Systematic analysis of NDUFAF6 in complex I assembly and mitochondrial disease

Andrew Y. Sung[1], **Rachel M. Guerra**[2], **Laura H. Steenberge**[3], **Charlotte L. Alston**[4,5], **Kei Murayama**[6,7], **Yasushi Okazaki**[7], **Masaru Shimura**[6,8], **Holger Prokisch**[8,9], **Daniele Ghezzi**[10,11], **Alessandra Torraco**[12], **Rosalba Carrozzo**[12], **Agnès Rötig**[13], **Robert W. Taylor**[4,5], **James L. Keck**[1], **David J. Pagliarini**[2,14,15,*]

[1]Department of Biomolecular Chemistry, University of Wisconsin School of Medicine and Public Health, Madison, WI 53706, USA

[2]Department of Cell Biology and Physiology, Washington University School of Medicine in St. Louis; St. Louis, MO 63110, USA

[3]Department of Biochemistry, University of Wisconsin-Madison; Madison, WI 53706, USA

[4]Mitochondrial Research Group, Translational and Clinical Research Institute, Faculty of Medical Sciences, Newcastle University; Newcastle upon Tyne, United Kingdom

[5]NHS Highly Specialised Service for Rare Mitochondrial Disorders, Newcastle upon Tyne Hospitals NHS Foundation Trust; Newcastle upon Tyne, United Kingdom

[6]Department of Metabolism, Chiba Children's Hospital; Chiba, Japan

[7]Diagnostics and Therapeutic of Intractable Diseases, Intractable Disease Research Center, Graduate School of Medicine, Juntendo University, Tokyo, Japan

[8]Institute of Neurogenomics, Computational Health Center, Helmholtz Zentrum Munchen; Neuherberg, Germany

[9]School of Medicine, Institute of Human Genetics, Technical University of Munich; Munich, Germany

[10]Department of Pathophysiology and Transplantation, University of Milan; 20122 Milan, Italy

[11]Medical Genetics and Neurogenetics Unit, Fondazione IRCCS Instituto Neurologico Carlo Besta; 20126 Milan, Italy

[12]Unit of Cell Biology and Diagnosis of Mitochondrial Disorders, Laboratory of Medical Genetics, Bambino Gesù Children's Hospital, IRCCS; Rome, Italy

*Correspondence and requests for materials should be addressed to D.J.P. pagliarini@wustl.edu (D.J.P).

[13]Université Paris Cité, Imagine Institute, INSERM UMR 1163; Paris, France

[14]Department of Biochemistry and Molecular Biophysics, Washington University School of Medicine; St. Louis, MO 63110, USA.

[15]Department of Genetics, Washington University School of Medicine; St. Louis, MO 63110, USA.

Abstract

Isolated complex I (CI) deficiencies are a major cause of primary mitochondrial disease. A substantial proportion of CI deficiencies are believed to arise from defects in CI assembly factors (CIAFs) that are not part of the CI holoenzyme. The biochemistry of these CIAFs is poorly defined, making their role in CI assembly unclear and confounding interpretation of potential diseasecausing genetic variants. To address these challenges, we devised a deep mutational scanning approach to systematically assess the function of thousands of *NDUFAF6* genetic variants. Guided by these data, biochemical analyses, and cross-linking mass spectrometry, we discovered that the CIAF NDUFAF6 facilitates incorporation of NDUFS8 into CI and reveal that NDUFS8 overexpression rectifies NDUFAF6 deficiency. Our data further provide experimental support of pathogenicity for seven novel *NDUFAF6* variants associated with human pathology and introduce functional evidence for over 5,000 additional variants. Overall, our work defines the molecular function of NDUFAF6 and provides a clinical resource for aiding diagnosis of NDUFAF6-related diseases.

Introduction

NADH:ubiquinone oxidoreductase, also known as complex I (CI), plays a vital role in oxidative phosphorylation (OxPhos). CI serves as the primary entry point for electrons into the mitochondrial respiratory chain and, together with complexes II-IV, generates and maintains the proton motive force that drives ATP synthesis[1]. The essentiality of CI in mitochondrial function is reflected in its disease relevance. Isolated CI deficiencies are collectively the most common cause of primary mitochondrial disease[2], which affects an estimated 1:5,000 people[3]. Strikingly, up to half of CI deficiencies are thought to arise from pathogenic variants affecting proteins that coordinate CI assembly rather than the structural subunits themselves[4,5]. Together, these CI assembly factors (CIAFs) assist the choreographed construction of the CI holoenzyme[6-10]. Despite their prominent role in CI disease pathogenesis, the precise molecular functions of these CIAFs are poorly defined.

A barrier to studying CI assembly is that many CIAFs are atypical members of their respective protein families. Some of these CIAFs are predicted to be pseudoenzymes — proteins possessing folds similar to known enzymes but lacking catalytic activity — whose molecular functions are notoriously difficult to define[11,12]. Thus, there is a need for effective, systematic analyses that can help reveal how CIAFs guide the assembly process. The lack of mechanistic clarity for CIAFs also translates to challenges in the clinical sphere. From a diagnostic standpoint, poor understanding of CIAF function complicates interpretation of disease gene variants. Therapeutically, lack of molecular insight hinders the ability to devise targeted interventions to treat these CI assembly defects[13,14].

To address these issues, we devised a deep mutational scanning (DMS) approach for the functional characterization of CIAFs. DMS is a high-throughput method to measure the fitness of thousands of variants in a protein of interest and has not yet been applied to the study of mitochondrial proteins[15]. Regions of the protein that are sensitive to mutations suggest functional relevance. In addition to functional insights, these DMS data also provide robust empirical evidence for evaluating variant pathogenicity and can therefore serve as a diagnostic resource.

We applied this approach to the CIAF NDUFAF6 (AF6), a predicted pseudoenzyme with a prenyltransferase fold[16,17] that has proven challenging to study through more traditional methods and was thus a particularly promising candidate for this systematic DMS approach. AF6 was first identified as a CIAF through phylogenetic profiling[18]. Subsequent work showed that AF6 has two splice isoforms, one mitochondrial and one cytoplasmic. The mitochondrial isoform is peripherally associated with the matrix face of the inner mitochondrial membrane[19]. This isoform is also sufficient to restore CI activity in patient fibroblasts[19] and implicated in early-stage CI assembly[19,20]. However, the molecular mechanism of AF6 and the exact assembly step it mediates remain elusive[6].

In this study, we used DMS as a launching point to characterize AF6. Guided by fitness data from thousands of AF6 variants, we designed a series of experiments to interrogate this CIAF's molecular function. We reveal that AF6 is a *bona fide* pseudoenzyme that directly binds to the CI subunit NDUFS8 and guides its incorporation into the 125 kDa assembly intermediate. We further show that loss of AF6 can be overcome by overexpression of NDUFS8, demonstrating the specificity of this interaction and suggesting a potential therapeutic avenue to bypass the assembly defect in cases of AF6 dysfunction. Finally, we provide the first experimental support of pathogenicity for seven novel AF6 variants and add phenotypic annotation for over 5,000 additional unannotated AF6 variants, thereby creating a comprehensive clinical resource for supporting the diagnosis of AF6-related diseases. Beyond the biochemical and clinical insights we provide for NDUFAF6, our work also establishes a framework for leveraging DMS to define orphan mitochondrial protein function and predict variant pathogenicity.

Results

Deep mutational scanning of AF6

To systematically measure the fitness of each single-amino-acid variant of AF6, we performed DMS using a pooled growth competition. In brief, AF6 knockout (KO) HEK293T cells were transduced with an AF6 variant library and passaged for three to six generations (roughly 4 and 8 doublings, respectively) in media containing galactose as the major carbon source. Growth in galactose media is reliant on OxPhos[21,22], which requires functional CI. Thus, cells expressing functional variants of AF6 can grow while cells expressing non-functional variants cannot. Deep sequencing was then used to determine the read counts for each variant before and after the galactose growth selection. Fitness scores for each variant were calculated as the natural log ratio between a variant's output count and input count, relative to the wild-type variant. Average fitness scores weighted by replicate-specific error estimates were then calculated for each variant[23]. The final dataset

comprised fitness estimates for 5,714 of 5,780 possible variants in the target DMS region (Fig. 1a, Extended Data Fig. 1, Table S1).

To set up the DMS system, we first created two-independent AF6 knockout (KO) cell lines in a HEK293T background. We demonstrated through Sanger sequencing, western blot, in-gel CI activity assay, and Seahorse analysis that these KO lines are deficient in AF6 and, consequently, CI (Fig. 1b-e). As validation for the galactose growth selection, we showed that these KO cells exhibit a growth defect in galactose media, which could be complemented by reintroduction of AF6 cDNA under the native promoter sequence (Fig. 1f). The native promoter sequence drove sufficient expression of AF6 to rescue CI assembly, but AF6 expression was difficult to detect by western blot when compared to control wild-type cells. Our DMS library was site-saturated (i.e., all possible amino acids at each position) and covered residues 45-333 of the mitochondrial isoform of AF6. The first 44 residues constitute the predicted mitochondrial targeting sequence (MTS)[24] and were therefore excluded. The DMS experiment was performed under the native AF6 promoter sequence, with five pre-selection biological replicates and ten post-selection biological replicates (Extended Data Fig. 2a).

We evaluated the robustness of our DMS data in several ways. First, we assessed the sequencing depth of the input samples. The input samples had a median read depth of >500x, with 5,635 of the 5,714 (99%) of the variants having a read depth of at least 100x in all five replicates (Extended Data Fig. 2b). Second, we calculated pairwise Pearson correlation coefficients of the fitness scores across the five replicates, showing that all five replicates are highly correlated with one another with an average Pearson's r of 0.897 (Extended Data Fig. 2c). Third, we examined the overall distribution of fitness scores, which showed a strong bimodal distribution, demonstrating successful separation of tolerated and deleterious substitutions. The fitness data were further modeled using a Gaussian mixture model (GMM) with three components corresponding to variants with A) strong, B) intermediate, or C) low functional impact (Extended Data Fig. 2d-e). We then evaluated the fitness scores of premature termination substitutions in the context of this model. Of the 286 nonsense variants, 274 fell within three standard deviations of the mean of the strong functional impact component [−1.729, −0.842] (Extended Data Fig. 2f). As premature termination appears to have a strong deleterious effect even at the C-terminal-most position (Extended Data Fig. 1, residue 333), all premature termination substitutions likely have a strong functional impact. The 12 nonsense variants outside the strong functional impact range thus likely represent a long tail of the distribution of variants with a strong functional impact rather than nonsense variants with intermediate functional impact. Altogether, these analyses ruled out obvious bottlenecks in the experimental pipeline and validated our overall experimental scheme.

We next analyzed the data in the context of a predicted AlphaFold model of AF6[25,26]. We used residues 56-333 of the AlphaFold AF6 model, of which 264 of 275 residues were classified as having very high model confidence (pLDDT > 90) and the remainder classified as having high confidence (pLDDT > 70), indicating a high-quality model[26]. For this analysis, we found it informative to first compute an aggregate measure of "mutational sensitivity" by tallying the number of deleterious substitutions at each residue

position and scaling it linearly to a value between 0 and 9 (Extended Data Fig. 3a, Table S2). To maximize the signal-to-noise ratio for this calculation, we considered deleterious substitutions to be those below the previously described threshold for variants with strong functional impact (fitness −0.842). To test whether our data were consistent with the predicted secondary structure of AF6, we then mapped these mutational sensitivity scores onto the predicted AF6 model. Residues in alpha helices with an exposed face and a buried face showed a periodicity in mutational sensitivity of 3-4 residues. This pattern aligns well with the ~3.6 residues-per-turn geometry of alpha helices and demonstrates an inverse relationship between mutational sensitivity and solvent accessible surface area (Extended Data Fig. 3b). Next, we assessed the effects of proline substitutions, well-known helix breakers[27] that should be poorly tolerated in alpha-helical regions of the structural model. Indeed, we see that the fitness scores of prolines are particularly decreased for residues involved in alpha helices (Extended Data Fig. 3c). Finally, comparison of mutational sensitivity and solvent-accessible surface area (SASA) for all residues shows an expected inverse proportionality between these two metrics. That is, buried residues that comprise the hydrophobic core of the protein (SASA 20%) tend to have high mutational sensitivity whereas surface-exposed residues (SASA > 20%) tend to have low mutational sensitivity (Extended Data Fig. 3d-e). Taken together, these results corroborate the predicted AlphaFold structure for AF6 and thus support our using it as a structural model for further analysis.

In addition to supporting the predicted structural model for AF6, our data recapitulated many other known biochemical principles (Extended Data Fig. 4). For example, alanines are predominantly found in the hydrophobic core of the protein and tolerated alanine substitutions are restricted to other residues with small side chains (Extended Data Fig. 4a). Aromatic residues tend to be restricted in mutational tolerance to other aromatic side chains (Extended Data Fig. 4b). Hydrophobic residues such as leucine, isoleucine, and valine are also largely found in the hydrophobic core and only tolerate substitutions to other residues with hydrophobic side chains (Extended Data Fig. 4c). Charged residues tend to tolerate substitutions with the same charge but not charge reversal (Extended Data Fig. 4d). This pattern is less stark, however, as most charged residues are surface residues, which tolerate substitutions better than buried residues. The fitness scores for each of these amino acids align well with expected scores for substitutions based on similarity matrices such as BLOSUM62[28]. We also observe that whereas mutational sensitivity is a good predictor for conservation, the reverse is not necessarily true (Extended Data Fig. 4e-f).

