressed in HeLa cells, stained with a marker for mitochondria (MitoTracker Red) and for nuclei (Hoechst, blue), and then analyzed by fluorescence microscopy. **a.** Nine novel Maestro predictions were analyzed, and based on these and additional images all but SLC35C1 showed clear 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 clear non-mitochondrial localization.



Method	Genome-scale dataset	Proteins predicted	% False discovery
Targeting signal	TargetP on human/mouse orthologs	4,532	69
Protein domain	Presence of 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 proliferation induced by PGC-1α	2,361	68
Maestro	Naive Bayes integration of eight datasets	**1,451**	**10**

Table 1. 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.

Disease (OMIM)	Clinical symptoms	Linkage region	Size Mb	Gene loci	Mitochondrial candidates
Hepatic mtDNA depletion	encephalomyopathy, liver failure, hepatocerebral mtDNA depletion	D2S2373-D2S2259[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[24]	18.0	70	MGC4825, ENSG00000182432, PDK3, GK, ACOT9, PRDX4
Friedreich ataxia 2 (601992)	autosomal recessive ataxia	D9S285-D9S1874[25]	21.1	147	HINT2, STOML2, NDUFB6, DNAJA1, ACO1
Paragangliomas 2 (601650)	nonchromaffin glomus body tumors of the head and neck	D11S956-PYGM[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[27]	8.6	44	ENSG00000119838, MDH1, CCT4, RAB1A
Striatonigral degeneration, infantile (271930)	choreoathetosis, abnormal eye movements, seizures, mental retardation	D19S596-D19S867[28]	1.3	65	BCAT2, BAX
Optic atrophy 4 (605293)	autosomal dominant optic atrophy	D18S34-D18S479[29]	8.8	39	ATP5A1, ACAA2
Wolfram Syndrome, mitochondrial form (604928)	insulin-dependent diabetes mellitus and optic atrophy	D4S1591-D4S3240[30]	7.6	35	HADHSC, PPA2
Total			**93**	**709**	**43**

Table 2. Novel candidates for mitochondrial diseases.

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 predictions underlined).

References

1. Andreoli, C. et al. MitoP2, an integrated database on mitochondrial proteins in yeast and man. *Nucleic Acids Res* **32**, D459-62 (2004).
2. Cotter, D., Guda, P., Fahy, E. & Subramaniam, S. MitoProteome: mitochondrial protein sequence database and annotation system. *Nucleic Acids Res* **32 Database issue**, D463-7 (2004).
3. Lopez, M.F. et al. High-throughput profiling of the mitochondrial proteome using affinity fractionation and automation. *Electrophoresis* **21**, 3427-40 (2000).
4. Spinazzola, A. et al. The Mpv17 gene encodes a protein of the inner mitochondrial membrane and is mutated in infantile hepatic mitochondrial DNA depletion. *Nature Genetics (in press)* (2006).
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**, 1005-16 (2000).
6. Mootha, V.K. et al. Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. *Cell* **115**, 629-40 (2003).
7. Taylor, S.W. et al. Characterization of the human heart mitochondrial proteome. *Nat Biotechnol* **21**, 281-6 (2003).
8. Jansen, R. et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302**, 449-53 (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 U S A* **101**, 6570-5 (2004).
11. Andersson, S.G. et al. The genome sequence of Rickettsia prowazekii and the origin of mitochondria. *Nature* **396**, 133-40 (1998).
12. Su, A.I. et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* **101**, 6062-7 (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**, 1785-94 (2004).
15. Kopp, E. et al. ECSIT is an evolutionarily conserved intermediate in the Toll/IL-1 signal transduction pathway. *Genes Dev* **13**, 2059-71 (1999).
16. Finsterer, J. Mitochondriopathies. *Eur J Neurol* **11**, 163-86 (2004).
17. Zeviani, M. Mitochondrial disorders. *Suppl Clin Neurophysiol* **57**, 304-12 (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**, 925-31 (2004).

20. Shoubridge, E.A. Nuclear gene defects in respiratory chain disorders. *Semin Neurol* **21**, 261-7 (2001).

21. Thorburn, D.R. Mitochondrial disorders: prevalence, myths and advances. *J Inherit Metab Dis* **27**, 349-62 (2004).

22. Zwacka, R.M. et al. The glomerulosclerosis gene Mpv17 encodes a peroxisomal protein producing reactive oxygen species. *Embo J* **13**, 5129-34 (1994).

23. Mootha, V.K. et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* **34**, 267-73 (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**, 201-6 (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**, 127-32 (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**, 386-96 (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**, 805-10 (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**, 1229-36 (2000).

Chapter 3

—

A mitochondrial protein compendium elucidates complex I disease biology

Sarah E. Calvo∗, David J. Pagliarini∗, Betty Chang, Sunil A. Sheth, Scott B. Vafai, Shao-En Ong, Geoffrey A. Walford, Canny Sugiana, Avihu Boneh, William K. Chen, David E. Hill, Marc Vidal, James G. Evans, David R. Thorburn, Steven A. Carr, and Vamsi K. Mootha

∗These authors constributed equally to this work.

This chapter is presented in the context of its contemporary science and originally appeared in *Cell* 134(1):112-23 (2008).

Corresponding Supplementary Material can be found in Appendix B.

A mitochondrial protein compendium elucidates complex I disease biology

Mitochondria are complex organelles whose dysfunction underlies a broad spectrum of human diseases. Identifying all the proteins resident in this organelle and understanding how they integrate into pathways represent major challenges in cell biology. Toward this goal, we performed mass spectrometry, GFP tagging, and machine learning to create a mitochondrial compendium of 1098 genes and their protein expression across 14 mouse tissues. We link poorly characterized proteins in this inventory to known mitochondrial pathways by virtue of shared evolutionary history. Using this approach we predict 19 proteins to be important for the function of complex I (CI) of the electron transport chain. We validate a subset of these predictions using RNAi, including *C8orf38*, which we further show harbors an inherited mutation in a lethal, infantile CI deficiency. Our results have important implications for understanding CI function and pathogenesis, and more generally, illustrate how our compendium can serve as a foundation for systematic investigations of mitochondria.

Introduction

Mitochondria are dynamic organelles essential for cellular life, death, and differentiation. Although they are best known for ATP production via oxidative phosphorylation (OXPHOS), they house myriad other biochemical pathways and are centers for apoptosis and ion homeostasis. Mitochondrial dysfunction causes over 50 diseases ranging from neonatal fatalities to adult onset neurodegeneration, and is a likely contributor to cancer and type II diabetes[1-3]. The 13 proteins encoded by the mitochondrial genome have been known since its sequencing[4] and have been linked to a variety of maternally inherited disorders. However, there may be as many as 1500 nuclear-encoded mitochondrial proteins[5], though less than half have been identified with experimental support. A complete protein inventory for this organelle across tissues would provide a molecular framework for investigating mitochondrial biology and pathogenesis.

Recent progress in defining the mitochondrial proteome has been driven by large-scale approaches, including mass spectrometry (MS) based proteomics in mammals[6-11] and yeast[12,13], epitope tagging combined with microscopy in yeast[14,15], and computation[16-18]. However each of these methods suffers from intrinsic technical limitations. MS-based approaches struggle with distinguishing genuine mitochondrial proteins from co-purifying contaminants, and published reports exhibit up to 41% false positive rates (Table S1). Additionally, these approaches tend to miss low abundance proteins or those expressed only in specific tissues or developmental states, and thus capture only 23-40% of known mitochondrial components (Table S1). Other experimental approaches such as epitope tagging are limited by the availability of cDNA clones, tag interference, and over-expression artifacts. While integrative machine-learning methods can be more comprehensive[18,19], they require subsequent experimental validation.

Here, we perform in-depth protein mass spectrometry, microscopy, and machine learning to construct a protein compendium of the mitochondrion. We perform MS-based proteomics on both highly purified and crude mitochondrial preparations to discover genuine mitochondrial proteins and distinguish them from contaminants based on enrichment. We integrate these MS data with six other genome-scale datasets of mitochondrial localization using a Bayesian framework and additionally perform the most extensive GFP tagging study focused on mammalian mitochondria. The resulting compendium consists of 1098 genes (Figure 1) and their protein expression across 14 mouse tissues. Although not complete, this represents the most comprehensive and accurate molecular characterization of the organelle to date.

Our compendium provides a framework for identifying novel proteins within pathways resident in the mitochondrion. Here, we focus on complex I (CI) of the electron transport chain, a macromolecular structure composed of ~45 subunits in mammals[20]. CI deficiency is the most common cause of rare, respiratory chain diseases[1] and has been implicated in Parkinson's disease[21]. Half of the patients with CI deficiency lack mutations in any known CI subunit, suggesting that yet unidentified genes crucial for maturation, assembly, or stability of CI are mutated in the remaining cases[22]. Multiple assembly factors for much smaller complexes IV and V have been identified in *S. cerevisiae*, and it is estimated that complex IV alone requires over 20 factors[23,24]. However, the absence of CI in *S. cerevisiae* has impeded similar studies and, to date, only three CI assembly and maturation factors have been identified[25-27].

To systematically discover proteins essential for CI function, we apply the technique of phylogenetic profiling which uses shared evolutionary history to highlight functionally related proteins[28]. This approach was recently used to identify the CI assembly factor NDUFA12L using five yeast species[25]. We apply this approach more broadly to our mitochondrial protein inventory and report that 19 of these proteins share ancestry with a large subset of CI proteins. We validate several of these predictions in cellular models and additionally report that one of these genes, *C8orf38,* harbors a causative mutation in an inherited CI deficiency.

Together, these studies illustrate the utility of an expanded mitochondrial inventory in advancing basic and disease biology of the organelle. Our compendium, called MitoCarta, is freely available at www.broad.mit.edu/publications/MitoCarta.

Results and Discussion

Discovery and Subtractive Proteomics of Mouse Mitochondria

As a first step toward establishing an experimentally supported inventory of mammalian mitochondrial proteins, we performed protein mass spectrometry on mitochondria from 14 diverse mouse organs (Figure 1). We designed our proteomic experiments in two phases in order to identify as many mitochondrial proteins as possible while systematically flagging co-purifying contaminants. In the *discovery* phase, we isolated highly purified mitochondria from cerebrum, cerebellum, brainstem, spinal cord, kidney, liver, heart, skeletal muscle, white adipose tissue, stomach, small intestine, large intestine, testis and placenta obtained from healthy C57BL/6 mice. Mitochondrial purity was assessed by western blots against selected mitochondrial and non-mitochondrial proteins, and intactness was verified by polarographic studies (data not shown) and electron microscopy (Figure 2A, S2). Each sample was separated by SDS-PAGE and then sectioned into 20 bands that were each analyzed by high performance, liquid chromatography tandem mass spectrometry (LC-MS/MS) using an LTQ Orbitrap Hybrid MS system. We captured 4.7 million tandem mass spectra and searched them against the mouse RefSeq protein database using stringent matching criteria, resulting in the confident identification of products from 3,881 genes (Table S3). The detected proteins are not biased by molecular weight, isoelectric point, or presence of transmembrane helices, but do show a slight bias against proteins whose transcripts exhibit low abundance (Figure S4). We estimate that we identify 85% of proteins within each sample (based on technical liver replicates), but we saturate detection of distinct proteins by sampling many tissues (Figure 2B). In total, we identify 88% of previously known mitochondrial proteins, including 93% of OXPHOS proteins.

In the *subtractive* proteomics phase, we applied in-solution LC-MS/MS on both crude and purified mitochondria from 10 of the above tissues. This approach is based on the observation that *bona fide* mitochondrial proteins should become enriched during the purification process, and likewise contaminants should become depleted (e.g., the loss of ER protein calreticulin in Figure 2A). This subtractive method is similar in concept to protein correlation profiling[10]. Of the 2,565 gene products detected in either crude or pure samples, 1,022 were more abundant in crude samples (crude-enriched), 709 more abundant in purified samples (pure-enriched), and the remainder inconclusive (see Experimental Procedures). The crude-enriched set contained many plasma membrane and extracellular proteins (likely as precursors in the ER) whereas the pure-enriched set was almost exclusively mitochondrial, validating that the subtractive proteomics approach can aid in distinguishing genuine mitochondrial proteins from contaminants (Figure 2C).

We next combined the data from the discovery and subtractive phases in order to assign a probability that each protein detected by discovery MS/MS was truly mitochondrial. To do so, we compiled training sets comprised of 591 known mitochondria genes (T_{mito}) and 2519 non-mitochondrial genes ($T_{\sim\mathrm{mito}}$), listed in Table S5. To avoid circularity, our curated T_{mito} list excludes mitochondrial proteins characterized solely by prior proteomic studies. Using our training data, we calculated the likelihood ratio that each protein is genuinely mitochondrial based on its discovery MS/MS protein abundance and its subtractive MS/MS enrichment (see Experimental Procedures and Figure S6). As shown in Figure 2D, the likelihood ratio quantifies the confidence that a protein detected by MS/MS is truly mitochondrial.

Integration of Mass Spectrometry Analysis with Genome-Scale Datasets

Our combination of discovery and subtractive proteomics is extremely powerful for discovering *bona fide* mitochondrial proteins, though this approach alone is not sufficiently sensitive or specific (Figure 3A). For example, these experiments miss proteins that are extremely low in abundance, lack tryptic peptides amenable to MS, or localize to mitochondria only under specific conditions. In order to approach a comprehensive mitochondrial inventory we need to integrate these data with other available information.

We therefore combined our MS/MS results with six complementary computational, homology-based, and experimental techniques to determine likelihood of mitochondrial localization (Figure 3A and Experimental Procedures). Using the Maestro naïve Bayes framework we developed previously[18], we used training data to convert each method's data values into log-likelihood scores of mitochondrial localization (Table S7). Since the seven methods are largely conditionally independent (Figure S8), we sum these individual log-likelihood scores into the combined Maestro score based on an independent probability model. Using Maestro, we systematically rank all mouse genes by their likelihood of mitochondrial localization (Table S5). We can assess accuracy at each score using a corrected false discovery rate statistic (cFDR), which accounts for the sizes of our training sets (see Experimental Procedures). At a Maestro score threshold of 4.56, corresponding to 10% cFDR, there are 951 mitochondrial gene predictions including 498/591 known mitochondrial genes (84% sensitivity). This Bayesian integration avoids overfitting the training data, as shown through 10-fold cross-validation (in rotation, training on 90% of the data and reserving 10% for testing) that achieves comparable 82% sensitivity at the same cFDR. As seen in Figure 3A, integration greatly increases prediction accuracy.

Large-Scale GFP-Microscopy of Mitochondrial Localization

We additionally undertook a large-scale microscopy study as a complementary experimental approach to confirm mitochondrial localization (Figure 1). We tested the human orthologs of our mouse predictions due to the availability of high quality clones from the human hORFeome v3.1 collection[29]. We created C-terminus GFP-fusion constructs and visualized subcellular localization in HeLa cells by fluorescence microscopy. This method showed clear mitochondrial localization of 12/21 positive controls and none of 18 negative controls, indicating that this technique is specific but has limited sensitivity. We then tested 470 genes that lacked prior experimental support of mitochondrial localization. These candidates were selected from an interim Maestro analysis and have an estimated 59% cFDR based on our final Bayesian analysis. Of the 404 candidates successfully transfected, we identified 131 genes with clear mitochondrial localization (representatives shown in Figure 3B and the complete set available at www.broad.mit.edu/publications/MitoCarta). The success rate of this approach matches our estimated cFDR and sensitivity rates – thus validating our Bayesian integration. The 273 constructs without clear mitochondrial localization were less informative since it is possible that the GFP tag interfered with mitochondrial import, the wrong splice form was tested, or HeLa cells lacked necessary chaperones/modifiers.

MitoCarta: an Inventory of 1098 Genes Encoding the Mitochondrial Proteome and their Protein Expression across 14 Tissues

Combining our discovery and subtractive proteomics with computation, microscopy and previous literature, we defined a high-confidence mitochondrial compendium of 1098 genes, termed MitoCarta (Figure 1). This inventory is estimated to be over 85% complete and contain ~10% false positives (see Supplemental Data). It contains 356 genes without previous mitochondrial annotation in Gene Ontology (GO) or MitoP2[30] databases, and distinguishes itself from other catalogs by providing strong experimental support for 87% of genes based on: mass spectrometry (70%), GFP studies (12%), and/or literature curation (54%). We conservatively estimate that at least 85 of the MitoCarta proteins are also resident in other cellular locations, based on crossing MitoCarta with two organelle-based proteomic surveys shown in Table S9[7,10].

The MitoCarta collection includes some notable components and highlights important regulatory features for the organelle. For example, the inventory includes several kinases, phosphatases, RNA-binding proteins and disease-related proteins (*MMACHC*, *ATIC*) not previously associated with the mitochondrion (Table S5B). Interestingly, as a collection the MitoCarta genes have significantly shorter UTRs and coding regions, and are more highly expressed, compared to all mouse genes (Figure S10). Their promoters tend to have CpG islands and lack TATA boxes, a feature shared with other "housekeeping" genes that may account for their higher expression[31]. Additionally MitoCarta promoters are enriched for the presence of eight conserved sequence motifs, including five known mitochondrial transcription factor binding sites and three novel elements (Figure S10).

In addition to expanding the number of known mitochondrial proteins, our inventory provides the opportunity to assess differences in mitochondrial protein expression across tissues (Figure 4A). We assessed the relative abundance of each MitoCarta protein across our 14 tissues using MS total peak intensity (see Experimental Procedures). This metric is highly reproducible across technical replicates (Figure S11) and correlates quite well with mRNA expression (see Supplemental Data). However, as our atlas contains only a single replicate per tissue, we note two caveats: first, it cannot be used to assess statistically significant differences in abundance across tissues; and second, due to stochastic sampling we estimate that we detect approximately 90% of proteins present in each tissue.

We utilize this protein atlas to investigate the differences in mitochondrial pathways between tissues. We find that approximately 1/3 of MitoCarta genes are core mitochondrial components present across all sampled tissues, including most OXPHOS subunits and the TCA cycle (Figure 4B). However, most MitoCarta genes show some degree of tissue specificity (Figure 4A). Interestingly, these include much of the mitochondrial ribosome and half the subunits of complex IV, several of which have previous verification of tissue-specific expression[32]. Additionally, the enzymes of the ketogenesis and urea cycle pathways are expressed in a broader set of tissues than expected, including brain and placenta (Figure S12). Typically, we find that mitochondria express an average of ~760 unique gene products per tissue (range 554-797, Figure 4C), with pairs of tissues typically sharing ~75% of proteins (range 63-88%). Moreover, using a cytochrome *c* ELISA, we estimate that mitochondrial *quantity* varies by a remarkable 30-fold amongst a panel of 19 tissues (Figure 4D). Together these analyses reveal the tissue diversity of mitochondrial quantity and composition, and demonstrate how our compendium can serve as a resource for future investigations into tissue-specific mitochondrial biology.

Identifying Complex I Associated Proteins Through Phylogenetic Profiling

The expanded mitochondrial compendium also provides an opportunity to discover novel components for pathways resident in the organelle. Nearly 300 genes—26% of our inventory—have no association with a GO biological process. To associate a subset of these with known pathways, we perform phylogenetic profiling, which uses shared evolutionary history to identify functionally related proteins[28]. This approach is likely to be particularly applicable to the mitochondrion, given its unique evolutionary history of descending from a Rickettsia-like endosymbiont early in eukaryotic evolution[33].

To explore the utility of phylogenetic profiling for mitochondria, we first identified homologs of mouse MitoCarta proteins in 500 fully sequenced species (Figure 5A, Table S13). We find that 75% of present-day mitochondrial components have clear bacterial ancestry (BlastP expect < 1e-3) and that 57% have bacterial best-bidirectional orthologs, which is more than three-fold higher than that of all mouse proteins (Figure 5C). The phylogenetic profiles confirm that functionally-related mitochondrial proteins tend to have similar evolutionary histories. For example, most proteins involved in fatty acid metabolism, the citric acid cycle, and folate metabolism have ancient origins (Figure 5B). Conversely, the mitochondrial protein import machinery and mitochondrial carriers are more recent innovations (Figure 5B). Thus, it may be possible to use shared evolutionary history to associate unannotated MitoCarta proteins with known pathways.

We focused this strategy on identifying factors essential to respiratory chain complex I (CI) because of its prominent role in energy metabolism and disease. Currently, there are only three known assembly factors for this large, macromolecular complex, though clinical data suggest that there are many unidentified factors needed for its assembly and activity[22]. These factors likely reside in the mitochondrion, and thus our MitoCarta compendium aids in prioritizing candidates. Additionally, the evolutionary history of CI across 5 yeast species has recently been proven useful in identifying the assembly factor NDUFA12L, supporting this phylogenetic approach[25,34].

In order to establish a broader phylogenetic profile for CI, we first built a rooted phylogenetic tree of 42 eukaryotes (Figure 6C, Experimental Procedures). This tree is robust to different phylogenetic reconstruction methods, except for some positioning uncertainty of three deep branching protist species (see Supplemental Data). We observed that a set of 15 CI proteins are not only absent from several yeast species, but are ancestral bacterial subunits that have been independently lost at least four times in eukaryotic evolution (Figure 6A, Table S14). It is probable that the species that lost CI also lost the proteins required for its assembly and function. Only 19 other MitoCarta proteins share this profile and now represent strong candidates for functional association with CI (Figure 6B). These 19 MitoCarta proteins, termed COPP (Complex One Phylogenetic Profile), as well as an expanded set with weaker phylogenetic signatures, are listed in Table S14. The COPP set includes two well-studied proteins involved in branched chain amino acid degradation (*Ivd*, *Mccc2*), and four proteins involved in lipid breakdown (*Dci*, *Phyh*, *Amacr*, *AF397014*), which raises the intriguing hypothesis of an association between these pathways and complex I activity.

We tested four of our COPP genes for an involvement in CI activity by creating stable knockdowns in human fibroblasts using lentiviral-mediated RNAi[35]. Given that we are interested in the clinical relevance of these predictions, we chose to test the human orthologs of our mouse candidates. We achieved ≥ 80% knockdown of 3 COPP genes and 50% knockdown of the fourth, as measured by quantitative real-time PCR (Figure 6E). We next assessed both CI abundance, using immunoblots against a CI subunit, and

CI activity, using immunocapture-based activity assays (see Experimental Procedures and Figure S15). Knockdown of *C8orf38* showed the strongest reduction of both CI abundance and activity, comparable to the known CI assembly factor NDUFAF1 (Figure 6D-F). These data strongly suggest that *C8orf38,* which previously had no prior association to any biological process or subcellular location, is crucial for activity and/or assembly of endogenous CI. The other three candidate knockdown lines showed 20-40% reduction of CI activity (Figure 6F) with variable effects on CI abundance (Figure 6D). The moderate reduction of CI activity does not offer definitive evidence of association with CI, however we note that the CI activity assay measures only the NADH dehydrogenase activity, which may still be largely intact even if other modules of CI are improperly assembled. Thus we experimentally validate the importance of a one COPP gene, show suggestive evidence for three other COPP genes, and prioritize more than one dozen additional proteins for future studies of complex I.

A Mutation in *C8orf38* Causes an Inherited Complex I Deficiency in Humans

The 19 MitoCarta COPP genes identified above represent strong candidates for genes underlying clinical CI deficiency. We used these candidates in combination with homozygosity mapping to search for a causative gene mutation in two siblings (female and male) with severe isolated CI deficiency, born to first cousin Lebanese parents (Figure 7A). The siblings presented at 10 and 7 months, respectively, with focal right hand seizures, decreased movement and strength, ataxia and evolving rigidity. Both had persistent lactic acidosis and neuroimaging was consistent with Leigh syndrome. The affected girl had isolated CI deficiency in muscle, liver and fibroblasts with normal or elevated activities of other complexes and citrate synthase (Figure 7B). She died at 34 months of age from a cardiorespiratory arrest following admission to hospital with pneumonia. The affected boy had an isolated CI defect confirmed in fibroblasts and is currently 22 months of age.

Since the underlying molecular defect is likely a recessive mutation, we performed homozygosity mapping on DNA isolated from the five family members and identified eight chromosomal regions of homozygosity shared only by the affected siblings (Figure 7C and Experimental Procedures). Collectively, these regions contain 857 genes, including 4 CI structural subunits and one COPP gene: *C8orf38* (Figure 7C). Sequencing of two CI structural subunit genes showed no mutations, however sequencing of *C8orf38* (NM_152416) revealed a c.296A>G mutation in exon 2 that segregated with the disease in the family (Figure 7D). This mutation causes a predicted Gln99Arg substitution in a residue fully conserved across vertebrates, and may also cause a splicing defect due to its position at the 3' end of exon 2 (Figure 7D). This mutation was not present in EST databases, SNP databases, or in 100 Lebanese chromosomes tested. The localization of C8orf38 to the mitochondrion, its RNAi phenotype of CI deficiency (Figure 6F), and the segregating *C8orf38* mutation at a highly conserved residue together strongly establish that *C8orf38* is a human CI disease gene.

Conclusion

We have constructed a high quality compendium of mitochondrial proteins, used comparative genomics to predict roles for unannotated proteins in CI biology, and validated these predictions using cellular models and human genetics. Our inventory of 1098 mitochondrial genes and their protein expression across 14 tissues represents the most comprehensive characterization of the organelle to date and provides a framework for addressing major questions in mitochondrial biology.

We leveraged our compendium to discover proteins essential for proper complex I activity. Despite CI's critical importance in energy production and broad role in rare and common human disease, many aspects of its structure, assembly and activity are poorly understood. Through phylogenetic profiling, we identified 19 additional genes likely to be associated with CI, most notably *C8orf38*, which we further show is mutated in an inherited CI deficiency. *C8orf38* was first shown to be mitochondrial in this study and was not previously associated with any biological function. The domain structure of C8orf38 suggests involvement in phytoene metabolism, potentially implicating it in branched chain lipid metabolism along with other COPP proteins Phyh, Amacr, and AF397014. The remaining COPP genes are now prime candidates for other CI deficiencies, and may help unravel the assembly and maturation program for CI.

In addition to fueling the discoveries we present here, the MitoCarta inventory can be used immediately in other disease related projects. As we have demonstrated in the current report, the mitochondrial compendium can help highlight specific candidates within linkage regions of any Mendelian mitochondrial disease. MitoCarta can also help elucidate the pathogenesis of common degenerative diseases, which have recently been associated with declining mitochondrial gene expression and rising reactive oxygen species production[36-38]. Importantly, MitoCarta can also serve as a foundation for basic mitochondrial biology. The orchestrated transcription, translation, and assembly of the mitochondrial components, encoded by two genomes, into functioning, tissue-specific organelles is a remarkable feat about which much remains unknown. Our protein compendium provides a framework with which these tissue-specific programs can be deciphered.

Experimental Procedures

Protein mass spectrometry

Discovery phase: Mitochondria were isolated from C57BL/6 mouse tissues by Percoll density gradient purification (see Supplemental Data for complete details), and assessed for purity with antibodies against Calreticulin (Calbiochem), VDAC1 (Abcam), and an 8

KDa CI subunit (Mitosciences). To further demonstrate purity, a more extensive set of organelle marker antibodies were used for a subset of the mitochondrial preparations (Figure S2). Each sample was size separated by 4-12% bis-Tris gradient SDS-PAGE, separated into 20 gel slices and then reduced, alkylated, and subjected to in-gel tryptic digestion. Extracted peptides from each slice were analyzed by reversed-phase LC-MS/MS using an LTQ-Orbitrap (Thermo Scientific). Data dependent MS/MS were collected in the LTQ for the top ten most intense ions observed in the Orbitrap survey scan, using dynamic exclusion to exclude re-sampling peaks recently selected for tandem MS/MS (within 60s intervals). MS/MS spectra were filtered for spectral quality, pooled from all 14 tissues, and searched against the RefSeq mouse protein database using the Spectrum Mill MS Proteomics Workbench. We required proteins to have ≥2 unique peptides detected, with at least one peptide that distinguished the matching gene from all other mouse Entrez genes. Data were aggregated at the gene level, using the highest MS values for any splice form. Abundance was measured by coverage (percent of amino acids with MS evidence) for cross-protein comparisons, and by total peak intensity (the sum of MS peak areas for all sequence identified peptides matching a protein) for cross-tissue comparisons.

Subtractive phase: Matched crude and highly purified mitochondria were collected from 10 tissues. Sample proteins were reduced, alkylated and then digested with trypsin in-solution. MS/MS spectra were obtained and searched as above, but proteins required only ≥1 peptide spectra, since these results affected only proteins detected via discovery MS/MS. Proteins found only in crude extracts, or found at ≥ twofold higher peak intensity in crude extracts compared to pure were considered crude-enriched (and similarly for pure-enriched).

Data combination: Proteins were assigned integrated MS/MS scores using the likelihood ratio $L(d,s)=P(d,s|T_{mito})/P(d,s|T_{\sim mito})$ where d is the discovery MS/MS abundance level (coverage), s is the subtractive MS/MS enrichment category, and T_{mito} and $T_{\sim mito}$ are training sets.

See Supplemental Data for complete details.

Mouse and human datasets

Mouse RefSeq Release 20 proteins were mapped to 23,640 NCBI Entrez gene identifiers (ftp.ncbi.nih.gov/gene/DATA/, 12/12/2006), excluding proteins mapped to non-reference assemblies or to pseudogenes (Entrez annotation, 6/21/07). Human-mouse orthologs were obtained from Homologene (ftp.ncbi.nih.gov/pub/HomoloGene, 1/26/2007). Training sets (Table S5) included T_{mito}: 591 genes with mitochondrial annotations from MitoP2 or Gene Ontology (GO) databases, that were manually curated for experimental evidence of mitochondrial localization in mammals, excluding genes with support solely from large scale proteomics surveys; $T_{\sim\text{mito}}$: all 2519 genes with GO subcellular localization annotations (type "inferred by direct assay"), excluding mitochondrial and uninformative categories[18]. Protein domains from Pfam (ftp.sanger.ac.uk/pub/databases/Pfam, 11/22/2006) were identified using HMMER (expect parameter=0.1, trusted threshold cutoffs).

Integration of genome-scale data sets

Seven methods for determining mitochondrial localization were integrated using the Maestro naïve Bayes classifier[18]. Training sets (T_{mito} and $T_{\sim mito}$) were used to convert each of the individual feature scores ($s_1..s_7$) into a log-likelihood ratio, defined as $\log_2[P(s_1..s_7| \; T_{mito}) \; / \; P(s_1..s_7| \; T_{\sim\text{mito}})]$. For transcript or protein level scores, the gene inherited the highest score of any splice form. The scores for the seven genomic features were calculated at predefined ranges (see Table S7) as follows (see Supplemental Data for details):

Proteomics: one of 12 categories shown in Figure 2D, or NA if not detected

Targeting sequence: TargetP v1.1 confidence score[16]

Protein domain: categorical score (M+, M-, M±, NA) representing presence of a protein domain that is exclusively mitochondrial, exclusively non-mitochondrial, ambiguous, or not present in any annotated SwissProt eukaryotic protein.

Yeast homology: 1 if the best *S. cerevisiae* homolog (BlastP expect < 1e-3, coverage >50% of longer gene) is mitochondrial (Saccharomyces Genome Database, 12/27/06), 0 otherwise

Ancestry: BlastP expect value from *R. prowazekii* homolog, or NA if expect > 1e-3

Coexpression: N50 score (number of T_{mito} genes found within the gene's 50 nearest transcriptional co-expression neighbors) within the GNF1M atlas of 61 mouse tissues[39]

Induction: fold-change of mRNA expression in cellular models of mitochondrial proliferation (overexpression of PGC-1α in mouse myotubes) compared to controls[18,40]

The corrected false discovery rate was used to assess accuracy of predictions since the sizes of the training sets T_{mito} and $T_{\sim\text{mito}}$ do not match our prior expectation of the proportion of mitochondrial to non-mitochondrial cellular proteins[18]. We define cFDR = (1 – SP) / (1 – SP + SN x O_{prior}), where TP, TN, FP, FN represent true/false positives and negatives, specificity SP = TN / (TN + FP), sensitivity SN = TP / (TP + FN), and O_{prior} = 1500/21000.

To compare performance of each method (Figure 3A), we chose the following thresholds: MS/MS pure-enriched, or inconclusive with coverage > 25%; TargetP ≥ 1; Induction ≥ 1.5; Domain M+; Coexpression ≥ 5; Yeast Homology 1; Ancestry ≤ 1e-3; Maestro ≥ 4.56.

Epitope tagging with GFP and microscopy.

cDNAs from the Human Orfeome collection[29] were cloned into the C-terminal GFP vector pcDNA6.2/C-EmGFP-DEST (Invitrogen). Approximately 4 X10^3 HeLa cells were seeded in 100 μL of medium (DMEM with 10% FBS, 1x GPS) in 96-well imaging plates (Falcon) 24 h before transfection using Lipofectamine LTX (Invitrogen). 48 h post transfection, cells were stained with medium containing 50 nM MitoTracker Red CMXRos and 1:1000 diluted Hoechst 33258 (Molecular Probes), washed, fixed, and imaged (see Supplemental Data). Mitochondrial localization was determined by overlap of GFP and MitoTracker signals.

Cytochrome c ELISA assays

Fresh mouse tissues were prepared in ice-cold PBS (see Supplemental Data). Following homogenization, tissue lysates were resuspended in PBS containing 0.5% Triton X-100 detergent and protease inhibitors (Roche) and spun at maximum speed in a table top centrifuge set to 4°C for 30 minutes. Supernatant was drawn off, flash frozen in liquid nitrogen and stored at -80°C until use. Cytochrome *c* levels were measured in duplicate using an ELISA kit (Quantikine) following the manufacturer's protocol.

Phylogenetic profiling

Homologs of mouse proteins within 500 fully sequenced species (Table S13) were defined by BlastP expect < 1e-3. Mouse genes with ≤1 bacterial homologs were called "eukaryotic innovations". We built a rooted phylogenetic tree of 42 eukaryotic species and a bacterial outgroup (*E. coli*) using PhyML[41] (JTT matrix, 4 substitution rate categories) based on ClustalW multiple alignments of 6 well-conserved mouse proteins (Rps16, Ak2, Drg1, Dpm1, Cct7, Psmc3) that were concatenated and manually edited to remove regions of poor alignment. COPP genes were identified using the following profile: absent in 11 species (*S. pombe, A. gossypii, C. glabrata, S. cerevisiae, C. hominis, C. parvum, P. falciparum 3D7, T. annulata, T. parva, G. lamblia, E. cuniculi),* present in a bacterial genome, present in ≥ 1 plant-like species (*A. thaliana, O. sativa, D. discoideum, C. merolae*) and present in ≥ 2 other yeasts (*Y. lipolytica, C. albicans, P. stipitis, D. hansenii*), where presence was defined by BlastP expect < 1e-3. See Supplemental Data and Table S14 for full details.

