bioRxiv preprint doi: https://doi.org/10.1101/798264; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Selective constraints and pathogenicity of mitochondrial DNA variants inferred from a novel database of 196,554 unrelated individuals.

Alexandre Bolze[1,*,#], Fernando Mendez[1,*], Simon White[1], Francisco Tanudjaja[1], Magnus Isaksson[1], Misha Rashkin[1], Johnathan Bowes[1], Elizabeth T. Cirulli[1], William J. Metcalf[2,3], Joseph J. Grzymski[2,3], William Lee[1], James T. Lu[1], Nicole L. Washington[1,#]

[1] Helix, San Mateo, California
[2] Desert Research Institute, Reno, Nevada
[3] Renown Institute of Health Innovation, Reno, Nevada

*Equal contribution
[#] Correspondence: alexandre.bolze@helix.com or nicole.washington@helix.com

Abstract

Robust characterization of mitochondrial variation provides an opportunity to map regions under high constraint, and identify essential functional domains. We sequenced the mitochondrial genomes of 196,554 unrelated individuals, and identified 15,035 unique variants. We found that 47% of the mitochondrial genome was invariant across the population, and generated a map of constrained mitochondrial regions. We find that the longest intervals in the mitochondrial genome without any variant were in the two rRNA genes (26 of 40 intervals >10bp long). We also showed that the 13 protein-coding genes in the mitochondrial genome did not tolerate loss-of-function variants. The only frameshift or nonsense variant observed at homoplasmic levels was a nonsense at the start codon of *MT-ND1*, which may be rescued by the methionine at amino acid position 3. Lastly, we applied these data to variants reported to be pathogenic for Leber's Hereditary Optic Neuropathy (LHON). We found that 42% of variants (19 of the 45) reported to be pathogenic have a frequency above the maximum credible population allele frequency for an LHON-causing variant, including the primary LHON mutation m.14484T>C, which suggests that m.14484T>C cannot be causing LHON by itself. This result showed that allele frequency information across a large unselected population is important to assess the pathogenicity of variants in the context of rare mitochondrial disorders. We made HelixMTdb -- the list of variants and their allele frequency in 196,554 unrelated individuals -- publicly available.

bioRxiv preprint doi: https://doi.org/10.1101/798264; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Introduction

Mitochondrial diseases are among the most common of inherited disorders, with an estimated combined prevalence of 1 in 5,000 [1,2]. Mitochondrial disorders can be caused by variants encoded in nuclear (nDNA) or mitochondrial DNA (mtDNA); we focus here specifically on mtDNA variants. The mitochondrial genome codes for 13 protein-coding genes, 22 transfer RNA (tRNA) genes, 2 ribosomal RNA (rRNA) genes, and the non-genic displacement (D)-loop [3]. Unlike the nuclear genome, there are no introns, and very few non-coding bases in between genes. Human mitochondria are inherited through the maternal line, and there are multiple copies (>>2 copies) in every cell. Mitochondrial DNA can be uniform in sequence (homoplasmic) or can have variable sequences (heteroplasmy) within an individual cell. In addition, their unequal inheritance during cell division may result in changes to mutational load (% of the mitochondria carrying the mutation) during both the lifespan of an individual and across different cell types or tissues[4]. For individuals with heteroplasmic deleterious variants, those differences in mutational load can lead to varying phenotypic presentations of the same disease [5,6].

The analysis of genetic variation in the population has been an efficient tool to understand the role and essentiality of genes, functional domains, and to assess the pathogenicity of variants underlying rare disease. For example, scientists have recently drawn maps of constrained coding regions using the Genome Aggregation Database (gnomAD) [7], which highlighted regions depleted of non-synonymous variants across a large adult population. These regions pointed to genes and functional domains (some of them without any known function) that may cause severe developmental phenotypes when mutated[7]. Another example was the development of a framework using population allele frequency information to assess whether a variant is “too common” to be pathogenic for a specific disease given the prevalence of this disease in the population, and its assumed genetic architecture [8]. These two examples illustrate how large databases such as Bravo [9] and gnomAD [10] have been used to help the interpretation of the human genome. However, at the time of writing this manuscript, these two databases did not have information on variants in the human mitochondrial genome.

At the time of writing this manuscript, MITOMAP [11] and HmtDB [12] were the two largest publicly available databases of human mtDNA variants which have been used to assess the pathogenicity of mtDNA variants. For example, MITOMAP was used to select candidate mtDNA variants that may cause tubulointerstitial kidney disease [13]. However, both databases gathered mtDNA variant information drawn from nearly the same ~50,000 full mitochondrial genomes reported in GenBank, which has some caveats. Mitochondrial genome sequences uploaded in GenBank come from different sources, and are of unequal quality. Moreover, these databases are affected by biases in recruitment and are enriched for samples derived from patients with inherited mitochondrial disease. This biases likely skew baseline rates of variation and

bioRxiv preprint doi: https://doi.org/10.1101/798264; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

estimates of allele frequencies. Lastly, these databases did not include heteroplasmic variants, which are essential when studying mitochondrial disorders[14].

