EBioMedicine 54 (2020) 102730

Contents lists available at ScienceDirect

EBioMedicine

journal homepage: www.elsevier.com/locate/ebiom

Research paper

# Lifetime risk of autosomal recessive mitochondrial disorders calculated from genetic databases

Jing Tan[a,b,1,2], Matias Wagner[a,c,d,*,1], Sarah L. Stenton[a,c], Tim M. Strom[a,c], Saskia B. Wortmann[a,c,e], Holger Prokisch[a,c], Thomas Meitinger[a,c], Konrad Oexle[h], Thomas Klopstock[b,f,g,**]

[a] Institute of Human Genetics, School of Medicine, Technische Universität München, Munich, Germany
[b] Friedrich-Baur-Institute, Department of Neurology, University Hospital, LMU Munich, Munich, Germany
[c] Institute of Human Genetics, Helmholtz Zentrum München, Neuherberg, Germany
[d] Institute of Neurogenomics, Helmholtz Zentrum München, Neuherberg, Germany
[e] Department of Pediatrics, University Children's Hospital, Paracelsus Medical University (PMU), Salzburg, Austria
[f] German Center for Neurodegenerative Diseases (DZNE), Munich, Germany
[g] Munich Cluster for Systems Neurology (SyNergy), Munich, Germany
[h] Institute of Neurogenomics, Neurogenetic Systems Analysis Unit, Helmholtz Zentrum München, Neuherberg, Germany

ARTICLE INFO

*Article History:*
Received 10 January 2020
Revised 25 February 2020
Accepted 5 March 2020
Available online xxx

*Keywords:*
Autosomal recessive mitochondrial disorders
Population genetics
Prevalence
Lifetime risk
POLG
SPG7

ABSTRACT

*Background:* Mitochondrial disorders are a group of rare diseases, caused by nuclear or mitochondrial DNA mutations. Their marked clinical and genetic heterogeneity as well as referral and ascertainment biases render phenotype-based prevalence estimations difficult. Here we calculated the lifetime risk of all known autosomal recessive mitochondrial disorders on basis of genetic data.

*Methods:* We queried the publicly available Genome Aggregation Database (gnomAD) and our in-house exome database to assess the allele frequency of disease-causing variants in genes associated with autosomal recessive mitochondrial disorders. Based on this, we estimated the lifetime risk of 249 autosomal recessive mitochondrial disorders. Three of these disorders and phenylketonuria (PKU) served as a proof of concept since calculations could be aligned with known birth prevalence data from newborn screening reports.

*Findings:* The estimated lifetime risks are very close to newborn screening data (where available), supporting the validity of the approach. For example, calculated lifetime risk of PKU (16·0/100,000) correlates well with known birth prevalence data (18·7/100,000). The combined estimated lifetime risk of 249 investigated mitochondrial disorders is 31·8 (20·9–50·6)/100,000 in our in-house database, 48·4 (40·3–58·5)/100,000 in the European gnomAD dataset, and 31·1 (26·7–36·3)/100,000 in the global gnomAD dataset. The disorders with the highest lifetime risk (> 3 per 100,000) were, in all datasets, those caused by mutations in the *SPG7*, *ACADM*, *POLG* and *SLC22A5* genes.

*Interpretation:* We provide a population-genetic estimation on the lifetime risk of an entire class of monogenic disorders. Our findings reveal the substantial cumulative prevalence of autosomal recessive mitochondrial disorders, far above previous estimates. These data will be very important for assigning diagnostic a priori probabilities, and for resource allocation in therapy development, public health management and biomedical research.

*Funding:* German Federal Ministry of Education and Research.



## 1. Background

Mitochondrial disorders (MDs) are a group of clinically and genetically extremely heterogeneous disorders. Here we use the term MDs to describe diseases with a primary defect in the entire route of the pyruvate oxidation process, including the pyruvate dehydrogenase complex, the citrate cycle and the respiratory chain including ATP synthase (respiratory chain complexes I-V). In this regard, the requisite

* Corresponding author at: Institute of Human Genetics, Technische Universität München, Trogerstr. 32, 81675 Munich, Germany.
** Corresponding author at: Friedrich-Baur-Institute, Department of Neurology, University of Munich, Ziemssenstr. 1, 80336 Munich, Germany.
*E-mail addresses:* matias.wagner@mri.tum.de (M. Wagner), tklopsto@med.LMU.de (T. Klopstock).
[1] These authors contributed equally.
[2] Present address: Department of Neurology, the Second Affiliated Hospital of Dalian Medical University, Dalian, China

https://doi.org/10.1016/j.ebiom.2020.102730

**Research in context**

*Evidence before this study*

Mitochondrial disorders (MDs), primary defects in mitochondrial energy metabolism, are a diagnostically challenging, clinically and genetically heterogeneous group of diseases, caused by mitochondrial or nuclear DNA mutations. Knowledge of disease prevalence is crucial in guiding physicians' attention in patient care and resource allocation in both public health management and biomedical research. Regarding MDs with autosomal inheritance, the most comprehensive study so far identified 62 clinically affected individuals in the population of North East England equating to a minimum point prevalence of 2.9 (95% CI 2.2−3.7) in 100,000 and an estimated lifetime risk of 5.9 (5.0−6.9) in 100,000.

*Added value of this study*

We utilised the publicly available gnomAD database comprising 123,136 exomes and 15,496 genomes, as well as our own 14,130 in-house exomes to provide the lifetime risk data for 249 nuclear-encoded autosomal recessive MDs. The combined lifetime risk estimate was up to 48·4 in 100,000, corresponding to almost 1 in 2000 people. This implies that of the appr. 5 million newborn babies per year in the European Union, around 2500 will develop a nuclear-encoded recessive MD during their lifetime. Moreover, we could rank all 249 disorders by their prevalence and define the disorders with the highest lifetime risk as being caused by mutations in the *SPG7, ACADM, POLG* and *SLC22A5* genes.