Complex I abundance and activity assays

Lentiviral vectors (pLKO.1) encoding short hairpin sequences were obtained from the Broad RNAi Consortium (TRC)[35]. These vectors were transfected with a packaging plasmid (pCMV-dR8.91) and VSV-G envelope plasmid (pMD2.G) into 293T cells using Fugene (Roche) following TRC protocols (www.broad.mit.edu/genome_bio/trc/publicProtocols.html). Virus-containing medium was harvested 24 and 48 hours post transfection. Approximately 30,000 MCH58 human fibroblasts were seeded onto 24-well plates the day prior to infection. To infect cells, 150 µl of virus-containing medium mixed with 350 µl of low antibiotic medium containing 8µg/ml polybrene was added to each well and the plate spun at 2250 rpm for 90 minutes at 37°C. Post spin, medium was replaced with DMEM (5% FBS, 1X GPS) for 12-24 hours and then switched to DMEM with 2 µg/ml puromycin for 1-2 weeks for selection of stably infected cells. RNA was extracted from each cell line (Qiagen RNAeasy) and used for 1st strand cDNA synthesis (Invitrogen). Knockdown efficiency was then assessed using real-time PCR (ABI Taqman Assays) using HPRT as an endogenous control. For immunoblot analysis of CI and actin, 10 µg of cleared whole cell lysate was separated on a 4-12% gel (Invitrogen) and transferred to pvdf membrane. Membranes were probed with antibodies against -actin (Sigma) and an 8kDa CI subunit (Mitosciences). CI activity assays were performed on 15 µg of cell lysate using immunocapture-based assays following the manufacturer's protocol (Mitosciences). Results were scanned using a BioRad GS-800 scanner and analyzed with Quantity One software.

Mitochondrial enzyme assays

Respiratory chain complexes I, II, III and IV plus the mitochondrial marker enzyme citrate synthase were assayed in skeletal muscle and liver homogenates and in enriched fibroblast mitochondrial preparations by spectrophotometric methods as described previously[42,43]. Respiratory chain enzyme assays measured NADH:coenzyme Q1 reductase (CI), succinate:coenzyme Q1 reductase (CII), decylbenzylquinol:cytochrome c reductase (CIII) and cytochrome c oxidase (CIV). Enzyme activities were expressed as a ratio relative to citrate synthase and then as a percentage of normal control mean value.

Homozygosity mapping

DNA from five family members was analyzed using Affymetrix GeneChip Mapping 50K XbaI SNP arrays. Loss of heterozygosity regions were detected using Affymetrix software (GDAS v.3.0.2.8, CNAT v.2.0.0.9 and IGB v.4.56).

Acknowledgements

We thank J. Jaffe, K. Clauser, P. Matsudaira, for advice; D. Arlow, S. Silver, O. Goldberger, T. Gilbert, and T. Hirozane-Kishikawa for technical assistance; M. McKee for performing electron microscopy; E. A. Shoubridge for providing MCH58 cell lines;

and A. Ting, D. Altshuler, and J. Hirschhorn for comments on the manuscript. Electron microscopy was performed in the Microscopy Core of the Center for Systems Biology, which is supported by an Inflammatory Bowel Disease Grant DK43351 and a Boston Area Diabetes and Endocrinology Research Center Award DK57521.This work was supported by a Principal Research Fellowship from the Australian NHMRC awarded to D.R.T., and a Burroughs Wellcome Fund Career Award in the Biomedical Sciences, an Early Career Award from the Howard Hughes Medical Institute, a Charles E. Culpeper Scholarship in Medical Science, and a grant from the National Institutes of Health (GM077465) awarded to V.K.M.

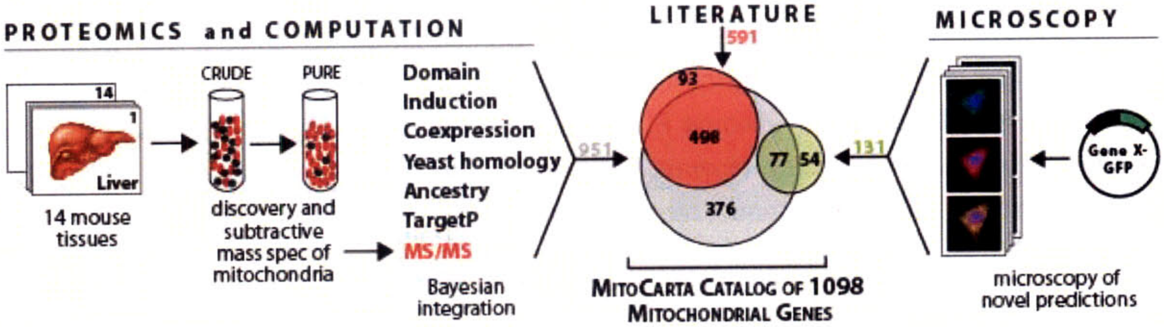


Figure 1: Building a Compendium of Mitochondrial Proteins

MitoCarta is a compendium of 1098 genes encoding proteins with strong support of mitochondrial localization. Each protein was determined to be mitochondrial by one or more of the following approaches: 1) an integrated analysis of seven genome-scale data sets, including in-depth proteomics of isolated mitochondria from 14 mouse tissues (gray circle), 2) large-scale GFP-tagging/microscopy (green circle), and 3) prior experimental support from focused studies (red circle). The union of genes from each approach comprises the MitoCarta compendium.

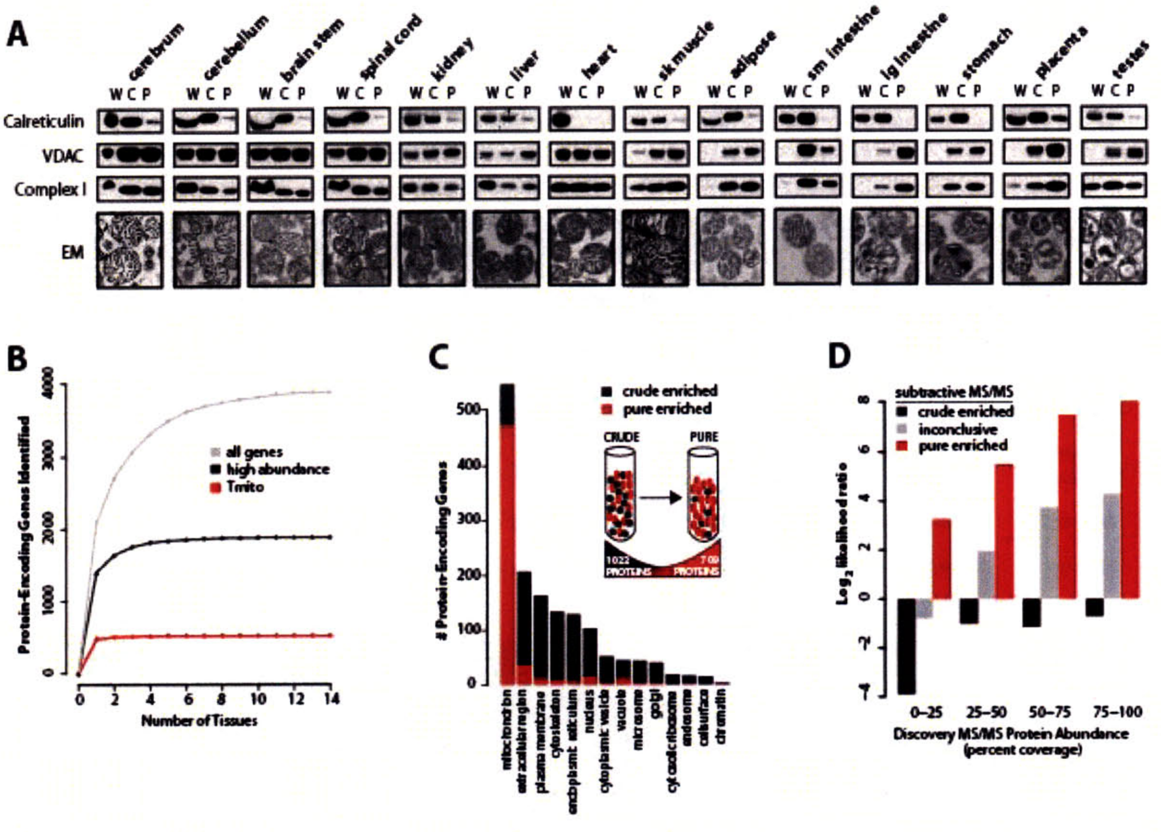


Figure 2: Discovery Proteomics and Subtractive Proteomics of Isolated Mitochondria

(A) Purification of mitochondria from 14 mouse tissues. Mitochondrial enrichment was tracked by the ratio of an ER protein (calreticulin) to mitochondrial proteins (VDAC and CI 8kDa subunit) at three stages of isolation (W, whole tissue lysate; C, crude mitochondrial extracts; P, purified mitochondrial extracts). Electron micrographs show intactness of the purified organelles.

(B) Saturation of protein identifications by discovery MS/MS is plotted for previously known mitochondrial proteins (T_{mito}), abundant proteins (>25% coverage), and all proteins.

(C) Gene Ontology annotations of proteins enriched in pure (red) or crude (black) mitochondrial samples based on subtractive MS/MS experiments. Inset: schematic overview of subtractive MS/MS method.

(D) Likelihood ratio of a protein being truly mitochondrial based on detection in discovery and subtractive MS/MS experiments.

A

Method	Description	# Genes Predicted	Sensitivity	cFDR	Ict1	Rmnd1	Lyrm2	2310016M24Rik	BC051227	5730469M10Rik	Comtd1	Fastk	Mmachc	Atic
MS/MS	Abundance and enrichment in our proteomics survey	1223	79%	36%	3.7	5.5	1.9	4.3	-0.8	7.5	7.5	-2.6	-2.6	-0.8
TargetP	N-terminal mitochondrial targeting signal	2896	64%	61%	9.0	4.8	9.0	-1.4	2.8	-1.4	-1.4	1.7	-1.4	-1.4
Induction	mRNA upregulation during mitochondrial proliferation	1760	48%	69%	2.0	2.4	-0.9	2.4	-0.2	-0.9	2.0	2.4	-0.2	-0.9
Domain	Mitochondrial-specific Pfam protein domain	645	47%	20%	5.8	-4.0	5.8	0.3	0.8	0.3	0.3	0.3	0.3	5.8
Coexpression	Coexpression with Tmito across 61 mouse tissues	736	41%	41%	-1.0	5.4	-1.0	8.8	2.8	2.8	-1.0	5.4	8.8	0.2
Yeast Homology	*S. cerevisiae* mitochondrial homolog	615	38%	38%	-0.7	-0.7	-0.7	-0.7	4.5	-0.7	-0.7	-0.7	-0.7	-0.7
Ancestry	Rickettsial homolog	1389	34%	73%	-0.5	2.4	-0.5	-0.5	-0.5	-0.5	-0.5	-0.5	-0.5	-0.5
Maestro	**Naive Bayes Integration**	**951**	**84%**	**10%**	18.4	15.9	13.6	13.2	9.4	7.0	6.2	6.0	3.7	1.7

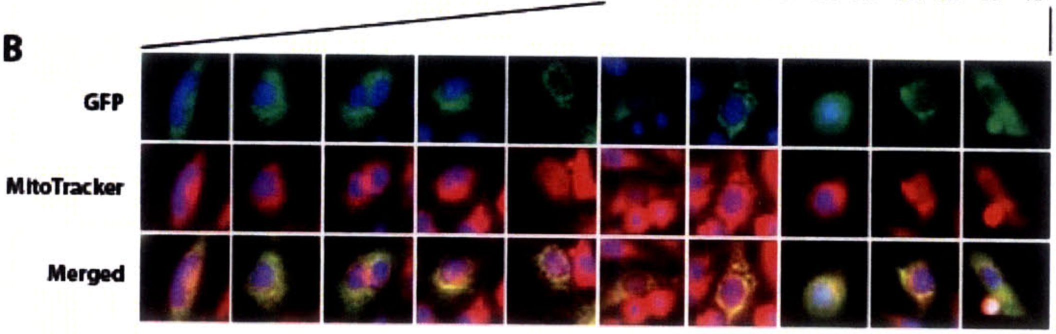


Figure 3: Data Integration and Validation by Microscopy

(A) Eight genome-wide methods for predicting mitochondrial localization, with sensitivity and corrected false discovery rates (cFDR) calculated from large training sets at predefined thresholds (Experimental Procedures). Rightmost columns show each method's log-likelihood score for a selection of mouse genes, which are summed to produce the Maestro log-likelihood of mitochondrial localization.

(B) Fluorescence microscopy images of 10 GFP-fusion constructs with clear mitochondrial localization, corresponding to examples in panel A. Images for all 131 constructs showing mitochondrial localization are available at www.broad.mit.edu/publications/MitoCarta.

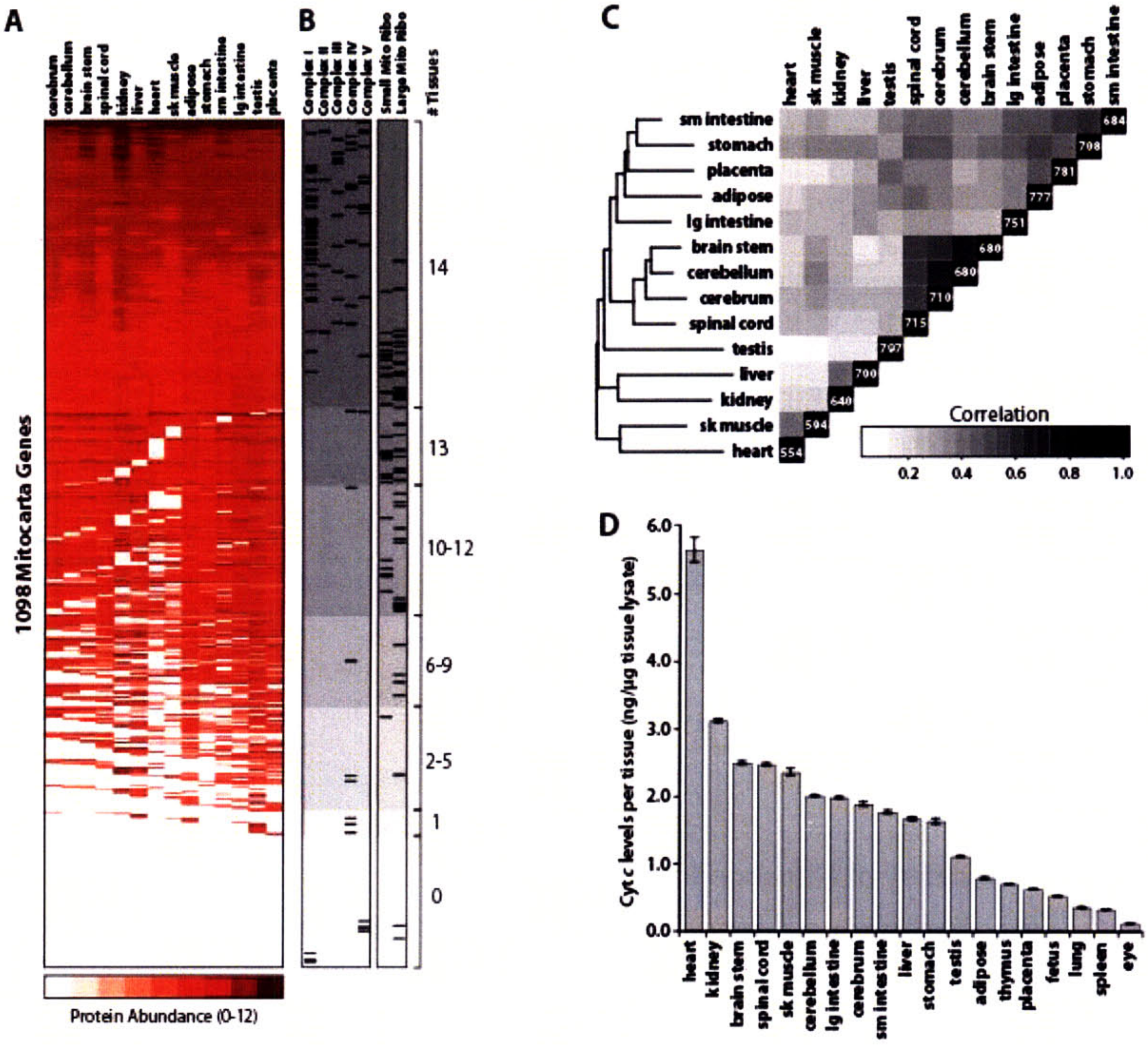


Figure 4: Mitochondrial Protein Expression Across 14 Mouse Tissues

(A) Heatmap of protein abundance, measured by $\log_{10}$ (total MS peak intensity), for 1098 MitoCarta genes across 14 tissues. Genes are ordered by number of tissues and total intensity. White background indicates genes whose protein product was not detected by MS/MS, but are mitochondrial based on prior annotation, computation, or microscopy.

(B) Tissue-distribution of proteins within selected pathways. Tick marks indicate locations of corresponding proteins within (A), and gray shading indicates the total number of tissues in which the protein was detected (0-14).

(C) Correlation matrix of MitoCarta proteins detected by MS/MS in each tissue, clustered hierarchically. Counts on diagonal indicate number of MitoCarta proteins identified by MS/MS.

(D) Mitochondrial quantity per tissue, assessed by ELISA measurements of cytochrome *c* from whole tissue lysates.

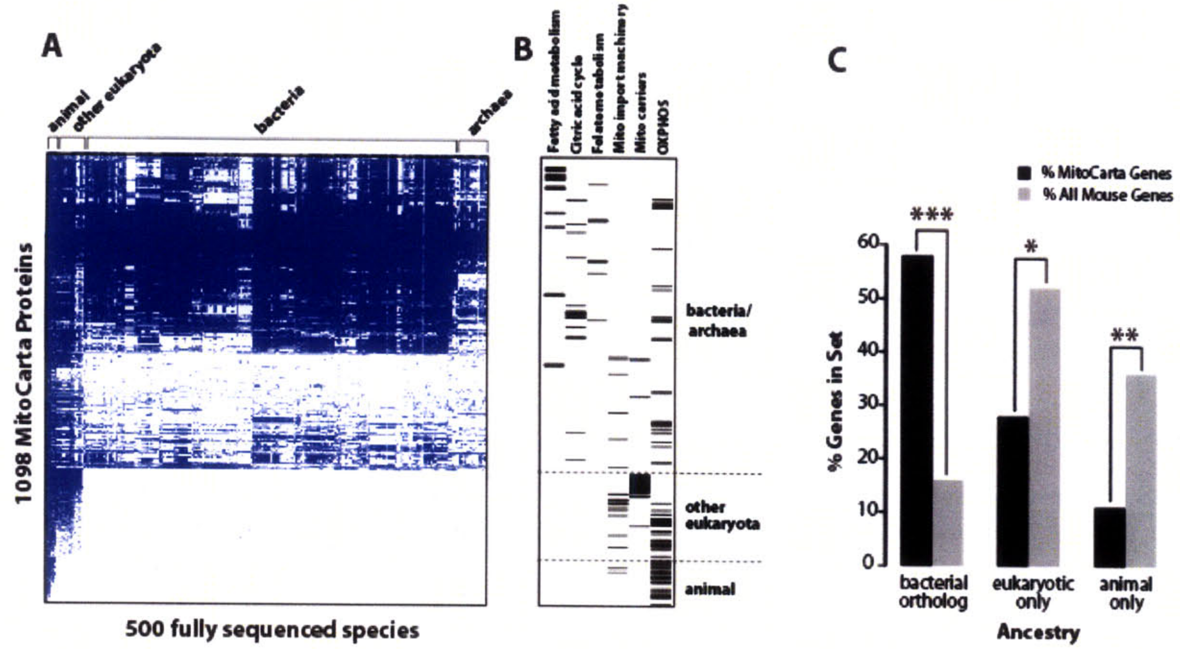


Figure 5: Ancestry of Mitochondrial Proteins

(A) Presence/absence matrix for the 1098 MitoCarta proteins across 500 fully sequenced organisms. Blue squares indicate homology of the mouse protein (row) to a protein within a target species (column).

(B) Ancestry of MitoCarta proteins from selected groups. Tick marks indicate location of proteins within (A).

(C) Comparison of MitoCarta protein ancestry to all mouse proteins, considering only best-bidirectional hits. P values based on hypergeometric distribution with Bonferroni multiple hypothesis correction: $^*p = 6e^{-64}$, $^{**}p = 4e^{-78}$, $^{***}p = 2e^{-232}$.

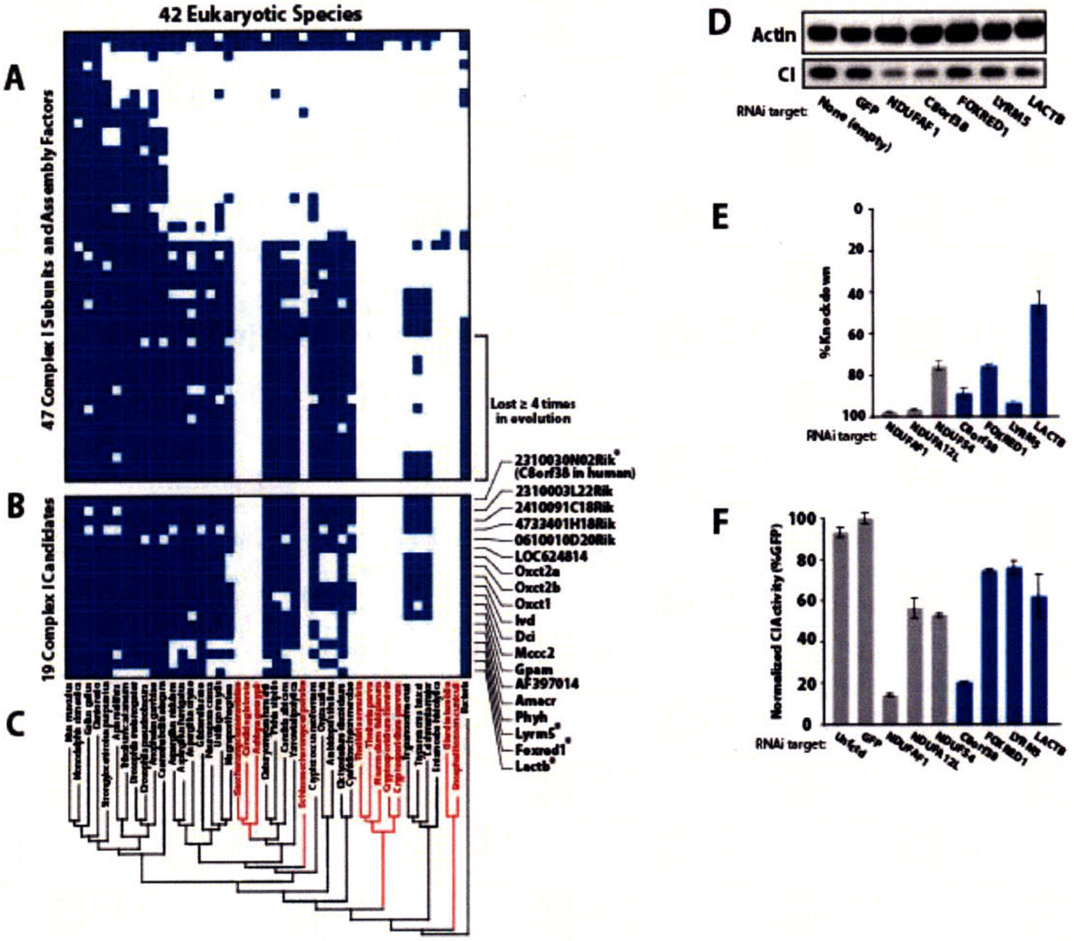


Figure 6: Identification of Complex I Associated Proteins through Phylogenetic Profiling

(A) Presence/absence matrix for 44 respiratory chain CI subunits and 3 assembly factors across 42 eukaryotic species. Blue squares indicate homology of the mouse protein (row) to a protein in a target species (column).

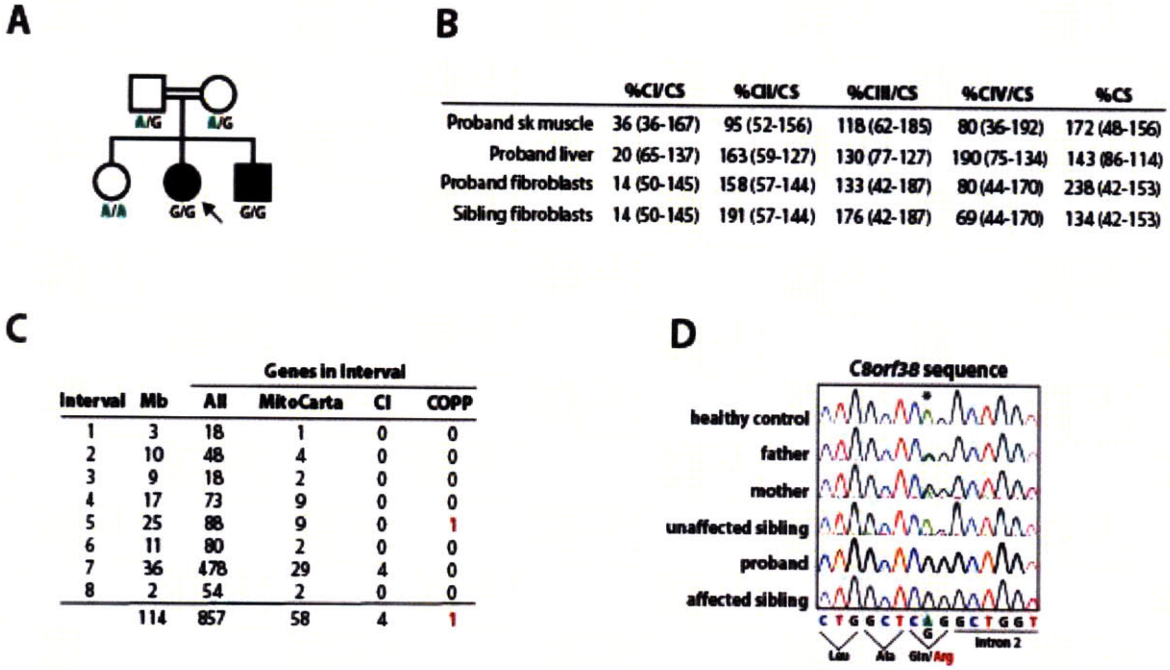
(B) MitoCarta proteins matching the phylogenetic profile of the subset of CI subunits lost independently at least four times in evolution. Asterisks indicate candidates tested by RNAi in (D-F).

(C) Reconstructed phylogenetic eukaryotic tree, with red text indicating species that have lost CI.

(D) Effect of candidate knockdown on CI levels in human fibroblasts. Immunoblots of actin and a CI subunit from whole cell lysates were performed following lentiviral-mediated delivery of an empty vector or hairpins targeted against GFP (negative control), NDUFAF1 (known CI assembly factor) and four CI candidates.

(E) Percent knockdown of mRNA expression achieved for controls (gray bars) or CI candidates (blue bars) as measured by real-time qPCR.

(F) CI activity assays from fibroblast lysates (as in D) for controls (gray bars) and four candidates (blue bars). Error bars represent the range of duplicate assays.



	%CI/CS	%CII/CS	%CIII/CS	%CIV/CS	%CS
Proband sk muscle	36 (36-167)	95 (52-156)	118 (62-185)	80 (36-192)	172 (48-156)
Proband liver	20 (65-137)	163 (59-127)	130 (77-127)	190 (75-134)	143 (86-114)
Proband fibroblasts	14 (50-145)	158 (57-144)	133 (42-187)	80 (44-170)	238 (42-153)
Sibling fibroblasts	14 (50-145)	191 (57-144)	176 (42-187)	69 (44-170)	134 (42-153)

		Genes in Interval			
Interval	Mb	All	MitoCarta	CI	COPP
1	3	18	1	0	0
2	10	48	4	0	0
3	9	18	2	0	0
4	17	73	9	0	0
5	25	88	9	0	1
6	11	80	2	0	0
7	36	478	29	4	0
8	2	54	2	0	0
	114	857	58	4	1

Figure 7: Discovery of a *C8orf38* Mutation in an Inherited Complex I Deficiency

(A) Pedigree from a consanguineous Lebanese family with two children affected by Leigh syndrome and complex I deficiency. Letters beneath each family member represent the genotype for a c.296A>G mutation in *C8orf38*. Proband indicated by arrow.

(B) Respiratory chain enzyme activities, standardized against the mitochondrial matrix marker enzyme citrate synthase, expressed as percentages of the mean value (normal ranges in parentheses). The final column lists citrate synthase activities (relative to total protein) as % of normal control mean (see Experimental Procedures).

(C) Results of homozygosity mapping using DNA from family members in (A). Eight intervals of homozygosity shared by the affected siblings but not the parents or unaffected sibling are listed along with the number of genes in various categories for each interval (CI, known complex I genes; COPP, Complex One Phylogenetic Profiling candidates).

(D) Sequence traces of *C8orf38* from each family member in (A) and one healthy control demonstrating homozygosity for a c.296A>G mutation in both affected siblings.

References

1. DiMauro, S. & Schon, E.A. Mitochondrial respiratory-chain diseases. *N Engl J Med* **348**, 2656-68 (2003).
2. Lowell, B.B. & Shulman, G.I. Mitochondrial dysfunction and type 2 diabetes. *Science* **307**, 384-7 (2005).
3. Wallace, D.C. A mitochondrial paradigm of metabolic and degenerative diseases, aging, and cancer: a dawn for evolutionary medicine. *Annu Rev Genet* **39**, 359-407 (2005).
4. Anderson, S. et al. Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457-65 (1981).
5. Lopez, M.F. et al. High-throughput profiling of the mitochondrial proteome using affinity fractionation and automation. *Electrophoresis* **21**, 3427-40 (2000).
6. Mootha, V.K. et al. Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. *Cell* **115**, 629-40 (2003).
7. Kislinger, T. et al. Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling. *Cell* **125**, 173-86 (2006).
8. Johnson, D.T. et al. Tissue heterogeneity of the mammalian mitochondrial proteome. *Am J Physiol Cell Physiol* **292**, C689-97 (2007).
9. Forner, F., Foster, L.J., Campanaro, S., Valle, G. & Mann, M. Quantitative proteomic comparison of rat mitochondria from muscle, heart, and liver. *Mol Cell Proteomics* **5**, 608-19 (2006).
10. Foster, L.J. et al. A mammalian organelle map by protein correlation profiling. *Cell* **125**, 187-99 (2006).
11. Taylor, S.W. et al. Characterization of the human heart mitochondrial proteome. *Nat Biotechnol* **21**, 281-6 (2003).
12. Reinders, J., Zahedi, R.P., Pfanner, N., Meisinger, C. & Sickmann, A. Toward the complete yeast mitochondrial proteome: multidimensional separation techniques for mitochondrial proteomics. *J Proteome Res* **5**, 1543-54 (2006).
13. Sickmann, A. et al. The proteome of Saccharomyces cerevisiae mitochondria. *Proc Natl Acad Sci U S A* **100**, 13207-12 (2003).
14. Kumar, A. et al. Subcellular localization of the yeast proteome. *Genes Dev* **16**, 707-19 (2002).
15. Huh, W.K. et al. Global analysis of protein localization in budding yeast. *Nature* **425**, 686-91 (2003).
16. 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**, 1005-16 (2000).

17. Guda, C., Fahy, E. & Subramaniam, S. MITOPRED: a genome-scale method for prediction of nucleus-encoded mitochondrial proteins. *Bioinformatics* **20**, 1785-94 (2004).

18. Calvo, S. et al. Systematic identification of human mitochondrial disease genes through integrative genomics. *Nat Genet* **38**, 576-82 (2006).

19. Jansen, R. et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science* **302**, 449-53 (2003).

20. Carroll, J. et al. Bovine complex I is a complex of 45 different subunits. *J Biol Chem* **281**, 32724-7 (2006).

21. Schapira, A.H. Mitochondria in the aetiology and pathogenesis of Parkinson's disease. *Lancet Neurol* **7**, 97-109 (2008).

22. Janssen, R.J., Nijtmans, L.G., van den Heuvel, L.P. & Smeitink, J.A. Mitochondrial complex I: structure, function and pathology. *J Inherit Metab Dis* **29**, 499-515 (2006).

23. Devenish, R.J., Prescott, M., Roucou, X. & Nagley, P. Insights into ATP synthase assembly and function through the molecular genetic manipulation of subunits of the yeast mitochondrial enzyme complex. *Biochim Biophys Acta* **1458**, 428-42 (2000).

24. Fontanesi, F., Soto, I.C., Horn, D. & Barrientos, A. Assembly of mitochondrial cytochrome c-oxidase, a complicated and highly regulated cellular process. *Am J Physiol Cell Physiol* **291**, C1129-47 (2006).

25. Ogilvie, I., Kennaway, N.G. & Shoubridge, E.A. A molecular chaperone for mitochondrial complex I assembly is mutated in a progressive encephalopathy. *J Clin Invest* **115**, 2784-92 (2005).

26. Saada, A. et al. C6ORF66 is an assembly factor of mitochondrial complex I. *Am J Hum Genet* **82**, 32-8 (2008).

27. Vogel, R.O. et al. Human mitochondrial complex I assembly is mediated by NDUFAF1. *Febs J* **272**, 5317-26 (2005).

28. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. & Yeates, T.O. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A* **96**, 4285-8 (1999).

29. Lamesch, P. et al. hORFeome v3.1: a resource of human open reading frames representing over 10,000 human genes. *Genomics* **89**, 307-15 (2007).

30. Prokisch, H. et al. MitoP2: the mitochondrial proteome database--now including mouse data. *Nucleic Acids Res* **34**, D705-11 (2006).