DMS highlights functionally relevant protein regions

Bioinformatic analyses predict AF6 to be a pseudoenzyme. Pseudoenzymes are found across nearly every protein family and have been shown to play key roles in protein complex assembly, allosteric regulation, substrate sequestration, and more[11,12]. Despite their ubiquity, the diverse molecular actions of these proteins are notoriously difficult to predict and define. Here, by integrating our DMS data with the structural model of AF6, we conducted a systematic search for protein regions that are sensitive to mutations, suggesting functional relevance. We began our analysis with the vestigial "active" site, the putative hydrophobic pocket, and surface-exposed residues (Fig. 2a).

AF6 possesses high sequence homology to members of the head-to-head (HH) synthase family of prenyltransferases[16], which have characteristic aspartate-rich active sites and a hydrophobic substrate binding pocket[29]. AF6 lacks three of the four key catalytic aspartate residues required for canonical HH condensation reactions[29] and, as such, is unlikely to retain canonical HH synthase activity (Fig. 2b). Our DMS data also show low mutational sensitivity in the vestigial active site motifs, supporting the loss of catalytic activity (Fig. 2a). We then examined the putative hydrophobic binding pocket of AF6. Residues near the entrance of this pocket also showed comparable mutational sensitivity relative to other buried residues (Fig. 2a). Furthermore, comparison of the putative binding pocket of AF6 with that of a structurally similar HH synthase revealed that the AF6 pocket is much shallower and unlikely able to accommodate similar substrates (Fig. 2c). Altogether, these data argue against a catalytic role for AF6 in CI assembly, novel or canonical, and support the pseudoenzyme hypothesis.

We next analyzed the mutational sensitivity of all surface residues (SASA > 20%). As noted previously, we observed a strong inverse relationship between the SASA and the mutational sensitivity of a given residue, with 264 out of 284 (92%) residues having either low mutational sensitivity (< 2) or low SASA (20%). We then examined the 22 surface residues that fell outside this pattern of inverse proportionality (i.e. residues with a mutational sensitivity 2 and SASA > 20%), which we reasoned may suggest protein-protein interaction (PPI) interfaces. We mapped these residues on the predicted structure of AF6 and identified two clusters on the protein surface comprising residues 173, 234, 239, 242, 244, 256, 305, and 310 for the first cluster and residues 77, 80, 125, 127, 139, 140, 191, 194, 197, and 281 for the second cluster (Fig. 2d). Interestingly, these two surface patches showed little correlation between mutational sensitivity and conservation, nor did they stand out from examining the surface electrostatics (Fig. 2e-f). This demonstrates the utility of experimental approaches such as DMS for finding regions of interest that may otherwise not be immediately apparent, particularly for proteins such as pseudoenzymes.

We concluded this analysis with an examination of the alpha-helix at the C-terminus of AF6. This C-terminal helix (CTH) is amphipathic[30] and forms part of the predicted membrane association interface of AF6[17]. Our DMS data show that AF6 does not tolerate any length of truncation at the C-terminus (Extended Data Fig. 1, residues 319-333). This suggests an important role for the CTH in CI assembly, likely by driving appropriate membrane localization. The AF6 CTH is explored further below.

AF6 enables NDUFS8 incorporation into CI

Strikingly, despite a plethora of large-scale PPI studies, no physical mitochondrial interactors have been reported for AF6[31]. We reasoned that AF6 may engage in a weaker or more transient interaction with a CI subunit, consistent with AF's absence in mature CI. As such, we used cross-linking mass spectrometry (XL-MS) to identify the potential binding partner(s) of AF6. We first isolated mitochondria from wild-type cells and AF6 KO cells overexpressing an AF6 construct with a C-terminal FLAG-tag. We then treated the mitochondria with disuccinimidyl sulfoxide (DSSO), a chemical crosslinker, performed FLAG immunoprecipitations, and analyzed the samples by mass spectrometry to identify

enriched proteins. Our results revealed significant and specific enrichment of the core CI subunit NDUFS8. Other CI subunits or CIAFs were only modestly enriched (log2 fold-change < 2) (Fig. 3a, Extended Data Fig. 5, Table S3).

NDUFS8 is a core subunit of the quinone binding module (Q module), one of three functional modules that comprise the CI holoenzyme. The two other modules are the NADH-oxidizing dehydrogenase module (N module) and the proton pump module (P module), which is further subdivided into the Pp-a, Pp-b, Pd-a, and Pd-b modules[32]. NDUFS8 coordinates two iron-sulfur (Fe-S) clusters that form part of the Fe-S cluster chain responsible for shuttling electrons from the N module to the Q module[1].

To further test the AF6-NDUFS8 interaction experimentally, we performed a yeast two-hybrid (Y2H) assay to assess the physical interaction between several Q module proteins: AF6, NDUFS8, NDUFS7, and NDUFAF3. The Y2H assay recapitulated the specific interaction between AF6 and NDUFS8 (Fig. 3b). We then tested the ability of NDUFAF6 constructs with mutations in the surface patches highlighted by DMS for their ability to interact with NDUFS8. Our results show that alanine mutations in either of the two surface patches disrupt binding of NDUFAF6 to NDUFS8, whereas alanine mutations in regions of the surface with low mutational sensitivity do not (Fig. 3c, d).

Given the interaction between AF6 and NDUFS8 shown by XL-MS and Y2H, as well as the lack of essential residues in the vestigial active site shown by DMS, we reasoned that AF6 likely mediates the physical incorporation of NDUFS8 into the Q module. Q module assembly proceeds through several defined subassemblies with predicted molecular weights of 86 kDa, 125 kDa, 159 kDa, and 224 kDa. As NDUFS8 first appears in the 125 kDa intermediate, we hypothesized that AF6 is necessary for the transition from the 86 kDa intermediate to the 125 kDa intermediate (Fig. 3e). Loss of AF6 should thus stall assembly at the 86 kDa intermediate.

To test this, we performed blue native western blots tracking the migration of several Q module subunits and assembly factors across HAP1 wild-type and CIAF KO cell lines (NDUFAF5 KO, AF6 KO, and TIMMDC1 KO). Loss of NDUFAF5 and TIMMDC1 has been shown to stall CI assembly at the 86 kDa and 159 kDa intermediates, respectively[33,34]. Our results show that in wild-type cells, Q module subunits are able to fully assembly into the CI holoenzyme. In NDUFAF5 and TIMMDC1 KO cells, those subunits are instead stalled at their respective expected intermediates. Loss of AF6 leads to the same pattern of subunit migration as seen in the NDUFAF5 KO cells (i.e., stalling at the 86 kDa intermediate) (Fig. 3f). These data support the model that AF6 facilitates assembly of NDUFS8 into CI and establishes it as a Q module assembly factor.

AF6 mediates assembly at the IMM

AF6 is a peripheral membrane protein localized to the matrix face of the inner mitochondrial membrane[19]. This membrane association is predicted to be driven by its amphipathic C-terminal helix (CTH)[17,30]. Our DMS data show that no C-terminal truncations are tolerated, suggesting an essential role for the CTH in the function of AF6. Intriguingly, despite poor conservation and generally high mutational tolerance, the surface-exposed residues of the

CTH show a sensitivity to negatively charged substitutions (Fig. 4a). We reasoned that if the CTH is involved in the membrane association of AF6, mutations of the surface-exposed residues in the CTH to negatively charged amino acids could disrupt membrane association and, consequently, CI assembly. Additionally, if the CTH is interfacing with the negatively charged phospholipid head groups of the IMM, then positively charged mutations may strengthen or have no effect on membrane association.

To test this, we generated cell lines expressing wild-type AF6, a negative CTH mutant, or a positive CTH mutant under either the native promoter sequence or an overexpression promoter. Mutations for the negative CTH mutant (L319D, L322E, Y325D, I326D, W329D, R330D) were selected based on the least tolerated substitutions in the surface-exposed CTH residues. Mutations for the positive CTH mutant (L319Q, L322K, Y323E, Y325K, I326S, W329K) were modeled after the electrostatically positive surface-exposed CTH residues of a structurally similar bacterial HH synthase (PDBID: 5IYS). We then measured growth of these cell lines in galactose media. Our results show that, under native expression, neither the negative nor positive CTH mutants were able to grow in galactose, suggesting some degree of functional impairment for both constructs. However, overexpression results show that the positive CTH mutant can still complement the KO whereas the negative CTH mutant cannot (Fig. 4b). Together, these results demonstrate that negatively charged CTH mutations are more disruptive than their positive counterparts.

To determine whether these mutations affected the ability of AF6 to associate with the peripheral membrane, we performed a membrane fractionation experiment. We first isolated mitochondria from cell lines overexpressing wild-type AF6, the negative CTH mutant, or the positive CTH mutant. We then lysed the mitochondria and separated the membrane and soluble fractions by ultracentrifugation. Our results show that cells expressing the negative CTH mutant have markedly reduced levels of AF6 and its target NDUFS8, possibly due to increased turnover. Most importantly, the AF6 that is expressed only appears in the soluble fraction, suggesting loss of membrane association and, consequently, function. Cells expressing the positive CTH mutant show a similar pattern of membrane association to the wild-type control (Fig. 4c). These results support the hypothesis that the CTH drives membrane association of AF6 and that this membrane association is essential for its role in CI assembly.

The Q module forms part of the matrix arm of the CI holoenzyme and, like AF6, is peripherally associated with the IMM. The current model of Q module assembly predicts that membrane association occurs when the matrix Q module joins with the membrane Pp-a module, several steps downstream of the 125 kDa intermediate[32]. Given that the peripheral membrane localization of AF6 is essential for its function and that its target, NDUFS8, forms part of the interface anchoring the Q module to the IMM, we hypothesized that assembly of the 125 kDa intermediate occurs at the IMM rather than the mitochondrial matrix. To test this, we performed a membrane fractionation experiment on TIMMDC1 KO cells, which stall assembly at the 125 kDa and 159 kDa intermediates. The predominant membrane localization of Q module subunits in the TIMMDC1 KO supports the model that the majority of Q module assembly occurs at the IMM, at least as early as assembly of the 125 kDa intermediate (Fig. 4d).

Overexpression of NDUFS8 complements AF6 KO cells

Thus far, our DMS data helped reveal that AF6 is a *bona fide* pseudoenzyme that interacts specifically with NDUFS8 via a defined interaction interface. We further showed that loss of AF6 prevents assembly of NDUFS8 into the Q module and that this incorporation relies of the peripheral membrane association of AF6. This ultimately leads us to a model where AF6 mediates CI assembly by physically recruiting and positioning NDUFS8 for incorporation into the Q module at the IMM.

Though the activity of pseudoenzymes can be difficult to establish unambiguously, we reasoned that if our model is correct, we may be able to overcome the assembly defect in AF6 KO cells by overexpressing NDUFS8 and driving its assembly into CI by mass action. To test this, we generated a cell line overexpressing NDUFS8 in an AF6 KO background and measured its growth in galactose media. Indeed, we observed that overexpression of NDUFS8 was able to complement the KO of AF6 (Fig. 4e). Western blot analysis showed loss of NDUFS8 in the AF6 KO cells, likely due to increased turnover of unincorporated NDUFS8, and that overexpression of either AF6 or NDUFS8 is sufficient to rescue loss of NDUFS8 in the AF6 KO cells (Fig. 4f). These data offer further confirmation of the AF6-NDUFS8 interaction and support our model that AF6 acts by increasing the local concentration of NDUFS8 and/or the efficiency of its incorporation into the Q module (Fig. 4g). Furthermore, these data argue against a role for AF6 in the maturation of NDUFS8. If AF6 were involved in NDUFS8 maturation (e.g., insertion of Fe-S clusters, post-translation modification, etc.), increasing expression of NDUFS8 would likely be insufficient to overcome defects in its maturation. These data also suggest a potential therapeutic avenue to bypass AF6 dysfunction by directly modulating levels of the presumably functional NDUFS8 already in the cell.

DMS provides a diagnostic resource for AF6-related diseases

Pathogenic, bi-allelic variants in AF6 are associated with isolated mitochondrial CI deficiency and a spectrum of clinical manifestations and sequelae dominated by Leigh syndrome[18,35-37]. A grand challenge in the field of mitochondrial medicine is establishing whether variants in a given mitochondrial gene are pathogenic or simply benign polymorphisms (i.e., the causal variant is located in another gene). Current biochemical and genetic methods to confirm diagnoses are costly, limited in throughput, and often yield equivocal results[3]. DMS provides a systematic and high-throughput framework for evaluating the functional impact of variants to support clinical interpretations of variant pathogenicity.

To begin testing the clinical applicability of our DMS data, we first examined the DMS fitness scores of four variants annotated as "benign/likely benign" and five annotated as "pathogenic/likely pathogenic" in the ClinVar database[38]. For these nine variants, we generated stable cell lines expressing the variants in an AF6 KO background and measured NDUFS8 levels by immunoblot as an orthogonal readout for AF6 function (Fig. 5a). The DMS fitness and immunoblot data were consistent with seven of the nine ClinVar annotations. One of the two "inconsistent" variants, NM_152416.4:c.296A>G:p.Gln99Arg, causes a splice site defect that abolishes transcript levels of AF6[18]. Here, our data support

that the pathogenicity of this splice site variant, which showed no evidence of functional impairment at the protein level by DMS or immunoblot, is not a consequence of the amino acid substitution. The second variant, p.Pro281Ala, which is currently annotated as "likely pathogenic," had an intermediate DMS fitness score and only mildly reduced NDUFS8 levels relative to the wild-type control. This suggests that variants in this intermediate fitness range may be hypomorphic in function but still sufficient to cause disease.

