Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Nat Genet. Author manuscript; available in PMC 2011 April 01.

Published in final edited form as:

Nat Genet. 2010 October ; 42(10): 851–858. doi:10.1038/ng.659.

High-throughput, pooled sequencing identifies mutations in *NUBPL* and *FOXRED1* in human complex I deficiency

Sarah E Calvo[1,2,3,*], **Elena J Tucker**[4,5,*], **Alison G Compton**[4,*], **Denise M Kirby**[4], **Gabriel Crawford**[3], **Noel P Burtt**[3], **Manuel A Rivas**[1,3], **Candace Guiducci**[3], **Damien L Bruno**[4], **Olga A Goldberger**[1,2], **Michelle C Redman**[3], **Esko Wiltshire**[6,7], **Callum J Wilson**[8], **David Altshuler**[1,3,9], **Stacey B Gabriel**[3], **Mark J Daly**[1,3], **David R Thorburn**[4,5,†], and **Vamsi K Mootha**[1,2,3,†]

[1] Center for Human Genetic Research, Massachusetts General Hospital, Boston, MA [2] Department of Systems Biology, Harvard Medical School, Boston, MA [3] Broad Institute of Harvard and MIT, Cambridge, MA [4] Murdoch Childrens Research Institute and Victorian Clinical Genetics Services, Royal Children's Hospital, Melbourne, Australia [5] Department of Paediatrics, University of Melbourne, Melbourne, Australia [6] Department of Paediatrics and Child Health, University of Otago Wellington, Wellington, New Zealand [7] Central Regional Genetics Service, Capital and Coast District Health Board, Wellington, New Zealand [8] National Metabolic Service, Starship Children's Hospital, Auckland, New Zealand [9] Department of Genetics, Harvard Medical School, Boston, MA

Abstract

Discovering the molecular basis of mitochondrial respiratory chain disease is challenging given the large number of both mitochondrial and nuclear genes involved. We report a strategy of focused candidate gene prediction, high-throughput sequencing, and experimental validation to uncover the molecular basis of mitochondrial complex I (CI) disorders. We created five pools of DNA from a cohort of 103 patients and then performed deep sequencing of 103 candidate genes to spotlight 151 rare variants predicted to impact protein function. We used confirmatory experiments to establish genetic diagnoses in 22% of previously unsolved cases, and discovered that defects in *NUBPL* and *FOXRED1* can cause CI deficiency. Our study illustrates how large-scale sequencing, coupled with functional prediction and experimental validation, can reveal novel disease-causing mutations in individual patients.

Users may view, print, copy, download and text and data- mine the content in such documents, for the purposes of academic research, subject always to the full Conditions of use: http://www.nature.com/authors/editorial_policies/license.html#terms

†Correspondence should be addressed to VKM (vamsi@hms.harvard.edu) or DRT (david.thorburn@mcri.edu.au).

*contributed equally to this work

AUTHOR CONTRIBUTIONS

This study was conceived and designed by SEC, DRT, and VKM with input from MJD and SBG. Enzyme diagnosis of the cohort was coordinated by DMK. EW and CJW provided clinical interaction and assisted with sample collection. Samples were collected by DMK, EW, and CJW and prepared by AGC and EJT. The pooled sequencing protocol was designed and established at the Broad Institute by DA, MJD and SBG. Project management was performed by SEC, NPB, and CG. GC performed pooling. MCR and CG performed the genotyping. SEC designed and performed the computational analyses, with assistance from EJT, AGC, and MR. All experiments were designed and performed by EJT, AGC, and OAG. Affymetrix array-based cytogenetic analysis was performed by DLB. Syzygy was developed and run by MR and MJD. The manuscript was written by SEC, EJT, AGC, DRT, and VKM. All aspects of the study were supervised by DRT and VKM.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Complex I (CI) of the mitochondrial respiratory chain is a large ~1MDa macromolecular machine composed of 45 protein subunits encoded by both the nuclear and mitochondrial (mtDNA) genomes. CI is the main entry point to the respiratory chain and catalyzes the transfer of electrons from NADH to ubiquinone while pumping protons across the mitochondrial inner membrane. Defects in CI activity are the most common type of human respiratory chain disease, which collectively has an incidence of 1 in 5000 live births[1]. CI deficiency can present in infancy or early adulthood and shows a wide range of clinical manifestations, including Leigh Syndrome, skeletal muscle myopathy, cardiomyopathy, hypotonia, stroke, ataxia, and lactic acidosis[2–4]. The diagnosis of CI deficiency is challenging given its clinical and genetic heterogeneity and usually relies on biochemical assessment of biopsy material[5,6]. Estimates suggest that roughly 15–20% of isolated CI deficiency cases are due to mutations in the mtDNA, while the rest are likely caused by nuclear defects[7,8], though most of these mutations remain unknown.

To date, 25 genes underlying human CI deficiency have been identified via candidate gene sequencing, linkage analysis, or homozygosity mapping. These include 19 subunits of the complex (7 mtDNA genes, 12 nuclear genes), and 6 nuclear-encoded accessory factors that are required for its proper assembly, stability, or maturation (Supplementary Table 1). Many more assembly factors are likely required, as suggested by the 20 factors necessary for assembly of the smaller complex IV[9] and by cohort studies that estimate that only half of CI patients have mutations in known genes[10–13].

Additional proteins required for CI activity are likely to reside in the mitochondrion and aid in its assembly and regulation. To systematically predict such proteins, we combined our recent MitoCarta inventory of mitochondrial proteins[14] with functional prediction through phylogenetic profiling[15,16]. Ogilvie and colleagues initially used phylogenetic profiling to identify the CI assembly factor *NDUFAF2*[17]. We generalized this method to identify 34 additional candidates[14], three of which have been shown to harbor mutations causing inherited forms of CI deficiency[14,18,19]. The remaining predictions, combined with all the known CI structural subunits and assembly factors, comprise a focused set of 103 candidate genes for human CI deficiency (Supplementary Table 1).

Recent technological advances[20] offer the prospect of sequencing all 103 candidate genes in a cohort of patients with clinical and biochemical evidence of CI deficiency. Such “massively parallel” sequencing technology yields a tremendous amount of sequence in each run, far greater than that needed to interrogate 103 candidate genes in a single patient. Therefore, we used a pooled sequencing approach to assess candidate gene exons across many individuals. We created pools of DNA from ~20 individuals, selected target regions, sequenced to high depth, and detected novel variants present within each pool (Figure 1). We then used genotyping technology to type these newly discovered variants, as well as previously reported pathogenic mutations, in all patients. Finally, we confirmed the pathogenicity of prioritized variants using molecular approaches including cDNA rescue in patient fibroblasts.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Here, we report the results of our project, which we term “Mito10K” reflecting the 103 candidate genes sequenced in 103 patients with CI deficiency.

RESULTS

Discovery of new variants in CI patients using pooled sequencing

Our cohort of 103 patients had “definite”, isolated CI deficiency based on biochemical assessment. The cohort included 60 patients who lacked a previous molecular diagnosis as well as 43 patient controls with established molecular diagnoses (Table 1 and Supplementary Table 2). We also sequenced 42 healthy controls from the European HapMap collection. We combined DNA from these individuals into 5 pools of CI patients and 2 pools of HapMap controls, with each pool containing DNA from 20 or 21 individuals. For each pool, we performed PCR amplification to capture the 145 Kb of target sequence, which included 653 nuclear-encoded exons (138 Kb) and two mtDNA regions (7 Kb). PCR reactions successfully captured 97% of targeted bases. The 952 successful PCR amplicons were combined in equimolar amounts, concatenated, and then sheared to construct libraries. The 7 libraries were sequenced using a single Illumina Genome Analyzer flow cell, with one pool per lane (see Methods).

High-throughput sequencing yielded large amounts of high quality data for each pool (Supplementary Table 3). We captured 90% of our nuclear target regions at ≥100X coverage and achieved 3,359X median coverage per pool, corresponding to an average of 168X per individual (Supplementary Fig. 1). Approximately 10% of nuclear target regions were poorly covered, largely due to skewed GC content (Supplementary Fig. 1). The mtDNA target regions showed substantially higher coverage (10,144X median coverage). However the mtDNA in the pooled samples was not uniformly distributed across patients, primarily due to biases introduced by whole genome amplification (Supplementary Fig. 2). In one pool, for example, 96% of the mtDNA came from a single patient. Nonetheless, the deep coverage of mtDNA permitted variant discovery even in some poorly represented samples.