31. Carninci, P. et al. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat Genet* **38**, 626-35 (2006).

32. Huttemann, M., Jaradat, S. & Grossman, L.I. Cytochrome c oxidase of mammals contains a testes-specific isoform of subunit VIb--the counterpart to testes-specific cytochrome c? *Mol Reprod Dev* **66**, 8-16 (2003).

33. Andersson, S.G. et al. The genome sequence of Rickettsia prowazekii and the origin of mitochondria. *Nature* **396**, 133-40 (1998).

34. Gabaldon, T., Rainey, D. & Huynen, M.A. Tracing the evolution of a large protein complex in the eukaryotes, NADH:ubiquinone oxidoreductase (Complex I). *J Mol Biol* **348**, 857-70 (2005).

35. Root, D.E., Hacohen, N., Hahn, W.C., Lander, E.S. & Sabatini, D.M. Genome-scale loss-of-function screening with a lentiviral RNAi library. *Nat Methods* **3**, 715-9 (2006).

36. Houstis, N., Rosen, E.D. & Lander, E.S. Reactive oxygen species have a causal role in multiple forms of insulin resistance. *Nature* **440**, 944-8 (2006).

37. Mootha, V.K. et al. PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat Genet* **34**, 267-73 (2003).

38. Schon, E.A. & Manfredi, G. Neuronal degeneration and mitochondrial dysfunction. *J Clin Invest* **111**, 303-12 (2003).

39. Su, A.I. et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* **101**, 6062-7 (2004).

40. 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 U S A* **101**, 6570-5 (2004).

41. Guindon, S. & Gascuel, O. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**, 696-704 (2003).

42. Kirby, D.M. et al. Respiratory chain complex I deficiency: an underdiagnosed energy generation disorder. *Neurology* **52**, 1255-64 (1999).

43. Rahman, S. et al. Leigh syndrome: clinical features and biochemical and DNA abnormalities. *Ann Neurol* **39**, 343-51 (1996).

Chapter 4

—

Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans

Sarah E. Calvo, David J. Pagliarini, and Vamsi K. Mootha

This chapter is presented in the context of its contemporary science and currently in press at *PNAS* 2009.

Corresponding Supplementary Material can be found in Appendix C.

Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans

Upstream open reading frames (uORFs) are mRNA elements defined by a start codon in the 5' untranslated region (UTR) that is out-of-frame with the main coding sequence. While uORFs are present in approximately half of human and mouse transcripts, no study has investigated their global impact on protein expression. Here, we report that uORFs correlate with significantly reduced protein expression of the downstream ORF, based on analysis of 11,649 matched mRNA and protein measurements from four published mammalian studies. Using reporter constructs to test 25 selected uORFs, we estimate that uORFs typically reduce protein expression by 30-80%, with a modest impact on mRNA levels. We additionally identify polymorphisms that alter uORF presence in 509 human genes. Finally, we report that five uORF-altering mutations, detected within genes previously linked to human diseases, dramatically silence expression of the downstream protein. Together, our results suggest that uORFs influence the protein expression of thousands of mammalian genes and that variation in these elements can influence human phenotype and disease.

Introduction

The regulation of gene expression is controlled at many levels, including transcription, mRNA processing, protein translation, and protein turnover. Post-transcriptional regulation is often controlled by short sequence elements in the untranslated regions (UTRs) of mRNA. One such 5' UTR element is the upstream open reading frame (uORF) depicted in Fig. 1*A*. Since eukaryotic ribosomes usually load on the 5' cap of mRNA transcripts and scan for the presence of the first AUG start codon, uORFs can disrupt the efficient translation of the downstream coding sequence (1, 2). Previous reports have shown that ribosomes encountering a uORF can either (i) translate the uORF and stall, triggering mRNA decay, (ii) translate the uORF and then, with some probability, reinitiate to translate the downstream ORF or (iii) simply scan through the uORF (2). uORFs have been shown to reduce protein levels in approximately 100 eukaryotic genes (Table S1). Additionally, mutations that introduce or disrupt a uORF have found to cause three human diseases (3-5). In several interesting cases the uORF–derived protein is functional, however in most cases the mere presence of the uORF is sufficient to reduce expression of the downstream ORF (1, 2, 6-8). Previous genomic analyses suggest that uORFs may be widely functional for several reasons: they correlate with lower mRNA expression levels (9), they are less common in 5' UTRs than would be expected by chance (6, 10), they are more conserved than expected when present (6), and several hundred have evidence of translation in yeast (11). However, no study has demonstrated that these elements have a widespread impact on cellular protein levels. Moreover, no study has investigated whether uORF presence varies in the human population. Here, we take advantage of recently-available datasets of protein abundance (12-17) and genetic variation (18, 19) to assess the impact and natural variation of mammalian uORFs.

Results

uORF Prevalence Within Mammalian Transcripts. We define a uORF as formed by a start codon within a 5' UTR, an in-frame stop codon preceding the end of the main coding sequence (CDS), and length at least 9 nucleotides including the stop codon. As shown in Fig. 1*A*, this definition includes uORFs both fully upstream and overlapping the CDS, since both types are predicted to be functional (20). We searched for uORFs within all human and mouse RefSeq transcripts with annotated 5' UTRs exceeding 10nt. Consistent with previous estimates (9, 10), we find that 49% of human and 44% of mouse transcripts contain at least one uORF (Fig. 1*B*). Interestingly, human and mouse uORF start codons (uAUGs) are the most conserved 5' UTR trinucleotide across vertebrate species (Fig. S1), consistent with a widespread functional role.

uORF Impact on Cellular Protein Levels. If uORFs cause widespread reduction in protein expression, as predicted by ribosome scanning models, we would expect uORF-containing transcripts to correlate with lower protein levels when compared to uORF-less transcripts. To test this hypothesis, we analyzed a total of 11,649 matched mRNA and protein abundance measurements from four published studies across a variety of mouse tissues and developmental stages. These included: 2484 genes expressed in liver (12), 722 genes expressed in six stages of lung development (13), 487 mitochondria-localized gene products expressed in 14 tissues (14), and 925 genes expressed in six tissues (15) (see Supporting Information for details). Proteins were detected via tandem mass spectrometry (MS/MS) and abundance was estimated by standard methods using the normalized number (12, 13, 15) or total peak area (14) of matching MS spectra. mRNA abundance in these conditions was measured by microarrays (21, 22). While neither technology provides absolute quantitation, these large-scale datasets can reveal trends across thousands of genes. Since MS/MS technology cannot reliably distinguish splice variants, we analyzed expression at the gene level and considered only those genes whose collective splice variants either all contain, or all lack, uORFs. Consistent with previous reports (23), we observed that the 10% most highly expressed transcripts based on microarray tissue atlases (21) tend to lack uORFs (Fig. S2 and Supporting Information) and therefore we conservatively excluded these genes to avoid overestimating uORF effects.

Despite differences in experimental methodology, all four independent datasets showed a reduced distribution of protein levels for genes containing versus lacking uORFs (Fig. 2*A-D*). Median protein levels were reduced, respectively, by 39% (p=1e-5), 29% (p=0.007), 34% (p=0.008), and 13% (p=0.36) where significance was determined by empirical permutation testing. mRNA levels were reduced to a lesser extent with only the liver dataset (12) showing a statistically significant median reduction (Fig. 2*E* and Fig. S3). Importantly, the ratio of protein to mRNA was significantly reduced for uORF-containing genes in three of four datasets (Fig. 2*E* and Fig. S3), suggesting that uORF presence likely inhibits translation of the main coding sequence. We observed the same trends when we modified the definition of a uORF by altering length and overlap criteria (data not shown), and when we included the 10% most highly expressed genes (Fig. S4). Analysis of two additional MS/MS studies of mouse adipocyte cells (16) and differentiating embryonic stem cells (17) also showed reduced protein levels for uORF-containing genes, although matched mRNA data were not available (Fig. S3). Collectively, these analyses across 3297 mouse genes demonstrated the first large-scale correlation of uORF presence with reduced protein levels.

To determine whether uORFs play a causal role in reducing protein levels, and to more accurately quantify their effect size, we performed a series of experiments on 15 uORF-containing genes using dual-luciferase reporter constructs (see Materials and Methods). Five genes were chosen randomly from the set of all mouse transcripts containing single uORFs and where, for technical ease, 5' UTR length exceeded 100nt (Fig. 3*B, F*). An additional 10 were selected from our mitochondrial study (14) where MS/MS and conservation data suggested functionality (Fig. 3*C, G*). We cloned the 5' UTR of each selected gene upstream of a luciferase reporter (Fig. 3*A*). HEK 293A cells were then transfected with uORF-containing luciferase constructs or control constructs where the uORF's start codon (ATG) was mutated to TTG. After 48 hours, cells were assayed for luciferase transcript levels by quantitative PCR and for luciferase activity by luminometry. These experiments showed that, on average, uORFs cause a 58% decrease in protein levels (Fig. 3*B,C*), and a 5% decrease in transcript levels (Fig. 3*F,G*). All individual protein differences and four mRNA differences were statistically significant (Fig. 3), and all protein/mRNA ratio differences were statistically significant except for gene *Hsdl2* (Table S2). The constructs with randomly selected uORFs showed higher protein levels compared to the uORFs selected with evidence of functionality (p=1e-5 based on t-test). Similar results were obtained using HEK 293T cells (data not shown). Together, the large-scale correlations and validation experiments demonstrate that uORFs cause blunted protein expression of downstream coding sequences.

Influence of uORF Context, Position, and Conservation. We next investigated whether specific uORF properties were associated with stronger translational inhibition. We analyzed uORF length, number, conservation, position relative to the cap, position relative to the CDS, and uAUG context (also called "Kozak sequence") (see Methods). We quantified uORF effects using the Kolmogorov-Smirnov D statistic within the largest dataset (liver), which offered statistical power for these analyses. All tested subsets of uORFs showed reduced protein levels compared to uORF-less genes (p<0.05), although certain properties modified the effect size (Fig. S5). As predicted by Kozak's classic experiments (1, 20, 24-26), increased inhibition correlated with strong versus weak uAUG context (p=0.04), long versus short cap-to-uORF distance (p=4e-4), presence of multiple uORFs in the 5' UTR (p=8e-6), and increased conservation (p=1e-6) (Fig. S5). Surprisingly, we observed no significant difference between uORFs fully upstream versus overlapping the CDS (p=0.9), between uORFs of different proximity to the CDS (p=0.5) or between uORFs of different lengths (p=0.3). These comparisons over hundreds of liver genes indicate that while all types of uORFs can reduce protein expression, four uORF properties are associated with greater inhibition: strong uAUG context, evolutionary conservation, increased distance from the cap, and multiple uORFs in the 5' UTR.

Polymorphic uORFs in Humans. Given that uORFs reduce protein expression, polymorphisms that create or delete uORFs could influence human phenotypes. Therefore we searched for uORF-altering variants within the 12 million single nucleotide polymorphisms (SNPs) in the human dbSNP database (18). We coin the term polymorphic uORF (puORF) to indicate a uORF that is created or deleted by a polymorphism. We identified puORFs in 509 unique genes (Table S3), of which 366 genes had multiple uORFs and 143 genes had single uORFs (Table 1). Using the cellular reporter constructs described above, we tested the functionality of five puORFs. In all cases the constructs with uORFs produced 30-60% less protein than those with the uORF-less SNP variant, with an average 3% decrease in mRNA levels (Fig. 3*D,H*). All individual protein and protein/mRNA reductions were statistically significant (Table S2). The impact of the puORFs was comparable to all other uORFs that were tested experimentally (Fig. 3). Thus naturally occurring uORF-altering polymorphisms are likely to alter cellular expression of the downstream protein.

puORF-Mediated Differences in Factor XII Protein Levels. One of the human uORF-altering SNPs (rs1801020) has previously been associated with differences in circulating plasma levels of clotting factor XII (*FXII*) in five independent studies (27-31) (Fig. 4). This SNP represents a common T/C polymorphism with prevalence of the T allele estimated at 20% in Caucasian and 70% in Asian populations (27-31). Kanaji and colleagues demonstrated that the T allele reduces protein levels, and proposed that the mechanism could be due to disruption of the Kozak consensus sequence or to the introduction of a uORF, though these hypotheses were not tested (30). To experimentally test the uORF hypothesis, we created eight reporter constructs that included all four possible nucleotide variants at the SNP site, three artificial uORF-generating mutations, and one mutation creating an alternate in-frame start site (Fig. 4*A*). All four uORF-containing UTR constructs showed >50% reduction in protein levels (p<2e-6), whereas the four constructs lacking uORFs did not show strong differences in protein levels (Fig. 4*B*). mRNA levels were altered by less than 30% (Table S2). These results strongly suggest that the presence of a puORF is responsible for the observed variation in human factor XII protein levels.

uORF-Altering Mutations Related to Human Disease. In addition to common puORFs, rare mutations that alter uORFs may cause disease, as has been shown for three genes (Table 2). To systematically identify additional cases, we searched the Human Gene Mutation Database (19) for mutations that introduce or eliminate uORFs. We found 11 additional mutations (Table 2), which were detected by re-sequencing known disease-related genes in affected patients (32-42). These uORF-altering mutations were not present in population controls (32-42), and were either the sole mutation detected in the sequenced exons, or were compound heterozygous with a missense/nonsense mutation (Table 2). The patient presentation was consistent with a recessive phenotype in three of the four compound heterozygous cases (37, 38, 42, 43), and was ambiguous in the remaining case (36). To our knowledge, the mechanistic link between the gene mutation and uORFs had not been previously proposed for SRY (32), IRF6 (33) or GCH1 (34).

To assess whether the uORF-altering mutations influenced protein expression, we used luciferase reporter constructs to test patient mutations in five genes (HBB, PRKAR1A, IRF6, SRY and SPINK1). The uORF-altering mutations in these genes reduced luciferase mRNA levels by less than 20% and luciferase activity levels by 70-100% (Fig. 3*E,I*). These effects on protein levels were highly significant (p<2e-12) and were larger than in the other uORFs experimentally tested (p=4e-4). Thus, these uORF-altering mutations cause dramatically reduced protein levels in our reporter assays, suggesting that they may indeed be responsible for the observed disease phenotypes.

Discussion

Our analyses provide the first assessment of the widespread impact of uORFs on mammalian protein expression. Many previous studies of individual genes demonstrated that the presence of uORFs can lead to reduced mRNA stability and protein translation. Here we show that approximately half of human and mouse protein-encoding genes contain uORFs and that uORF presence correlates with reduced protein expression across thousands of mammalian genes in a variety of tissues and conditions (Fig. 2). We quantify uORF effects using mutation experiments on 25 selected 5' UTRs (Fig. 3), which have typical length, context, position, and conservation features (Fig. S6). These experiments indicate that uORFs typically affect mRNA levels by under 30% and reduce protein levels by 30-80%, although complete protein suppression is possible (Fig. 3). While our mutation experiments focused chiefly on 5' UTRs containing single uORFs, our MS/MS data suggest that multiple uORFs lead to greater reduction of protein expression (Fig. S5*E*). Collectively, these data suggest that uORFs cause reduced protein levels of thousands of mammalian genes.

Our data provide insight into the mechanism by which uORFs influence protein expression. Without exception, uORF-containing reporter constructs exhibit more pronounced reduction of protein compared to mRNA levels (Fig 3), in agreement with the trend observed in large-scale datasets (Fig. 2*E*). This suggests that uORFs act primarily by reducing translational efficiency, and more modestly by affecting mRNA levels. Additionally, since uORF effects do not correlate with the distance between the uORF and CDS (Fig. S5*D*), it is likely that CDS translation generally proceeds from ribosomes that scan through the uORF rather than from ribosomes that reinitiate after uORF translation – at least in genes that contain only a single uORF.

Given that uORFs reduce translation, variants that create or delete uORFs are likely to alter cellular protein levels and in some cases may influence phenotype. uORF-altering variants are likely to be widespread, since each human transcript contains on average 28 nucleotides that could be mutated to introduce a uORF. We identified 509 human genes with polymorphic uORFs (puORFs), although more are likely to be identified as genome variation databases expand. Our data suggest that puORFs will typically alter cellular protein levels by 30-80% in cases where the 5' UTR contains a single uORF. When these puORFs cause physiologically relevant changes in protein levels, as we showed for factor XII, they may cause phenotypic variation. Indeed, the factor XII puORF has been associated with several thromboembolic conditions, although the associations are in contention due to small sample sizes (44). We speculate that other puORFs in our collection (Table S3) may also affect phenotype. For instance, the puORF in chemokine receptor CCR5 might mediate susceptibility to HIV-1 infection, as previous studies showed that variants affecting CCR5 expression alter susceptibility to HIV-1 infection and progression of AIDS (45). Similarly, the puORFs in bitter taste receptors TAS2R5 and TAS2R3 might lead to common variation in taste perception, and puORFs in receptors for ACTH, serotonin, and oxytocin may modulate neurohormonal response (Table 1).

In addition to common polymorphisms, rare uORF-altering mutations that alter levels of essential proteins can cause human disease. To date, three such mutations have been reported. First, a hereditary form of thrombocythaemia is caused by a mutation in THPO mRNA that eliminates a uORF through a splicing defect, and thus causes increased translation of thrombopoetin (3). Second, a mutation introducing a uORF into CDKN2A causes a familial predisposition to melanoma (4). Third, disruption of uORF presence and coding sequence in gene HR causes Marie Unna hereditary hypotrichosis (5). Additional uORF-altering mutations detected in patients with 11 diseases have been reported in the literature, although they were not followed up experimentally (Table 2). In each case, the patient mutation was present within a gene known to underlie the disease when disrupted, and was the sole mutation detected or was compound heterozygous with a nonsynonymous variant. Using reporter assays, we tested five patient mutations in genes associated with disease: Gonadal dysgenesis (*SRY*), Van der Woude syndrome (*IRF6*), Carney complex type 1 (*PRKAR1A*), Hereditary pancreatitis (*SPINK1*) and Thalassaemia beta (*HBB*). We found that the uORF-altering patient mutation caused severely reduced protein levels, and in two cases almost no reporter protein was detected (SRY and SPINK1, Fig. 3*E*). In these two cases, the patient mutation created a second uORF within the gene 5' UTR, rather than creating a single uORF. The strong suppression of protein expression by these five patient mutations offers a simple mechanistic basis for their pathogenicity. These cases add to the growing list of uORF-altering mutations linked to disease and highlight the importance of searching for uORF changes in addition to coding changes underlying disease.

In summary, our analyses demonstrate that uORFs have a widespread impact on the expression of human and mouse genes, and that the human genome contains hundreds of polymorphic uORFs. With the routine application of newer generation sequencing technologies, an important challenge will be to link variation in genome sequences to physiology and disease – and puORFs may represent an important class of functional variants that can be readily linked to phenotype. Although the current analyses focused on the constitutive effects of uORFs on steady-state protein levels, an important next step is to determine whether the influence of uORFs is widely regulated by environmental conditions or signaling pathways, as been shown for a handful of examples (2).

Materials and Methods

Human and mouse uORFs. RefSeq transcripts for human (hg18) and mouse (mm9) were obtained from the UCSC Genome Browser Database (46) (5/20/2008), along with 28-vertebrate species alignments (47). Custom perl scripts annotated uORFs and computed features: uORF context ("strong" indicates a -3 purine and +4 guanine relative to uAUG, otherwise "weak"), cap-to-uORF distance (length between mRNA cap and uAUG), uORF length (including start and stop codon), uORF-to-CDS distance (length between uORF stop codon and CDS start codon), uORF number (number of distinct uORFs in a transcript, where uORFs may overlap but not in the same frame), and conservation (number of species with aligned start codons within 28-species alignments). The first four features were analyzed on transcripts containing single uORFs. uORF properties were compared using a Bonferroni-corrected, one-sided KS test. 5' UTR trinucleotide conservation was scored by number of identities in 28-species alignments.

Matched mRNA and protein datasets. MS/MS protein abundance measurements were obtained from published studies (12-17). Matched mRNA data were available in three studies (13-15). For the liver study (12) we used mean mRNA expression from GNF1M liver replicates (21). All data were mapped to Entrez Gene identifiers with the gene inheriting the highest score from any splice form. We excluded genes with poorly quantified mRNA values (expression values below 40) and the top 10% most highly expressed genes, based on mean mRNA expression values from the GNF1M atlas. We analyzed mouse genes with annotated 5' UTRs (>10nt), where all splice forms contained a uORF (6933 genes) or lacked a uORF (9343 genes). Differences in median protein expression were measured as percent reduction from uORF-less genes, using 10,000 permutations of gene uORF labels to assess significance. See Supporting Information for details.

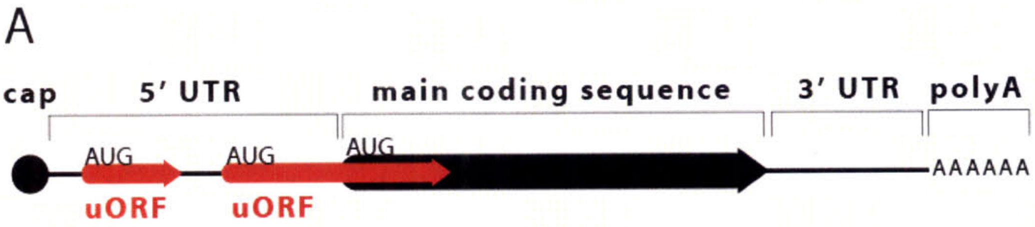
Luciferase assays. UTR sequences, up to and including the primary ATG initiation codon, were synthesized (IDT), cloned and ligated into the NheI site directly preceeding the *Renilla* luciferase gene in the dual-luciferase vector psiCHECK-2 (Promega) (Table S4). Prior to cloning, the ATG of the *Renilla* luciferase was mutated to TTG so that the *Renilla* luciferase expression would be driven by the primary ATG initiation codon of the gene under investigation. HEK 293A cells were seeded at 6000 cells/well in 96-well opaque white cell culture plates (Nunc). After overnight incubation, cells were transfected with 20-100ng of each construct using Fugene 6 (Roche). Forty-eight hours post transfection, cells were washed with PBS and lysed with Passive Lysis Buffer (Promega). *Renilla* and Firefly luciferase signals were generated using Promega's Dual-Luciferase Assay System according to the manufacturer's protocol. For each construct, *Renilla* luciferase signal was normalized to the Firefly luciferase internal control signal. Plates were read using a Victor^3 plate reader (Perkin Elmer) and the data analyzed using Wallac 1420 Workstation software. Two-sided, homoscedastic t-tests assessed significance.

Real-time PCR. HEK 293A cells were seeded at $2x10^5$ cells/well in 6-well cell culture plates 24 hrs before transfection. One µg of each construct was transfected per well using Fugene 6 as above. Forty-eight hours post transfection, cells were washed with PBS, and RNA was harvested using a Qiagen RNAeasy kit. First-strand cDNA synthesis was performed using SuperScript III (Invitrogen) using one µg of RNA from each transfection as starting material. Real-time PCR was performed using custom TaqMan Assays (ABI) designed against *Renilla* luciferase (target) and Firefly luciferase (endogenous control). Two-sided, homoscedastic t-tests assessed significance.

uORF-altering variants. Human dbSNP version 128 (18) was obtained from UCSC (46) and filtered for SNPs (class "single") that mapped to single locations within hg18 and overlapped annotated RefSeq 5' UTRs, excluding those that overlapped RefSeq CDSs. Custom perl scripts mapped SNPs onto mRNA transcripts and determined those that altered uORF presence. The Human Gene Mutation Database professional release 2008.2 (19), was searched for all non-coding substitutions or micro-lesions which altered presence of ATG codons and which overlapped 5' UTRs based on manual inspection of Blat alignments to hg18.

Acknowledgements

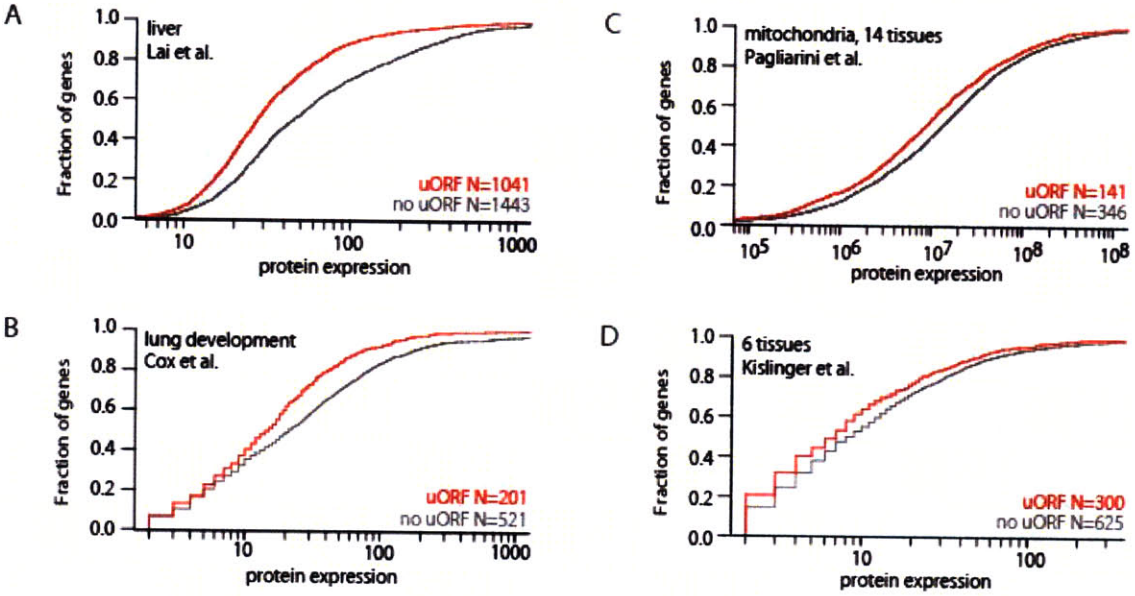
We thank S.E. Ong for advice on mass spectrometry; M. Garber for advice on alignments; O. Goldberger for technical assistance; J. Dixon for technical resources; and D. Altshuler, J. Hirshhorn, M. Springer, D. Neafsey, B. Voight, S. Carter, J. Avruch and E. Lander for comments on the manuscript and project. We thank the Broad Institute Sequencing Platform, Washington University Genome Sequencing Center, and Baylor Human Genome Sequencing Center for the vertebrate genome sequences used in comparative analyses. This work was supported by a grant from the National Institute of General Medical Science (GM077465) awarded to V.K.M.



B

# Transcripts with:	Human	Mouse
annotated 5' UTR	23775	18663
≥1 uORF	11670	8253
≥2 uORFs	6268	4197
≥1 uORF fully upstream	9879	6935
≥1 uORF overlapping CDS	4275	2872
Median Length (nt):		
5' UTR	170	139
uORF	48	48

Fig. 1. uORF definition and prevalence. (*A*) Schematic representation of mRNA transcript with two uORFs (red arrows), one fully upstream and one overlapping the main coding sequence (black arrow). uORFs are defined by a start codon (AUG) in the 5' UTR, an in-frame stop codon (arrow head) preceding the end of the main coding sequence, and length ≥9 nucleotides. (*B*) Number and length of uORFs in human and mouse RefSeq transcripts.



E

median reduction of expression for genes with vs without uORFs

	protein	mRNA	protein/mRNA
liver	39% (1e-5)	18% (1e-5)	21% (1e-4)
lung dev.	29% (0.007)	17% (0.08)	9% (0.26)
mitochondria	34% (0.008)	8% (0.12)	30% (0.03)
6 tissues	12% (0.361)	-3% (0.63)	24% (0.03)

Fig. 2. Protein expression of uORF-containing genes. (*A-D*) Cumulative distribution of protein expression for mouse genes containing uORFs (red curve) or lacking uORFs (gray curve) in each of four independent MS/MS studies (12-15). N indicates the number of unique genes in each set. (*E*) Median reduction of protein and mRNA expression for genes containing uORFs compared to genes lacking uORFs, with p-values (in parentheses) computed by empirical permutation testing.

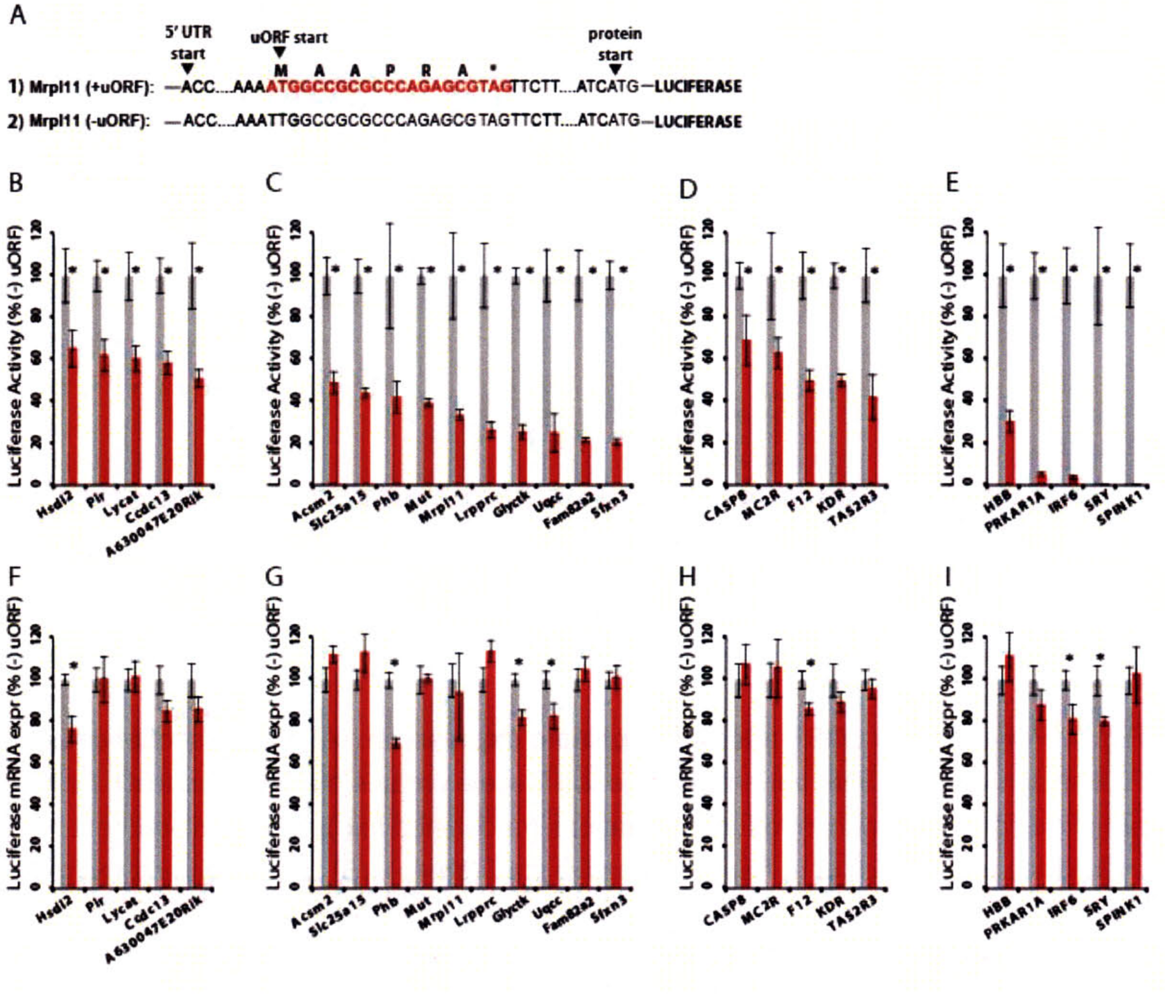


Fig. 3. Luciferase assays of uORF effects on protein and mRNA levels. Experimental design of reporter constructs with and without uORFs is shown for example *Mrpl11* (*A*). Normalized luciferase activity (*B-E*) and mRNA expression (*F-I*) are shown for reporter constructs that contain a uORF (red) or lack a uORF (gray) due to a mutation that disrupts the uORF start codon. The constructs contain 5' UTRs from: five mouse genes chosen randomly (*B,F*), ten mouse genes with proteomic and conservation signatures of functional uORFs (*C,G*), five human genes with polymorphic uORFs (*D,H*), and five human disease genes with uORF-altering mutations detected in patients (*E,I*). Error bars represent ± standard error of ≥ six biological replicates (*B-E*) and ≥ four technical replicates (*F-I*). Asterisks indicate significant difference (p<0.01).

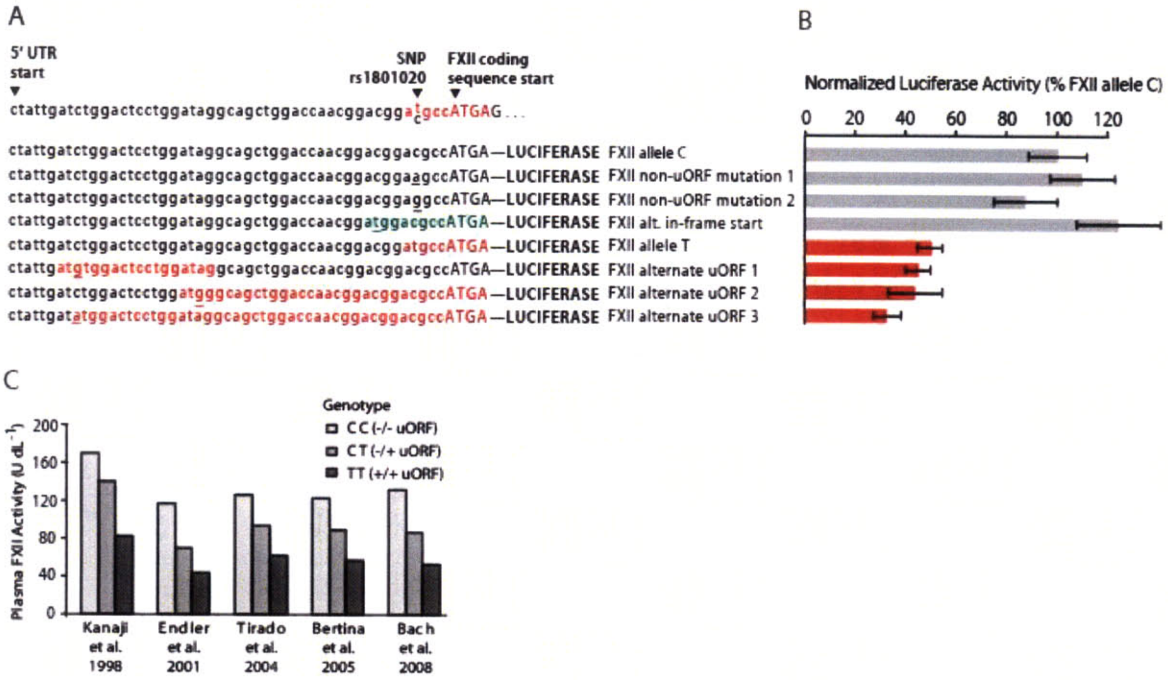


Fig. 4. Polymorphic uORF alters FXII protein expression. (*A*) 5' UTR sequence of FXII shown with two SNP variants, where the T allele creates a uORF (red text). Below are eight constructs with introduced mutations (underlined text), where colored text indicates a uORF (red) or in-frame alternative start (green). (*B*) Luciferase activity from reporter constructs listed in (*A*). Error bars represent ± standard deviation of ≥ six biological replicates. (*C*) Meta-analysis of plasma FXII activity levels measured by five independent studies, stratified by genotype of SNP rs1801020.