Here we present a characterization of the mitochondrial genomes of 196,554 individuals, sequenced in the same clinical laboratory, without bias towards individuals with a mitochondrial disorder. We provide a research resource of all mtDNA variants identified, including both homoplasmic and heteroplasmic calls, together with a map of constrained mitochondrial regions. Lastly, we demonstrate the utility of this resource for interpretation of disease-causing variants by analyzing those reported to be pathogenic for Leber's Hereditary Optic Neuropathy (LHON, OMIM:535000), the genetics of which has been studied for the past 30 years [15].

Results

Mitochondrial DNA variation in 196,554 individuals

We sequenced and analyzed the mitochondrial genomes of 196,554 unrelated individuals (**Methods**). 91.2% of the mitochondrial haplogroups were part of the Eurasian N lineages. The distribution of all haplogroups of the 196,554 individuals is shown in **Table S1**. After filtering out calls of poor quality (GQ < 21 or DP <10), we find a set of 15,035 mtDNA variants observed in at least 1 individual. The full database, HelixMTdb, can be downloaded (link in **Data availability** section). It has information on the identity of variants, the number of times they were observed as homoplasmic or heteroplasmic, and the haplogroup lineage(s) on which the variants were found. The majority of variants (66%, n=9,952) were found to be present in less than 1 in 10,000 individuals in this cohort, with 24% of the variants (n=3,615) observed in only one individual (**Figure 1A**). Only 0.3% of the variants (n=42) were present in more than 10% of the individuals. We identified 13,702 single nucleotide variants (SNVs), 1,000 insertions, and 333 deletions (**Figure 1B**), and observed a higher abundance of transitions (66% of unique variants) than transversions (25% of unique variants) (**Figure 1C**). Lastly, 50% of the variants (n=7,518) were observed both as homoplasmic and heteroplasmic in the population, whereas 28% of the variants (n=4,144) were only observed in homoplasmic calls, and 22% of the variants (n=3,373) were only observed in heteroplasmic calls (**Figure 1D**).

Compared to MITOMAP, we found 5,758 novel variants that were not previously reported (**Figure 1E**). We also compared the allele frequencies of variants in both HelixMTdb and MITOMAP as an additional quality check for the accuracy of our calls. Although the distribution of haplogroups is not exactly the same between HelixMTdb and MITOMAP [16] (**Table S1**), variants private to one database should have low allele frequency in the other database, and

bioRxiv preprint doi: https://doi.org/10.1101/798264; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

common variants in one database should also be common in the other database. The majority of variants unique to HelixMTdb are singletons (**Figure 1A, Figure S1**). However a detailed analysis of the exceptions (for example, **Figure S1** insertions panel) revealed that nearly all of them mapped between positions 300 and 316, a region with homopolymer tracts known to be associated with misalignment errors [4,17]. Other homopolymer tracts are at mt positions: 66-71, 300-316, 513-525, 3106-3107, 12418-12425 and 16182-16194 [4]. Calls in these regions may be inaccurate in both HelixMTdb and MITOMAP, and should be considered with caution. These calls remain in the database, as they may help other researchers visualize the diversity of variants and calls obtained in these regions when using short-read sequencing technology. Even with these variants, the overall concordance of allele frequencies between the two databases was high (rho=0.82, Spearman Rho) (**Figure 1F**), especially for SNVs (**Figure S1**), providing confidence that our variant calling pipeline was accurate, and that HelixMTdb is a reliable resource with which to annotate and interpret clinical mitochondrial DNA variants.

Constrained regions in the mitochondrial genome

Inspired by the work to map the coding constrained regions in the nuclear genome[7], we looked for regions of the mitochondrial genome without any variation, hypothesizing that highly constrained regions may be functionally important. When restricting our variant list to only homoplasmic calls, we observed that 7,936 bases were without any variation in this cohort (**Table 1**). When restricting to only homoplasmic calls plus heteroplasmic calls with a Raw Alternate allele Fraction (RAF) >=0.5, we observed that 7,712 bases were invariable in this cohort (**Table 1**). When considering all homoplasmic and heteroplasmic calls, we observed that 6,092 bases were invariable in this cohort (**Table 1**). The full lists of invariant bases are reported in **Tables S2**, **S3**, and **S4**.

