Genetics in Medicine (2026) 28, 101620

ELSEVIER

Genetics in Medicine
An Official Journal of the ACMG

www.journals.elsevier.com/genetics-in-medicine

ARTICLE

Charting the phenotypic landscape of mitochondrial diseases through a systematic evaluation of pathogenic mitochondrial DNA and nuclear gene variants

Thiloka Ratnaike[1,2,3,*], Siddharth Ramanan[4], Nour Elkhateeb[5], Ramya Narayanan[6], Jenny Yang[6,7], Eszter Sara Arany[6], Manya Mirchandani[8], Rachael Piper[8], Katherine Schon[5,7], M. Eren Kule[6,9], Christopher Gilmartin[10], Angela Lochmüller[6], Emogene Shaw[11], Rita Horváth[7], Patrick F. Chinnery[7]

ARTICLE INFO

Article history:
Received 5 May 2025
Received in revised form
7 October 2025
Accepted 9 October 2025
Available online 24 October 2025

Keywords:
HPO
Mitochondrial disease
Phenotype similarity
Rare disease
UMAP

ABSTRACT

Purpose: Primary mitochondrial diseases (PMD) arise from variants in the mitochondrial or nuclear genomes. Phenotype-based recognition of specific PMD genotypes remains difficult, prolonging the diagnostic odyssey. We expanded the *MitoPhen* database to characterize phenotypic variation across PMD more systematically.

Methods: Individual-level data on mitochondrial DNA disorders, nuclear-encoded mitochondrial diseases, and single large-scale mitochondrial DNA deletions were manually curated with Human Phenotype Ontology (HPO) terms to produce *MitoPhen v2*. Principal-component analysis summarized system-level abnormalities; HPO-level enrichment and mean phenotype-similarity scores were then used to distinguish common PMD genotypes.

Results: *MitoPhen v2* adds 3940 individuals to the original release, now encompassing 1597 publications, 10,626 individuals, and 117 genotypes. Among 7586 affected cases, 72,861 HPO terms were recorded. Principal-component analysis revealed 6 phenotype dimensions capturing most system-level variance. At the HPO level, we observed genotype-specific enrichments and identified 111 gene-phenotype links absent from the current HPO database. Using *MT-TL1*, single large-scale mitochondrial DNA deletions, and *POLG* as exemplars, phenotype-similarity scores reliably separated individuals with these genotypes from those without.

Conclusion: *MitoPhen v2* enabled systematic, genotype-aware analysis of heterogeneous PMD phenotypes and highlighted the diagnostic value of structured, individual-level data. Phenotype-similarity metrics from such data sets can refine variant interpretation in large rare-disease cohorts and provide a transferable framework for other phenotypically complex genetic disorders.

The Article Publishing Charge (APC) for this article was paid by University of Cambridge.

Thiloka Ratnaike and Siddharth Ramanan are joint first authors who contributed equally to this work.

*Correspondence and requests for materials should be addressed to Thiloka Ratnaike, School of Physiology, Pharmacology and Neuroscience, University of Bristol, Bristol BS8 1TD, United Kingdom. *Email address:* sh25676@bristol.ac.uk

Affiliations are at the end of the document.

doi: https://doi.org/10.1016/j.gim.2025.101620

Introduction

Primary mitochondrial diseases (PMD) are among the most common inherited metabolic disorders and are estimated to affect 1 in 4300 people, manifesting across the age spectrum.[1] The clinical heterogeneity of these conditions is reflective of the complex genetic control of mitochondrial maintenance and function. Mitochondria are the only components of the human body that are under dual genetic control. Nuclear genes and mitochondrial DNA (mtDNA) contribute to the effective functioning of the oxidative phosphorylation system housed within the inner mitochondrial membrane, which generates the bulk of adenosine triphosphate, required for cellular processes.[2] Nuclear genes such as *POLG* (HGNC:9179, OMIM *174763), *TWNK* (HGNC:1160, OMIM *606075) and *RRM2B* (HGNC: 17296, OMIM *604712) are important for mtDNA replication and maintenance; others are required for mitochondrial fusion and fission, such as *OPA1* (HGNC:8140, OMIM *605290). Mitochondrial aminoacyl transfer RNA (tRNA) synthetases are essential for the charging of the specific mitochondrial tRNA with their cognate amino acid, whereas *SPG7* (HGNC:11237, OMIM *602783) regulates ribosome assembly within mitochondria. Most of the structural subunits of the oxidative phosphorylation system are encoded by nuclear genes.[3] In contrast, there are 37 mtDNA genes—13 of which encode subunits of OXPHOS, 22 which encode tRNA molecules, and 2 that encode ribosomal RNA molecules. Wild-type and mutant mtDNA molecules can coexist in a state of heteroplasmy, which is known to give rise to cellular and tissue dysfunction once a certain threshold of heteroplasmy is reached, at which the mitochondria are likely to be ineffective at adenosine triphosphate production. This threshold can vary between tissues and between variants of the mtDNA.[4] Additionally, certain pathogenic variants exist in homoplasmic states in which 100% of the mtDNA contains the variant; however, the penetrance of disease can be low and tissue specific, for example, in Leber's hereditary optic neuropathy (LHON, OMIM 535000).[5]

Affected individuals with PMDs can present with a large array of symptoms varying between individuals with the same disorder and overlapping with other distinct conditions.[6] For example, siblings with the same genetic defect, despite sharing 50% genetics and a shared environment, can manifest with nonoverlapping symptoms.[7] Similarly, PMDs caused by nuclear gene variants are also clinically heterogeneous, and a main limitation to accurate diagnosis in large, rare disease sequencing projects is the interpretation of variants of uncertain significance (VUS). Typically, 30 to 40 rare variants are detected per human genome,[8] and in the context of a rare disease in which the phenotypic spectrum continues to evolve, it becomes challenging to confidently exclude VUS or consider their contribution to the individual's phenotype. The blurred boundaries within and between conditions contribute to challenges in accurate diagnosis, as well as in predicting disease progression and survival patterns, developing tailored prognostication plans, and understanding modifiers of disease phenotype.[9] A new approach to characterization is needed that leverages phenotypic heterogeneity to clinical advantage, captures the range and nature of phenotypes, and maps the landscape of clinical variation across disorders. Recent publications tackling challenges posed by interpretation of large quantities of phenotype data have utilized the human phenotype ontology (HPO), which enables the systematic quantification of clinical data through a structured language.[10,11] This has led to international collaborations to provide genetic diagnoses, through inclusion of phenotype similarity measures with the established HPO database, for individuals with rare diseases.[12]

Our previous proof-of-concept publication highlighted the value of a rich "MitoPhen" (Mitochondrial Phenotype) database of genotype and phenotype data in the form of HPO terms in mtDNA diseases caused mainly by single-nucleotide variants.[13] Therefore, we aimed to expand this database ("MitoPhen v2") by including single large-scale mtDNA deletions (SLSD) and nuclear genes associated with PMDs to not only gain an overall understanding about the phenotypes across different genotypes but also to understand finer details using HPO that can distinguish between genotypes of PMDs.

SLSD occur as de novo sporadic variants and individuals may present at different ages with variable symptoms, which may not be easily categorized into syndromes associated with SLSD.[14] Examples of such syndromes include chronic progressive external ophthalmoplegia (CPEO), Kearns-Sayre syndrome (KSS, OMIM 530000), which is a progressive multisystem syndrome that leads to death in early adulthood, and Pearson syndrome (OMIM 557000), which can be fatal because it is associated with bone marrow failure and refractory anemia in neonates and infants (however, survivors may go on to develop KSS).[15] A muscle biopsy that demonstrates cytochrome c oxidase deficiency and/or ragged-red fibers due to high SLSD heteroplasmy levels in these deficient fibers was essential for diagnosis in adult-onset SLSD-associated mitochondrial diseases.[6] However, genome sequencing technologies are now being developed to study the detection of SLSD using muscle and other tissues in the research setting.[16] It may be challenging to detect SLSD in blood after early childhood because of declining heteroplasmy levels with age.[17] However, SLSD can accumulate in postmitotic tissues over time or because of tissue-segregation of disease. Consequently, mtDNA analysis from muscle is often required from adults.[18] A comprehensive phenotypic reference data set to consider the likelihood of SLSD-associated PMDs can aid decision making in the context of obtaining invasive tissue samples and can improve the annotation of SLSD detected using genome sequencing.

Additionally, phenocopies of mitochondrial disorders complicate the diagnostic odyssey for affected individuals

and their families.[19,20] These reports show the need for robust, structured ongoing exploration of phenotypic spectrums associated with these autosomal dominant, autosomal recessive, X-linked, or de novo variants. For example, *POLG* (OMIM *174763) is one of the most prevalent monogenic causes of PMDs in different adult cohorts[1,21] and has a broad phenotypic spectrum. *POLG*-related diseases can present in the neonatal period with childhood myocerebrohepatopathy spectrum, in infancy as Alpers-Huttenlocher syndrome (OMIM 203700), which consists of explosive seizures, developmental regression, and liver failure, or later in childhood and early adulthood with different systemic presentations ranging from neurological, eye, and muscle phenotypes to gastrointestinal failure.[22] It is important to diagnose *POLG*-related diseases early to tailor treatment of seizures and ensure drugs such as sodium valproate are avoided as this is associated with significant mortality in these individuals.[23] Given the frequency of *POLG* variants found in European populations,[24] it is likely that these variants will be found in rare disease cohorts as VUS and having a systematic methodology by which to assess the phenotypic similarity would ensure further investigation of the *POLG* variants is tailored to the likelihood of pathogenicity.

To achieve a more comprehensive understanding of the phenotypic landscape of PMDs, we pooled 117 genotypes in a combined analysis of MitoPhen v2. Specifically, we investigated 3 questions: (1) "at system-level abnormalities, can we map the landscape of clinical variations between individuals from mtDNA, SLSD and nuclear gene groups?"; (2) "at the gene level, how phenotypically similar are clinically affected individuals"; and (3) "are there phenotypic and/or demographic features that could help classify one causative gene from other mitochondrial disease-associated genes, with relevance to clinical diagnostics?" To answer these questions, we used dimension reduction methodologies and focused phenotype similarity evaluations.

