npj | genomic medicine

Article

Published in partnership with CEGMR, King Abdulaziz University

https://doi.org/10.1038/s41525-024-00429-5

Biallelic GGGCC repeat expansion leading to *NAXE*-related mitochondrial encephalopathy

Kokoro Ozaki[1,2,12], Yukiko Yatsuka[2,12], Yoshinobu Oyazato[3], Atsushi Nishiyama[3], Kazuhiro R. Nitta[2], Yoshihito Kishita[2,4], Takuya Fushimi[5], Masaru Shimura[5], Shohei Noma[1], Yohei Sugiyama[6], Michihira Tagami[1], Moe Fukunaga[1], Hiroko Kinoshita[1], Tomoko Hirata[1], Wataru Suda[7], Yasuhiro Murakawa[8], Piero Carninci[9,10], Akira Ohtake[11], Kei Murayama[2,5] & Yasushi Okazaki[1,2] ✉

Repeat expansions cause at least 50 hereditary disorders, including Friedreich ataxia and other diseases known to cause mitochondrial dysfunction. We identified a patient with *NAXE*-related mitochondrial encephalopathy and novel biallelic GGGCC repeat expansion as long as ~200 repeats in the *NAXE* promoter region using long-read sequencing. In addition to a marked reduction in the RNA and protein, we found a marked reduction in nascent RNA in the promoter using native elongating transcript-cap analysis of gene expression (NET-CAGE), suggesting transcriptional suppression. Accordingly, CpG hypermethylation was observed in the repeat region. Genetic analyses determined that homozygosity in the patient was due to maternal chromosome 1 uniparental disomy (UPD). We assessed short variants within *NAXE* including the repeat region in the undiagnosed mitochondrial encephalopathy cohort of 242 patients. This study identified the GGGCC repeat expansion causing a mitochondrial disease and suggests that UPD could significantly contribute to homozygosity for rare repeat-expanded alleles.

Mitochondrial diseases are a group of diseases that arise from dysfunctions in mitochondrial oxidative phosphorylation[1,2]. They are the most prevalent diseases of hereditary metabolic disorders, especially during infancy, and are known to occur in 1 out of 5000 births[3]. To date, over 425 of nuclear and mitochondrial genome-coded genes have been reported to be associated with mitochondrial diseases[4]. Single nucleotide variants (SNVs), small indels, and structural variants cause mitochondrial diseases, while Friedreich ataxia (FRDA), a neurodegenerative disorder caused by biallelic GAA repeat expansion in *FXN*, causes mitochondrial dysfunction. It develops during adolescence and leads to loss of frataxin, which in turn causes abnormal accumulation of mitochondrial iron, compromised mitochondrial oxidative phosphorylation, and overproduction of free oxygen radicals which damage diverse cellular functions[5].

At least 50 diseases have been reported to be caused by pathogenic short tandem repeat expansions, and several diseases such as Huntington disease, C9orf72 associated frontotemporal dementia/amyotrophic lateral sclerosis, and myotonic dystrophy type 1, as well as FRDA are known to cause mitochondrial dysfunction[6]. Moreover, repeat expansion diseases often involve the central nervous system (CNS)[7]. Long-read sequencers by PacBio or Oxford Nanopore are useful for studying repeat expansion diseases as they can sequence long segments of repeat expansions, which can amount to several hundreds of base pairs to kilobases[8].

Mitochondrial diseases often affect the CNS. A group of mitochondrial diseases especially characteristic of prominent CNS lesions is called "mitochondrial encephalopathy". One type of mitochondrial encephalopathy called *NAXE*-related mitochondrial encephalopathy is an autosomal recessive hereditary disorder caused by biallelic pathogenic variants in NAD(P)HX epimerase (*NAXE*, alias: *APOA1BP*)[9–17]. Patients with *NAXE*-related mitochondrial encephalopathy exhibit developmental delay, cognitive regression, altered consciousness, abnormalities in eye movement (including nystagmus), muscle weakness, respiratory failure, seizure, ataxia, and gait disturbance at the age of 1 to 2 years. The disease is characterized by fluctuating symptoms over time, often exacerbated by febrile illness; it is progressive and fatal in the long term. NAD(P)HX epimerase is an important enzyme that catalyzes the toxic substance NAD(P)HX in the metabolism of NADPH ("NAD(P)HX repair system"). NADPH is used for various biosynthesis and cellular processes and whose main pools are cytosol and mitochondria[18]. NAD(P)HX epimerase deficiency, typically caused by missense,

A full list of affiliations appears at the end of the paper. ✉e-mail: ya-okazaki@juntendo.ac.jp

nonsense, frameshift, or splicing variants in *NAXE*, compromises the NAD(P)HX repair system and leads to mitochondrial dysfunctions.

In our study of a cohort of undiagnosed mitochondrial disease patients, we identified a patient with *NAXE*-related mitochondrial encephalopathy and novel biallelic GGGCC repeat expansion in the promoter region of *NAXE* using long-read sequencing. Genetic analyses showed that homozygosity in the patient was due to maternal chromosome 1 uniparental disomy (UPD). A search for pathogenic GGGCC repeat expansion using repeat-primed polymerase chain reaction (RP-PCR) in an undiagnosed mitochondrial encephalopathy cohort yielded negative results. Furthermore, amplicon long-read sequencing used to assess short variants and GGGCC repeats in our cohort showed that among the 484 alleles observed, none was longer than 7 repeats, except for the patient and the mother who carried pathogenic repeat expansion. When repeat genotypes in the *NAXE* GGGCC repeat locus were searched in the TR-gnomAD[19], a dedicated catalogue of short tandem repeat, less than 0.15% alleles were not within 2–7 repeats. Consequently, this study provides an important perspective that UPD, occurring in 0.05% of live births in the general population[20], could significantly contribute to homozygosity for exceedingly rare repeat-expanded alleles.

Results

Isolation of the patient with markedly reduced *NAXE* expression and biallelic GGGCC repeat expansion in the *NAXE* promoter

As part of our strategy to achieve molecular diagnoses on undiagnosed 2932 patients of mitochondrial diseases in Japan, we conducted RNA sequencing of RNA samples from fibroblasts obtained from 400 patients, in addition to genomic sequencing including panel sequencing, whole exome sequencing, and whole genome sequencing (WGS). A total of 303 (75.8%) of the 400 patients were biochemically confirmed to have mitochondrial dysfunctions (Fig. 1a). To isolate the individuals with markedly dysregulated candidate causative gene expression, we utilized "OUTRIDER", a tool that deciphers singleton samples with outlying dysregulation of gene expression in a batch of samples (for example, a batch of 32 samples)[21]. In the sixth batch of the 11 batches examined (400 cases in total), among 32 patients, we found a case of a 3 year-old-girl with outstandingly reduced *NAXE* expression (Fig. 1a–c).

This patient (Pt2359, the proband in Family A) without an apparent familial history of mitochondrial, rare, or neurodegenerative diseases, presented with truncal ataxia at the age of 1 year and 1 month, along with mild psychomotor developmental delay with regression, hepatomegaly, seizure, and difficulty hearing in the right ear. The patient was not dysmorphic. The clinical course was fluctuating with deterioration often preceded by febrile events and progressive in neurological deficits (Supplementary Fig. 1a). At age of 3 years, the patient was unable to sit, a skill she had acquired at 9 months old. Instead, she could only pull herself up to stand, exhibiting trunk instability. Febrile events such as upper respiratory infection or urinary tract infection for several days often preceded clinical deteriorations, which could last for weeks and partially recover over the course of several weeks or months. Laboratory screening revealed elevated lactate and pyruvate in cerebrospinal fluid (lactate: 30.6 mg/dl, pyruvate: 1.36 mg/dl). No inflammatory changes were observed in the cerebrospinal fluid. Although repeated episodes of seizures were observed at the age of 1 year and 3 months, the electroencephalograms showed only scarce intermittent nonspecific slow waves, but no apparent epileptic discharges or other specific findings. Supplementation therapy with vitamins B1, B6, B12, C, E, biotin, carnitine, and ketogenic diet was initiated at the age of 1 year and 3 months, and continued thereafter. Clinical stabilization and neurological improvement were subsequently observed at least for 9 months. However, several events of neurological exacerbation, often preceded by seemingly infectious episodes, ensued.

At the age of 5 years, an episode of severe neurological exacerbation occurred. Gait disturbances, decreased level of consciousness, respiratory failure, dysautonomia such as highly volatile blood pressure and hypersalivation, and skin symptoms resembling heat burn on the extremities and trunk were accompanied by urinary tract infection upon admission. The exacerbation was so rapid that the patient became comatose in several days. Brain magnetic resonance imaging (MRI) was initially normal but two weeks later, during the rapid exacerbation, showed evolution of high signal intensities of the brainstem and cerebellum in fluid attenuated inversion recovery (FLAIR) as well as restricted diffusion in the diffusion weighted imaging (DWI) / apparent diffusion coefficient (ADC) map (Fig. 1b) compatible with cytotoxic edema. Later, narrow areas of the cerebrum showed similar lesions (Supplementary Fig. 1b). All these brain lesions revealed by MRI persisted for several weeks at least and resulted in severe neurological sequelae.

Fluctuating but thus progressive neurological symptoms and lab tests suggestive of mitochondrial diseases were compatible with *NAXE*-related mitochondrial encephalopathy. For the analysis of mitochondrial oxidative phosphorylation, the oxygen consumption rate (OCR) of the patient's fibroblasts was measured, revealing a moderate decrease (48% of the controls and 43% in the galactose-containing medium). Furthermore, the enzyme activity of oxidative phosphorylation in the patient's fibroblasts was reduced in complex II and III (complex I: 61.6%, II: 49.1%, II + III: 34.7%, III: 113.7%, and IV: 102.5% in the Citrate Synthase (CS) ratio), whereas the changes were milder in the muscle (I: 69.2%, II: 79.6%, II + III: 85.5%, III: 44.0%, and IV: 49.4%). Loss of the NAXE protein in Pt2359 was also confirmed by western blotting (Fig. 1d and its original gel blot image in Supplementary Fig. 2), in comparison with the controls and another *NAXE*-related mitochondrial encephalopathy patient (Pt2659), by compound heterozygous splicing and missense variants. However, short read WGS with conventional variant calling in Pt2359 revealed no causative variants in *NAXE*.