We then focused on the most constrained regions, which we defined as the longest stretches of mtDNA without any variation, when taking into account homoplasmic calls and heteroplasmic calls with a RAF>=0.5. We found 40 intervals of 11 bases or longer (**Table 2)**. We hypothesized that haplogroup markers should not be located within these constrained regions, and could be used as a control to verify the observed constraint. We obtained a list of 1,495 unique haplogroup markers from MITOMAP, using markers found at >=80% in haplogroups (Letter-Number-Letter). Indeed, we found that no haplogroup markers -- even those from haplogroups not represented in our dataset -- were mapped to these highly constrained regions (**Table 2**). In addition, no variants from PhyloTree Build 17 mapped to one of these highly constrained regions. Of note, the majority (26 out of 40) of these highly constrained regions were located in the 2 rRNA genes, and all but three of the remaining (11) were located in tRNAs. This map of highly constrained regions will be helpful to decipher the role of specific domains of rRNA or tRNA genes, and will provide an additional annotation to interpret variants in noncoding regions, in tRNA and rRNA genes.

bioRxiv preprint doi: https://doi.org/10.1101/798264; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Allele frequency of loss-of-function variants

In order to better understand the impact of the variants observed, we first grouped the variants by their genomic feature: protein-coding, ribosomal RNA (rRNA), transfer RNA (tRNA), or non-genic (including the D-loop and intergenic regions) (**Figure 2A**). We hypothesized that variants predicted to be damaging should be less frequent than non-damaging variants in the general population, and should be seen more often heteroplasmic (higher ratio (heteroplasmic calls) / (heteroplasmic + homoplasmic calls)).

Of the 9,752 unique variants in protein-coding genes, only 135 (1.4%) were putative loss-of-function (LoF): 94 frameshift variants, 31 stop-gain variants, and 10 stop-loss variants (**Table S5)**. They were found in all genes (**Figure 2B**). These LoF variants were extremely rare in the population, with a mean allele frequency 0.0046%, which is far below 0.14% the average allele frequency of variants in protein-coding genes. Moreover, there was significant enrichment of heteroplasmy among calls for variants predicted to be LoF: 26% of all the calls for these variants were heteroplasmic, compared to ~1% of the calls for variants predicted to be of medium or low severity (26% vs 1%; p=9.1E-278, fisher exact test) (**Table S5**). In particular, all 94 frameshift and 30/31 stop-gained variants were only observed in the heteroplasmic state in the population at low levels of heteroplasmicity (the average max RAF observed across these variants was 0.16), suggesting that they may not be tolerated when homoplasmic (**Table S6**).

The only nonsense variant observed was p.M1* in the *MT-ND1* gene. A few lines of evidence suggested that p.M1* in *MT-ND1* may not be a true loss-of-function: (i) it was observed in 89 individuals at homoplasmic levels, and none at heteroplasmic levels, (ii) it was only observed on haplogroup T, suggesting that this is a common polymorphism in a specific haplogroup, (iii) it was also observed in 6 individuals in MITOMAP, all of them belonging to the T1a haplogroup, (iv) there is a common missense variant at this position (m.3308T>C), and (v) the next methionine is at amino acid position 3, which may serve as an alternate start codon. These results indicate that all protein-coding genes in the mitochondrial genome were highly intolerant to LoF variants, especially at homoplasmic levels.

There were 1,033 unique variants that mapped to the 22 tRNA genes. We classified the predicted pathogenicity of each tRNA variant using the scoring model from MitoTip [18]. There were 82 (7.9%) observed variants classified as known (P), likely (LP), or possibly (PP) pathogenic. These variants were very rare in the population as the total number of counts in the population was 408 including both homoplasmic and heteroplasmic calls (0.3% of the total counts of tRNA variants) (**Table S5**). Moreover, 51% of the calls for variants predicted to be of high-severity were heteroplasmic calls, which is significantly more compared to ~33% for variants predicted to be of medium severity (p=0.0001, fisher-exact test), 3% for variants predicted to be of low severity (p=1.4E-163, fisher-exact test), and 26% for variants of unknown severity (p=1.3E-06, fisher-exact test) (**Table S5**). These results show that tRNA genes were also intolerant to predicted damaging variants at homoplasmic levels.

bioRxiv preprint doi: https://doi.org/10.1101/798264; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

There were 1,850 unique variants that mapped to the 2 rRNA genes. We classified the variants in rRNA genes into three groups depending on the predicted severity by the heterologous inferential analysis (HIA) technique [19,20] (**Methods**). The application of this method is not yet fully automated, thus we decided to only annotate a small number of variants that were previously annotated with this method. This resulted in 1,807 rRNA variants (97%) having no interpretation regarding their potential impact (**Figure 2C**, **Table S5**). With 34% of these rRNA variants being singletons, and only 1% of the calls being heteroplasmic, these variants displayed a similar pattern to the medium severity variants in protein-coding genes (i.e. missense, inframe indels and start-loss variants).

Assessing classification of LHON variants reported in MITOMAP and ClinVar

In addition to identifying highly constrained regions that can help prioritize variants involved in severe developmental disorders, large databases with population allele frequencies of variants can help discriminate variants for researchers or physicians interested in rare diseases (even those with adult age of onset, or non-lethal phenotype). Inclusion in or exclusion from HelixMTdb was not based on any clinical phenotype. This database can therefore be used to assess whether a mtDNA variant is a good candidate variant for a rare mitochondrial disorder. If the frequency of a variant in HelixMTdb is above the maximum credible population allele frequency, then the variant is unlikely to cause a mitochondrial disorder by itself. For a mitochondrial disorder assumed to be caused by a homoplasmic variant, the maximum credible population allele frequency can be calculated with the equation: *Maximum credible population AF = prevalence x maximum allelic contribution x 1/penetrance* [8]. It is then possible to estimate the upper bound of the allele count expected in a population database given this maximum credible population AF (**Methods**).

