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

Sarah E Calvo1–3,10, Elena J Tucker4,5,10, Alison G Compton4,10, Denise M Kirby4, Gabriel Crawford3, Noel P Burtt3, Manuel Rivas1,3, Candace Guiducci3, Damien I Bruno4, Olga A Goldberger1,2, Michelle C Redman3, Esko Wiltshire6,7, Callum J Wilson8, David Altshuler1,3,9, Stacey B Gabriel3, Mark J Daly1,3, David R Thorburn4,5 & Vamsi K Mootha1–3

Discovering the molecular basis of mitochondrial respiratory chain disease is challenging given the large number of both mitochondrial and nuclear genes that are 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 disorders. We created seven pools of DNA from a cohort of 103 cases and 42 healthy controls and then performed deep sequencing of 103 candidate genes to identify 151 rare variants that were predicted to affect protein function. We established genetic diagnoses in 13 of 60 previously unsolved cases using confirmatory experiments, including cDNA complementation to show that mutations in NUBPL and FOXRED1 can cause complex I deficiency. Our study illustrates how large-scale sequencing, coupled with functional prediction and experimental validation, can be used to identify causal mutations in individual cases.

Complex I of the mitochondrial respiratory chain is a large ~1-MDa macromolecular machine composed of 45 protein subunits encoded by both the nuclear and mitochondrial (mtDNA) genomes. Complex I 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 complex I activity are the most common type of human respiratory chain disease, which collectively has an incidence of 1 in 5,000 live births5. Complex I 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 acidosis3–4. The diagnosis of complex I deficiency is challenging given its clinical and genetic heterogeneity and usually relies on biochemical assessment of biopsy material5,6. Estimates suggest that roughly 15–20% of isolated complex I deficiency cases are due to mutations in the mtDNA, and the rest are probably caused by nuclear defects7,8, though most of these mutations remain unknown.

Twenty-five genes underlying human complex I deficiency have been identified by candidate gene sequencing, linkage analysis or homozygosity mapping. These include 19 subunits of the complex (7 mtDNA genes and 12 nuclear genes) and 6 nuclear-encoded assembly factors that are required for the proper assembly, stability or maturation of complex I (Supplementary Table 1). Many more assembly factors are probably required, as suggested by the 20 factors necessary for assembly of the smaller complex IV9 and by cohort studies that estimate that only half of individuals with complex I deficiency have mutations in known genes10–13.

Additional proteins that are required for complex I activity are likely to reside in the mitochondrial and aid in the assembly and regulation of complex I. To systematically predict such proteins, we combined the MitoCarta inventory of mitochondrial proteins14 with functional prediction through phylogenetic profiling15,16. Phylogenetic profiling was previously used to identify the complex I assembly factor NDUFAF217. We generalized this method to identify 34 additional candidate genes14, three of which have been shown to harbor mutations causing inherited forms of complex I deficiency14,18,19. The remaining predictions, combined with the known complex I structural subunits and assembly factors, comprise a targeted set of 103 candidate genes for human complex I deficiency (Supplementary Table 1).

Recent technological advances20 offer the prospect of sequencing all 103 candidate genes in a cohort of individuals with clinical and biochemical evidence of complex I 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. 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 these regions to high depth, and detected new variants in each pool (Fig. 1). We then used genotyping technology to type these newly discovered variants, as well as previously reported pathogenic mutations, in all subjects. Finally, we confirmed the pathogenicity of prioritized variants using molecular approaches including cDNA rescue in subject fibroblasts.

Here, we report the results of our project, which we term ‘Mito10K’ to reflect the 103 candidate genes sequenced in 103 individuals with complex I deficiency.

RESULTS

Rare variant discovery by pooled sequencing

Our cohort of 103 cases had ‘definite’, isolated complex I deficiency shown by biochemical assessment. The cohort included 60 individuals who lacked a previous molecular diagnosis as well as 43 controls with established molecular diagnoses (Table 1 and Supplementary Table 2). We also sequenced 42 healthy control subjects from the European HapMap collection. We combined DNA into 5 pools from cases and 2 pools from 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 sheared to construct libraries. The seven libraries were sequenced using a single Illumina Genome Analyzer flowcell, with one pool per lane (see Online 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 ≥100× coverage and achieved 3,359× median coverage per pool, corresponding to an average of 168× per individual (Supplementary Fig. 1). Around 10% of nuclear target regions were poorly covered, largely owing to skewed GC content (Supplementary Fig. 1). The mtDNA target regions showed substantially higher coverage (10,144× median coverage).