Furthermore, we used PacBio long-read sequencing to identify biallelic 1 kb stretch of almost pure GGGCC repeat expansion (~200 repeats) in the promoter region of *NAXE* in the patient's fibroblasts, compared to the three repeats in the reference sequence (GRCh38) (Fig. 1e (upper part), f and g, and Supplementary Table 1).

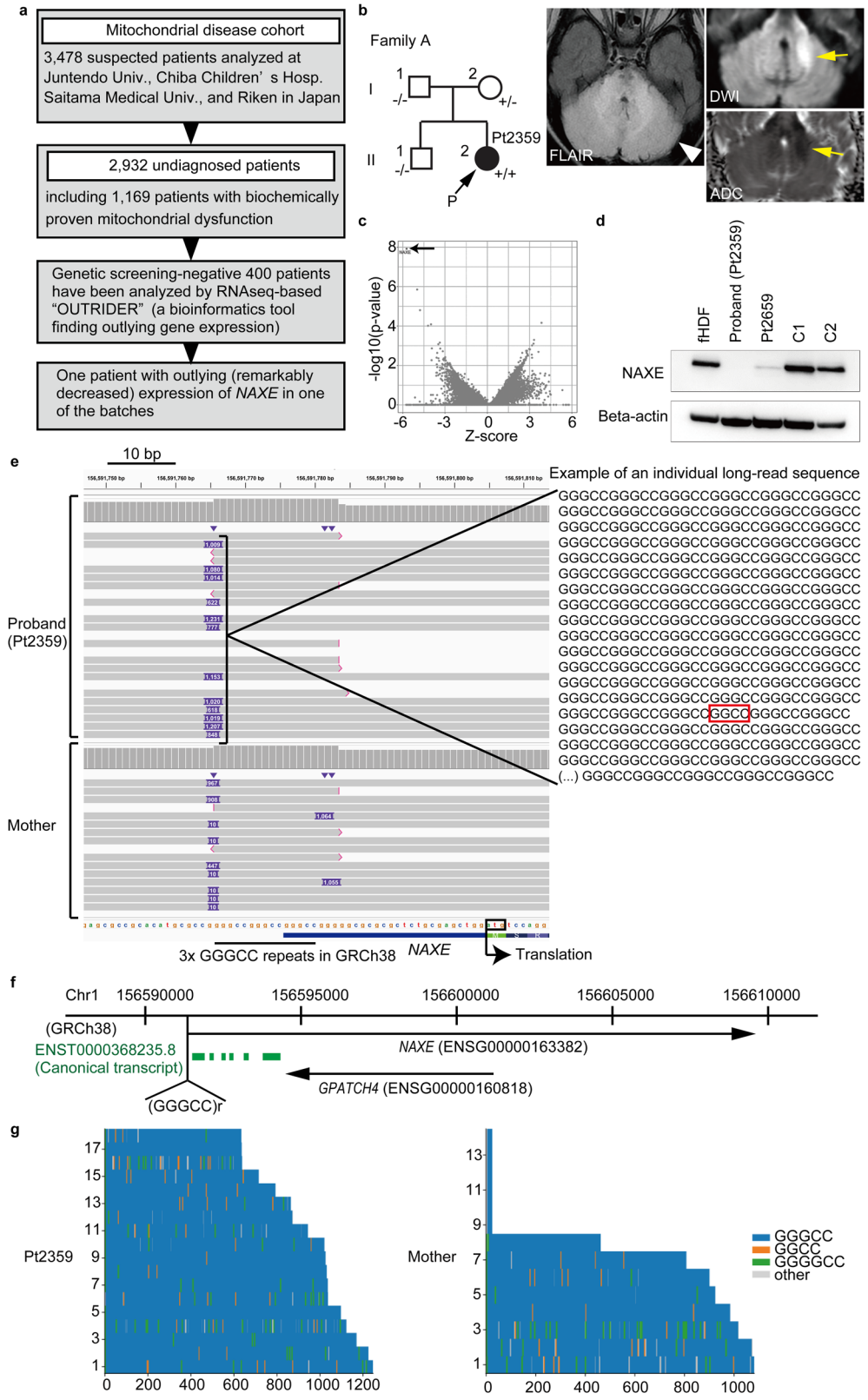
We confirmed that no other pathogenic variants were detected in the genes, which was previously reported to cause encephalopathies[22,23] (Supplementary Table 5) using the data of WGS and RNA sequencing. For the mutations in the mitochondrial genome, we identified no clear pathogenic point mutations such as m.3243 A > G or single/multiple deletions in the mitochondrial genome, using WGS data by short-read and long-read sequencing.

The chromosome 1 UPD in the proband

We developed a protocol for RP-PCR targeting the GGGCC repeat expansion in *NAXE*, which can detect existence of repeat expanded alleles for evaluation and cost-effective screening (Fig. 2a). We first evaluated Family A and detected the repeat expanded allele in Pt2359 and the mother, but not in the father or older brother (Fig. 2b). Using long-read sequencing, we confirmed heterozygous repeat expansion in the mother's peripheral blood leukocytes (Fig. 1e, lower part). Southern blotting (Fig. 2c, d, and its original gel blot image in Supplementary Fig. 3) also showed that only the mother was heterozygous for the repeat expansion. Analysis of short-read WGS data by evaluation of homozygous SNV ratios using AutoMap[24] revealed that Pt2359 harbored whole chromosome 1 UPD (Fig. 2e). We searched for additional homozygous variants in other genes within chromosome 1 UPD regions which could lead to mitochondrial dysfunction or symptoms using WGS data of Pt2359. We found no variants that can reasonably explain the symptoms of Pt2359 (Supplementary Table 2).

RP-PCR screening for pathogenic GGGCC repeat expansion in *NAXE* in an undiagnosed mitochondrial encephalopathy cohort

We utilized RP-PCR to screen 461 undiagnosed patients with signs of encephalopathy in the Japanese mitochondrial disease cohort



(Group1, $n = 314$ and Group2, $n = 147$) (Table 1). RP-PCR detected no alleles of extreme repeat expansion in *NAXE*, but it was difficult to evaluate alleles with 3 to 4 repeats longer than the reference (repeat Number=3) because of inherent limited precision at single repeat unit resolution in RP-PCR.

Evaluation of GGGCC repeat expansion by bioinformatics tools using short read whole genome sequencing data

For future screening for GGGCC repeat expansion using short read whole genome sequencing data, we examined ExpansionHunter, a bioinformatics tool for detecting repeat expansion using short read sequencing data[25].

Fig. 1 | Biallelic GGGCC repeat expansion in an *NAXE*-related mitochondrial encephalopathy patient. a Our strategy for the identification of patients with outlying gene expression related to mitochondrial dysfunction. A total of 400 patients, including 303 patients with proven biochemical deficits, who were negative for pathogenic variants in the genetic screening, were analyzed by OUTRIDER. One of such OUTRIDER batches isolated a patient (Pt2359) with a marked decrease in *NAXE* expression. **b** "Family A" with markedly decreased *NAXE* expression in the proband. An arrow with "P" denotes the proband. -/-, +/-, and +/+ denote homozygote for wild type, heterozygote for repeat expansion, and homozygote, respectively. FLAIR MRI showed hyperintensity in brainstem and cerebellum (arrowhead). DWI and ADC showed restricted diffusion in part (arrows). **c** Volcano plot showing a remarkably decreased *NAXE* expression in the proband (arrow). **d** Western blotting showing loss of NAXE protein in the proband's fibroblasts. NAXE positive controls (fHDF, fetal human dermal fibroblast (normal); C1 and C2, mitochondrial disease controls (Pt1615YS and Pt1753) who have no causative variants in *NAXE* and normal *NAXE* expression); Pt2659, another patient with a novel splicing variant and a missense variant; Beta-actin, a loading control. **e** Left panel, long-read sequencing of the proband showed the expansion of the GGGCC repeat (indicated as insertion (a dark violet tag with a number denoting inserted length of bases)) in each read (upper half) and the mother in a heterozygous state (lower half). Right panel, an example of such long-read sequence with rare inserts of slightly different repeat units (the red rectangle). **f** The site of repeat expansion with respect to *NAXE* gene body. **g** Waterfall plot showing repeat contents of Pt2359 and the mother for each long-read.