We tested the utility of HelixMTdb using this approach on Leber's Hereditary Optic Neuropathy (LHON), which is one of the most studied mitochondrial disorders, with many references available to calculate the prevalence of the disease, genetic homogeneity, and penetrance [21]. With a model aimed at providing an upper estimate of the allele frequency in the population, we estimated the *LHON maximum credible population AF = 1/30,000 x 0.7 x 1/(0.1) = 0.00023*. Assuming that the number of observed variant instances in HelixMTdb follows a Poisson distribution[8], the expected allele count for LHON in HelixMTdb is 46, based on 196,554 mitochondrial genomes, with a maximum tolerated allele count (MTAC) for an LHON-causing variant of 57 (95% confidence). Variants reported to be LHON-causing in the literature should have allele counts below the maximum tolerated allele count calculated.

MITOMAP and ClinVar are two databases that catalog variants reported to be pathogenic for many mitochondrial diseases. As of July 2019, there were a total of 45 variants linked to LHON in either MITOMAP or ClinVar (**Table S7**). We grouped these variants based on the amount of evidence that supported the impact of the variant for LHON: (i) the 3 primary LHON variants, (ii)

bioRxiv preprint doi: https://doi.org/10.1101/798264; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

26 additional variants reported as pathogenic in ClinVar and linked to LHON in MITOMAP, (iii) 9 variants reported as pathogenic in ClinVar but not reported in MITOMAP, and (iv) 7 variants on the MITOMAP LHON page[22], but not reported as pathogenic in ClinVar (**Table S7**). We compared the observed counts for homoplasmic calls for these known LHON variants in HelixMTdb to the MTAC (summarized with their quality metrics in **Table S7)**. On average, the read depth (DP) was 168, the genotype quality (GQ) was above 95, the mapping quality (MQ) was 60 (see **Methods**), and the strand odds ratio (SOR) was 0.83, which altogether indicate that the calls for LHON variants were of high quality. Homoplasmic counts in HelixMTdb were above the MTAC for LHON for 19 (of 45) reportedly pathogenic LHON variants (**Figure 3A, Table S7**). These 19 variants are unlikely to be pathogenic by themselves, assuming the estimates regarding the prevalence of LHON, genetic homogeneity, and penetrance of the most common LHON variant are accurate.

One example of a variant whose frequency in this unselected cohort challenges existing literature is m.14484T>C, one of the three primary mutations for LHON [23–25]. This variant was present in 172 individuals, with 144 homoplasmic calls and 28 heteroplasmic calls, out of 196,554 individuals (AF_{hom}: ~9 in 10,000). Electronic medical records (EMR) were available for 18 of the 144 individuals with a homoplasmic m.14484T>C call in HelixMTdb. None of these 18 individuals had an ICD10 code starting with H47.2 in their electronic health record, which represents all optic atrophies, including hereditary optic atrophy (code H47.22) (**Table 3**). We then tested whether this result would replicate by looking at the allele frequency of m.14484T>C in the UK Biobank (UKB) cohort. The m.14484T>C variant and 8 other known LHON variants were directly genotyped with the UKB genotyping array. The allele frequency was AF_{hom}: ~8 in 10,000 in the entire cohort (n=392 individuals out of 486,036), and it was AF_{hom}: ~9 in 10,000 in a subset of unrelated individuals of European ancestry (n=291 individuals out of 335,840). These results confirmed the relatively high frequency of m.14484T>C variant in the population (**Figure 3B, Table S7**). Looking at the ICD10 codes in all UKB medical records, 97 participants had at least one ICD10 code H47.2 in their health records (optic atrophies have been recorded); however, none of the participants with the m.14484T>C variant had an ICD10 code starting with H47.2 (**Table 3**). Altogether, these analyses strongly suggest that the m.14484T>C variant does not cause LHON by itself.

Discussion

Here we present a genomic resource that can be used to decipher the genetic etiology of rare mitochondrial disorders. HelixMTdb reflects the aggregated and de-identified mitochondrial DNA variants of 196,554 unrelated individuals. This is about 4 times more full mitochondrial genomes than what is currently available in MITOMAP or HmtDB [11,12], two prominent mtDNA variant databases. Unique properties of HelixMTdb are that: (i) it is not enriched for patients with

bioRxiv preprint doi: https://doi.org/10.1101/798264; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

mitochondrial disorders; (ii) it is less prone to batch effects since all samples were processed through the same lab protocol and variant calling pipeline; and (iii) it includes heteroplasmic calls and statistics on the allele fraction for these calls. We showed that ~25% of the variants were only observed at heteroplasmic levels, which would be missed if heteroplasmic calls were not included. This resource is limited in the following ways. First, the average read depth per sample was 180 for the mitochondrial genome, which did not allow confident calls for extremely low-fraction heteroplasmies (<10%), the majority of which were excluded from HelixMTdb. Secondly, since the mitochondrial DNA in this study was extracted from saliva, heteroplasmic levels may not reflect those present in typical mitochondria from phenotype-affected tissues such as muscle. Lastly, the diversity of mitochondrial genomes represented in HelixMTdb is relatively low. For example, a smaller percentage of the individuals came from the L lineages (African) or M lineages (Asian) in HelixMTdb compared to MITOMAP.