*Implications of all the available evidence*

The study highlights the substantial cumulative prevalence of autosomal recessive MDs and suggests that previous phenotype-based epidemiological investigations largely underestimated their prevalence. These data are valuable for clinical, genetic and pharmaceutical applications alike. Knowledge of disease prevalence is crucial in guiding physicians' attention in patient care and resource allocation in both public health management and biomedical research. The data are especially relevant in view of the substantial and further growing number of MDs amenable to specific treatment strategies.

"support machinery" of mitochondrial DNA (mtDNA)-related protein synthesis (including mtDNA replication, mitochondrial RNA metabolism, and mitochondrial translation), the metabolism of mitochondrial cofactors and finally mitochondrial homeostasis (including protein import into mitochondria, lipid metabolism, fusion and fission, quality control etc.) is accordingly taken into account. MDs can result from mutations in both the mtDNA and the nuclear DNA (nDNA) genes, since both genomes code for mitochondrial proteins [1]. While the mtDNA codes for two rRNAs, 22 tRNAs, and 13 poly-peptides, it is estimated that mutations in the mtDNA are responsible for 15−30% of childhood-onset and over 50% of adult-onset cases of MD [2,3]. The proteins encoded by the nDNA comprise the large majority of respiratory chain complex I - V subunits, their assembly factors, and proteins involved in the "support machinery" [4−6]. To date, pathogenic variants in more than 306 nDNA genes have been identified and the number continues to grow on a high pace [5].

Currently, there is only limited epidemiological data on MDs. The point prevalence of childhood-onset and adult-onset MDs combined was estimated to be at least 20 in 100,000 [7]. For childhood-onset disease, the point prevalence was reported as 4·7 - 6·2/100,000 in studies where the clinical diagnosis was confirmed by biochemical or genetic evidence [8,9]. For adult-onset MD, a combined point prevalence of 12·5/100,000 was reported in northern England, with a prevalence of 9·6/100,000 for mtDNA mutations and 2·9/100,000 for nDNA mutations [10]. To date, prevalence estimations were based on clinical diagnoses, potentially generating an ascertainment bias towards patients with classical symptoms of MD and overlooking those with atypical presentations.

Therefore, taking a different approach, we calculated the lifetime risk of nuclear-encoded MDs based on the frequency of pathogenic and likely pathogenic variants in genetic databases under the assumption of the Hardy-Weinberg equilibrium. This is the first study providing lifetime risk data for all autosomal recessive MDs using a population-based genotype approach. These data are valuable for clinical, genetic and pharmaceutical considerations alike.

## 2. Methods

Fig. 1. shows a schematic diagram of the experimental design. Our study was performed following the GATHER statement guidelines [11].

All raw data used for this study has been uploaded to figshare (https://figshare.com/; DOI: 10.6084/m9.figshare.11366027).

### 2.1. Ethics

The study was conducted within a research project approved by the local ethics committee of the Technical University in Munich (#5360/12S).

### 2.2. Defining the gene list

We thoroughly reviewed the literature available via PubMed to collect a comprehensive list of MD disease genes. We included all genes coding for enzymes involved in the (1) pyruvate oxidation process, including the pyruvate dehydrogenase complex, (2) the citrate cycle, (3) the respiratory chain including ATP synthase (respiratory chain complex I-V), (4) mtDNA replication, mitochondrial RNA metabolism and mitochondrial translation, (5) the metabolism of mitochondrial cofactors and their metabolism and, finally (6) mitochondrial homeostasis (including protein import into mitochondria, lipid metabolism, fusion and fission, quality control etc.) that were reported in association with MD in two or more independent patients [12].

In total, 306 genes were selected of which 249 are encoded by the nuclear DNA and have been associated with autosomal recessive MDs. These were used to estimate the total calculated lifetime risk (Supplementary Table 1).

### 2.3. Defining the set of pathogenic variants

First, we assessed the publicly available databases ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/), Human Gene Mutation Database (HGMD, http://www.hgmd.cf.ac.uk/ac/index.php), and Human DNA Polymerase Gamma Mutation Database (https://tools.niehs.nih.gov//polg/, as of May 2018) and collected all variants which have been submitted as "pathogenic" or "likely pathogenic" in at least one of these databases. According to the American College of Medical Genetics and Genomics (ACMG), "likely pathogenic" correlates to a probability of 90% that a variant will be disease-causing [13]. We assumed that the subgroup of variants of uncertain significance (VUS) that are in fact disease-causing would compensate for the proportion of likely pathogenic variants that are benign polymorphisms. Additionally, our in-house exome database was queried for variants not listed in any of the above databases but classified as pathogenic or likely pathogenic according to the guidelines for the interpretation of sequence variants developed by the ACMG [13]. All frameshift, nonsense and splice variants were included as they most likely result in a loss of function.