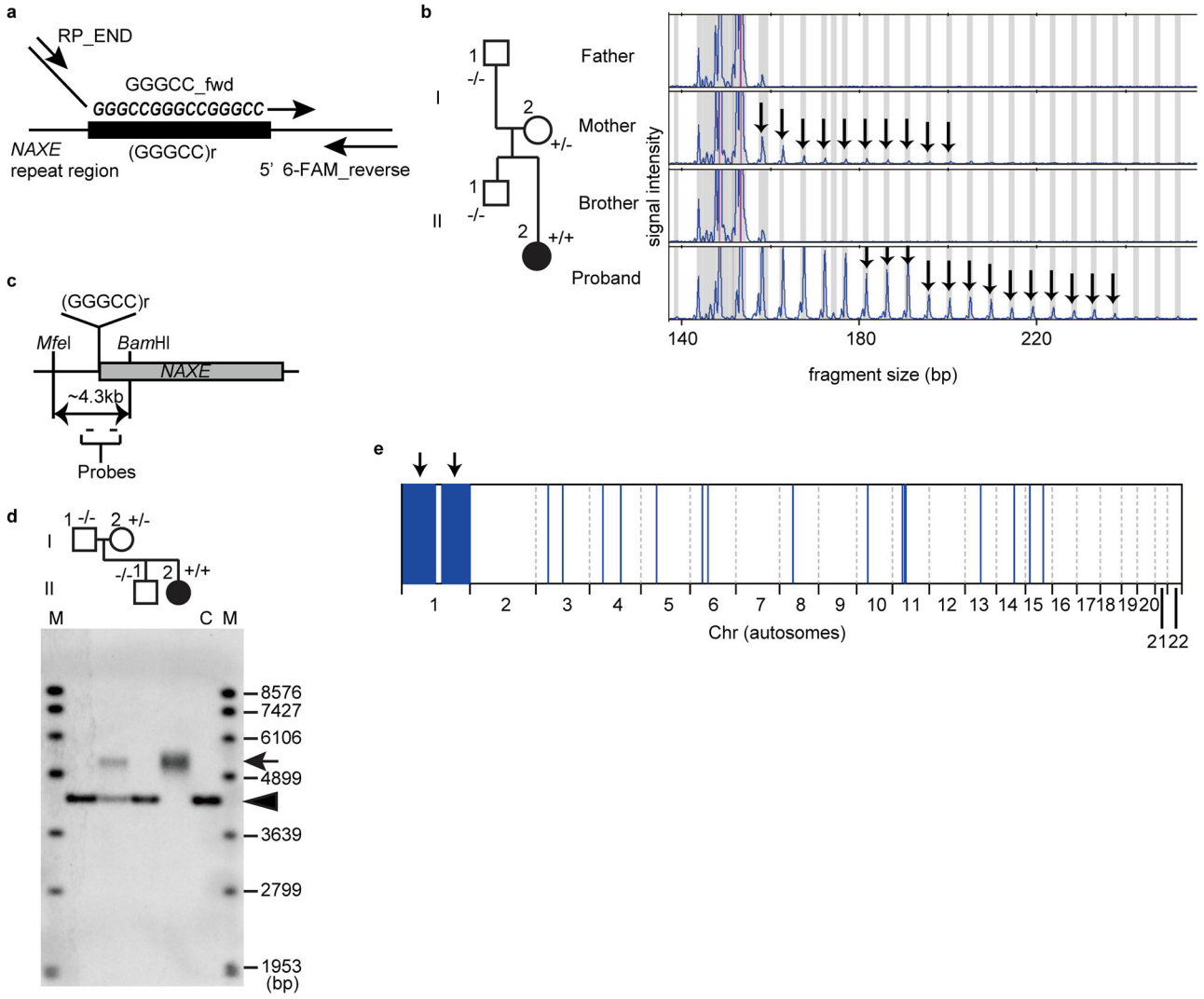


Fig. 2 | Genetic evaluation supporting the status of the *NAXE* locus in the pedigree. a Design of repeat-primed polymerase chain reaction (RP-PCR) showing three primers (GGGCC_fwd primer and 5' 6-FAM_reverse were used for initiating amplification between the repeat stretch and the adjacent genomic region, and RP_END primer was used for further efficient amplification) for the *NAXE* GGGCC repeat. **b** Segregation study using RP-PCR, showing Family A (left panel) and corresponding RP-PCR electropherogram detecting fluorescently labeled amplicons. Arrows in the mother and proband (Pt2359) indicate the presence of pathogenic GGGCC repeat expansion. **c** Probe design for southern blotting. Two small bars are probes that recognize the ~4.3 kb fragment created by *Mfe*I and *Bam*HI digestion of genomic DNA. (GGGCC)r denotes the GGGCC repeat in the *NAXE* promoter region. **d** Southern blotting confirmed the presence of a repeat expansion allele in the mother (I-2, in a heterozygous state) and proband (in a homozygous state). C, HapMap GM12878 sample with a normal range of repeat stretch as the control; M, molecular markers for southern blotting; arrow, expanded repeat allele; arrowhead, normal allele. **e** Homozygous stretch of single nucleotide variants (SNVs) using AutoMap on short read WGS data of the proband shows that most of the chromosome 1 regions were homozygous, suggesting uniparental disomy.

When the parameters were set to find GGGCC repeat expansion in the *NAXE* promoter region, homozygous repeat expansion was identified in Pt2359, while mild signs of heterozygous repeat expansion were identified in the mother (Supplementary Table 3). In addition, 366 patients with undiagnosed mitochondrial diseases were examined. Four cases showed slightly expanded repeat in contrast to 362 other cases. However, when reviewed in the Integrative Genomic Viewer (IGV)[24–26], one case had the very frequent genotype with 5× GGGCC in one allele and GGGTC and 4× GGGCC in another allele. In the other three cases, visual inspection in the IGV could not determine the repeat length because of low coverage in the repeat region. We further checked them using RP-PCR, and obtained negative results for all the three cases. Therefore, Pt2359, who has homozygous GGGCC repeat expansion in *NAXE* was reliably detected by ExpansionHunter, while the mother, who has heterozygous GGGCC repeat expansion, showed mild signs of expansion according to ExpansionHunter.

To further examine whether the detection of GGGCC repeat expansion in *NAXE* without pre-setting the locus as the specific target site of repeat expansion ("de novo detection of repeat expansion") using whole genome sequencing data, genome-wide evaluation by ExpansionHunter Denovo[27] and STRling[28] were performed. Both ExpansionHunter Denovo and STRling failed to detect the GGGCC repeat expansion in *NAXE* in Pt2359 and the mother.

Table 1 | Patient characteristics of the mitochondrial encephalopathy cohort

Features		Group1	Group2
	Total number of patients	314	147
Sex	Male	169	73
	Female	143	69
	Unknown	2	5
Age of onset	Prenatal	6	0
	0 day to <1 month	78	14
	1 month to < 1 year	85	35
	1 year to < 3 y	47	21
	3 y to < 10 y	26	20
	10 y to <20 y	9	8
	20 y or older	5	2
	Unclear	58	47
Family history	Present	19	5
	Unremarkable within 3 degrees	295	142
Clinical diagnosis	Cardiomyopathy	4	0
	Hepatic Disease	1	1
	Leigh Disease	117	27
	Lethal Mitochondrial Infantile Disorder	2	0
	Mitochondrial Cytopathy	176	105
	Neurodegenerative Disorder	9	6
	Sudden Unexpected Death	4	8

We listed the mitochondrial encephalopathy patients in chronological order and grouped them into the two groups (Group1, $n = 314$; Group2, $n = 147$), for screening by RP-PCR. Group1 was investigated by subsequent ultralong PCR and long-read amplicon sequencing. As for the "Clinical diagnosis", a list of pre-defined disease types (Cardiomyopathy, Hepatic Disease, Leigh Disease, Lethal Mitochondrial Infantile Disorder, Mitochondrial Cytopathy, Neurodegenerative Disorder, Sudden Unexpected Death) was used for expressing the chief clinical state of each patient.

Amplicon long-read sequencing for evaluation of short GGGCC repeats, SNVs, small indels, and structural variants

To evaluate the composition, length, and frequency of short GGGCC repeats, along with SNVs, small indels, and structural variants at the *NAXE* locus, we further analyzed 314 patients (Group1) using amplicon long-read sequencing with PacBio for two overlapping amplicons (14086 and 13786 bp) covering the *NAXE* genomic region (Fig. 3a, b). We added four members of Family A, Pt2659, and HapMap sample GM12878 as controls. Following a quality assessment of the sequencing data, 242 samples were selected for further analysis. We initially examined if specific variations in the GGGCC repeat region are associated with mitochondrial encephalopathies. Most of the alleles were with 5× GGGCC repeats or 1× GGGTC + 4× GGGCC repeats (chr1: 156591765 G > GGGGCCGGGCC, allele frequency in the cohort 0.483 (234/484); chr1: 156591765 G > GGGGTCGGGCC, 0.498 (241/484), respectively). Only five alleles of 6× GGGCC repeats (0.010 (5/484)) and 1 allele of 7× GGGCC repeats (0.002 (1/484)) were observed. These allele frequencies were comparable with those reported in the TR-gnomAD v1.0, the dedicated tandem repeat catalogue in 338,963 human genomes[19], where only 0.15% of the alleles in all the ancestries correspond to the repeat length that is not within the common 2–7 repeats in the locus (Fig. 3b). *NAXE* expression was determined in one mitochondrial encephalopathy patient with 6× GGGCC allele and another with 7× GGGCC allele whose fibroblasts were available and underwent RNAseq. The patients showed *NAXE* expression levels comparable to those of 70 other patients in the same batches (data not shown).

To identify short variants (SNVs and small indels outside the GGGCC repeat region) that could be pathogenic for *NAXE*-related mitochondrial encephalopathy, we focused on the rare (allele frequencies of 0.1 or lower) variants that may impact the NAXE protein function (see the criteria in the Methods). Using this criteria, five variants (two missense variants, one splicing variant, and two deep intronic variants) were identified, including the splicing variant and the missense