However, the mtDNA in the pooled samples was not uniformly distributed across subjects, primarily owing to biases introduced by whole-genome amplification (Supplementary Fig. 2). In one pool, for example, 96% of the mtDNA came from a single individual. Nonetheless, the deep coverage of mtDNA allowed us to discover variants 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. The estimated 1% error rate of individual Illumina reads makes it difficult to detect alleles present in 1:40 chromosomes. Therefore, we applied a method called Syzygy to empirically estimate error rates at each base in order to confidently identify rare variants (M.J.D. and M.R., personal communication, and Supplementary Note). Using this method, we detected 652 high-confidence variants in the case pools (Table 2). To improve sensitivity, we applied an ad hoc approach to identify 246 low-confidence variants supported by at least 3 reads on each strand (Table 2). We identified 898 high- and low-confidence variants.

Next, we assessed the accuracy of these 898 variants using known genotypes available from our case and HapMap controls. Overall, we achieved 92% sensitivity and 99.6% specificity for control SNVs at nuclear DNA sites with ≥100 reads (see Online Methods and Supplementary Table 3). This high sensitivity is due to the deep sequence coverage and the relatively high allele frequency for many HapMap control variants (Supplementary Fig. 3). However, as expected, we achieved lower sensitivity for rare 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 case controls (32%) owing to the nonuniform distribution of mtDNA within each pool. The minor allele frequencies estimated from read counts correlated strongly with expected frequencies in HapMap pools (R² = 0.96), indicating that the pooled sequencing protocol had high fidelity (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 (Fig. 2a). Briefly, we filtered out: (i) variants that were filtered out by the Syzygy approach; (ii) common variants that are not pathogenic; (iii) very rare variants that might not be present in our HapMap controls; and (iv) variants that do not segregate with the phenotype. The remaining variants were prioritized based on their frequency in our cohort and their known role in complex I deficiency.

Table 1: Clinical, molecular and biochemical features of the cohort

<table>
<thead>
<tr>
<th>Clinical diagnosis</th>
<th>mtDNA mutations</th>
<th>Nuclear mutations</th>
<th>Unknown mutations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leigh syndrome</td>
<td>11</td>
<td>6</td>
<td>15</td>
</tr>
<tr>
<td>Other mitochondrial encephalopathy</td>
<td>3</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>Cardiomyopathy/encephalopathy</td>
<td>0</td>
<td>2</td>
<td>12</td>
</tr>
<tr>
<td>LIMD</td>
<td>2</td>
<td>6*</td>
<td>9</td>
</tr>
<tr>
<td>MELAS</td>
<td>6</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Mitochondrial myopathy</td>
<td>2</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>Mitochondrial cytopathy</td>
<td>1</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Mitochondrial hepatopathy</td>
<td>0</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>VCFS/George plus</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Total</td>
<td>25</td>
<td>18</td>
<td>60</td>
</tr>
</tbody>
</table>

Note: LIMD, lethal infantile mitochondrial disease; MELAS, mitochondrial encephalopathy, lactic acidosis, stroke-like episodes, VCFS, velo-cardio-facial syndrome.

*Two subjects represent prenatal diagnoses that were terminated and diagnosis was assumed to be the same as the proband.

**Family history consistent with a mitochondrial disorder.

I The reversion of mtDNA defect present in subject fibroblasts.

Figure 1: Schematic overview of the Mito10K project.
were present in healthy individuals, based on HapMap controls, dbSNP\textsuperscript{23}, mtDB\textsuperscript{24} and pilot data from the 1,000 genomes project; (ii) synonymous variants; and (iii) non-coding variants, unless they corresponded to tRNA or splice sites. We selected 8 splice site positions using training data of 8,189 disease-associated splice variants in the Human Gene Mutation Database (HGMD)\textsuperscript{24} (Fig. 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 (Fig. 2c; see Online Methods). 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 complex I controls. The approach missed 4 nuclear and 17 mtDNA variants in the discovery screen, and filtered out 1 nuclear splice variant located 4 bp into an intron and 1 mtDNA missense variant at a poorly conserved site (Supplementary Table 2).