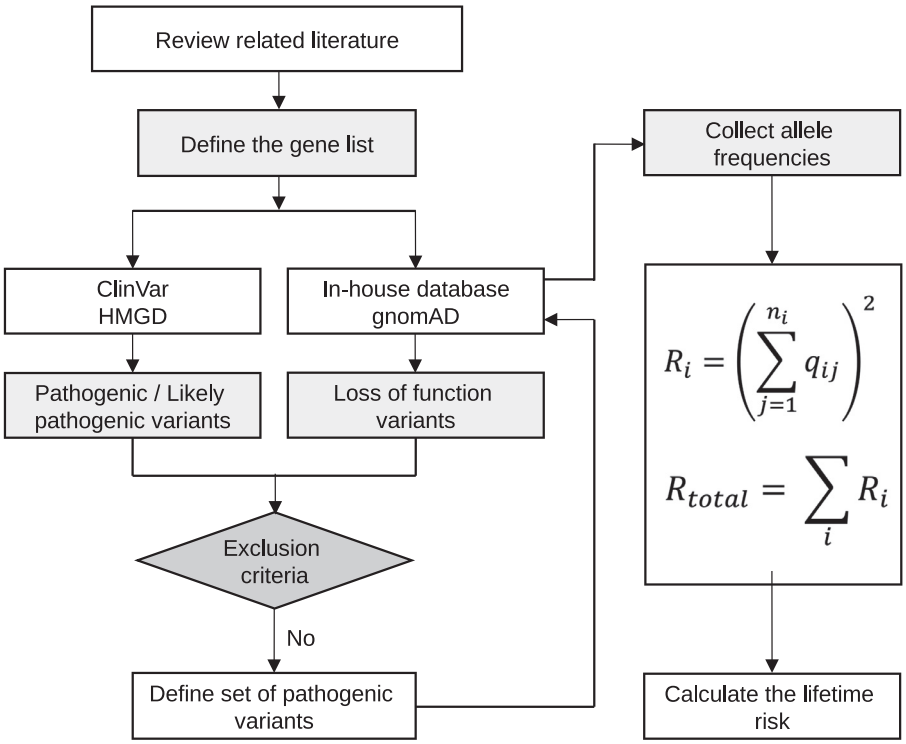


**Fig. 1.** Analysis diagram showing the experimental design. A comprehensive set of genes for autosomal recessive mitochondrial disorders was defined based on a review of the literature. ClinVar and HGMD were queried to collect the total number of disease causing variants for each gene. In addition, loss of function variants in our in-house database and gnomAD were considered pathogenic. These variants were evaluated for their pathogenicity according to the ACMG guidelines. The lifetime risk of mitochondrial disorders was calculated based on the allele frequencies of these variants in the gnomAD dataset as well as in our in-house database.

Second, in order to avoid an overestimation, we took a number of precautions. The pathogenicity of each variant from our list was re-evaluated with the information available in public databases. Variants were excluded if they were 1) listed as VUS in ClinVar, 2) found in a homozygous state in gnomAD or ExAC and 3) when publications listed in PubMed raised doubt on pathogenicity. For variants with conflicting interpretation of pathogenicity in ClinVar we reviewed the literature and evaluated the pathogenicity according to the ACMG criteria.

For the phenylalanine hydroxylase gene *(PAH)*, we first collected 837 variants from ClinVar, HMGD and our in-house database, and later excluded 277 (33·1%) variants, mostly because of uncertain significance, summing up to 560 likely pathogenic or pathogenic variants. For the 249 MD genes analysed in the present paper, we collected 37,857 variants and later excluded 530 (1·4%) of the variants (as listed in Supplement 2).

### 2.4. Estimation of the lifetime risks of diseases

The lifetime risk is defined as the proportion of a population that at some point in life will develop the disease. The expected lifetime risk $R_i$ for an autosomal recessive MD caused by mutations in nDNA gene $i$ was calculated from the sum of the allele frequencies $q_{ij}$ of the $n_j$ disease-causing variants in the respective gene under the assumption of Hardy-Weinberg equilibrium and mutual independence of these rare variants. Biallelic combinations of variants were considered to be fully penetrant. Accordingly:

$$R_i = (q_{i1} + q_{i2} + \ldots + q_{in_i})^2 = \left(\sum_{j=1}^{n_i} q_{ij}\right)^2$$

The combined lifetime risk $R_{total}$ for developing one of the assessed MDs was calculated by summation of the lifetime risks of each of the diseases, assuming that these rare disorders are independent of each other and do not occur together.

$$R_{total} = \sum_i R_i = \sum_i \left(\sum_{j=1}^{n_i} q_{ij}\right)^2$$

Due to low numbers, 95% confidence intervals (95%CI) were calculated using the Clopper-Pearson Exact method [14].

Two databases were utilized to assess the allele frequencies of disease-causing variants in the general population. First, the genome Aggregation Database (gnomAD, http://gnomad.broadinstitute.org/) comprising 123,136 exomes and 15,496 genomes from unrelated individuals of various disease-specific and population genetic studies was queried. In gnomAD, allele frequencies of all identified variants are provided for different ethnic backgrounds. We assessed the prevalence in both the European (Non-Finnish) population and in the overall dataset encompassing European (Non-Finnish), Finnish European, African, Latino, Ashkenazi Jewish and East and South Asian individuals. Second, the in-house database of the Institute of Human Genetics, containing exome sequencing data of healthy unrelated individuals and patients with various genetic disorders, was used as an independent data source. To prevent selection bias in the analysis of our in-house database, we excluded individuals with homozygous or compound heterozygous variants, which were causative for the respective disease as well as their parents, leaving 14,130 individuals (28,260 alleles) in May 2018.

To verify the validity and feasibility of our method, we used phenylketonuria (PKU) as proof of concept. PKU is an autosomal recessive disease resulting from deficiency of phenylalanine hydroxylase (PAH). There are no predominant common mutations in PKU, which ensures that pathogenic variants segregate independently of each other. The lifetime risk of PKU is equivalent to the birth prevalence in population-wide metabolic newborn screening programs. There is a significant variability of country-specific prevalence of PKU in Europe [15]. As the European gnomAD dataset is mainly based on individuals living in North-America

**Table 1**
Comparison of the calculated lifetime prevalence with the birth prevalence of PKU, MCADD, BD and VLCADD per 100,000 according to the German National Screening Report.