Altogether, we identified 15,035 unique variants, including variants overlapping homopolymer tracts. We found that 47% of the bases of the mitochondrial genome did not even have one homoplasmic or heteroplasmic call at a level higher than 50% across the entire cohort. This result shows the high constraint on the mitochondrial genome despite the fact that the mutation rate of mtDNA is higher than the mutation rate of nuclear DNA [26]. Notably, the 2 rRNA genes were under highest constraint with 66% of their bases invariant. This high level of constraint is potentially the result of the absence of redundancy for the mitochondrial rRNA genes, unlike the rRNA genes in the nuclear genome that are present in >100 copies located in 5 rDNA clusters[27]. Most of the known modifications of 16S rRNA and 12S rRNA fall within the most highly constrained regions[28]. We hope that this map will allow molecular biologists hypothesize and design experiments to study translation and regulation of protein expression in the mitochondria [28]. The tRNA genes also showed strong constraints, and the smaller representation of the tRNAs in the longest stretches without any variant may be explained by the smaller size of tRNA genes compared to the 2 rRNA genes.

At the opposite of rRNA and tRNA genes, most of the non-coding bases were variable in the population. The one exception to this rule was a short stretch mapping to the mitochondrial light strand origin of replication. The fact that few intervals under high constraint mapped to protein-coding genes can be explained by the fact that synonymous variants are unlikely to have a strong impact and are therefore well tolerated. On the other hand, our results confirmed that the protein-coding genes in the mitochondrial genome were intolerant to loss-of-function variants.

We have also shown that this resource can be used to better evaluate and prioritize variants suspected to cause rare mitochondrial disease, as long as some assumptions on the prevalence, and genetic architecture of the disease could be made. Through a comparison of the allele frequencies of LHON variants to disease prevalence, we showed that the m.14484T>C variant is likely not pathogenic for LHON by itself. We were able to replicate these results using the UK Biobank cohort, and we showed that the frequency of the variant in unselected cohorts is high, with very low penetrance (0/144 and 0/392 individuals had a LHON diagnosis in their health record). These results are consistent with previously reported pedigree

bioRxiv preprint doi: https://doi.org/10.1101/798264; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

analyses finding that this variant exhibits a low LHON penetrance in a non-haplogroup-J background [24,25,29,30]. Of note, the m.14484T>C variant was present in 13 different haplogroup lineages in HelixMTdb (**Table S7**), and the ratio of (haplogroup J m.14484T>C carriers) / (all haplogroup J) = 4 / 16,030 was the lowest compared to the ratio for the 12 other haplogroups. There may be a branch of haplogroup J where a combination of variants and m.14484T>C is pathogenic[23,31]. Overall, our analysis of variants reported to be pathogenic in ClinVar and MITOMAP for a well-characterized mitochondrial disorder highlights the clinical utility of HelixMTdb. We believe this resource will be instrumental in improving clinical classification of variants, similar to the role that large nuclear DNA variation databases play in clinical interpretation today[32].

bioRxiv preprint doi: https://doi.org/10.1101/798264; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Methods

Individuals

The HelixMTdb database reflects aggregated and de-identified mitochondrial DNA variants observed in individuals sequenced at Helix. The cohort is skewed slightly female at 55%, with a normal distribution of samples aged 18-85+ (mean age=46-50). All individuals sequenced resided in the United States at the time of providing their saliva sample. Importantly, these individuals have not been sequenced based on the presence or absence of any medical phenotype (i.e. there are no inclusion or exclusion criteria in the registration process based on any medical phenotype). Nine percent of Helix users in this study were also participants in the Healthy Nevada Project under the University of Nevada Reno IRB protocol: #7701703417. Electronic medical records were available for most of the Healthy Nevada Project participants, and these records showed no enrichment for classic mitochondrial diseases as shown in **Table S9**.

The replication study for the primary LHON variants was based on the UK Biobank resource [33], under application number 40436.

Sample preparation, Sequencing, and Variant Calling

Library Preparation and Enrichment was performed in the Helix clinical laboratory (CLIA #05D2117342 , CAP #9382893). Samples were sequenced using the Exome+ assay, a proprietary exome that combines a highly performant medical exome, the mitochondrial genome, and a microarray-equivalent SNP backbone into a single sequencing assay (www.helix.com). Read length was 75 bp. Base calling and alignment were run on BaseSpace servers. For mitochondria, we first extracted read pairs in which both reads were mapped and at least one was mapped to the mtDNA. This enabled us to map regions that might otherwise be discarded due to multimapping regions of homology with nuclear sites (NUMTs). Reads were mapped to the rCRS (GenBank: J01415.2) using BWA mem [34], and were deduplicated and realigned using the Sentieon implementation of the GATK algorithms[35] [36]. VCF files were generated using haplotyper with `emit_mode=confident`.