We next aimed to identify low frequency single nucleotide variants (SNVs) and small insertion/deletion variants (indels) in the pooled samples. Given the estimated 1% error rate of individual Illumina reads, detecting alleles present in 1:40 chromosomes is intrinsically challenging. Therefore, we developed a method called Syzygy to empirically estimate error rates at each base in order to confidently identify rare variants (Rivas *et al*., manuscript in preparation, and Supplementary Note). Using this method, we detected 652 high-confidence variants within the patient pools (Table 2). To improve sensitivity, we additionally applied an *ad hoc* approach to identify 246 low-confidence variants supported by at least 3 reads on each strand (Table 2). We identified a total of 898 high and low confidence variants.

Next, we assessed accuracy of these 898 variants using known genotypes available from our patient controls and HapMap controls[21]. Overall, we achieved 92% sensitivity and 99.6% specificity for control SNVs present at nuclear DNA sites with ≥100 reads (see Methods and Supplementary Table 3). We note that this high sensitivity is due to the deep sequence coverage and to the relatively high allele frequency for many HapMap control variants (Supplementary Fig. 3). However, as expected, we observed lower sensitivity for rare

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

nuclear variants: 86% for doubletons and 66% for singletons in a pool. For mtDNA variants, we achieved high sensitivity and specificity in genomic DNA of HapMap controls (96% and 100%, respectively) but much lower sensitivity for patient controls (32%) due to the non-uniform distribution of mtDNA within each pool. The minor allele frequencies estimated from read counts correlated strongly with expected frequencies in HapMap pools (R^2=0.96), indicating high fidelity of the pooled sequencing protocol (Supplementary Fig. 3).

Next we prioritized the 898 discovered variants to focus our attention on those that are likely to underlie a rare and devastating phenotype (Figure 2a). Briefly, we filtered out: (i) variants that were present in healthy individuals, based on HapMap controls, dbSNP[22], mtDB[23], and pilot data from the 1000 genomes project, (ii) synonymous variants, and (iii) non-coding variants, unless they corresponded to tRNA or splice sites. 8 splice sites positions were selected using training data of 8189 disease-associated splice variants within the Human Gene Mutation Database (HGMD)[24] (Figure 2b). In addition, we filtered out missense variants at sites with low evolutionary conservation, as these sites had a reduced frequency of pathogenic mutations based on training data (Figure 2c). See Methods for details. Using these filters, we prioritized for genotyping 109 high-confidence variants and 107 low-confidence variants that were deemed ‘likely deleterious’.

Together, the discovery screen and stringent definition of ‘likely deleterious’ variants captured 18/23 (78%) of the causal nuclear variants and 7/25 (28%) of causal mtDNA variants within our CI patient controls. The approach missed 4 nuclear and 17 mtDNA variants in the discovery screen, and filtered out 1 nuclear splice variant located 4bp into an intron and 1 mtDNA missense variant at a poorly conserved site (Supplementary Table 2).

Genotyping previously known and newly discovered variants in CI patients

Our next goal was to genotype the discovered ‘likely deleterious’ variants, as well as previously known disease variants, in each patient sample (Supplementary Table 4 and Methods). The genotyping served multiple purposes. First, it was necessary to validate novel variants from the pooled discovery screen. Second, it enabled us to search for previously known mutations underlying CI deficiency, which were not detected in our discovery screen due to a lack of power (e.g., mtDNA variants). Third, it allowed us to assign the variants to individual patients.

Of the newly discovered ‘likely deleterious’ variants, we validated 84% of high-confidence variants, and as expected, only 11% of low-confidence variants (Supplementary Table 4). ‘Less likely deleterious’ variants had a higher 96% validation rate, based on 101 additional high-confidence variants genotyped (Supplementary Table 4). We further validated SNVs of particular interest using Sanger sequencing, since Sequenom genotypes showed an estimated 11% false positive rate for extremely rare variants (Supplementary Note). In a subset of instances where we identified heterozygous variants of interest, we used Sanger sequencing to fully resequence the gene.

In total, we validated 151 ‘likely deleterious’ patient variants corresponding to 115 unique loci (91 high-confidence, 12 low-confidence, and 12 pathogenic variants missed in the discovery screen). Detailed data are provided in Supplementary Table 2. We detected a

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

higher frequency of 'likely deleterious' variants in our patient cohort compared to European controls, although this enrichment might be due to differences in ancestry (Supplementary Note).

Newly discovered mutant alleles in CI patients

With the Mito10K sequence data in hand, we next looked for homozygous, compound heterozygous and pathogenic mtDNA variants within our cohort of 60 undiagnosed patients (Figure 3). We expected that many patients would have homozygous or two heterozygous variants in known disease-related genes, consistent with recessive inheritance. We refer to these variants as 'recessive-type'.

Only 3 patients had previously reported pathogenic mtDNA mutations and only 8 patients had recessive-type mutations in known disease genes, including 5 novel and 2 previously reported mutations (Table 3). Of interest, 2 patients had recessive-type mutations in candidate disease genes (*NUBPL*, *FOXRED1*) (Table 3). The remaining patients included 3 with mtDNA 'likely deleterious' variants of unknown clinical significance, 17 with heterozygous 'likely deleterious' nuclear variants of unknown clinical significance, and 27 with no 'likely deleterious' variants (Supplementary Table 2).

Establishing 11 patient diagnoses in known disease genes

We next assessed the pathogenicity of variants detected within the 3 patients with causal mtDNA mutations (in *ND3*[25], *ND5*[26], and *MT-TW*[26]) and the 8 patients with recessive-type variants in previously reported CI disease genes: *NDUFS4*[10],[27–31], *NDUFAF2*[17], *NDUFV1*[32], and *NDUFS8*[33] (Table 3). The discovered patient mutations were absent from all other patient and HapMap samples sequenced, except as noted below.

We identified one novel and two previously reported *NDUFS4* mutations in 3 patients with Leigh Syndrome (Table 3 and Supplementary Fig. 4). Siblings, DT37 and DT38, were compound heterozygous for the reported mutations c.462delA (p.K154NfsX34)[30] and c. 99-1G>A (p.S34IfsX4)[10]. The unrelated patient DT107 was compound heterozygous for the same c.99-1G>A mutation and a novel mutation c.351-2A>G, inherited from his father and mother, respectively. *In silico* and RT-PCR analyses indicated that both the c.99-1G>A and c.351-2A>G mutations alter *NDUFS4* splicing. The heterozygous c.351-2A>G mutation was detected in DT107 genomic DNA, however it was undetectable in cDNA +/−cycloheximide (CHX) suggesting a high level of mRNA instability. Western blot analysis on fibroblasts from patients DT38 and DT107 showed no detectable NDUFS4 protein. This is the second report of the c.99-1G>A mutation[10] and the third of the c.462delA mutation[28],[30] suggesting not only that *NDUFS4* shows recurrent mutations underlying Leigh Syndrome but also that several previously unrecognized founder mutations may exist in this gene.

We also identified novel homozygous mutations in *NDUFAF2* in 3 patients presenting with Leigh Syndrome (Table 3 and Supplementary Fig. 5). A consanguineous patient, DT16, harbored a homozygous c.221G>A mutation (p.W74X) within a 6.3Mb region of homozygosity (determined by Affymetrix 250K *Nsp* SNP chip). Siblings, DT67 and DT68, harbored a homozygous c.103delA mutation (p.I35SfsX17). Analysis of cDNA from patient

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

fibroblasts demonstrated that *NDUFAF2* transcripts containing these mutations were stable. Additionally, the c.221G>A nonsense mutation in DT16 (located 4bp into exon 3) resulted in occasional exon 3 skipping, which generates a transcript also encoding a truncated protein (p.A73GfsX5). All three patients lacked any detectable NDUFAF2 protein by western blot analysis indicating that the truncated protein products are unstable.

A novel homozygous *NDUFV1* mutation (c.1129G>A, p.E377K) was identified in a 2.1Mb region of homozygosity (determined by Affymetrix 250K *Nsp* SNP chip) in a consanguineous Lebanese patient, DT3, who presented with lethal infantile mitochondrial disease (LIMD) (Table 3 and Supplementary Fig. 6). Both unaffected parents were heterozygous carriers. This mutation introduces a positively charged residue within the consensus motif for the iron sulfur binding site (pfam10589) which is highly conserved across eukaryotic species.

We identified a novel homozygous *NDUFS8* mutation (c.460G>A, p.G154S) in a Sudanese patient, DT61, presenting with mitochondrial encephalopathy (Table 3 and Supplementary Fig. 7). This mutation affects a highly conserved amino acid and alters polarity within the highly conserved Fer4 4Fe-4S iron-sulfur cluster binding domain (pfam00037). This mutation segregated with disease in this family, with an affected sibling also homozygous while both unaffected parents were heterozygous carriers.

Novel genes underlying CI deficiency: *NUBPL* and *FOXRED1*

Within our 60 patients, we also discovered recessive-type mutations in two genes not previously linked to CI deficiency: *NUBPL* and *FOXRED1*.