Materials and Methods

Data set

This study combined data on published individuals with mtDNA diseases within the original MitoPhen database (MitoPhen v1), with additional data on individuals with pathogenic or likely pathogenic mtDNA, nuclear gene, and SLSD mitochondrial diseases into the MitoPhen v2 database. Briefly, MitoPhen v1 is a curated database derived from 676 publications with clinical phenotypes captured as 26,348 HPO terms across 6688 published individuals with 89 pathogenic mtDNA variants.[13] The MitoPhen v2 data set, is an extension of the previous work with data collected on nuclear genes using literature reviews conducted in PubMed updated to June 1, 2022 using the search strategy: [Gene name] AND 'mutation', with a second search for [Gene name] AND 'clinical'. We used epidemiology data to decide on the frequent nuclear genes to focus our search strategy,[1] namely, *POLG*, *OPA1*, *TWNK*, *SPG7*, and *RRM2B*. We focused on additional nuclear genes encoding structural components and assembly genes of the respiratory chain complex by searching for each complex deficiency in PubMed. However, we did not carry out focused literature reviews on each gene associated with complex deficiencies because of time constraints. The Genomics England PanelApp high evidence ("green") gene list (https://panelapp.genomicsengland.co.uk/) was used to check genes associated with respiratory chain complex deficiencies.[25] Some genes on the PanelApp moderate evidence ("amber") gene list were included if there was at least 1 family described alongside functional work to evidence pathogenicity of the gene variants. Variants with likely pathogenic or pathogenic status within Varsome[26] and/or Franklin databases[27] were included. If there was a discrepancy between published literature and pathogenicity status in these databases, we performed further evaluations of the evidence for pathogenicity using American College of Medical Genetics guidelines[28] by (1) checking allele frequency using gnomAD v.3.1,[29] (2) aggregating evidence from other databases such as ClinVar[30] and the Human DNA Polymerase Gamma Mutation Database (https://tools.niehs.nih.gov/polg/), and (3) through expert consensus between coauthors T.R., N.E., K.S., and R.H. A similar search strategy was used to extract articles relevant to SLSD. Finally, we updated the mtDNA variants data set after searching for variants with 5 or fewer probands listed in MitoPhen v1 and added pathogenic mtDNA variants using MITOMAP "confirmed" variants, which have also been curated by the ClinGen Mitochondrial Disease Nuclear and Mitochondrial Variant Curation Expert Panel classification.[31,32] Individual-level data curation for MitoPhen v2 followed the same procedures as in MitoPhen.[13] However, we have now included age at death or follow-up, as well as the Human Genome Variation Society nomenclature with relevant reference sequences for the nuclear gene variants. Additional details on the data curation and quality control measures undertaken to ensure data collectors used similar approaches for mapping free text to HPO terms, are given in the Supplementary Methods. A full breakdown of the dataset by gene group is displayed in Figure 1.

Data preparation

Characterizing mitochondrial and nuclear gene disorders

Demographic and clinical characteristics for individuals with mtDNA, SLSD, and nuclear gene associated PMDs were computed. Frequencies for categorical variables (eg, sex, number of HPO terms per variant) and averages and ranges for continuous measures (eg, age, age of onset) were calculated. No statistical comparisons of variant-level group differences on demographic and clinical variables were conducted, given the large number of variants.

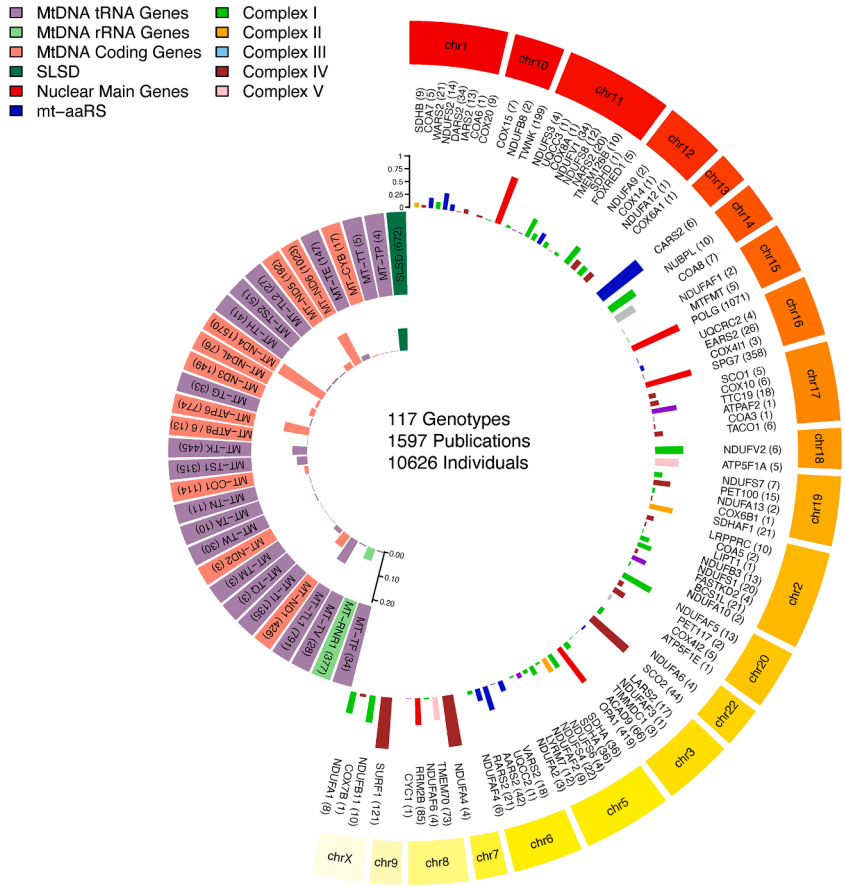

Figure 1 MitoPhen v2 data set, which has combined newly curated data with the previously published mtDNA disease data set. There are 117 genotype categories overall with individual-level data captured across 1597 articles. MtDNA genotypes are shown in the inner tracks, colored by gene characteristic—whether they are tRNA, rRNA, coding, or single large-scale deletions (SLSD). The bar plot corresponding to the mtDNA gene data set represents the proportions of individuals across mtDNA genotypes, which includes affected and unaffected individuals, the scale for proportion is shown on the left. Nuclear genotype data are displayed in the outer tracks, with genes grouped by chromosomes. The nuclear gene bar plots represent the proportion of individuals per chromosome (scale at the top), colored by the nuclear gene data set. Main genes (red bars) refer to the curations conducted on genes: *POLG*, *OPA1*, *TWNK*, *SPG7*, and *RRM2B*. Mitochondrial aminoacyl tRNA synthetase (mt-aaRS) genes (blue bars) refer to the 10 frequent genes from this group that were curated. The remaining nuclear genes are associated with complex I-V impairments.

Mapping genotype-phenotype relationships with system-level abnormalities

We created an "analytical data set" with the steps detailed below to investigate the HPO terms at the level of system-level abnormalities (SLA), this is summarized as a flow-chart in Supplemental Figure 1. The focus of the study was on symptomatic individuals:

- Retaining individuals with PMDs: of all the individuals in the MitoPhen v1 data set ($N = 6688$), we excluded (1) all asymptomatic individuals ($n = 2992$); (2) individuals with mtDNA m.4317A>G variant ($n = 2$), whose pathogenicity has been refuted[33]; (3) digenic variants ($n = 31$); (4) and those with no recorded HPO terms ($N = 14$), leading to a total of $N = 3649$ retained individuals with mtDNA diseases in the MitoPhen v1 data set. Of the new mtDNA disease individuals in the MitoPhen v2 data set ($N = 143$), we excluded asymptomatic individuals ($n = 41$) and retained a total $N = 102$ individuals with mtDNA diseases. All individuals with SLSD ($n = 672$) and nuclear gene variants ($n = 3123$) had at least 1 recorded HPO term. Ultimately, across both MitoPhen and MitoPhen v2 data sets, these steps led to a total dataset of $N = 7546$ included individuals from a total of 10,626 individuals.
- Reducing HPO terms and HPO-to-SLA transformation: to map the landscape of phenotypic heterogeneity, as indexed via the various HPO terms, the next step focused on tackling missing data and redundancies in HPO terms by transforming HPOs to system-level abnormality—an umbrella "ancestor term" capturing multiple HPOs together.[10] As a first step, we reduced the frequency of rarely occurring HPO terms (any HPO term that appeared <5 times in the MitoPhen v2 [n = 1847 HPO terms]) by replacing

it with its next relevant ancestor to account for noise in the data set with infrequent HPO terms that may not be related to the PMD. Then, we reduced the full set of HPO terms that applied to an individual by deriving a minimal, deduplicated set of HPO terms (using the *ontologyX* R package).[34] We then operated at the level of SLAs by deriving the corresponding SLA for each HPO term. For individuals whose phenotype encompassed multiple SLAs, we summarized proportions of HPOs per SLA, resulting in 21 SLAs that captured all HPO terms in the data set. In cases in which HPO terms had multiple parent SLAs ($n = 268$ HPO terms), an expert clinician (T.R.) verified the most applicable SLA to be carried forward for that term. After the SLA transformation steps, a total of $n = 21$ individuals were excluded because of the lack of HPO/SLA data after transformation.

- Identifying single-system mtDNA disorders: several individuals with mtDNA variants had a set of HPO terms that corresponded with a single-SLA, by virtue of their phenotype of having a single-system disorder, such as LHON. These individuals were labeled as having a "single-SLA" condition (ie, if all applicable HPO terms fell under 1 main SLA; Supplemental Table 1), a classification that was further verified by an expert clinician (T.R.). The remaining individuals with mtDNA diseases, whose HPO terms fell across multiple SLAs, were classed as "oligo-SLA" mtDNA disease. For single-SLA conditions, all applicable HPO terms for the relevant single-SLA were converted to a maximal proportion value of 1. For example, an individual with LHON had a proportion value of 1 on Eye SLA and 0 on all others, because they presented only with ocular-related complaints.

The above steps led to a total of 7525 individuals (1761 single-SLA mtDNA; 1975 oligo-SLA mtDNA; 671 SLSD; 3118 nuclear gene) with 21 SLAs. This "analytical data set" was carried forward to subsequent dimension reduction analyses.

To address the first aim, we conducted 3 complementary analyses to map the landscape of phenotypic presentations and its relationship to overlaps between different genetic groups. First, we used classical 2D multidimensional scaling (MDS) to visualize the clustering of the SLAs as a proxy to understand the relationships between phenotypic presentations of mtDNA, SLSD, and nuclear gene variants. Using a Spearman's decorrelation "distance" matrix of the SLA data, MDS leveraged the interfeature distances as a proxy of similarity/dissimilarity to provide insights into the similarity between SLAs in the data set[35] that we then interpreted based on the prevalence of each SLA in the entire cohort.