The mean read depth across the mitochondria for an individual was DP=180.

Before including samples and calls into HelixMTdb, we used the following filters: removed samples with mean mtDNA coverage <20; removed calls at positions covered by less than 10 reads; removed calls with genotype quality (GQ) below 20.

Variants in regions with homopolymer tracts (mtDNA positions 66-71, 300-316, 513-525, 3106-3107, 12418-12425, and 16182-16194) were kept in the HelixMTdb and in all analyses.

bioRxiv preprint doi: https://doi.org/10.1101/798264; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Haplogroup Calling

We collapsed heteroplasmic calls into either ALT or REF homoplasmic calls whenever the majority call consisted of at least 75% of the total reads. The remaining sites were left as heteroplasmic, although they are ignored (assumed as reference) by the haplogroup caller. Some indels in a VCF can be left- or right-aligned, meaning that they could be expressed in more than one fashion, changing the coordinates of the positions that are affected. For instance, a change in the number of repeats of a microsatellite can be expressed as an indel at the beginning or at the end of the microsatellite. Both of these can be used to reconstruct the same sequence, but might be used differently by haplogroup callers. Our pipeline originally makes a left-alignment, which is the way the calls are represented in HelixMTdb. We changed it to a right-alignment to be able to use Haplogrep [37]. We removed in advance sites and mutations that were not incorporated in Phylotree v17 [38], because they are not commonly used for haplogroup assignment. We then ran Haplogrep, using rCRS as reference, and kept the first (ranked) 40 hits for further analysis. A number of steps were taken to further reduce the number of haplogroups under consideration: (i) the quality call (for haplogroup) had to be at least 0.94 of that of the maximum, and (ii) at least as high as the value of the third ranked quality (ties might result in more than three haplogroup passing this filter). The most recent common ancestor of these haplogroups was selected as representative for the sample. When the lineage falls in a haplogroup that is similar to that of rCRS, Haplogrep tends to provide inaccurate results. For instance, a VCF file without any position, does not provide a haplogrep call, but corresponds to a sequence that matches the rCRS. Also, lineages similar to the rCRS that share a mutation with a different part of the tree might be assigned to an incorrect haplogroup, but are characterized by a large number of missing mutations for the haplogroup. These were all corrected afterwards.

For comparative representation in HelixMTdb, we combined haplogroups into higher-level haplogroups that matched those shown in MITOMAP (**Table S1**). For HelixMTdb and **Tables S6** and **S7**, we further grouped higher-level haplogroups with less than 10 individuals with other higher-level haplogroups to avoid providing an individual's full mitochondrial DNA sequence. This resulted in the grouping together of 'L4 + L5 + L6' and 'X + S'.

Relatedness analysis

In addition to calling mitochondrial DNA variants, reads from the entire Exome+ were mapped to Human Reference GRCh38 for non-mitochondrial variant calling, using a custom version of the Sentieon align and calling algorithms [39] following GATK best practices. For allele frequency analysis, we further reduced the sample set by removing individuals related at the 2nd-degree or closer.

Briefly, we calculated kinship using the Hail pc_relate method [40] using 11,772 representative common SNPs spread across the genome. The method pc_relate was run with the 10 principal

bioRxiv preprint doi: https://doi.org/10.1101/798264; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

components and a kinship cutoff of 0.0884.

From clusters of family members, we first kept both halves of the father - child relationships. For the other relationships, we randomly selected one representative to retain. 21,074 samples were removed at this stage. We labeled this the unrelated dataset, and proceeded with our analysis based on this cohort of 196,554 individuals.

Analysis of Allele Frequency and Heteroplasmy

All allele frequency and heteroplasmy analysis, as well as all of the analyses listed after this paragraph, were performed in Hail (Hail Team. Hail 0.2.13-81ab564db2b4. https://github.com/hail-is/hail/releases/tag/0.2.13.) on Amazon HPC clusters. Briefly, batches of 500 gVCF files were combined into multi-sample gVCF files using GenomicsDB (https://github.com/Intel-HLS/GenomicsDB/wiki). Multi-sample VCF files were extracted from the resulting gVCF at sites deemed to be informative and then combined into large pVCF files for ingest into Hail using Bcftools (https://samtools.github.io/bcftools/bcftools.html). All variants were left-aligned.

Levels of heteroplasmy play important roles in causing a mitochondrial disease, as well as modulating the strength of phenotypes. To provide as much information as possible regarding the levels of heteroplasmy observed for each heteroplasmic call, we defined *RAF = Raw Alternate allele Fraction = (counts of reads supporting the alternate allele) / (count of all reads at this position).*

Comparison with MITOMAP database

We downloaded MITOMAP GenBank FL ID set (all of the full-length sequences) on June 16, 2019 (http://www.mitomap.org). The MITOMAP database at the time was based on 47,412 full-length mitochondrial sequences.