Genotyping rare variants
Our next goal was to genotype the discovered ‘likely deleterious’ variants, as well as previously known disease variants, in each case sample (Supplementary Table 4 and see Online Methods). The genotyping served multiple purposes. First, it was necessary to validate newly identified variants from the pooled discovery screen. Second, it enabled us to search for known mutations underlying complex I deficiency that were not detected in our discovery screen owing to a lack of power (for example, mtDNA variants). Third, it allowed us to assign the variants to individuals.

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 validation rate, based on 101 additional high-confidence variants genotyped (Supplementary Table 4). We further validated SNVs of particular interest using Sanger sequencing, as Sequenom genotypes showed an estimated 11% false positive rate for extremely rare variants (Supplementary Note). In a subset of instances in which we identified heterozygous variants of interest, we used Sanger sequencing to fully resequence the gene.

In total, we validated 151 ‘likely deleterious’ 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 higher frequency of ‘likely deleterious’ variants in our cases than in European controls, although this enrichment might be due to differences in ancestry (Supplementary Note).

Prioritizing variants for complex I deficiency
With the Mito10K sequence data in hand, we next searched our 60 undiagnosed cases for individuals harboring either known pathogenic mtDNA mutations or two mutant alleles in the same
nuclear gene (Fig. 3). We refer to the latter as ‘recessive-type’ variants, which include homoygous and compound heterozygous variants, consistent with a recessive mode of inheritance. Of course, compound heterozygosity can only be ascertained after confirmatory phasing.

Only three subjects had previously reported pathogenic mtDNA mutations and only eight subjects had recessive-type mutations in known disease genes, including five undescribed and two previously reported mutations (Table 3). Two subjects had recessive-type mutations in candidate disease genes (NUBPL and FOXRED1; Table 3). The remaining subjects included three individuals with ‘likely deleterious’ mtDNA 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 genetic diagnoses in known disease genes

We next assessed the pathogenicity of variants detected in the three individuals with causal mtDNA mutations (in ND323, ND526 and MT-TW20) and the eight individuals with recessive-type variants in previously reported complex I disease genes: NDUF5410,27–31, NDUF427,27, NDUFV132 and NDUF833 (Table 3). The discovered mutations were absent from all other cases and HapMap controls sequenced, except as noted below.

We identified one undescribed and two previously reported NDUF54 mutations in three individuals with Leigh syndrome (Table 3 and Supplementary Fig. 4). Two siblings, DT37 and DT38, were compound heterozygous for the reported mutations c.462delA (p.Lys154AsnfsX34)30 and c.99-1G>A (p.Ser34IlefsX4)10. The unrelated individual DT107 was compound heterozygous for the same c.99-1G>A mutation and a new mutation c.351-2A>G, which were 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 NDUF54 splicing. The heterozygous c.351-2A>G mutation was detected in genomic DNA from DT107, however, it was undetectable in cDNA with or without cycloheximide (CHX), suggesting that the mRNA was unstable. Protein blot analysis on fibroblasts from individuals DT38 and DT107 showed no detectable NDUF54 protein. This is the second report of the c.99-1G>A mutation10 and the third of the c.462delA mutation28,30, suggesting not only that recurrent mutations in NDUF54 underlie Leigh syndrome but also that several previously unrecognized founder mutations may exist in this gene.

We also identified new homozygous mutations in NDUF54 in three individuals with Leigh syndrome (Table 3 and Supplementary Fig. 5). A consanguineous individual, DT16, harbored a homozygous c.221G>A mutation (p.Trp74X) within a 6.3-Mb region of homozygosity (determined by Affymetrix 250K Nsp SNP chip). Two siblings, DT67 and DT68, harbored a homozygous c.103delA mutation (p.Ile35SerfsX17). Analysis of cDNA from subject fibroblasts showed that NDUF54 transcripts containing these mutations were stable. In addition, the c.221G>A nonsense mutation in DT16 (located 4 bp into exon 3) resulted in occasional exon 3 skipping, which generates a transcript that also encodes a truncated protein (p.Ala73GlyfsX5). All three subjects lacked any detectable NDUF54 protein by protein blot analysis, which indicates that the truncated protein products are unstable.