We next expanded our analysis to all 14 non-VUS variants deposited in ClinVar[38] (Table S4) as well as the 286 nonsense variants in the DMS dataset. Examining the DMS fitness scores of these control variants in the context of our GMM components showed that 1) the benign/ likely benign variants all had fitness scores under the low functional impact component and 2) the pathogenic/likely pathogenic variants, as well as the nonsense variants, had fitness scores spanning both the strong and intermediate functional impact components (Fig. 5b, Extended Data Fig. 2f). This suggests that while pathogenic variants may have different degrees of functional impact, as evidenced by our DMS fitness scores, their clinical manifestations are of similar severity. For clinical purposes, we therefore decided to classify variants with either intermediate or strong functional impact under the single umbrella of "abnormal" function and to classify variants with low functional impact as "normal" function.

For this classification, we first calculated a probability of abnormal function ($P_{abnormal}$) for each variant in our DMS library. We used the low functional impact component of our GMM to model the scores of variants with normal function and the strong and intermediate functional impact components to model the scores of variants with abnormal function. The overall model was the sum of the normal and abnormal function models. The $P_{abnormal}$ for a variant was then computed as the likelihood of observing its fitness score given the abnormal function model divided by the likelihood of observing its fitness score given the overall model (Fig. 5c). Pabnormal ranges of [0, 0.25), [0.25, 0.75), [0.75, 1] were then used to classify variants as "normal," "uncertain," or "abnormal" function, respectively (Fig. 5d, Table S1). These cutoffs are concordant with the 14 control variants and classify all nonsense variants as abnormal function. The area of uncertainty is limited by the rather small number of control variants and can likely be refined as more variants are described.

According to the American College of Medical Genetics and Genomics (ACMG) guidelines for sequence variant interpretation, several criteria must be met in order to classify a variant as pathogenic or benign. In this framework, DMS data fall under the PS3 (pathogenic evidence) and BS3 (benign evidence) criteria corresponding to "well established *in vitro* or *in vivo* functional studies"[39]. For large scale functional studies, the strength of PS3 or BS3 evidence is determined by the odds of pathogenicity (OddsPath) value, a metric of how well the proposed classifier performs on established control variants[40]. Our classification scheme correctly assigned 9 out of 9 pathogenic/likely pathogenic variants to the functionally abnormal class and 5 out of 5 benign/likely benign variants to the functionally normal class (for the purposes of this calculation, the p.Gln99Arg splice site defect variant was treated as a "benign" variant, as it was demonstrated to be functionally normal at the protein level). This resulted in OddsPath values of 5.0 for pathogenic assessment and 0.11 for benign assessment (Table S5), corresponding to moderate evidence strength for both PS3 and BS3

criteria[40]. Both OddsPath scores were limited by the small number of benign control variants and will likely be improved as more benign variants become available. That said, application of even moderate strength PS3 and BS3 evidence was already sufficient to improve the ACMG classification of six out of the 14 control ClinVar variants (Table S4).

Previous and ongoing studies have nominated several variants in AF6 as possible causal variants for CI deficiency, but these variants have not been evaluated by experimental analyses. One immediate use of the DMS data is to provide functional evidence for the interpretation of these candidate variants. We examined a cohort of 18 patients with both primary mitochondrial disease and variants in AF6, including 13 patients with candidate pathogenic variants (Table S6). The incorporation of our DMS data in the ACMG sequence variant interpretation framework helped classify two existing variants with conflicting interpretations of pathogenicity and seven candidate variants as either pathogenic or likely pathogenic. As additional validation, we measured NDUFS8 levels by immunoblot from cell lines expressing several of these candidate variants, which all showed loss of NDUFS8 by immunoblot and DMS fitness scores below the cutoff for abnormal function (Fig. 5e). Beyond demonstrating that these variants have abnormal function in our cell culture system, we also performed respiratory chain enzyme activity assays on fibroblast or muscle biopsy samples from patients harboring these genetic variants, confirming their CI deficiency (Table S7). Together, these data support the robustness of our DMS approach and its potential for aiding clinical variant interpretation.

Discussion

AF6 is a pseudoenzyme whose molecular role in CI assembly was unclear. Guided by a DMS approach, we demonstrated that AF6 mediates CI assembly through a protein-protein interaction. This led us to identify the Q module subunit NDUFS8 as the target interactor of AF6 and to show that loss of AF6 prevents incorporation of NDUFS8 into CI. Our DMS results and subsequent analyses also showed that the C-terminal helix plays an important role in membrane association of AF6. These findings ultimately provide new insights into how the matrix modules of CI are assembled onto the transmembrane pump modules. Based on our new model of Q module assembly, we determined that the loss of AF6 can be bypassed by overexpression of NDUFS8. This both confirms our model for AF6 in CI assembly and provides a potential therapeutic strategy for bypassing AF6 dysfunction.

If AF6 defects can be overcome by simply overexpressing NDUFS8, why is AF6 essential? CI subunits have dual origins — seven of the core subunits are encoded in the mitochondrial genome while the remaining 38 in human, including NDUFS8, are encoded in the nuclear genome. Regulatory mechanisms coordinate the synthesis of CI subunits from both the cytoplasm and mitochondria[41] and ensure that subunits are co-expressed at stoichiometrically appropriate levels[42]. Thus, under normal conditions, it is unlikely that cells can rescue assembly defects through upregulation of any single subunit. Our results suggest that AF6 plays a key role in increasing the efficiency of NDUFS8 incorporation — likely an otherwise energetically unfavorable assembly step. It is important to note that even though overexpression of NDUFS8 remedies loss of AF6, it may also introduce other, albeit lesser, problems such as disruption of mitochondrial proteostasis and Fe-S metabolism.

Our understanding of the mechanisms mediated by CIAFs has not kept pace with the steady rate of their discovery. Characterization of CIAFs frequently stops after initial validation, often because the immediate next steps for further functional characterization are difficult or unclear. There is thus a need for systematic approaches to navigate this nebulous stage of CIAF characterization and link CIAFs with the appropriate downstream lines of experimentation. In such cases, DMS can be a valuable tool for adding more mechanistic depth to the study of CIAFs. The benefits of this systematic mutational analysis are manifold. First, this approach is agnostic to the molecular function of the protein, since it relies on a generalizable, downstream readout of OxPhos function (i.e. growth in galactose media). This is important because these assembly factors are a diverse group of proteins, multiple of which appear to be pseudoenzymes that have repurposed known protein folds for unpredictable new functions in CI assembly (e.g. AF6, NDUFAF5). A framework for assessing CIAF function without needing to know the exact function *a priori* can thus serve as a launching point for more targeted analyses. Our studies here show that DMS results allow one to generate testable, mechanistic hypotheses for further biochemical characterization.

Finally, despite the increasing number of known mitochondrial disease genes, the success rate of molecular diagnosis of mitochondrial diseases typically range between 30-60%[3,43-48]. This is in part due to the heterogeneous presentation of mitochondrial diseases as well as difficulty in interpreting genetic variants[13]. For AF6-related diseases, our DMS results help address the latter challenge by providing empirical evidence for the functional consequences of 5,714 AF6 single-missense variants. Incorporation of these data into the ACMG classification framework improved several existing variant interpretations and established seven new pathogenic variants. These results highlight the clinical value of DMS and demonstrate its potential for clarifying unresolved cases of mitochondrial diseases. Indeed, large-scale DMS efforts on CIAFs and other mitochondrial proteins may be an opportunity to advance both our molecular diagnostic capabilities for mitochondrial disorders and our understanding of their disease pathogenesis.

Methods:

Cell Culture:

Unless otherwise specified, HEK293T wild-type (WT) and knockout (KO) cells were cultured in DMEM (Thermo, 11965118) containing 10% heat-inactivated FBS (Biotechne, S11550) and 1x penicillin-streptomycin (Thermo, 15140122) at 37°C and 5% CO_2. HAP1 WT and KO cells (Horizon Discovery) were cultured in IMDM (Thermo, 12440053) with 10% heat-inactivated FBS (Biotechne, S11550) and 1x penicillin-streptomycin (Thermo, 15140122) at 37°C and 5% CO_2. HEK293T cells were negative for mycoplasma contamination as tested using a commercial test kit (LiLiF 25235). See Table S8 for the complete list of cell lines used in this study.

Generation of NDUFAF6 KO cell lines: CRISPR KOs of NDUFAF6 were generated using the IDT Alt-R CRISPR-Cas9 system. Two KO clones were generated independently using different target guide-RNAs.

- Guide 1 (Hs.Cas9.NDUFAF6.1.AA): GAGGGCCTTTAATGTGGAAC

- Guide 2 (Hs.Cas9.NDUFAF6.1.AB): TTCCATAGTTCAATGGCCAC

- Negative control: IDT negative control crRNA #1 (IDT 1072544)

One clone from each guide was expanded and validated through Sanger sequencing around the target cut site, western blot, and Seahorse analysis.

Generation of stable lines.: Constructs of interest were cloned into lentiviral vectors under either the EF1α promoter (pLV-OE) or the endogenous NDUFAF6 promoter (chr8: 95,024,000-95,024,989, pLV-NE) using standard restriction enzyme and ligation methods. Variant constructs were generated using the NEB Q5 Site-Directed Mutagenesis Kit Protocol. See Table S9, S10 for the constructs and oligonucleotides, respectively, used in this study. Plasmids were purified using Nucleospin Plasmid Transfection-grade kit (Takara, 740490). Lentiviral particles were produced using the TransIT Lenti system (Mirus, MIR 6650). Viral media was harvested from each well 48 hours post-transfection and either filtered through a 0.45 μm PES filter (Cytiva, 6896-2504) or centrifuged for 10 minutes at 600xg before aliquoting and flash freezing the filtrate or supernatant, respectively, and storing at −80°C.

To transduce cells, NDUFAF6 KO cells were seeded in 6-well plates, in the absence of penicillin-streptomycin, and incubated overnight or until ~40% confluent. Media was supplemented with 16 μg/ml of TransduceIT reagent (Mirus, MIR 6620) and 200-400 μl of lentivirus was added to each well. Cells were incubated at for 48 hours. Cells were then passaged in DMEM with 10% heat-inactivated FBS and 0.5 μg/ml puromycin (Thermo, A1113803) for 5-7 days. Cells were then passaged and expanded as normal and frozen down using freezing media (90% DMEM, 10% DMSO).

Galactose growth assay.: Cell lines were seeded with 6e4 cells/well in 12-well plates in triplicate (three separate wells per cell line) and incubated overnight to allow cells to adhere to the plate. Cells were then washed with dPBS and the medium was replaced with glucose-free DMEM supplemented with 10 mM galactose, 10% hiFBS, and 1x penicillin/ streptomycin. Cell counts were tracked over the next 48 hours with a Sartorius Incucyte S3, with images taken every 4 hours. To calculate the growth constant for the cells, total cell counts from each well were collected using the Sartorius IncuCyte S3 live-cell analysis software v2019A. Cell counts were normalized relative to time zero and log-transformed using Python v3.10 and Pandas v.1.5.3, and fit to a linear regression using Scipy v1.9.0[49]. Growth constants were visualized using Altair v4.2.2[50].

Seahorse CI activity assay.: Cellular OCR measurements were performed using an Agilent Seahorse XFe96 Analyzer. XFe96 microplates (Agilent, 102416-100) were first treated with poly-D-lysine (Thermo, A3890401) for 1 hour at 37°C and then seeded with HEK293T cells at a density of 1.5e4 cells per well in DMEM (Thermo, 11965118) containing 10% heat-inactivated FBS (Biotechne, S11550) and 1x penicillin-streptomycin (Thermo 15140122) and incubated at 37°C and 5% CO_2 overnight. The next day, the growth medium was exchanged for XF assay medium (Agilent, 102365-100) supplemented with 10 mM glucose,

1 mM pyruvate, and 2 mM glutamine. OCR measurements were done in 3 cycles of 3 minutes of mixing followed by 3 minutes of measuring. Cells were treated with 0.5 μM rotenone (Sigma, R8875). OCR measurements were normalized to relative cell number determined by crystal violet staining of cell nuclei[7]. Data analysis was performed using Agilent Seahorse Wave Desktop Software v2.6.1 and visualized using GraphPad Prism v9.5.1.

Mitochondrial isolation:

Mitochondria were isolated from cultured HEK293T and HAP1 cell lines by differential centrifugation[51]. In brief, cells were collected from 3-5 confluent 15 cm plates and washed with ice-cold PBS. Cell pellet was then resuspended in 10 ml of isolation buffer (220 mM D-mannitol, 70 mM D-sucrose, 5 mM HEPES-KOH, 1 mM EGTA, 0.5% fatty-acid free BSA, 1X cOmplete EDTA-free Protease Inhibitor Cocktail (Roche 11697498001)). HEK293Ts were lysed using 15 passes through a 20G needle. HAP1 cells were lysed via nitrogen cavitation with a cell disruption vessel (Parr Instruments) pressurized at ~800 psi for 10 minutes. Cellular debris in the lysate was pelleted by centrifugation at 600xg, 10 min, 4°C. The supernatant was transferred to fresh tubes and crude mitochondria were pelleted by centrifugation at 7,000xg, 10 min, 4°C. Mitochondrial pellets were finally washed with isolation buffer without BSA. Total mitochondrial protein content was measured by either Pierce BCA Protein Assay Kit (Thermo, 23225) or Qubit Protein Broad Range Assay Kit (Thermo Q33211). Crude mitochondria were aliquoted, flash frozen, and stored at −80°C for further use.