In the next step, we explored the concurrent phenotypic similarities and differences at the individual level using Uniform Manifold Approximation and Projection (UMAP). Unlike the MDS that mapped similarity between SLAs at the group-level, UMAP concurrently maps similarity between features and individuals[36,37] and is further explained in Supplemental Information. This approach has been used by previous studies for mapping the landscape of graded phenotypic variations in clinical disorders, including rare diseases.[9,38] As per recommended approaches, we initialized the UMAP with principal component analysis (PCA) of SLA data.[39] Three convergent methods were used to decide the optimal number of PCs to extract, after which a varimax-rotated PCA on the SLA data was performed (full details in Supplemental Information). The rotated components were then fed into the UMAP, and 2 key hyperparameters were tuned (number of neighbors that constrains the number of neighboring points [value of 20] and minimum distance that controls the distance between each point [value of 0.5]) to provide a good balance of local vs global structure. To briefly explain these hyperparameters, the "number of neighbors" hyperparameter controls how UMAP balances local vs global structure in the data by constraining the number of neighboring points. A lower "number of neighbors" value forces the UMAP to concentrate on a highly local structure (potentially causing a lack of focus on the big picture), whereas a larger value pushes points away from each other to focus on a global neighborhood at the cost of giving up finer details, analogous to the community structure of a village street versus a large residential community. The minimum distance parameter controls the minimum distance between the points in the low-dimensional space. Smaller values result in clumpier embeddings of smaller connected components, whereas larger values give an overarching view of the data at the cost of detailed topological structure. Both parameters together give a balanced view of global versus local structure of the data.[9,36] Detailed examples of the effects of combinations of various parameters on UMAP embeddings are presented in the UMAP explanatory vignettes: https://umap-learn.readthedocs.io/en/latest/parameters.html. The transformations underpinning UMAP can limit the direct correspondence between input and output data[37]; therefore, we projected various clinical variables (sex distribution, generation of individual relative to the proband, heteroplasmy from blood and muscle, clinical syndromes [as detailed in Supplemental Table 2], 25 most frequent genes in each group, and mitochondrial respiratory chain complex deficiency disorder labels) to the UMAP plots and generated an interactive graph in Plotly[40] for clinical dissemination (see Supplemental Information for Plotly figure).

Examining within-gene phenotypic similarity

To address the second aim, we examined the within-variant phenotypic similarity by computing a phenotype similarity index (PSI), similar to that computed in the MitoPhen v1 publication.[13] This analysis retained all the HPO data in the MitoPhen v2 database. Using the *ontologyX* and *ontologySimilarity* R packages[34] for each individual, we (1)

computed the similarity in their list of applicable HPO terms to the 5 most similar individuals within their variant group (excluding self-comparisons), (2) assigned the mean similarity value to this test individual, and (3) compiled this information into a data set of mean PSI (of HPO terms) for every individual. A high PSI score suggested high similarity in the "overall phenotype" (as assessed by available HPO terms) between the test person and others with the same variant. We visualized these PSI values for the 25 most frequent mtDNA and nuclear gene groups using violin plots, using *ggplot2*.[41] The HPO was used to compute pairwise phenotype similarities between probands using the *get_sim_grid* function from the *ontologySimilarity* package.[34] Nonredundant HPO terms were extracted per individual, and similarity was calculated across all probands to generate a similarity matrix. Genes with $\geq$2 associated probands were analyzed. For each gene, the statistical significance of within-gene phenotype similarity was assessed using the *get_sim_p* function, which compares observed similarity within the gene to a null distribution. To improve computational efficiency, the gene list was divided into chunks of 50 genes, and parallel processing was implemented using the *parallel* package.[42] The resulting *P* values were aggregated and ranked to identify genes with significant within-group phenotype similarity ($P < .05$). We also visualized the frequently occurring HPO terms within MitoPhen v2, seen in $\geq$5 individuals with the same gene, which were not seen in association with the respective genes within the latest version of the HPO database (downloaded on 26/01/2025 from https://hpo.jax.org). To assess for genetic category-specific HPO enrichment, we labeled the genotypes according to the 11 categories shown in Figure 1 and reshaped the individual-level HPO annotations into binary presence-absence data. Fisher's exact tests for each HPO term were performed in R, and significance was determined using false discovery rate (FDR) correction to identify HPO terms overrepresented in each genetic category compared with all others.

Classifying *POLG, MT-TL1*, and SLSD from other PMD genotypes using PSI and other characteristics

To address the third aim, we chose *POLG, MT-TL1* (HGNC:7490) and SLSD as examples of frequent genotypes across mtDNA and nuclear gene causes of PMDs with diverse phenotype presentations. We tested the ability for PSI, computed as per the step above, to differentiate the underlying genotype from other PMDs within MitoPhen v2. We calculated the PSI using the diagnosed individuals with *POLG*, *MT-TL1*, or SLSD separately as the reference data sets. We then explored the predictive utility of the above PSI values, and other variables (age at onset, sex, and inheritance pattern) in classifying *POLG* from non-*POLG* variants, *MT-TL1* from non-*MT-TL1* variants, and SLSD from non-SLSD genotypes using separate cross-validated logistic regressions. We performed an 80:20 train-test split and used 5-fold cross-validated logistic regression with receiver operator characteristic (ROC) area under the curve as the performance metric. Model classification accuracy was evaluated using a confusion matrix from the test fold. We used *caret*[43] and *pROC* R packages[44] to carry out the analyses and used *ggplot2*[41] for data visualization.

Phenotypic coclustering analysis was performed leveraging *POLG*, *MT-TL1*, and SLSD data sets retaining individuals with high PSI scores, to ensure high phenotypic similarity (see Supplemental Methods for details). Nonredundant HPO term sets were extracted, and terms occurring in fewer than 3 individuals were excluded. A co-occurrence matrix was generated, retaining terms shared by at least 5 individuals, and a weighted network graph was constructed. The Walktrap community detection algorithm[45] identified the 5 largest clusters, and the top-10 most frequent HPO terms per cluster were selected. A Sankey diagram[46] was used to visualize phenotypic relationships, with HPO terms categorized by top-level biological systems to enhance interpretability. The "Other" category captured less frequent terms to maintain proportional representation. R packages *networkD3* and *htmlwidgets* were used for these visualizations.[47]

Statistical analysis

All statistical analyses were conducted using R v4.4.1.

Results

Demographic and clinical characteristics

MitoPhen v2 contains 10,626 individuals: 6831 individuals with mtDNA variants (affected $N = 3798$), which includes the additional 143 individuals and data on 17 new mtDNA variants across 3 mt-genes; individuals with SLSD ($N = 672$), and nuclear gene variants across 86 genes ($N = 3123$). We reviewed 1305 separate published articles and included 912 articles with sufficient individual-level phenotype and genotype data. Therefore, overall, MitoPhen v2 now contains data from 1597 peer-reviewed articles in total (Figure 1).

Demographic and clinical features of all groups, per the "analytical data set," are displayed in Table 1 with accompanying frequency, mean, and range data where applicable. Sex distribution, age at onset, age at follow-up at which data were available, and the number of HPO terms collected were relatively balanced in the full data set. The average age at death for the mtDNA disease group was 14.5 years (SD 17.35, range 0-51), 6.9 years (SD 8.8, range 0-41) in SLSD, and 9.9 years (SD 17.21, range 0-84) in the nuclear gene group. It should be noted, however, that only a small proportion of the mtDNA disease group had data on age at death because of MitoPhen v1 data set not capturing these data.

Table 1 Clinical and demographic characteristics of analytical data set

Variable	MtDNA $N = 3736$	SLSD $N = 671$	Nuclear Gene $N = 3118$	All $N = 7525$
Demographic and data set characteristics[a]				
Sex: Male, *n* (%)	1934 (51.8)	251 (37.4)	1327 (42.6)	3512 (46.7)
Sex: Female, *n* (%)	1710 (45.8)	285 (42.5)	1238 (39.7)	3233 (43.0)
Sex: Missing data, *n* (%)	93 (2.5)	135 (2.0)	553 (17.7)	781 (10.4)
Mean age at onset in years (SD, range)	16.6 (15.04, 0-76)	14.3 (15.43, 0-78)	15.4 (18.34, 0-75)	15.8 (16.77, 0-78)
Mean age at follow-up in years (SD, range)	36.6 (20.22, 1-85)	25.5 (18.25, 0-79)	33.0 (23.16, 0-104)	32.0 (22.59, 0-104)
Mean age at death in years (SD, range)	14.5 (17.35, 0-51)	6.95 (8.80, 0-41)	9.9 (17.21, 0-84)	9.6 (16.54, 0-84)
Mean number of HPO terms (SD, range)	6.4 (6.84, 1-44)	9.0 (8.08, 1-61)	10.6 (8.15, 1-59)	8.4 (7.79, 1-61)
Distribution of mitochondrial complex disorders based on gene groupings				
Complex I, *n* (%)	276 (48.5)	28 (5)	264 (46.4)	568 (7.5)
Complex II, *n* (%)	2 (2.6)	1 (1.2)	75 (96.2)	78 (1)
Complex III, *n* (%)	49 (29.5)	20 (12)	97 (58.5)	166 (2.2)
Complex IV, *n* (%)	153 (32.7)	28 (6)	287 (61.3)	468 (6.2)
Complex V, *n* (%)	25 (25)	0	79 (75)	104 (1.3)
Mitochondrial DNA maintenance disorder, *n* (%)	0	0	2,285 (100)	2,285 (30.3)
Distribution of clinical syndromes				
Leigh syndrome, *n* (%)	321 (76.8)	5 (1.2)	92 (22)	418 (5.5)
MELAS, *n* (%)	158 (95.7)	0 (0)	7 (4.3)	165 (2.1)
MERRF, *n* (%)	85 (90.4)	2 (2.1)	7 (7.5)	94 (1.2)
LHON, *n* (%)	1209 (100)	0	0	1209 (16)
CPEO ± MM, *n* (%)	200 (5.3)	516 (76.9)	914 (29.3)	1630 (21.6)
NARP, *n* (%)	28 (100)	0	0	28 (0.3)
Alpers-Huttenlocher syndrome, *n* (%)	0	0	135 (100)	135 (1.8)
Pearson syndrome, *n* (%)	0 (0)	128 (100)	0 (0)	128 (1.7)
Kearns-Sayre syndrome, *n* (%)	0 (0)	186 (99.5)	1 (0.5)	187 (2.4)
Inheritance pattern				
De novo, *n* (% of respective group)	0 (0)	671 (100)	87 (2.8)	758 (10)
Autosomal dominant *n* (% of respective group)	0 (0)	0 (0)	638 (20.5)	638 (8.4)
Autosomal recessive *n* (% of respective group)	0 (0)	0 (0)	2198 (70.5)	2198 (29.2)
Maternal inheritance, *n* (% of respective group)	3737 (100)	0 (0)	0 (0)	3737 (49.6)
X-linked inheritance, *n* (% of respective group)	0 (0)	0 (0)	18 (0.5)	18 (0.2)
Missing, *n* (% of respective group)	0 (0)	0 (0)	177 (5.6)	177 (2.3)

ADOA, autosomal dominant optic atrophy; *CPEO*, chronic progressive external ophthalmoplegia; *HON*, hereditary optic neuropathy; *LHON*, Leber's hereditary optic neuropathy; *MELAS*, mitochondrial encephalomyopathy with lactic acidosis and stroke-like episodes; *MERRF*, myoclonic epilepsy with ragged-red fibers; *MM+CPEO*, mitochondrial myopathy with CPEO; *mtDNA*, mitochondrial DNA; *NARP*, neuropathy, ataxia and retinitis pigmentosa syndrome; *NE*, necrotizing encephalopathy; *SLSD*, single large-scale deletions.

[a]For continuous variables, mean (standard deviation; range) displayed. For categorical variables, number (percentage of all individuals or within-group, where applicable) displayed.