#	SNP	avHet	Gene	Gene description
1	rs1801020	50%	F12	coagulation factor XII (Hageman factor)
2	rs12272467	50%	TRIM6	tripartite motif-containing 6
3	rs1108842	50%	GNL3	guanine nucleotide binding protein-like 3 (nucleolar)
4	rs6460054	50%	CLDN3	claudin 3
5	rs1046188	50%	SCAMP3	secretory carrier membrane protein 3
6	rs13104310	49%	C4orf21	chromosome 4 open reading frame 21
7	rs7667298	49%	KDR	kinase insert domain receptor
8	rs7331765	49%	RASL11A	RAS-like, family 11, member A
9	rs2001216	49%	RCCD1	RCC1 domain containing 1
10	rs12975585	48%	HNRNPUL1	heterogeneous nuclear ribonucleoprotein U-like 1
11	rs2838343	46%	HSF2BP	heat shock transcription factor 2 binding protein
12	rs765007	46%	TAS2R3	taste receptor, type 2, member 3
13	rs17499247	45%	CREM	cAMP responsive element modulator
14	rs1048371	42%	MUCL1	mucin-like 1
15	rs1800070	*	CFTR	cystic fibrosis transmembrane conductance regulator
16	rs34704828	*	HBB	hemoglobin, beta
17	rs28926176	0.2%	MC2R	melanocortin 2 (ACTH hormone) receptor
18	rs41409645	4%	CCL3	chemokine (C-C motif) ligand 3
19	rs2856759	*	CCR5	chemokine (C-C motif) receptor 5
20	rs34819868	*	HAVCR1	hepatitis A virus cellular receptor 1
21	rs41275166	*	CD59	CD59 molecule, complement regulatory protein
22	rs6057688	*	DEFB119	defensin, beta 119
23	rs2234011	*	TAS2R5	taste receptor, type 2, member 5
24	rs1091826	*	OXTR	oxytocin receptor
25	rs6781226	*	HTR1F	5-hydroxytryptamine (serotonin) receptor 1F

Table 1. Notable human variants that create polymorphic uORFs. List contains common SNP variants (#1-14) and genes associated with monogenic disease (#15-17), immune response (#18-22), and receptor activity (#23-25 and 7, 12, 17, 19, 20). Table S3 contains complete list. AvHet indicates SNP's average heterozygosity (* indicates data not available).

#	Gene	Disease	Mutation	uORF link
1	THPO	Thrombocythaemia	splice site (3)	known
2	CDKN2A	Melanoma	G-34T (4)	known
3	HR	Marie Unna hereditary hypotrichosis	A-321G (5)	known
4	SRY	Gonadal dysgenesis	G-75A (32)	novel*
5	IRF6	Van der Woude syndrome	A-48T (33)	novel*
6	GCH1	DOPA-responsive dystonia	C-22T (34)	novel
7	HAMP	Juvenile haemochromatosis	G-25A (35)	predicted
8	KCNJ11	Hyperinsulinemic hypoglycemia, 2	C-54T (36) †	predicted
9	LDLR	Familial hypercholesterolaemia	delC-22 (37) †	predicted
10	PEX7	Rhizomelic chondrodysplasia punctata	C-45T (38) †	predicted
11	POMC	Proopiomelanocortin deficiency	C-11A (39)	predicted
12	PRKAR1A	Carney complex type 1	G-97A (40)	predicted*
13	SPINK1	Hereditary pancreatitis	C-53T (41)	predicted*
14	HBB	Thalassaemia beta	G-29A (42) †	predicted*

Table 2. uORF-altering mutations linked to disease. uORF-altering mutations detected in patients but not population controls. Mutation column includes 5' UTR position relative to translation start and literature reference (in parentheses). Cross indicates compound heterozygous mutations. The links between mutations and uORFs were previously known, predicted, or novel. Asterisk indicates mutations tested experimentally in this study.

References

1. Kozak, M. (1991) Structural features in eukaryotic mRNAs that modulate the initiation of translation. *The Journal of biological chemistry* **266,** 19867-19870.
2. Morris, D. R. & Geballe, A. P. (2000) Upstream open reading frames as regulators of mRNA translation. *Molecular and cellular biology* **20,** 8635-8642.
3. Wiestner, A., Schlemper, R. J., van der Maas, A. P., & Skoda, R. C. (1998) An activating splice donor mutation in the thrombopoietin gene causes hereditary thrombocythaemia. *Nature genetics* **18,** 49-52.
4. Liu, L., *et al.* (1999) Mutation of the CDKN2A 5' UTR creates an aberrant initiation codon and predisposes to melanoma. *Nature genetics* **21,** 128-132.
5. Wen, Y., *et al.* (2009) Loss-of-function mutations of an inhibitory upstream ORF in the human hairless transcript cause Marie Unna hereditary hypotrichosis. *Nature genetics*.**41,** 228-233.
6. Neafsey, D. E. & Galagan, J. E. (2007) Dual modes of natural selection on upstream open reading frames. *Molecular biology and evolution* **24,** 1744-1751.
7. Meijer, H. A. & Thomas, A. A. (2002) Control of eukaryotic protein synthesis by upstream open reading frames in the 5'-untranslated region of an mRNA. *Biochem J* **367,** 1-11.
8. Vilela, C. & McCarthy, J. E. (2003) Regulation of fungal gene expression via short open reading frames in the mRNA 5'untranslated region. *Mol Microbiol* **49,** 859-867.
9. Matsui, M., Yachie, N., Okada, Y., Saito, R., & Tomita, M. (2007) Bioinformatic analysis of post-transcriptional regulation by uORF in human and mouse. *FEBS letters* **581,** 4184-4188.
10. Iacono, M., Mignone, F., & Pesole, G. (2005) uAUG and uORFs in human and rodent 5'untranslated mRNAs. *Gene* **349,** 97-105.
11. Ingolia, N. T., Ghaemmaghami, S., Newman, J. R., & Weissman, J. S. (2009) Genome-Wide Analysis In Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. *Science*. Published Online February 12, 2009.
12. Lai, K. K., Kolippakkam, D., & Beretta, L. (2008) Comprehensive and quantitative proteome profiling of the mouse liver and plasma. *Hepatology* **47,** 1043-1051.
13. Cox, B., *et al.* (2007) Integrated proteomic and transcriptomic profiling of mouse lung development and Nmyc target genes. *Molecular systems biology* **3,** 109.
14. Pagliarini, D. J., *et al.* (2008) A mitochondrial protein compendium elucidates complex I disease biology. *Cell* **134,** 112-123.
15. Kislinger, T., *et al.* (2006) Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling. *Cell* **125,** 173-186.

16. Adachi, J., Kumar, C., Zhang, Y., & Mann, M. (2007) In-depth analysis of the adipocyte proteome by mass spectrometry and bioinformatics. *Mol Cell Proteomics* **6,** 1257-1273.

17. Williamson, A. J.*, et al.* (2008) Quantitative proteomics analysis demonstrates post-transcriptional regulation of embryonic stem cell differentiation to hematopoiesis. *Mol Cell Proteomics* **7,** 459-472.

18. Sherry, S. T.*, et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic acids research* **29,** 308-311.

19. Stenson, P. D.*, et al.* (2003) Human Gene Mutation Database (HGMD): 2003 update. *Human mutation* **21,** 577-581.

20. Kozak, M. (1987) Effects of intercistronic length on the efficiency of reinitiation by eucaryotic ribosomes. *Molecular and cellular biology* **7,** 3438-3445.

21. Su, A. I.*, et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America* **101,** 6062-6067.

22. Mariani, T. J., Reed, J. J., & Shapiro, S. D. (2002) Expression profiling of the developing mouse lung: insights into the establishment of the extracellular matrix. *American journal of respiratory cell and molecular biology* **26,** 541-548.

23. Kochetov, A. V.*, et al.* (1998) Eukaryotic mRNAs encoding abundant and scarce proteins are statistically dissimilar in many structural features. *FEBS letters* **440,** 351-355.

24. Kozak, M. (1986) Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. *Cell* **44,** 283-292.

25. Kozak, M. (1991) An analysis of vertebrate mRNA sequences: intimations of translational control. *The Journal of cell biology* **115,** 887-903.

26. Kozak, M. (2001) Constraints on reinitiation of translation in mammals. *Nucleic acids research* **29,** 5226-5232.

27. Bach, J.*, et al.* (2008) Coagulation factor XII (FXII) activity, activated FXII, distribution of FXII C46T gene polymorphism and coronary risk. *J Thromb Haemost* **6,** 291-296.

28. Bertina, R. M., Poort, S. R., Vos, H. L., & Rosendaal, F. R. (2005) The 46C-->T polymorphism in the factor XII gene (F12) and the risk of venous thrombosis. *J Thromb Haemost* **3,** 597-599.

29. Endler, G.*, et al.* (2001) A common C-->T polymorphism at nt 46 in the promoter region of coagulation factor XII is associated with decreased factor XII activity. *Thrombosis research* **101,** 255-260.

30. Kanaji, T.*, et al.* (1998) A common genetic polymorphism (46 C to T substitution) in the 5'-untranslated region of the coagulation factor XII gene is associated with low translation efficiency and decrease in plasma factor XII level. *Blood* **91,** 2010-2014.

31. Tirado, I., *et al.* (2004) Association after linkage analysis indicates that homozygosity for the 46C-->T polymorphism in the F12 gene is a genetic risk factor for venous thrombosis. *Thrombosis and haemostasis* **91,** 899-904.

32. Poulat, F., *et al.* (1998) Mutation in the 5' noncoding region of the SRY gene in an XY sex-reversed patient. *Human mutation* **Suppl 1,** S192-194.

33. Kondo, S., *et al.* (2002) Mutations in IRF6 cause Van der Woude and popliteal pterygium syndromes. *Nature genetics* **32,** 285-289.

34. Tassin, J., *et al.* (2000) Levodopa-responsive dystonia. GTP cyclohydrolase I or parkin mutations? *Brain* **123 (Pt 6),** 1112-1121.

35. Matthes, T., *et al.* (2004) Severe hemochromatosis in a Portuguese family associated with a new mutation in the 5'-UTR of the HAMP gene. *Blood* **104,** 2181-2183.

36. Huopio, H., *et al.* (2002) Acute insulin response tests for the differential diagnosis of congenital hyperinsulinism. *The Journal of clinical endocrinology and metabolism* **87,** 4502-4507.

37. Sozen, M. M., *et al.* (2005) The molecular basis of familial hypercholesterolaemia in Turkish patients. *Atherosclerosis* **180,** 63-71.

38. Braverman, N., *et al.* (2002) Mutation analysis of PEX7 in 60 probands with rhizomelic chondrodysplasia punctata and functional correlations of genotype with phenotype. *Human mutation* **20,** 284-297.

39. Krude, H., *et al.* (1998) Severe early-onset obesity, adrenal insufficiency and red hair pigmentation caused by POMC mutations in humans. *Nature genetics* **19,** 155-157.

40. Groussin, L., *et al.* (2002) Molecular analysis of the cyclic AMP-dependent protein kinase A (PKA) regulatory subunit 1A (PRKAR1A) gene in patients with Carney complex and primary pigmented nodular adrenocortical disease (PPNAD) reveals novel mutations and clues for pathophysiology: augmented PKA signaling is associated with adrenal tumorigenesis in PPNAD. *American journal of human genetics* **71,** 1433-1442.

41. Witt, H., *et al.* (2000) Mutations in the gene encoding the serine protease inhibitor, Kazal type 1 are associated with chronic pancreatitis. *Nature genetics* **25,** 213-216.

42. Oner, R., *et al.* (1991) The G----A mutation at position +22 3' to the Cap site of the beta-globin gene as a possible cause for a beta-thalassemia. *Hemoglobin* **15,** 67-76.

43. Cai, S. P., *et al.* (1992) Two novel beta-thalassemia mutations in the 5' and 3' noncoding regions of the beta-globin gene. *Blood* **79,** 1342-1346.

44. Bersano, A., Ballabio, E., Bresolin, N., & Candelise, L. (2008) Genetic polymorphisms for the study of multifactorial stroke. *Human mutation* **29,** 776-795.

45. Navratilova, Z. (2006) Polymorphisms in CCL2&CCL5 chemokines/chemokine receptors genes and their association with diseases. *Biomedical papers of the Medical Faculty of the University Palacky, Olomouc, Czechoslovakia* **150,** 191-204.

46. Karolchik, D., *et al.* (2003) The UCSC Genome Browser Database. *Nucleic acids research* **31,** 51-54.

47. Miller, W., *et al.* (2007) 28-way vertebrate alignment and conservation track in the UCSC Genome Browser. *Genome research* **17,** 1797-1808.

Chapter 5

—

Implications and Future Directions

Implications and Future Directions

Through the integration of diverse genomic datasets, we have deduced most of the protein components of human mitochondria, measured their amounts in numerous tissues, linked several newly discovered components to function in complex I assembly, identified disease-causing mutations, and demonstrated that uORFs cause widespread blunting of translation of mitochondrial and other cellular proteins. This work has important implications for mitochondrial medicine and for the systematic understanding of mitochondrial function.

Implications

Discovery of mitochondrial disease genes

Our mitochondrial protein compendium has already been used to identify six novel disease-related genes underlying inherited mitochondrial disorders. These represent a substantial contribution to the 86 total nDNA mitochondrial disease genes identified to date.

First, in collaboration with Antonella Spinazzola, Massimo Zeviani and colleagues, we showed that defects in *MPV17* cause infantile hepatic mitochondrial DNA depletion in three unrelated families[1]. This devastating disorder presents within six months of birth and is characterized by persistent vomiting, failure to thrive, hypotonia, hypoglycemia and progressive neurological symptoms[2]. Pathogenic mutations in *MPV17* were confirmed using yeast complementation, whereby the wild type but not mutant alleles of *MPV17* rescued an ethanol growth phenotype in yeast cells lacking the homolog of *MPV17, Sym1*[1]. Our discovery led to the subsequent identification of *MPV17* mutations underlying Navajo neurohepatopathy, a disease having a relatively high prevalence within the western Navajo Reservation due to a founder effect (1 in 1600 live births)[3]. In addition, homozygous or compound heterozygous mutations in *MPV17* have since been found in seven unrelated families[4-8]. These subsequent studies have helped define the mutational spectrum of this disease (nine unique mutations) and relate genotype to clinical phenotype[2,8]. It is estimated that *MPV17* defects account for 22% of hepatocerebral mtDNA depletion cases[2]. Moreover, a knockout mouse strain (Mpv17-/-) is now being used to model the human disease pathogenesis, as it exhibits severe mtDNA depletion in liver and skeletal muscle[9].

Second, our phylogenetic profiling approach has led to the discovery of three genes underlying complex I deficiency: *C8orf38*[10], *C20orf7*[11], and *FOXRED1* (manuscript in preparation). Complex I deficiency is the most common form of respiratory chain disease. The patients with complex I deficiency presented neonatally or within 10 months of birth, and showed biochemical complex I deficiency, elevated lactate, and, for two siblings, focal right-hand seizures, ataxia, and evolving rigidity[10,11]. We demonstrated that RNAi knockdown of both *C8orf38* and *C20orf7* in MCH58 fibroblast cells caused reduced complex I activity, consistent with patient phenotypes[10,11]. Validation of *FOXRED1* mutations are still pending.

Third, we used MitoCarta to select candidate genes for an autosomal dominant muscle myopathy, leading to discovery of pathogenic mutations in *CHCHD10* in collaboration with Senda Ajroud-Driss and colleagues (manuscript in progress).

Lastly, our published Maestro compendium was used in the discovery of *TMEM70* mutations underlying isolated ATP synthase deficiency and neonatal mitochondrial encephalocardiomyopathy[12]. Although we were not directly involved in this discovery, we hope our Maestro and MitoCarta inventories will be used routinely in this manner to highlight mitochondrial gene candidates within regions of the genome to which mitochondrial disease phenotypes have been linked.

Development of molecular diagnostics and potential therapeutics

Since the discovery of defects in *MPV17* in 2006, there has been progress on development of diagnostics and therapies for mtDNA depletion. Starting in late 2006, Baylor college of Medicine added *MPV17* sequencing to its list of routine diagnostic tests for mitochondrial diseases. This diagnostic test requires 2-3cc of blood in an EDTA tube, takes 4 weeks turnaround time, and costs $300[13]. Therapeutically, a diet-based treatment was developed by Zeviani and colleagues and assessed in two individuals with severe *MPV17* mutations[8]. These two patients showed improved liver function tests upon continuous IV glucose infusion or when fed every 3h[8]. With additional validation, this diet-based approach may prove to be an effective therapy to slow disease progression for patients with *MPV17* mutations.

For the other disease gene discoveries, published after July 2008, there has not yet been time for development of clinical applications, although we anticipate that diagnostic tests will be available soon.

Pathogenic uORF-altering mutations

In addition to discovery of novel disease-related genes, above, we have also generated evidence supporting the pathogenicity of five mutations that create uORFs. These mutations in HBB[14], SRY[15], IRF6[16], PRKAR1A[17], and SPINK1[18] were previously found to be associated with disease, but there was no evidence supporting their causality. Using reporter constructs in cell models, we showed that these single mutations that create uORFs cause severe reduction of protein encoded by downstream open reading frames. These experiments provide evidence that these five uORF-altering mutations may cause disease. When added to the three previously known uORF-related disease mutations[19-21], these cases suggest that mutations in upstream open reading frames may be a more common cause of disease than previously suspected. For this reason, resequencing projects aimed at discovering disease-causing mutations should not be limited to coding exons, but should also include upstream regions.

Future directions

Mito10K: systematic search for genes underlying complex I deficiency

While we have successfully applied our compendium to discover gene defects underlying inherited forms of mitochondrial disease, we also plan to tackle mitochondrial diseases that arise sporadically. Advances in sequencing technology make it possible to systematically address all forms of complex I deficiency using a candidate gene approach. In collaboration with Dr. David Thorburn, we have initiated a project to sequence 100 gene candidates, detected by MitoCarta and phylogenetic profiling[10], in 100 patients having well-characterized complex I deficiency. We term this project "Mito10K", reflecting the 10,000 total genes that require sequencing. Using massively parallel Illumina technology, we are currently sequencing 831 candidate gene exons at an average coverage of 80X. While much computational and experimental work is required to identify and validate pathogenic mutations, this methodology is especially promising because it should systematically uncover sporadic mutations leading to complex 1 deficiency. If successful, this approach could transform the diagnosis of this mitochondrial disease.

Decoding components of mitochondrial pathways

In addition to helping elucidate the molecular basis of disease, our MitoCarta compendium also provides a foundation for systematic investigation of mitochondrial function. Current work in the Mootha laboratory is directly applying this inventory to (i) annotate gene function based on mRNA coexpression across thousands of microarray datasets; (ii) identify regulators of mtDNA copy number through an RNAi screen; and (iii) identify protein regulators of mitochondrial calcium signaling through an RNAi screen. Additionally, kinases and phosphotases present within MitoCarta can help decipher the signaling networks within mitochondria.

Identifying mitochondrial regulatory and targeting elements

The inventory also provides a resource for the identification of transcriptional, post-transcriptional, and targeting elements present in mitochondrial genes. Transcriptional regulatory elements can be identified by detecting sequence motifs specifically enriched in subsets of MitoCarta gene promoters. Similarly, sequence elements statistically enriched in MitoCarta gene 5' UTRs and 3' UTRs may lead to discovery of post-transcriptional regulatory elements. Additionally, analysis of protein sequences of MitoCarta proteins that lack canonical targeting signals may elucidate novel targeting motifs that direct mitochondrial localization.

Systems biology of the mitochondrion

The work described here takes a classic reductionist approach to characterize the location and function of individual molecular components. However, by defining the mitochondrial proteome, this work lays the foundation for a systems approach to understanding how these thousands of parts are coordinately regulated and assembled into functional networks that respond to changing cellular needs. The network properties of pathways residing within mitochondria, and between mitochondria and the rest of the cell, will ultimately enable the understanding of complex diseases involving mitochondrial dysfunction.

References

1. Spinazzola, A. et al. MPV17 encodes an inner mitochondrial membrane protein and is mutated in infantile hepatic mitochondrial DNA depletion. *Nat Genet* **38**, 570-5 (2006).
2. Spinazzola, A. et al. Clinical and molecular features of mitochondrial DNA depletion syndromes. *J Inherit Metab Dis* (2008).
3. Karadimas, C.L. et al. Navajo neurohepatopathy is caused by a mutation in the MPV17 gene. *Am J Hum Genet* **79**, 544-8 (2006).
4. Sarzi, E. et al. Mitochondrial DNA depletion is a prevalent cause of multiple respiratory chain deficiency in childhood. *J Pediatr* **150**, 531-4, 534 e1-6 (2007).
5. Wong, L.J. et al. Mutations in the MPV17 gene are responsible for rapidly progressive liver failure in infancy. *Hepatology* **46**, 1218-27 (2007).
6. Navarro-Sastre, A. et al. Lethal hepatopathy and leukodystrophy caused by a novel mutation in MPV17 gene: description of an alternative MPV17 spliced form. *Mol Genet Metab* **94**, 234-9 (2008).
7. Spinazzola, A. et al. Hepatocerebral form of mitochondrial DNA depletion syndrome: novel MPV17 mutations. *Arch Neurol* **65**, 1108-13 (2008).
8. Parini, R. et al. Glucose metabolism and diet-based prevention of liver dysfunction in MPV17 mutant patients. *J Hepatol* **50**, 215-21 (2009).
9. Viscomi, C. et al. Early-onset liver mtDNA depletion and late-onset proteinuric nephropathy in Mpv17 knockout mice. *Hum Mol Genet* **18**, 12-26 (2009).
10. Pagliarini, D.J. et al. A mitochondrial protein compendium elucidates complex I disease biology. *Cell* **134**, 112-23 (2008).
11. Sugiana, C. et al. Mutation of C20orf7 disrupts complex I assembly and causes lethal neonatal mitochondrial disease. *Am J Hum Genet* **83**, 468-78 (2008).
12. Cizkova, A. et al. TMEM70 mutations cause isolated ATP synthase deficiency and neonatal mitochondrial encephalocardiomyopathy. *Nat Genet* **40**, 1288-90 (2008).
13. Baylor College of Medicine Medical Genetics Laboratories.
14. Oner, R. et al. The G----A mutation at position +22 3' to the Cap site of the beta-globin gene as a possible cause for a beta-thalassemia. *Hemoglobin* **15**, 67-76 (1991).
15. Poulat, F. et al. Mutation in the 5' noncoding region of the SRY gene in an XY sex-reversed patient. *Hum Mutat* **Suppl 1**, S192-4 (1998).
16. Kondo, S. et al. Mutations in IRF6 cause Van der Woude and popliteal pterygium syndromes. *Nat Genet* **32**, 285-9 (2002).
17. Groussin, L. et al. Molecular analysis of the cyclic AMP-dependent protein kinase A (PKA) regulatory subunit 1A (PRKAR1A) gene in patients with Carney complex and primary pigmented nodular adrenocortical disease (PPNAD) reveals novel

mutations and clues for pathophysiology: augmented PKA signaling is associated with adrenal tumorigenesis in PPNAD. *Am J Hum Genet* **71**, 1433-42 (2002).

18. Witt, H. et al. Mutations in the gene encoding the serine protease inhibitor, Kazal type 1 are associated with chronic pancreatitis. *Nat Genet* **25**, 213-6 (2000).
19. Liu, L. et al. Mutation of the CDKN2A 5' UTR creates an aberrant initiation codon and predisposes to melanoma. *Nat Genet* **21**, 128-32 (1999).
20. Wen, Y. et al. Loss-of-function mutations of an inhibitory upstream ORF in the human hairless transcript cause Marie Unna hereditary hypotrichosis. *Nat Genet* **41**, 228-33 (2009).
21. Wiestner, A., Schlemper, R.J., van der Maas, A.P. & Skoda, R.C. An activating splice donor mutation in the thrombopoietin gene causes hereditary thrombocythaemia. *Nat Genet* **18**, 49-52 (1998).

Appendix A

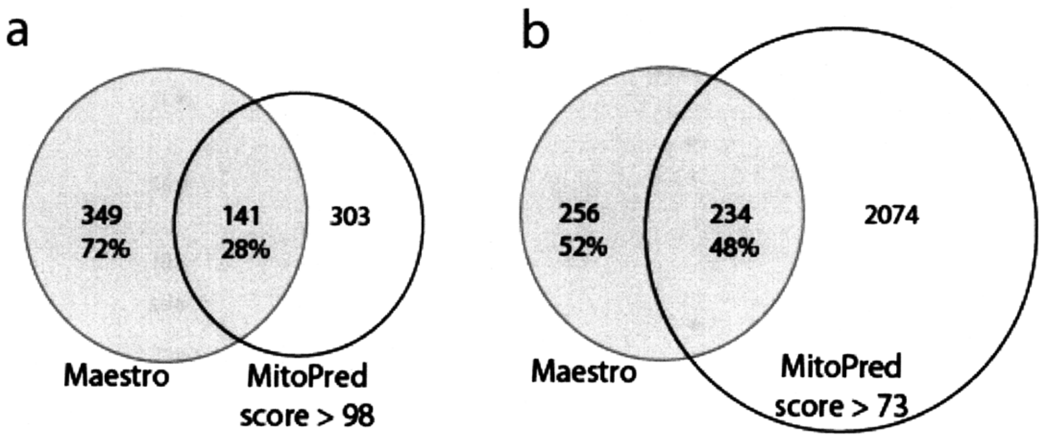
—

Supplementary Material for Chapter 2

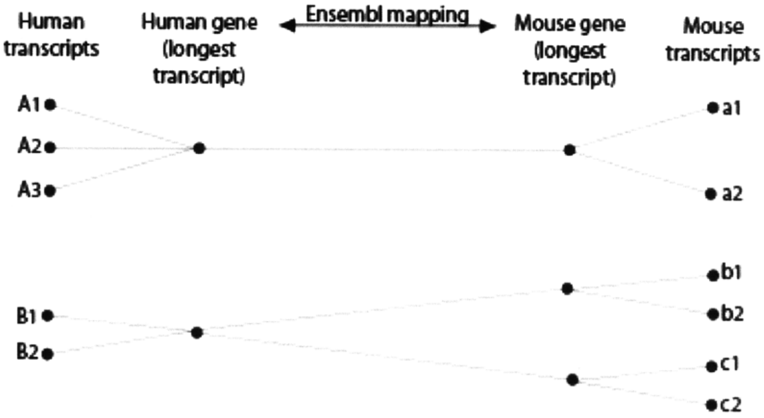
Supplementary Data: Systematic identification of human mitochondrial disease genes through integrative genomics

Supplementary Figures and Tables

Below are Supplementary Figures S1-S3 and Supplementary Table S1. Supplementary Tables 2-5 are large Excel files that are available online at: www.nature.com/ng/journal/v38/n5/suppinfo/ng1776_S1.html.



Supplementary Figure 1. Maestro and MitoPred Overlap.
While a direct comparison between prediction methods is not possible since the algorithms were trained on different datasets, we show the overlap of Maestro novel protein predictions (threshold 5.65) with the novel predictions of MitoPred using both stringent (panel a) and relaxed (panel b) threshold criteria.



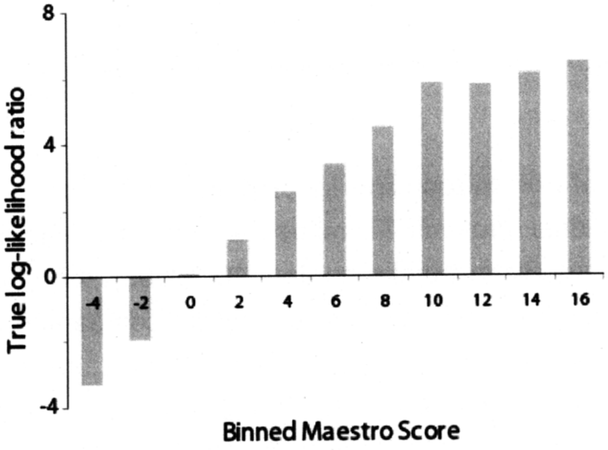
Supplementary Figure 2. Mouse to human orthology mapping.

For experiments performed on mouse models (mass spec, induction, GNF mouse tissue coexpression), mouse gene transcripts were mapped to human counterparts based on an Ensembl homology mapping that uses synteny and gene sequence similarity (See EnsMart www.ensembl.org). Since the Ensembl orthology mapping is performed at the gene level (using the longest transcript for each gene locus), we computed a transcript level orthology mapping with each transcript inheriting all orthologs from its gene locus. For example, transcript A1 was mapped to two mouse transcripts (a1,a2), as were A2 and A3. Transcript B1 was assigned four mouse orthologs (b1, b2, c1, c2) from two different mouse gene loci.

a b

A	B	C	D	E	F	G	H				A	B	C	D	E	F	G	H	
1.00	0.14	0.06	0.03	-0.04	0.18	0.17	0.13	A	**Target Signal**	A	1.00	-0.01	-0.02	0.03	0.02	0.06	0.00	0.02	A
	1.00	0.04	0.33	0.01	0.04	0.13	0.05	B	**Domains**	B		1.00	-0.02	0.09	0.02	-0.03	0.05	0.02	B
		1.00	0.05	-0.06	0.15	0.27	0.08	C	**Motif**	C			1.00	-0.02	0.04	-0.02	-0.02	0.04	C
			1.00	-0.04	0.03	0.05	0.06	D	**Yeast**	D				1.00	-0.01	0.05	0.03	0.00	D
				1.00	-0.09	0.00	-0.01	E	**Ancestry**	E					1.00	-0.03	-0.01	0.01	E
					1.00	0.50	0.16	F	**Coexpression**	F						1.00	0.01	0.09	F
						1.00	0.20	G	**Mass Spec**	G							1.00	0.01	G
							1.00	H	**Induction**	H								1.00	H

c



Supplementary Figure 3. Correlation between 8 datasets.

a. Pairwise correlation coefficient of scores between different genomic features for gold standard mitochondrial proteins Tmito. **b.** Pairwise correlation coefficient of scores between different genomic features for gold standard nonmitochondrial proteins T~mito. **c.** Log-likelihood ratio of Maestro scores (y-axis) computed as P(score in interval| Tmito) / P(score in interval | T~mito) for each range of Maestro scores (x-axis). This plotshows that Maestro scores are linearly related to computed log-likelihoods of mitochondrial localization until score 10.

GO identifier	Cellular location	GO identifier	Cellular location
GO:0009986	cell surface	GO:0005794	Golgi apparatus
GO:0000785	chromatin	GO:0005911	intercellular junction
GO:0005929	cilium	GO:0042470	melanosome
GO:0005905	coated pit	GO:0042579	microbody
GO:0016023	cytoplasmic vesicle	GO:0005792	microsome
GO:0005856	cytoskeleton	GO:0005634	nucleus
GO:0005830	cytosolic ribosome	GO:0005886	plasma membrane
GO:0005783	endoplasmic reticulum	GO:0009579	thylakoid
GO:0005768	endosome	GO:0005923	tight junction
GO:0031012	extracellular matrix	GO:0000151	ubiquitin ligase complex
GO:0005576	extracellular region	GO:0005773	vacuole
GO:0009434	flagellum (Eukaryota)		

Supplementary Table 1. Computation of non-mitochondrial protein set T~mito. The gold-standard set of non-mitochondrial proteins (T~mito) was created from all Ensembl human and mouse proteins (www.ensembl.org, Jan 10, 2005) that had SWISSPROT entries and Gene Ontology (GO) annotations to specific compartments outside of the mitochondrion. GO represents the cellular compartment structure as a complex tree allowing nodes to have multiple ancestors (i.e. a directed acyclic graph). Ensembl associates proteins with identifiers (GO IDs) to the most specific nodes within the GO tree. A protein was considered "non-mitochondrial" if the annotated GO ID was within the subtree of any of the compartments listed above. Because of the mixed quality of GO annotations, we conservatively considered only those proteins for which the human and mouse orthologs (see Methods) were both "non-mitochondrial".

Supplementary Methods

Protein domain analysis

In order to identify sets of mitochondrial-specific proteins domains across eukaryotic species, we first created sets of mitochondrial and non-mitochondrial proteins from the set of SwissProt eukaryotic proteins (release 48.8, 1/23/06), as described in Methods. For the entries with high confidence 'subcellular location' (excluding 'by similarity', 'potential','probable',and 'possible' entries), we partitioned the proteins into two sets based on the annotations in the 'subcellular location' field:

- S_{mito} contained the keyword Mitochondria, and its variants (Mitochondrial, Mitochondrion)
- $S_{\sim mito}$ started with one of the following 36 locations: Cell membrane, Cell surface, Cell wall, Centrosomal, Centrosome, Cis-Golgi complex, Coated vesicle, Cytoskeletal, Cytoskeleton, Cytosol, Endoplasmic reticulum, Endosomal, Endosomal, Flagellar, Focal adhesion, Glycosomal, Glyoxysomal, Golgi, Lysosomal, Microsomal, Microsomes, Mitotic spindle, Neuromuscular junction, Nuclear, Nucleolar, Nucleoplasmic, Nucleus, Peroxisomal, Plasma membrane, Secreted, Surface protein, Synapse, Synaptic vesicles, Vacuolar, Vacuole, Vesicular. All entries containing the keyword mitochondria (or variant mitochondrial, mitochondrion) were filtered out.