To be able to compare the MITOMAP database with HelixMTdb, we took the file downloaded from MITOMAP, and grouped all identical variants together and then summed the counts across different haplogroups. We also annotated each variant with information on AF_bin. Seven different bins were defined:

- 0 counts
- Count of 1
- Count 1 < AF < AF =0.0001
- 0.0001 =< AF < 0.001
- 0.001 =< AF < 0.01
- 0.01 =< AF < 0.1
- AF >=0.1

We compared the variants, their counts and their allele frequencies in multiple ways. We looked at all calls, homoplasmic SNVs, homoplasmic insertions, and homoplasmic deletions. We

bioRxiv preprint doi: https://doi.org/10.1101/798264; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

plotted the results using a scatter plot, and calculated the Spearman rho coefficient. We also created a 2D histogram to highlight results for the many variants with extremely small counts (**Figure S1**).

The most notable differences in variant calls were observed in the homopolymer stretch between position m.302 and m.315. We think that it is likely that both HelixMTdb and MITOMAP have inaccurate calls at this locus. In addition, some differences in variant frequencies between the two databases may be due to differences in left- or right- alignment in homopolymer stretches.

Annotation of feature type

Genomic feature locations were annotated using the list from MITOMAP (https://www.mitomap.org/foswiki/bin/view/MITOMAP/GenomeLoci), and further curated into four groups: protein-coding, rRNA, tRNA, and non-coding (all remaining sites including the D-loop).

Moreover, a few positions overlap multiple features (e.g. positions 4329-4331 overlapping *MT-TI* and *MT-TQ*, or positions 5721-5729 overlapping MT-TN and the noncoding L strand origin MT-OLR). In these cases, we made arbitrary decisions to avoid overlapping annotations that may impact some future analyses. The positions and their associated feature type are represented in **Table S8**.

List of constrained intervals in the mitochondrial genome

To calculate invariable positions in HelixMTdb, we defined a position as being variable if at least one SNV, or one deletion was overlapping this position. **Table 1** provides the results taking into account (i) only homoplasmic calls, or (ii) homoplasmic calls and heteroplasmic variants where at least one individual was observed with a RAF >=0.5, or (iii) all homoplasmic and heteroplasmic calls. We used BEDTools [41] to `sort` and `merge` the list of SNVs and positions deleted, and defined the final list of positions that were variable.

We then used `bedtools complement` to obtain the list of constrained intervals in the mitochondrial genome.

PhyloTree variants in highly constrained intervals

To test that the regions identified as highly constrained are invariable in the main structure of the mitochondrial phylogenetic tree, we collected the list of mutations from the official phylotree page. After trimming the characters that do not identify position (e.g. ref base, their character of recurrent, deletion, insertion), we generated a list of all positions, and a BED file spanning those positions. Likewise, we generated a BED file for the intervals indicated in **Table 2**. Using

bioRxiv preprint doi: https://doi.org/10.1101/798264; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

BEDTools we assessed the intersection of these BED files, and the result is that the intersection was empty.

Annotation of impact and predicted severity

All variants were classified using Variant Effect Predictor (VEP) against ENSEMBL e!95. Conservation scores were reported from phastCons 100way_vertebrate, obtained from UCSC for GRCh38 (http://hgdownload.cse.ucsc.edu/goldenpath/hg38/phastCons100way/hg38.100way.phastCons/chrM.phastCons100way.wigFix.gz).

For protein-coding variants, severity was determined using the VEP `most_severe_consequence` annotation, and grouped as follows in **Table S5**:

- High: frameshift, stop_gained, stop_loss
- Medium: inframe indel, missense, start loss
- Low: synonymous, stop_retained

tRNA variants were classified by submitting variants in tabular format to the Mitomaster Web Service API, obtaining their raw MitoTip [18] score, then converting the score to a predicted pathogenicity, using the MitoTip scoring matrix (https://www.mitomap.org/foswiki/bin/view/MITOMAP/MitoTipInfo) as follows: >16.25=Likely Pathogenic (LP); 12.66-16.25=Possibly Pathogenic (PP); 8.44-12.66=Possibly Benign (PB); <8.44=Likely Benign (LB). Mitotip tRNA pathogenicity scores are a result of a combination of conservation, frequency in databases, and predicted secondary structure disruption. In **Table S5**, these tRNA variants were combined as follows:

- High severity: known Pathogenic (P), likely pathogenic (LP), and probably pathogenic (PP)
- Low severity: probably benign (PB), likely benign (LB), and Benign (B)

rRNA variants were annotated using the list of variants and their severity categories determined by Heterologous Inferential Analysis (HIA) published in [19,20]. Briefly, this technique maps rRNA variants onto the crystal structure for Human 12S and 16S subunits to understand likely structural defects and leverages functional assay results from highly conserved homologs in multiple species to assign pathogenicity. Of the 113 variants derived from Genbank sequences collected from these two papers, we were able to annotate the predicted severity for only 43 matching variants. In addition, we found 1,807 novel variants that have not been previously classified in the literature or reported in MITOMAP. We left these as “Unknown” severity.

