Supplementary experimental protocol:

Characterization of the Human Heart Mitochondrial Proteome

Methodology and Bioinformatics. Given the limitations of 2D gel maps for the identification of hydrophobic, low molecular weight, acidic and basic proteins, we recently reported an alternative strategy to elucidate the mitochondrial proteome that has been further developed in the current study. Mitochondria were solubilized in detergent and proteins were initially separated as multimeric complexes by sucrose gradients. The high purity of these mitochondria was assessed by Western analysis with antibodies directed against selected proteins of the cytoplasm, plasma membrane and other organelles (Supplementary Fig. 1). The sucrose gradient fractions and the n-dodecyl-β-D-maltoside-insoluble protein pellet were resolved on 1D gels and proteins were identified by high throughput MALDI peptide mass fingerprinting (PMF). In the current study, more protein has been loaded onto each gel lane, entire gel lanes have been processed by manually slicing the gel into strips from the bottom of the gel to the loading wells (see images at http://www.mitokor.com/files/), and LC/MS/MS has been added for protein identification from in-gel tryptic digests. The additional dimension of separation provided by sucrose gradient fractionation permitted higher protein loads per gel lane and, coupled with the large number of gel slices processed (~65 × ~1 mm slices per lane), these factors have enhanced our ability to detect proteins compared with a recent study on yeast mitochondria. Those advances exemplify a recurrent theme in proteomic analysis whereby the number of protein identifications appears to be a function of the amount of starting material, number of degrees of separation and the extent of sampling.
Some proteins are resistant to PAGE resolution even in a single dimension. For these proteins, non-gel based strategies such as multidimensional protein identification technology (“MudPIT”, for example, strong cation exchange followed by reversed phase chromatography) offer some advantages when applied to protein digests. Nevertheless, by using the combination of sucrose gradients and 1D gels, two significant advantages over MudPIT are achieved. Not only can confidence be gained in the identification of a protein based on the relationship between its molecular weight and gel mobility (See Supplementary Fig. 2), but the sucrose fraction in which a protein is identified suggests whether it forms a multimeric protein complex. As an illustration, we previously confirmed that the 30 kDa protein, prohibitin, participates in a high molecular weight complex called PHB which acts as a molecular chaperone for respiratory complex assembly, and additional examples are reported in the main text.

Interestingly, the number of SEQUEST peptide matches was found to be directly proportional to the distribution of the complexes in a given sucrose gradient fraction. The tryptic peptide profile presented in Supplementary Figure 3 is almost identical to our previously published distribution the oxidative phosphorylation complexes as detected by their Western blot intensities using monoclonal antibodies directed at individual subunits. These data indicate partial resolution of the intact complexes based on their size, with the maximal concentration of Complexes I and II in fractions 3 and 7, respectively. While a relationship between the relative amounts of proteins in complex mixtures and the number of peptides identified has been noted previously, we are not aware of any previous studies that so clearly demonstrate this relationship as validated by Western data. Thus, this correlation is used as a rough measure of relative protein abundances. We do this with the caveat that the complexities of protein biochemistry, including varying electrophoretic behavior, protease susceptibility, and peptide ionization may be confounding issues.
The choice of protein database used for MALDI and LC/MS/MS searches was dictated by the goal to interrogate as many isoforms, splice-variants and processed versions of each human gene product as possible. In particular, a number of recent reports have indicated that alternative splicing or post-translational modification events can result in protein targeting to the mitochondrion. Therefore, a human subset of the GenBank NR protein database was selected, despite a number of clear disadvantages relative to smaller, more highly curated datasets such as SwissProt and Ensembl in terms of quality of annotations, sequence quality and ease of bioinformatic analysis. Justification for this type of approach may be seen in Supplementary Table 3, which comprises a list of 24 different proteins for which there was significant mass-spectral evidence for the existence of 2 or more sequence variants.

The identification of proteins was based upon MALDI PMF using the Protein Prospector search algorithm and LC/MS/MS using the SEQUEST and SonarMSMS algorithms. The informatics strategy employed to reduce the initial dataset of over 3000 unique GenBank identifications to the final dataset of 615 proteins contained a number of distinct phases. Firstly, a set of more stringent scoring thresholds was applied to both SEQUEST and SonarMSMS results in order to eliminate most of the poorly supported database hits. Secondly, proteins identified on the basis of 1 or 2 SEQUEST peptide matches, whose molecular weight differed significantly from the mean value for that particular region of the PAGE gel in which they were identified (See Supplementary Fig. 2) were eliminated, using a spreadsheet-based screening method (see under “Data Integration” below). However, those protein identifications with highly reliable spectral data in this data set were retained, because the possibility existed that splice-variants or truncated versions of proteins that have not been deposited in public databases might be detected in the course of this study. All other proteins identified on the basis of lower confidence peptide SEQUEST data were subject to rigorous manual inspection and in some cases synthetic analogues were prepared and relative retention times and
fragmentograms were compared (data available at http://www.mitokor.com/files/).