Mass spectrometry validation

In gel digestion protocol – 2 lanes each of 100ug of liver homogenate and purified liver mitochondria were size separated by a 10-20% Tris-HCl gradient. Gel slices were incubated in 50% acetonitrile, 50mM ammonium bicarbonate for extraction of coomassie stain, and then dried by vacuum centrifugation and subjected to reduction in 10mM DTT for 45min at 60°C, followed by alkylation in 55mM iodoacetamide for 60 minutes at room temperature. Each gel slice was then subjected to in-gel tryptic digestion (immersed with 1ug trypsin in 100mM ammonium bicarbonate solution overnight at 370°C). Peptides were extracted from gel slices using three cycles of vortexing in 70% acetonitrile, 1% trifluoroacetic acid in water and 1% trifluoroacetic acid in water. Extracted material were lyophilized by vacuum centrifugation and resuspended in 20uL of 3% acetonitrile/5% formic acid in water.

Reverse phase chromatography – Peptides extracted from the gel slices were then analyzed by reverse phase liquid chromatography tandem mass spectrometry (LC-MS/MS) using an LTQ-Orbitrap (Thermo, San Jose, CA). Analysis using a linear acetonitrile gradient on 3µm 200Å C18 reverse phase material packed to 30cm into a 10µm tip, 75µm ID picofrit™ column (New Objective, Woburn, MA). Gradient performed from 7% B to 40% B in 27.5 minutes using mobile phases A: 0.1% formic acid in water and B: 90% acetonitrile/0.1% formic acid in water.

Mass spectrometry –We performed a LTQ-Orbitrap survey scan over the range 300 to 1500 m/z, followed by eight data dependent tandem MS (from a set inclusion list) using dynamic exclusion (m/z precursor selection for MS/MS repeat count of 2 within 20sec, exclusion for 60sec, early expiration from exclusion list at signal to noise threshold of 4). The inclusion list of 303 masses was derived from a set of 30 proteins, representing 224 distinct peptides; with accurate mass selection criteria for tandem MS set at 15ppm to left and 25 ppm to right of the monoisotopic peak.

Database searching – Using Spectrum Mill™ Extractor, spectra were filtered (requiring sequence tag length > 1) and redundant spectra (identical m/z observed within 20sec windows) were merged. These spectra were then searched against the Ensembl mouse protein database (www.ensembl.org, Jan 10, 2005). Search parameters included:

- carbamidomethylation of cysteines,
- 50% minimum matched peak intensity,
- precursor mass tolerance of +/- 0.05 Da,
- production ion mass tolerance of +/- 0.7 Da,
- 1 missed tryptic cleavage,
- electrospray ionization trap scoring settings

Autovalidation scoring threshold criteria:
(a) *protein details mode, charge-dependent*:

- peptide charge +1: (score > 7, SPI (scored peak intensity) >70%)
- peptide charge +2: (score > 8, SPI>70%) or (score > 6, SPI>90%)
- peptide charge +3: (score > 9, SPI>70%)
- peptide charge +4: (score > 9, SPI>70%)

(b) *peptide mode*: peptide charge +1,+2,+3,+4: score >13, SPI > 70%

Inclusion List - A mass list was generated from *in silico* tryptic digestion of 30 proteins, allowing for no miscleavages, charge states 2, 3, 4 and a mass range between 300 to 1500 m/z.

Comparison of integration approaches

In addition to integrating the eight genome-wide scores with a naïve Bayes network, we also explored several other methods:

- *decision trees.* We applied the CART decision tree software (Salford Systems) to our eight genomic datasets. Decision trees output easily interpretable results showing the contribution of each type of data. CART has been successfully used in other genomics applications [1], however it performed less well on our large gold-standard training data than naïve Bayes networks.

- *boosting.* Boosting is a method of improving the performance of any machine learning algorithm based on the assumption that a series of weak algorithms can be "boosted" into a strong algorithm by successively training on different subsets of the initial data. We evaluated this method by integrating our data using the Boostexter[2] (www.cs.princeton.edu/~schapire/BoosTexter) implementation of the Schapire and Singer AdaBoost algorithm[3], as well as implementing a version of this algorithm based on the simple naïve Bayes integration. In neither implementation did the boosting methodology improve the training results.

References

1. Davuluri, R.V., Suzuki, Y., Sugano, S. & Zhang, M.Q. CART classification of human 5' UTR sequences. *Genome Res* **10**, 1807-16 (2000).
2. Schapire, R.E. & Singer, Y. BoosTexter: A Boosting-based System for Text Categorization. *Machine Learning* **39**, 135-168.
3. Schapire, R.E. A Brief Introduction to Boosting. *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, 1999.*

Appendix B

—

Supplementary Material for Chapter 3

Supplementary Data:

A mitochondrial protein compendium elucidates complex I disease biology

Supplementary Figures and Tables

Below are Supplementary Figures S1-S8 and Supplementary Tables S1, S4, and S6. Supplemental Tables S2, S3, S5, S7, and S8 are Excel spreadsheets that are available online at www.cell.com/supplemental/S00928674(08)00768-X.

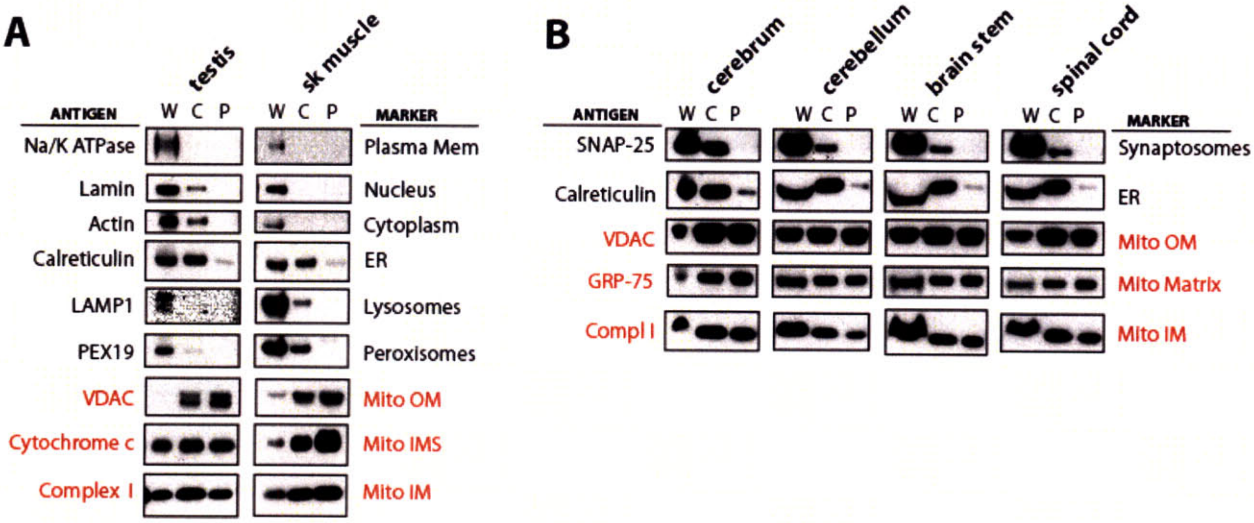


Figure S1. Assessment of Mitochondrial Purity.

(A) Enrichment of mitochondria isolated from testis or skeletal muscle assessed at three stages of the isolation procedure (W: whole tissue lysate; C: crude mitochondrial extracts; P: purified mitochondrial extracts) by immunoblot against various protein markers (contamination markers in black font, mitochondrial markers in red font).

(B) Same as (A) for mitochondria isolated from four brain regions. Here, immunoblots against SNAP-25 was included to ensure loss of synaptosomes from the mitochondrial preparations.

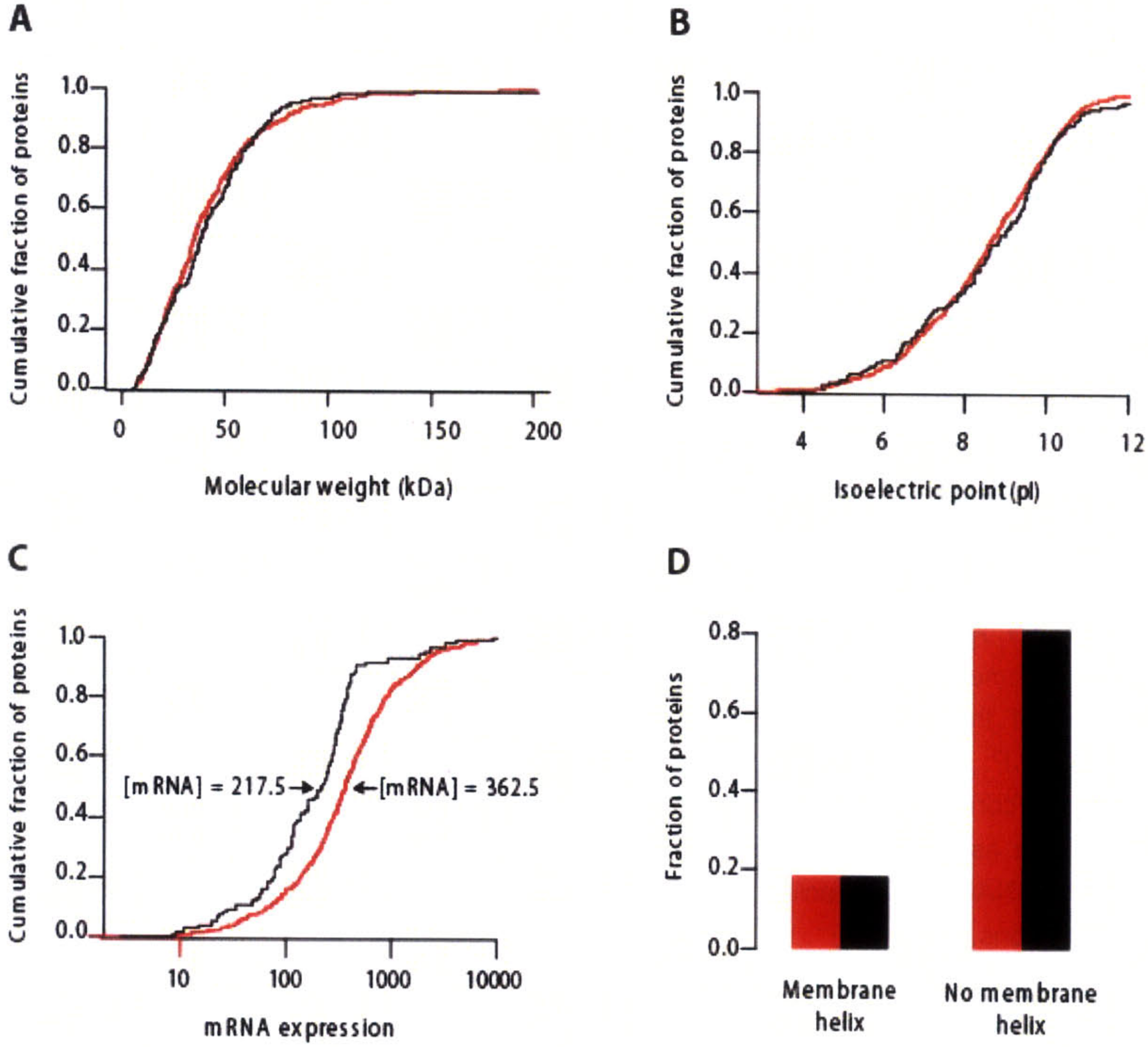


Figure S2. Analysis of MS/MS Detection Biases.

(A-C) The cumulative distributions for molecular weight (A), isoelectric point (B), and mRNA abundance (C) are plotted for MitoCarta proteins detected (red) or not detected (black) by our proteomic survey. Molecular weight and isoelectric points were calculated from the underlying protein sequences. Median mRNA expression levels for the fourteen tissues sampled was obtained from the GNF tissue atlas (Su et. al. 2004).

(D) Fraction of genes containing one or more membrane helices (based on TMHMM prediction, www.cbs.dtu.dk/services/TMHMM) are plotted for MitoCarta proteins detected (red) or not detected (black) by MS/MS.

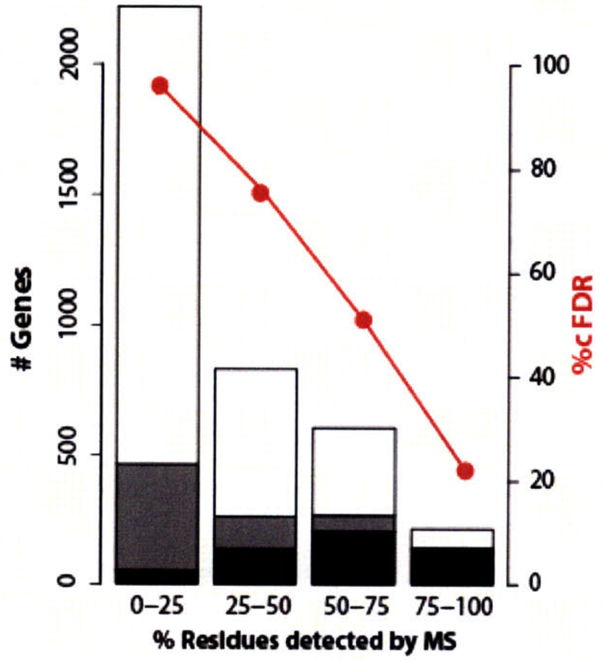


Figure S3. Assessment of Discovery MS/MS.

Number of genes detected by discovery MS/MS at four levels of protein abundance, measured by coverage (black, known mitochondrial proteins (T_{mito}); gray, non-mitochondrial proteins ($T_{\sim\mathrm{mito}}$); white, other proteins). Overlaid in red is the corrected false discovery rate for each coverage range estimated from T_{mito} and $T_{\sim\mathrm{mito}}$. The high cFDRs for discovery MS/MS data alone motivate the need to distinguish co-purifying contaminants using subtractive techniques, shown in Figure 2C-D.

A

A	B	C	D	E	F	G		
1	-0.06	0.14	0.06	0.05	-0.05	0.04	A	TargetP
	1	0.23	0.02	0.03	0.32	0.05	B	YeastMitoHomolog
		1	0.01	0.05	0.21	-0.05	C	RpExpect
			1	0.21	0.04	0.17	D	PGC_induction
				1	0.06	0.31	E	Coexpression
					1	0.14	F	MitoDomain
						1	G	MS/MS

B

	A	B	C	D	E	F	G
A	1	0	0.02	0.02	0	-0.01	0
B		1	0.04	0	0	0.18	0.24
C			1	0	0.06	0.17	0.09
D				1	0.03	0	0.01
E					1	-0.02	0.05
F						1	0.06
G							1

Figure S4. Correlation between Seven Genome-Scale Data Sets.

(A) Pairwise Pearson correlation coefficient of scores between seven genomic features, within the training set of known mitochondrial proteins $T_{\text{mito.}}$

(B) Pairwise Pearson correlation coefficient of scores between seven genomic features, within the training set of known non-mitochondrial proteins $T_{\text{~mito.}}$

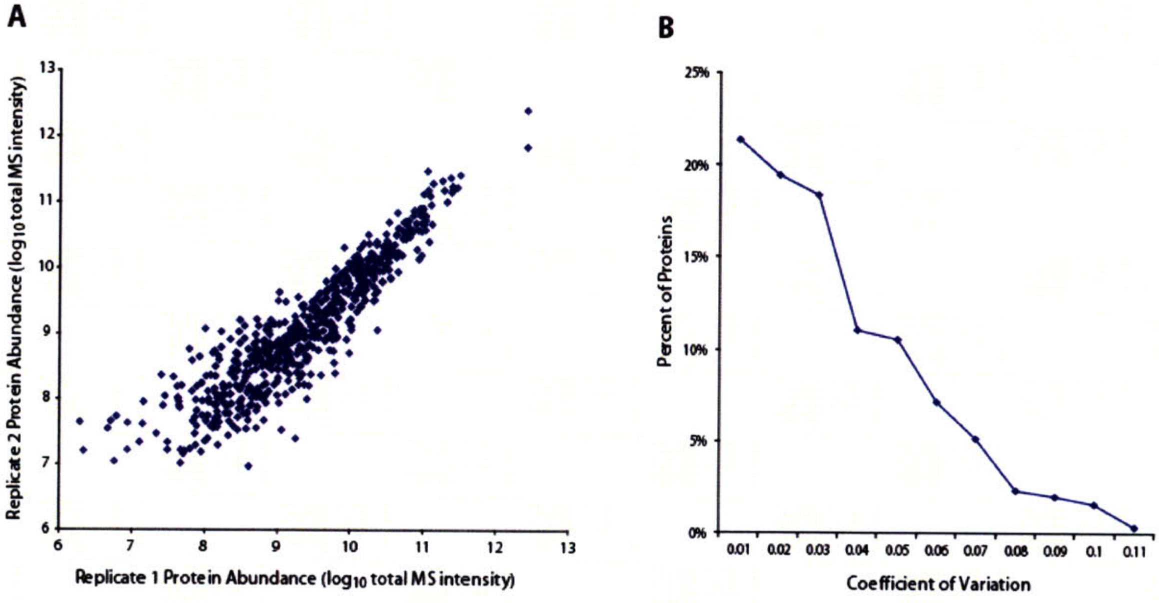


Figure S5. Discovery Proteomics Reproducibility.

(A) Protein abundance measurements of the 643 MitoCarta proteins detected by mass spectrometry within two technical replicates of liver mitochondrial samples (R^2=0.83).

(B) Histogram of coefficient of variation (sd/mean) for proteins identified in both replicates.

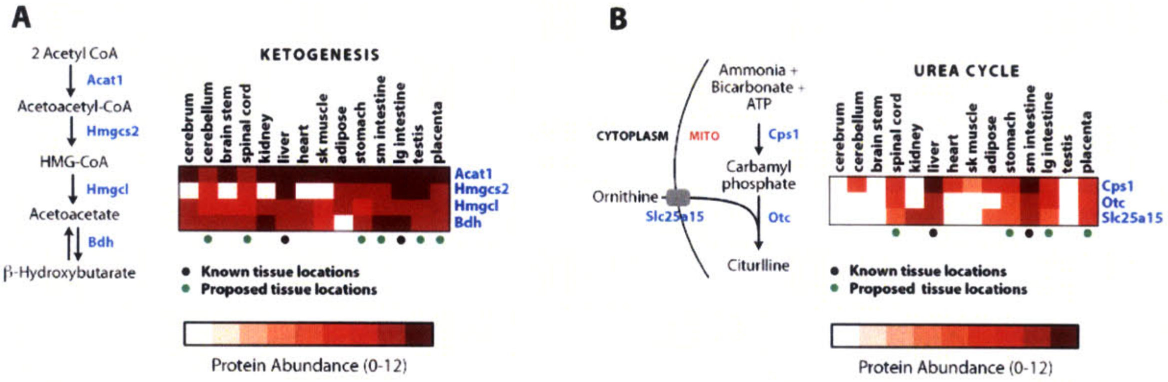


Figure S6. Tissue Distribution of Ketogenesis and Urea Cycle Pathways.

Tissues containing all pathway enzymes for ketogenesis (A) or the urea cycle (B) are marked with dots colored black (previously known to exist in this tissue) or green (novel). Key: protein abundance from proteomic analyses, as in Figure 4A.

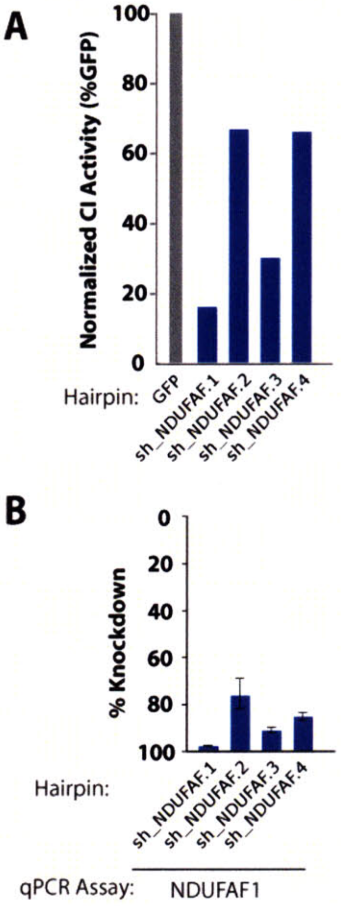
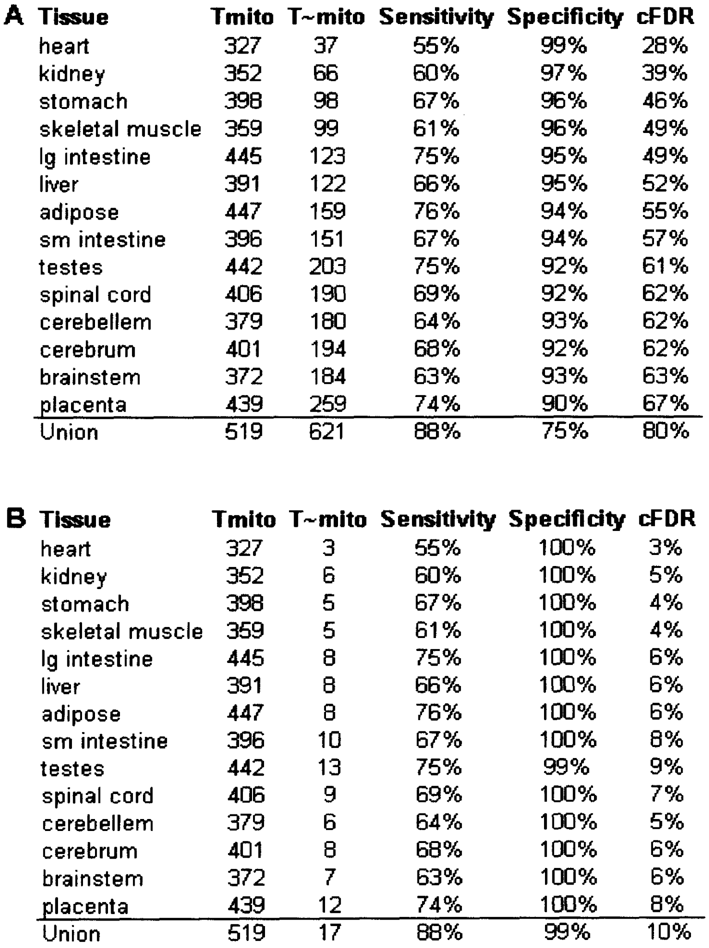


Figure S7. Sensitivity of Immunocapture Assays for Assessing CI Activity.

(A) CI activity assays from fibroblasts following lentiviral-mediated delivery of four distinct hairpins targeted against the known CI assembly factor *NDUFAF1,* or against *GFP* (negative control).

(B) Assessment of knockdown efficiency of each hairpin in (A) by real-time qPCR, showing correlation of knockdown efficiency to CI activity.



A

Tissue	Tmito	T~mito	Sensitivity	Specificity	cFDR
heart	327	37	55%	99%	28%
kidney	352	66	60%	97%	39%
stomach	398	98	67%	96%	46%
skeletal muscle	359	99	61%	96%	49%
lg intestine	445	123	75%	95%	49%
liver	391	122	66%	95%	52%
adipose	447	159	76%	94%	55%
sm intestine	396	151	67%	94%	57%
testes	442	203	75%	92%	61%
spinal cord	406	190	69%	92%	62%
cerebellem	379	180	64%	93%	62%
cerebrum	401	194	68%	92%	62%
brainstem	372	184	63%	93%	63%
placenta	439	259	74%	90%	67%
Union	519	621	88%	75%	80%

B

Tissue	Tmito	T~mito	Sensitivity	Specificity	cFDR
heart	327	3	55%	100%	3%
kidney	352	6	60%	100%	5%
stomach	398	5	67%	100%	4%
skeletal muscle	359	5	61%	100%	4%
lg intestine	445	8	75%	100%	6%
liver	391	8	66%	100%	6%
adipose	447	8	76%	100%	6%
sm intestine	396	10	67%	100%	8%
testes	442	13	75%	99%	9%
spinal cord	406	9	69%	100%	7%
cerebellem	379	6	64%	100%	5%
cerebrum	401	8	68%	100%	6%
brainstem	372	7	63%	100%	6%
placenta	439	12	74%	100%	8%
Union	519	17	88%	99%	10%

Figure S8. Estimated False Discovery Rates of Discovery MS/MS.

Corrected false discovery rates (cFDRs) calculated from the numbers of training set gene products detected by discovery MS/MS in each tissue. (A) All detected gene products listed, including MitoCarta and all other mouse genes. (B) Only MitoCarta gene products listed, showing improvements in specificity achieved by Bayesian integration.

Publication	Current study	Foster et. al. 2006	Kislinger et. al. 2006	Johnson et. al. 2006	Forner et. al. 2006	Mootha et. al. 2003
Method to filter out contaminants	**MitoCarta**	**PCP**	**KNN mito (>0.8)**	**lit & targeting signal**		
# tissues analyzed	14	1	6	4	3	4
# mouse gene loci detected	1097	257	301	260	533	396
Sensitivity	84%	28%	29%	23%	37%	40%
cFDR	10%	2%	26%	30%	41%	NA*

Table S1. Mass Spectrometry Studies of Mammalian Mitochondria.

Proteins detected in each study were mapped to mouse Entrez gene identifiers and assessed using training data of 591 mitochondrial genes and 2519 non-mitochondrial genes. We note that mapping between protein annotations and between species is difficult and we failed to map some detected proteins to mouse Entrez genes. Asterisk (*) indicates published data that contributed to the gene ontology “mitochondrial” annotation, which were excluded from the non-mitochondrial training set, and thus cannot be used to calculate cFDR.

Dataset	Bin	logodds	specificity	sensitivity	Tmito found	T~mito found
MitoDomain	M+	5.8	99%	47%	280	21
MitoDomain	M+/-	0.8	82%	30%	179	445
MitoDomain	M-	-4.0	34%	4%	25	1671
MitoDomain	null	0.3	85%	18%	107	382
Coexpression	30	8.8	100%	18%	107	1
Coexpression	20	5.4	100%	7%	41	4
Coexpression	10	4.0	100%	6%	34	9
Coexpression	5	2.8	99%	10%	58	36
Coexpression	<5	-1.0	19%	40%	239	2052
Coexpression	null	0.2	83%	19%	112	417
Mass Spectrometry	75-100pure	8.0	100%	20%	119	2
Mass Spectrometry	50-75pure	7.5	100%	28%	165	4
Mass Spectrometry	25-50pure	5.5	100%	14%	84	8
Mass Spectrometry	0-25pure	3.2	100%	2%	11	5
Mass Spectrometry	75-100ambig	4.3	100%	2%	9	2
Mass Spectrometry	50-75ambig	3.7	100%	6%	34	11
Mass Spectrometry	25-50ambig	1.9	98%	7%	43	48
Mass Spectrometry	0-25ambig	-0.8	89%	6%	38	281
Mass Spectrometry	75-100crude	-0.7	100%	0%	1	7
Mass Spectrometry	50-75crude	-1.1	98%	1%	5	47
Mass Spectrometry	25-50crude	-1.0	97%	1%	8	69
Mass Spectrometry	0-25crude	-3.9	95%	0%	2	128
Mass Spectrometry	null	-2.6	24%	12%	72	1907
PGC induction	3	3.7	99%	13%	79	26
PGC induction	2.5	3.4	99%	9%	54	22
PGC induction	2	2.4	98%	9%	52	41
PGC induction	1.5	2.0	96%	17%	99	104
PGC induction	<1.5	-0.9	17%	44%	259	2096
PGC induction	null	-0.2	91%	8%	48	230
Rickettsia Ancestry	<1e-10	2.4	94%	34%	202	162
Rickettsia Ancestry	<1e-5	0.1	100%	0%	1	4
Rickettsia Ancestry	null	-0.5	7%	66%	388	2353
TargetP	1	9.0	100%	20%	120	1
TargetP	2	4.8	99%	23%	135	21
TargetP	3	2.8	99%	10%	59	36
TargetP	4	1.7	98%	7%	40	52
TargetP	5	0.7	97%	4%	26	70
TargetP	null	-1.4	7%	36%	211	2339
YeastMitoHomolog	1	4.5	98%	38%	226	42
YeastMitoHomolog	0	-0.7	2%	62%	365	2477

Table S4. Log-likelihood Scores for Seven Genome-Scale Data Sets.

Log-likelihood scores for each of the seven genome-scale data sets were calculated at the predefined ranges listed. We counted the number of T_{mito} and $T_{\sim\text{mito}}$ genes with scores within each range (Tmito found, T~mito found columns) to compute the sensitivity, specificity and log-likelihood odds ratio (base 2) as defined in Experimental Procedures.

	mRNA features	MitoCarta genes	Mouse genes	pvalue	motif(s)
A	5' UTR length (nt)	83	129	<2e-16	
	CDS length (nt)	966	1062	2e-12	
	3' UTR length (nt)	484	631	1e-8	
	mRNA abundance	322	167	<2e-16	
B	**Promoter features**				
	CpG	72%	41%	4e-101	
	TATA	9%	19%	3e-18	
C	**Conserved promoter motifs**				
	ERR1/SF1	10%	4%	8e-17	TGACCTY_V_ERR1_Q2/TGACCTTG_V_SF1_Q6
	SP1	14%	10%	1e-5	GGGCGGR_V_SP1_Q6
	NRF1	5%	3%	8e-5	RCGCANGCGY_V_NRF1_Q6
	NRF2/ELK1	8%	4%	3e-10	V_NRF2_01/SCGGAAGY_V_ELK1_02
	MYC/USF	6%	3%	3e-3	CACGTG_V_MYC_Q2/V_USF_Q6
	▶ NF-Y	6%	4%	1e-2	GATTGGY_V_NFY_Q6_01
	▶ TMTCGCGANR	1%	0.5%	1e-3	TMTCGCGANR_UNKNOWN
	▶ ACTAYRNNNCCCR	3%	1%	2e-2	ACTAYRNNNCCCR_UNKNOWN

Table S6. MitoCarta Transcript and Promoter Features.

(A) Frequency of genomic features in MitoCarta genes compared to all mouse genes, with significance assessed by Mann Whitney test. mRNA abundance was measured by median expression value within the mouse GNF tissue atlas.

(B) Frequency of promoter features in MitoCarta genes compared to all mouse genes, with significance assessed by the Bonferroni corrected hypergeometric distribution.

(C) Frequency of conserved motifs in MitoCarta promoters compared to promoters of all mouse genes, with significance assessed by the Bonferroni corrected hypergeometric distribution. We searched for enrichment within all 68 motifs from MSigDB for which conserved motif instances were found in ≥ 15 MitoCarta promoters. Mouse MitoCarta genes were associated with conserved motifs using the MSigDB annotation. Listed are the eight unique motifs with corrected pvalues < 1e-2. Since MSigDB contains redundant motifs, we collapsed sequence motif variants and reported results for the genes containing any of the motif variants listed (E.G. MYC/USF indicates promoters containing CACGTG_V_MYC_Q2 or V_USF_Q6).

Supplemental Experimental Procedures

Mitochondrial Purification: C57BL/6 mice were euthanized by CO_2 asphyxiation followed by cervical dislocation. Organs were quickly excised and placed into ice-cold isolation buffer (buffer A: 220 mM mannitol, 70 mM sucrose, 5 mM HEPES-KOH, pH 7.4, 1 mM EGTA-KOH). All subsequent steps were performed either on ice or at 4°C in a cold room. Tissues were carefully trimmed and washed twice in isolation buffer A and then twice in buffer B (buffer A plus 1.0 mg/ ml bovine serum albumin (BSA) and protease inhibitor cocktail (Roche Complete EDTA free tablets)). Tissues were resuspended to ~ 0.2 mg/ml in buffer B and homogenized with four strokes of a Potter-Elvehjem glass/Teflon homogenizer attached to a Eurostar power-visc motorized stirrer set at 1000 rpm. The homogenate was diluted up to 20 ml with buffer B and immediately decanted into a 45 ml Parr bomb attached to a pressurized N_2 tank. While stirring at medium speed, the bomb was pressurized to 800 psi for 10 min. After rapid depressurization, the tissue homogenate was decanted into a 50 ml conical tube (homogenates from stomach, large intestine and small intestine were stirred for 10 min with DEAE celluose (1.0 mg/ 10 ml), 75 U/ml heparin, and 2 mM DTT) [1]. A small sample was taken (whole tissue lysate- W), and the remaining homogenate was centrifuged at 1000 rpm for 10 min. The supernatant was carefully decanted and saved. The pellet was transferred to a Kontes glass/ teflon homogenizer and rehomogenized in buffer B with 2 strokes at 1000 rpm as above. This second homogenate was again centrifuged at 1000 rpm for 10 min. The supernatants from the two spins were combined, filtered through nylon mesh, and spun at 8,000 x g for 10 min. The supernatant was decanted and any loose or lighter colored material surrounding the pellet was suctioned away. A small amount of buffer B (200- 300 µl) was then added to the tube and the pellet was carefully dislodged using a polished glass rod. The pellet was thoroughly resuspended until homogeneous. A small sample was taken (crude mitochondria- C). The remaining mitochondria were carefully layered on top of a stepwise density gradient of 0.5 ml 80%, 1.5 ml 52%, and 1.5 ml 26% Percoll in a 4 ml Beckman Ultra-Clear centrifuge tube (note: Percoll is supplied as a 23% colloidal suspension in water. Here, this solution is considered "100% Percoll." This solution was made isotonic by the addition of 1 part 5X buffer A to 4 parts 100% Percoll to create 80% Percoll (aka SIP- stock isotonic Percoll). 52% and 26% Percoll were made by further diluting SIP with 1X buffer A). The gradient was spun at 20,000 rpm for 45 min in a Beckman SW 60 Ti rotor. Mitochondria were collected from the interface of the 26%52% interface, diluted to capacity in a 2 ml eppendorf tube with buffer A, and spun at full speed in a refrigerated table top centrifuge for 10 min. The supernatant was carefully discarded, and the mitochondria were washed with an additional 2 ml of buffer A and respun. The resulting pellet was resuspended in a small volume of buffer A. A small sample was taken (P- pure mitos). The protein concentration of the remaining mitochondrial solution was determined by Bradford assay, and the mitochondria were divided into eppendorf tubes in 100 μg aliquots and flash frozen in liquid nitrogen.

The use of animals outlined in this proposal has been reviewed and approved by the Massachusetts General Hospital (MGH) Subcommittee on Research Animal Care (Protocol #: 2005N000332/2).