Patient DT35 presented with mitochondrial encephalomyopathy and was found to contain an apparent homozygous *NUBPL*:c.166G>A mutation (Supplementary Fig. 8). We did not detect this mutation in the 204 other patient chromosomes or 84 HapMap control chromosomes sequenced. This mutation is predicted to cause substitution of a highly conserved glycine residue with arginine (p.G56R), 18 amino acids from the mitochondrial targeting sequence cleavage site predicted by TargetP (Supplementary Fig. 8). Although the patient's father was heterozygous for this mutation, the mother did not carry the mutation (Supplementary Fig. 8). In order to determine whether the mother may have transmitted a deletion involving this portion of exon 2, we performed Affymetrix array-based cytogenetic analysis on DNA from DT35. We detected a complex chromosomal rearrangement including a ~240Kb deletion spanning exons 1–4 of *NUBPL* and a ~130Kb duplication involving exon 7 of *NUBPL* as shown in Supplementary Fig. 8. Next, we assessed *NUBPL* mRNA species present in DT35. RT-PCR showed very low expression of the full-length transcript, while the predominant mRNA species was a shorter fragment (Supplementary Fig. 8). Sequencing revealed that the shorter fragment resulted from exon 10 skipping, and that it contained the c.166G>A mutation, suggesting it was the paternal allele. There was no evidence of expression of the maternal allele. To determine the cause of exon 10 skipping, we performed Sanger sequencing of exon 10 and the flanking intronic regions (an area of previous poor high throughput sequence coverage). A c.815-27T>C mutation was identified that is predicted to ablate a consensus branch sequence. This mutation was present in 2/232 Caucasian control chromosomes. Thus, DT35 contains one *NUBPL* allele harboring a

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

deletion spanning exons 1–4 and a second allele that harbors both a p.G56R missense mutation and a c.815-27T>C mutation that likely causes exon 10 skipping.

We performed a complementation experiment to assess whether the introduction of wildtype cDNA into patient fibroblasts rescued the defect in CI activity. Fibroblasts from this patient exhibit a strong CI defect, with only 19% residual CI activity when assayed by spectrophometric enzyme assay and 40% residual CI activity when assayed by dipstick enzyme assay. Using a lentiviral expression system, we transduced patient fibroblasts with wildtype cDNA. Expression of wild-type *NUBPL* rescued CI activity in the patient with *NUBPL* mutations but did not alter CI activity of the control fibroblasts or patient fibroblasts with *FOXRED1* mutations (Figure 4a), establishing *NUBPL* as the causal gene in this case.

Although we have proven that *NUBPL* underlies complex I deficiency in this patient, we have not established the pathogenicity of individual mutations. Due to its prevalence in controls, the c.815-27T>C branch site mutation may be a pseudo-deficiency allele, that if homozygous generates sufficient full-length *NUBPL* transcript for NUBPL functionality. However, this mutation may be pathogenic when inherited with a null allele such as in DT35. Alternatively, the p.G56R missense mutation may abolish *NUBPL* function or may act in synergy with the branch-site mutation to cause disease.

Patient DT22 presented with Leigh Syndrome and was found to be compound heterozygous for two mutations in *FOXRED1*, c.694C>T (p.Q232X) and c.1289A>G (p.N430S) (Supplementary Fig. 9). The c.694C>T mutation was detected in the discovery screen and was not detected in 204 other patient chromosomes or 84 HapMap control chromosomes. The c.1289A>G mutation was in an area of low coverage but was subsequently identified by Sanger sequencing of *FOXRED1* and was not present in 102 control chromosomes of European ancestry screened by RFLP analysis. Analysis of cDNA from fibroblasts treated with CHX to inhibit nonsense mediated decay demonstrated the presence of both mutations. In the absence of CHX, however, the transcript containing the c.694C>T (p.Q232X) was undetectable leaving the transcript containing the c.1289A>G mutation as the predominant species, consistent with compound heterozygosity (Supplementary Fig. 9). The c.1289A>G mutation was inherited from the patient's mother, and is predicted to cause the substitution of a highly conserved asparagine residue with a serine (p.N430S) (Supplementary Fig. 9). Paternal DNA was not available for genotyping. RT-PCR analysis of patient cDNA also shows occasional skipping of exon 6 (containing c.694C>T), which results in a transcript predicted to lack 40 internal residues (Supplementary Fig. 9).

As above, we performed a complementation experiment in patient fibroblasts to assess the role of *FOXRED1* in CI activity. Fibroblasts from this patient exhibit a striking CI defect, with only 9% residual CI activity when assayed by spectrophometric enzyme assay and 15% residual CI activity when assayed by dipstick enzyme assay. We were able to rescue the defect in these fibroblasts using lentiviral-mediated cDNA rescue with the wild-type *FOXRED1* cDNA, and this rescue was specific to this cell line (Figure 4b).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Together, the mutation data and complementation experiments support *NUBPL* and *FOXRED1* as *bona fide* CI disease-related genes in individuals DT35 and DT22, respectively.

The mutational spectrum of CI deficiency

The large-scale discovery and validation studies for 60 patients reported here, in addition to the previous molecular diagnosis of all 43 other patients with definite isolated CI deficiency seen at our diagnostic laboratory, provide the largest systematic sequencing study of CI deficiency to date. Our cohort of 103 patients includes 94 unrelated individuals; 52% of them now have firm genetic diagnoses, including diagnoses due to mtDNA mutations (29%), recessive-type mutations (22%), and X-linked mutations (1%) (Figure 5). These represent 33% with mutations in CI structural subunits, 6% with mutations in established CI assembly factors (including *NUBPL*), 7% with tRNA mutations required for mtDNA translation, 4% with mutations in other auxiliary factors (mtDNA replication proteins *POLG* and *C10orf2*, and the *TAZ* protein required for CI stability via the maintenance of cardiolipin pools within the mitochondrial inner membrane)[34], and 1% with mutations in an uncharacterized gene (*FOXRED1*).

DISCUSSION

Advances in genome sequencing technology offer a new opportunity to solve the genetic basis of disease even beginning with individual cases. Perhaps the major challenge of human genetics moving forward will be distinguishing pathogenic alleles from the plethora of benign sequence differences between individuals. Even within the protein coding portion of the genome, each person carries an estimated 400–500 protein-modifying rare variants[35],[36]. Several recent whole-exome sequencing projects have detected causal variants for Mendelian disease by using multiple affected individuals to hone in on regions of interest, and established pathogenicity by identifying different mutations in unrelated individuals with the same phenotype[36],[37]. While this approach has broad utility, it may not be readily applicable to individual, sporadic cases of disease.

In the current Mito10K project, we have demonstrated an alternate approach. We prioritized candidate genes based on functional clues, performed pooled DNA sequencing of a patient cohort, and identified novel variants that we predict to be deleterious. Key to success of our approach was the availability of cellular models of disease, with which we could establish pathogenicity of novel mutations in single patients. This strategy can be applied in principle to any disorder for which a cellular phenotype exists.

Our approach successfully discovered novel pathogenic roles for *NUBPL* and *FOXRED1*. *NUBPL* (nucleotide binding protein-like), also known as *IND1*, was recently shown to be an assembly factor for CI[38]. Similar to its role in the yeast *Y. lipolytica*, human *NUBPL* is essential for the incorporation of Fe/S clusters into CI subunits, and its knockdown causes improper assembly of the peripheral arm of CI, reduced CI activity, and abnormal mitochondrial morphology[38],[39]. We now report the first *NUBPL* mutations in a patient with CI deficiency, a male who presented at 2 years of age with developmental delay, leukodystrophy and elevated CSF lactate (see Supplementary Note for complete clinical

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

description). Marked CI deficiency was observed in muscle biopsy and skin fibroblasts (37% and 19% normalized activity relative to controls). Sequencing of DNA from this patient revealed an apparent homozygous

NUBPL:p.G56R missense mutation in an amino acid that has been conserved across all 36 aligned vertebrate species. However, further analysis indicated that this patient was actually compound heterozygous: one allele contained both the p.G56R missense mutation and a branch site mutation that caused skipping of exon 10, and the other allele contained a complex chromosomal rearrangement involving deletion of exons 1–4 and duplication of exon 7 of *NUBPL*. This patient highlights the limitations of 2nd generation sequencing. Large deletions are not detected and variants such as branch site mutations may be missed or overlooked. Nevertheless, the CI defect in patient fibroblasts was rescued by expression of a wildtype allele of *NUBPL*, thus establishing a pathogenic role for *NUBPL* mutations in CI deficiency.