Blue-Native PAGE Western Blot:

In brief, 200 μg aliquots of crude mitochondria were resuspended in solubilization buffer (50 mM imidazole, 500 mM 6-hexaminocaproic acid, 1 mM EDTA) containing 6.0 g/g digitonin (detergent/protein ratio), incubated for 20 min on ice, then centrifuged at 15,000xg, 20 min, 4°C to remove insoluble material. 30 μg protein, as determined by BCA assay (Thermo, 23225) was combined with 5% glycerol and G-250 Coomassie Brilliant Blue to a final detergent/dye ratio of 8g/g and separated on a 4-16% Native PAGE[52] with a protein standard (Thermo, LC0725). The gel was then incubated with denaturing buffer (300 mM Tris, pH 8.6, 100 mM acetic acid, 1% SDS), for 20 minutes at RT prior to transferring to PVDF membrane and destaining with methanol. The membranes were probed with primary antibodies and analyzed using a Bio-Rad ChemiDoc MP and BioRad Image Lab Touch Software (version 3.0.1.14) with secondary antibodies (Cell Signaling, 7076S, 7074S) and ECL substrate (Thermo, 37071). See Table S11 for detailed antibody information.

In-gel CI Activity Assay:

Crude mitochondria were solubilized and separated on BN-PAGE as described above. Following electrophoresis, the gel was incubated in complex I substrate solution (0.1 M Tris-HCl pH 7.4, 0.14 mM NADH, 0.1% (w/v) nitroblue tetrazolium[53]) for 20 minutes, RT. The reaction was stopped with 10% acetic acid, then washed with water and scanned.

SDS-PAGE Immunoblot:

Whole cell lysates and mitochondrial lysates were prepared by resuspending samples in RIPA buffer (Invitrogen, 89900) and protease inhibitors (Sigma, 11836170001). The samples were solubilized on ice for 20 minutes and clarified by centrifugation (20,000xg, 20 min, 4°C). Supernatants were transferred to fresh tubes. Lysate protein content was quantified by Pierce BCA Protein Assay Kit (Thermo, 23225) or Qubit Protein Broad Range Assay Kit (Thermo Q33211). For each sample, 15-30 μg of protein was mixed with sample buffer (Invitrogen, NP0007) and separated on a NuPAGE 10% Bis-Tris gel (Thermo, NP0303BOX) with a protein standard (Licor, 928-60000). Resolved proteins were then transferred to PVDF membrane (Fisher, IPFL00010), probed with primary antibodies followed by HRP-conjugated secondary antibody (Cell Signaling Technologies, 7074S, 7076S) and ECL substrate (Thermo, 34577, 34579, 34094). See Table S11 for detailed antibody information. Blots were imaged and analyzed using either a LI-COR Odyssey CLx Imaging System and LI-COR Image Studio Software (version 5.2.5) or an Azure 600 (version 1.9.0.0406).

Cross-linking mass spectrometry (XL-MS)

Crude mitochondria were isolated from cells as described above and subjected to chemical crosslinking (0.5 mM disuccinimidyl sulfoxide (DSSO) (Thermo, A33433), 1 hour, RT). Crosslinking was quenched with 100 mM Tris pH 8.0 followed by centrifugation at 15,000xg, 5 min, 4°C. The mitochondria were then solubilized with 50 mM imidazole, 500 mM 6-hexaminocaproic acid, 1 mM EDTA, and 6 g/g digitonin. The bait protein and cross-linked interactors were then enriched by FLAG immunoprecipitation (IP) using magnetic anti-FLAG beads (Sigma, M8823-1ML), washed, and subjected to on-bead tryptic digest. The on-bead crosslinked proteins were denatured with 2M urea in 200 mM Tris pH 8.0, then reduced with 5 mM DTT for 30 min at 56°C and alkylated with 15 mM iodoacetamide for 30 min at RT in the dark. The proteins on-bead were digested overnight at RT with 1 μg trypsin (Promega, V5113). The digested supernatant was acidified with 10% TFA to a pH of 2 and desalted with 10 mg StrataX solid phase extraction columns (Phenomenex), then dried under vacuum using a SpeedVac (Thermo Scientific) and stored at −80°C until MS analysis.

Samples were resuspended in 0.2% formic acid and subjected LC-MS analysis. LC separation was performed using the Thermo Ultimate 3000 RSLCnano system. A 15 cm EASY-Spray™ PepMap™ RSLC C18 column (150 mm × 75 μm, 3 μm) was used at 300 nL/min flow rate with a 90 min gradient using mobile phase A consisting of 0.1% formic acid in H2O, and mobile phase B consisting of 0.1% formic acid in ACN/H2O (80/20, v/v). EASY-Spray source was used and temperature was at 35°C. Each sample run was held at 4.0% B for 5 min and increased to 50% B over 65 min, followed by 8 min at 95% B and back to 4% B for equilibration for 10 min. An Acclaim PepMap C18 HPLC trap column (20 mm × 75 μm, 3 μm) was used for sample loading. MS detection was performed with Thermo Exploris 240 Orbitrap mass spectrometer in positive mode. The source voltage was set to 1.8 kV, ion transfer tube temperature was set to 275°C, RF lens was at 70%. Full MS spectra were acquired from m/z 350 to 1,400 at the Orbitrap resolution of 60,000, with the normalized AGC target of 300% (3E6). Datadependent acquisition (DDA) was performed

for the top 20 precursor ions with the charge state of 2-6 and an isolated width of 2. Intensity threshold was 5E3. Dynamic exclusion was 30 s with the exclusion of isotopes. Other settings for DDA include Orbitrap resolution of 15,000 and HCD collision energy of 30%.

Raw files were analyzed by SequestHT Search Engine incorporated in Proteome Discoverer v2.5.0.400 software against human databases downloaded from Uniprot. Label-free quantification was enabled in the searches. The resulting data was analyzed by Perseus v1.6.15.0 software[54] and visualized using Altair v4.2.2[50].

Yeast two-hybrid (Y2H) assay

AF6, NDUFS8, and other proteins of interest were N-terminally tagged with either the DNA binding domain (BD) or activation domain (AD) of the Gal4 transcription factor. The resultant constructs were transformed[55] into *Saccharomyces cerevisiae* strain PJ69-4a[56] and plated on synthetic dextrose (SD) -Leu/-Trp agar plates. For the assay, overnight liquid cultures (SD -Leu/-Trp) with transformants were incubated overnight at 30°C, 225 rpm. Overnight cultures were diluted to an OD_{600nm} of 1.0 in a 2% glucose (w/v) solution and 10-fold serial dilutions in 2% glucose were spotted onto SD His- and Ade-deficient plates. Plates were then incubated at 30°C for 2-3 days before imaging. Variant constructs were generated using the NEB Q5 Site-Directed Mutagenesis Kit Protocol. See Table S9, S10 for the constructs and oligonucleotides, respectively, used in this study.

Membrane fractionation

In brief, 200 μg of crude mitochondria were resuspended in 400 μl of lysis buffer (10 mM Tris, pH 7.6, 100 mM NaCl) to a concentration of 0.5 μg/μl and incubated on ice for 30 minutes. Mitochondria were then lysed by sonication (4× 5s pulses, 20% amplitude, 30s between pulses, 1/16" microtip). Membrane fraction was pelleted by ultracentrifugation at 100,000xg, 1 hour, 4°C in a Beckman Optima Max XP Ultracentrifuge. The supernatant was collected and soluble protein was TCA precipitated and resuspended in 1x LDS (Invitrogen NP0007) + 50 mM DTT. The membrane pellet was washed with lysis buffer, then resuspended in the same volume of 1x LDS + 50 mM DTT as the soluble fraction. Samples were boiled for 10 minutes and western blot analysis performed as described above.

Deep Mutational Scanning

Library Preparation—The site saturation variant library (SSVL) was synthesized by Twist Biosciences as linear dsDNA fragments. Variants included all possible single amino acid substitutions, including termination codons, along residues 45-333 of NDUFAF6. The substituted amino acid sequences were codon optimized. The SSVL was first cloned into a pUC19 vector backbone for library storage using HindIII and EcoRI restriction sites. SSVL fragments were combined in pools of 8 positions (e.g. all variants for residues 45-52, 53-60, etc.) prior to restriction digest. Restriction reactions were heat-inactivated at 80°C for 10 minutes, according to NEB protocol, ligated with pUC19 backbone (HindIII/ EcoRI digested, CIP-treated), and transformed into high efficiency chemically competent cells (Lucigen E. cloni 10G). >20x transformants per variant was observed, estimated by plating 1/20 of transformation reaction on LB agar plates. The library was purified from liquid culture using a miniprep kit (Thermo GeneJET Miniprep Kit). The library was

then amplified out from the pUC19 backbone and cloned into lentiviral vectors under either the EF1a overexpression promoter or the endogenous NDUFAF6 promoter (chr8: 95,024,000-95,024,989). An indexed wild-type sequence was added to the library as a benchmark for variant performance. The indexed wild-type sequence contained a two-bp 'CT' insertion after the termination codon. Libraries were purified from liquid culture using Nucleospin Plasmid Transfection-grade kit (Takara, 740490). Our library covers 5714/5780 (98.9%) of possible NDUFAF6 variants. (Missing variants include residues 108, 116, and 328 as well as Y70P, S86F, D165P, L198P, I235P, and F306P).

Production of Lentiviral Particles for variant library—The lentiviral plasmid libraries were packaged into lentiviral particles using the TransIT Lenti System (Mirus, MIR 6650). Viral supernatant was passed through a 0.45 μm PES filter (Cytiva, 6896-2504). Virus was concentrated using the Lenti-X Concentrator (Takara, 631232). Lentivirus containing control GFP plasmid was used to determine the appropriate amount of virus for ~40% transduction efficiency.

Generation of variant cell populations and functional selection through growth in galactose media—For each replicate, 2.5e7 cells were transduced with the appropriate lentiviral library at an estimated efficiency of 40% for an estimate coverage of 1,500-2,000 cells per variant. Cell numbers for each replicate were maintained above 1.0e7 throughout the experiment. Puromycin selection was started 48 hours post-transduction and carried out for 5-7 days. Prior to starting the galactose growth selection, cell pellets were collected and frozen in aliquots of 2.0e7 cells. The remaining cells (~1.5e7) were then grown in galactose media for six passages (~12-14 days). Cell pellets were collected and frozen in aliquots of 2.0e7 cells after three and six passages in galactose.

Next generation sequencing—Genomic DNA was isolated from 2e7 cells with the QIAGEN QIAamp Blood Midi Kit (QIAGEN 13343), following the manufacturer's protocol. Isolated gDNA was dissolved and resuspended in 5 mM Tris/HCl pH 8.5. Yield was quantified by NanoDrop. Amplification of gDNA was performed using the NEBNext Ultra II Q5 polymerase system[57]. See Table S10 for oligonucleotide sequences used for generating sequencing samples. For the first PCR (PCR1), 12×100 μl PCRs were set up for each sample using 5 μg of gDNA as template per PCR (total of 60 μg gDNA per sample) and primers complimentary to the integrated lentiviral backbone for 18 cycles. The PCR1 products were purified using a Thermo GeneJET PCR Purification Kit. For the second PCR (PCR2), 1×100 μl PCRs were set up using 20 μl of PCR1 product as template and primers flanking the target DMS region (residues 45-333) for 15 cycles. PCR2 were purified using a Thermo GeneJET PCR Purification Kit. Purified PCR2 product were concatenated with T4 Polynucleotide Kinase (NEB) and T4 ligase (NEB) in 15% PEG-8000 (w/v). Concatemers were purified using a Thermo GeneJET PCR Purification Kit and sent to the Genome Technology Access Center at the McDonnell Genome Institute for further library prep and sequencing. In brief, concatenated amplicons were then fragmented using a Covaris E220 sonicator using peak incident power 175, duty factor 10%, cycles per burst 200 for 240 seconds. DNA was blunt ended, had an A base added to the 3' ends, and then had Illumina sequencing adapters ligated to the ends. Ligated fragments were then amplified for 8 cycles

using primers incorporating unique index tags. Fragments were sequenced on an Illumina NovaSeq-6000 using paired end reads extending 150 bases and demultiplexed into pairs of fastq files for each sample with Illumina's bcl2fastq2 software v2.2.

DMS Data Analysis—Raw sequencing reads were first aligned to a wild-type NDUFAF6 cDNA reference using Bowtie2 v.2.4.4[58]. Aligned reads were then merged using FLASH v1.2.11[59]. A custom pipeline using Python v3.10, Pandas v1.5.3, Biopython v1.81[60], and SnakeMake v7.22.0[61] was then used to filter the merged reads for reads containing only a single expected codon substitution and determine the number of reads for each variant. Full-length reads for each variant were then reconstructed and fitness estimates were obtained for each variant using the DiMSum pipeline v1.3[23]. Pearson correlation coefficients (r) were calculated using SciPy v1.9.0[49].