Thirdly, proteins identified on the basis of lower confidence SonarMSMS scores whose fragmentogram data survived careful manual inspection were also retained. Fourthly, a comprehensive BLAST\textsuperscript{o} analysis was performed on the remaining set of proteins to eliminate identical or near-identical sequences (due to the fact that a later version of the GenBank protein database was used for SonarMSMS searches) and also to identify a significant number of GenBank entries whose functional annotation was uncertain or unknown. Fifthly, a careful analysis of the evidence for the identification of multiple isoforms or sequence variants of the same gene product was undertaken with the assistance of a customized SQL-based algorithm. Cases where the distinguishing peptides differed by 1 u or less as the result single or double amino acid substitutions were eliminated due to the mass-accuracy limitation of the ion-trap LC/MS/MS instrument used. All peptides satisfying the above criteria and that have been used to identify proteins have been listed on our website; see http://www.mitokor.com/files/.

The extent of sequence coverage and of distinct peptide assignments provides a measure of confidence in protein identification.

Two key observations were made regarding the relative abilities of the three different search algorithms employed to identify the final tally of 615 proteins: (a) all proteins in the final list were identified by at least one of the two LC/MS/MS algorithms, rendering the MALDI data largely redundant, and (b) a significant number of protein identifications were detected by only one LC/MS/MS algorithm (123 for SEQUEST, 93 for SonarMSMS), suggesting an advantage in combining the more statistically-based SonarMSMS approach with the SEQUEST method which was seen to have greater sensitivity in detecting reliable single peptide hits.
Experimental protocols

Sample Preparation. Immediately prior to processing and analysis by mass spectrometry, the concentrated gradient fractions and the solubilized pellet were subjected to electrophoretic resolution on NuPAGE gels using ultraclean reagents. Buffers were made using HPLC grade water, and the electrophoresis apparatus and staining box were set aside for these samples. Aliquots (25 μg) of each concentrated gradient fraction were loaded on a 4-12% NuPAGE gel and run at 25 mA for 1 h, then 35 mA for another 1 h 20 min. Gels were fixed for 10 min (40% methanol, 10% acetic acid), washed three times for 5 min in HPLC grade water, stained with colloidal Coomassie for 10-15 sec, and then partially destained in water.

Although we have previously used robotics for gel band excision, destaining, reduction, alkylation and digestion, in the current investigation we opted for a manual approach as described by Simpson and co-workers to comprehensively analyze entire gel lanes irrespective of the position of Coomassie-stained bands. To avoid keratin contamination, all manipulations were performed while wearing latex gloves, shower caps and lab coats. After imaging, the lightly Coomassie-stained electrophoretic gels were imaged, placed on a light box in a laminar flow hood on a plastic cutting mat with a 65 × 1 mm grid placed underneath. Starting at the bottom of the gel, approximately 1 mm slices were excised across the entire width of a gel lane with a clean razor, further cut into approximately 1 mm cubes and transferred to 500 μL microcentrifuge tubes that had been prewashed with 50:50 water: acetonitrile. This procedure was progressively continued to the top of the gel to ensure comprehensive coverage of all proteins in the gel lane. Although most gel slices were 1 mm wide, discrete bands were selectively excised whenever they were encountered. Slightly wider slices were taken near the top of the gel to compensate for lower protein concentration. This procedure resulted in 50-64 slices for each of the 12 lanes processed (corresponding to sucrose fractions 1-11 and
the pellet, images available on the authors’ website; see http://www.mitokor.com/files/). The gel pieces were incubated with 200 µL destain (25 mM ammonium bicarbonate, 25% acetonitrile) at 37°C for 45 min. The destain was decanted and another cycle of destaining performed if necessary to remove residual coloration. The gel pieces were then dried on a Genevac concentrator using the “cool heat” setting (about 30 min). To minimize handling, and the risk of keratin contamination, losses in transfer or manipulation errors on this large set of samples (~700 × ~1mm gel pieces), reduction and alkylation steps were not performed. The dried gel pieces were slightly moistened with 5 µL 50 mM ammonium bicarbonate, 5% acetonitrile and 5 µL of freshly prepared ice-cold Promega modified trypsin (0.1 mg/mL in 50 mM ammonium bicarbonate, 5% acetonitrile) was added. The gel pieces were allowed to soak up the trypsin solution for 10 min, and then were fully hydrated with a 65 µL aliquot of 50 mM ammonium bicarbonate, 5% acetonitrile. After an overnight incubation at 37°C, the digestion was terminated by addition of 7.5 µL 10% acetic acid, followed by brief vortexing and light centrifugation in a microcentrifuge. The digest supernatants were subsequently transferred to secondary pre-washed 500 µL microcentrifuge tubes and carefully concentrated using the Genevac to final volumes of 10-20 µL. At no stage were the digests taken to dryness to avoid irreversible adsorption of low abundance peptides to the walls of the tubes. The concentrated digests were then carefully decanted to avoid particulates and transferred to the wells of a V-bottom 220 µL polypropylene microtiter 96 well plate.