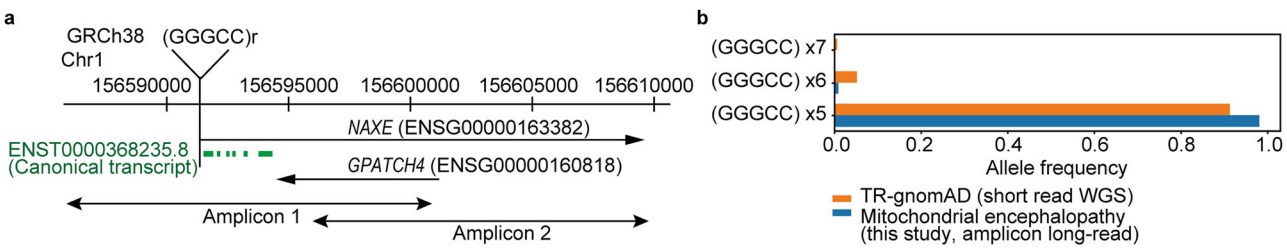
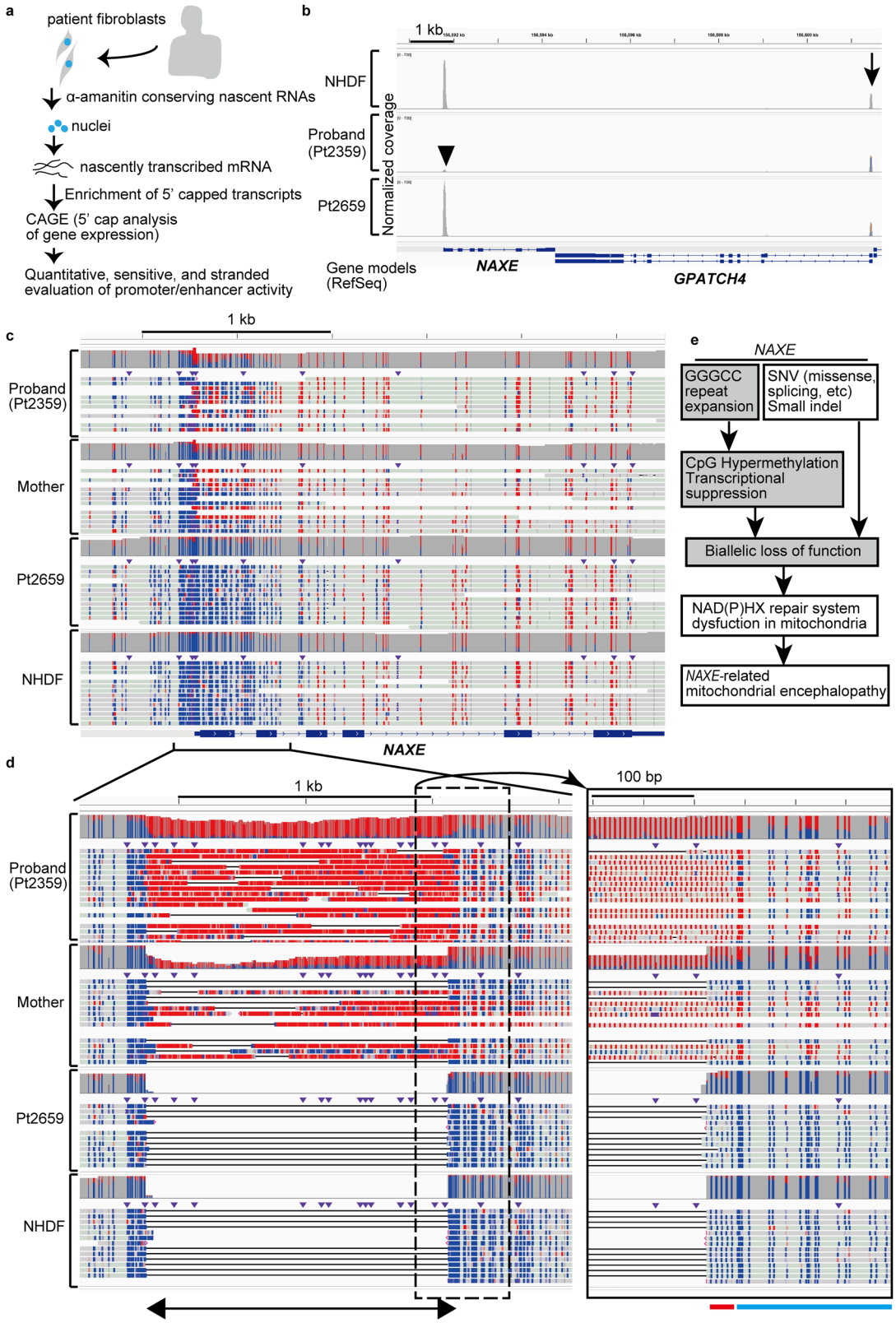


Fig. 3 | Amplicon long-read sequencing in the unidentified mitochondrial encephalopathy cohort. a Design of amplicons for long-read sequencing. Two overlapping long-PCR amplicons covering the *NAXE* genomic region were designed. (GGGCC)r denotes the site of pathogenic GGGCC repeat expansion in the promoter region of *NAXE*. **b** Allele frequencies of variations in the GGGCC repeat region in the examined cohort and in TR-gnomAD, a dedicated tandem repeat catalogue, based on short-read whole genome sequencing (WGS). (GGGCC) ×5 denotes ×5 repeat and its derivatives. For details of each allele with different GGGCC repeat length or internal sequence, see Supplementary Table 4.



variant in Pt2659 (Supplementary Table 4). Among the five variants, another heterozygous missense variant (chr1:156593531 A > G, c.640 A > G, p.Ile214Val), absent in the public databases such as ToMMo54KJPN[29] and Gnomad v4.0[30], exhibited a Combined Annotation Dependent Depletion (CADD) phred scale score of 17.07[31]. The two deep intronic variants were found to have low SpliceAI[32] delta and low CADD scores.

For structural variations, we called three structural variations in total, but manual inspection in IGV revealed that the two following variations (SV1 in one patient ("Pt0552") and SV2 in another patient ("Pt1753")) are

Fig. 4 | Functional consequence of GGGCC repeat expansions in *NAXE*. **a** Experimental principles of NET-CAGE. The nuclei were isolated in the presence of α-amanitin, which maintains the nascently transcribed mRNAs. The nascent transcripts were enriched and processed using 5' cap trapping technology to capture the full-length transcripts for subsequent sequencing (cap analysis of gene expression (CAGE)). **b** NET-CAGE of the proband fibroblasts compared to the controls (NHDF, normal neonatal human dermal fibroblast; Pt2659, patient with splicing and missense variants in *NAXE*) showing markedly suppressed nascent transcripts in the promoter region of *NAXE* (arrowhead). Note the comparable level of signals at *GPATCH4* (arrow), which is reciprocally transcribed. **c** CpG hypermethylation (methylation indicated in red) was detected within the *NAXE* promoter region by analysis of long-read sequencing data in the proband as well as in the mother, whereas not in Pt2659 and NHDF. Reads were aligned to the reference genome (GRCh38). **d** Left panel, long-reads mapped against a sequence with expanded repeat showing hypermethylation in the GGGCC repeat per se as well as in the region downstream of the repeat stretch. The bidirectional arrow at the bottom denotes the repeat stretch. Right panel, magnified view of the dotted rectangle area in the left panel. Red and light blue lines below the right panel show the sites of normal length repeat in the controls (Pt2659 and NHDF) and the promoter region downstream of the repeat stretch, respectively. **e** Diagram of the suggested pathologic mechanism. GGGCC repeat expansion causes hypermethylation and suppressed transcription.

likely real variations in the highly polymorphic region upstream of the *NAXE* locus: SV1 (in "Pt0552"): heterozygous deletion of sequence GTTTCACCATGTTGGCCAGGCTGGTCTCGAACTCCTGACATCAGGTGATCCACCTGCCTTGGCCTCCCAAAGTGTTGGGATTACAGGTGTGAGCCAGTGCATCCAGCCCTAATTTTTGTATTTTTAGTAGAGGTGGT between chr1:156590050-156590187 (length=137 bp) and SV2 (in "Pt1753"): heterozygous insertion of sequence GTTTCACCATGTTGGCCAGGCTGGTCTCGAACTCCTGACATCAGGTGATCCACCTGCCTTGGCCTCCCAAAGTGTTGGGATTACAGGTGTGAGCCAGTGCATCCAGCCCTAATTTTTGTATTTTTAGTAGAGGTGGT at chr1:156590050 (length=137 bp). Within this region upstream of the *NAXE* gene body, we could not detect an enhancer/promoter element described in ENCODE cis regulatory elements (cCREs) using SCREEN (see Supplementary Note 1). Therefore, we have not identified any significant structural variants in the *NAXE* locus in the screening.

Variabilities of repeat internal sequences in the patient and the mother

We investigated the variations in repeat internal sequences of every long read in the patient's (Pt2359) fibroblasts and mother's (I-2 in Family A) leukocyte genome and found slight variations in repeat unit sequences as well as moderate variations in the number of units (patient/mother: 96.8/96.1% of all repeat units were pure GGGCC in 12/5 reads with repeat unit numbers of 196.1 ± 41.5/180.4 ± 51.1 (average±standard deviation), respectively) (Fig. 1g and Supplementary Table 1).

Transcriptional suppression of *NAXE* in the promoter

To understand the molecular pathologic mechanism underlying the loss of *NAXE* transcripts, we conducted a Native Elongating Transcript Cap Analysis of Gene Expression (NET-CAGE)[33] to quantitatively and sensitively evaluate the promoter/enhancer activity by specific enrichment of nascently transcribed and 5' capped RNAs in the patient's fibroblasts (Fig. 4a) in a strand-aware manner. NET-CAGE showed a marked decrease in the nascent transcripts in the *NAXE* promoter in Pt2359 compared to that in the normal control and Pt2659 (Fig. 4b). Furthermore, bioinformatics analysis of the kinetics features of PacBio long-read sequencing data detected CpG hypermethylation at the repeat sequence (GGGCC) per se and in the region downstream of the repeat sequence (Fig. 4c, d) around the promoter region of the patient's genome, compared to those in the controls. The long-read sequencing in the mother also showed hypermethylation in the allele with repeat expansion, whereas the allele with a normal range of GGGCC repeat did not show hypermethylation in the repeat and its downstream region (Fig. 4d).

Discussion

In this study, we described a *NAXE*-related mitochondrial encephalopathy patient whose condition was caused by biallelic GGGCC repeat expansion in the *NAXE* promoter region (ACMG criteria[34], PVS1). To date, repeat expansions leading to mitochondrial disease have not been reported, except in FRDA. Moreover, this is the first case of the GGGCC/GGCCC repeat expansion leading to a human disorder. Analysis of nascent transcripts by NET-CAGE and CpG hypermethylation by long-read sequencing suggested that the loss of *NAXE* was caused by transcriptional suppression supported by CpG hypermethylation in the promoter containing the GGGCC repeat (Fig. 4e).

As the search for an allele of pathologic repeat expansion by RP-PCR showed that there was not an additional patient with the *NAXE* GGGCC repeat expansion in the mitochondrial encephalopathy cohort of 461 undiagnosed patients, very low frequency of the repeat expanded allele was revealed, consistently with the public repeat database TR-gnomAD. Further evaluation of short variants as well as structural variants by amplicon long-read sequencing in the mitochondrial encephalopathy cohort of 314 undiagnosed patients, which were conducted in search of candidate pathogenic variants for *NAXE*-related mitochondrial encephalopathy, identified several variants including a novel missense variant. Further studies will be needed to evaluate the pathogenicity of the variants. Variations in the GGGCC repeat within the *NAXE* promoter in the mitochondrial encephalopathy cohort exhibited allele frequencies similar to those in the Japanese public database ToMMo54KJPN or the repeat catalogue TR-gnomAD. The data also showed that a sharp decrease in allele frequencies with increasing GGGCC repeats (for example, from 0.010 for 6 repeats to 0.002 for 7 repeats). It is therefore expected that alleles with pathogenic and expanded GGGCC repeats would be exceedingly rare in the general population, with frequencies as low as 0.0001 or even lower.