The MitoPhen v2 "analytical data set" capturing the 21 SLAs after HPO-to-SLA transformation (Figure 2), across 7525 affected individuals, contained on average 8.4 HPO terms per affected individual (SD 7.69, range 1-61). There was a mean of 6.4 HPO terms (SD 6.84, range 1-44) in mtDNA due to the high proportion of individuals with LHON and therefore a narrow phenotypic spectrum; 9 (SD 8.08, range 1-61) in SLSD, and 10.6 (SD 8.15, range 1-59) HPO terms in the nuclear gene group.

Considering the frequency of mitochondrial complex I-V disorders (based on genes), although complex I disorders were more frequent in the mtDNA group, complex II-V disorders were most frequent in the nuclear gene group. The labeling of clinical syndromes using HPO terms and free-text columns showed that Leigh syndrome and CPEO with and without mitochondrial myopathy were noted across all genotypes, whereas other syndromes, such as LHON, were specific to genotype, as expected. The highest

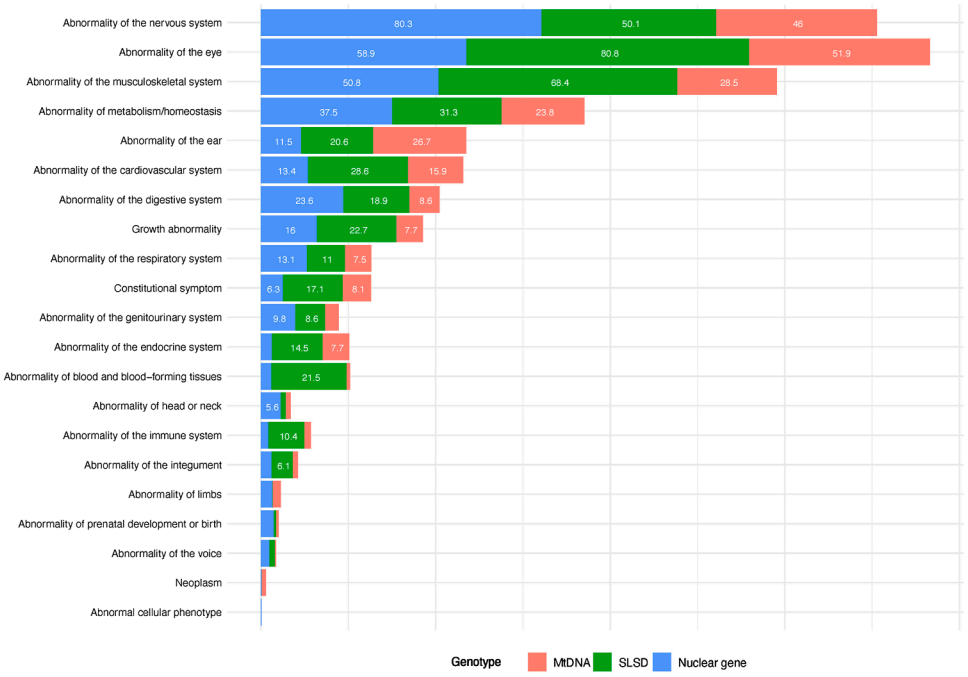

Figure 2 Distribution of 21 system-level abnormalities and their proportion in individual genotype groups. The MitoPhen v2 "analytical data set" of 7525 affected individuals captured phenotypes belonging to 21 system-level abnormalities. Values indicate percentage of individuals within that group with that particular system-level abnormality. Proportion values below 5% are not presented. MtDNA, mitochondrial DNA; SLSD, single large-scale deletions.

frequency of clinical syndromes in the MitoPhen v2 analytical data set belonged to CPEO, followed by LHON.

Mapping genotype-phenotype relationships with SLA

MDS across the analytical data set and in individual groups revealed patterns of co-occurrence of musculoskeletal, nervous, eye, ear and metabolism/homeostasis SLAs in the data set, in particular, of the musculoskeletal, nervous and metabolism/homeostasis SLAs (Figure 3). These were all highly prevalent in mtDNA, SLSD and nuclear gene groups, with some key differences (Figure 3). Eye and ear SLAs were located on adjacent quadrants, suggesting patterns of mutual exclusivity from other highly prevalent SLAs. These patterns are better explained when looking at the individual groups. A similar pattern to the analytical data set was noted in in individuals labeled "mtDNA oligo-SLA," with close clustering of nervous, musculoskeletal, and eye SLAs (Supplemental Figure 2A). In contrast, in the mtDNA single-SLA group, the MDS was largely driven by the mutual exclusivity and high prevalence of eye and ear SLAs (Supplemental Figure 2B). In SLSD, the *x*-axis of the MDS plot was driven largely by prevalence, with highly prevalent SLAs tending to co-occur as phenotypic features in this group (Supplemental Figure 2C). Similarly, in the nuclear gene group, the *y*-axis of the MDS plot reflected prevalence, with multiple clusters of (1) nervous and musculoskeletal changes (both highly prevalent in the nuclear gene group), (2) metabolism/homeostasis, gastrointestinal, and respiratory SLAs (moderately prevalent SLAs in this group), and (3) cardiovascular, head/neck and perinatal changes (which may be prevalent to a smaller extent in the nuclear gene group as a whole) (Supplemental Figure 2D).

To examine the landscape of phenotypic heterogeneity concurrently at the level of individual and group, we then distilled the SLAs into 6 PCs explaining 45.2% of the overall variance (Supplemental Figure 3A-D). The feature: sample ratio was acceptable (Kaiser-Meyer-Olkin statistic = 0.72). A plot of feature loadings on each PC component is presented in Supplemental Figure 4. Briefly, PC1 loaded highly on neuromuscular and metabolic SLAs; PC2 on respiratory, gastrointestinal system, integument, and head/neck SLAs; PC3 on blood, immune, and genitourinary SLAs; PC4 on eye and cardiovascular SLAs; PC5 on limbs and ear SLAs; and PC6 on prenatal development or birth SLAs.

Distilling these phenotypic components into a 2D low-dimensional space, the UMAP revealed that the landscape of phenotypic heterogeneity in mitochondrial disorders is underscored by marked overlap between different groups (Figure 4). Visually, exceptions were noted for (1) individuals with Pearson syndrome (Figure 5) and (2) individuals labeled with single-SLA mtDNA diseases, who exhibited homogeneity (ie, self-similarity by clustering close to each other) driven largely by persons with LHON who are highly phenotypically alike irrespective of their contributing mtDNA variant (Figure 5, Supplemental Figures 5), no clear patterns of phenotypic differentiation

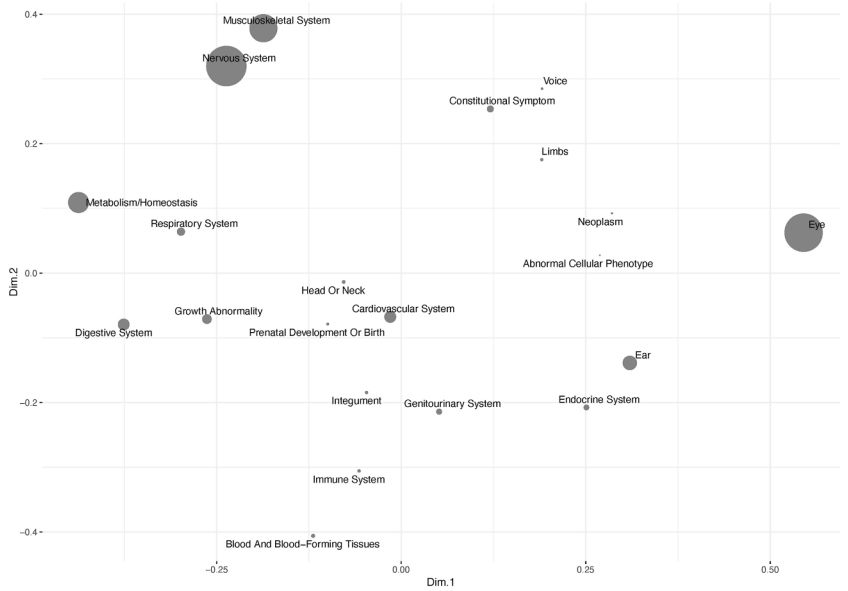

Figure 3 Multidimensional scaling (MDS) analysis of system-level abnormalities in mtDNA, SLSD, and nuclear gene groups. Classical 2D MDS was performed using a Spearman's decorrelation distance matrix on the data. Each system-level abnormality (SLA) is scaled by its prevalence in the analytical data set, with larger dots indicating higher prevalence of that particular SLA in the analytical data set. In this plot, x and y axes (reflecting dimensions 1 and 2, respectively) correspond to mathematical dimensions of a reduced space capturing the largest and second largest variation in the data, respectively. Relative positions of each feature (dot) along these axes indicates its relationship to other points; points located close together in the plot are more similar or related based on the data, whereas those further apart are more dissimilar. The overall pattern of dots, including clusters or outliers, provides insight into the underlying structure or relationships within the data.

between groups or most clinical syndromes, stratified by contributing genotype, were observed (Supplemental Figures 6-8).

Furthermore, no clear and distinct patterns of phenotypic similarity and differentiation emerged that were associated: (1) most frequent mtDNA and nuclear genes (Supplemental Figures 7 and 8); (2) sex distribution (Supplemental Figure 9); (3) type of mitochondrial complex disorder (Supplemental Figure 10); (4) generation relative to proband; or (5) inheritance pattern (Supplemental Figure 11). Taken together, although the MDS and PCA suggested that phenotypic features (SLAs) are not necessarily individually occurring but display a systematic pattern of clustering and covariance, the UMAP added a second granular layer to

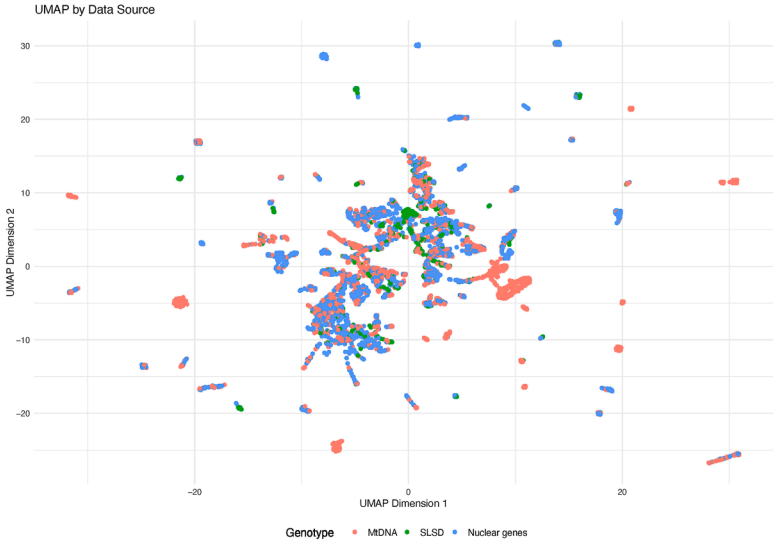

Figure 4 Low-dimensional UMAP embedding of phenotypic presentation in mtDNA, SLSD, and nuclear gene groups. Uniform Manifold Approximation and Projection (UMAP) embedding initialized using principal component analysis, with genotype projected into each displayed data point. Considerable overlap seen between broad genotype groups with mitochondrial DNA (mtDNA) shown in pink, single large-scale mtDNA deletions (SLSD) shown in green, and nuclear genes causing primary mitochondrial diseases shown in blue.