We identified a previously undescribed homozygous NDUF54 mutation (c.1129G>A, p.Glu377Lys) in a 2.1-Mb region of homozygosity (determined by Affymetrix 250K Nsp SNP chip) in a consanguineous Lebanese individual, 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 in the consensus motif for the iron sulfur binding site (pam10589), which is highly conserved across eukaryotic species.

We identified a new homozygous NDUF548 mutation (c.460G>A, p.Gly154Ser) in a Sudanese subject, DT61, who presented with mitochondrial encephalomyopathy (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 (pam00037). This mutation segregated with disease in this family: an affected sibling was also homozygous whereas both unaffected parents were heterozygous carriers.

NUBPL and FOXRED1 in complex I deficiency

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

Subject DT35 presented with mitochondrial encephalomyopathy and was found to carry an apparent homozygous c.166G>A mutation in NUBPL (Supplementary Fig. 8). We did not detect this mutation in the 204 other subject chromosomes or the 84 HapMap control chromosomes sequenced. This mutation is predicted to cause substitution of a highly conserved glycine residue with arginine (p.Gly56Arg), 18 amino acids from the mitochondrial targeting sequence cleavage site predicted by TargetP (Supplementary Fig. 8). Although the subject's father was heterozygous for this mutation, the mother did not carry the mutation (Supplementary Fig. 8). To determine whether the mother could have transmitted a deletion involving this portion of exon 2, we performed Affymetrix array-based cytogenetic analysis on DNA from individual DT35. We detected a complex chromosomal rearrangement including a ~240-Kb deletion spanning exons 1–4 of NUBPL and a ~130-Kb duplication involving exon 7 of NUBPL (Supplementary Fig. 8). Next, we assessed NUBPL mRNA species
in individual DT35. RT-PCR showed very low expression of the full-length transcript, and 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 that 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). We found a c.815–27T>C mutation that is predicted to ablate a consensus branch sequence. This mutation was present in 2 out of 232 control chromosomes from individuals of European ancestry. Thus, DT35 contains one NUBPL allele harboring a deletion that spans exons 1–4 and a second allele that harbors both a p.Gly56Arg missense mutation and a c.815–27T>C mutation that probably causes exon 10 skipping.

We performed a complementation experiment to assess whether the introduction of wild-type cDNA into subject fibroblasts rescued the defect in complex I activity. Fibroblasts from this individual showed a strong complex I defect, with only 19% residual complex I activity when assayed by spectrophotometric enzyme assay and 40% residual complex I activity when assayed by dipstick enzyme assay. Using a lentiviral expression system, we transduced subject fibroblasts with wild-type cDNA. Expression of wild-type NUBPL rescued complex I activity in fibroblasts from subject DT35 but not from subject DT22 who harbored FOXRED1 mutations (Fig. 4a), establishing NUBPL as the causal gene in this case.

Although we have shown that NUBPL underlies complex I deficiency in this subject, we have not established the pathogenicity of individual mutations. Owing to its prevalence in controls, the c.815–27T>C branch site mutation may be a pseudo-deficiency allele, that is, it confers increased fitness when inherited with a null allele, as in DT35. Alternatively, the p.Gly56Arg missense mutation might abolish NUBPL function or act in synergy with the branch-site mutation to cause disease.

Subject DT22 presented with Leigh syndrome and was found to be compound heterozygous for two mutations in FOXRED1, c.694C>T (p.Gln232X) and c.1289A>G (p.Asn430Ser) (Supplementary Fig. 9). The c.694C>T mutation was detected in the discovery screen and was not detected in 204 other case 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.