SDS-PAGE and In-Gel Protein Digestion Protocol: 100 μg of protein sample was dissolved in 1% SDS and 100 mM NH_4CO_3 to a final volume of 20 μL and subjected to reduction in 10 mM DTT for 45 min at 37°C, followed by alkylation in 55 mM iodoacetamide for 60 minutes at room temperature. Each sample was separated by protein molecular weight using a 4-12% bis-Tris gradient SDS-PAGE gel and MES-SDS running buffer (Invitrogen, Carlsbad, CA). Gels were stained with SimplyBlue™ Safestain (Invitrogen) and each lane was excised into 20 slices and placed in separate tubes. Each gel slice was further excised into 1 mm cube pieces and incubated in a solution of 100 mM NH_4CO_3 in 50% acetonitrile/50%water for Coomassie stain extraction and gel dehydration. After removal of the extraction solution, gel pieces were immersed in a solution of 100 mM NH_4CO_3 in water containing 1 μg trypsin (modified sequencing grade, Promega, Madison, WI) and incubated at 37°C overnight for in-gel tryptic digestion. Digested peptides were extracted from gel pieces by vortexing the gel pieces in 75 μL of peptide extraction buffer (1% trifluoroacetic acid in 70% acetonitrile/30% water); the solution was removed from the gel pieces and reserved. 25 μL of rehydration buffer (1% trifluoroacetic acid in water) is then added to the gel pieces. The cycle of peptide extraction and rehydration was performed two more times and the extracted material was dried by vacuum centrifugation and reconstituted in 6.5 uL of 3% acetonitrile/5% formic acid in water.

In-Solution Protein Digestion Protocol: 100 μg of protein sample was dissolved in 7 M Urea/ 2 M Thiourea and 100 mM NH_4CO_3 to a final volume of 20 μL and subjected to reduction in 10 mM DTT for 45 min at 37°C, followed by alkylation in 55 mM iodoacetamide for 60 minutes at room temperature. Each sample was incubated with 1 μg Lys-C (Roche Diagnostics, Indianapolis, IN) for 4 hrs at 37°C and then diluted 9-fold in a solution of 100 mM NH_4CO_3 in water. 1 μg of trypsin was added and the sample was incubated at 37°C overnight. The digested sample was concentrated and desalted using an Oasis ® HLB 5mg cartridge (Waters Corporation, Milford, MA). Peptides were eluted from the HLB cartridge with 1 mL of 0.1% formic acid in 70% acetonitrile /30% water and dried by vacuum centrifugation and reconstituted in 100μL of 3% acetonitrile/5% formic acid in water.

Reverse Phase Chromatography and Mass Spectrometry: Peptide samples prepared from in-gel digestion were analyzed by reverse phase liquid chromatography tandem mass spectrometry (LC-MS/MS) using an LTQ-Orbitrap (Thermo Scientific, San Jose, CA) equipped with a custom nanospray source (James A. Hill Instrument Services, Arlington, MA). 6 µL of each 6.5 µL peptide extract was injected on to a 15 cm ReproSil-Pur C18-AQ 3 µm reverse phase resin (Dr. Maisch GmbH, Ammerbuch-Entrigen, Germany) packed into a 15 µm tip, 75 µm inner diameter fused silica column prepared in-house and analyzed by a linear gradient of mobile phase A (0.1% formic acid in water) and mobile phase B (0.1% formic acid in 90% acetonitrile/10% water), performed from 5% B/95% A to 65% B/35% A in 70 minutes at 200nL/min using an 1100 Series Nanoflow Pump (Agilent Technologies, Santa Clara, CA). Data dependent MS/MS were collected in the LTQ for the top ten most intense ions observed in the LTQ-Orbitrap survey scan, taken over an m/z range of 300 to 1800 at resolution 60,000 (for m/z at 400) and lock mass injection of polydimethylcyclosiloxane for internal calibration at m/z 445.120025. Dynamic exclusion was employed and set to a precursor repeat selection of 2 within 20 seconds and exclusion for 60 seconds, with an early expiration from the exclusion list set for a detection of a signal to noise threshold of 4 in 2 consecutive scans.

Peptide samples prepared from in-solution digestion were also analyzed by LC-MS/MS. 2 µL of each 100 µL peptide sample was injected on to a reverse phase column as described above, and analyzed using a linear reverse phase gradient from 5% B/95% A to 65% B/35% A in 100minutes. Dynamic exclusion was set to a set to a precursor repeat selection of 1 within 20 seconds and exclusion for 60 seconds, with an early expiration from the exclusion list set for a detection of a signal to noise threshold of 4 in 2 consecutive scans.

Mass Spectra Database Matching and Validation: Using Spectrum Mill™ Rev B.03.03.070 Extractor (Agilent, Santa Clara, CA), all captured MS/MS spectra were filtered (requiring sequence tag length > 1) and redundant spectra (identical m/z observed within a 40 second window) were merged. Spectrum Mill search parameters included:

- carbamidomethylation of cysteines,
- 50% minimum matched peak intensity,
- precursor mass tolerance of +/- 0.05 Da for orbitrap scans,
- product ion mass tolerance of +/- 0.7 Da,
- $\leq$ 3 missed tryptic cleavages,
- variable modification mode - allowing for multiple modifications of oxidized methionine, conversion of glutamine to pyroglutamic acid, deamidation of asparagine, acetylated lysine, phosphorylation of tyrosine, threonine and serine.

- reverse database search, based on the target-decoy strategy [2] and described below.

MS/MS spectra matched via these search parameters were marked as valid using Spectum Mill Autovalidation using 4 parameters:

1. charge state (+1, +2, +3, or +4)
2. scored peak intensity (SPI)
3. peptide scores (representing the match between the observed spectra and a theoretical spectrum from the input protein database), where *rank1* indicates the score of the best matching peptide and *rank2* indicates the score of the second best matching peptide.
4. reverse database peptide scores (representing the match between the observed spectra and a theoretical spectrum from the reversed input protein database, where each sequence is reversed), where *revrank1* indicates the score of the best matching peptide from the reverse database

Proteins were searched in Spectrum Mill *protein details* and *peptide* mode as described below. In *protein details* mode, a protein group must have an additive peptide score of ≥ 20, where individual peptides must meet the following criteria:

1. Peptides of charge state +1, +3 or + 4 must have:*rank1* ≥8 **and** SPI ≥70% **and** *rank1-revrank1* ≥2 **and** *rank1-rank2* ≥2.
2. Peptides of charge state +2 must have: [*rank1* ≥8 **and** SPI ≥70% **and** *rank1-revrank1* ≥2 **and** *rank1-rank2* ≥2] **or** [*rank1* ≥6 **and** SPI ≥90% **and** *rank1-revrank1* ≥1 **and** *rank1-rank2* ≥1]

In *peptide* mode, a peptide must have: *rank1* ≥13 **and** SPI ≥ 70% **and** *rank1-revrank1* ≥ 2 **and** *rank1- rank2* ≥2.

Discovery Proteomics of Purified Mitochondrial Extracts: Pure mitochondria from 14 mouse tissues were analyzed according to the in-gel digestion protocol described above. A total of 285 LTQ-Orbitrap instrument files for 280 gel slices were collected, representing 4,475,842 tandem mass spectra. In five instances due to instrument failure, analyses were continued without re-injection of sample; data from these partial analyses were included in this study. The instrument files for all gel slices and tissues were concatenated prior to database search. Spectrum Mill Extractor merging of identical spectra yielded 4,241,196 spectra and quality filtering yielded 1,643,930 spectra. These spectra were then searched against the mouse Refseq database as described above. Validation was performed using protein grouping in *protein details* mode, enabling the validation of peptides grouped across all 14 tissue mitochondrial samples. This allows for the consideration, for example, of protein identifications that do not meet the validation criteria in an individual tissue. The validation in *peptide* mode was then performed as described above. Validation across all 14 tissue mitochondria yielded 416,432 validated spectra and 48,206 unique peptide identifications. In addition, a technical replicate of liver samples was performed.

Corrected false discovery rates (cFDR) are shown in Figure S16 for each tissue, based on our training sets.

Subtractive Proteomics of Crude and Pure Mitochondrial Extracts: Crude and pure mitochondria from ten tissues were purified and in-solution digestion was performed as described above. Four technical replicates of crude and pure cerebellum samples were performed, while single samples were analyzed for the other nine tissues (cerebrum, brainstem, spinal cord, kidney, liver, heart, skeletal muscle, testis and placenta). Spectra from each tissue were separately searched from databases matches using SpectrumMill. We considered proteins to be crude-enriched if they were found only in crude extracts, or if they were found at ≥ twofold higher peak intensity in crude extracts compared to pure in a tissue (and similarly for pure-enriched). For the few genes found to be crude-enriched in some tissues and pure-enriched in others, the label was assigned based on a majority vote. Genes not detected by subtractive proteomics, or those with under twofold difference in protein abundance, were considered inconclusive in terms of enrichment.

Mass Spectrometry Protein Abundance Measurements: We used two estimates of protein abundance generated by SpectrumMill: coverage (the percent of protein residues identified by MS/MS spectra) and total peak intensity (the sum of peak areas from extracted ion chromatograms (XIC) for all sequence identified peptides). While coverage is useful for estimating relative abundance between proteins, it quickly saturates for abundant molecules. In contrast, total peak intensity correlates with relative abundance of a given protein across samples, but it is difficult to compare between proteins since the number of potential peptides can vary widely. Thus we use coverage for cross-protein comparisons and total peak intensity for cross-sample comparisons. Total peak intensity values were $\log_{10}$ transformed in data provided in Supplemental Data and Tranche.

Mass Spectrometry Data Access: MS RAW files and identified peptides for both the discovery and subtractive proteomics are available in Tranche (www.proteomecommons.org/dev/dfs). In addition, peptide level information is provided in Table S3 and protein level information is provided in Table S5 (aggregated by gene locus, with each gene inheriting the highest score of any splice form).

Integration of Genome-Scale Datasets: Genome-scale datasets of mitochondrial localization were created and integrated using a Naïve Bayes approach as previously described [3]. Below we provide specific details on the scoring of these datasets.

Proteomics: one of 12 categories combining discovery proteomics abundance (coverage) with subtractive proteomics (Figure 2D) or NA if not detected

Targeting sequence: 0 if no mitochondrial targeting signal was detected by TargetP v1.1 [4], otherwise a confidence score of 1-5 (1 is most confident)

Protein domain: Following MitoPred's methodology [5] for identifying mitochondrial domains, we used ~99,000 SwissProt [6] eukaryotic proteins (release 54.1) containing annotations for 'subcellular location', which we filtered (excluding low-confidence annotations containing 'by similarity', 'potential', 'probable', and 'possible') and partitioned into two sets: S_{mito} containing 3,852 mitochondrial proteins and $S_{\sim mito}$ containing 27,873 non-mitochondrial proteins. Pfam domains for each Swissprot protein were obtained from Swissprot annotations. We assigned each Pfam domain a categorical score (M+, M-, M±, NA) based on whether the SwissProt proteins containing the domain were exclusively from S_{mito}, exclusively from $S_{\sim mito}$, found in both, or not present in either set. Each mouse gene received a categorical score based on the best score of any of its Pfam domains. Protein domains for all mouse proteins were determined using HMMER (with expect parameter=0.1 and using pfam TC trusted threshold cutoffs) to search 7,973 Pfam domains (ftp.sanger.ac.uk/pub/databases/Pfam/current_release, 11/22/2006). Note that for cross-validation studies, all mouse proteins were removed from S_{mito} to avoid overestimating sensitivity.

Yeast homology: 1 if the best *S.cerevisiae* homolog (BlastP expect < 1e-3, coverage >50% of longer gene) was annotated as mitochondrial (825 mitochondrial genes within the Saccharomyces Genome Database, 12/27/06), 0 otherwise

Ancestry: best BlastP expect value from *R. prowazekii* homolog, or NA if expect > 1e-3

Transcriptional coexpression: Genes were assigned a score 0-50 based on transcriptional co-expression with Tmito across the GNF1 survey of gene expression across 61 mouse tissues [7] (GEO accession GSE1133). The score represents the number of T_{mito} genes found within a gene's 50 nearest neighbors (Euclidean distance) [8]. Probe set IDs were mapped to RefSeq IDs via SymAtlas (symatlas.gnf.org, 12/28/06), excluding probe sets matching more than one protein. Mouse genes were assigned the highest score of any splice form, or NA if not available. Microarray rows were clipped to smooth low intensity values (any expression level < 20 was replaced with 20) and normalized to mean=0, variance=1. Rows with no post-normalization value > 1.5 were excluded. A total of 15,778 mouse genes had probes meeting the filtering requirements.

Transcriptional activation during mitochondrial proliferation: We used time-course microarray studies (GEO accession GSE4330) to assess gene transcription in mouse myoblasts during overexpression of PGC-1α [3,8], which is known to induce mitochondrial proliferation. Expression intensities were sample normalized via linear fit to the median scan. Each gene was assigned an induction score measured in fold-change; dividing average intensity in PGC-1α treated cells (average of replicates on days 2,3) by average intensity in GFP control cells. Only those probes showing significant difference between case and control (p<0.05, measured by 1-tailed heteroscedastic student t-test) were considered (3,438 probe-sets corresponding to 2,944 genes).

The scores for each method are available for all 23,640 mouse genes in Table S5C. The dataset scores were converted to log-likelihood ratios at predefined ranges shown in Table S7. In calculating log-likelihood values, we added small pseudocounts (1e-4) in order to avoid mathematical errors when the denominator was zero.

Epitope-Tagging with GFP and Microscopy: cDNAs cloned into the Gateway entry vector pDONR 223 were obtained from the Human Orfeome collection and robotically arrayed into BioRad HardShell 96 well plates. (Rual et al., Genome Research 2004) These clones were recombined into the C-terminal GFP Gateway Destination vector pcDNA6.2/C-EmGFP-DEST (Invitrogen) in 96 well plate format by adding 0.5 µL of LR clonase recombination enzyme (Invitrogen) to wells containing 1 µL of destination vector DNA (75 ng/µL), 1.5 µL entry vector, 1 µL of 5x LR buffer, and 1 µL of TE. These mixtures were incubated overnight at 25 °C.

Competent DH5α cells were arrayed in Marsh 96-well plates in 8 µl aliquots and transformed with 2 µl of the LR reaction mixture. After 30 minutes of incubation on ice, these plates were heat shocked for 45 seconds at 42 °C using a thermal cycler and incubated on ice for 5 minutes. We then added 105 µL of SOC to each well, incubated the plates at 37 °C and 225 rpm for 1 hour, and then transferred the entire transformation mixture to deep-well plates containing 1 mL of LB with ampicillin. Cultures were grown overnight at 37 °C and 225 rpm. DNA was extracted using the PrepEase 96-well Plasmid Kit (USB Corp.).

The quality of these GFP-labeled clones was then assessed using three independent metrics. First, DNA concentrations and OD260/280 ratios were measured using a Nanodrop spectrophotometer. 418 out of 470 clones (89%) had both a yield of > 1.0 µg of DNA and an OD 260/OD 280 ratio between 1.6 and 2.2, indicating a high overall quality of DNA preparations. In addition, 60 clones (12 clones per 96 well plate) were selected and subjected to PCR analysis using a T7 forward primer and a GFP reverse primer. Analysis of the PCR products by gel electrophoresis revealed that 56/60 clones (93%) had the correct insert size. Finally, the same 60 clones were analyzed by forward and reverse sequencing using the T7 and GFP primers. We assessed our sequences for the presence of an intact start codon, the absence of a premature stop codon before the C-terminal GFP, and for the correct identity of the gene insert by BLAST. 57 out of 60 clones (95%) passed all sequencing criteria.

Approximately $4x10^3$ HeLa cells in 100 µL medium (DMEM with 10% FBS, 1x GPS) were seeded in Falcon 96-well Imaging Plates using a MultiDrop Combi robot and incubated at 37 °C in a humidified 5% CO2 atmosphere (Thermo Scientific). The following day, the medium was replaced and transfection reagents were added to each well. We created our transfection reagents by first diluting Lipofectamine LTX (Invitrogen) 1:10 in OptiMEM I Reduced Serum Medium (Invitrogen). 2 µL of this solution was then added to a mixture of 1 µL DNA that had been diluted with 17 µL OptiMEM I. The complete solution was incubated at RT for 30 minutes prior to transfection.

48 hours post transfection, we stained our cells by replacing the media with our staining solution, which consisted of full medium with 50 nM MitoTracker Red CMXRos and 1:1000 diluted Hoechst 33258 (Molecular Probes). Cells were stained in this solution for 30 minutes at 37 °C in a humidified 5% CO2 atmosphere. Following staining, the cells were washed twice with medium and then fixed with 3.7% formaldehyde in PBS for 15 minutes at RT. The cells were then washed twice more with PBS before continuing with microscopy.

Fluorescence microscopy was performed with a 63x oil-immersion objective on a Zeiss wide-field microscope. Multiple images using DAPI, FITC, and Texas Red filters were captured for the constructs using a 12-bit CCD camera and reviewed for colocalization of GFP and MitoTracker red signals.

Creation of MitoCarta

MitoCarta was created as the union of the 951 Maestro predictions (scores $\geq$ 4.56, corresponding to 10% cFDR), the 131 genes validated by GFP-tagging and microscopy, and the 591 T_{mito} genes with experimental evidence of mitochondrial localization from literature. The full MitoCarta dataset is available at www.broad.mit.edu/publications/MitoCarta. We estimate the completeness of MitoCarta by the following steps. Our Maestro set includes 498/591 T_{mito} genes (84% sensitivity) and 453 predictions with an estimated 29.6% cFDR, indicating that 319 predictions should be *bona fide* mitochondrial genes (note that the 29.6% cFDR of predictions exceeds the 10% cFDR for training data due to their lower average log-likelihood scores). Based on this sensitivity, in addition to T_{mito} there should be an additional 538 mitochondrial genes (538=453/84%) – or a total of 1129 *bona fide* mitochondrial genes (1129=591+538). We validate by GFP-tagging 54 genes not Maestro predicted. Thus out of the estimated 1129 *bona fide* mitochondrial genes, we capture 964 (591 + 319 + 54) or 85% in MitoCarta, and we miss an estimated 165 genes. We estimate that the remaining 134 Maestro predictioned genes (1098-964=134) are false positives, and thus the estimated false discovery rate of MitoCarta is ~10%.

Transcript and Promoter Analysis: Promoter features for RefSeq NM annotations (UCSC mm8 assembly, 8/23/07) were annotated as follows: CpG islands within +/- 200bp of transcription start (UCSC annotation); TATA motif weight matrix [9] within +/- 50bp of transcription start. mRNA abundance was measured by median expression value within the mouse GNF tissue atlas. Transcription factor motif assignments to human gene promoters were downloaded from MSigDB [10] based on 4-species conservation, and mapped to mouse via HomoloGene. All 67 motifs with ≥15 MitoCarta members were then tested for enrichment in MitoCarta promoters compared to all gene promoters using the hypergeometric distribution, with Bonferroni correction for multiple hypotheses (Table S10).

Concordance of mRNA Expression and Protein Expression: MitoCarta protein expression values were compared to mRNA expression values from matched tissues using the pairwise concordance statistic developed previously [11]. mRNA expression data (gcRMA) from the GNF1M tissue atlas [7] was obtained for our fourteen tissues by calculating the mean expression value for technical replicates of the matching tissues (hypothalamus and substantia.nigra were averaged together for "brainstem", and spinal.cord.lower and spinal.cord.upper were averaged together for "spinal cord"). Microarray rows were clipped to smooth low intensity values (any expression level < 40 was replaced with 40) and rows with no value > 40 were excluded. The protein expression (peak intensity) was quantile normalized across tissues based on total peak intensity for all MitoCarta genes detected in each tissue. Next, the protein expression values were divided by the protein length (number of amino acids). We cannot compare our proteomics expression data (generated per mass of mitochondria) directly to mRNA expression (generated per tissue) since different tissues have extremely different mitochondrial quantity. Thus, using the ELISA assay results described in Experimental Procedures, we normalize our protein expression values by the mitochondrial quantity per tissue. For each gene, we computed the pairwise concordance statistic (number of pairs of tissues for which mRNA and protein expression values are concordant, divided by number of all possible tissue pairs). A pair of tissues (*a, b*) is concordant if [(mRNA(*a*) > mRNA(*b*)) and (prot(*a*) > prot(*b*))] or [(mRNA(*a*) < mRNA(*b*)) and (prot(*a*) < prot(*b*))], where mRNA(a) and prot(a) represent the normalized mRNA and protein levels respectively. Using the two technical replicates for each tissue from the mRNA atlas, we also computed pairwise concordance for mRNA duplicates. By the above definition, we find over 60% pairwise concordance between protein and mRNA expression [7] across the fourteen tissues. This is particularly high given that technical replicates of mRNA expression show only 80% pairwise concordance.

Eukaryotic phylogeny: The shown phylogeny of eukaryotic species in Figure 6C is based on a phylogenetic reconstruction method previously described [12]. The resulting tree was robust to several phylogenetic reconstruction methods using different aligned protein sequences and using small subunit RNA, except for the positioning of three deep-branching protest species: *E. histolytica, G. lamblia,* and *E. cuniculi.* These branches also received low scores in bootstrap analysis. Because of the uncertainty in their phylogeny, we didn't count the loss of CI in *E. histolytica* as a separate evolutionary loss event even though this would be implied in the phylogeny displayed in Figure 6C.

References

1. Lawrence, C.B. & Davies, N.T. A novel, simple and rapid method for the isolation of mitochondria which exhibit respiratory control, from rat small intestinal mucosa. *Biochim Biophys Acta* **848**, 35-40 (1986).
2. Elias, J.E. & Gygi, S.P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat Methods* **4**, 207-14 (2007).
3. Calvo, S. et al. Systematic identification of human mitochondrial disease genes through integrative genomics. *Nat Genet* **38**, 576-82 (2006).
4. 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**, 1005-16 (2000).
5. Guda, C., Fahy, E. & Subramaniam, S. MITOPRED: a genome-scale method for prediction of nucleus-encoded mitochondrial proteins. *Bioinformatics* **20**, 1785-94 (2004).
6. Wu, C.H. et al. The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Res* **34**, D187-91 (2006).
7. Su, A.I. et al. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* **101**, 6062-7 (2004).
8. 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 U S A* **101**, 6570-5 (2004).
9. Bucher, P. Weight Matrix Descriptions of Four Eukaryotic RNA Polymerase II Promoter Elements Derived from 502 Unrelated Promoter Sequences. *J. Mol. Biol.* **212**, 563-578 (1989).
10. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-50 (2005).
11. Mootha, V.K. et al. Integrated analysis of protein composition, tissue diversity, and gene regulation in mouse mitochondria. *Cell* **115**, 629-40 (2003).
12. Ciccarelli, F.D. et al. Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**, 1283-7 (2006).

Appendix C

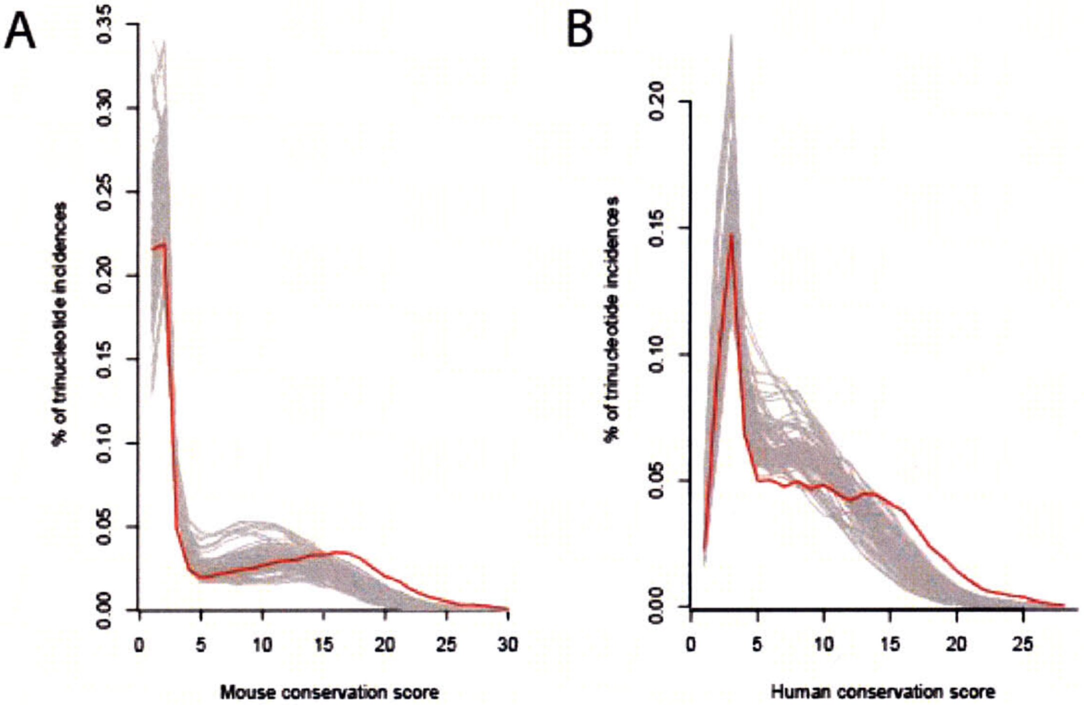
—

Supplementary Material for Chapter 4

Supplementary Data: Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans

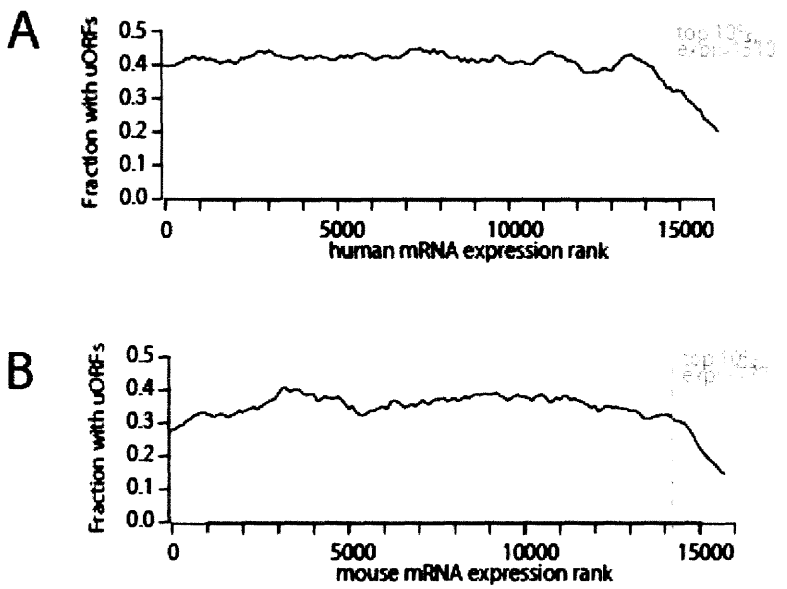
Supplementary Figures, Tables, and Notes

Below are Supplementary Figures S1-S6, Supplementary Tables S1, S2, S4, and Supplementary Notes. Supplementary Table S3 is a large excel file that is available online.



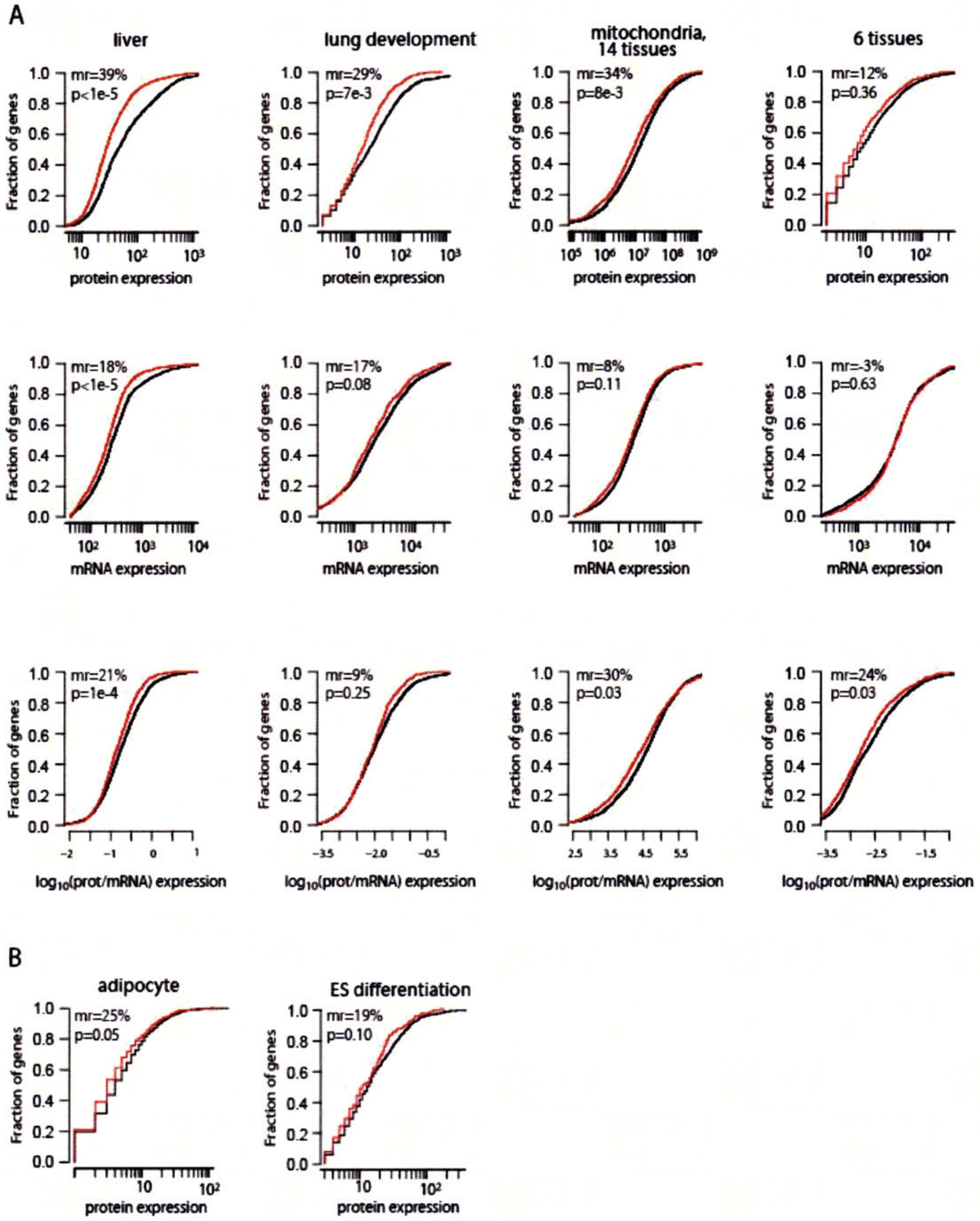
Supplementary Figure 1: uAUG conservation.

(*A-B*) Histograms of trinucleotide conservation scores are plotted for all 64 trinucleotides (gray curves) and for the ATG trinucleotide (red curve) within mouse (*A*) and human (*B*) 5' UTRs. The conservation score is the number of species with the given trinucleotide aligned and conserved within 5' UTR multiple alignments of 30 species (*A*) or 28 species (*B*). Note that >97% of ATG trinucleotides define a uORF, whereas the remaining ~3% of ATGs are in-frame with the main coding sequence without an intervening stop codon.

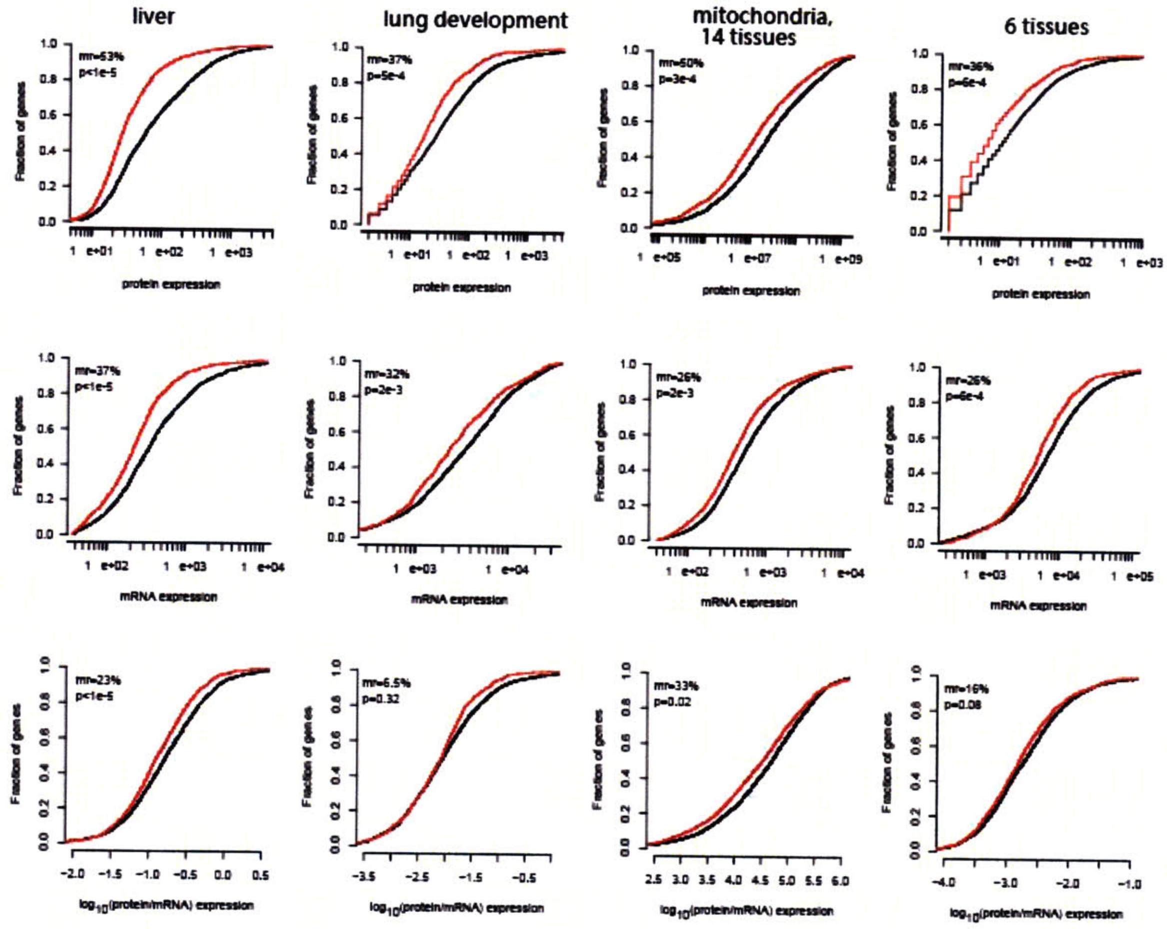


Supplementary Figure 2: Distribution of uORFs by mRNA expression.