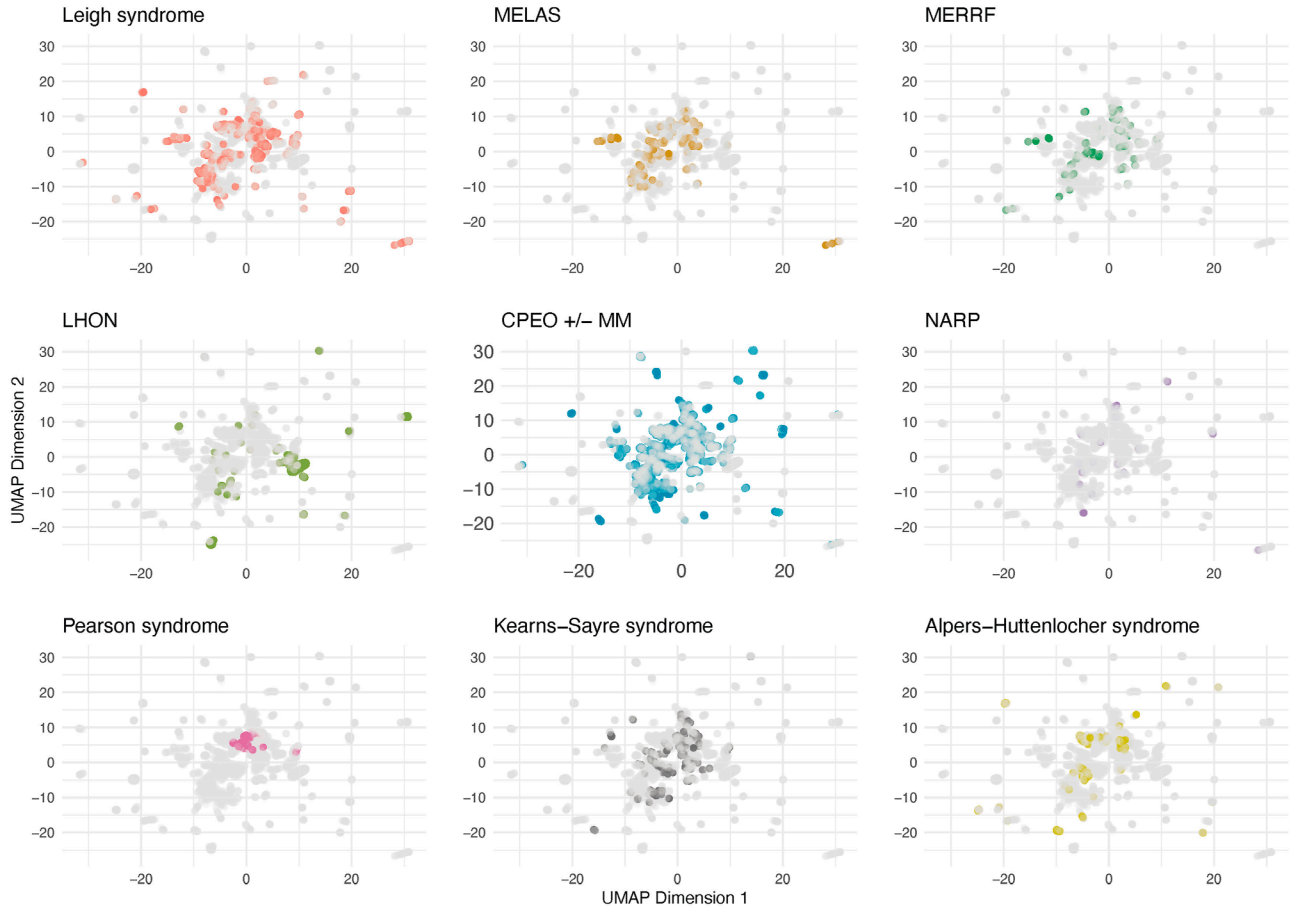

Figure 5 Projection of clinical syndromes in low-dimensional UMAP embeddings of phenotypic presentation in primary mitochondrial disease genotypes. UMAP embedding initialized using principal component analysis, with clinical syndrome projected into each displayed data point from the MitoPhen v2 "analytical data set" of 7525 affected individuals. Light gray data points denote no available data/applicable clinical syndrome. CPEO, chronic progressive external ophthalmoplegia; LHON, Leber's hereditary optic neuropathy; MELAS, mitochondrial encephalomyopathy with lactic acidosis and stroke-like episodes; MERRF, myoclonic epilepsy with ragged-red fibers; MM and CPEO, mitochondrial myopathy with CPEO; NARP, neuropathy, ataxia and retinitis pigmentosa syndrome; NE, necrotizing encephalopathy; UMAP, Uniform Manifold Approximation and Projection.

indicate that the association between these distinct clinical features is not necessarily an outcome of the specific disease or group that the person belongs to. Except for a select few conditions (eg, Pearson syndrome, LHON), the majority of mitochondrial and nuclear gene disorders fall along a mosaic of graded, not categorically distinct, phenotypic variation, ultimately complicating their diagnosis and differentiation based on system-level or syndromic phenotype alone.

Examining within-gene phenotypic similarity

The SLA-driven approach revealed no clear distinguishing features across most PMDs; therefore, we further evaluated the use of phenotype similarity scoring with all available nonredundant HPO terms in MitoPhen v2. First, we evaluated 99 genes with data on 2 or more probands each. This analysis highlighted genes where shared phenotypes are more similar than expected by chance. Eighty-five genes, which included SLSD considered as a single genetic entity, demonstrated statistically significant within-group phenotype similarity ($P < .05$, Supplemental Table 3). There were $\leq$5 probands for 9 out of 14 genes without significant within-group phenotype similarity, in 3 out of these 14 genes (*MT-TG* [HGNC:7486], *MT-CYB* [HGNC:7427], *NDUFV2* [HGNC:7717]) there were 6 to 8 probands but variant-specific phenotypic differences. In the 2 genes with $\geq$10 probands: *MT-TI* [HGNC:7488] and *MT-TV* [HGNC:7500], recent updates by the ClinGen Mitochondrial Disease Nuclear and Mitochondrial Variant Curation Expert Panel[48] suggest that 6 of 8 variants recorded for these genes do not reach likely pathogenic status. Therefore, the lack of significant phenotypic homogeneity with the data available is supportive of these variants not showing phenotypic segregation, which is required to meet likely pathogenic status. These results are reflected in Figure 6, which shows overall PSI above 0.25 for most of the frequent mtDNA and nuclear genes in the data set.

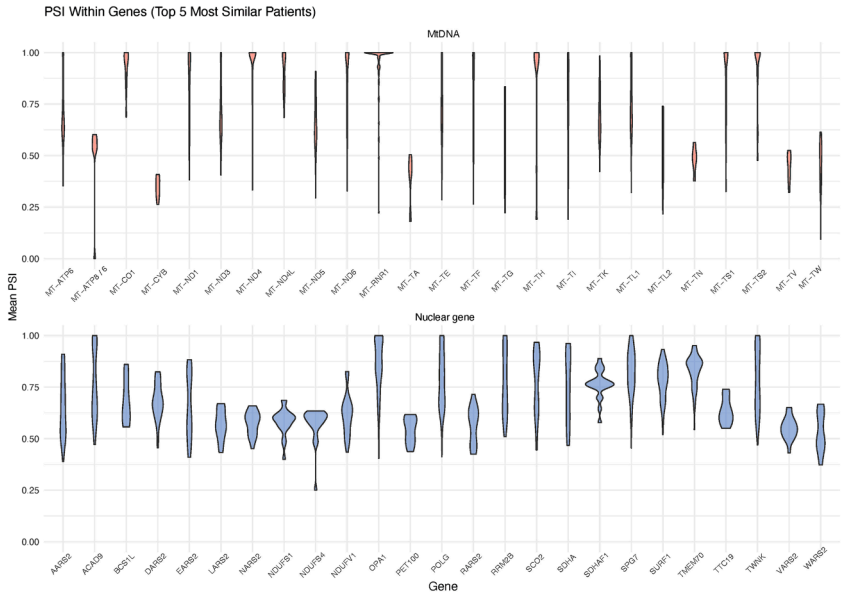

Figure 6 Within-gene phenotype similarity index calculations for 25 most common mtDNA and nuclear gene groups in the full data set. A. Mitochondrial DNA (mtDNA) genes show variability in the mean phenotype similarity scores signifying phenotypic heterogeneity. B. Frequent nuclear genes within MitoPhen v2 show higher PSI values overall, compared with mtDNA genes. PSI, phenotype similarity index.

There appeared to be more heterogeneity within the mtDNA gene group with broader PSI ranges, compared with the nuclear gene group, and this could be explained by the influence of variant heteroplasmy levels on phenotypic spectrums in mtDNA diseases. The data shown in Figure 6 includes nonprobands, which means that milder phenotypic spectrums are expected in affected relatives of probands with mtDNA diseases, therefore generating lower PSI values. Genes with PSI ranges $\geq$0.5 suggest stronger clustering of phenotypes within these groups.

In view of the PSI results, we investigated for HPO-gene association differences between the current HPO database (downloaded 26/01/2025 from https://hpo.jax.org), and the MitoPhen v2 dataset. We found 111 terms seen at least 5 times (after excluding the SLSD genotype) across 64 genes, which were not seen in association with the genes within the current HPO database (Supplemental Figure 12). A HPO-level enrichment analysis within MitoPhen v2 found 1148 significantly enriched HPO terms (false-discovery-rate-adjusted $P < .05$) across the 11 genetic categories shown in Figure 1, with the number of enriched HPO terms ranging from 13 for the mtDNA rRNA gene category, to 219 for the nuclear main genes category (Supplemental Table 4). This highlighted the potential to identify specific PMD genotypes within the data set using individual-level PSI.

Classifying *POLG*, *MT-TL1*, and SLSD from other PMD genotypes using PSI and clinical characteristics

To further evaluate the predictive value of PSI, we considered a logistic regression model using *POLG*, *MT-TL1*, and SLSD genotypes as exemplars because of the high overall within-genotype PSI (Figure 6) but also the phenotypic diversity associated with these groups (Supplemental Figures 2, 7, and 8). ROC analyses confirmed strong predictive performance of PSI, with area under the curve values of 0.9279 (95% CI: 0.9121-0.9438) for *POLG*, 0.8455 (95% CI: 0.8077-0.8832) for *MT-TL1*, and 0.9297 (95% CI: 0.9068-0.9527) for SLSD (Figure 7). The optimal PSI, based on Youden's index,[49] ranged between 0.47 and 0.52 to achieve sensitivity > 80% and specificity > 70% in all 3 scenarios (Figure 7). The confusion matrix results are shown in Supplemental Table 5.