### Table 3 New genetic diagnoses for cases of complex I deficiency

<table>
<thead>
<tr>
<th>Subject</th>
<th>Clinical diagnosis</th>
<th>Genetic diagnosis</th>
<th>Homozygous variants</th>
<th>Heterozygous variants</th>
<th>Supporting evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>DT58</td>
<td>Mt. enc.</td>
<td>Firm (ND3 het.)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT55</td>
<td>LS</td>
<td>Firm (ND5 het.)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT20</td>
<td>LIMD</td>
<td>MT-TW hom.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT38</td>
<td>LS</td>
<td>Firm (NDUFS4 cmpd. het.)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT107</td>
<td>LS</td>
<td>Firm (NDUFS4 cmpd. het.)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT68</td>
<td>LS</td>
<td>Firm (NDUFS4 hom.)a</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT16</td>
<td>LS</td>
<td>Firm (NDUFS4 hom.)a</td>
<td>NDUF2c.221G&gt;A, p.Trp74Xa</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT3</td>
<td>LIMD</td>
<td>Probable (NUFV1 hom.)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT61</td>
<td>Mt. enc.</td>
<td>Probable (NDUFS8 hom.)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT35</td>
<td>Mt. enc.</td>
<td>Firm (NUBPL cmpd. het.)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DT22</td>
<td>LS</td>
<td>Firm (FOXRED1 cmpd. het.)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Bold indicates likely causal variants. Mt. enc., mitochondrial encephalopathy; LS, Leigh syndrome; LIMD, lethal infantile mitochondrial disease; hom., homozygous/homoplasmic; heter., heterozygous/heteroplasmic; cmpd. het., compound heterozygous; rescue, pathogenicity confirmed by rescue of complex I defect in subject fibroblasts; NDP, no detectable protein, by SDS-PAGE and protein blot; seg., variant segregates with disease in family; reseq., variant confirmed by Sanger sequencing of genomic DNA; splice, splicing defect observed in subject fibroblast cDNA with or without CHX; conservation, amino acid conserved in ≥30/44 vertebrate species; 250K SNP, region of homozygosity from Affymetrix 250K Nsp SNP chip.

a,bAffected sibling pairs. cNovel variant, not previously reported.
Paternal DNA was not – + Supplementary Fig. 9 mtDNA (29% b.

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 showed that both mutations were present. However, in the absence of CHX the transcript containing the c.694C>T (p.Gln232X) mutation 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 subject’s mother, and is predicted to cause the substitution of a highly conserved asparagine residue with a serine (p.Asn430Ser; Supplementary Fig. 9). Paternal DNA was not available for genotyping. RT-PCR analysis of subject cDNA also shows occasional skipping of exon 6 (containing c.694C>T), which results in a transcript that is predicted to lack 40 internal residues (Supplementary Fig. 9).

As above, we performed a complementation experiment in subject fibroblasts to assess the role of FOXRED1 in complex I activity. Fibroblasts from this subject show a striking complex I defect, with only 9% residual complex I activity when assayed by spectrophotometric enzyme assay and 15% residual complex I activity when assayed by dip-stick 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 (Fig. 4b).

Together, the mutation data and complementation experiments provide evidence that NUBPL and FOXRED1 are bona fide complex I disease-related genes in individuals DT35 and DT22, respectively.

Mutational spectrum of complex I deficiency

The large-scale discovery and validation studies for 60 cases reported here, in addition to the previous molecular diagnosis of all 43 other individuals with definite isolated complex I deficiency seen at our diagnostic laboratory, provide the largest systematic sequencing study of complex I deficiency to date. Our cohort of 103 subjects 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%; Fig. 5). Of these mutations, 33% are in complex I structural subunits, 6% are in established complex I assembly factors (including NUBPL), 7% are tRNA mutations required for mtDNA translation, 4% are in other auxiliary factors (mtDNA replication proteins POLG and C10orf2, and the TAZ protein required for complex I stability by maintaining cardiolipin pools within the mitochondrial inner membrane)34, and 1% are in an uncharacterized gene (FOXRED1). In total, the previous and new genetic diagnoses in our cohort correspond to 47 unique mutations in 20 genes, highlighting the allelic and locus heterogeneity of complex I deficiency.