Plotted are the distributions of genes with uORFs based on mean mRNA expression across 79 human tissues (*A*) and 61 mouse tissues (*B*) from the GNF tissue atlas, using sliding windows of 1000 genes. X-axis shows rank of mRNA expression from lowest (left) to highest (right), and Y-axis shows the fraction of genes in the window that contain uORFs. Dashed lines indicate the 90th percentile, which is equivalent to expression values of 1510 and 777 for human and mouse atlases, respectively.

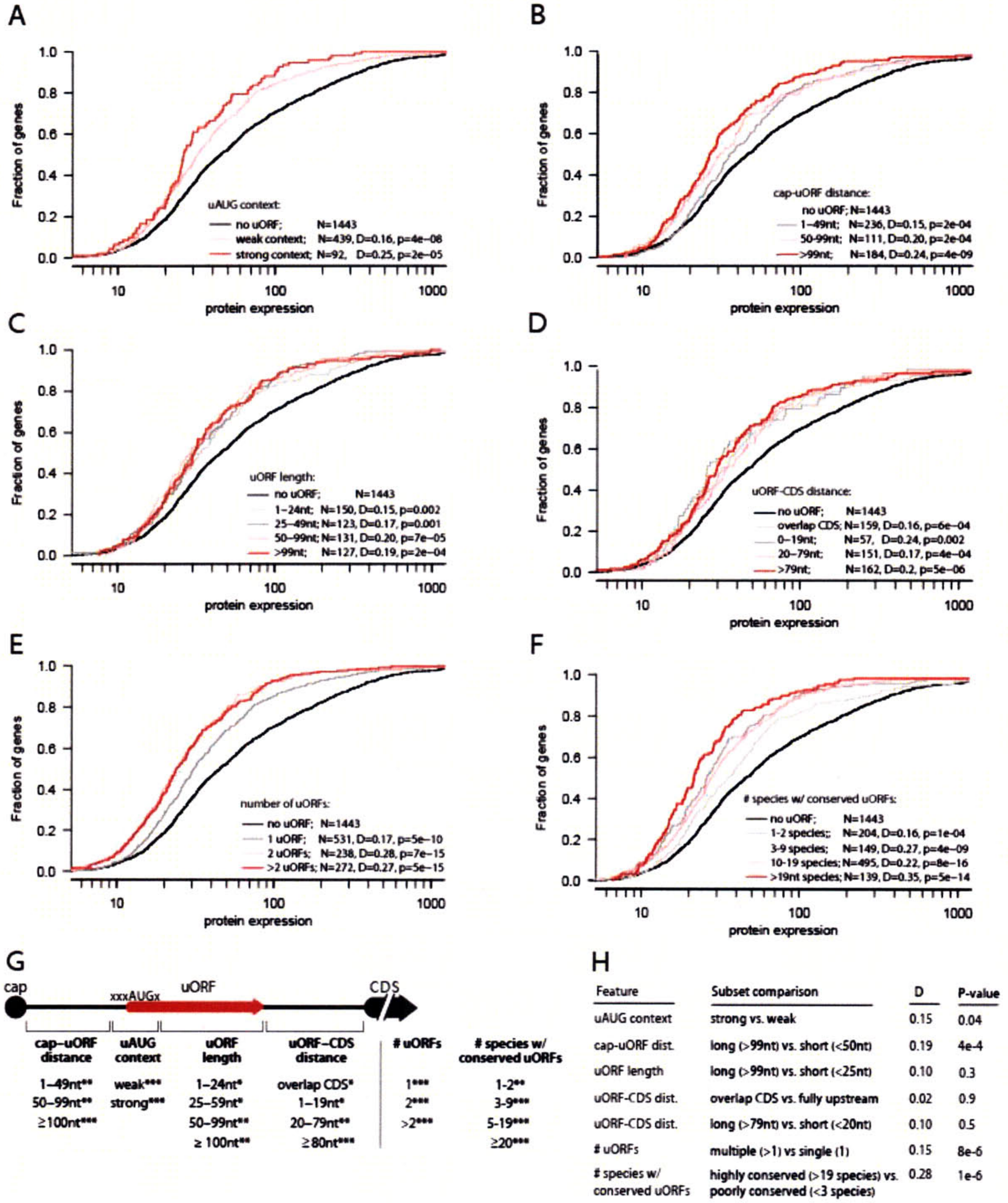


Supplementary Figure 3: Cumulative distribution of expression values.
Plotted are the empirical cumulative distribution of expression of genes with uORFs (red curves) and genes without uORFs (black curves) for six independent datasets. Statistic mr indicates the median reduction of expression from uORF-containing vs. uORF-less genes, with p-values computed by permutation testing (10,000 permuations of uORF labels). (*A*) Expression of proteins (first row), mRNA (second row), and protein/mRNA ratios (third row) are shown for genes detected by MS/MS and microarrays in 4 studies (columns). (*B*) Expression of proteins detected by MS/MS in 2 additional studies of mouse adipocyte cells and embryonic stem cell differentiation.



Supplementary Figure 4: Cumulative distribution of expression values, all genes. Plotted are the empirical cumulative distribution of expression of genes with uORFs (red curves) and genes without uORFs (black curves) for four independent datasets. Unlike Fig. S3, these plots include the 10% most highly expressed genes. Statistic mr indicates the median reduction of expression from uORF-containing vs uORF-less genes, with p-values computed by permutation testing (10,000 permuations of uORF labels). Expression of proteins (first row), mRNA (second row), and protein/mRNA ratios (third row) are shown for genes detected by MS/MS and microarrays in 4 studies (columns).

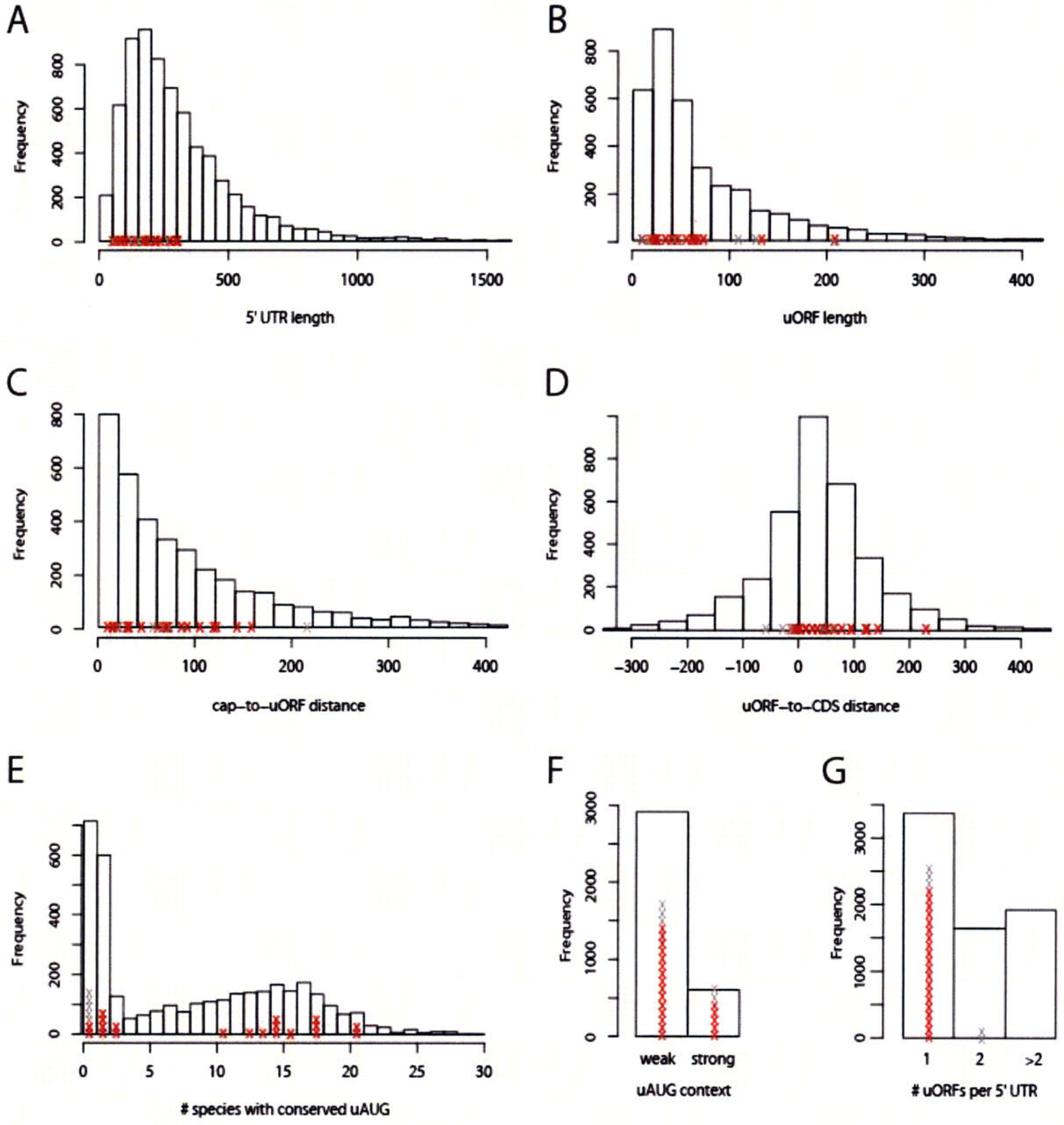
As discussed in the Supporting Information, the bias against uORFs in the most highly expressed genes causes a shifted distribution of mRNA expression for uORF-containing genes.



Feature	Subset comparison	D	P-value
uAUG context	strong vs. weak	0.15	0.04
cap-uORF dist.	long (>99nt) vs. short (<50nt)	0.19	4e-4
uORF length	long (>99nt) vs. short (<25nt)	0.10	0.3
uORF-CDS dist.	overlap CDS vs. fully upstream	0.02	0.9
uORF-CDS dist.	long (>79nt) vs. short (<20nt)	0.10	0.5
# uORFs	multiple (>1) vs single (1)	0.15	8e-6
# species w/ conserved uORFs	highly conserved (>19 species) vs. poorly conserved (<3 species)	0.28	1e-6

Supplementary Figure 5: Effects of uORF properties on liver protein levels. (*A-F*) Cumulative distributions of protein expression are shown for mouse gene loci lacking uORFs (black curve) or with uORFs of different properties (red, pink, and gray curves). In each plot, the legend contains the feature name, the number of gene loci in each subset (N), the one-sided Kolmogorov-Smrinov (KS) statistic (D) and p-value (p), where D measures the maximum deviation in cumulative distribution between uORF-less genes and the uORF subset. (*G*) Schematic of uORF features, where feature subsets are listed along asterisks indicating statistical significance of the reduced protein distribution compared to uORF-less genes (*p< 0.05, **p< 0.01, ***p< 0.001 based on

Bonferroni-corrected KS p-values). (H) KS test D statistics and associated p-values for feature subsets compared against each other. We note that other MS/MS datasets were too small to provide significant results for uORF subsets.



Supplementary Figure 6: Properties of uORFs assayed by mutation experiments. The 25 uORFs assessed by mutation experiments (Xs) are overlaid on histograms of all mouse uORFs, with regard to seven 5' UTR or uORF properties (*A-G*). Color indicates disease-related genes (gray) and all other constructs (red). Histograms in panels *B-F* show only genes with a single uORF. In panels *E-G*, Xs are offset vertically to show the number of selected uORFs in each bin.

Supplementary Table 1: List of previously known eukaryotic genes with functional uORFs, including selected references. These were manually curated from PubMed articles referring to "upstream open reading frames" or "uORFs".

#	Gene	Clade(s)	References	Pubmed ID
1	ABI3	plant	(Ng et al., 2004)	15159632
2	ADH5	mammal	(Kwon et al., 2001)	11368338
3	AdoMetDC	mammal, plant	(Law et al., 2001; Mize et al., 1998; Raney et al., 2000; Nishimura et al., 1999; Chang et al., 2000; Franceschetti et al., 2001; Raney et al., 2002; Ruan et al., 1996)	11489903; 9829983; 10829027; 10570962; 11029703; 11139406; 11741992; 8939886
4	AN0244	fungi	(Galagan et al., 2005)	16372000
5	AN1179	fungi	(Galagan et al., 2005)	16372000
6	arg-2	fungi	(Freitag et al., 1996; Luo and Sachs, 1996; Wang et al., 1998; Wang et al., 1999; Wang and Sachs, 1997a, b)	8770589; 8636015; 9819438; 10608810; 8995256; 8995256
7	ARG-2	fungi	(Shen and Ebbole, 1996)	10679190
8	AS	mammal	(Pendleton et al., 2005)	15851478
9	AT(1)R	mammal	(Martin et al., 2006)	16504375
10	ATB2/AtbZIP11	plant	(Wiese et al., 2004)	15208401
11	ATF4	mammal	(Harding, et al., 2000; Harding, et al., 2003; Blais, et al., 2004)	11106749; 12667446; 15314157
12	ATF5	mammal	(Watatani, et al., 2008)	18055463
13	AtMHX	plant	(David-Assael et al., 2005)	15710632
14	BACE1	mammal	(Rogers et al., 2004)	14981268
15	BCKD-kinase	mammal	(Muller and Danner, 2004)	11073965
16	bcl-2	mammal	(Harigai et al., 1996; Salomons et al., 1998)	8649841; 9645350
17	brlA	fungi	(Han and Adams, 2001)	12586880
18	C/EBPalpha	mammal	(Lincoln et al., 1998)	9545285
19	C/EBPbeta	mammal	(Lincoln et al., 1998)	9545285
20	Cat-1	mammal	(Fernandez et al., 2002; Yaman et al., 2003)	11684693; 12757712
21	CBS1	fungi	(Tzagoloff and Dieckmann, 1990)	18155664
22	CD36	mammal	(Griffin et al., 2001)	11433350
23	CHOP	mammal	(Jousse et al., 2001)	11691921
24	Cited2	mammal	(van den Beucken et al., 2007)	17499866
25	CLN3	fungi	(Polymenis and Schmidt, 1997)	11753385
26	c-mos	mammal	(Steel et al., 1996)	8891345
27	CPA1	fungi	(Messenguy et al., 2002; Nyunoya and Lusty, 1984; Gaba et al., 2005; Linz et al., 1997)	12172963; 6086650; 16285926; 9083042
28	cpcA (GCN4)	fungi	(Wanke et al., 1997; Hoffmann et al., 2001)	9004217; 11553722
29	Cx41	animal	(Meijer et al., 2000)	10896676
30	Cyc1	fungi	(Pinto, et al., 1992)	1327957
31	CYP27	mammal	(Lodhi et al., 2003)	12909643
32	DCD1	fungi	(McIntosh and Haynes, 1986)	17692847
33	ERalpha	mammal	(Kos et al., 2002; Pentecost et al., 2005)	12147702; 15607532
34	Esi47	plant	(Shen, et al., 2001)	11244122
35	ETT	plant	(Nishimura et al., 2005)	16227452

36	Fli-1	mammal	(Sarrazin et al., 2000)	10757781
37	FOL1	fungi	(Zhang and Dietrich, 2005)	1625572
38	GADD34	mammal	(Lee, et al., 2009)	19131336
39	GATA-6	mammal	(Takeda et al., 2004)	15173203
40	GCN4	fungi	(Gaba et al., 2001)	11707416
41	GDNF	mammal	(Tanaka, et al., 2001)	11457495
42	Gld	C. elegans	(Lee, et al., 2004)	15105376
43	gna-2	animal	(Lee and Schedl, 2004)	11489903
44	GR (NR3C1)	mammal	(Diba et al., 2001)	11180405
45	H(+)-ATPase	plant	(Lukaszewicz et al., 1998)	9670558
46	HAC1	fungi	(Saloheimo et al., 2003)	12581366
47	HAP4	fungi	(Forsburg and Guarente, 1989)	15596718
48	HCS1	plant	(Puyaubert et al., 2008)	18156294
49	HER-2/neu	mammal	(Child et al., 1999; Mehta et al., 2006; Spevak et al., 2006)	10446211; 16598037; 17045969
50	HIAP2	mammal	(Warnakulasuriyarachchi et al., 2003)	12867997
51	HOL1	fungi	(Wright et al., 1996)	8955402
52	Huntington	mammal	(Lee et al., 2002)	12466534
53	INO2	fungi	(Eiznhamer et al., 2001)	11251853
54	LBP	protist	(Mittag et al., 1997)	9271214
55	Lc	plant	(Wang and Wessler, 2001)	12890013
56	LEU4	fungi	(Beltzer et al., 1986)	2420798
57	LPA	mammal	(B. R. Zysow, et al., 1995)	7749816
58	MDM2	mammal	(Brown et al., 1999; Jin et al., 2003)	10523842; 12730202
59	MKK1	fungi	(Zhang and Dietrich, 2005)	1625572
60	Mona (aka Gads)	mammal	(Guyot et al., 2002)	12487779
61	MOR	mammal	(Song et al., 2007)	17284463
62	MP	plant	(Nishimura et al., 2005)	16227452
63	Mrp2	mammal	(Zhang et al., 2007)	17065236
64	MS	mammal	(Col et al., 2007)	17683808
65	MtHAP2-1	plant	(Combier et al., 2008)	18519645
66	MVP	mammal	(Holzmann et al., 2001)	11297743
67	myb-7	plant	(Locatelli et al., 2002)	11855732
68	MYEOV	mammal	(de Almeida et al., 2006)	16275643
69	NOD2	mammal	(Rosenstiel et al., 2007)	18096043
70	ODC	plant, mammal	(Kwak and Lee, 2001; Ivanov et al., 2008)	1782674; 18626014
71	Opaque-2	plant	(Lohmer, et al., 1993)	8439744
72	P27 (Kip1)	mammal	(Gopfert et al., 2003)	12837699
73	PEAMT	plant	(Tabuchi et al., 2006)	16960350
74	PET111	fungi	(Strick and Fox, 1987)	16679454
75	PPR1	fungi	(Kammerer et al., 1984)	6096561
76	PR-39	mammal	(Wu et al., 2002)	12213322
77	R genes	plant	(Wang and Wessler, 1998)	12890013
78	RAR beta 2	mammal	(Reynolds et al., 1996)	8769409
79	RPC11	fungi	(Zhang and Dietrich, 2005)	1625572
80	RPL24	plant	(Nishimura et al., 2004)	15270688

81	SAC51	plant	(Imai et al., 2006)	16936072
82	SCHO9	fungi	(di Blasi et al., 1993)	8442384
83	SCO1	fungi	(Krummeck et al., 1991)	1782674
84	SEPT9	mammal	(McDade, et al., 2007)	17468182
85	SLC	mammal	(Calkhoven et al., 2003)	12704079
86	SOCS-1	mammal	(Schluter et al., 2000)	10679190
87	Sp3	mammal	(Sapetschnig, et al., 2004)	15247228
88	StuAp	fungi	(Wu and Miller, 1997)	8955402
89	Tie	mammal	(Park et al., 2006)	16457819
90	TIF 4631	fungi	(Goyer et al., 1993)	8336723
91	TPK1	fungi	(Zhang and Dietrich, 2005)	1625572
92	TPO	mammal	(Stockklausner et al., 2006)	16679454
93	UCP2	mammal	(Hurtaud et al., 2006; C. Pecqueur, et al., 2001)	16845607; 11098051
94	V(1b)	mammal	(Nomura et al., 2001; Rabadan-Diehl et al., 2007)	11287361; 17355321
95	VAR2CSA	parasite	(Amulic, et al., 2009)	19119419
96	VEGF-A	mammal	(Bastide et al., 2008)	18304943
97	Vigilin	mammal	(Rohwedel et al., 2003)	14504658
98	Wnt13	mammal	(Tang et al., 2008)	18155664
99	WSC3	fungi	(Zhang and Dietrich, 2005)	1625572
100	Yap1p	fungi	(Vilela et al., 1998; Vilela et al., 1999)	9469820; 10357825
101	Yap2p	fungi	(Vilela et al., 1998; Vilela et al., 1999)	9469820; 10357825

Supplementary Table 2: Listed are the 25 gene UTRs for which uORFs were tested using reporter constructs. Corresponding sequences for each construct are listed in Table S4. Columns 5 and 7 report the average protein and mRNA expression for the uORF-containing construct as a percent of the uORF-less construct expression values. Column 6 and 8 represent significance of expression difference between the construct containing vs lacking the uORF (based on two-sided t-test). Column 9 represents the mean protein expression value divided by the mean mRNA expression value, and column 10 represents significance (one-sided t-test). Columns 11-17 (next page) list the construct uORF properties, which are described in the Methods section. Construct numbers 18.1-18.8 show the eight constructs for FXII, of which the first four lack uORFs and the last four contain uORFs (see Table S4).

#	Category	Entrez ID	Symbol	Protein Expr: % -uORF	Protein Expr: pval	mRNA Expr: % -uORF	mRNA Expr: pval	Prot/ mRNA Expr: % -uORF	Prot/ mRNA Expr: p-val
1	random	72479	Hsdl2	65	3.E-04	76	2.E-06	85	6.E-02
2	random	69656	Pir	62	2.E-12	101	8.E-01	62	5.E-06
3	random	225010	Lycat	60	4.E-10	102	7.E-01	59	2.E-05
4	random	434446	Ccdc13	58	1.E-06	85	2.E-02	69	3.E-03
5	random	271981	A630047E20Rik	51	2.E-05	86	2.E-02	59	3.E-04
6	MS/MS	233799	Acsm2	50	2.E-07	112	1.E-02	44	6.E-05
7	MS/MS	18408	Slc25a15	45	2.E-08	113	3.E-02	40	4.E-07
8	MS/MS	18673	Phb	42	3.E-04	69	6.E-06	60	5.E-04
9	MS/MS	17850	Mut	39	9.E-12	101	9.E-01	39	9.E-04
10	MS/MS	66419	Mrpl11	33	1.E-05	94	7.E-01	35	9.E-05
11	MS/MS	72416	Lrpprc	26	5.E-07	114	2.E-02	23	4.E-03
12	MS/MS	235582	Glyctk	27	7.E-12	82	3.E-04	33	6.E-08
13	MS/MS	56046	Uqcc	25	3.E-07	83	7.E-04	30	1.E-04
14	MS/MS	67809	Fam82a2	21	2.E-08	105	2.E-01	20	8.E-06
15	MS/MS	94280	Sfxn3	20	7.E-11	101	4.E-01	20	5.E-04
16	puORF	841	CASP8	69	2.E-04	108	1.E-01	64	3.E-03
17	puORF	3791	MC2R	63	2.E-03	106	4.E-01	59	7.E-03
18	puORF	2161	F12	50	2.E-06	86	1.E-05	58	9.E-06
19	puORF	4158	KDR	50	5.E-09	89	3.E-02	56	6.E-04
20	puORF	50831	TAS2R3	42	7.E-06	96	2.E-01	44	2.E-05
21	disease-related	3043	HBB	30	4.E-13	112	1.E-01	27	3.E-04
22	disease-related	5573	PRKAR1A	5	6.E-19	88	2.E-02	5	1.E-09
23	disease-related	3664	IRF6	4	1.E-17	81	1.E-03	5	2.E-05
24	disease-related	6736	SRY	0	2.E-12	80	3.E-04	1	4.E-04
25	disease-related	6690	SPINK1	0	8.E-17	103	4.E-01	0	3.E-05
18.1	F12 exp	2161	F12_C_allele	100	NA	100	NA	100	NA
18.2	F12 exp	2161	F12_non_uORF1	110	7.E-01	100	5.E-01	111	3.E-04
18.3	F12 exp	2161	F12_non_uORF2	87	4.E-02	99	5.E-01	87	3.E-05
18.4	F12 exp	2161	F12_alt_start	124	1.E-02	95	1.E-01	131	3.E-03
18.5	F12 exp	2161	F12_T_allele	50	2.E-06	86	1.E-05	58	2.E-05
18.6	F12 exp	2161	F12_alt_uORF1	45	8.E-07	128	2.E-07	35	2.E-07
18.7	F12 exp	2161	F12_alt_uORF2	44	4.E-06	113	9.E-02	39	2.E-06
18.8	F12 exp	2161	F12_alt_uORF3	32	1.E-07	109	1.E-01	29	8.E-07

#	Symbol	# uORFs in 5' UTR	5' UTR len	uAUG sequence context	cap-uORF distance	uORF len.	uORF-CDS distance	# species w/ conserved uORF
1	Hsdl2	1	168	0	91	66	12	12
2	Pir	1	274	1	9	39	227	1
3	Lycat	1	188	0	61	9	119	14
4	Ccdc13	1	214	0	66	72	77	17
5	A630047E20Rik	1	235	0	121	21	94	10
6	Acsm2	1	111	0	19	42	50	1
7	Slc25a15	1	174	0	118	60	-4	15
8	Phb	1	80	0	29	18	33	17
9	Mut	1	95	0	85	24	-14	14
10	Mrpl11	1	78	1	14	21	43	20
11	Lrpprc	1	273	0	69	63	141	1
12	Glyctk	1	137	1	104	21	12	20
13	Uqcc	1	297	0	71	132	94	17
14	Fam82a2	1	56	1	31	33	-8	14
15	Sfxn3	1	220	1	142	54	24	13
16	CASP8	1	303	0	157	24	122	0
17	MC2R	1	302	0	31	207	64	2
18	F12	1	49	0	44	9	-4	1
19	KDR	1	177	0	71	60	46	0
20	TAS2R3	1	61	0	15	45	1	2
21	HBB	1	50	0	21	42	-13	0
22	PRKAR1A	1	153	1	56	126	-29	0
23	IRF6	1	263	0	214	108	-59	0
24	SRY	2	148	0	73	15	60	0
25	SPINK1	2	120	1	66	9	45	0
18.1	F12_C_allele	0	49	NA	NA	NA	NA	NA
18.2	F12_non_uORF1	0	49	NA	NA	NA	NA	NA
18.3	F12_non_uORF2	0	49	NA	NA	NA	NA	NA
18.4	F12_alt_start	0	49	NA	NA	NA	NA	NA
18.5	F12_T_allele	1	49	0	44	9	-4	1
18.6	F12_alt_uORF1	1	49	0	6	18	25	NA
18.7	F12_alt_uORF2	1	49	0	20	33	-4	NA
18.8	F12_alt_uORF3	1	49	1	8	45	-4	NA

Supplementary Table 4: UTR sequences synthesized for reporter construct experiments. Column 2 lists the Gene Symbol and RefSeq ID of the uORF-containing gene. Column 3 lists the two sequences synthesized; the two tested variants are listed in parentheses, the uORFs are highlighted in red, and the main ORF ATG or alternate in-frame ATGs are in bold. Each sequence is flanked on either end with the XhoI cut site 'gctagc' and, when necessary, 1-2 additional residues to keep the main ORF ATG in-frame with the luciferase sequence. For the 8 FXII constructs (18.2-18.8), the modified residue is underlined.

#	Gene / RefSeq	Sequence
1	Hsdl2 NM_024255	gctagcTCTTGTAAGCTCTCGGTCTTGTAGGATCTCGGTCTTGTAGGAGGGCCGGTCCCCGAGCGGGTCTCGGGGCGGGGCCCGGGGCGCGGCTAA(T/A)TGCGGAGAGAAAGTTCGCCTGTCACTCACAGCTCGCTGCTGTTCCACTGCCACCAAGTTCTCTGAACTGCGAAGGTC**ATG**TTgctagc
2	Pir NM_027153	gctagcAGAGAGCC(T/A)TGGGGCGGAGTCAGAGAATCCAAACCCAGGGTAAATAAAGGGTGTGTCCCCCGGTCCAAGGCCCCTGAGACCTAGAGACTCCCGCCTTCTGGAGGCCCGCAACAAAGCGCTGAGTCACGCTGACTGACGGCTGGCGCTCCGCAAACCTGTCCTCCCTTTGCTGTGTCCCTGCCGCTGTGGAGTCTAGGCACCTCCGACTGTGGCCTCCCTCCTGGGCCTGCGCTACTGCCCACTCAACATCAGAACTTTCTTTATAAACTTCAAA**ATGTT**gctagc
3	Lycat NM_001081071	gctagcGGTGGAGCCCGGAGGGACGCATTACTAAGGCGACGGGGCTCGACGCCTCCCCGCTTCGGG(T/A)TGAATTAGCGGCGGGTTCTTCTCCCAGGTTGTGACCCCCGTGGAGTCGCTGTCTCGCCCCAGTGTCCCCGCGACGTCGCGCGTTTCCCCCGGTGTGTCCCAGGTGCGCGGCCCGCGGTGTCAGAATC**ATG**TTgctagc
4	Ccdc13 NM_001033784	gctagcGAGGCCTGGGCACTGTTGCCTGTGAGGAGCAGTCGGTCCACCTTTCCCTTGCAGGGTGTTGGAAG(T/A)TGAAATCCAGCATCTTCGCAGTGAGCTCCGGGAGACAGTCGATGAGAATGGGCGACTATATAAGCTGCTGAAGGAGAGGGACTTTGAGATCAAACACCTCAAGAAGAAAATAGAAGAGGACAGATTCGCCTTCACAGGGGCGTCTGGG**ATG**TTgctagc
5	A630047E20Rik NM_173032	gctagcGGAGCGAGGTGCGCGCTCACCCTCGGGTTCTGAGCCGGGGACCCGGGAGGCCCGCGCGGACGGAGGCGTGTCGCTCCGGCTGTCACCTTCCCTAGGCGGCCTCTTGTAGACGCGCCTGAA(T/A)TGCAGGACACGGCCCACTAAGCACTGCCAGGCATAGCTTCCGCGCTTAGGAGCTCAACCCCCCAATCTCTGCGTCCCCAGATTCAGATCATCTTAAAGGATCCAAGACTACAAA**ATG**TTgctagc
6	Acsm2 NM_146197	gctagcAGTGCTCTTCTCTCCACTG(T/A)TGCTGCAGGAATTCTCAAACGACTTCCCAAGAGGGAAATAGACAAAGCTACTGTCAGAGAGAGTTCCAGAAGAAGACCAGAAGCTCTGACT**ATG**TTgctagc
7	Slc25a15 NM_011017	gctagcGGAGGGCCCGAGGGCCGGACCCACAGGGCGACCTTAAAAATTGCCCGGGGGCCCGCGGCCGCCAGCGGAGCGCGCGGGCGGCAGAGGGCGAGCCGGACACAGCCAGGCGCGGCGGGAT(T/A)TGTGGTACATTCCCTTCCCTCCAGAGAATCGCCTTCCACAGAAACCAGTAACGCC**ATG**AAgctagc
8	Phb NM_008831	gctagcCACGCGCAGTATCCGGAGCTGGGGAATTC(T/A)TGTGGAGGTCAGAGTGAAAGCAGGTGTGAGTGGAAGCAACAGAAGGAGTC**ATG**TTgctagc
9	Mut NM_008650	gctagcGGCGTCTGGGTTCGGTTCTGAAGTCCGGGCTTGGTCCGGGTGATCCCCAAACACTGACCGTTCTCATTTCCTTTTGGGAGTATTC(T/A)TGCTTCACC**ATG**TTGAGAGCTAAgctagc
10	Mrpl11 NM_025553	gctagcACCTCTGACCCAAA(T/A)TGGCCGCGCCCAGAGCGTAGTTCTTTGCTTCTCCGAGGCGAGCTAAGATTAAAATCCTACATC**ATG**TTgctagc
11	Lrpprc NM_028233	gctagcGCCAGGCTCGCTGAGAGCCGGGGCGCTGGACAAGGGAACAGGGACACTGTGAACGGAAGACAAAAAACA(T/A)TGTATAGTAAAACAGAAGGCGGATCCAGGGTATCCCGCGAACTCGGCTTGCCTCTCCCGTGAGCCTTGGCGGGGATCTGCCTCCTCCTCCGCTGGACGCCCTCGGTCCTTAGTTTGTCCCACTAGGGGCGACCGGGTCGTCACGTGCTCCTCCAGGTCCTCTGCAGGAGCGTGCATCCTATCCTGCCTTGGGTACGCTGAGCC**ATG**TTgctagc
12	Glyctk NM_001039586	gctagcGGACTTGGAGCAGAGTGTCCAGAAGACTGGTGTGTGGTCTACAGAAGAGGACGTGGCTGGGCATCAGGGCTGGGCTTTCAGGGCTGGTGTAAGATCAGGCAGCT(T/A)TGGACACGAGGCTGTGCTGAGAGAAGCGGGGC**ATG**TTgctagc
13	Uqcc NM_018888	gctagcTTGTGGAGGAACATGGCGGCACGAGGTTGCTGGTGCGAGTCCTTAGGAACCAGACTACCATCTCTCAGTGGGTTCCAGT(T/A)TGCAGCCAACTGGTATCAGTGTCTCCTACCCAAAGACAGTGGAGCAGCACTTCCCAGTGGCTCCAGAAGAATCAGTCTCGAGTGTGTTTGGGATCAGAACAGACTGTTGGAGCGGACACAGCGCAGAGTAGGAAATACCACAACACAAGTAAGCTTCTTACTACCCAAGATTTCCCACAGCCTGTTGAAGAGAAGGTCGGTCCCTTCACGAAGATAATAGAAGCC**ATG**TTgctagc
14	1200015F23Rik NM_001033136	gctagcACAGAGTGAAAGCGCGGTGCCTGCTGCTGCC(T/A)TGGTGGCTGGTGGCTGGGTACACT**ATG**TCTAGgctagc
15	Sfxn3 NM_053197	gctagcGTCTCAACGGCCTGGTCTGGGAGAATCACTCTGGACATCCACTGTCTCGGAACTCTGCCAAGAGGGGGGGTGGGTCAGGCGAACGAGCTCAGGGAGCCCCCGCCCTTCCCTGCTGCTCAGCGTCACGCGTGACGTCTCGGTG(T/A)TGGCTGGGAGGAAAGCGGAGAGCGGTGAGGAAGGCGGGTCTGAGAGCTTCTAGAGGCTGAAAACCCCGGAAAGCAAG**ATG**TTgctagc
16	CASP8 NM_001228	gctagcGCTCTGAGTTTTTGGTTTCTGTTTCACCTTGTGTCTGAGCTGGTCTGAAGGCTGGTTGTTCAGACTGAGCTTCCTGCCTGCCTGTACCCCGCCAACAGCTTCAGAAGAAGGTGACTGGTGGCTGCCTGAGGAATACCAGTGGGCAAGAGAATTAGCAT(T/G)TCTGGAGCATCTGCTGTCTGAGCAGCCCCTGGGTGCGTCCACTTTCTGGGCACGTGAGGTTGGGCCTTGGCCGCCTGAGCCCTTGAGTTGGTCACTTGAACCTTGGGAATATTGAGATTATATTCTCCTGCCTTTTAAAAAG**ATG**TTgctagc
17	MC2R NM_002253	gctagcATTCCTTCTCATTCATTTTGCCCAGAAAGTTCCTGCTTCAGAGCTGAAGGTGATTGGGAGATTTTAACTTAGATCTCCAGCAA(G/A)TGCTACAAGGAAGAAAAGATCCTGAAGAATCAATCAAGTTTTCCGTGAAGTCAAGTCCAAGTAACATCCCCGCCTTAACCACAAGCAGGAGAA**ATG**AAGCACATTATCAACTCGTATGAgctagc
18	F12	gctagcCTATTGATCTGGACTCCTGGATAGGCAGCTGGACCAACGGACGGACGCC**ATG**ACgctagc