In the pedigree described, genetic analyses determined that homozygosity of repeat expansion in the proband was due to maternal chromosome 1 UPD. Supporting the concept, it is known that UPDs occur most frequently on chromosomes 1, 4, 16, 21, 22, and X in the general population[20]. In the study, 105 incidences (approximately 0.05%) of UPD were detected among 214,915 trios. This implies that UPD events leading to a homozygous state might occur as frequently as, or even more frequently than, homozygosity in offspring from unrelated heterozygous parents for exceedingly rare repeat-expanded alleles.

GGGCC repeat expansion in *NAXE* has common characteristics with fragile X syndrome (FXS) caused by *FMR1* CGG repeat expansion in the 5' untranslated region (UTR)[35] in that both display transcriptional suppression and CpG hypermethylation associated with repeat expansion, leading to the loss of gene product. It would be interesting to examine if the R-loop mediated mechanism, which presumably mediates FXS suppression[36], is present in *NAXE* GGGCC repeat expansion.

Notably, GGGCC repeat expansion in *NAXE* might be similar to GGGGCC repeat expansion in C9orf72, which is responsible for familial and sporadic amyotrophic lateral sclerosis (ALS). GGGGCC repeat in C9orf72 leads to DNA- and RNA-G-quadruplexes, which are non-canonical structures, by Hoogsteen base pairing between guanines[37,38]. While DNA G-quadruplex affects transcription and replication, RNA G-quadruplex affects translation and sequesters RNA-binding proteins as well as produces RNA foci[37]. Because GGGCC repeat in *NAXE* matches the consensus sequence for G-quadruplex formation [5'-G(> = 3)N(1-7)G(> = 3)N(1-7)G(> = 3)N(1-7)G(> = 3)-3'], the biological consequences mentioned may arise. Furthermore, repeat-associated non-AUG translation (RAN translation), another known pathologic mechanism of GGGGCC in C9orf72, may occur in GGGCC repeat in *NAXE*. In RAN translation, peptide repeats are translated from genomic repeat sequences in the manner

ribosomes start translation without an AUG initiation codon[38]. Peptide repeats are often toxic to cells, comprising one of the gain-of-function mechanisms by pathogenic repeat expansion. Peptide sequences starting from any amino acid residue of "[-Gly-Pro-Gly-Arg-Ala-]", translated from [-GGGCCGGGCCGGGCC-] in *NAXE*, may play a role in the pathogenesis of a case with GGGCC repeat expansion. However, in our case of GGGCC repeat expansion in *NAXE*, our observation that *NAXE* mRNA was extremely decreased, may minimize the impact of such gain-of-function mechanisms compared to loss-of-function mechanisms.

When considering the secondary structure in the GGGCC repeat sequence by a bioinformatics tool, mfold[39], DNA- and RNA- 10xGGGCC repeat sequences both gave a long stalk with a loop structure (Supplementary Fig. 4). Although the biological consequence of the structure is unknown, further investigation may reveal a functional role of the structures, particularly in RNA level.

This study had several limitations. One is that we found only a single case (or two carriers) of repeat expansion in the pedigree, which warrants further search of more cases and carriers in appropriate cohorts, and another is the limited availability of human samples for tissue types, as only the fibroblasts, but not CNS cells, were examined for molecular analysis. In examining the undiagnosed mitochondrial encephalopathy cohort for variants in the *NAXE* genomic region, we investigated pathogenic GGGCC repeat expansion by RP-PCR, and analyzed PCR-amplifiable fragments through amplicon long-read sequencing. Thus, PCR-resistant structural variants including large insertions may be overlooked. Another limitation is that we have not tried to measure the toxic metabolites such as cyclic-NADHX, which is a direct consequence of the dysfunction of NAD(P)HX epimerase encoded by *NAXE*.

Finally, it is worth considering checking the *NAXE* GGGCC repeat expansion in cases of unexplained mitochondrial encephalopathy, especially when *NAXE* expression is decreased. In addition, this study added GGGCC repeat to the list of short tandem repeats whose abnormal expansion can cause human diseases. This work underscores the importance of genetic scrutiny using modern technologies such as long-read sequencing and NET-CAGE in discovering and evaluating novel repeat expansions even in previously defined hereditary diseases which are mostly caused by non-repeat variants, and suggests that UPD could significantly contribute to homozygosity for rare repeat-expanded alleles.

Methods

Subjects

The study was conducted under ethical agreement and permission of the review board in the involved facilities (Riken, Juntendo University, Saitama-Medical University, Chiba Children's Hospital in Japan). Patients or parents of patients gave written informed consent. All procedures were conducted following the relevant rules and guidelines and in accordance with the Declaration of Helsinki.

In our alliance of multiple hospitals and institute (Juntendo University, Chiba Children's Hospital, Saitama Medical College Hospital, and Riken) in Japan, we recruited a cohort of 3478 suspected mitochondrial disease patients under an ethical agreement by Institutional Review Board from 2007 to 2024. Patients were of Japanese origin except for one American-Japanese, one Brazilian, one Korean, two Sudanese, and one Vietnamese. Entry and molecular diagnosis were requested by pediatricians and physicians nationwide in Japan with signed informed consents from patients or their parents, motivated by clinical diagnoses of suspected mitochondrial diseases based on clinical characteristics suggestive of mitochondrial diseases, laboratory tests results such as elevated pyruvate or lactate levels in blood and/or cerebrospinal fluid, suggestive pathological findings in muscle biopsy, or compatible imaging abnormalities by brain computed tomography (CT) or magnetic resonance imaging (MRI). Assessments were done in candidate cases by panel sequencing on a Human Gene Mutation Database (HGMD (Supplementary Note 1))-based mitochondrial disease-related in-house list of genes, whole exome sequencing, or whole genome sequencing as well as biochemical evaluation of the oxygen consumption rate (OCR) and enzyme complex activity assays. Of the 2932 undiagnosed patients, 1169 were biochemically proven to have mitochondrial dysfunctions either in OCR or enzyme complex activity assays. As part of our strategy to diagnose such undiagnosed mitochondrial disease cases, 400 cases in total (303 (75.8%) of which were biochemically proven) were investigated by RNA sequencing-based search for candidate causative genes using "OUTRIDER".

In the undiagnosed mitochondrial disease cohort of 2932 patients, we defined "mitochondrial encephalopathy" patients for this study by filtering our database by clinician's diagnosis of any kind of "encephalopathy", "Leigh disease", or "mitochondrial cytopathy" referring to mitochondrial encephalo-myopathies, but lacking molecular diagnosis after either panel sequencing, whole exome sequencing, or whole genome sequencing. We listed the patients in chronological order and grouped them into two groups because of preparation timing (Group1, $n = 314$; Group2, $n = 147$) (We screened the Group1 for the subsequent ultralong PCR and long-read amplicon sequencing) (Table 1). We conducted RP-PCR for both groups along the protocol described in the Methods section. As the controls for the RNA analyses, including NET-CAGE, we utilized data of an *NAXE*-related mitochondrial encephalopathy patient (Pt2659) verified by panel sequencing to have a novel splicing variant (NM_144772.3:exon6:c.402+3_402+6del) and a known missense variant (NM_144772.3:exon6:c.733 A > C:p.Lys245Gln) in *NAXE*. Pt2659 was also analyzed by amplicon long-read sequencing as well as western blotting. The two variants were proven to be in different alleles by long-read sequencing (data not shown) and in state of compound heterozygosity. Pt2359 (homozygous for GGGCC repeat expansion) and Pt2659 (compound heterozygous for splicing and missense variants), both having *NAXE*-related mitochondrial encephalopathy, were unrelated and not from the same local region. For western blotting, two additional patients, namely Pt1615YS (*NDUFA4* deficiency) and Pt1753 (*NDUFA8* deficiency), who were negative for *NAXE* pathogenic variants and had normal expression levels, were used as positive controls as well. The family members (the proband (Pt2359), mother, father, and older brother) of Family A, in which *NAXE* GGGCC repeat expansion was identified, were all subjected to amplicon long-read sequencing and RP-PCR as well.

Patient fibroblast culture and DNA/RNA isolation

Patient skin fibroblasts and control fibroblasts (fetal human dermal fibroblast (CA10605f, HDF-fetal, Toyobo, Japan) as well as neonatal normal human dermal fibroblasts (NHDF, Cell Applications, USA)) were cultured in Dulbecco's modified Eagle's Medium (DMEM) with 10% fetal bovine serum and 1% penicillin and streptomycin. DNA was purified using the MagAttract HMW DNA kit (Qiagen, Germany). For quality assessment, NanoDrop One (ThermoFisher, USA), Qubit dsDNA kit (ThermoFisher, USA), and TapeStation 4200 (Agilent Technologies, USA) were used. RNA was purified using the Maxwell RSC simplyRNA Cells Kit (Promega, USA).

Short-read whole genome sequencing and data analysis

WGS libraries were prepared from 200 ng of genomic DNA using an MGIEasy FS DNA Library Prep Kit v2.0 (MGItech, China) in the proband or v2.1 in the other subjects following the manufacturer's instructions. Paired-end 150-bp sequencing was performed on DNBSEQ-G400. Single nucleotide variations were assessed along an in-house pipeline. Briefly, after quality check and removal of low-quality reads, the reads were mapped against GRCh38/hg38 using Burrows-Wheeler Aligner[40]. Recalibration and variant calling was conducted using GATK[41]. Annotation was carried out using ANNOVAR[42]. AutoMap was used to detect the homozygous regions of each chromosome. To rule out the pathogenic variants in the genes that cause encephalopathies, the list of genes (Supplementary Table 5) was screened. For evaluation of mutations in the mitochondrial genome, we used Matchclips[43] for detecting mitochondrial deletions as well as referring long-read data.