DISCUSSION

Advances in genome sequencing technology offer a new opportunity to solve the genetic basis of disease even in individual cases. Perhaps the most important challenge of human genetics moving forward will be to distinguish 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 variants35,36. Several recent whole-exome sequencing projects have detected causal mutations for Mendelian disease by using multiple affected individuals to hone in on regions of interest, and have established pathogenicity by identifying different mutations in these regions in unrelated individuals with the same phenotype36,37. Although this approach has broad utility, it may not be readily applicable to individual, sporadic cases of disease.

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

Our approach successfully identified pathogenic roles for NUBPL and FOXRED1. NUBPL, also known as IND1, is an assembly factor for complex I38. Similar to its role in the yeast Yarrowia lipolytica, human NUBPL is essential for the incorporation of Fe/S clusters into complex I subunits, and its knockdown causes improper assembly of the peripheral arm of complex I, decreased complex I activity and abnormal mitochondrial morphology38,39. We now report NUBPL mutations in an individual with complex I deficiency, a male who presented at 2 years of age with developmental delay, leukodystrophy and elevated CSF lactate (see Supplementary Note for a complete clinical description). Muscle biopsy and skin fibroblasts showed marked complex I deficiency (37% and 19% normalized activity, respectively, relative to controls). Sequencing of DNA from this individual revealed an apparent homozygous p.Gly56Arg missense mutation in NUBPL in an amino acid that has been conserved across all 36 aligned vertebrate species. However, further analysis indicated that this individual was
compound heterozygous: one allele contained both the p.Gly56Arg 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 individual highlights the limitations of second-generation pooled sequencing. Large deletions are not detected and variants such as branch site mutations may be missed or overlooked. Nevertheless, the complex I defect in fibroblasts was rescued by expression of a wild-type allele of NUBPL, thereby establishing a pathogenic role for NUBPL mutations in complex I 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 on the basis of its mitochondrial localization and shared phylogenetic profile with complex I subunits. 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 a complete clinical description). Muscle biopsy and fibroblasts showed severe complex I deficiency (9% of normal control mean in both samples relative to citrate synthase). Sequencing samples from this subject revealed compound heterozygous FOXRED1 mutations: a p.Gln232X nonsense mutation and a p.Asn430Ser missense mutation in a conserved acidic amino acid. As with NUBPL above, cDNA rescue established FOXRED1 as a disease-related gene. The function of FOXRED1 is not clear, although its four human homologs (DMGDH, SARDH, PIPOX and PDDR) perform redox reactions in amino acid catabolism, suggesting a potential link between amino acid metabolism and complex I.

Although the Mito10K project successfully identified or confirmed pathogenic mutations in half of the 103 subjects with complex I deficiency (Fig. 5), 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. Although in some of the undiagnosed complex I individuals we detected ‘likely deleterious’ variants that may contribute to pathogenesis, most carry 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 owing to lack of sensitivity, especially in 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. It is also possible that in some individuals, the disease is caused by complex inheritance or epigenetic mechanisms. Broader sequencing, combined with functional validation, will be required to fully elucidate the molecular bases of these remaining cases.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturegenetics/.

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 NDUFA2 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 O.A.G. for the lentiviral expression vector, S. Flynn for assistance with computational analyses, M. Cabili for tool evaluation, J. Flannick for human subjects protocols, R. Onofrio for designing PCR primers, K. Ardlie and S.B.G. for assistance with evolutionary conservation analyses, J. Pirrucchello, R. Do and S. Kathiresan for data and analysis of control data, and the many physicians who referred subjects and assisted with these studies. This work was supported by a grant (456901) and Principal Research Fellowship from the Australian National Health & Medical Research Council awarded to D.R.T., an Australian Postgraduate Award to E.J.T. and a grant from the US National Institutes of Health (GM077465) to V.K.M. 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.