We also discovered pathogenic mutations in *FOXRED1*, which is an uncharacterized protein that derives its name from a FAD dependent oxidoreductase protein domain. This gene was selected as a candidate solely based on its mitochondrial localization[40] and shared phylogenetic profile with CI subunits[14]. We detected *FOXRED1* mutations in a male infant who presented at birth with congenital lactic acidosis and was diagnosed with Leigh Syndrome at 6 years of age (see Supplementary Note for complete clinical description). Severe CI deficiency was observed in muscle biopsy and fibroblasts (9% of normal control mean in both samples relative to citrate synthase). Sequencing this patient revealed compound heterozygous *FOXRED1* mutations: a p.Q232X nonsense mutation and a p.N430S missense mutation in a conserved amino acid. As with *NUBPL* above, cDNA rescue established *FOXRED1* as a novel disease-related gene. At present the function of *FOXRED1* is not clear, though its four human homologs (*DMGDH*, *SARDH*, *PIPOX*, *PDPR*) perform redox reactions in amino acid catabolism, suggesting a potential link between amino acid metabolism and CI.

While the Mito10K project successfully identified or confirmed pathogenic mutations in half of the 103 patients with CI deficiency (Figure 5), it is notable that we were unable to identify “smoking gun” mutations for the remaining half. Our results are comparable to a recent sequencing study of X-linked mental retardation[41]. While in some of the undiagnosed CI patients we detected ‘likely deleterious’ variants that may contribute to pathogenesis, most contain no such variants. It is likely that the true causal variants in the unsolved cases (i) reside in a non-targeted gene, (ii) reside in a non-targeted region, such as a regulatory region or un-annotated exon, (iii) were not detected due to lack of sensitivity, especially within the mtDNA, (iv) contain full exon or gene deletions, which our approach cannot detect, or (v) were present in our discovery screen but filtered out by our stringent criteria. Additionally, it is possible that in some patients the disease is caused by complex inheritance or epigenetic mechanisms. Broader sequencing, combined with functional validation, will be required to fully elucidate the molecular basis of these remaining cases.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

ONLINE METHODS

CI patients and controls

The 60 patients plus 43 patient controls had a definite diagnosis of isolated CI deficiency, based on spectrophotometric enzyme assays interpreted by previously described criteria[5],[42]. Briefly, the ratio of CI activity to citrate synthase or relative to Complex II, was required to be ≤25% of normal, and the normalized activity of complexes II, III, and IV were required to be at least twofold higher than CI activity (Supplementary Fig. 10). The cohort includes all such patients diagnosed in Melbourne from 1992 to 2007, with the exception of 9 patients from whom no suitable DNA was available for sequencing.

DNA preparation and pooling

DNA was isolated from cultured cells using a Nucleon DNA Extraction kit or from patient tissues (skeletal or cardiac muscle and liver) by proteinase K digestion followed by salting-out. Each patient sample was whole-genome amplified using a QIAGEN REPLI-g™ Kit with 100ng input DNA. HapMap samples were not whole-genome amplified. DNA concentration was measured by Quant-iT™ PicoGreen® dsDNA reagent detected on a Thermo Scientific Varioskan Flash. DNA concentration was normalized to 20ng/μL based on two rounds of quantification and dilution, yielding mean 19.2ng/μl concentration (1.56 standard deviation). We allowed for 10% variance as that is the accuracy limit of PicoGreen® quantitation. The normalization steps were automated using the Packard Multiprobe II HT EX. The same robotic automation was used across the entire set and in all steps in order to guarantee a uniform pipetting error. 20 or 21 samples where then pooled in equimolar amounts. Each patient pool contained patients with unknown diagnoses, known mtDNA mutations, and known nuclear mutations, with the following counts: Pool1=12, 5, 4; Pool2=13, 5, 3; Pool3=12, 5, 4; Pool4=12, 5, 3; Pool5=11, 5, 4. See Supplementary Note for HapMap sample identifiers.

Target selection

Targets included 2 mtDNA regions and coding and UTR exons of 111 RefSeq transcripts (release 29) from 103 gene loci (Supplementary Table 1). Primers were iteratively designed using PRIMER3 software on the hg17 reference sequence (150–600bp amplicon length, no buffer) and validated on 3 HapMap CEU samples, using three design iterations. *NotI* tails were added to provide a recognition site for downstream concatenation. Target regions were PCR-amplified using 20 ng of whole-genome-amplified DNA, 1× HotStar buffer, 0.8 mM dNTPs, 2.5 mM $MgCl_2$, 0.2 units of HotStar Enzyme (Qiagen), and 0.25 μM forward and reverse primers in a 10-μl reaction volume. PCR cycling parameters were: one cycle of 95°C for 15 min; 35 cycles of 95°C for 20 s, 60°C for 30 s, and 72°C for 1 min; followed by one cycle of 72°C for 3 min. The PCR products were separately quantified, normalized and pooled as described above. Secondary confirmation was ascertained by testing one column of PCR product per plate on 2% agarose E-gel against a 1kb DNA ladder to visualize PCR product size. The PCR products were then pooled by DNA sample pool using Packard Multiprobe II HT EX.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Sequencing

The PCR products for each pooled sample were concatenated using *NotI* adapters and sheared into fragments as previously described[43]. Libraries were constructed by a modified Illumina single-end library protocol, with 225–275bp gel size selection and PCR enrichment using 14 cycles of PCR, and then single-end sequenced with 76 cycles on an Illumina Genome Analyzer. 76bp reads were aligned to the genome using MAQ algorithm[44] within the Picard analysis pipeline, and further processed using the SAMtools software[45] and custom scripts.

Variant discovery

High-confidence SNVs were detected within each pooled sample using the Syzygy algorithm on targeted bases with a minimum of 100 high-quality aligned reads (base quality ≥20, mapping quality >0, ≥30 reads on each strand). High confidence SNVs had log odds (LOD) scores ≥3, with the strand-specific LOD>-1.5 or a Fisher's exact test of strand bias >0.1 (see Supplementary Note). Low-confidence SNVs were supported by at least 3 reads on each strand (base quality ≥20, mapping quality >0, ≥200 reads on each strand). Indels were identified from within unaligned reads, and were supported by ≥10 unaligned reads on each strand that contained an insertion/deletion preceding an exact 20bp match to a targeted exon, excluding indels adjacent to homopolymer runs (see Supplementary Note).

Discovery screen sensitivity was estimated from genotype data using sites where ≥1 individual in the pool contained a variant compared to hg18, whereas specificity was calculated at sites where all individuals contained the hg18 reference allele.

Variants were annotated as 'likely deleterious' based on any of the following criteria: i) previously reported as a disease variant, based on manual curation and the Human Gene Mutation Database (HGMD) professional version 2009.1[24]; ii) present in a mitochondrial tRNA gene; iii) present in 5′ UTR and altering the presence of an upstream ORF[46]; iii) present at a splice site (splice acceptor sites −1, −2, −3, and splice donor sites −1,1,2,3,5 selected based on training data consisting of all 8189 HGMD disease-associated splice variants); iv) coding indel; v) nonsense variant; vi) missense variant at an amino acid conserved in ≥10 aligned vertebrate species, based on the multiz44way genome alignments downloaded from UCSC genome browser[47] (see Supplementary Note), or predicted as 'damaging' by PolyPhen-2.0 (HumVar training data)[48] (see Supplementary Note). Variants that were not previously associated with disease were excluded if present in 42 HapMap controls, dbSNP[22], 1000 genomes pilot 1, or present at >0.005 minor allele frequency in mtDB[23] based on the frequency of asymptomatic carriers of pathogenic mtDNA mutations[49]. Conservation thresholds were selected from training data: all disease-associated missense variants in HGMD version 2009.1, and all dbSNP128 sites annotated as nonsynonymous, excluding those present in HGMD.

Genotyping

SNVs were assayed within whole-genome-amplified DNA from the 103 CI patients using Sequenom MassARRAY® iPLEX™ GOLD chemistry[50]. Oligos were synthesized and mass-spec QCed at Integrated DNA Techologies, Inc. All SNVs were genotyped in

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

multiplexed pools of 20–38 assays, designed by AssayDesigner v.3.1 software, starting with 10ng of DNA per pool. ~7 nl of reaction was loaded onto each position of a 384-well SpectroCHIP preloaded with 7 nl of matrix (3-hydroxypicolinic acid). SpectroCHIPs were analyzed in automated mode by a MassArray MALDI-TOF Compact system with a solid phase laser mass spectrometer (Bruker Daltonics Inc., 2005). We obtained high quality data (>95% genotype call rate, HWE P-value >0.001 and MAF >1%) in all samples that had at least one SNV. Variants were called by real-time SpectroCaller algorithm, analyzed by SpectroTyper v.4.0 software and manually reviewed for rare variants.

Deletions and selected SNVs were validated by Sanger resequencing, performed on genomic DNA, using ABI 3130XL and BigDye v3.1 Terminators (Applied Biosystems) as per manufacturer's protocols.