The distribution of DMS fitness scores was fit to a Gaussian mixture model (GMM) using the Scikit-learn v1.1.2[62]. A three-component GMM was chosen by minimizing the Bayesian information criterion (BIC) score of GMMs with the number of component models ranging between 1 and 6. Three- and four-component GMMs produced similarly low BIC scores and a three-component model was chosen to reduce the risk of overfitting the data.

The "mutational sensitivity" of each position in NDUFAF6 was computed as the number of missense substitutions (excluding premature termination substitutions) with a fitness below −0.842 (three standard deviations from the mean of the gaussian distribution covering variants with a strong functional impact), divided by 2, and rounded down to the nearest integer. This produced an aggregate metric with values between 0 and 9, analogous to the ConSurf scale.

OddsPath calculations were performed as outlined by the Clinical Genome Resource (ClinGen) Sequence Variant Interpretation (SVI) Working Group[40]. See Table S5 for detailed calculations.

DMS data were visualized using Altair v4.2.2[50].

Clinical Data

All patients gave their informed consent for the use of their diagnostic data in this study. These data as part of the patients' diagnostic workups and were not part of a clinical intervention or trial. As such, there are no study design considerations, participant compensation, or sex and gender analyses to report.

Patient Genotyping—Candidate pathogenic AF6 variants were identified through diagnostic work-up of probands with suspected mitochondrial disease using various next-generation sequencing strategies, including panel-based target capture[63,64] and whole exome sequencing[35,65]. The exact procedure varied based on the center at which the sequencing studies were performed. All listed variants were confirmed by Sanger sequencing.

Respiratory Complex Activity Assays—Respiratory complex and citrate synthase (CS) activity assays were performed on patient-derived skin fibroblasts or patient muscle

biopsy samples[66,67]. The exact procedure for measuring respiratory chain enzyme activity varied based on the center at which the assays were performed. Enzyme activity is expressed as a percentage of the mean of the control CS activity.

Clinical Databases—Variant data from ClinVar were accessed on Oct. 24, 2023 and filtered for only missense variants. Allele frequency data downloaded from the gnomAD browser v4.0.0 on Nov. 19, 2023. Precomputed REVEL scores were downloaded from the REVEL v1.3[68] web server on Jan. 6, 2024.

Computational Tools:

Structural Prediction—Predicted structure for AF6 isoform 1 (UniProt ID: Q330K2) was downloaded from the AlphaFold Protein Structure Database[25,26] on Aug. 2, 2022.

Conservation Analysis—Conservation analysis was performed using the ConSurf server[69]. ConSurf scores for AF6 were calculated for AF6 isoform 1 using default parameters.

Solvent Accessible Surface Area—Relative and absolute solvent accessible surface area (SASA) was calculated from the AF6 AlphaFold model using the built-in *get_sasa_relative* and *get_area* commands, respectively, in PyMOL v2.5.2.

Electrostatic Potential—Electrostatic potential of structural models were calculated using the APBS electrostatics plugin[70] in PyMOL v2.5.2.

Sequence Logos—Sequence logos for the HH synthase DxxxD motifs and the analogous residues in AF6 were constructed using the Skylign tool[71]. Multiple sequence alignment for HH synthases was obtained using 88 seed sequences from Pfam SQS_PSY domain (PF00494). Multiple sequence alignment for AF6 homologs was obtained using the top 100 BLASTp hits with AF6 isoform 1 as the query sequence.

Statistics and Reproducibility:

All quantitative experiments were performed in at least biological triplicate, as indicated in the figure legends. No statistical method was used to predetermine sample size. No data were excluded from the analyses. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Extended Data
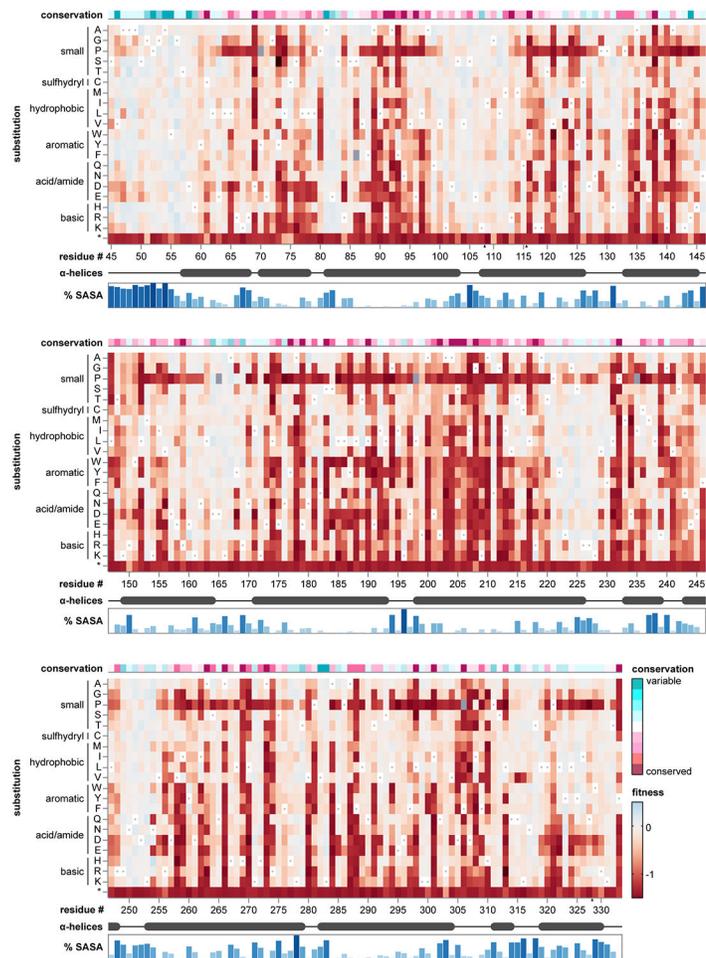
**Extended Data Fig. 1 |. Full DMS dataset.**
Heatmap representation of the DMS data. The residue position is labeled along the horizontal axis (missing residues in the dataset denoted by a black triangle), the amino acid substitution is labeled on the vertical axis. The color of the rectangles represents the fitness score of the variant. Gray dots mark the wild-type amino acid at each position. Additional annotations to aid interpretation are shown above and below the fitness score heatmaps. Conservation at each position (calculated using ConSurf) is shown using the ConSurf color scale. Predicted alpha-helical secondary structure based on the AlphaFold AF6 model is shown as a thick, grey line. Percent solvent-accessible surface area (% SASA) at each residue position, calculated from the AlphaFold AF6 model is shown as a bar graph.
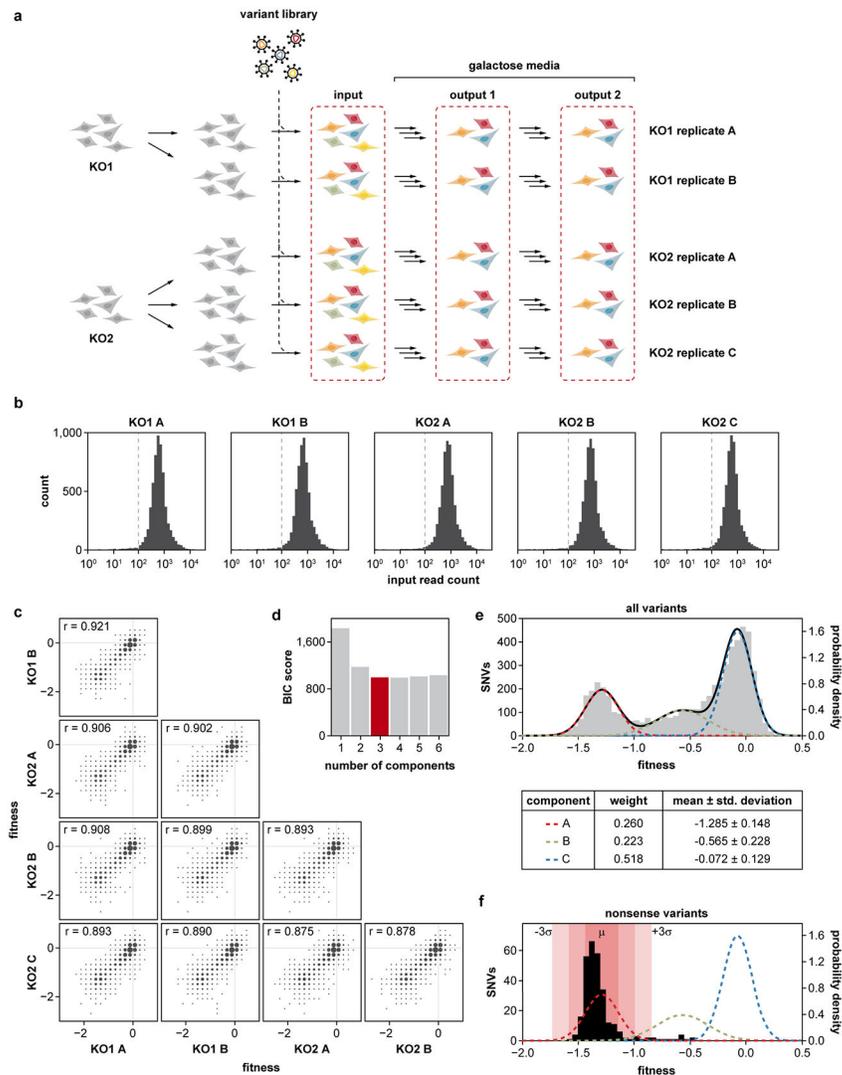
**Extended Data Fig. 2 |. Experimental design and quality control.**

a. Experimental design of DMS experiment. AF6 KO1 and AF6 KO2 cells are split into five replicates and separately transduced with the AF6 variant library. Input samples are collected 6–8 days after transduction (with 4-5 days of selection with puromycin). Cells are then cultured in galactose media to select for cells expressing functional variants of AF6. Output samples are collected afterthree and six passages in galactose media. **b**. Histogram of input read count distribution of the variants by replicate. The dashed line represents a read depth of 100x. **c**. Binned scatterplot of fitness scores from individual replicates. The dot size represents the number of variants in each fitness score bin. Pearson correlation coefficients (r) for each pairwise comparison is provided in the upper left corner of each plot. **d**. Bayesian information criterion (BIC) scores for Gaussian mixture models (GMM) with 1–6 components. A three-component model (shown in red) was chosen for this analysis. **e**. Histogram of DMS fitness scores overlaid with a three-component GMM. The three components are represented as colored dashed lines while the overall model is represented by a solid black line. The component weight, mean, and standard deviation are shown in the

table below. **f**. Histogram of DMS fitness scores of nonsense variants overlaid with the three components from the GMM. The red component likely represents variants with a strong functional impact. The red shaded region represents three standard deviations from the mean of the red component, [−1.729, −0.842], and encompasses the fitness scores for 274 out of 286 nonsense variants.
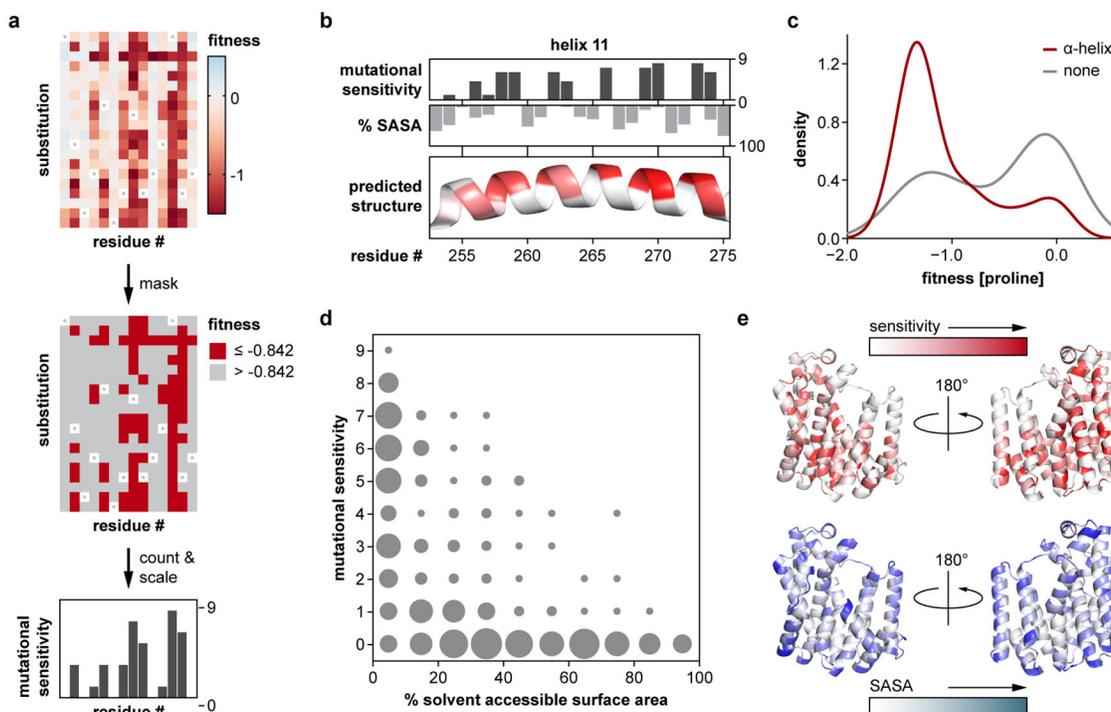
Extended Data Fig. 3 |. Comparisons of DMS data to predicted NDUFAF6 structure.