| | | PKU | MCADD | BD | VLCADD |
|---|---|---|---|---|---|
| Calculated liftetime risk | gnomAD dataset (European) | 16·0 (14·5–17·6) | 7·9 (7·0–8·9) | 2·1 (1·8–2·5) | 1·3 (1·1–1·6) |
| | gnomAD dataset (worldwide) | 8·2 (7·6–8·9) | 3·3 (3·0–3·6) | 1·2 (1·0–1·3) | 0·64 (0·55–0·74) |
| | in-house database | 12·3 (9·8–15·3) | 4·8 (3·6–6·3) | 0·35 (0·20–0·60) | 0·45 (0·26–0·75) |
| Average birth prevalence in Germany[16] | | 18·7 (16·9–21·6) | 9·8 (7·6–12·4) | 3·9 (2·6–5·7) | 1·2 (0·5–2·3) |

that largely descend from immigrants from Britain and Germany we used the newborn screening results made available by the German Society for Newborn Screening ("Deutsche Gesellschaft für Neugeborenenscreening", DGNS e.V., http://www.screening-dgns.de), from 2004 to 2015 (Table 1) as a control dataset [16]. These data were compared to the calculated lifetime risk for PKU, based on the variant allele frequencies in datasets of European origin in gnomAD and in our in-house database. In addition, biotinidase deficiency (BD), medium chain Acyl-CoA dehydrogenase deficiency (MCADD) and very-long-chain-acyl-CoA-dehydrogenase deficiency (VLCADD) which are mitochondrial disease genes and for which newborn screening data is available were used as further validation.

Overestimation of the lifetime risk could occur in this approach in benign variants being falsely interpreted as pathogenic (non-sampling error) or in the setting of pathogenic variants conferring reduced penetrance (parameter uncertainty). Conversely, underestimation of the lifetime risk could occur as a consequence of exclusion of variants not identified by exome sequencing such as pathogenic large deletions or insertions as well as intronic splice variants (non-sampling error). In addition, not all coding variants, which are predicted to be pathogenic, were considered as we limited screening to those which have already been published or are present on well-established platforms such as ClinVar (non-sampling error). To date, there are no reports of polygenic inheritance for MD, which could lead to specification uncertainty if present. Of note, our calculations are based on the gnomAD as well as our in-house database, which do not represent an ideal random sample therefore leading to sampling error. This however is reduced by comparing the results from two databases.

### 2.5. Statistical analysis

Statistical analysis was performed using R 3.6.2. Comparison between the calculated lifetime risks based on the European gnomAD population and the overall dataset was done using a 2-tailed Wilcoxon signed ranks test using the R package "MASS". Lin's Concordance Correlation Coefficient was calculated using the R package "DescTools". Bland-Altman-Plots were drawn using the R package "BlandAltmanLeh".

## 3. Results

### 3.1. Estimation of the lifetime risk of PKU, MCADD, BD and VLCADD as a proof of concept

A total number of 174 (likely) pathogenic variants in *PAH* was identified in the gnomAD dataset. The sum of their allele frequencies was 0·0111 in the European (Non-Finnish) and 0·0091 in the total gnomAD dataset resulting in a calculated PKU lifetime risk of 16·0 (14·5–17·6 95% CI)/100,000 in the European (Non-Finnish) population and of 8·2 (7·6–8·9)/100,000 in the total dataset. In the in-house database, 82 variants were identified collectively with 313 disease-causing alleles among the 14,130 individuals. Accordingly, the combined frequency of these variants was 0·0111 resulting in a calculated PKU prevalence of 12·3 (9·8–15·4)/100,000. These estimated lifetime risks for PKU, in particular the one for the European gnomAD dataset (16·0 (14·5–17·6)/100,000) are very close to the birth prevalence in the German newborn screening (18·7 (16·9–21·6)/100,000). Moreover, birth prevalences from newborn screening were also available for three recessive MDs. Again, the calculated lifetime risks were very similar to the birth prevalences in medium chain Acyl-CoA dehydrogenase deficiency (MCADD, disease gene *ACADM*), biotinidase deficiency (BD, *BTD*) and very-long-chain-acyl-CoA-dehydrogenase deficiency (VLCADD, *ACADVL*) (see below, Table 1 and Fig. 2a).

### 3.2. Estimation of the lifetime risk of autosomal recessive MDs

The numbers of all (likely) pathogenic variants in the most frequent MD genes are depicted in Table 2 and Fig. 2 and in all MD genes in Supplementary Table 3. A detailed list of these variants and their associated allele frequencies in our in-house database and gnomAD can be found in the online material on figshare (https://figshare.com/).

As the most frequent autosomal recessive MDs we found adult-onset neurological disorders due to mutations in *SPG7* and *POLG*. *SPG7* encodes a mitochondrial metalloprotease, and biallelic mutations of *SPG7* cause *spastic paraplegia 7* (MIM #607259) [17]. Based on the gnomAD dataset, the calculated lifetime risk is 8·4 (7·5–9·4)/100,000 in Europe (Non-Finnish), and 3·7 (3·4–4·1)/100,000 worldwide. Based on the in-house database, the calculated lifetime risk is 5·2 (3·8–6·6)/100,000.

*POLG* encodes mitochondrial DNA polymerase gamma. Mutations of *POLG* cause a broad range of autosomal recessive disorders ranging from an often early lethal *mitochondrial DNA depletion syndromes 4A/B* (MIM #203700/#613662), *progressive external ophthalmoplegia* (MIM #258450) to *mitochondrial recessive ataxia syndrome* (MIM #607459). Additionally, autosomal dominant progressive external ophthalmoplegia due to *POLG* mutations has been described (MIM #157640). Genotype-phenotype comparison shows that different variants cause different conditions depending on their location and their potential to reduce exonuclease or polymerase activity, therefore complicating our assessment of the lifetime risk of *POLG*-related disorders [18,19]. In this study, we analysed all pathogenic variants independent of the inheritance mode, as the combination of these variants in a biallelic state, though linked with different phenotypes, will all be disease-causing. Our method of estimating the lifetime risk of a genetic disorder cannot be applied to autosomal dominant disorders as population databases will be depleted of their pathogenic variants. Therefore, the *POLG*-associated autosomal dominant disorders are not included in the analyses. The allele frequency of *POLG* mutations was 0·0083 in the European (Non-Finnish) gnomAD dataset, 0·0061 in the total gnomAD dataset and 0·0074 in the in-house database, resulting in lifetime risk estimates for *POLG*-associated disorders of 6·9 (6·1–7·8), 3·7 (3·4–4·0) and 3·7 (3·3–4·1)/100,000, respectively. Previously, the point prevalence of *POLG*-related mitochondriopathies was reported as 0·3 (0·1–0·6)/100,000, not including early onset mitochondrial depletion syndrome 4A which is often lethal in infancy [10].