Cloning

The *FOXRED1* open reading frame (ORF) was purchased in a pDONR223 vector (Clone ID: 3956972, Open Biosystems) and cloned into pLEX TRC970 (V5 C-terminal tag) via Gateway cloning (Invitrogen). Initial experiments using this vector did not rescue CI activity so site-directed mutagenesis was performed to change codon 343 from CCA (proline) (dbSNP rs17855445) to the hg18 reference codon GCA (alanine) using QuikChange II XL site-directed mutagenesis kit (Stratagene) according to manufacturer's instructions (primers listed in Supplementary Table 5) to generate the RefSeq FOXRED1-V5 pDest vector. Full-length NUBPL ORF was amplified from MCH58 cells by RT-PCR incorporating Gateway adaptors, then was cloned into into pLEX TRC970 (V5 C-terminal tag) via Gateway cloning to generate the NUBPL-V5 pDest vector.

Viral particle production and transduction

HEK-293T cells were grown on 10cm plates to 60% confluence and cotransfected with a packaging plasmid (pCMV-δ8.91), a pseudotyping plasmid (pMD2-VSVg) and either NUBPL-V5-pDest or FOXRED1-V5-pDest. Transfection was performed using Effectene reagents (Qiagen) according to the manufacturer's protocol. Fresh media was applied to the cells 16 hours post-transfection and, following 24 hours incubation, supernatants containing packaged virus were harvested and filtered through a 0.45 μM membrane filter.

Patient fibroblasts were grown to 80% confluence in 6 well plates before addition of 62.5 μL of NUBPL-V5 or 125 μL FOXRED1-V5 viral particles and polybrene at final concentration of 5 μg/ml in 8.75mL total media. Plates were spun at 2500rpm for 90 minutes and incubated for 24 hours at 37°C before replacing media. Cells were grown in antibiotic-free media for 30 hours before applying selection media containing 1 μg/mL puromycin. Following 12–20 days of selection, cells were harvested for dipstick assays.

Dipstick enzyme activity assays

CI and Complex IV (CIV) dipstick activity assays were performed on 10 μg and 15 μg, respectively, of cleared cell lysates according to the manufacturer's protocol (Mitosciences). A Hamamatsu ICA-1000 immunochromatographic dipstick reader was used for densitometry. Two-way Repeated Measures Analysis of Variance (ANOVA) was used for

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

comparisons of groups followed by post hoc analysis using the Bonferroni method to determine statistically significant differences.

Homozygosity mapping

Homozygosity was determined using SNP Mapping GeneChip *Nsp* 250 k Array (Affymetrix), performed by the Australian Genome Research Facility. Data were analyzed using the Loss of Heterozygosity (LOH) Analysis Tool of GCOS Client software (Affymetrix).

RT-PCR

RNA was extracted from cultured patient fibroblasts using the RNAspin Mini Kit (Illustra) and cDNA was generated using the SuperScript III First strand synthesis kit (Invitrogen) as per manufacturers' protocols. For analysis of nonsense mediated decay and mRNA splicing, fibroblasts were cultured in media containing 100ng/μL CHX for 24 hours prior to RNA preparation[51]. PCR primers (Supplementary Table 5) were designed to amplify the entire cDNA in either one PCR product or in overlapping segments. PCR products were either directly sequenced using ABI 3130XL and BigDye v3.1 Terminators (Applied Biosystems) as per manufacturer's protocols or following gel-purification using the MinElute Gel Extraction kit (Qiagen).

SDS-PAGE and western blot

Primary control and patient fibroblasts were lyzed in RIPA buffer (50mM Tris pH 8.0, 150mM NaCl, 1% NP-40, 0.5% sodium deoxycholate and 0.1% SDS) containing protease inhibitor cocktail (Roche). 25–50 μg of cleared lysate were run per lane on 10% NuPAGE Bis-Tris gels (Invitrogen), proteins were transferred to PVDF membranes (Millipore), blocked (PBS containing 5% skim milk powder, 0.05% Tween-20) and incubated with primary antibodies overnight at 4 °C (for primary antibody details and concentrations see Supplementary methods). After washing, membranes were incubated in anti-mouse or rabbit^{HRP} secondary antibodies (DakoCytomation used at 1:10000) at room temperature for 1 hour and developed using ECL or ECL Plus detection reagents (Amersham Bioscience).

RFLP Screen (FOXRED1:c.1289A>G and NUBPL:c.815-27T>C)

Exon 11 of *FOXRED1* or exon 10 of *NUBPL* was PCR-amplified (Supplementary Table 5) from 100ng of patient gDNA, the products were checked by gel electrophoresis, digested overnight with *AflIII* or *NlaIV* respectively (New England Biolabs) as per manufacturer's protocol, then resolved on 1% agarose gels.

Antibodies for western blotting

Antibodies included NDUFS4 (MS104, Mitosciences) at 1:1000, Porin (529534, Calbiochem) at 1:10000, Complex II 70kD subunit (A-1142. Molecular Probes) at 1:1000, and NDUFAF2 (kind gift from Dr. Mat McKenzie and Prof. Michael Ryan, La Trobe University, Bundoora, Victoria) at 1:5000.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Microarray DNA Copy Number Analysis

Genome-wide microarray analysis was conducted using the Affymetrix GeneChip 2.7M array, according to the manufacturer's instructions. Data analysis was performed using Chromosome Analysis Suite (ChAS) software v1.2 (Affymetrix).

Data availability

Supplementary Table 2 provides detailed data on all validated patient variants, and the 7 pooled sequence data files (BAM format) are available upon request.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank S. Tregoning, A. Laskowski and S. Smith for assistance with enzyme assays and DNA preparation, M. McKenzie and M. Ryan for the NDUFAF2 antibody, J. Boehm for the lentiviral expression vector, S. Flynn for assistance with human subjects protocols, R. Onofrio for designing PCR primers, K. Ardlie and S. Mahan for assistance in DNA sample preparation, J. Wilkinson and L. Ambrogio for Illumina sequence project management, T. Fennel for sequence alignment, L. Ziaugra for genotyping assistance, M. Cabili for tool evaluation, J. Flannick for assistance with pooled sequence analysis, I. Adzhubei and S. Sunyaev for kindly providing PolyPhen-2.0 predictions, M. DePristo, E. Banks, A. Sivachenko for advice on sequence data analysis, M. Garber for assistance with evolutionary conservation analyses, J. Pirruccello, R. Do, and S. Kathiresan for data and analysis of control data, and the many physicians who referred patients and assisted with these studies. This work was supported by a grant (436901) and Principal Research Fellowship from the Australian National Health & Medical Research Council awarded to DRT, an Australian Postgraduate Award to EJT and a grant from the National Institutes of Health (GM077465) awarded to VKM. The authors wish to dedicate this article to the memory of our co-author Denise Kirby, an outstanding scientist and dear colleague who died during the preparation of this manuscript.

References

1. Skladal D, Halliday J, Thorburn DR. Minimum birth prevalence of mitochondrial respiratory chain disorders in children. Brain. 2003; 126:1905–12. [PubMed: 12805096]
2. Distelmaier F, et al. Mitochondrial complex I deficiency: from organelle dysfunction to clinical disease. Brain. 2009; 132:833–42. [PubMed: 19336460]
3. Janssen RJ, Nijtmans LG, van den Heuvel LP, Smeitink JA. Mitochondrial complex I: structure, function and pathology. J Inherit Metab Dis. 2006; 29:499–515. [PubMed: 16838076]
4. Lazarou M, Thorburn DR, Ryan MT, McKenzie M. Assembly of mitochondrial complex I and defects in disease. Biochim Biophys Acta. 2009; 1793:78–88. [PubMed: 18501715]
5. Bernier FP, et al. Diagnostic criteria for respiratory chain disorders in adults and children. Neurology. 2002; 59:1406–11. [PubMed: 12427892]
6. Morava E, et al. Mitochondrial disease criteria: diagnostic applications in children. Neurology. 2006; 67:1823–6. [PubMed: 17130416]
7. McFarland R, et al. De novo mutations in the mitochondrial ND3 gene as a cause of infantile mitochondrial encephalopathy and complex I deficiency. Ann Neurol. 2004; 55:58–64. [PubMed: 14705112]
8. Dimauro S, Davidzon G. Mitochondrial DNA and disease. Ann Med. 2005; 37:222–32. [PubMed: 16019721]
9. Fontanesi F, Soto IC, Horn D, Barrientos A. Assembly of mitochondrial cytochrome c-oxidase, a complicated and highly regulated cellular process. Am J Physiol Cell Physiol. 2006; 291:C1129–47. [PubMed: 16760263]
10. Benit P, et al. Genotyping microsatellite DNA markers at putative disease loci in inbred/multiplex families with respiratory chain complex I deficiency allows rapid identification of a novel