a. Schematic for the calculation of mutational sensitivity. At each residue position, the number of missense substitutions with a fitness ≤ −0.842 (threshold for variants with strong functional impact, see Extended Data Fig. 2f) are counted and scaled linearly to a value between 0 and 9. **b**. Mutational sensitivity and percent solvent-accessible surface area (% SASA) of helix 11 (residues 253–275) are shown as bars. A cartoon representation of this alpha helix from the AF6 AlphaFold model is shown below and colored by mutational sensitivity. **c**. Density plot of DMS fitness scores for proline substitutions grouped by predicted secondary structure. **d**. Binned scatterplot showing the distribution of percent solvent accessible surface area and mutational sensitivity of residues in AF6. The dot size represents the number of residues in each bin. **e**. Cartoon representation of the AF6 AlphaFold model colored by mutational sensitivity (top) and % SASA (bottom). A front view and a back view are shown.
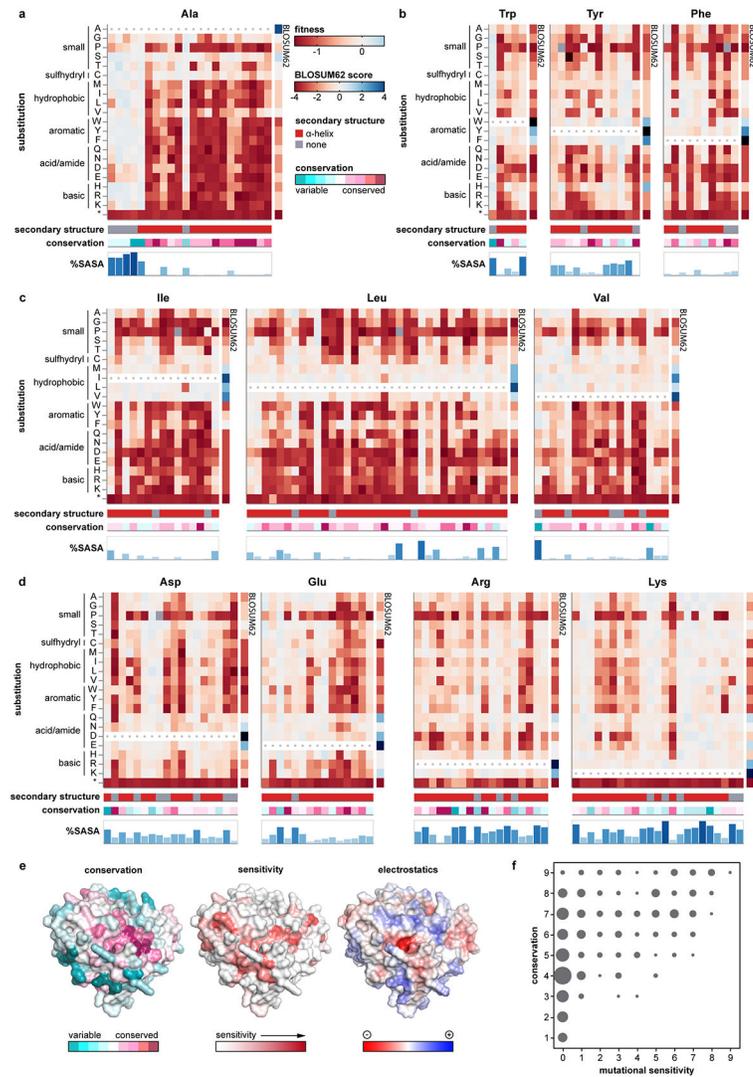
Extended Data Fig. 4 |. Biochemical principles recapitulated by DMS.
a-d. DMS fitness scores of substitutions in alanine residues (a), aromatic residues (b),
hydrophobic residues (c), and charged residues (d). BLOSUM62 similarity scores for
each substitution is shown along the right. Predicted secondary structure is shown below,
conservation (calculated using ConSurf), and percent solvent accessible surface area (%
SASA) are shown below. **e**. Surface representation of AF6 AlphaFold model colored by
conservation, mutational sensitivity, and electrostatics. **f**. Binned scatterplot showing the
distribution of conservation (calculated by ConSurf) and mutational sensitivity of residues in
AF6. The dot size represents the number of residues in each bin.

Extended Data Fig. 5 |. Additional XL-MS hits.

Volcano plot showing the mean log2 fold-change and the −log10 of the two-tailed Student's t-test p-value (no adjustment for multiple comparisons) of proteins in mitochondria overexpressing FLAG-tagged AF6 compared to wild-type mitochondria, n = 4 biologically independent samples per condition. CI subunits and assembly factors are marked as red dots and labeled. The most significant hits are labeled in the magnified portion of the volcano plot.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments:

Data Availability:

All unique/stable reagents generated in this study are available upon request with a materials transfer agreement. The raw next-generation sequencing data from the DMS experiment have been deposited in the Sequence Read Archive (BioProject accession: PRJNA1007392). DMS variant and count data have been deposited on MaveDB (MaveDB accession: urn:mavedb:00000663-a). The raw proteomics data from the XL-MS experiment have been deposited in the MassIVE repository (MassIVE ID: MSV000094276). All deposited data are publicly available. Unprocessed image files and statistical data underlying the figures can be found in the source data files published alongside this manuscript.

Code Availability:

All original code has been deposited on GitHub (https://github.com/AYSung/af6-dms) and is publicly available.

References:

1. Sazanov LA A giant molecular proton pump: structure and mechanism of respiratory complex I. Nat Rev Mol Cell Bio 16, 375–388 (2015). [PubMed: 25991374]

2. Kirby DM et al. Respiratory chain complex I deficiency: an underdiagnosed energy generation disorder. Neurology 52, 1255–64 (1999). [PubMed: 10214753]

3. Schon KR et al. Use of whole genome sequencing to determine genetic basis of suspected mitochondrial disorders: cohort study. BMJ 375, e066288 (2021). [PubMed: 34732400]

4. Nouws J, Nijtmans LGJ, Smeitink JA & Vogel RO Assembly factors as a new class of disease genes for mitochondrial complex I deficiency: cause, pathology and treatment options. Brain 135, 12–22 (2012). [PubMed: 22036961]

5. Swalwell H. et al. Respiratory chain complex I deficiency caused by mitochondrial DNA mutations. Eur J Hum Genet 19, 769–775 (2011). [PubMed: 21364701]

6. Formosa LE, Dibley MG, Stroud DA & Ryan MT Building a complex complex: Assembly of mitochondrial respiratory chain complex I. Semin Cell Dev Biol 76, 154–162 (2018). [PubMed: 28797839]

7. Rensvold JW et al. Defining mitochondrial protein functions through deep multiomic profiling. Nature 382–388 (2022) doi:10.1038/s41586-022-04765-3. [PubMed: 35614220]

8. Formosa LE et al. Optic Atrophy-associated TMEM126A is an assembly factor for the ND4-module of Mitochondrial Complex I. *Biorxiv* 2020.09.18.303255 (2020) doi:10.1101/2020.09.18.303255.

9. Jackson TD et al. Sideroflexin 4 is a complex I assembly factor that interacts with the MCIA complex and is required for the assembly of the ND2 module. Proc National Acad Sci 119, e2115566119 (2022).

10. Dibley MG et al. The mitochondrial acyl-carrier protein interaction network highlights important roles for LYRM family members in complex I and mitoribosome assembly. Mol Cell Proteomics 19, mcp.RA119.001784 (2019).

11. Ribeiro AJM et al. Emerging concepts in pseudoenzyme classification, evolution, and signaling. Sci Signal 12, (2019).

12. Jeffery CJ The demise of catalysis, but new functions arise: pseudoenzymes as the phoenixes of the protein world. Biochem Soc T 47, 371–379 (2019).

13. Frazier AE, Thorburn DR & Compton AG Mitochondrial energy generation disorders: genes, mechanisms, and clues to pathology. J Biol Chem 294, 5386–5395 (2019). [PubMed: 29233888]

14. Schon KR, Ratnaike T, van den Ameele J, Horvath R & Chinnery PF Mitochondrial Diseases: A Diagnostic Revolution. Trends Genet 36, 702–717 (2020). [PubMed: 32674947]

15. Fowler DM & Fields S Deep mutational scanning: a new style of protein science. Nat Methods 11, 801–807 (2014). [PubMed: 25075907]

16. Marchler-Bauer A. et al. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Res 45, D200–D203 (2017). [PubMed: 27899674]

17. Lemire BD Evolution, structure and membrane association of NDUFAF6, an assembly factor for NADH:ubiquinone oxidoreductase (Complex I). Mitochondrion 35, 13–22 (2017). [PubMed: 28476317]

18. Pagliarini DJ et al. A mitochondrial protein compendium elucidates complex I disease biology. Cell 134, 112–23 (2008). [PubMed: 18614015]

19. McKenzie M. et al. Mutations in the gene encoding C8orf38 block complex I assembly by inhibiting production of the mitochondria-encoded subunit ND1. J Mol Biol 414, 413–26 (2011). [PubMed: 22019594]

20. Stroud DA et al. Accessory subunits are integral for assembly and function of human mitochondrial complex I. Nature 538, 123–126 (2016). [PubMed: 27626371]

21. Robinson BH, Petrova-Benedict R, Buncic JR & Wallace DC Nonviability of cells with oxidative defects in galactose medium: a screening test for affected patient fibroblasts. Biochem Med Metab B 48, 122–6 (1992).

22. Gohil VM et al. Nutrient-sensitized screening for drugs that shift energy metabolism from mitochondrial respiration to glycolysis. Nat Biotechnol 28, 249–255 (2010). [PubMed: 20160716]

23. Faure AJ, Schmiedel JM, Baeza-Centurion P & Lehner B DiMSum: an error model and pipeline for analyzing deep mutational scanning data and diagnosing common experimental pathologies. Genome Biol 21, 207 (2020). [PubMed: 32799905]

24. Fukasawa Y. et al. MitoFates: improved prediction of mitochondrial targeting sequences and their cleavage sites. Mol Cell Proteomics 14, 1113–26 (2015). [PubMed: 25670805]

25. Jumper J. et al. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589 (2021). [PubMed: 34265844]

26. Varadi M. et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res 50, D439–D444 (2021).

27. Imai K & Mitaku S Mechanisms of secondary structure breakers in soluble proteins. Biophysics 1, 55–65 (2005). [PubMed: 27857553]

28. Henikoff S & Henikoff JG Amino acid substitution matrices from protein blocks. Proc National Acad Sci 89, 10915–10919 (1992).

29. Lin FY et al. Mechanism of action and inhibition of dehydrosqualene synthase. Proc Natl Acad Sci USA 107, 21337–42 (2010). [PubMed: 21098670]

30. Gautier R, Douguet D, Antonny B & Drin G HELIQUEST: a web server to screen sequences with specific α-helical properties. Bioinformatics 24, 2101–2102 (2008). [PubMed: 18662927]

31. Oughtred R. et al. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci 30, 187–200 (2021). [PubMed: 33070389]

32. Guerrero-Castillo S. et al. The Assembly Pathway of Mitochondrial Respiratory Chain Complex I. Cell Metab 25, 128–139 (2017). [PubMed: 27720676]

33. Alston CL et al. Pathogenic Bi-allelic Mutations in NDUFAF8 Cause Leigh Syndrome with an Isolated Complex I Deficiency. Am J Hum Genet 106, 92–101 (2020). [PubMed: 31866046]

34. Guarani V. et al. TIMMDC1/C3orf1 Functions as a Membrane-Embedded Mitochondrial Complex I Assembly Factor through Association with the MCIA Complex. Mol Cell Biol 34, 847–861 (2014). [PubMed: 24344204]

35. Kohda M. et al. A Comprehensive Genomic Analysis Reveals the Genetic Landscape of Mitochondrial Respiratory Chain Complex Deficiencies. PLoS Genet 12, e1005679 (2016). [PubMed: 26741492]

36. Lake NJ, Compton AG, Rahman S & Thorburn DR Leigh syndrome: One disorder, more than 75 monogenic causes. Ann Neurol 79, 190–203 (2016). [PubMed: 26506407]

37. McCormick EM et al. Expert panel curation of 113 primary mitochondrial disease genes for the Leigh syndrome spectrum. Ann Neurol (2023) doi:10.1002/ana.26716.