AUTHOR CONTRIBUTIONS

This study was conceived and designed by S.E.C., D.R.T. and V.K.M. with input from M.J.D. and S.B.G. Enzyme diagnosis of the cohort was coordinated by D.M.K. E.W. and C.J.W. provided clinical interaction and assisted with sample collection. Samples were collected by D.M.K., E.W. and C.J.W. and prepared by A.G.C. and E.J.T. The pooled sequencing protocol was designed and established at the Broad Institute by D.A., M.J.D. and S.B.G. Project management was performed by S.E.C., N.P.B. and C.G. G.C. performed pooling. M.C.R. and C.G. performed the genotyping. S.E.C. designed and performed the computational analyses, with assistance from E.J.T., A.G.C. and M.R. All experiments were designed and performed by E.J.T., A.G.C. and O.A.G. Affymetrix array-based cyto genetic analysis was performed by D.L.B. Syzygy was developed and run by M.R. and M.J.D. The manuscript was written by S.E.C., E.J.T., A.G.C., D.R.T. and V.K.M.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturegenetics/

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

ARTICLES

ONLINE METHODS

Complex I deficiency cases. The 60 cases plus 43 case controls had a definite diagnosis of isolated complex I deficiency, based on spectrophotometric enzyme assays interpreted by published criteria42,43. Briefly, the ratio of complex I 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 was required to be at least twofold higher than that of complex I (Supplementary Fig. 10). The cohort includes all such individuals diagnosed in Melbourne from 1992 to 2007, with the exception of nine individuals 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 subject tissues (skeletal or cardiac muscle and liver) by proteinase K digestion followed by salting-out. Each subject sample was whole-genome amplified using a QiAGEN REPLi-g Kit with 100 ng 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 20 ng μl⁻¹ based on two rounds of quantification and dilution, yielding a mean concentration of 19.2 ng μl⁻¹ (1.56 s.d.). We allowed for 10% variance as that is the accuracy limit of PicoGreen quantification. 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. Twenty or twenty-one samples were then pooled in equimolar amounts. Each case pool contained individuals with unknown diagnoses, known mtDNA mutations, and known nuclear mutations, with the following counts: Pool 1 = 12, 5, 4; pool 2 = 13, 5, 3; pool 3 = 12, 5, 4; pool 4 = 12, 5, 3; pool 5 = 11, 5, 4. See Supplementary Note for HapMap sample identifiers.

Target selection. Targets included 2 mtDNA regions and coding and UTR SNVs were assayed in whole-genome-amplified DNA from HapMap (vii) missense variant at an amino acid conserved in ≥10 aligned vertebrate species, based on the multiz44way genome alignments downloaded from UCSC genome browser47 (see Supplementary Note), or predicted as ‘damaging’ by PolyPhen-2.0 (Human Var training data)48 (see Supplementary Note). Variant sets that were not previously associated with disease were excluded if present in 42 HapMap controls, dbsNP127, 1,000 genomes pilot 1, or present at >0.005 minor allele frequency in mtDB49 based on the frequency of asymptomatic carriers of pathogenic mtDNA mutations45. 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 in whole-genome-amplified DNA from the 103 individuals with complex I deficiency using Sequenom MassARRAY iPLEX GOLD chemistry50. Oligos were synthesized and mass-spec QCed at Integrated DNA Technologies. All SNVs were genotyped in multiplexed pools of 20–38 assays, designed by AssayDesigner v3.1 software, starting with 10 ng of DNA per pool. Around 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 (Brucker Daltonics Inc.). We obtained high quality data (≥95% genotype call rate, HWE P > 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 v4.0 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) according to the 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 TRC70 (V5 C-terminal tag) by Gateway cloning (Invitrogen). Initial experiments using this vector did not rescue complex I 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. The full-length NUBPL ORF was amplified from MCH58 cells by RT-PCR incorporating Gateway adaptors, then was cloned into pLEX TRC970 (V5 C-terminal tag) by Gateway cloning to generate the NUBPL-V5 pDest vector.

Viral particle production and transduction. HEK-293T cells were grown on 10-cm plates to 60% confluence and cotransfected with a packaging plasmid (pCMV-88.91), 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 medium was applied to the cells 16 h after transfection and, after 24 h incubation, supernatants containing packaged virus were harvested and filtered through a 0.45-μm membrane filter. Subject 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 a final concentration of 5 μg ml⁻¹ in 8.75 ml total medium. Plates were spun at 2,500 r.p.m. for 90 min and incubated for 24 h at 37 °C before medium was replaced. Cells were grown in antibiotic-free medium for 30 h.
Supplementary Table 5

Antibodies included NDUFS4 (MS104, -1) were designed to

Primary control and case fibroblasts were

Genome-wide microarray analysis doi:10.1038/ng.659

Complex I and complex IV dipstick activity

provides detailed data on all vali

The Sequence Alignment/Map format and SAMtools.