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

nonsense mutation (IVS1nt -1) in the NDUFS4 gene in Leigh syndrome. Hum Genet. 2003; 112:563–6. [PubMed: 12616398]

11. Bugiani M, et al. Clinical and molecular findings in children with complex I deficiency. Biochim Biophys Acta. 2004; 1659:136–47. [PubMed: 15576045]
12. Lebon S, et al. Recurrent de novo mitochondrial DNA mutations in respiratory chain deficiency. J Med Genet. 2003; 40:896–9. [PubMed: 14684687]
13. Smeitink J, Sengers R, Trijbels F, van den Heuvel L. Human NADH:ubiquinone oxidoreductase. J Bioenerg Biomembr. 2001; 33:259–66. [PubMed: 11695836]
14. Pagliarini DJ, et al. A mitochondrial protein compendium elucidates complex I disease biology. Cell. 2008; 134:112–23. [PubMed: 18614015]
15. Marcotte EM, Pellegrini M, Thompson MJ, Yeates TO, Eisenberg D. A combined algorithm for genome-wide prediction of protein function. Nature. 1999; 402:83–6. [PubMed: 10573421]
16. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A. 1999; 96:4285–8. [PubMed: 10200254]
17. Ogilvie I, Kennaway NG, Shoubridge EA. A molecular chaperone for mitochondrial complex I assembly is mutated in a progressive encephalopathy. J Clin Invest. 2005; 115:2784–92. [PubMed: 16200211]
18. Saada A, et al. Mutations in NDUFAF3 (C3ORF60), encoding an NDUFAF4 (C6ORF66)-interacting complex I assembly protein, cause fatal neonatal mitochondrial disease. Am J Hum Genet. 2009; 84:718–27. [PubMed: 19463981]
19. Sugiana C, et al. Mutation of C20orf7 disrupts complex I assembly and causes lethal neonatal mitochondrial disease. Am J Hum Genet. 2008; 83:468–78. [PubMed: 18940309]
20. Bentley DR, et al. Accurate whole human genome sequencing using reversible terminator chemistry. Nature. 2008; 456:53–9. [PubMed: 18987734]
21. Frazer KA, et al. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007; 449:851–61. [PubMed: 17943122]
22. Sherry ST, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001; 29:308–11. [PubMed: 11125122]
23. Ingman M, Gyllensten U. mtDB: Human Mitochondrial Genome Database, a resource for population genetics and medical sciences. Nucleic Acids Res. 2006; 34:D749–51. [PubMed: 16381973]
24. Stenson PD, et al. The Human Gene Mutation Database: 2008 update. Genome Med. 2009; 1:13. [PubMed: 19348700]
25. Kirby DM, et al. NDUFS6 mutations are a novel cause of lethal neonatal mitochondrial complex I deficiency. J Clin Invest. 2004; 114:837–45. [PubMed: 15372108]
26. Valente L, et al. Identification of novel mutations in five patients with mitochondrial encephalomyopathy. Biochim Biophys Acta. 2009; 1787:491–501. [PubMed: 18977334]
27. Budde SM, et al. Combined enzymatic complex I and III deficiency associated with mutations in the nuclear encoded NDUFS4 gene. Biochem Biophys Res Commun. 2000; 275:63–8. [PubMed: 10944442]
28. Leshinsky-Silver E, et al. NDUFS4 mutations cause Leigh syndrome with predominant brainstem involvement. Mol Genet Metab. 2009; 97:185–9. [PubMed: 19364667]
29. Petruzzella V, et al. A nonsense mutation in the NDUFS4 gene encoding the 18 kDa (AQDQ) subunit of complex I abolishes assembly and activity of the complex in a patient with Leigh-like syndrome. Hum Mol Genet. 2001; 10:529–35. [PubMed: 11181577]
30. Anderson SL, et al. A novel mutation in NDUFS4 causes Leigh syndrome in an Ashkenazi Jewish family. J Inherit Metab Dis. 2008
31. van den Heuvel L, et al. Demonstration of a new pathogenic mutation in human complex I deficiency: a 5-bp duplication in the nuclear gene encoding the 18-kD (AQDQ) subunit. Am J Hum Genet. 1998; 62:262–8. [PubMed: 9463323]
32. Schuelke M, et al. Mutant NDUFV1 subunit of mitochondrial complex I causes leukodystrophy and myoclonic epilepsy. Nat Genet. 1999; 21:260–1. [PubMed: 10080174]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

33. Loeffen J, et al. The first nuclear-encoded complex I mutation in a patient with Leigh syndrome. Am J Hum Genet. 1998; 63:1598–608. [PubMed: 9837812]
34. McKenzie M, Lazarou M, Thorburn DR, Ryan MT. Mitochondrial respiratory chain supercomplexes are destabilized in Barth Syndrome patients. J Mol Biol. 2006; 361:462–9. [PubMed: 16857210]
35. Choi M, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. Proc Natl Acad Sci U S A. 2009; 106:19096–101. [PubMed: 19861545]
36. Ng SB, et al. Targeted capture and massively parallel sequencing of 12 human exomes. Nature. 2009; 461:272–6. [PubMed: 19684571]
37. Ng SB, et al. Exome sequencing identifies the cause of a mendelian disorder. Nat Genet. 42:30–5. [PubMed: 19915526]
38. Sheftel AD, et al. Human ind1, an iron-sulfur cluster assembly factor for respiratory complex I. Mol Cell Biol. 2009; 29:6059–73. [PubMed: 19752196]
39. Bych K, et al. The iron-sulphur protein Ind1 is required for effective complex I assembly. Embo J. 2008; 27:1736–46. [PubMed: 18497740]
40. Calvo S, et al. Systematic identification of human mitochondrial disease genes through integrative genomics. Nat Genet. 2006; 38:576–82. [PubMed: 16582907]
41. Tarpey PS, et al. A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation. Nat Genet. 2009; 41:535–43. [PubMed: 19377476]
42. Kirby DM, et al. Respiratory chain complex I deficiency: an underdiagnosed energy generation disorder. Neurology. 1999; 52:1255–64. [PubMed: 10214753]
43. Gnirke A, et al. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. Nat Biotechnol. 2009; 27:182–9. [PubMed: 19182786]
44. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008; 18:1851–8. [PubMed: 18714091]
45. Li H, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009; 25:2078–9. [PubMed: 19505943]
46. Calvo SE, Pagliarini DJ, Mootha VK. Upstream open reading frames cause widespread reduction of protein expression and are polymorphic among humans. Proc Natl Acad Sci U S A. 2009; 106:7507–12. [PubMed: 19372376]
47. Karolchik D, Hinrichs AS, Kent WJ. The UCSC Genome Browser. Curr Protoc Bioinformatics. 2009; Chapter 1(Unit1 4)
48. Dimmic MW, Sunyaev S, Bustamante CD. Inferring SNP function using evolutionary, structural, and computational methods. Pac Symp Biocomput. 2005:382–4. [PubMed: 15759643]
49. Cree LM, Samuels DC, Chinnery PF. The inheritance of pathogenic mitochondrial DNA mutations. Biochim Biophys Acta. 2009; 1792:1097–102. [PubMed: 19303927]
50. Gabriel S, Ziaugra L, Tabbaa D. SNP genotyping using the Sequenom MassARRAY iPLEX platform. Curr Protoc Hum Genet. 2009; Chapter 2(Unit 2):12. [PubMed: 19170031]
51. Lamande SR, et al. Reduced collagen VI causes Bethlem myopathy: a heterozygous COL6A1 nonsense mutation results in mRNA decay and functional haploinsufficiency. Hum Mol Genet. 1998; 7:981–9. [PubMed: 9580662]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