Other disorders with a high calculated lifetime risk were inborn errors of metabolisms (IEMs) with secondary impairment of

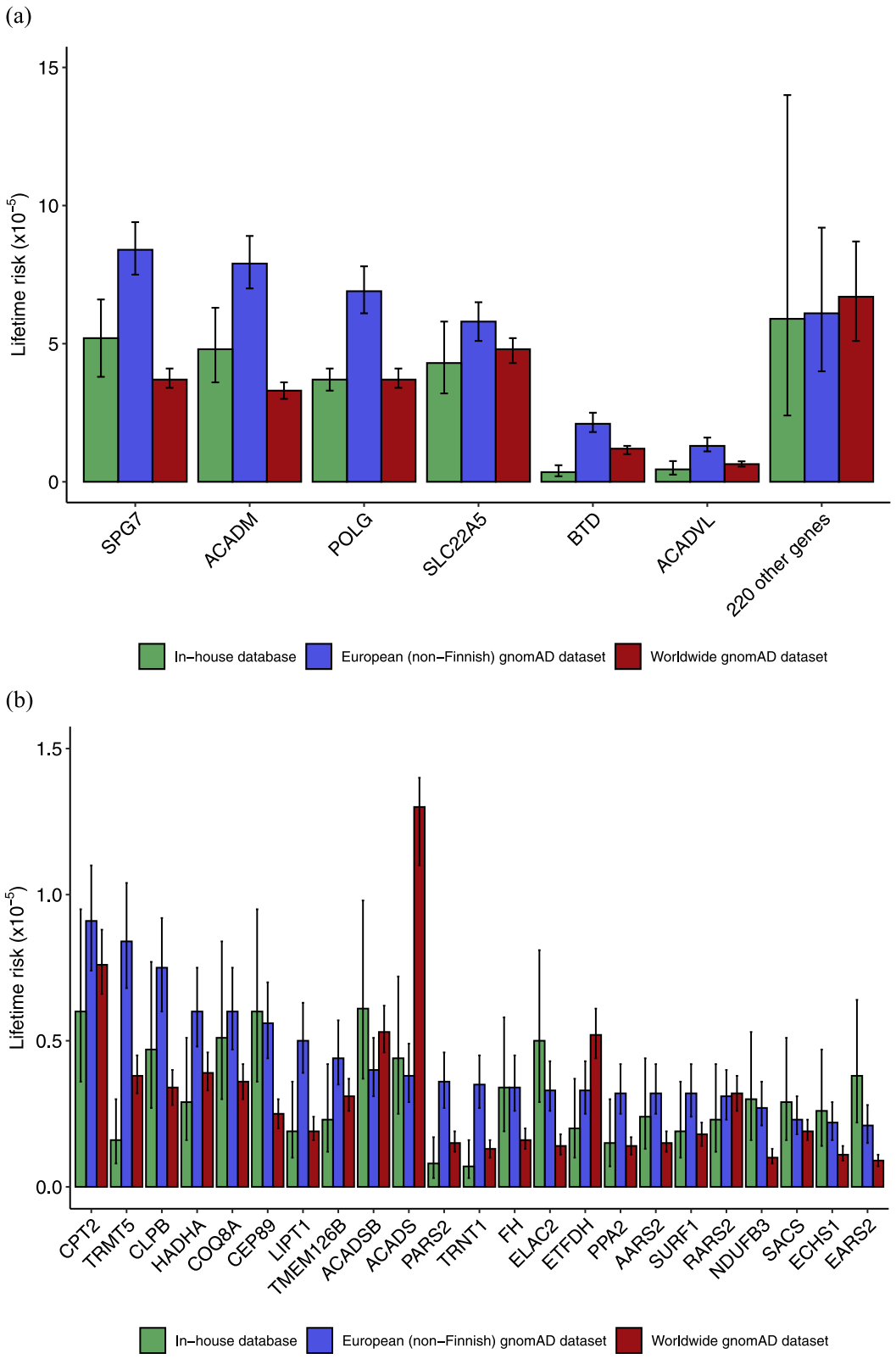


**Fig. 2.** Calculated life time risk for the most prevalent autosomal recessive mitochondrial disorders. Comparison of the lifetime risks of different monogenic nuclear mitochondrial diseases according to the in-house database and gnomAD dataset calculated independently for the European population and the overall dataset. Error bars represent 95%-confidence intervals. (a) depicts the lifetime risk for genes >1/100,000 in any population as well as for the remaining 220 genes. (b) shows the lifetime risk for genes with a lifetime risk >0.02/100,000 in any population.

mitochondrial function due to mutations in *ACADM, SLC22A5, BTD* and *ACADVL*.

For medium-chain-acyl-CoA-dehydrogenase deficiency (MCADD) caused by *ACADM* mutations, one of the most common inborn errors of fatty acid metabolism, 80 pathogenic/likely pathogenic variants were verified in the gnomAD dataset and the combined frequency of these variants is 0·0089 in the European (Non-Finnish) and 0·0057 in the total dataset, leading to a lifetime risk of 7·9 (7·0−8·9)/100,000 in