We performed generalized linear models to evaluate the predictive utility of PSI and clinical variables (age at onset, sex, and inheritance pattern) for identifying *POLG*, *MT-TL1*, and SLSD genotypes. Across the models, PSI remained the strongest predictor for all genotypes ($P <$ .001), with effect estimates ranging from 9.83 (*MT-TL1*) to 25.90 (SLSD). Age at onset exhibited genotype-specific effects: earlier onset was significantly associated with *POLG* (estimate $= -0.0278$, $P <$ .001) and SLSD (estimate $= -0.0683$, $P <$ .001), whereas in the model with *MT-TL1*, there was a weaker positive association with later onset (estimate $= 0.0105$, $P = .00753$). Sex was only significant for *POLG*, for which males had lower odds of this genotype (estimate $= -0.409$, $P = .00137$). Inheritance pattern was a significant predictor for *POLG*, particularly for autosomal recessive (estimate $= 1.13$, $P <$ .001) and sporadic patterns (estimate $= -3.44$, $P <$ .001) but had no significant effect in the *MT-TL1* and SLSD models. Model performance was highest for *POLG* and SLSD (Δ

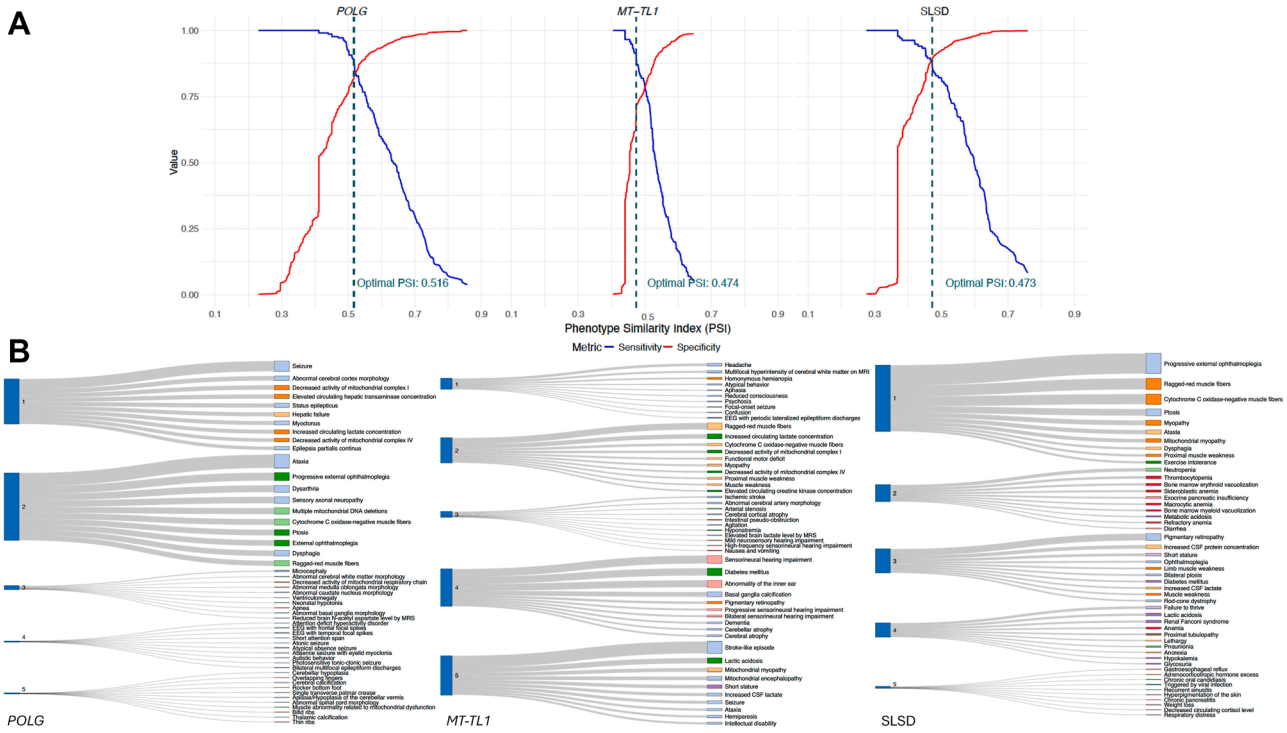

Figure 7 Phenotype similarity index for predicting *POLG, MT-TL1*, and SLSD genotypes and for co-clustering of phenotypes. A. Phenotype similarity index (PSI) was calculated as a mean phenotype similarity score based on the 5 most similar individuals within the reference data sets to the query individual. Youden's index was used to calculate the optimal PSI for each genotype, optimizing sensitivity and specificity for correctly predicting the genotype. B. Sankey diagrams visualize the top 10 co-occurring HPO terms across the 5 largest phenotype clusters for *POLG*, *MT-TL1*, and SLSD. Each Sankey plot was generated using PSI > 0.5 per genotype to generate data from highly similar individuals. The HPO terms are colored by their top-level system involvement.

Deviance = 3130.8 and 3218.0, respectively), whereas the *MT-TL1* model demonstrated weaker predictive power (Δ Deviance = 931.5).

To further explore the role of phenotype similarity in HPO co-occurrence, we used Sankey plots to visualize the most frequently co-occurring HPO terms across the 5 largest phenotype clusters for *POLG*, *MT-TL1*, and SLSD (Figure 7). Each Sankey plot was constructed using PSI > 0.5, ensuring that highly phenotypically similar individuals contributed to the network. The distribution of HPO terms differed substantially across genes. *POLG*-associated clusters included cluster 1 with terms in keeping with Alpers-Huttenlocher syndrome (OMIM 203700), such as "seizure" and "hepatic failure," other clusters were dominated by terms "progressive external ophthalmoplegia," "ataxia," and epilepsy-related features, aligning with its known neuromuscular involvement. However, there were additional terms related to syndromic features, such as "overlapping fingers," and skeletal abnormalities noted in the smaller *POLG*-related cluster 5. *MT-TL1*-associated phenotype clusters showed a strong enrichment of metabolic phenotypes, including "lactic acidosis," "mitochondrial myopathy," and "stroke-like episodes," consistent with its role in mitochondrial encephalomyopathy with lactic acidosis and stroke-like episodes (cluster 5) but also clusters with more multisystem features, including, brain atrophy related findings, hearing impairment, and diabetes mellitus. SLSD-associated clusters 1 and 2 were marked by muscle-related terms as expected with CPEO and features in keeping with Pearson syndrome (OMIM 557000), respectively, but also clusters with several systems, including terms related to infection, reflected the breadth of SLSD-related conditions. Across all genotypes, a substantial proportion of co-occurring terms were found less frequently (ie, not in the top 10 per cluster) and are not visualized in Figure 7; however, these represented a diverse range of associated phenotypes, seen in several individuals. These findings highlight distinct genotype-phenotype relationships and demonstrate how PSI can be leveraged to uncover co-occurring phenotypic networks with clinical relevance.

Discussion

Mitochondrial diseases are genetically and phenotypically heterogeneous, often progressive, single or multisystem conditions without current curative treatments. An early diagnosis is crucial for identifying organ systems at risk and for improving quality of life for individuals, as well as providing genetic counseling for families. However, untangling the graded phenotypic variations in PMDs has

posed continuing clinical and analytical challenges, with major implications to clinical diagnosis and management.[6] Current diagnostic pipelines rely on clinicians recognizing features that are in keeping with a presentation of mitochondrial disease. Often this would be a combination of symptoms or signs affecting multiple systems or features of a well-defined syndrome, such as LHON or Leigh syndrome, prompting initiation of genetic sequencing with/without mtDNA analysis. However, families often face a prolonged diagnostic odyssey because of the complexity of these conditions. Large, rare disease genetic testing studies using mainly blood samples have been established to reduce invasive investigations and to limit the protracted time spent without a diagnosis.[50-52] Yet, the diagnostic yield in these large cohorts can be limited by phenotype mimics with other conditions, quality and quantity of HPO data gathered by recruiting researchers, and, importantly, known gene-phenotype relationships.[19,53]

This study addressed the diagnostic issue posed by limited gene-HPO associations by combining a large, well-characterized data set of mitochondrial disorders and a suite of advanced analytics to facilitate interrelations between clinical phenotype and genotype. The clinical variability of PMDs was underpinned by 6 fundamental dimensions approximating changes in (1) neuromuscular and metabolic, (2) respiratory and gastrointestinal, (3) blood and immune-related, (4) eye-related and cardiac, (5) ear-related and limbs, and (6) perinatal-related functions. Individual SLAs, each of which loaded on to 1 of these 6 dimensions, tend to cluster in the form of symptom complexes, patterns of which vary across mtDNA, SLSD and nuclear genes. Irrespective of the underlying genetic variation, we showed that individuals largely demonstrated features of multiple axes and located across a graded multidimensionally defined phenotypic mosaic. We were able to show, using SLAs, that individuals with common mitochondrial disease syndromes, such as Leigh syndrome, Kearns-Sayre syndrome, mitochondrial encephalomyopathy with lactic acidosis and stroke-like episodes, and MERRF, do not cluster together in terms of phenotypic presentations. The overlap of discrete variants in terms of SLAs, highlights the complex phenotypic landscape, in which there may be limited direct 1-to-1 correspondence between genotype and phenotype and the limited utility of syndrome names in describing clinical feature spectrums. Overall, our findings show that SLA-based clustering does not yield distinct genotype groupings, limiting its standalone diagnostic utility.