For evaluation of repeat expansion, ExpansionHunter, ExpansionHunter Denovo, and STRling were used. ExpansionHunter was conducted

in 366 undiagnosed mitochondrial patients and the four members of Family A. The settings for the calculation were as follows: [LocusID: NAXE, ReferenceRegion: chr1:156591740-156591800, LocusStructure: GGGCC, and VariantType: Repeat]. ExpansionHunter Denovo was conducted in the "case-control mode" comparing Pt2359 and the mother against 10 other patients with undiagnosed mitochondrial disease. STRling was conducted in 366 patients with undiagnosed mitochondrial disease and the four members of Family A.

RNA sequencing and data analysis

The mRNA was purified from total RNA using oligo(dT)-attached magnetic beads. After fragmentation, first-strand cDNA was generated using random hexamers, followed by second-strand cDNA synthesis. The synthesized cDNA was subjected to end repair, 3' adenylation, and adaptor ligation to complete the library. The dsDNA libraries were heat-denatured and circularized by the splint oligo sequence to generate a single strand circle DNA, followed by rolling circle replication to create DNA nanoballs. Sequencing was performed on an MGI DNBSEQ-T7 platform (BGI, China) using 150-bp paired-end reads. Fastq files were aligned to the GRCh38/hg38 genome using STAR. Gene counting was computed using STAR quantMode as GeneCounts. The outlier mRNA expression analysis was performed using OUTRIDER. Outliers were detected by comparing the mitochondrial disease cases that underwent RNA-seq analysis on the same platform.

Oxygen consumption rate measurement

Oxygen consumption rate (OCR) was using a with Seahorse XF96 extracellular flux analyzer (Agilent Technologies, USA). Patient skin-derived fibroblasts as well as control fibroblasts were cultured in Dulbefcco's modified Eagle's Medium (DMEM) with fetal bovine serum and 1% penicillin and streptomycin. The cells were seeded at 2×10^4 cells/well in a 96-well plate at 37 °C and 5% CO_2. After 24 h of culture, the medium was changed for 25 mM glucose or 10 mM galactose- containing DMEM for 1 h. Subsequently, the medium was replaced with an oligomycin (2 μM), carbonyl cyanide 4-(trifluoromethoxy) phenylhydrazone (FCCP, 0.4 μM), and rotenone (1 μM)-containing medium to control the ATP synthesis of the cells to determine the maximal rate of oxygen consumption[44].

Measurement of enzyme activity in oxidative phosphorylation

Oxidative phosphorylation enzyme activities in the fibroblasts were measured using a Cary 300 spectrophotometer (Agilent Technologies, USA) following the manufacturer's instructions[45]. The enzyme activities of each complex were presented as the percentage of normal control mean relative to the appropriate reference enzyme activities, such as that of citrate synthase (CS).

Western blotting

To prepare the total cell lysate, cell pellets were lysed with 1× RIPA buffer (08714-04, Nacalai Tesque, Japan) and kept on ice for 15 min. They were then centrifuged at 10,000 g for 10 min at 4 °C and the supernatants were collected. Protein concentration was determined using the BCA Protein Assay Kit (Thermo Fisher Scientific, USA) according to the manufacturer's instructions. Prepared samples were denatured for 5 min at 95 °C and separated by SDS-PAGE on a 5–20% gradient polyacrylamide gel (EHR-T520L, ATTO, Japan) with the Prestained XL-Ladder Broad (SP-2120, Pharma Foods International, Japan) as a molecular marker. Proteins were transferred to a PVDF membrane and subjected to western blotting. The NAXE (HPA048164, Sigma-Aldrich, USA) and beta-actin (A5441, Sigma-Aldrich, USA) primary antibodies were used.

Repeat-primed PCR (RP-PCR)

RP-PCR was carried out as follows: 10 μl of ExPremier 2× premix (TaKaRaBio, Japan), 1 μl of 5′ 6-FAM_reverse primer (10 uM), 1 μl of RP_END primer (10 uM), 1 μl of GGGCC_fwd primer (3.2 uM), 20 ng of template genomic DNA, nuclease-free water up to 20 μl of the total volume in a PCR-tube per sample. Primers were synthesized by Integrated DNA Technologies (IDT, USA) in HPLC purification grade (5′ 6-FAM_reverse primer) in 100 nmol scale or desalt grade (RP_END and GGGCC_fwd primers). The thermal cycles were as follows: 94 °C for 1 min, [98°C for 10 s, 60 °C for 15 s, 68 °C for 1 min] × 35 cycles, 68 °C for 1 min, 4 °C until subsequent dilution for fragment analysis. We used VeritiPro Thermal cycler or GeneAmp PCR system 9700 (Thermo Fisher Scientific, USA). Fragment analysis was conducted by adding HiDi-formamide and GeneScan 600 LIZ dye Size Standard v2.0 (Thermo Fisher Scientific, USA), followed by electrophoresis on GeneticAnalyzer 3500 (Thermo Fisher Scientific, USA). Data analysis was performed using the GeneMapper software (Thermo Fisher Scientific, USA). Primer sequences are described in Supplementary Table 6.

PacBio Sequel2 whole genome long-read sequencing and data analysis

Long DNA purified as described above were then sheared at 10-20 kb with Megaruptor 2 (Diagenode, USA) and further used for library construction using the SMARTbell prep kit 3.0 (Pacific bioscience, USA) following the manual's instructions. The libraries were then sequenced on a PacBio Sequel2 sequencer on 2 single molecule realtime (SMRT) sequencing cells on the proband and 1 on the mother sample. Acquired raw data (subreads.bam) were then transformed into circular consensus (ccs) reads.bam by pbccs 6.4.0 with or without the kinetics information. For the evaluation of CpG methylation status, primrose was used for processing the ccs reads file (.bam). The ccs reads were then mapped against the GRCh38 reference genome. To evaluate methylation in the internal sequences of the repeat stretch, the reads were further mapped against a synthetic sequence derived from pure 244 GGGCC repeats and contiguous sequences. Visual inspection was carried out using IGV. Methylation level at each CpG site was calculated as probability (0 to 100%) and illustrated in blue (0 to 49%) or red (50% to 100%) in IGV. For repeat internal sequence analysis, tandem repeat finder[46] was employed. PacBio RepeatAnalysisTools was used to draw the waterfall plot (see Supplementary Note 1).

PacBio Sequel2 amplicon sequencing of the *NAXE* region

We used ultralong PCR covering *NAXE* locus combined with PacBio Sequel2 amplicon sequencing to screen for the candidate pathogenic variants leading to *NAXE*-related mitochondrial encephalopathy as well as variations in the GGGCC repeat region as candidate risk factors. We aimed to screen 314 samples (Group1) from 461 mitochondrial encephalopathy patients initially examined by RP-PCR. In addition, we added four samples (the proband (Pt2359), older brother, mother, and father) of the *NAXE* repeat expansion pedigree (Family A), another case (Pt2659) with *NAXE*-related mitochondrial encephalopathy caused by compound heterozygous variants (missense and splicing; see also the Subjects section), and HapMap sample GM12878 as the control. As a result, 277 out of the 314 Group1 samples, four members of *NAXE* repeat expanded pedigree, Pt2659, and GM12878 were successfully amplified on either or both of the two 13 to 14 kb-long ultralong PCR amplicons covering the *NAXE* locus, whereas the other samples were not amplifiable, probably due to slight disintegration of genomic DNA during DNA preservation. A total of 242 samples with 10 or more reads covering the amplicon regions were further analyzed.

We conducted amplicon sequencing using PacBio Sequel2 long-read sequencing with the PacBio M13 targeted amplicon sequencing protocol. This protocol utilizes two-step PCR to add M13 primer sequences in the first PCR and barcode sequences at both ends in the second PCR.

The following amplicons were designed to cover all the gene regions of *NAXE* based on entry of ENSG00000163382 (Chr1: 156,591,756-156,609,507 on GRCh38.p14): Amplicon 1: chr1:156586942-156601027, 14086 bp, Amplicon 2: chr1:156595861-156609646, 13786 bp.

The first PCR was carried out as follows: ExPremier PCR enzyme (TaKaRaBio, Japan) was used with 10 μl of 2× ExPremier premix, final 0.25 μM of each primer, and 20 ng of template genomic DNA in a total

reaction volume of 20 µl, with the following thermal cycles: 94 °C 1 min, [98 °C 10 s, 68°C 7 min] × 30 cycles.

After purification, second step PCR was carried out to add barcodes on both sides using M13_barcode primers as second PCR primers. Each of the M13_bcXXXX_F and _R primers was used to comprise different 16 × 24 (= 384) sets of primer pairs. The second PCR was carried out as follows: ExPremier PCR enzyme (TaKaRaBio, Japan) was used with 10 µl of 2× ExPremier premix, final 0.25 µM of each primer and 1 ng of template DNA in a total reaction volume of 20 µl, with the following thermal cycles: 94 °C 1 min, [98 °C 10 s, 60 °C 15 s, 68 °C 7 min] × 8 cycles (see Supplementary Table 7 and 8 for information on first PCR and second PCR primers, respectively).

After the second PCR, the PCR products were pooled, purified, repaired, A-tailed, followed by adapter ligation and cleaning up, after which they were nuclease-treated and cleaned up following the PacBio SMARTbell prep kit 3.0 user instruction manual. The completed libraries were pooled and sequenced on 1 single molecule real time (SMRT) cell with the PacBio Sequel2 sequencer (Pacific bioscience, USA). For data analysis, circular consensus reads were produced with the pbccs 6.4.0 tool, demultiplexed, and mapped against GRCh38 using pbmm2. For the analysis of structural variants, sniffles2 was used[47]. For the analysis of SNVs and small indels, DeepVariant was used, and the variants were annotated using snpEff[48].

For single nucleotide variations and small indels, we used DeepVariant[49] for variant calling as well as visual inspection by IGV specifically on variants in the GGGCC repeat region of *NAXE*. We selected the variants with one or more of the following characteristics: 1) "MODARATE" or "HIGH" impact by snpEff annotation, 2) splice region by snpEff annotation, and 3) Delta score of 0.1 or higher in SpliceAI, and 4) variants affecting the *NAXE* GGGCC repeat sequence. Only variants with allele frequencies of 0.1 or lower in ToMMo54KJPN were considered for those meeting the criteria of 1), 2), and 3). For functional prediction by bioinformatics tools, CADD phred scale score GRCh38, delta score of SpliceAI, and ClinVar classification (if registered) were described based on the Gnomad database (if not registered, hand search by CADD GRCh38-v1.7 Web resources was used). Allele frequencies in public databases (ToMMo54KJPN and Gnomad v4.0) were described. To statistically compare the frequencies of variations in the GGGCC repeat region, chi-square test of independence was conducted by python scripts with significance threshold of $p < 0.05$.