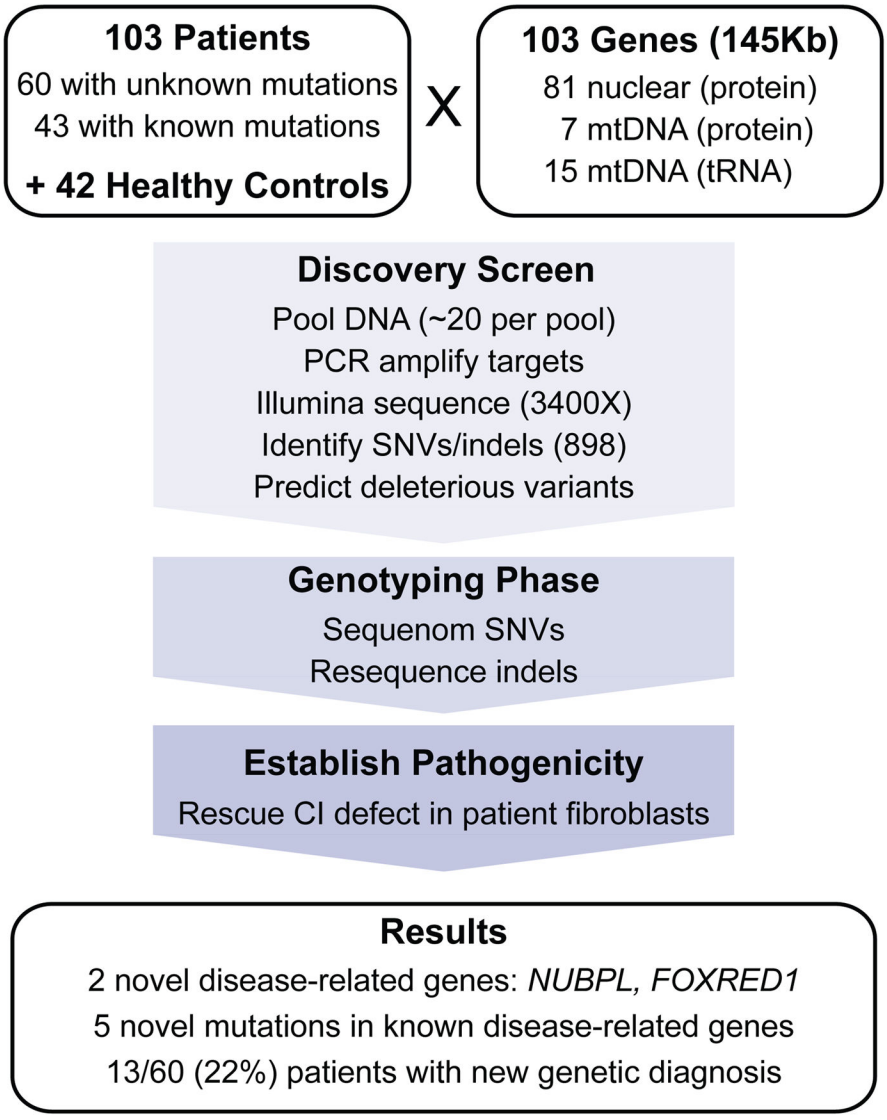

Figure 1.
Schematic overview of the Mito10K project.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

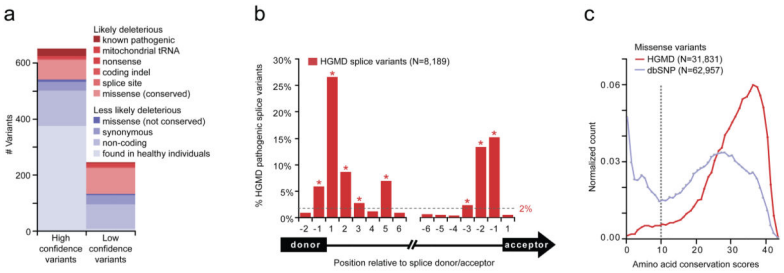

Figure 2.

Definition of ‘likely deleterious’ variants detected in pooled sequencing discovery screen. (a) Barplot of high-confidence and low-confidence variants, categorized by predicted deleterious consequences. (b) Histogram of known disease-associated splice variants, annotated in HGMD[24], by position relative to nearest splice donor and splice acceptor exons (black rectangles). Dashed line indicates frequency threshold and asterisk indicates splice positions considered ‘likely deleterious’. (c) Histogram of amino acid conservation score (# species with identical amino acid, out of 44 aligned vertebrate exons) shown for training data: missense variants annotated as disease-associated in HGMD (red curve) or present in dbSNP128 (blue curve). Dashed line indicates minimum conservation required for ‘likely deleterious’ variants.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

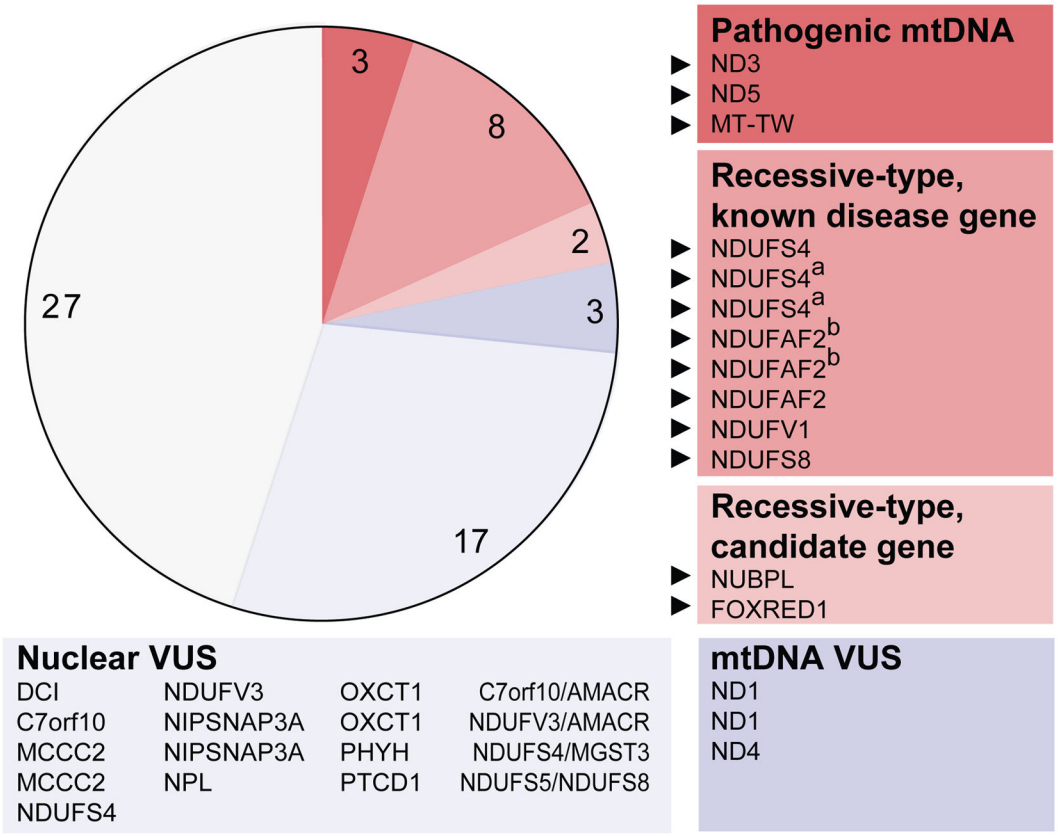

Figure 3.

60 patients with CI deficiency without a prior genetic diagnosis, categorized by type of 'likely deleterious' variants detected per gene. Red indicates patients with likely pathogenic variants, blue indicates patients with variants of uncertain significance (VUS), and gray indicates patients without 'likely deleterious' variants. Boxes list genes containing 'likely deleterious' variants in each patient. Black triangles indicate new experimentally established genetic diagnoses. [a,b] indicate pairs of affected siblings.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

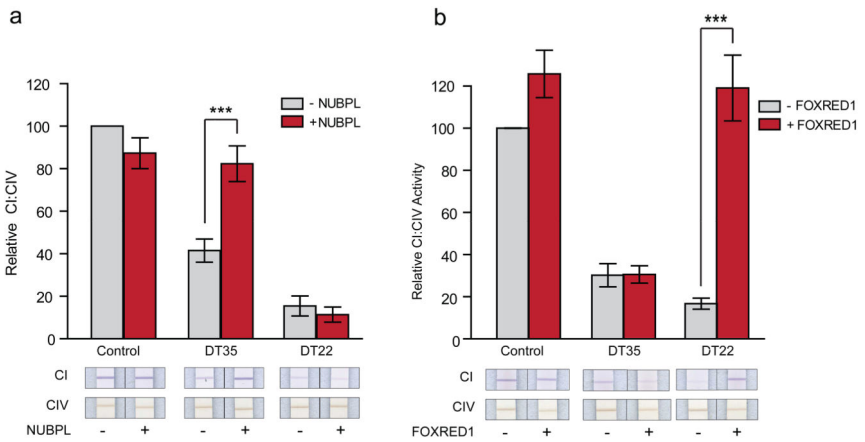

Figure 4.