**Analysis of MALDI data.** MALDI spectra were analyzed with the Protein Prospector algorithm, using a human subset of the GenBank NR protein database (11/28/01 release). Searches were automated using the Proteomics Solution 1 software on Windows NT. Two types of searches were performed as described previously. Briefly, for method A, proteins for which at least 4 peptides were matched within a
mass tolerance of 100 ppm and a MOWSE score of at least 10,000 were considered hits. In method B, a first-pass step was performed in which proteins for which at least 3 peptides were matched within a mass tolerance of 150 ppm were selected for recalibration using the IntelliCal algorithm. A subsequent second-pass threshold search was applied to obtain protein hits in which at least 4 peptides were matched within a mass tolerance of 15 ppm and a MOWSE score of at least 100. Results from all searches were migrated from Microsoft Access data files to a centralized set of Oracle database tables using parsing programs written in Visual Basic, Perl and PHP. Data from 5,112 Protein Prospector database matches was collected in this manner. A total of 654 distinct proteins were found by the IntelliCal search method (Method B above) and an additional 95 distinct proteins were identified by non-IntelliCal search (Method A).

**Analysis of LC/MS/MS Data.** LC/MS/MS data was analyzed with the two independent searching algorithms SEQUEST\(^1\) and SonarMSMS\(^m\), respectively.

**A: SEQUEST:** A total of 870,534 spectra were collected from 701 LC/MS/MS runs and were searched with the SEQUEST algorithm running on Windows NT, using a human subset of the GenBank NR protein database (11/28/01 release). A set of web-based batch programs, written in Perl, were designed to enable automated raw data preparation, searching and reporting of SEQUEST runs. All results were migrated to a centralized set of Oracle database tables, yielding a total of 263,590 peptide matches in which the SEQUEST \(\Delta\)corr value (a measure of the difference in cross-correlation (Xcorr) between the best and next best peptide match) was \(\geq 0.1\). A set of Oracle PL/SQL queries was used to generate a summary report of the entire SEQUEST dataset, and to organize the results based on the number of unique peptides detected for each protein identification which satisfied the following search criteria: Charge State 1, Xcorr...
≥ 1.7; Charge State 2, Xcorr ≥ 2.0; and Charge State 3, Xcorr ≥ 3.0. A total of 2,059 unique protein identifications that satisfy the above Xcorr criteria were obtained, 509 of which matched 2 or more different peptides. The remaining 1550 protein identifications were based on a single peptide hit.

B: SonarMSMS. Spectral data from each of 701 LC/MS/MS runs were first merged into a set of input files, and were subjected to SonarMSMS algorithm running on Linux, using a human subset of the GenBank NR protein database (5/9/02 release), with oxidation of methionine specified as a differential modification. A total of 2.2 GB of spectral data was thus processed in batch mode, using a web-client emulation program written in Perl. Data corresponding to the top protein hits (Protein Expect value <1) for each LC/MS/MS run corresponding to a gel slice, as well as results for all peptide identifications within those protein hits, were extracted from the SonarMSMS results and migrated to an Oracle database. A total of 2026 unique protein identifications were obtained by this search strategy, and these were very similar to those obtained with SEQUEST searching.

Data Integration. Protein matches from Protein Prospector, SEQUEST and SonarMSMS searches were compared by GenBank identification to establish an initial level of cross-validation. This large “unedited” dataset of approximately 3,000 different GenBank identifications was initially filtered by applying the following set of criteria:

1. High confidence mass spectral data for proteins including (a) ≥ 3 peptide matches obtained by SEQUEST for a given protein identification; (b) Protein E value ≤ 1 x 10^{-3}, 99.9% confidence) by SonarMSMS search; (c) High scoring single or double peptide hits by SEQUEST (Xcorr ≥ 3.0 for (+2) charge state, Xcorr ≥ 4.0 for (+3) charge state);
(2) 1 and 2-peptide matches in the correct molecular weight range with lower SEQUEST scores which satisfied rigorous manual inspection of MS/MS fragmentograms.