Southern blotting

Probe design: Probes were designed to be >200 bp in length, having no off-target sequence in the human genome by blastn and being within an appropriate fragment, using restriction enzymes (*Mfe*I at chr1:156589300 and *Bam*HI at chr1:156563563). As a result, the following two probes were selected: p2_9F/R: chr1:156591431-156591637 (GRCh38), p3_9F/R: chr1:156592713-156592960 (GRCh38) (For primers for probe preparation, see Supplementary Table 9).

Template preparation: Artificially synthesized respective DNA fragments were cloned in pUC vectors and midi-prepped (Eurofins Japan, Japan). Midi-prepped plasmids were used as templates for probe preparation. PCR DIG Probe Synthesis Kit (Sigma-Aldrich Japan, Japan, Cat. No. 11 636 090 910) was used for DIG-probe preparation from 10 pg of template for each probe following the user instruction manual. PCR products were confirmed by agarose electrophoresis in 1×TBE buffer.

Restriction enzyme digestion, electrophoresis, transfer to membrane, and hybridization: From each sample, 1 µg of genomic DNA was digested by *Bam*HI (NEB) and *Mfe*I-HiFi (NEB) in rCutSmart buffer (NEB) at 37 °C for 2 h and electrophoresed in 1×TAE agarose (0.8%) gel made with SeaKem LE agarose (Lonza, Switzerland) of 14 cm length × 12 cm width format. As molecular markers, 1 kb DNA ladder PLUS (Nippon Genetics, Japan) and DIG-labeled marker VII (Sigma-Aldrich Japan, Japan) were used. After 35 V × 18 h of electrophoresis at room temperature, the gel was taken out and soaked in 0.25 M HCl 200 ml for 40 min for depurination. After two washes in 500 ml MilliQ (Merck-Millipore, USA) water for 3 min each, the gel was soaked in 250 ml of denaturation solution (0.5 M NaOH, 1.5 M NaCl) twice for 15 min each. The gel was then washed with 500 ml of MilliQ water for 5 min and soaked in 250 ml of neutralization solution (0.5 M Tris-Cl pH 7.5, 1.5 M NaCl) for 30 min. For blotting, gravity-assisted blotter (G capillary blotter C set, TAITEC, Japan, cat 0014675-000) was used for overnight blotting at room temperature using 10× SSC buffer to transfer to a positively charged nylon membrane (Sigma-Aldrich, USA). The membrane was then washed in 2× SSC buffer and dried at room temperature. UV crosslinking was conducted for irradiating 120,000 µJ of ultraviolet (UV) light (UV crosslinker CL1000, UVP (Analytik Jena, USA) on the membrane. The membrane was rinsed with 2× SSC and prehybridized in DIG Easy hyb buffer (Sigma-Aldrich, USA) for more than 3 h at 68 °C, followed by hybridization in probes in DIG Easy hyb buffer for 18 h at 42°C. Subsequently, washes with 100 ml of 2× SSC and 0.1% SDS for 15 min twice, 100 ml of 0.5× SSC with 0.1% SDS for 15 min once, and 100 ml of 0.15× SSC with 0.1% SDS for 15 min once were conducted in a hybridization bottle in a rotating incubator set to 68 °C. For detection of hybridized fragments, the membrane was taken out from the hybridization bottle, placed in a hybridization bag, and washed with 10 ml of washing buffer (0.1 M maleic acid, 0.15 M NaCl, pH 7.5, 0.3% Tween20). The buffer was discarded, and 10 ml of 1× blocking solution were used for blocking for 2 h at room temperature. The blocking solution was made from 10× blocking solution containing a blocking reagent (Sigma-Aldrich, USA, cat 11096176001) 10 g in 100 ml of 0.1 M maleic acid, and 0.15 M NaCl; pH 7.5; the mixture was stirred with heating, autoclaved, aliquoted in 15 ml centrifuge tubes, and preserved at -20 °C before use. Anti-DIG-AP Fab fragments (Sigma-Aldrich, USA cat 11093274910) were diluted 1:10,000 in 1× blocking buffer for immunodetection. After 30 min of mild shaking in the antibody solution at room temperature, the membrane was washed three times with 10 ml of washing buffer for 15 min each and rinsed with a detection buffer (0.1 M Tris-HCl, 0.1 M NaCl, pH 9.5) twice. Finally, ready-to-use CDP-star (Sigma-Aldrich, USA, cat 12041677001) was used for chemiluminescence. Images were obtained by FusionSolo S (Vilbert Lourmet, France). The raw image was linearly rescaled between intensities between 4 and 4000 in the software FusionCapt Advance in the gel imager.

NET-CAGE

1×10^7 cells per sample were used for the extraction of nascently transcribed RNA. Pelleted cells were dissolved in 1 ml of a cell lysis buffer (1527 µl of EZ lysis buffer (Sigma-Aldrich, USA), 40 µl of 1 mM α-amanitin (Fujifilm, Japan), 32 µl of 50× cOmplete protease inhibitor 50× (Sigma-Aldrich, USA), and 1 µl of 20 U/ul SUPERase•In (Thermofisher, USA) for preparation of 1600 µl of cell lysis buffer), incubated on ice for 10 min, and centrifuged for 800 × g 5 min at 4 °C. The pellets were dissolved in 600 µl of the cell lysis buffer and centrifuged for 800 × g 5 min at 4 °C. Nuclear lysis buffer (200 µl; final concentrations of 1% NP-40, 20 mM HEPES, 300 mM NaCl, 2 M urea, 0.2 mM EDTA, 1 mM DTT, 25 µM α-amanitin, 1× cOmplete protease inhibitor, and 20 U SUPERase in nuclease-free water) was added to the pellet and pipet mixed, incubated for 10 min on ice, and centrifuged for 3000 × g 4 min at 4 °C. The pellets were then suspended in 100 µl of nuclease lysis buffer and centrifuged for 3000 × g 2 min at 4 °C, followed by the addition of 50 µl of DNase I mix (5 µl of ×10 DNase buffer (Thermofisher, USA), 5 µl of DNase I (Thermofisher, USA), and 1 µl of 20 U/µl SUPERase•In, and 39 µl of nuclease free water in total volume of 50 µl per sample) and mixed with ThermoMixer (Thermofisher, USA) at 1400 rpm, 30 to 90 min (until pellet was solubilized) at 37 °C with gentle agitation with pipetting. Finally, 600 µl of RLT buffer (QIAGEN, Germany) were added and gently mixed. The resulting solution contained enriched nascently transcribed RNAs (nascent RNA) and was further utilized for the construction of the CAGE library. For the preparation of single strand CAGE library on nascent RNA (3 to 5 µg of nascent RNA per library), a previously described protocol was used (see details in Supplementary Note 1). Libraries were sequenced on Illumina HiSeq 2500 (Illumina, USA) using the 50 bp single-end mode. The sequenced reads were then filtered, mapped against

the GRCh38 genome, and normalized using MOIRAI[50] and in-house scripts.

Code availability

The codes used in this study are available from the corresponding author upon reasonable request.

Received: 19 March 2024; Accepted: 16 September 2024;
Published online: 25 October 2024