NUBPL and *FOXRED1* cDNA rescue of CI defects in patient fibroblasts. Barplots show CI activity, normalized by CIV activity, measured in control and patient fibroblasts, before and after transduction with wild-type *NUBPL-V5* mRNA (a) or wild-type *FOXRED1-V5* mRNA (b). Bars show mean of 3 biological replicates, and error bars indicate ±1 s.e.m. Asterisks indicate $p<0.01$. Representative dipstick assays shown below.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

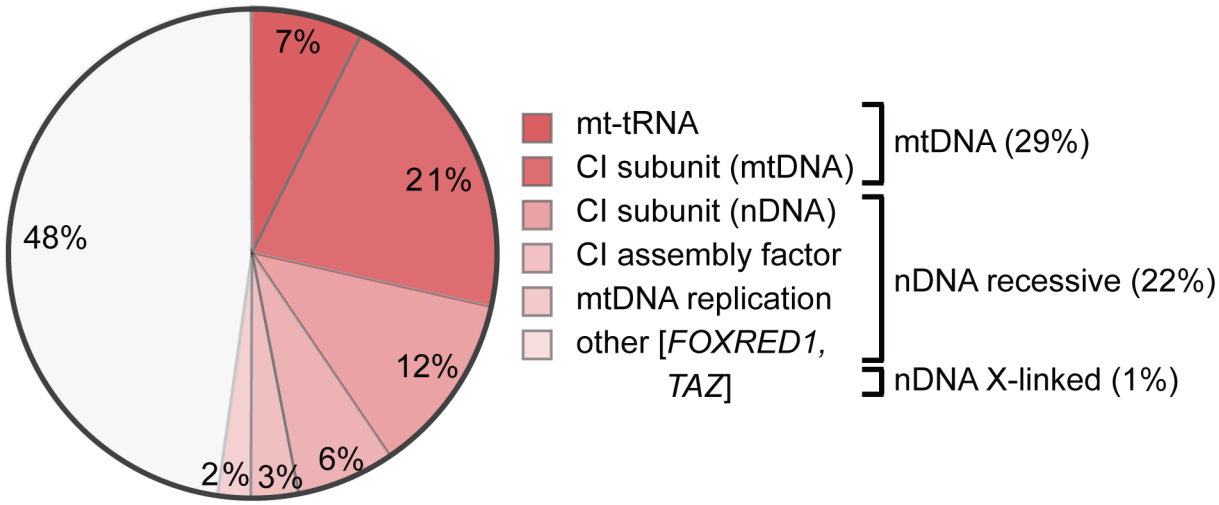

Figure 5.

Genetic diagnosis of 94 unrelated patients with definite, isolated complex I deficiency grouped by function of underlying gene. Red indicates patients with confirmed genetic diagnosis, and gray indicates absence of genetic diagnosis. Patients are representative cohort, selected as all unrelated individuals within the 103 patients sequenced.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1

Clinical and other features of patient cohort

	Patients with:		
Clinical Diagnosis	**mtDNA mutations**	**nuclear mutations**	**unknown mutations**
Leigh Syndrome	11	6	15
Other mitochondrial encephalopathy	3	1	13
Cardiomyopathy/encephalopathy	0	2	12
LIMD	2	6[a]	9
MELAS	6	0	0
Mitochondrial myopathy	2	0	5
Mitochondrial cytopathy	1	0	3
Mitochondrial hepatopathy	0	3	2
VCFS/DiGeorge Plus	0	0	1
Total	**25**	**18**	**60**
Consanguinity	0	7	6
Family history[b]: definite, possible	7, 9	9, 0	9, 9
Fibroblast defect[c] (# tested)	17 (20)	10 (15)	18 (32)

[a] 2 patients were affected prenatal diagnoses that were terminated and diagnosis was assumed to be the same as the proband.

[b] Family history consistent with a mitochondrial disorder

[c] CI enzyme defect present in patient fibroblasts

Abbreviations: LIMD, Lethal Infantile Mitochondrial Disease; MELAS, Mitochondrial Encephalopathy, Lactic Acidosis, Stroke-like episodes; VCFS, Velo-Cardio-Facial Syndrome;

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Nat Genet. Author manuscript; available in PMC 2011 April 01.

Table 2

Number of variants detected in pooled sequencing discovery screen.

Variant type	High Confidence Variant Calls			Low Confidence Variant Calls		
	Detected in Patients	Likely Deleterious	Validated	Detected in Patients	Likely Deleterious	Validated
nDNA						
nonsense	3	2	1	5	5	1
missense	131	60	51	97	86	9
splice	78	28	22	40	16	2
synonymous	92	0	0	33	0	0
UTR	214	0	0	71	0	0
coding indels	3	3	3	0	0	0
mtDNA						
nonsense	0	0	0	0	0	0
missense	37	14	12	0	0	0
synonymous	85	0	0	0	0	0
noncoding	9	2	2	0	0	0
Total	**652**	**109**	**91**	**246**	**107**	**12**

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 3

New genetic diagnoses for 13 patients with CI deficiency

Patient	Clinical diagnosis	Genetic diagnosis	Homozygous variants	Heterozygous variants	Supporting Evidence
DT58	Mt enc	firm (ND3 het.)		**ND3:m.10197G>A,p.A47T**	Known disease variant[25], ~90% mutant load in blood
DT55	LS	firm (ND5 het.)		**ND5:m.13094T>C,p.V253A,** C2orf56:c.208C>G,p.P70A	Known disease variant[26], ~60% mutant load in muscle
DT20	LIMD	firm (MT-TW hom.)	**MT-TW:m.5567T>C,** ND2:m.4890A>G,p.I141V, ND5:m.13676A>G,p.N447S	TMEM22:c.500G>A,p.R167Q	Known disease variant[26], 100% homoplasmic in blood, muscle, liver and fibroblasts
DT37[a]	LS	firm (NDUFS4 cmpd het.)	DCI:c.392T>C,p.L131P	**NDUFS4:c.462delA,p.K154NfsX34, NDUFS4:c.99-1G>A,p.S34IfsX4,** NDUFS2:c.96-3C>T, GAD1:c.990A>T,p.E330D	Known disease variants[10,30], reseq, splice
DT38[a]	LS	firm (NDUFS4 cmpd het.)		**NDUFS4:c.462delA,p.K154NfsX34, NDUFS4:c.99-1G>A,p.S34IfsX4,** GAD1:c.990A>T,p.E330D DCI:c.392T>C,p.L131P	Known disease variants[10,30], reseq, splice, NDP
DT107	LS	firm (NDUFS4 cmpd het.*)		**NDUFS4:c.351-2A>G*, NDUFS4:c.99-1G>A,p.S34IfsX4**	Known disease variant[10], seg, reseq, splice, conservation, NDP
DT67[b]	LS	firm (NDUFAF2 hom.*)	**NDUFAF2:c.103delA,p.I35SfsX17***	GPAM:c.1340C>T,p.T447M	NDP, reseq, splice, conservation
DT68[b]	LS	firm (NDUFAF2 hom.*)	**NDUFAF2:c.103delA,p.I35SfsX17***	GPAM:c.1340C>T,p.T447M	NDP, reseq, splice, conservation
DT16	LS	firm (NDUFA2 hom.*)	**NDUFAF2:c.221G>A,p.W74X***		NDP, 250K SNP, reseq, splice
DT3	LIMD	probable (NDUFV1 hom.*)	**NDUFV1:c.1129G>A,p.E377K***	C20orf7:c.412G>A,p.V138I	250K SNP, reseq, conserv. in NADH 4Fe- 4S domain
DT61	Mt enc	probable (NDUFS8 hom.*)	**NDUFS8:c.460G>A,p.G154S***	NDUFV3:c.826G>A,p.E276K	seg, reseq, conservation in Fer4 domain
DT35	Mt enc	firm (NUBPL cmpd het.*)		**[NUBPL:c.166G>A,p.G56R*+ NUBPL:c.815-27T>C,p.D273QfsX31*], [chr14:g.(30,932,976_30,953,766)_ (31,193,278_31,194,846)del*+ chr14:g.(31,211,800_31,212,780)_ (31,345,080_31,350,225)dup*],** NDUFB9:c.290A>G,p.Y97C	Rescue, reseq, conservation, splice
DT22	LS	firm (FOXRED1 cmpd het.*)		**FOXRED1:c.694C>T,p.Q232X*, FOXRED1:c.1289A>G,p.N430S*,** NIPSNAP1:c.215A>G,p.Y72C	Rescue, reseq, conservation, splice

[a,b] affected sibling pairs

* novel variant, not previously reported

Bold indicates likely causal variants.

Abbreviations: Mt enc, mitochondrial encephalopathy; LS, Leigh Syndrome; LIMD, lethal infantile mitochondrial disease; hom., homozygous/homoplasmic; het., heterozygous/heteroplasmic, cmpd het., compound heterozygous; Rescue, pathogenicity confirmed by rescue of CI defect in patient fibroblasts; NDP, no detectable protein, by SDS-PAGE and western blot; Seg, variant segregates with disease in family; Reseq, variant confirmed by Sanger sequencing of genomic DNA; Splice, splicing defect observed in patient fibroblast cDNA +/−CHX; Conservation, amino acid conserved in ≥30/44 vertebrate species; 250K SNP, region of homozygosity from Affymetrix 250K *Nsp* SNP chip.