(3) Lower confidence SonarMSMS scores in the correct molecular weight range with Peptide E values $\leq 1 \times 10^{-2}$ (99% confidence), and which satisfied rigorous manual inspection of MS/MS fragmentograms.

Comments on Criteria (2) and (3): As it was observed that the relationship between the gel slice and the mean molecular weight of the most confidently identified proteins was almost linear, especially in the low-medium molecular weight range (Supplementary Fig. 2), proteins identified by LC/MS/MS for each sucrose density gradient fraction were sorted by gel slice (i.e. gel mobility) in a Microsoft Excel spreadsheet. A macro program was written to calculate the mean molecular weight at each position in the gel by examining the 20 closest protein identifications, removing the 5 highest and lowest molecular weight outliers, and calculating the average molecular weight values of the remaining 10 proteins. Comparison of the molecular weight of each protein identification with the corresponding mean value for that region of the gel enabled convenient identification of outliers differing by more than 15kD. In general, these proteins tended to be in the higher molecular weight range (>100 kDa), with the 1-3 peptides randomly distributed throughout the sucrose gradient fractions or gel slices. A smaller subset of these outliers, in contrast, was supported by strong spectral evidence, indicating that these proteins may represent truncated sequences or oligomers.

Also, in addition to manual evaluation of the MS/MS data, a random set of 56 synthetic peptides (Mimotopes, Clayton, Victoria, Australia (55 peptides) and ResGen, Huntsville, AL (1 peptide)) was generated corresponding to tryptic peptides whose initial identification fell under criteria (1c) and (2) or initial SEQUEST criteria. These peptides were subjected to similar LC/MS/MS conditions as used for the tryptic peptides and relative retention times and fragmentograms examined. Near identical
MS/MS data validated 25 peptide identifications and 3 were further validated by subsequently satisfying criteria (1a) and/or (1b). Disparate data invalidated 15 identifications and the remaining identifications were neither validated nor invalidated by the synthetic peptide data because similarities in the overall appearance of the fragmentograms were countered by one or more non-matching ions (see examples on our website http://www.mitokor.com/files/). These ions may be the result of either sequence differences or co-eluting isobaric peptides.

(4) The dataset was next sorted by GenBank annotation, and protein sequences were subjected to a comprehensive BLAST analysis in order to identify essentially identical protein sequences containing different annotations, as well as to identify a number of sequences whose annotation inferred that it was an “unknown” protein.

(5) All single peptide hits (singletons) identified by SEQUEST were subjected to BLAST searches to establish whether any other proteins contained the same peptide.

Comments on Criterion (5): A customized program was written to determine whether LC/MS/MS results could identify more than one isoform or variant of a particular protein, based on evidence for unique peptide(s) characteristics of a given variant. For this, an Oracle database table was constructed which contained all possible tryptic peptides for proteins in the human subset of GenBank. A web-based query interface was designed to enable computation of unique tryptic peptides corresponding to any 2 protein identifications of interest (isoforms, splice-variants etc). Secondly, the program identified any of those unique tryptic peptides that were also detected by LC/MS/MS searches. Thus, this software tool was used to efficiently identify multiple variants of a single protein in those cases where sufficient MS evidence was available. Isoform specific peptides were examined and those differing by single or double amino acid substitutions corresponding to a 1 mass unit or less difference in molecular mass were identified (e.g., L/I, Q/K, D/N, E/Q etc.). The corresponding proteins were then
eliminated as less than unit mass accuracy is not attainable on our Ion Trap mass spectrometer and thus the MS/MS fragmentation patterns for these peptide pairs are indistinguishable.

Motif Analysis. All protein sequences corresponding to the edited list of MS hits were subjected to Hidden Markov Model (HMM) motif searches against the Pfam database, using a Perl batch script running on Linux.

Calculations based on primary structure. Protein molecular weight and hydrophobicity (Kyte-Doolittle) calculations were written in Perl. The “iep” module from the EMBOSS suite was used to calculate isoelectric point (pI) values.

Mitochondrial proteome data storage and retrieval system. Results of all MALDI (Protein Prospector) and LC/MS/MS (SEQUEST, SonarMSMS) searches were compiled and stored in a schema of Oracle database tables. A set of web-based querying and reporting interfaces was assembled, using the PHP programming language in conjunction with Apache web server (Linux). Mass spectrometry search scores for each protein identification were linked to corresponding data for motif (HMM) searches, LocusLink information, primary protein sequence and original search result outputs.
References for Supplementary Information.


h. Tolkunova, E., Park, H., Xia, J., King, M.P. & Davidson, E. The human lysyl-tRNA synthetase gene encodes both the cytoplasmic and mitochondrial enzymes