**Table 2**
Lifetime risk of the most frequent mitochondrial diseases per gene.

| Gene | Number of disease-causing variants in-house database | Number of disease-causing alleles in-house database | Number of disease-causing variants in gnomADdataset | Number of disease-causing alleles in gnomAD dataset (European, Non-Finnish population) | Number of disease-causing alleles in gnomAD dataset (worldwide) | Lifetime risk in-house database per 100,000 (95%CI) | Lifetime risk in European (Non-Finnish) population (gnomAD dataset) per 100,000 (95%CI) | Lifetime risk in worldwide population (gnomAD dataset) per 100,000 (95%CI) |
|---|---|---|---|---|---|---|---|---|
| ***SPG7*** | 31 | 195 | 107 | 1160 | 1684 | 5·2 (3·8–6·6) | 8·4 (7·5–9·4) | 3·7 (3·4–4·1) |
| ***ACADM*** | 24 | 195 | 80 | 1126 | 1584 | 4·8 (3·6–6·3) | 7·9 (7·0–8·9) | 3·3 (3·0–3·6) |
| ***POLG*** | 35 | 174 | 137 | 1050 | 1697 | 3·7 (3·3–4·1) | 6·9 (6·1–7·8) | 3·7 (3·4–4·1) |
| ***SLC22A5*** | 27 | 186 | 77 | 961 | 1911 | 4·3 (3·2–5·8) | 5·8 (5·1–6·5) | 4·8 (4·3–5·2) |
| ***BTD*** | 26 | 53 | 95 | 585 | 945 | 0·35 (0·20–0·60) | 2·1 (1·8–2·5) | 1·2 (1·0–1·3) |
| ***ACADVL*** | 25 | 261 | 108 | 789 | 1183 | 0·45 (0·26–0·75) | 1·3 (1·1–1·6) | 0·64 (0·55–0·74) |
| ***CPT2*** | 16 | 69 | 72 | 382 | 766 | 0·60 (0·36–0·95) | 0·91 (0·74–1·10) | 0·76 (0·66–0·88) |
| ***TRMT5*** | 5 | 35 | 23 | 368 | 542 | 0·16 (0·08–0·30) | 0·84 (0·68–1·04) | 0·38 (0·32–0·45) |
| ***CLPB*** | 19 | 61 | 33 | 346 | 510 | 0·47 (0·27–0·77) | 0·75 (0·60–0·92) | 0·34 (0·28–0·40) |
| ***HADHA*** | 14 | 48 | 56 | 310 | 546 | 0·29 (0·16–0·51) | 0·60 (0·48–0·75) | 0·39 (0·33–0·46) |
| ***COQ8A*** | 26 | 64 | 66 | 309 | 523 | 0·51 (0·30–0·84) | 0·60 (0·47–0·75) | 0·36 (0·30–0·42) |
| ***CEP89*** | 19 | 69 | 41 | 300 | 434 | 0·60 (0·36–0·95) | 0·56 (0·44–0·70) | 0·25 (0·20–0·30) |
| ***LIPT1*** | 8 | 39 | 31 | 282 | 385 | 0·19 (0·10–0·36) | 0·50 (0·39–0·63) | 0·19 (0·16–0·24) |

the European (Non-Finnish) population and of 3·3 (3·0–3·6)/100,000 in the total dataset. In the in-house database, 24 different disease-causing variants were verified in 195 disease-causing alleles among the 14,130 individuals, implying a combined frequency of 0·0083, and leading to a calculated MCADD lifetime risk of 4·8 (3·6–6·3)/100,000.

Systemic primary carnitine deficiency (CDSP, MIM #212140) is a metabolic disorder of the carnitine cycle due to SLC22A5 mutations. The estimated lifetime risk of CDSP was 4·3 (3·2–5·8)/100,000 based on the in-house database, 5·8 (5·1–6·5)/100,000 in Europe (Non-Finnish), and 4·8 (4·3–5·2)/100,000 worldwide based on the gnomAD dataset.

*BTD* mutations lead to a disorder of biotin metabolism - biotinidase deficiency (BD, MIM #253260), and *ACADVL* mutations result in very long-chain acyl-CoA dehydrogenase deficiency (VLCADD, MIM #201475) which is an inborn error of fatty acid oxidation. The calculated lifetime risk of BD was 0·35 (0·20–0·60)/100,000 according to our in-house database, which was less than the risk of 2·1 (1·8–2·5)/100,000 according to the European (Non-Finnish) and 1·2 (1·0–1·3)/100,000 in the whole gnomAD dataset. The lifetime risk of VLCADD was 0·45 (0·26–0·75)/100,000 based on our in-house database, 1·3 (1·1–1·6)/100,000 and 0·64 (0·55–0·74)/100,000 according to the European and overall gnomAD dataset, respectively. The calculated lifetime risks of MCADD, VLCADD and BD were compared with data from the German newborn screening report further validating the method used in this study (see above and Table 1). MCADD has a calculated lifetime risk of 7·9 (7·0–8·9)/100,000 according to the European gnomAD dataset whereas the incidence is 9·8 (7·6–12·4) according to the newborn screening data. The calculated lifetime risks of BD and VLCADD were 2·1 (1·8–2·5)/100,000 and 1·3 (1·1–1·6)/100,000, respectively, comparing to 3·9 (2·6–5·7)/100,000 and 1·2 (0·5–2·3)/100,000.

Next, we calculated the minimal combined lifetime risk of all autosomal recessive MDs, resulting in a combined lifetime risk of 48·4 (40·3–58·5)/100,000 based on the European (Non-Finnish) population in the gnomAD database, 31·1 (26·7–36·3)/100,000 based on the worldwide gnomAD dataset, and 31·8 (20·9–50·6)/100,000 based on our in-house database. Notably, there were 29 MDs with a calculated risk of more than 0·02/100,000 according to the European (Non-Finnish) population (Fig. 2b). The combined risk of these 29 diseases was 25·9 (18·5–36·6)/100,000 based on our in-house database, 42·3 (36·3–49·2)/100,000 in Europe (Non-Finland), and 24·4 (21·6–27·5)/100,000 in total gnomAD, corresponding to 81·5%, 87·4% and 78·5% of the overall risks, respectively.