38. Landrum MJ et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res 46, gkx1153- (2017).

39. Richards S. et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 17, 405–423 (2015). [PubMed: 25741868]

40. Brnich SE et al. Recommendations for application of the functional evidence PS3/BS3 criterion using the ACMG/AMP sequence variant interpretation framework. Genome Med 12, 3 (2019). [PubMed: 31892348]

41. Couvillion MT, Soto IC, Shipkovenska G & Churchman LS Synchronized mitochondrial and cytosolic translation programs. Nature 533, 499–503 (2016). [PubMed: 27225121]

42. van der Lee R. et al. Transcriptome analysis of complex I-deficient patients reveals distinct expression programs for subunits and assembly factors of the oxidative phosphorylation system. BMC Genomics 16, 691 (2015). [PubMed: 26369791]

43. Sung AY, Floyd BJ & Pagliarini DJ Systems Biochemistry Approaches to Defining Mitochondrial Protein Function. Cell Metab 31, 669–678 (2020). [PubMed: 32268114]

44. Forny P. et al. Diagnosing Mitochondrial Disorders Remains Challenging in the Omics Era. Neurol Genet 7, e597 (2021). [PubMed: 34056100]

45. Pronicka E et al. New perspective in diagnostics of mitochondrial disorders: two years' experience with whole-exome sequencing at a national paediatric centre. J Transl Med 14, 174 (2016). [PubMed: 27290639]

46. Taylor RW et al. Use of Whole-Exome Sequencing to Determine the Genetic Basis of Multiple Mitochondrial Respiratory Chain Complex Deficiencies. JAMA 312, 68–77 (2014). [PubMed: 25058219]

47. Theunissen TEJ et al. Whole Exome Sequencing Is the Preferred Strategy to Identify the Genetic Defect in Patients With a Probable or Possible Mitochondrial Cause. Front Genet 9, 400 (2018). [PubMed: 30369941]

48. Wortmann SB, Koolen DA, Smeitink JA, Heuvel L & Rodenburg RJ Whole exome sequencing of suspected mitochondrial patients in clinical practice. J Inherit Metab Dis 38, 437–443 (2015). [PubMed: 25735936]

49. Virtanen P. et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat Methods 17, 261–272 (2020). [PubMed: 32015543]

50. VanderPlas J. et al. Altair: Interactive Statistical Visualizations for Python. J Open Source Softw 3, 1057 (2018).

51. Frezza C, Cipolat S & Scorrano L Organelle isolation: functional mitochondria from mouse liver, muscle and cultured fibroblasts. Nat Protoc 2, 287–95 (2007). [PubMed: 17406588]

52. Wittig I, Braun HP & Schagger H Blue native PAGE. Nat Protoc 1, 418–28 (2006). [PubMed: 17406264]

53. Schertl P & Braun H-P Plant Mitochondria, Methods and Protocols. Methods Mol Biol 1305, 131–138 (2015). [PubMed: 25910731]

54. Tyanova S. et al. The Perseus computational platform for comprehensive analysis of (prote)omics data. Nat Methods 13, 731–740 (2016). [PubMed: 27348712]

55. Gietz RD & Schiestl RH High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. Nat Protoc 2, 31–34 (2007). [PubMed: 17401334]

56. James P, Halladay J & Craig EA Genomic Libraries and a Host Strain Designed for Highly Efficient Two-Hybrid Selection in Yeast. Genetics 144, 1425–1436 (1996). [PubMed: 8978031]

57. Yau EH & Rana TM Next Generation Sequencing, Methods and Protocols. Methods Mol Biol 1712, 203–216 (2017).

58. Langmead B & Salzberg SL Fast gapped-read alignment with Bowtie 2. Nat Methods 9, 357–359 (2012). [PubMed: 22388286]

59. Mago T & Salzberg SL FLASH: fast length adjustment of short reads to improve genome assemblies. Bioinformatics 27, 2957–2963 (2011). [PubMed: 21903629]

60. Cock PJA et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 25, 1422–1423 (2009). [PubMed: 19304878]

61. Mölder F. et al. Sustainable data analysis with Snakemake. F1000Research 10, 33 (2021). [PubMed: 34035898]

62. Pedregosa F. et al. Scikit-learn: Machine Learning in Python. (2012).

63. Alston CL et al. A recurrent mitochondrial p.Trp22Arg NDUFB3 variant causes a distinctive facial appearance, short stature and a mild biochemical and clinical phenotype. J Med Genet 53, 634 (2016). [PubMed: 27091925]

64. Legati A. et al. New genes and pathomechanisms in mitochondrial disorders unraveled by NGS technologies. Biochim Biophys Acta Bioenerg 1857, 1326–1335 (2016).

65. Alston CL et al. Bi-allelic Mutations in NDUFA6 Establish Its Role in Early-Onset Isolated Mitochondrial Complex I Deficiency. Am J Hum Genet 103, 592–601 (2018). [PubMed: 30245030]

66. Ogawa E. et al. Clinical validity of biochemical and molecular analysis in diagnosing Leigh syndrome: a study of 106 Japanese patients. J Inherit Metab Dis 40, 685–693 (2017). [PubMed: 28429146]

67. Frazier AE, Vincent AE, Turnbull DM, Thorbum DR & Taylor RW Chapter 5 Assessment of mitochondrial respiratory chain enzymes in cells and tissues. Methods Cell Biol 155, 121–156 (2020). [PubMed: 32183956]

68. Ioannidis NM et al. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. Am JHum Genet 99, 877–885 (2016). [PubMed: 27666373]

69. Yariv B. et al. Using evolutionary data to make sense of macromolecules with a "face-lifted" ConSurf. Protein Sci 32, e4582 (2023). [PubMed: 36718848]

70. Jurrus E. et al. Improvements to the APBS biomolecular solvation software suite. Protein Sci 27, 112–128 (2018). [PubMed: 28836357]

71. Wheeler TJ, Clements J & Finn RD Skylign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models. BMC Bioinformatics 15, 7 (2014). [PubMed: 24410852]
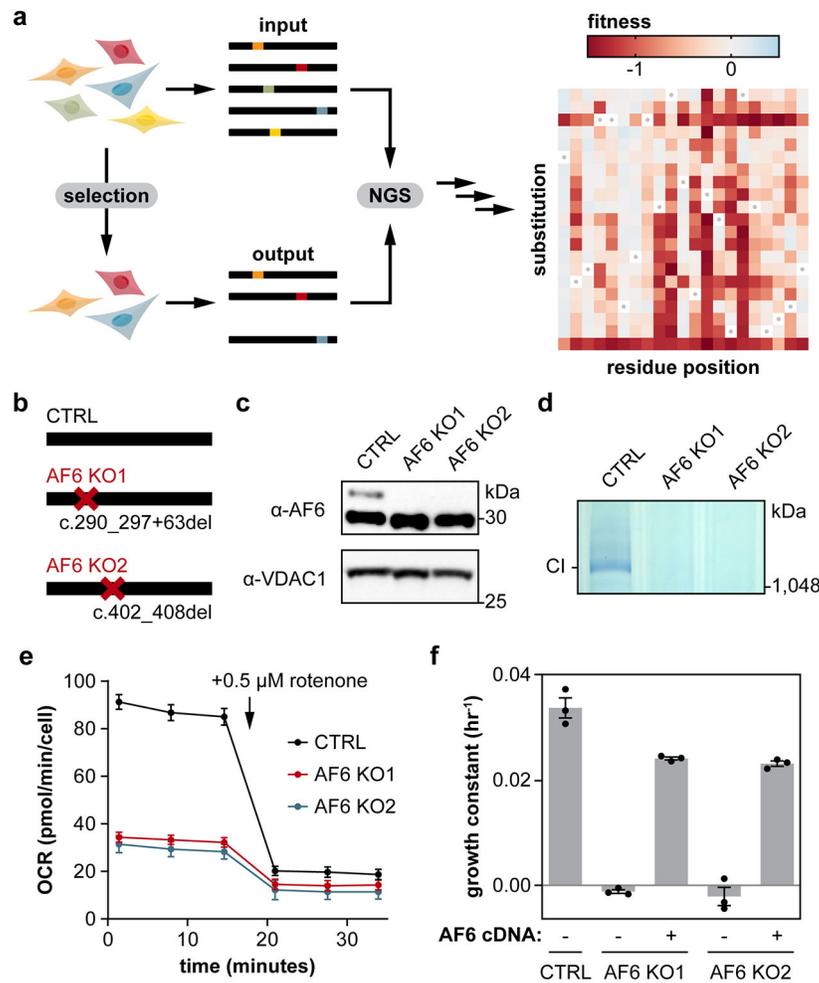
Figure 1: Deep mutational scanning of AF6

a. Schematic of DMS experiment. Cells expressing functional variants of AF6 were selected for by culturing in galactose media. Genomic DNA was extracted from samples collected before and after the selection and sequenced. Read counts from before and after selection were used to calculate fitness scores for each variant.

b. AF6 KO cell genotypes as determined by Sanger sequencing around the target Cas9 cut site. CTRL cells were treated with a non-targeting guide RNA.

c. Western blot of crude mitochondria isolated from CTRL, AF6 KO1 and AF6 KO2 cells, probing for AF6. VDAC1 is used as the loading control. The band at 30 kDa represents non-specific signal from the anti-AF6 antibody. This experiment was repeated three times with similar results.

d. In-gel activity assay measuring CI activity in crude mitochondria lysate isolated from CTRL, AF6 KO1, and AF6 KO2 cells. Band corresponding to CI is marked on the left. This experiment was performed once.

e. Seahorse oxygen consumption rate (OCR) measured before and after addition of 0.5 µM rotenone in CTRL, AF6 KO1, and AF6 KO2 cells. Results represent the mean of four replicates and error bars represent the standard error of the mean.

f. Growth constants in galactose media of CTRL cells and AF6 KO cells with and without expression of wild-type AF6. Bars represent the mean ± SEM, n=3 biologically independent experiments (replicate values shown as dots).
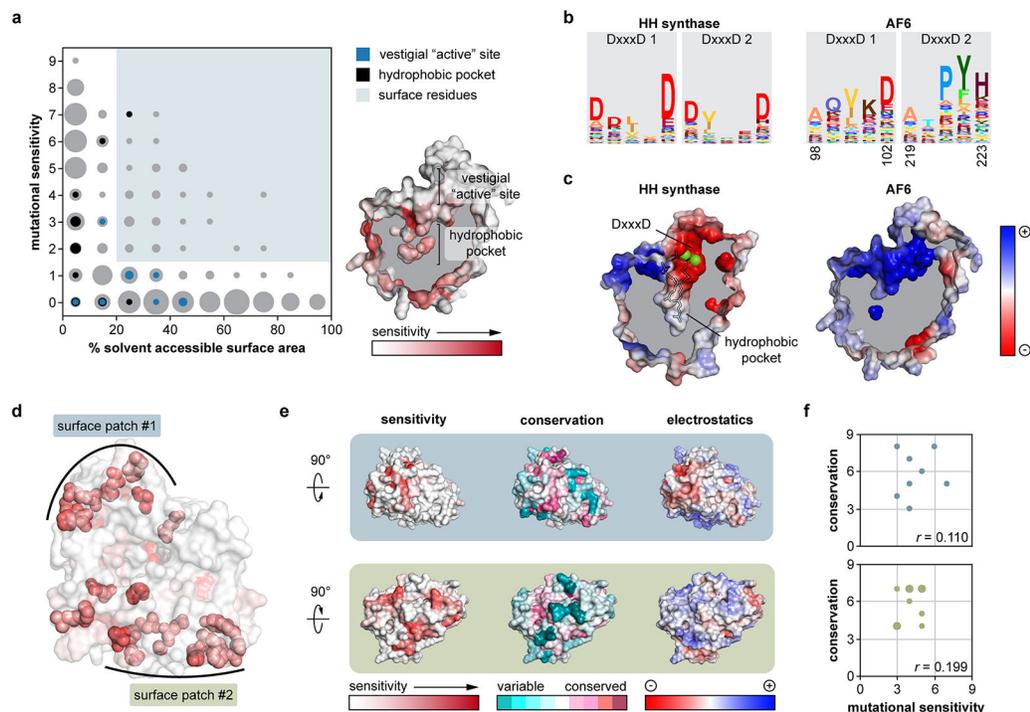
Figure 2: DMS highlights functionally relevant protein regions

a. Binned scatterplot showing the distribution of percent solvent accessible surface area and mutational sensitivity of residues in AF6. The dot size represents the number of residues in each bin. Residues comprising the vestigial "active" site motifs are highlighted in blue. Residues lining the putative hydrophobic binding pocket are highlighted in orange. Residues with a higher mutational sensitivity than expected for the given percent solvent-accessible surface area (mutational sensitivity 2 and SASA > 20%) are highlighted in light blue. A cut-away view of the AF6 AlphaFold model, colored by mutational sensitivity, is shown for reference. Regions comprising the vestigial "active" site and putative hydrophobic pocket are marked.

b. Sequence logos representing multiple sequence alignments for HH synthase active site compared to AF6 "active site". Residue numbers for AF6 residues corresponding to the catalytic aspartates in HH synthases are indicated below the sequence logos.

c. Cut-away view of the hydrophobic pocket in the *E. hirae* dehydrosqualene synthase structure (PDBID: 5IYS) compared to the AF6 AlphaFold model. Mg^{2+} ions depicted as green spheres, two farnesyl pyrophosphate substrates shown as black outlines. Surface is colored by electrostatics.