References

1. DiMauro, S. & Schon, E. A. Mitochondrial respiratory-chain diseases. *N. Engl. J. Med.* **348**, 2656–2668 (2003).
2. Ng, Y. S. & Turnbull, D. M. Mitochondrial disease: genetics and management. *J. Neurol.* **263**, 179–191 (2016).
3. Skladal, D., Halliday, J. & Thorburn, D. R. Minimum birth prevalence of mitochondrial respiratory chain disorders in children. *Brain* **126**, 1905–1912 (2003).
4. Schlieben, L. D. & Prokisch, H. The Dimensions of Primary Mitochondrial Disorders. *Front. Cell Dev. Biol.* **8**, 600079 (2020).
5. Lynch, D. R. & Farmer, G. Mitochondrial and metabolic dysfunction in Friedreich ataxia: update on pathophysiological relevance and clinical interventions. *Neuronal Signal.* **5**, NS20200093 (2021).
6. Giménez-Bejarano, A., Alegre-Cortés, E., Yakhine-Diop, S. M. S., Gómez-Suaga, P. & Fuentes, J. M. Mitochondrial Dysfunction in Repeat Expansion Diseases. *Antioxid. (Basel)* **12**, 1593 (2023).
7. Depienne, C. & Mandel, J. L. 30 years of repeat expansion disorders: What have we learned and what are the remaining challenges? *Am. J. Hum. Genet.* **108**, 764–785 (2021).
8. Logsdon, G. A., Vollger, M. R. & Eichler, E. E. Long-read human genome sequencing and its applications. *Nat. Rev. Genet.* **21**, 597–614 (2020).
9. Kremer, L. S. et al. NAXE Mutations Disrupt the Cellular NAD(P)HX Repair System and Cause a Lethal Neurometabolic Disorder of Early Childhood. *Am. J. Hum. Genet.* **99**, 894–902 (2016).
10. Manor, J. et al. NAXE deficiency: A neurometabolic disorder of NAD(P) HX repair amenable for metabolic correction. *Mol. Genet. Metab.* **136**, 101–110 (2022).
11. Chiu, L. W. et al. NAXE gene mutation-related progressive encephalopathy: A case report and literature review. *Med. (Baltim.)* **100**, e27548 (2021).
12. Incecik, F. & Ceylaner, S. Early-onset progressive encephalopathy associated with NAXE gene variants: a case report of a Turkish child. *Acta Neurol. Belg.* **120**, 733–735 (2020).
13. Spiegel, R., Shaag, A., Shalev, S. & Elpeleg, O. Homozygous mutation in the APOA1BP is associated with a lethal infantile leukoencephalopathy. *Neurogenetics* **17**, 187–190 (2016).
14. Ding, L. et al. De novo mutation of NAXE (APOAIBP)-related early-onset progressive encephalopathy with brain edema and/or leukoencephalopathy-1: A case report. *World J. Clin. Cases* **11**, 3340–3350 (2023).
15. Yu, D., Zhao, F. M., Cai, X. T., Zhou, H. & Cheng, Y. [Clinical and genetic features of early-onset progressive encephalopathy associated with NAXE gene mutations]. *Zhongguo Dang Dai Er Ke Za Zhi* **20**, 524–258 (2018).
16. Trinh, J. et al. Novel NAXE variants as a cause for neurometabolic disorder: implications for treatment. *J. Neurol.* **267**, 770–782 (2020).
17. Mohammadi, P., Heidari, M., Ashrafi, M. R., Mahdieh, N. & Garshasbi, M. A novel homozygous missense variant in the NAXE gene in an Iranian family with progressive encephalopathy with brain edema and leukoencephalopathy. *Acta Neurol. Belg.* **122**, 1201–1210 (2022).
18. Marbaix, A. Y. et al. Occurrence and subcellular distribution of the NADPHX repair system in mammals. *Biochem. J.* **460**, 49–58 (2014).
19. Cui, Y. et al. A genome-wide spectrum of tandem repeat expansions in 338,963 humans. *Cell* **187**, 2336–2341.e2335 (2024).
20. Nakka, P. et al. Characterization of Prevalence and Health Consequences of Uniparental Disomy in Four Million Individuals from the General Population. *Am. J. Hum. Genet.* **105**, 921–932 (2019).
21. Brechtmann, F. et al. OUTRIDER: A Statistical Method for Detecting Aberrantly Expressed Genes in RNA Sequencing Data. *Am. J. Hum. Genet.* **103**, 907–917 (2018).
22. Parissis, D., Dimitriou, M. & Ioannidis, P. Genetic causes of acute encephalopathy in adults: beyond inherited metabolic and epileptic disorders. *Neurol. Sci.* **43**, 1617–1626 (2022).
23. Yang, L. et al. Clinical features and underlying genetic causes in neonatal encephalopathy: A large cohort study. *Clin. Genet.* **98**, 365–373 (2020).
24. Quinodoz, M. et al. AutoMap is a high performance homozygosity mapping tool using next-generation sequencing data. *Nat. Commun.* **12**, 518 (2021).
25. Dolzhenko, E. et al. ExpansionHunter: a sequence-graph-based tool to analyze variation in short tandem repeat regions. *Bioinformatics* **35**, 4754–4756 (2019).
26. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
27. Dolzhenko, E. et al. ExpansionHunter Denovo: a computational method for locating known and novel repeat expansions in short-read sequencing data. *Genome Biol.* **21**, 102 (2020).
28. Dashnow, H. et al. STRling: a k-mer counting approach that detects short tandem repeat expansions at known and novel loci. *Genome Biol.* **23**, 257 (2022).
29. Tadaka, S. et al. jMorp: Japanese Multi-Omics Reference Panel update report 2023. *Nucleic Acids Res.* **52**, D622–D632 (2024).
30. Karczewski, K. J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
31. Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J. & Kircher, M. CADD: predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Res.* **47**, D886–D894 (2018).
32. Jaganathan, K. et al. Predicting Splicing from Primary Sequence with Deep Learning. *Cell* **176**, 535–548.e524 (2019).
33. Hirabayashi, S. et al. NET-CAGE characterizes the dynamics and topology of human transcribed cis-regulatory elements. *Nat. Genet.* **51**, 1369–1379 (2019).
34. Richards, S. et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).
35. Salcedo-Arellano, M. J., Dufour, B., McLennan, Y., Martinez-Cerdeno, V. & Hagerman, R. Fragile X syndrome and associated disorders: Clinical aspects and pathology. *Neurobiol. Dis.* **136**, 104740 (2020).
36. Groh, M., Lufino, M. M., Wade-Martins, R. & Gromak, N. R-loops associated with triplet repeat expansions promote gene silencing in Friedreich ataxia and fragile X syndrome. *PLoS Genet.* **10**, e1004318 (2014).
37. Asamitsu, S. et al. Perspectives for Applying G-Quadruplex Structures in Neurobiology and Neuropharmacology. *Int. J. Mol. Sci.* **20**, 2884 (2019).
38. Teng, Y., Zhu, M. & Qiu, Z. G-Quadruplexes in Repeat Expansion Disorders. *Int. J. Mol. Sci.* **24**, 2375 (2023).
39. Zuker, M. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31**, 3406–3415 (2003).
40. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
41. McKenna, A. et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
42. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).

43. Wu, Y., Tian, L., Pirastu, M., Stambolian, D. & Li, H. MATCHCLIP: locate precise breakpoints for copy number variation using CIGAR string by matching soft clipped reads. *Front. Genet.* **4**, 157 (2013).
44. Shimura, M. et al. Effects of 5-aminolevulinic acid and sodium ferrous citrate on fibroblasts from individuals with mitochondrial diseases. *Sci. Rep.* **9**, 10549 (2019).
45. Kirby, D. M., Thorburn, D. R., Turnbull, D. M. & Taylor, R. W. Biochemical assays of respiratory chain complex activity. *Methods Cell Biol.* **80**, 93–119 (2007).
46. Benson, G. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.* **27**, 573–580 (1999).
47. Smolka, M. et al. Detection of mosaic and population-level structural variants with Sniffles2. *Nat. Biotechnol.*, https://doi.org/10.1038/s41587-023-02024-y (2024).
48. Cingolani, P. et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly. (Austin)* **6**, 80–92 (2012).
49. Poplin, R. et al. A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.* **36**, 983–987 (2018).
50. Hasegawa, A., Daub, C., Carninci, P., Hayashizaki, Y. & Lassmann, T. MOIRAI: a compact workflow system for CAGE analysis. *BMC Bioinforma.* **15**, 144 (2014).

Acknowledgements

This work was supported by Grant-in-Aid from the Japan Agency for Medical Research and Development (AMED) Practical Research Project for Rare/Intractable Diseases (JP22ek0109485 (P.C.), JP23ek0109625 (Y.O. and K.M.), JP23ek0109672 (Y.O.)), JSPS KAKENHI (JP23H00424 (Y.O.), JP22K15950 (Y.Y.)), Health Labour Sciences Research Grant (JP23FC1034 (Y.O.)), Program for Promoting Platform of Genomics based Drug Discovery from AMED (JP23kk0305024 (Y.O.)), Mitochondrial preemptive medicine, Moonshot Research and Development Program (JP23zf0127001 (K.O. and Y.O.)), and Matching Fund Subsidy for Private Universities from MEXT (Y.O.). We thank Dr. Kiyohiro Kin (Hyogo Prefectural Amagasaki General Medical Center) for providing clinical information. We appreciate the administrative support of Yumiko Yamamoto, Miyuki Abe, and Hiroo Inaba, technical assistance from Yuki Yasuoka, Chitose Takahashi, Riichiroh Manabe, Nozomi Moritsugu, Kayuri Kadoya, and Tsugumi Kawashima, and scientific comments from Norihito Hayatsu, Ryota Teramoto, and Yuuri Yasuoka in Riken. We thank Kanako Oyama (Juntendo University), Mari Ohiwa (Chiba Children's Hospital), Rie Takeuchi (Saitama Medical University), Kaoru Kaida, and Akiko Oguchi (Riken) for technical assistance. We thank the Laboratory of Molecular and Biochemical Research, Biomedical Research Core Facilities, Juntendo University Graduate School of Medicine, for technical assistance. We also thank IMS center for Integrative Medical Sciences Genome Platform service in Riken for sequencing.

Author contributions

K.O.: conceptualized the study, performed the genetic analyses including the development of RP-PCR, carried out the dry analyses (including all the long-read analyses), and wrote the manuscript. Y.Y.: biochemically analyzed the patient, carried out the genetic study along with K.O., and wrote the manuscript along with K.O. Y Oyazato, A.N., T.F., M.S., Y.S., A.O., K.M.: carried out the clinical study. K.R.N., Y.K.: carried out the genetic study with K.O. and Y.Y. and performed the dry analysis of short read sequencing. S.N., W.S.: performed long-read sequencing. M.T.: carried out data processing along with K.O. M.F., T.H.: carried out the genetic study. H.K., Y.M.: carried out NET-CAGE. P.C.: supervised the study. Y Okazaki: conceptualized and supervised the study.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41525-024-00429-5.

Correspondence and requests for materials should be addressed to Yasushi Okazaki.

[1]Laboratory for Comprehensive Genomic Analysis, RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan. [2]Diagnostics and Therapeutics of Intractable Diseases, Intractable Disease Research Center, Graduate School of Medicine, Juntendo University, 2-1-1 Hongo, Bunkyo-ku, Tokyo 113-8421, Japan. [3]Department of Pediatrics, Kakogawa Central City Hospital, 439 Hon-machi, Kakogawa-cho, Kakogawa, Hyogo 675-8611, Japan. [4]Department of Life Science, Faculty of Science and Engineering, Kindai University, 3-4-1 Kowakae, Higashi-Osaka, Osaka 577-8502, Japan. [5]Center for Medical Genetics and Department of Metabolism, Chiba Children's Hospital, 579-1 Hetacho, Midori-ku, Chiba, Chiba 266-0007, Japan. [6]Department of Pediatrics, Faculty of Medicine, Juntendo University, 3-1-3 Hongo, Bunkyo-ku, Tokyo 113-8431, Japan. [7]Laboratory for Symbiotic Microbiome Sciences, RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan. [8]RIKEN-IFOM Joint Laboratory for Cancer Genomics, RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan. [9]Laboratory for Transcriptome Technology, RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan. [10]Human Technopole, via Rita Levi Montalcini 1, Milan 20157, Italy. [11]Department of Clinical Genomics and Pediatrics, Faculty of Medicine, Saitama Medical University, 38 Morohongo, Moroyama, Saitama 350-0495, Japan. [12]These authors contributed equally: Kokoro Ozaki, Yukiko Yatsuka. ✉e-mail: ya-okazaki@juntendo.ac.jp