### *3.3. Calculating the estimated lifetime risk based in different populations based on loss of function variants*

Calculated lifetime risks for the individual monogenic disorders as well as the combined lifetime risk for autosomal recessive MD in general are depicted in Table 2. Bland-Altman plot and concordance correlation coefficient of the log-transformed risk values (Suppl. Fig. 1) revealed a high degree of correlation ($\rho_c$ = 0·938) between the risks derived from the European dataset and those derived from the worldwide dataset. On average, the native, i.e. non-transformed risks appeared to be lower ($p$ = 0·013, $Z$ = −2.5, Wilcoxon signed rank test) when based on the worldwide dataset (Fig. 2). We hypothesized that this discrepancy arises from a bias in the queried databases HGMD and ClinVar for ethnicity-specific pathogenic variants as most of the submitters are based in the USA and Europe (https://www.ncbi.nlm.nih.gov/clinvar/docs/submitter_list/). Therefore, we compared the estimated lifetime risks by only including loss of function variants (namely nonsense, frameshift and splice variants) in the calculations as these are considered disease-causing regardless of a listing in any mutation database. Apparently, there was no significant difference anymore between the lifetime risks based on the European and the worldwide dataset (Wilcoxon signed rank test: $p$ = 0.32, $z$ = −1.0).

We conclude, that the actual lifetime risks of most MDs are therefore closer resembled by the calculations based on the European gnomAD dataset.

## 4. Discussion

The gnomAD dataset provides allele frequencies for different ethnical backgrounds, allowing the calculation of lifetime risks for each of the selected MDs based on the European (Non-Finnish) population only, or on the entire dataset additionally including Finnish European, African, Latino, Ashkenazi Jewish, East Asian and South Asian individuals. It should be noted that the ethnical composition of gnomAD does not reflect the world population and the data therefore do not yield worldwide lifetime risks. By comparison, our in-house database comprises 14,130 exomes from mostly Caucasian individuals. The exact ethnical composition of the in-house database is however unknown, as the ethnical background was not routinely documented.

The method of the present study has been used previously to assess lifetime risk for single genes [20–24]. For conditions covered by newborn screening, the lifetime risk equals the biochemical prevalence at birth, therefore enabling us to compare our calculated PKU data with the newborn screening reports. The small difference of 2·7/100,000 between the calculated lifetime risk in Europe and the newborn screening data in Germany for PKU likely results from the fact that our lists of pathogenic mutations cannot be 100% complete and due to our inability to detect non-coding mutations as well as pathogenic deletions and duplications with exome sequencing data which are responsible for a significant subset of disease causing variants [25]. Unavoidably, there will always remain pathogenic mutations that have evaded identification.

We were also able to validate our method by comparing the lifetime risks of medium-chain-acyl-CoA-dehydrogenase deficiency (MCADD), very-long-chain-acyl-CoA-dehydrogenase deficiency (VLCADD) and biotinidase deficiency (BD) with data from the national newborn screening report. As depicted in Table 1, the calculated lifetime risk of these conditions based on the European gnomAD dataset approximates the data published by the German National Screening program, further validating the applied method. We conclude that our method for assessing the lifetime risk of autosomal recessive disorders provides reliable data, with minimal risk of overestimation due to stringent variant exclusion criteria.

Of note, we calculated the lifetime risk assuming full penetrance. There are no reports about reduced penetrance for autosomal recessive MDs. However, there are examples such as Thrombocytopenia absent radius (TAR, MIM #274000) syndrome where alleles are not disease-causing when present in homozygosity but in combination with a null allele [26]. We cannot exclude the presence of variants that convey reduced penetrance but our method and framework can serve as a basis to detect these when thoroughly comparing calculated data with data from rare disease registries that are being established at the moment.

Our data provides the first estimation of the minimal overall lifetime risk of recessive MDs collectively. This lifetime risk is 48·4 (40·3–58·5) and 31·1 (26·7–36·3)/100,000 based on the European and overall gnomAD dataset, respectively. One limitation of our study is that the employed method of calculating lifetime risks is only applicable for autosomal recessive diseases. We feel, however, that these data provide a very valuable complement to the epidemiological data known for mtDNA-related MDs. In accordance with the overall gnomAD dataset, our in-house database calculates a lifetime risk of 31·8 (20·9–50·6). The difference between the risks based on the European and the overall gnomAD data is significant ($p$ = 0·013, $Z$ = −2.5, Wilcoxon signed rank test). As the main reason for this difference we assume the European and American predominance in genetic research and diagnostics which likely has identified already most of the frequent Caucasian pathogenic variants while a number of variants from other ethnical backgrounds probably have not been published yet. This hypothesis is supported by the fact that there is no statistical difference between the calculated lifetime risks based on the European and the overall dataset when only considering loss of function variants listed in gnomAD (Wilcoxon signed rank test: $p$ = 0.32, $z$ = −1.0). Additionally, the selective pressure of consanguinity in certain communities could reduce the frequencies of carriers for autosomal recessive disorders.

In summary, we provide estimation of the lifetime risk of all known autosomal recessive MDs based on population genotypes. The combined lifetime risk estimate is up to 48·4 in 100,000, corresponding to almost 1 in 2000 people. This implies that around 2500 of the approximately 5 million newborn babies per year in the European Union will develop a nuclear-encoded recessive MD during their lifetime.

## Declaration of Competing Interest

TK reports grants from the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG), the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF) and the European Commission, outside the submitted work. TK also reports grants, personal fees, and non-financial and other support from ApoPharma Inc, Retrophin Pharmaceuticals, Santhera Pharmaceuticals, GenSight Biologics and Stealth Biotherapeutics, outside the submitted work. HP reports grants from the German Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung, BMBF). All other authors do not report any conflict of interests.