Temperature for 1 h and developed using ECL or ECL Plus detection reagents

rabbit

0.05% Tween-20) and incubated with primary antibodies overnight at

membranes (Millipore), blocked (PBS containing 5% skim milk powder,

NuPAGE Bis-Tris gels (Invitrogen), and proteins were transferred to PVDF

tail (Roche). Next, 25–50 μg of cleared lysate were run per lane on 10%

SDS-PAGE and protein blot. Primary control and case fibroblasts were

lyzed in RIPA buffer (50 mM Tris pH 8.0, 150 mM NaCl, 1% NP-40, 0.5%

sodium deoxycholate and 0.1% SDS) containing protease inhibitor cock-
tail (Roche). Next, 25–50 μg of cleared lysate were run per lane on 10%

NuPAGE Bis-Tris gels (Invitrogen), and 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. After washing, membranes were incubated in anti-mouse or

secondary antibodies (DakoCytomation used at 1:10,000) at room

RT-PCR. RNA was extracted from cultured subject fibroblasts using the

RNAAspin 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 medium containing 100 ng ml⁻¹ puromycin. After

12–20 days of selection, cells were harvested for dipstick assays.

Dipstick enzyme activity assays. Complex I and complex IV 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 com-

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

Homogygosity mapping. Homogygosity 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).

Supplementary Table 2

Data availability. Supplementary Table 2 provides detailed data on all vali-
dated patient variants, and the seven pooled sequence data files (BAM format)
are available upon request.

Reduced collagen VI causes Bethlem myopathy: a heterozygous

COL6A1 nonsense mutation results in mRNA decay and functional haploinsufficiency.


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 100 ng of subject genomic DNA. The products were checked by gel electro-
phoresis, digested overnight with AflIII or NlaIV, respectively (New England
Biolabs) as per manufacturer’s protocol, and resolved on 1% agarose gels.

Antibodies for protein blotting. Antibodies included NDUFS4 (MS104, Mitosciences) at 1:1,000, Porin (529534, Calbiochem) at 1:10,000, complex II

70-kD subunit (A-1142, Molecular Probes) at 1:1,000, and NDUFV2 (kind gift from M. McKenzie and M. Ryan, La Trobe University) at 1:5,000.

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).

SDS-PAGE and protein blot. Primary control and case fibroblasts were

lyzed in RIPA buffer (50 mM Tris pH 8.0, 150 mM NaCl, 1% NP-40, 0.5%
sodium deoxycholate and 0.1% SDS) containing protease inhibitor cock-
tail (Roche). Next, 25–50 μg of cleared lysate were run per lane on 10%

SDS-PAGE Bis-Tris gels (Invitrogen), and 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. After washing, membranes were incubated in anti-mouse or

rabbit secondary antibodies (DakoCytomation used at 1:10,000) at room

temperature for 1 h and developed using ECL or ECL Plus detection reagents
(Amersham Bioscience).

42. Kirby, D.M. et al. Respiratory chain complex I deficiency: an underdiagnosed energy

43. Gnarke, A. et al. Solution hybrid selection with ultra-long oligonucleotides for

44. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling

45. Li, H. et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics

46. Calvo, S.E., Pagliarini, D.J. & Mootha, V.K. Upstream open reading frames cause
widespread reduction of protein expression and are polymorphic among humans.

Bioinformatics Chapter 1: Unit 1.4 (2009).

48. Dimmic, M.W., Sunyaev, S. & Bustamante, C.D. Inferring SNP function using

49. Cree, L.M., Samuels, D.C. & Chinnery, P.F. The inheritance of pathogenic mitochondrial

50. Gabriel, S., Ziaugra, L. & Tabbaa, D. SNP genotyping using the Sequenom MassARRAY

51. Lamandé, S.R. et al. Reduced collagen VI causes Bethlem myopathy: a heterozygous
COL6A1 nonsense mutation results in mRNA decay and functional haploinsufficiency.