d, e, f. The surface residues in (a) are highlighted as spheres on the structural model of AF6 and colored by mutational sensitivity. These residues map to two surface patches that may be important for protein-protein interaction (d). Head-on views showing the mutational sensitivity, conservation, and electrostatics of the two surface patches, respectively (e). Scatter plots showing correlation between mutational sensitivity and conservation for the two surface patches, with overlapping points marked by increased dot size. Pearson's *r* is shown for each plot (f).
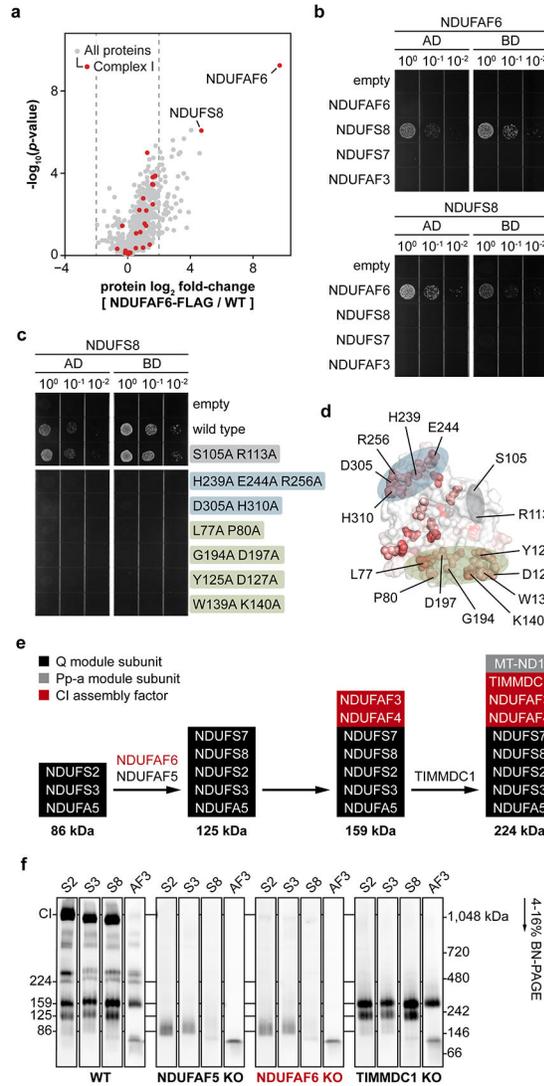
Figure 3: AF6 enables NDUFS8 incorporation into CI

a. Volcano plot showing the mean log2 fold-change and the −log10 of the two-tailed Student's t-test p-value (no adjustment for multiple comparisons) of proteins in mitochondria overexpressing FLAG-tagged AF6 compared to wild-type mitochondria, n=4 biologically independent samples per condition. Dashed lines mark log2 foldchanges of −2 and 2. CI subunits and assembly factors are marked as red dots. All other proteins marked as grey dots.

b. Yeast two-hybrid assay assessing the interaction between AF6, NDUFS8, and Q module subunits / assembly factors. Serial dilutions are marked along the top of the spot plate. The bait proteins are labeled above and the prey proteins are labeled on the side. Image is representative of three biological replicates.

c. Yeast two-hybrid assay assessing the interaction of various double and triple alanine variants in the three patches highlighted in (d). Serial dilutions are marked along the top of the spot plate. The bait proteins are labeled above and the prey proteins are labeled on the side. Image is representative of three biological replicates.

d. Residues in the surface patches identified by DMS targeted for mutagenesis are labeled and highlighted in blue and green for surface patches 1 and 2, respectively. Residues highlighted in grey represent a surface region that is not predicted to be important for interaction with NDUFS8 based on the DMS data.

e. Q/Pp-a module assembly pathway with known assembly intermediates and assembly factors. The predicted molecular weight of each assembly intermediate is marked. The assembly steps mediated by NDUFAF5 and TIMMDC1 are labeled above the arrows.

f. Blue native western blots of crude mitochondrial lysate from wild-type, NDUFAF5 KO, AF6 KO, and TIMMDC1 KO HAP1 cells, probing for NDUFS2, NDUFS3, NDUFS8, and NDUFAF3. Bands corresponding to Q/Pp-a module assembly intermediates and the CI holoenzyme are marked along the left of the blot. Molecular weights on the right indicate migration of the soluble protein ladder. Lanes were rearranged and grouped by cell line to facilitate interpretation. This experiment was repeated independently two times with similar results.
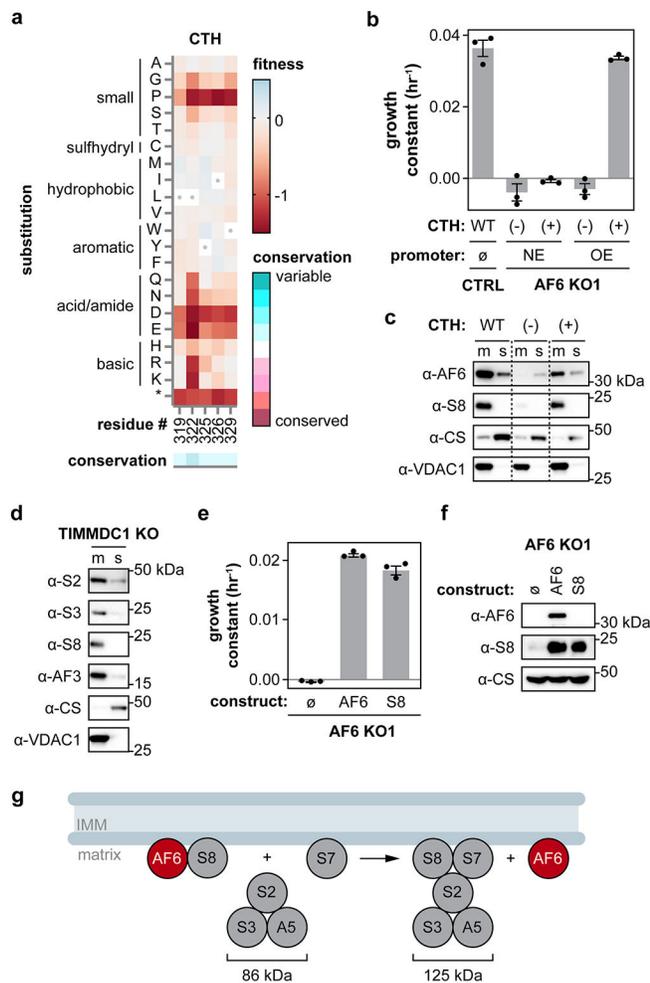
Figure 4: AF6 mediates assembly at the IMM

a. DMS fitness scores in the solvent-facing residues in the C-terminal helix (CTH). Conservation at each position is shown using the ConSurf color scale.

b. Growth constants in galactose media of CTRL cells and AF6 KO cells expressing the negative CTH mutant [(−)] or positive CTH mutant [(+)] under either the native expression (NE) promoter or the overexpression (OE) promoter. Bars represent the mean ± SEM, n=3 biologically independent experiments (replicate values shown as dots).

c. Western blots of membrane (m) and soluble (s) fractions of crude mitochondria isolated from AF6 KO cells overexpressing wild-type AF6, the negative CTH mutant, or the positive CTH mutant probing for AF6 and NDUFS8. VDAC1 is used as a membrane fraction marker and loading control. Citrate synthase (CS) is used as a soluble fraction marker. This experiment was repeated independently two times with similar results.

d. Western blots of membrane (m) and soluble (s) fractions of crude mitochondria isolated from TIMMDC1 KO HAP1 cells, probing for NDUFS2, NDUFS3, NDUFS8, and NDUFAF3. VDAC1 is used as a membrane fraction marker and loading control. Citrate synthase (CS) is used as a soluble fraction marker. See Fig. 3f for native PAGE migration pattern of blot targets in TIMMDC1 KO HAP1 cells. This experiment was repeated independently three times with similar results.

e. Growth constants in galactose media of AF6 KO cells and AF6 KO cells overexpressing either AF6 or NDUFS8. Bars represent the mean ± SEM, n=3 biologically independent experiments (replicate values shown as dots).

f. Western blots of whole cell lysate from AF6 KO cells and AF6 KO cells overexpressing either AF6 or NDUFS8, probing for AF6 and NDUFS8. Citrate synthase (CS) is used as a loading control. This experiment was performed once.

g. Model of AF6 role in CI assembly. AF6 binds to NDUFS8 and mediates its incorporation into the 125 kDa Q module intermediate at the inner mitochondrial membrane (IMM).
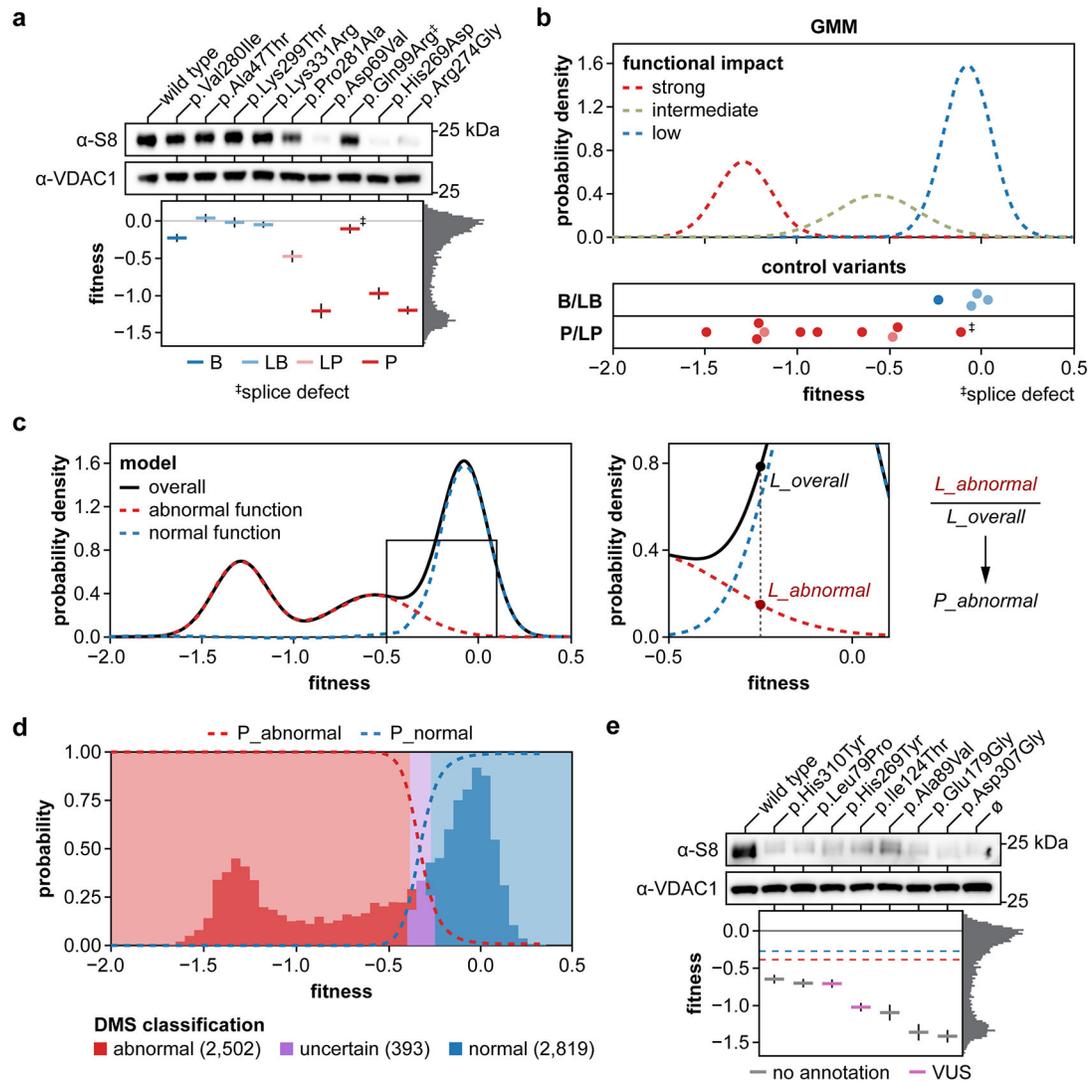
Figure 5: DMS provides a diagnostic resource for AF6-related diseases

a. Western blots of whole cell lysate from AF6 KO cells expressing control AF6 variants, probing for NDUFS8. VDAC1 is used as a loading control. DMS fitness scores for the variants are presented below the western blots as mean ± DiMSum error estimate, n=5 biologically independent samples. The color of the bar represents the current ClinVar annotations: benign (B), likely benign (LB), pathogenic (P), and likely pathogenic (LP). The histogram on the side represents the distribution of fitness scores in the full DMS dataset.

b. The three components of the GMM corresponding to variants with strong, intermediate, and low functional impact, respectively. DMS fitness scores for benign/likely benign (B/LP) and pathogenic/likely pathogenic (P/LP) variants from ClinVar are shown below as individual points.

c. Schematic for the calculation of probability of abnormal function ($P_{abnormal}$). The models derived from the GMM are shown on the left. The boxed region of the graph is magnified on the right, with an example calculation. For a given fitness score, the $P_{abnormal}$ is calculated as

the likelihood of observing that fitness score given the abnormal function model ($L_{abnormal}$) divided by the likelihood of observing that fitness score given the overall model ($L_{overall}$).

d. Histogram showing the distribution of DMS fitness scores overlaid with probabilities of abnormal or normal function (dashed lines). The blue, purple, and red shaded regions represent the variants classified as "normal," "uncertain," or "abnormal" function, respectively. The number of variants in each category is shown below.

e. Western blots of whole cell lysate from AF6 KO cells expressing candidate pathogenic AF6 variants, probing for NDUFS8. VDAC1 is used as a loading control. DMS fitness scores for the variants are presented below the western blots as mean ± DiMSum error estimate, n=5 biologically independent samples. The color of the bar represents the current ClinVar annotations. The blue and red dashed lines represent the fitness score cutoffs for normal and abnormal function, respectively. The histogram on the side represents the distribution of fitness scores in the full DMS dataset.