## Data sharing

Individual participant data will not be made available. Aggregate data, specifically the final list of variants that were rated as “pathogenic” and “likely pathogenic” and the respective number of carriers within the gnomAD dataset and the Munich in-house database will be made available. No other documents will be made available. Anyone who wishes to access the data for any purpose can access the data immediately following publication without an end date from Figshare (https://figshare.com/articles/Untitled_ItemLIFETIME_RISK_OF_AUTOSOMAL_RECESSIVE_MITOCHONDRIAL_DISORDERS_CALCULATED_FROM_GENETIC_DATABASES/11366027).

## Acknowledgements

This work was facilitated by the German Federal Ministry of Education and Research (BMBF, Bonn, Germany) through a grant to the German Network for Mitochondrial Disorders (mitoNET, 01GM1906A to TK and a grant for the E-Rare project GENOMIT (01GM1603 and 01GM1207 to HP and TK)). None of the funders had any role in the study design, data collection, data analysis, interpretation, or writing of the manuscript.

## Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.ebiom.2020.102730.

## References

[1] Wagner M, Berutti R, Lorenz-Depiereux B, et al. Mitochondrial DNA mutation analysis from exome sequencing-A more holistic approach in diagnostics of suspected mitochondrial disease. J Inherit Metab Dis 2019;42(5):909–17.
[2] Gorman GS, Chinnery PF, DiMauro S, et al. Mitochondrial diseases. Nat Rev Dis Primers 2016;2(1):16080.

[3] Chinnery PF. Mitochondrial disease in adults: what's old and what's new? EMBO Mol Med 2015;7(12):1503–12.
[4] Koopman WJ, Willems PH, Smeitink JA. Monogenic mitochondrial disorders. The N Engl J Med 2012;366(12):1132–41.
[5] Wortmann SB, Mayr JA, Nuoffer JM, Prokisch H, Sperl W. A guideline for the diagnosis of pediatric mitochondrial disease: the value of muscle and skin biopsies in the genetics era. Neuropediatrics 2017;48(4):309–14.
[6] Alston CL, Rocha MC, Lax NZ, Turnbull DM, Taylor RW. The genetics and pathology of mitochondrial disease. J Pathol 2017;241(2):236–50.
[7] Schaefer AM, Taylor RW, Turnbull DM, Chinnery PF. The epidemiology of mitochondrial disorders–past, present and future. Biochim. Biophys. Acta 2004;1659(2–3):115–20.
[8] Darin N, Oldfors A, Moslemi AR, Holme E, Tulinius M. The incidence of mitochondrial encephalomyopathies in childhood: clinical features and morphological, biochemical, and DNA abnormalities. Ann Neurol 2001;49(3):377–83.
[9] Skladal D, Halliday J, Thorburn DR. Minimum birth prevalence of mitochondrial respiratory chain disorders in children. Brain: A J Neurol 2003;126(Pt 8):1905–12.
[10] 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):753–9.
[11] Stevens GA, Alkema L, Black RE, et al. Guidelines for accurate and transparent health estimates reporting: the gather statement. Lancet 2016;388(10062):e19–23.
[12] Mayr JA, Haack TB, Freisinger P, et al. Spectrum of combined respiratory chain defects. J. Inherit. Metab. Dis. 2015;38(4):629–40.
[13] 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 Off J Am College Med Genet 2015;17(5):405–24.
[14] Clopper CJ, Pearson ES. THE use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika 1934;26(4):404–13.
[15] Loeber JG. Neonatal screening in Europe; the situation in 2004. J Inherit Metab Dis 2007;30(4):430–8.
[16] The National Screening Report (Germany 2015). Deutsche Gesellschaft für Neugeborenenscreening (DGNS e.V.); 2015.
[17] Casari G, De Fusco M, Ciarmatori S, et al. Spastic paraplegia and OXPHOS impairment caused by mutations in paraplegin, a nuclear-encoded mitochondrial metalloprotease. Cell 1998;93(6):973–83.
[18] Luoma PT, Luo N, Loscher WN, et al. Functional defects due to spacer-region mutations of human mitochondrial DNA polymerase in a family with an ataxia-myopathy syndrome. Hum Mol Genet 2005;14(14):1907–20.
[19] Sohl CD, Kasiviswanathan R, Copeland WC, Anderson KS. Mutations in human DNA polymerase gamma confer unique mechanisms of catalytic deficiency that mirror the disease severity in mitochondrial disorder patients. Hum Mol Genet 2013;22(6):1074–85.
[20] Fitterer B, Hall P, Antonishyn N, Desikan R, Gelb M, Lehotay D. Incidence and carrier frequency of SANDHOFF disease in Saskatchewan determined using a novel substrate with detection by tandem mass spectrometry and molecular genetic analysis. Mol. Genet. Metab. 2014;111(3):382–9.
[21] Appadurai V, DeBarber A, Chiang PW, et al. Apparent underdiagnosis of cerebrotendinous xanthomatosis revealed by analysis of ~60,000 human exomes. Mol Genet Metab 2015;116(4):298–304.
[22] Brezavar D, Bonnen PE. Incidence of PKAN determined by bioinformatic and population-based analysis of ~140,000 humans. Mol Genet Metab 2019;128(4):463–9.
[23] Gao J, Brackley S, Mann JP. The global prevalence of Wilson disease from next-generation sequencing data. Genet Med Off J Am College Med Genet 2019;21(5):1155–63.
[24] Del Angel G, Hutchinson AT, Jain NK, Forbes CD, Reynders J. Large-scale functional LIPA variant characterization to improve birth prevalence estimates of lysosomal acid lipase deficiency. Hum Mutat 2019;40(11):2007–20.
[25] Groselj U, Tansek MZ, Kovac J, Hovnik T, Podkrajsek KT, Battelino T. Five novel mutations and two large deletions in a population analysis of the phenylalanine hydroxylase gene. Mol Genet Metab 2012;106(2):142–8.
[26] Albers CA, Paul DS, Schulze H, et al. Compound inheritance of a low-frequency regulatory snp and a rare null mutation in exon-junction complex subunit RBM8A causes TAR syndrome. Nat Genet 2012;44(4):435–9 s1-2.