Furthermore, when we evaluated the phenotype data at the HPO level, we found over 100 HPO terms associated with 64 genes occurring at least 5 times in each case and not seen within the current online HPO database, adding to the current knowledgebase for disease-gene relationships. It was also possible to demonstrate significantly enriched HPO terms per genetic category compared with each other, which suggested that the individual HPO level would be helpful to delineate between PMDs. In the clinical setting, these findings can be used, for example, to evaluate an individual with a VUS in a known PMD gene by considering their deep phenotype alongside the MitoPhen v2 data set. This can also be done using phenotype similarity assessments because 85 genes showing significant within-gene phenotype similarity overall. The genes that did not reach statistical significance were either limited in data, or the variants did not reach likely pathogenic status according to recent evaluations. This signified the utility of phenotype similarity assessments of individuals characterized within MitoPhen v2 because the lack of phenotypic homogeneity in particular genes with sufficient published data can evidence the clinical annotation of variants as VUS. The findings, together, allude to the potential need to reconceptualize current views of genotype-phenotype interrelationships in mitochondrial diseases in which data that capture multidimensional associations between genotype and phenotype may offer an additional, robust explanatory framework to clinical heterogeneity, over and above categorical approaches linking specific genetic variants and their clinical presentations.[9]

From a clinical perspective, our findings underscore the necessity for comprehensive phenotype capture, through structured assessments using current tools including the HPO. Our results contribute to a growing body of evidence supporting the use of phenotype mapping indices, such as PSI, in refining genetic classification and improving diagnostic precision in mitochondrial diseases. In this work, we demonstrated that PSI can robustly and consistently predict mitochondrial genotypes: *POLG*, *MT-TL1*, and SLSD from other mitochondrial genotypes, irrespective of variable contributions from other clinico-demographic features. These genotypes were chosen because of their phenotypic breadth. We showed that it is possible to delineate top co-occurring HPO clusters in highly phenotypically similar individuals with these genotypes. Although these clusters broadly fit with known syndromic presentations, there were additional findings of terms related to dysmorphic features and skeletal abnormalities noted in the *POLG* group: brain atrophy terms co-occurring with diabetes mellitus and sensorineural hearing impairment in the *MT-TL1* group and infections such as sinusitis and candidiasis along with gastro-esophageal reflux disease being noted in a small cluster within the SLSD group. Therefore, the approach to retain all phenotype terms may offer value in situations in which genetic VUS are identified, as computed phenotype similarity measures against rich data sets, such as MitoPhen v2, could help prioritize or exclude variants in large rare disease data sets. On this front, it would be important for future work to explore the integration of artificial-intelligence-assisted matching and extraction of HPO terms for individuals, facilitating the tagging of a full set of HPO terms associated with their entire phenotype (in the form of an individual "phenopacket")[10] to understand the multidimensional clinical profile of each genetic variant.

A number of limitations, however, warrant discussion. The power of large data sets is often tempered by missing and incomplete data. In the case of MitoPhen v2, this

pertains to the number of coded HPO terms that fully represent the individual's phenotype reported in the original source publications. It would be important for future studies to validate the current findings against single-center and/or prospectively collected data sets with carefully coded HPO terms. A key contributor to the phenotype may also be disease severity, which we could not assess directly because relevant measures had not been consistently reported in source publications. However, capturing all phenotypic terms available in publications meant that within-gene HPO clustering in frequent genotypes (*POLG*, *MT-TL1*, and SLSD) could be visualized, which would reflect individuals with conditions of similar severity. Although our data-driven analyses, specifically UMAP, visualized the overlap and differences between phenotypes at a more generalized level, the nature of such data-driven analyses limit direct correspondence with the input data. This study circumvented some of these issues by projecting various labels and values into the UMAP. However, future work should examine the predictive capacity of such low-dimensional spaces in predicting the phenotypic similarity between a test case and current data across different rare diseases. We further evaluated phenotypic similarity using PSI and detailed the evaluation of *POLG*, *MT-TL1*, and SLSD as exemplars of genotypes with heterogeneous phenotypic presentations. However, the limitation of this work exists in the comparison of published literature with itself. The applicability of the PSI in "real world" rare disease data sets is currently ongoing, we have recently shown the utility of applying PSI in the context of aiding mtDNA disease diagnoses within Solve-RD,[54] a large Europe-wide study of rare diseases.[50] We have also been evaluating VUS and diagnosed individuals across mitochondrial genotypes within large cohorts such as Genomics England 100,000 Genomes Project,[51] and RD-Connect[12] (work ongoing). These evaluations are required to further enhance the computational interpretation of variants that may be causative of PMDs. The other limitation of applying PSI values to several genes in the data set is the scarcity of published literature on these rare disease genes; however, this also corresponds to the epidemiology of PMDs, the data for which will need to be continuously updated. Given the time-consuming nature of manual curation, this data set lends itself to research utilizing large language models to extract individual-level genotype-phenotype data from published literature. Finally, additional components of MitoPhen v2, such as data on reported ethnicity and treatments used, are subject to reporting bias but could be utilized in future studies of long-term mortality and morbidity associated with PMDs across different groups of individuals.

Overall, MitoPhen v2 captures data across 1597 publications of 10,626 individuals with 117 genotypes of PMDs. The resulting individual-level phenotype-genotype data set has enabled system-level computation approaches to phenotype mapping, and HPO-level phenotype similarity evaluations. PSI was able to differentiate common genotypes associated with PMDs from others; we demonstrated phenotypic homogeneity within the majority of genes, which can be harnessed for variant interpretation in large data sets. This work indicated that PSI generated using individual-level data sets is a clinically important measure to integrate into large, rare disease data sets for the evaluation of gene variants associated with PMDs and is a framework that could be utilized to evaluate other similarly heterogenous rare genetic diseases. MitoPhen v2 is a rich manually curated publicly available data source, and research is ongoing to model the interaction of genetic and clinical variations in mitochondrial diseases.

Data Availability

The MitoPhen database is freely accessible through the website https://www.mitophen.org/. We have separately attached the new data that was added to create MitoPhen v2, as a Supplemental table. We would be grateful for acknowledgement and citation of this work in research involving the dataset.

Acknowledgments

The authors are grateful to Dr Daniel Greene who has maintained the website www.mitophen.org.

Funding

T.R. was an academic clinical lecturer supported by Health Education England, East Suffolk and North Essex NHS Foundation Trust, and the University of Cambridge during the project. T.R. was also supported by the Elizabeth Blackwell Institute with funding provided by the Franklin-Adams endowment. S.R. is an employee of Costello Medical, Cambridge, UK. M.M. is an employee of Costello Medical, London, UK. R.P. is an employee of Costello Medical, London, UK. R.H. is supported by the Wellcome Discovery Award (226653/Z/22/Z), the Medical Research Council (UK) (MR/V009346/1), the Hereditary Neuropathy Foundation, the AFM-Telethon, the Ataxia UK, the Action for AT, the Muscular Dystrophy UK, the Rosetrees Trust (PGL23/100048), the LifeArc Centre to Treat Mitochondrial Diseases (LAC-TreatMito) and the UKRI/Horizon Europe Guarantee MSCA Doctoral Network Programme (Project 101120256: MMM). She is also supported by an MRC strategic award to establish an International Centre for Genomic Medicine in Neuromuscular Diseases (ICGNMD) MR/S005021/1. This research was supported by the NIHR Cambridge Biomedical Research Centre (NIHR203312). P.F.C. is currently funded by a Wellcome Discovery Award (226653/Z/22/Z) and a Wellcome Collaborative Award (224486/Z/21/Z), the Medical Research Council Mitochondrial Biology Unit

(MC_UU_00028/7), the MRC International Centre for Genomic Medicine in Neuromuscular Disease (MR/S005021/1), and the Biological and Biotechnology Research Council (BB/Y003209/1), and the LifeArc Centre to Treat Mitochondrial Diseases (LAC-TreatMito) under grant no. 10748. LifeArc is a charity registered in England and Wales under no. 1015243 and in Scotland under no. SC037861. The views expressed are those of the authors and not necessarily those of the NIHR or the Department of Health and Social Care.

Author Contributions

Conceptualization: T.R., S.R.; Data Curation: T.R., N.E., R. N., J.Y., E.S.A., K.S., M.E.K, C.G., A.L., E.S.; Formal Analysis: T.R., S.R., M.M., R.P.; Methodology: T.R., S.R., N.E.; Project Supervision: T.R., S.R.; Supervision: T.R., S. R., R.H., P.F.C.; Visualization: T.R., S.R.; Writing-original draft: T.R., S.R.; Writing-review and editing: T.R., S.R., N. E., R.N., J.Y., E.S.A., K.S., M.E.K, C.G., A.L., R.H., P.F.C.

ORCIDs

Thiloka Ratnaike: http://orcid.org/0000-0001-5400-104X
Siddharth Ramanan: http://orcid.org/0000-0002-8591-042X
Nour Elkhateeb: http://orcid.org/0000-0002-3076-3178
Eszter Sara Arany: http://orcid.org/0000-0002-3846-1596
Manya Mirchandani: http://orcid.org/0000-0003-2564-1042
Katherine Schon: http://orcid.org/0000-0001-8054-8954
Eren Kule: http://orcid.org/0009-0006-4614-2698
Christopher Gilmartin: http://orcid.org/0000-0001-5845-9429
Angela Lochmüller: http://orcid.org/0009-0001-3763-5881
Emogene Shaw: http://orcid.org/0000-0002-2126-8458
Rita Horváth: http://orcid.org/0000-0002-9841-170X
Patrick F. Chinnery: http://orcid.org/0000-0002-7065-6617

Ethics Declaration

No ethical approval was required for this study because all individual-level data have previously been published and remain deidentified. The authors did not access any unredacted protected health information for this study.

Conflict of Interest

The authors declare no conflicts of interest.

Declaration of AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, Dr Ratnaike used ChatGPT 4.0 to check R scripts to generate Figures 1 and 7. After using the tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Additional Information

The online version of this article (https://doi.org/10.1016/j.gim.2025.101620) contains supplemental material, which is available to authorized users.

Affiliations

[1]Department of Pediatric Neurology, University Hospitals Bristol and Weston NHS Foundation Trust, Bristol, United Kingdom; [2]Department of Pediatrics, University of Cambridge, Cambridge, United Kingdom; [3]Bristol Medical School, University of Bristol, Bristol, United Kingdom; [4]Costello Medical, Cambridge, United Kingdom; [5]Department of Clinical Genetics, Cambridge University Hospitals NHS Foundation Trust, Cambridge, United Kingdom; [6]School of Clinical Medicine, University of Cambridge, Cambridge Biomedical Campus, Cambridge, United Kingdom; [7]Department of Clinical Neurosciences, University of Cambridge, Cambridge, United Kingdom; [8]Costello Medical, London, United Kingdom; [9]Koç University, School of Medicine, Istanbul, Turkey; [10]Department of Medical Genetics, University of Cambridge, Cambridge, United Kingdom; [11]Population Health Sciences Institute, Newcastle University, Newcastle, United Kingdom

References

1. Gorman GS, Schaefer AM, Ng Y, et al. Prevalence of nuclear and mitochondrial DNA mutations related to adult mitochondrial disease. *Ann Neurol*. 2015;77(5):753-759. http://doi.org/10.1002/ana.24362
2. Smeitink J, van den Heuvel L, DiMauro S. The genetics and pathology of oxidative phosphorylation. *Nat Rev Genet*. 2001;2(5):342-352. http://doi.org/10.1038/35072063
3. Pizzamiglio C, Hanna MG, Pitceathly RDS. Chapter 4. Primary mitochondrial diseases. In: Lynch DS, Houlden H, eds. *Handbook of Clinical Neurology*. Amsterdam, The Netherlands: Elsevier; 2024:53-76.
4. Burr SP, Chinnery PF. Origins of tissue and cell-type specificity in mitochondrial DNA (mtDNA) disease. *Hum Mol Genet*. 2024;33(R1):R3-R11. http://doi.org/10.1093/hmg/ddae059