	NM_000505	
19	KDR NM_000529	gctagcCTGAGTCCCGGGACCCCGGGAGAGCGGTCA(G/A)TGTGTGGTCGCTGCGTTTCCTCTGCCTGCGCCGGGCATCACTTGCGCGCCGCAGAAAGTCCGTCTGGCAGCCTGGATATCCTCTCCTACCGGCACCCGCAGACGCCCCTGCAGCCGCGGTCGGCGCCCGGGCTCCCTAGCCCTGTGCGCTCAACTGTCCTGCGCTGCGGGGTGCCGCGAGTTCCACCTCCGCGCCTCCTTCTCTAGACAGGCGCTGGGAGAAAGAACCGGCTCCCGAGTTCTGGGCATTTCGCCCGGCTCGAGGTGCAGG**ATG**TTgctagc
20	TAS2R3 NM_016943	gctagcCAGTGAGGAGATTCTA(C/T)GTATCAACAGAAAGAACAAAGATCAGGGCTGCCTAATTGCTGAC**ATG**TTgctagc
21	HBB NM_000518	gctagcACATTTGCTTCTGACACAACT(G/A)TGTTCACTAGCAACCTCAAACAGACACC**ATG**GTGCATCTGATgctagc
22	PRKAR1A NM_002734	gctagcGATTGGCTGCGGCCAGGCCGTTTCCGGTGGAGCTGTCGCCTAGCCGCTATCGCAGA(G/A)TGGAGCGGGGCTGGGAGCAAAGCGCTGAGGGAGCTCGGTACGCCGCCGCCTCGCACCCGCAGCCTCGCGCCCGCCGCCGCCCGTCCCCAGAGAACC**ATG**GAGTCTGGCAGTACCGCCGCCAGTGAgctagc
23	IRF6 NM_006147	gctagcGAGCTCGGCGCACCTGGGCTGGGCAGGTAAGGGCTGGTGCGGGACGGGGAGAGGAACCTGCAGTCCCTACTTGGGTAGAGCCAGGCGCCCCTTGGCTAAGACGTCGAGGAGCGTGGTAGCGACGGGTGATCTTCGCTGCGGACTTGGTTCGGAGGGACGTCCGCTTCTGGTGGACAGATTGAGCAAAGAATCTTTGAGCGGTCAAGGGAAAGACA(A/T)GCCGACTCTTCAGATCCCTGTGGACACACTGCCTGCTCTTCCATATC**ATG**GCCCTCCACCCCCGCAGAGTCCGGCTAAAGCCCTGGCTGGTGGCCCAGGTGGATAGgctagc
24	SRY NM_003140	gctagcGTTGAGGGGGTGTTGAGGGCGGAGAAATGCAAGTTTCATTACAAAAGTTAACGTAACAAAGAATCTGGTAGAA(G/A)TGAGTTTTGGATAGTAAAATAAGTTTCGAACTCTGGCACCTTTCAATTTTGTCGCACTCTCCTTGTTTTTGACA**ATG**TTgctagc
25	SPINK1 NM_003122	gctagcAGCCCAGTAGGTGGGGCCTTGCTGCCATCTGCCATATGACCCTTCCAGTCCCAGGCTTCTGAAGAGA(C/T)GTGGTAAGTGCGGTGCAGTTTTCAACTGACCTCTGGACGCAGAACTTCAGCC**ATG**TTgctagc
18.2	F12 NM_000505	gctagcCTATTGATCTGGACTCCTGGATAGGCAGCTGGACCAACGGACGGAAGCC**ATG**ACgctagc
18.3	F12 NM_000506	gctagcCTATTGATCTGGACTCCTGGATAGGCAGCTGGACCAACGGACGGAGGCC**ATG**ACgctagc
18.4	F12 NM_000507	gctagcCTATTGATCTGGACTCCTGGATAGGCAGCTGGACCAACGG**ATG**GACGCC**ATG**ACgctagc
18.5	F12 NM_000508	gctagcCTATTGATCTGGACTCCTGGATAGGCAGCTGGACCAACGGACGGATGCC**ATG**ACgctagc
18.6	F12 NM_000509	gctagcCTATTGATGTGGACTCCTGGATAGGCAGCTGGACCAACGGACGGACGCC**ATG**ACgctagc
18.7	F12 NM_000510	gctagcCTATTGATCTGGACTCCTGGATGGGCAGCTGGACCAACGGACGGACGCC**ATG**ACgctagc
18.8	F12 NM_000511	gctagcCTATTGATATGGACTCCTGGATAGGCAGCTGGACCAACGGACGGACGCC**ATG**ACgctagc

Supplementary Notes

Details of matched protein and mRNA datasets

We analyzed large-scale protein and mRNA datasets from four published studies (1-4) across a variety of mouse tissues and developmental stages. Below we provide details of the datasets and mapping to Entrez Gene identifiers.

Lai and colleagues (1) investigated the mouse liver proteome using a 3-dimensional separation of intact liver proteins from adult mice. They used a combination of centrifugation, 2D HPLC separation of supernatant proteins, and SDS PAGE to separate proteins and then used LC-MS/MS with ProteinProphet software to identify a total of 7090 unique proteins. They approximated protein abundance by peptide counts normalized by protein length. We downloaded the protein abundance data and mapped their 7090 IPI protein identifiers to 5036 unique Entrez Gene identifiers, excluding IPI identifiers that mapped to more than one gene locus. Although Lai and colleagues used microarrays to quantify mRNA abundance from the same tissues, their mRNA data were not available either in supplemental materials or upon request. Thus we obtained mRNA abundance data using the GNF1M tissue atlas(5) by averaging the measurements from the two liver sample replicates and mapping probe level data to Entrez Gene identifiers. As described in Methods, we then excluded the top 10% most highly expressed genes, genes with poorly quantified mRNA values (<40), and genes with discordant uORF presence (i.e. splice forms that contained and lacked uORFs), leading to a total of 2484 genes with well-quantified protein and mRNA measurements in liver.

Cox and colleagues (2) investigated the proteome of mouse lung at six stages development (embryonic day 13.5, 16.5, 18.5, and post-natal day 2, 14, and 56). They performed MudPIT analyses on cellular fractions and used SEQUEST software searches to identify 3330 proteins. Abundance was estimated by spectral counts, summed over all cellular fractions. mRNA expression for equivalent time points was provided by microarray studies (6) using the Affymetrix Mu11K A and B chip sets. We downloaded their matched protein and mRNA abundance data and mapped their 1383 SwissProt identifiers to 1266 unique Entrez Gene identifiers, excluding SwissProt identifiers that mapped to more than one gene locus. As described in Methods, we then excluded the top 10% most highly expressed genes, genes with poorly quantified mRNA values (<40), and genes with discordant uORF presence, leading to a total of 2569 well-quantified protein and mRNA measurements across 722 unique gene loci.

In our previous study (3), we investigated the mitochondrial proteome across 14 mouse tissues. Mitochondria were purified from each tissue by a combination of centrifugation and Percoll density gradients. We used LC-MS/MS and SpectrumMill software to identify products from 3881 genes. Protein abundance was approximated by total peak intensity (sum of MS peak areas for all sequence identified peptides matching a protein) normalized by protein length. An integrated analysis was performed to separate the truly mitochondrial proteins from the co-purifying contaminants, leading to 591 mitochondrial proteins with MS/MS quantification. We obtained mRNA abundance data for these gene loci using the GNF1M tissue atlas (5) by averaging the measurements from equivalent tissue samples. Because protein abundance measurements were obtained per mass of mitochondria, and mRNA abundance was measured per tissue, we normalized protein measurements by the quantity of mitochondria within each tissue so that each measurement represents tissue-level data (5). As described in Methods, we then excluded the top 10% most highly expressed genes, genes with poorly quantified mRNA values (<40), and genes with discordant uORF presence, leading to a total of 5060 well-quantified protein and mRNA measurements across 487 unique gene loci.

Kislinger and colleagues (4) investigated proteins from six mouse organs (brain, heart, kidney, liver, lung, and placenta). Proteins were fractionated into four subcellular compartments (cytosol, membranes, mitochondria, and nuclei), and analyzed by MudPIT. A total of 4768 proteins were detected using SEQUEST software. Abundance was estimated by spectral counts, summed over all cell fractions. The authors matched their protein data to mRNA abundance from the GNF1M tissue atlas (5) and the Zhang et al. microarray study (7), using stringent normalization methods. We downloaded their matched protein and mRNA abundance data and mapped their 1758 SwissProt mouse identifiers to 1336 unique Entrez Gene identifiers, excluding SwissProt identifiers that mapped to more than one gene locus. As described in Methods, we then excluded the top 10% most highly expressed genes, genes with poorly quantified mRNA values (<40), and genes with discordant uORF presence, leading to a total of 2377 well-quantified protein and mRNA measurements across 925 unique gene loci. We analyzed matched data both from the GNF1M atlas and the Zhang et al. microarray studies with similar results (data shown only for GNF1M expression).

In addition to the above matched mRNA and protein studies, we also analyzed two additional proteomics datasets (8, 9).

Adachi and colleagues (8) analyzed proteins from 3T3-L1 adipocytes, which were fractionated into four subcellular compartments (nuclei, mitchondria, membrane, and cytosol) and analyzed by LC-MS/MS. A total of 3287 unique proteins were detected using Mascot software. Abundance was approximated by number of observed peptides. We downloaded their protein data and mapped their 3287 IPI proteins to 2805 unique Entrez Gene identifiers, excluding IPI identifiers that mapped to more than one gene locus. As described in Methods, we then excluded the top 10% most highly expressed genes, and genes with discordant uORF presence, leading to a total of 2563 well-quantified protein measurements. The authors also analyzed matched mRNA microarrays from 3T3-L1 adipocytes, however the raw mRNA expression values for these experiments were not provided in the supplemental materials. Therefore, we simply analyzed the protein data without matching mRNA values.

Williamson and colleagues (9) analyzed proteins expressed during differentiation of embryonic stem cells to hemangioblasts following the expression of *Bry*$^{GFP/+}$ and *Flk1* genes. Cells were sorted using flow cytometry, enriched for nuclei, and then analyzed by LC-MS/MS and ProQUANT software to detect 2389 proteins. Abundance was approximated by number of observed peptides. We downloaded their protein data and mapped their 2389 SwissProt and RefSeq proteins to 1306 unique Entrez Gene identifiers, excluding identifiers that mapped to more than one gene locus. As described in Methods, we then excluded the top 10% most highly expressed genes, and genes with discordant uORF presence, leading to a total of 800 well-quantified protein measurements. The authors also analyzed matched mRNA microarrays from the same cells. However we did not utilize the mRNA data because the protein abundance was measured only for nuclei and the mRNA was harvested from whole cells. Therefore, we simply analyzed the protein data without matching mRNA values.

Highly expressed genes tend to lack uORFs

As others have previously noted (10), the most highly expressed genes tend to lack uORFs (see Fig. S2). Because of this skewed distribution, the set of uORF-containing genes have lower mean mRNA expression compared to uORF-less genes. Matsui and colleagues (11) argue that uORFs *cause* widespread reduced mRNA expression, through mRNA instability or nonsense mediated decay. While this explanation is possible, an alternate explanation is that evolutionary selection ensures that uORFs are not present in transcripts that the cell requires in highest abundance. This would not be surprising, since the 5' UTRs of the most highly expressed transcripts not only lack uORFs, but also tend to be short, unstructured and have low GC content – and thus selection is acting differently on this set of genes (10). Because of the biases inherent in this class of highly expressed transcripts, we excluded the 10% most highly expressed genes from our analyses (Fig. 2). However, we do provide figures showing the distribution of protein expression for all transcripts (Fig. S4), as well as the distribution excluding the top 10% most highly expressed genes (Fig. S3).

References

1. Lai, K. K., Kolippakkam, D., & Beretta, L. (2008) Comprehensive and quantitative proteome profiling of the mouse liver and plasma. *Hepatology (Baltimore, Md* **47,** 1043-1051.
2. Cox, B., *et al.* (2007) Integrated proteomic and transcriptomic profiling of mouse lung development and Nmyc target genes. *Molecular systems biology* **3,** 109.
3. Pagliarini, D. J., *et al.* (2008) A mitochondrial protein compendium elucidates complex I disease biology. *Cell* **134,** 112-123.
4. Kislinger, T., *et al.* (2006) Global survey of organ and organelle protein expression in mouse: combined proteomic and transcriptomic profiling. *Cell* **125,** 173-186.
5. Su, A. I., *et al.* (2004) A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Sciences of the United States of America* **101,** 6062-6067.
6. Mariani, T. J., Reed, J. J., & Shapiro, S. D. (2002) Expression profiling of the developing mouse lung: insights into the establishment of the extracellular matrix. *American journal of respiratory cell and molecular biology* **26,** 541-548.
7. Zhang, W., *et al.* (2004) The functional landscape of mouse gene expression. *Journal of biology* **3,** 21.
8. Adachi, J., Kumar, C., Zhang, Y., & Mann, M. (2007) In-depth analysis of the adipocyte proteome by mass spectrometry and bioinformatics. *Mol Cell Proteomics* **6,** 1257-1273.
9. Williamson, A. J., *et al.* (2008) Quantitative proteomics analysis demonstrates post-transcriptional regulation of embryonic stem cell differentiation to hematopoiesis. *Mol Cell Proteomics* **7,** 459-472.
10. Kochetov, A. V., *et al.* (1998) Eukaryotic mRNAs encoding abundant and scarce proteins are statistically dissimilar in many structural features. *FEBS letters* **440,** 351-355.
11. Matsui, M., Yachie, N., Okada, Y., Saito, R., & Tomita, M. (2007) Bioinformatic analysis of post-transcriptional regulation by uORF in human and mouse. *FEBS letters* **581,** 4184-4188.

Appendix D

—

Machine learning predictions of the human mitochondrial proteome

Machine learning predictions of the human mitochondrial proteome

Chapter 2 details a naïve Bayes prediction of the human mitochondrial proteome. However, naïve Bayes is just one of many supervised classification algorithms. Here, I compare five alternative machine learning algorithms for predicting which of the ~22,000 genes in the human genome code for mitochondrial proteins. Comparison of naïve Bayes, support vector machines, decision trees, boosting, and bagging algorithms shows that naïve Bayes provides the most accurate and interpretable prediction of the human mitochondrial proteome.

Methods

Genomic datasets and training sets

Each classifier relies on clues of mitochondrial localization (features) and is trained on large sets of known mitochondrial and non-mitochondrial genes. Table 1 lists 21 computed features of mitochondrial localization. These features are heterogeneous (categorical and scalar), conditionally dependent, and contain up to 63% missing values.

The training data consist of 654 known mitochondrial proteins (termed gold+) and 2847 known non-mitochondrial proteins (gold-) described in Chapter 2. The training data are complicated by several factors: (i) the ratio of training examples (gold+/gold- = 23%) does not match the prior belief of actual class size ratio (7%); (ii) approximately ~5% of gold- labels are likely to be incorrect; (iii) gold- is missing several large subcategories of actual non-mitochondrial proteins.

Assessment criteria

All algorithms and parameter settings are compared on the same 10-fold cross-validation splits of training data. A commonly used alternative is "leave one out" cross validation, however this approach likely overestimates performance. A second alternative approach uses theoretical models of complexity, however these models may not match the actual data. I define the best predictor as the one with the highest sensitivity at a 10% corrected FDR.

Definitions

The following definitions and abbreviations are utilized: TP=true positives; FN=false negatives; FP=false positives; TN=true negatives; Sensitivity SN = TP/(TP+FN); Specificity SP=TN/(TN+FP); False discovery rate FDR= FP/(FP+TP). O_{prior}=prior odds of mitochondrial localization; Corrected false discovery rate cFDR = (1 - SP)/(1 - SP + SN * O_{prior}).

Results

Naïve Bayes

Naive Bayes classification relies on the simplifying assumption that the input features are conditionally independent. That is, for feature vector *x* and classification category *y*, P(**x**,*y*) = $P(y)\prod_{i=1}^{d} P(x_i|y)$. The weights $P(x_i|y)$ can be learned by simple counts[1]. Categorical and missing data are readily handled, by assigning weights to each category, and scalar values can be divided into relevant bins. Although this method can result in a biased classifier, the benefit is low variance[2].

I performed a Naive Bayes classification on 8 of 21 features that were selected to be conditionally independent, with correlation coefficients below 0.3[3]. Ten-fold cross-validation showed a range of sensitivity and cFDR values depending on the log-likelihood score thresholds selected (Fig. 1 brown diamonds). Naïve Bayes classification with 8 features achieves 71% sensitivity at a 10% cFDR threshold. Naïve Bayes classification using all 21 non-independent features (Fig. 1, brown cross) achieved worse performance at the 10% cFDR threshold, due to the violated assumption of independence.

Support Vector Machines (SVMs)

SVMs learn a linear boundary designed to maximally separate training classes. For classes that are not linearly separable, misclassifications can be allowed with a penalty. For data that is not linear, it is possible to transform the input into a higher dimensional space and perform linear classification in that higher space. Two common transformation types are polynomial and radial basis, and kernel methods are available to simplify the computation. However, more complex models (i.e. higher dimensional spaces) will always be able to better separate data than simpler models, and thus a complexity regularization penalty must be introduced to avoid overfitting. SVMs define the exact decision boundary only using a small subset of the training examples that are near the boundary (i.e. support vectors), and thus a single outlier or misclassified example will radically affect the solution[1]. This is in contrast to linear discriminant analysis in which the decision boundary is determined using the covariance of the class distributions and the position of the class centroids[2]. Other SVM limitations include difficulty in handling categorical or missing data, and equal dependence on all features despite the fact that some features are likely to be irrelevant or particularly noisy.

I applied SVM classification using the SVMlight package[4]. I compared SVM predictions using all 21 features to SVM predictions using 8 relevant features. I assessed a wide range of SVM parameters for cost-factor $j \in \{0.1, 0.3, 1, 2\}$, margin-tradeoff $c \in \{0.5, 1, 10\}$, gamma $g \in \{0.05, 0.1, 0.2, 0.5, 1, 10\}$, and polynomial level $d \in \{1\text{-}6\}$. For radial basis kernels, high gamma values concentrate on examples close to the decision boundary and thus create "wiggly" boundaries that overfit data, thus small values are likely to be more robust. The cost-factor parameter measures how much a positive training example should be weighted compared to a negative training example, therefore I predicted a cost parameter of 0.3 would best compensate for the discrepancy of my training sets. The margin-tradeoff parameter determines the trade-off between training error and margin, which should be high to avoid overfitting.

Not surprisingly, SVMs performed better on 8 relevant features compared to all 21 features (Fig. 1 pink diamonds and cross, respectively) since the extra features contained some irrelevant information. The optimal set of parameters for the radial basis kernel (j=0.3, g=0.05, c=1) achieved a 65% sensitivity at 11% cFDR, similar to the 63% sensitivity achieved by the best polynomial kernel (j=0.3, d=1, c=10) (Fig. 1 aqua diamonds). Surprisingly, polynomial degrees exceeding 1 performed poorly (cFDR > 0.8, off the chart in Fig. 1) while the linear polynomial performed well. Compared to Navie Bayes, the SVM overfit the data.

Decision trees

Decision tree classifiers automatically learn a tree of simple classification rules, where each node represents a simple rule (that is, a single threshold for a single feature) and the tree leaves hold the classification results. The appealing simplicity of the model mimics a person's rule-based protocol for making decisions but the sequential greedy implementation does not take joint feature probabilities into account when choosing new nodes. Like Naive Bayes classifiers, decision trees easily handle categorical data and are readily interpretable, however they do not require conditional independence and they automatically handle feature selection.

I utilized the GML AdaBoost Matlab[5] implementation of decision trees with a variety of tree levels {1-30}. Since the training relies on optimizing the error rate, I triplicated the gold- examples so that the size of the training sets matched prior expectations. The decision tree performed better on 8 features than 21 due to overfitting (Fig. 1, blue diamond vs cross). With 8 features, the trees converged at 10 levels and achieved 64% sensitivity at 13% FDR.

Boosting using decision trees.

Combining different learners tends to decrease variance[1]. Boosting is a technique to train an ensemble of learners in a greedy iterative algorithm, where the first learner is trained on all data equally, and successive learners are sequentially added and trained on weighted data (more weight on misclassified examples), to decrease the overall weighted logistic loss training error. Boosting tends to work well even with simple base learners, such as decision stumps (trees with one level). The advantages of boosting are good generalizability, easy handling of heterogeneous or missing data, and automatic feature selection. The tradeoff can be interpretability of results.

I used the GML AdaBoost Matlab[5] package with a variety of decision tree levels (1-4). Since boosting automatically selects the most informative features, it performed better with 21 features compared to 8 (Fig. 1 green cross vs diamonds). Boosting achieved 72% sensitivity at 10% cFDR.

Bagging

Bagging (**b**ootstrap **agg**regat**ing**) is another ensemble method where each classifier is trained on a sampling of the original set with replacement. Bagging is useful when the learning algorithms are unstable (i.e. small input changes cause different classifications, such as in decision trees and SVM)[6]. Like boosting, the ensemble reduces variance without affecting bias[6].

Using MATLABArsenal[7], I tested bagging using 3 weak learners: decision stumps, decision trees, and SVMs (parameters optimized from above). SVM weak learners showed the best performance, as the ensemble likely compensated for the overfitting tendency. Unlike the underlying base SVMs, the 21-feature learner slightly outperformed the 8-feature learner (Fig. 1, orange cross vs diamond), and radial basis kernel outperformed the polynomial kernel (data not shown).

Discussion

To compare performance of the five machine learning approaches, sensitivity is plotted against the corrected false discovery rate based on 10-fold cross validation (Fig. 1). At a 10% FDR threshold, the most sensitive algorithms were Naïve Bayes, boosting, and bagging (Fig. 1, red circle). SVMs and decision trees performed similarly to each other and tended to overfit the data (data not shown). Since Naïve Bayes with 8 features shows similar accuracy to the ensemble methods of boosting and bagging, it is preferable as the results are easily interpretable. However, unlike Naïve Bayes, the ensemble methods do not require careful selection of conditionally independent features, and thus are the easiest to implement.

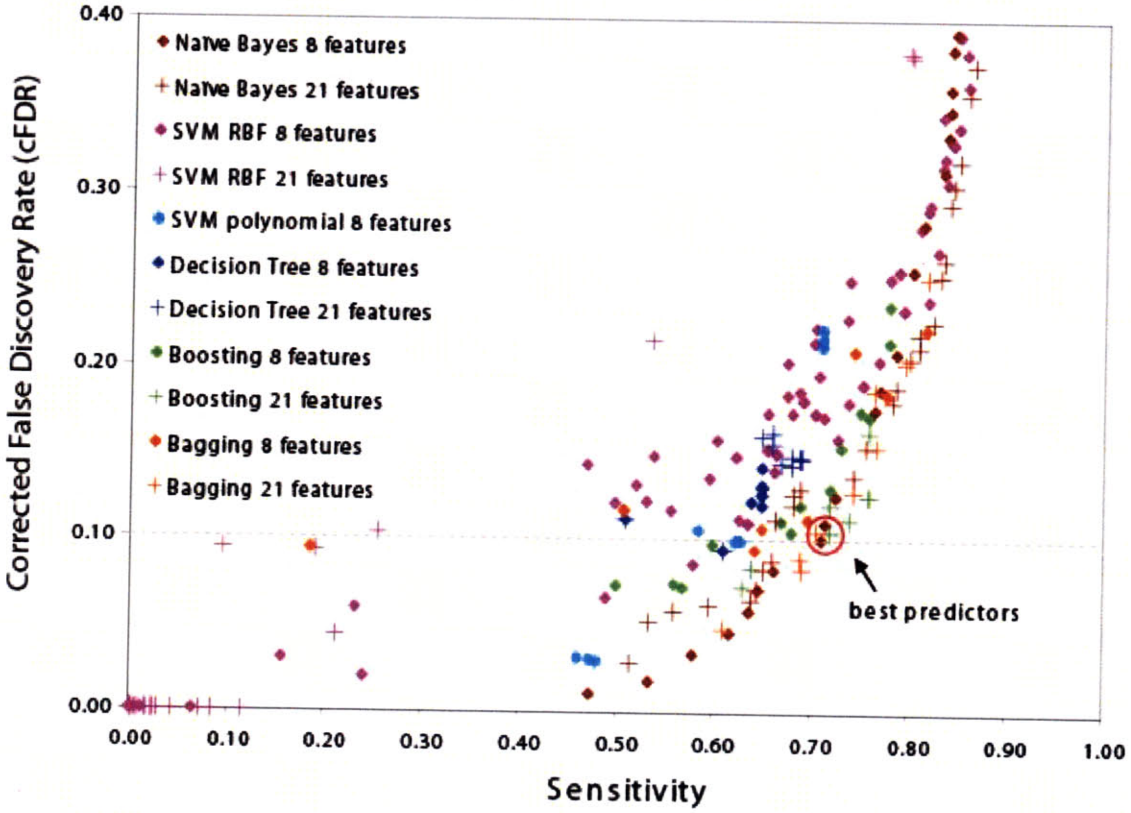
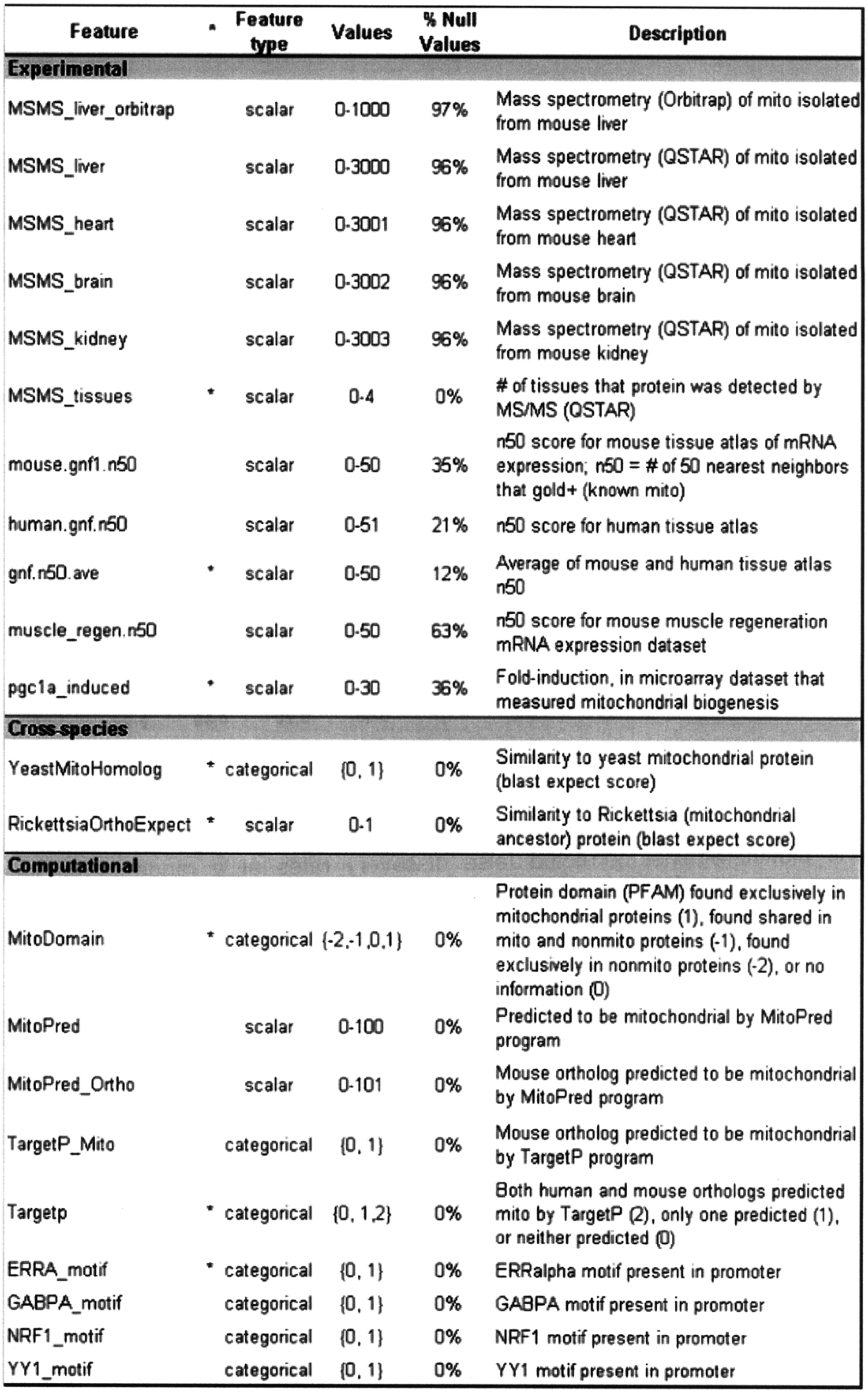


Figure 1: Assessment of machine learning algorithms.

Sensitivity is plotted against corrected false discovery rates at a range of model parameters and score thresholds. In the legend, SVM indicates support vector machine and RBF indicates radial basis kernel.



Feature	*	Feature type	Values	% Null Values	Description
Experimental					
MSMS_liver_orbitrap		scalar	0-1000	97%	Mass spectrometry (Orbitrap) of mito isolated from mouse liver
MSMS_liver		scalar	0-3000	96%	Mass spectrometry (QSTAR) of mito isolated from mouse liver
MSMS_heart		scalar	0-3001	96%	Mass spectrometry (QSTAR) of mito isolated from mouse heart
MSMS_brain		scalar	0-3002	96%	Mass spectrometry (QSTAR) of mito isolated from mouse brain
MSMS_kidney		scalar	0-3003	96%	Mass spectrometry (QSTAR) of mito isolated from mouse kidney
MSMS_tissues	*	scalar	0-4	0%	# of tissues that protein was detected by MS/MS (QSTAR)
mouse.gnf1.n50		scalar	0-50	35%	n50 score for mouse tissue atlas of mRNA expression; n50 = # of 50 nearest neighbors that gold+ (known mito)
human.gnf.n50		scalar	0-51	21%	n50 score for human tissue atlas
gnf.n50.ave	*	scalar	0-50	12%	Average of mouse and human tissue atlas n50
muscle_regen.n50		scalar	0-50	63%	n50 score for mouse muscle regeneration mRNA expression dataset
pgc1a_induced	*	scalar	0-30	36%	Fold-induction, in microarray dataset that measured mitochondrial biogenesis
Cross-species					
YeastMitoHomolog	*	categorical	{0, 1}	0%	Similarity to yeast mitochondrial protein (blast expect score)
RickettsiaOrthoExpect	*	scalar	0-1	0%	Similarity to Rickettsia (mitochondrial ancestor) protein (blast expect score)
Computational					
MitoDomain	*	categorical	{-2,-1,0,1}	0%	Protein domain (PFAM) found exclusively in mitochondrial proteins (1), found shared in mito and nonmito proteins (-1), found exclusively in nonmito proteins (-2), or no information (0)
MitoPred		scalar	0-100	0%	Predicted to be mitochondrial by MitoPred program
MitoPred_Ortho		scalar	0-101	0%	Mouse ortholog predicted to be mitochondrial by MitoPred program
TargetP_Mito		categorical	{0, 1}	0%	Mouse ortholog predicted to be mitochondrial by TargetP program
Targetp	*	categorical	{0, 1,2}	0%	Both human and mouse orthologs predicted mito by TargetP (2), only one predicted (1), or neither predicted (0)
ERRA_motif	*	categorical	{0, 1}	0%	ERRalpha motif present in promoter
GABPA_motif		categorical	{0, 1}	0%	GABPA motif present in promoter
NRF1_motif		categorical	{0, 1}	0%	NRF1 motif present in promoter
YY1_motif		categorical	{0, 1}	0%	YY1 motif present in promoter

Table 1: Features related to mitochondrial localization.

Asterisk indicates selection in 8 feature set.

References

1. Jaakola, T. 6.867 Machine Learning Lecture Notes. (2006).
2. Hastie, T., R, T. & H., F.J. The Elements of Statistical Learning. (July 30, 2003).
3. Calvo, S. et al. Systematic identification of human mitochondrial disease genes through integrative genomics. *Nat Genet* **38**, 576-82 (2006).
4. Joachims, T. Making large-Scale SVM Learning Practical. Advances in Kernel Methods - Support Vector Learning. *B. Schölkopf and C. Burges and A. Smola (ed.) MIT-Press.* (1999).
5. Vezhnevets, A. GML AdaBoost Matlab Toolbox. *research.graphicon.ru/machine-learning/gml-adaboost-matlab-toolbox.html* (Dec 2006).
6. Alexandre, L. Boosting and Bagging. *gnomo.fe.up.pt/~nnig/papers/boo_bag.pdf* (May 4, 2004).
7. Yan, R. MATLABArsenal: A MATLAB Package for Classification Algorithms. *finalfantasyxi.inf.cs.cmu.edu/MATLABArsenal/MATLABArsenal.htm* (2006).