5. Esmaeil A, Ali A, Behbehani R. Leber's hereditary optic neuropathy: update on current diagnosis and treatment. *Front Ophthalmol (Lausanne)*. 2022;2:1077395. http://doi.org/10.3389/fopht.2022.1077395
6. Schon KR, Ratnaike T, van den Ameele J, Horvath R, Chinnery PF. Mitochondrial diseases: a diagnostic revolution. *Trends Genet*. 2020;36(9):702-717. http://doi.org/10.1016/j.tig.2020.06.009
7. Maeda K, Kawai H, Sanada M, et al. Clinical phenotype and segregation of mitochondrial 3243A>G mutation in 2 pairs of monozygotic twins. *JAMA Neurol*. 2016;73(8):990-993. http://doi.org/10.1001/jamaneurol.2016.0886
8. 1000 Genomes Project Consortium, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. http://doi.org/10.1038/nature15393
9. Ramanan S, Akarca D, Henderson SK, et al. The graded multidimensional geometry of phenotypic variation and progression in neurodegenerative syndromes. *Brain*. 2025;148(2):448-466. http://doi.org/10.1093/brain/awae233
10. Gargano MA, Matentzoglu N, Coleman B, et al. The Human Phenotype Ontology in 2024: phenotypes around the world. *Nucleic Acids Res*. 2024;52(D1):D1333-D1346. http://doi.org/10.1093/nar/gkad1005
11. Lagorce D, Lebreton E, Matalonga L, et al. Phenotypic similarity-based approach for variant prioritization for unsolved rare disease: a preliminary methodological report. *Eur J Hum Genet*. 2024;32(2):182-189. http://doi.org/10.1038/s41431-023-01486-7
12. Laurie S, Piscia D, Matalonga L, et al. The RD-Connect Genome-Phenome Analysis Platform: accelerating diagnosis, research, and gene discovery for rare diseases. *Hum Mutat*. 2022;43(6):717-733. http://doi.org/10.1002/humu.24353
13. Ratnaike TE, Greene D, Wei W, et al. MitoPhen database: a human phenotype ontology-based approach to identify mitochondrial DNA diseases. *Nucleic Acids Res*. 2021;49(17):9686-9695. http://doi.org/10.1093/nar/gkab726
14. Pitceathly RD, Rahman S, Hanna MG. Single deletions in mitochondrial DNA–molecular mechanisms and disease phenotypes in clinical practice. *Neuromuscul Disord*. 2012;22(7):577-586. http://doi.org/10.1016/j.nmd.2012.03.009
15. Anteneová N, Kelifová S, Kolářová H, et al. The phenotypic spectrum of 47 Czech patients with single, large-scale mitochondrial DNA deletions. *Brain Sci*. 2020;10(11):766. http://doi.org/10.3390/brainsci10110766
16. Frascarelli C, Zanetti N, Nasca A, et al. Nanopore long-read next-generation sequencing for detection of mitochondrial DNA large-scale deletions. *Front Genet*. 2023;14:1089956. http://doi.org/10.3389/fgene.2023.1089956
17. Ganetzky R, Stanley KD, MacMullen LE, et al. Recognizing the evolution of clinical syndrome spectrum progression in individuals with single large-scale mitochondrial DNA deletion syndromes (SLSMDS). *Genet Med*. 2025;27(5):101386. http://doi.org/10.1016/j.gim.2025.101386
18. Macken WL, Falabella M, Pizzamiglio C, et al. Enhanced mitochondrial genome analysis: bioinformatic and long-read sequencing advances and their diagnostic implications. *Expert Rev Mol Diagn*. 2023;23(9):797-814. http://doi.org/10.1080/14737159.2023.2241365
19. Schon KR, Horvath R, Wei W, et al. Use of whole genome sequencing to determine genetic basis of suspected mitochondrial disorders: cohort study. *BMJ*. 2021;375:e066288. http://doi.org/10.1136/bmj-2021-066288
20. Parikh S, Karaa A, Goldstein A, et al. Diagnosis of "possible" mitochondrial disease: an existential crisis. *J Med Genet*. 2019;56(3):123-130. http://doi.org/10.1136/jmedgenet-2018-105800
21. Woodbridge P, Liang C, Davis RL, Vandebona H, Sue CM. POLG mutations in Australian patients with mitochondrial disease. *Intern Med J*. 2013;43(2):150-156. http://doi.org/10.1111/j.1445-5994.2012.02847.x
22. Rahman S, Copeland WC. POLG-related disorders and their neurological manifestations. *Nat Rev Neurol*. 2019;15(1):40-52. http://doi.org/10.1038/s41582-018-0101-0
23. Ratnaike TE, Elkhateeb N, Lochmüller A, et al. Evidence for sodium valproate toxicity in mitochondrial diseases: a systematic analysis. *BMJ Neurol Open*. 2024;6(1):e000650. http://doi.org/10.1136/bmjno-2024-000650
24. Hakonen AH, Davidzon G, Salemi R, et al. Abundance of the POLG disease mutations in Europe, Australia, New Zealand, and the United States explained by single ancient European founders. *Eur J Hum Genet*. 2007;15(7):779-783. http://doi.org/10.1038/sj.ejhg.5201831
25. Martin AR, Williams E, Foulger RE, et al. PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat Genet*. 2019;51(11):1560-1565. http://doi.org/10.1038/s41588-019-0528-2
26. Kopanos C, Tsiolkas V, Kouris A, et al. VarSome: the human genomic variant search engine. *Bioinformatics*. 2019;35(11):1978-1980. http://doi.org/10.1093/bioinformatics/bty897
27. genoox. Franklin by genoox. Accessed January 6, 2024. https://franklin.genoox.com/clinical-db/home
28. Richards S, Aziz N, Bale 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*. 2015;17(5):405-424. http://doi.org/10.1038/gim.2015.30
29. Chen S, Francioli LC, Goodrich JK, et al. A genomic mutational constraint map using variation in 76,156 human genomes. *Nature*. 2024;625(7993):92-100. http://doi.org/10.1038/s41586-023-06045-0
30. Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res*. 2018;46(D1):D1062-D1067. http://doi.org/10.1093/nar/gkx1153
31. MITOMAP: a human mitochondrial genome database. MITOMAP. Accessed May 15, 2019. http://www.mitomap.org
32. Rehm HL, Berg JS, Brooks LD, et al. ClinGen–the clinical genome resource. *N Engl J Med*. 2015;372(23):2235-2242. http://doi.org/10.1056/NEJMsr1406261
33. Wong LJC, Chen T, Wang J, et al. Interpretation of mitochondrial tRNA variants. *Genet Med*. 2020;22(5):917-926. http://doi.org/10.1038/s41436-019-0746-0
34. Greene D, Richardson S, Turro E. ontologyX: a suite of R packages for working with ontological data. *Bioinformatics*. 2017;33(7):1104-1106. http://doi.org/10.1093/bioinformatics/btw763
35. Kruskal JB, Wish M. *Multidimensional Scaling*. Sage Publications, Inc; 1978. http://doi.org/10.4135/9781412985130
36. Healy J, McInnes L. Uniform manifold approximation and projection for dimension reduction. *Nat Rev Methods Primers*. 2024 Nov 21;4(1):82. http://doi.org/10.1038/s43586-024-00363-x
37. Han H, Li W, Wang J, Qin G, Qin X. Enhance explainability of manifold learning. *Neurocomputing*. 2022;500:877-895. http://doi.org/10.1016/j.neucom.2022.05.119
38. Schmidt A, Danyel M, Grundmann K, et al. Next-generation phenotyping integrated in a national framework for patients with ultrarare disorders improves genetic diagnostics and yields new molecular findings. *Nat Genet*. 2024;56(8):1644-1653. http://doi.org/10.1038/s41588-024-01836-1
39. Kobak D, Linderman GC. Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nat Biotechnol*. 2021;39(2):156-157. http://doi.org/10.1038/s41587-020-00809-z
40. Sievert C. *Interactive Web-Based Data Visualization with R, Plotly, and Shiny*. Chapman & Hall/CRC; 2020. http://doi.org/10.1201/9780429447273
41. Wickham H. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag; 2022.

42. Team R.C. Package "Parallel". Accessed February 1, 2025. 2025. https://www.rdocumentation.org/packages/parallel/versions/3.6.2
43. Kuhn M. Building Predictive Models in R Using the caret Package. *J Stat Softw*. 2008;28(5):1-26. http://doi.org/10.18637/jss.v028.i05
44. Robin X, Turck N, Hainard A, et al. Proc an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics*. 2011;12(1):77.
45. Pons P, Latapy M. *Computing communities in large networks using random walks. In: International symposium on computer and information sciences*. Berlin, Heidelberg: Springer; 2005:284-293.
46. Rosvall M, Bergstrom CT. Maps of random walks on complex networks reveal community structure. *Proc Natl Acad Sci U S A*. 2008;105(4):1118-1123. http://doi.org/10.1073/pnas.0706851105
47. Allaire J, Ellis P, Gandrud C, et al. Package "networkD3". D3 JavaScript Network Graphs from R. Accessed March 3, 2025. 2017. https://cran.r-project.org/web/packages/networkD3/index.html
48. McCormick EM, Lott MT, Dulik MC, et al. Specifications of the ACMG/AMP standards and guidelines for mitochondrial DNA variant interpretation. *Hum Mutat*. 2020;41(12):2028-2057. http://doi.org/10.1002/humu.24107
49. Martínez-Camblor P, Pardo-Fernández JC. The Youden index in the generalized receiver operating characteristic curve context. *Int J Biostat*. 2019;15(1). /j/ijb.2019.15.issue-1/ijb-2018-0060/ijb-2018-0060.xml https://doi.org/10.1515/ijb-2018-0060
50. Laurie S, Steyaert W, de Boer E, et al. Genomic reanalysis of a pan-European rare-disease resource yields new diagnoses. *Nat Med*. 2025;31(2):478-489. http://doi.org/10.1038/s41591-024-03420-w
51. 100,000 Genomes Project Pilot Investigators, Smedley D, Smith KR, et al. 100,000 genomes pilot on rare-disease diagnosis in health care – preliminary. *N Engl J Med*. 2021;385(20):1868-1880. http://doi.org/10.1056/NEJMoa2035790
52. Stenton SL, Laricchia K, Lake NJ, et al. Mitochondrial DNA variant detection in over 6,500 rare disease families by the systematic analysis of exome and genome sequencing data resolves undiagnosed cases. *HGG Adv*. 2025;6(3):100441. http://doi.org/10.1016/j.xhgg.2025.100441
53. Bullich G, Matalonga L, Pujadas M, et al. Systematic collaborative reanalysis of genomic data improves diagnostic yield in neurologic rare diseases. *J Mol Diagn*. 2022;24(5):529-542. http://doi.org/10.1016/j.jmoldx.2022.02.003
54. Ratnaike T, Paramonov I, Olimpio C, et al. Mitochondrial DNA disease discovery through evaluation of genotype and phenotype data: the Solve-RD experience. *Am J Hum Genet*. 2025;112(6):1376-1387. http://doi.org/10.1016/j.ajhg.2025.04.003