Maximum tolerated allele count

Our main objective was to filter out variants that could not be disease-causing given a pre-defined genetic architecture. The objective was not to prove the pathogenicity of any given

bioRxiv preprint doi: https://doi.org/10.1101/798264; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

variant. Therefore, we opted for a conservative model that would minimize the number of variants discarded and provide a high estimate of the maximum credible population AF. The method and calculations used here are almost identical to a method previously published to calculate the maximum credible population AF, and maximum tolerated allele count for dominant disorders [8]

Maximum credible population AF = prevalence x maximum allelic contribution x 1/penetrance

For LHON:

- Genetic architecture: disease is caused by a homoplasmic mtDNA variant.
- Prevalence in the population: 1 in 30,000. Reports have shown that prevalence is about 1/31,000 in the North East of England, and 1 in 50,000 in Finland [21,30,42].
- Maximum allelic contribution: 0.7. The three primary LHON mutations (m.3460G>A, m.11778G>A, and m.14484T>C) explain the majority of reported LHON cases. Among these, the m.11778G>A is accounting for approximately 70% of cases among northern European populations [21,43]. Overall, we felt like one variant accounting for 70% of LHON cases in our cohort from all parts of the United States was a very high estimate for maximum allelic contribution.
- Penetrance: 0.1. Penetrance is probably the harder number to estimate for this equation. Of note, the penetrance for LHON is sex-specific [21]. Males have a much higher risk of developing symptoms than females. The ranges of the risk of developing symptoms were 32-57% for males and 8-28% for females [21,42]. To be conservative, we selected a penetrance number on the lower end of these ranges.
- Result: **Maximum credible population AF for LHON = 0.00023**.

To calculate the maximum tolerated allele count (MTAC), we calculated the allele count at the upper bound of the one-tailed 95% confidence interval for the established maximum allele frequency, given the number of alleles in the population database. An approximation using a Poisson distribution has been previously reported [8], and we used the same method in R.
*MTAC = qpois(quantile_limit, an*af)*
where *an* is the number of total alleles in the database, and *af* is the maximum credible population allele frequency.

For LHON in HelixMTdb:

```
MTAC = qpois(0.95, 196554*0.00023) = 57
```

We also looked at LHON variants in the UK Biobank cohort. Genotyping information was available for 265 mtDNA positions, for 488,377 samples.

For LHON in UK Biobank:

```
MTAC = qpois(0.95, 488377*0.00023) = 130
```

bioRxiv preprint doi: https://doi.org/10.1101/798264; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

LHON variants

The list of LHON variants from MITOMAP was copied and pasted in July 2019 from this address: https://www.mitomap.org/foswiki/bin/view/MITOMAP/MutationsLHON. The list of LHON variants from ClinVar was obtained using the following steps:

- Started from clinvar_20190603.vcf.gz (obtained here: ftp://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh38/)
- Selected mitochondrial DNA variants
- Kept variants that included 'Leber's_optic_atrophy' in the CLNDN field.
- Selected variants that were labeled as Pathogenic (of note, there were no Likely Pathogenic variants).

Analysis of LHON phenotype in electronic medical records

Electronic medical records were analyzed by parsing the ICD10 codes. No filters were applied based on the source or date of entry. The code H47.2 was used for all Optic Atrophies, which includes the LHON phenotype H47.22). Other ICD10 codes related to eye diseases or other mitochondrial diseases were used as controls. Details are in **Table S9**.

bioRxiv preprint doi: https://doi.org/10.1101/798264; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Acknowledgements

We thank all Helix users, all participants in the Healthy Nevada Project, as well as all research participants in the UK Biobank project. This research has been conducted using the UK Biobank Resource under Application Number 40436. We acknowledge Dr. Ekaterina Yonova-Doing for initial work on the frequency of the m.14484T>C variant in the UK Biobank. We also thank Dr. Agnel Sfeir for discussions about mitochondrial biology. We thank Dr. Shishi Luo and Dr. Ruomu Jiang for helpful discussions about this work. We acknowledge the entire Helix Bioinformatics team for their contributions to the production Exome+ sequencing pipeline, and the entire Laboratory Operations team for the production of clinical exomes. We thank M. Henderson, T. Curreri and all the ambassadors of the Healthy Nevada Project (HNP). We thank Renown Health and DRI marketing for helping to launch the HNP project.

Declaration of Interests

AB, FM, SW, FJ, MR, JB, EC, MI, WL, JL and NW are employees of Helix.

bioRxiv preprint doi: https://doi.org/10.1101/798264; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Data availability

This database is published under a Creative Commons Attribution-NonCommercial-ShareAlike 4.0 License, and may be used, shared and redistributed appropriately. Please cite this paper when using this database.

HelixMTdb can be downloaded using this link:

https://s3.amazonaws.com/helix-research-public/mito/HelixMTdb_20190926.tsv

bioRxiv preprint doi: https://doi.org/10.1101/798264; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Figures

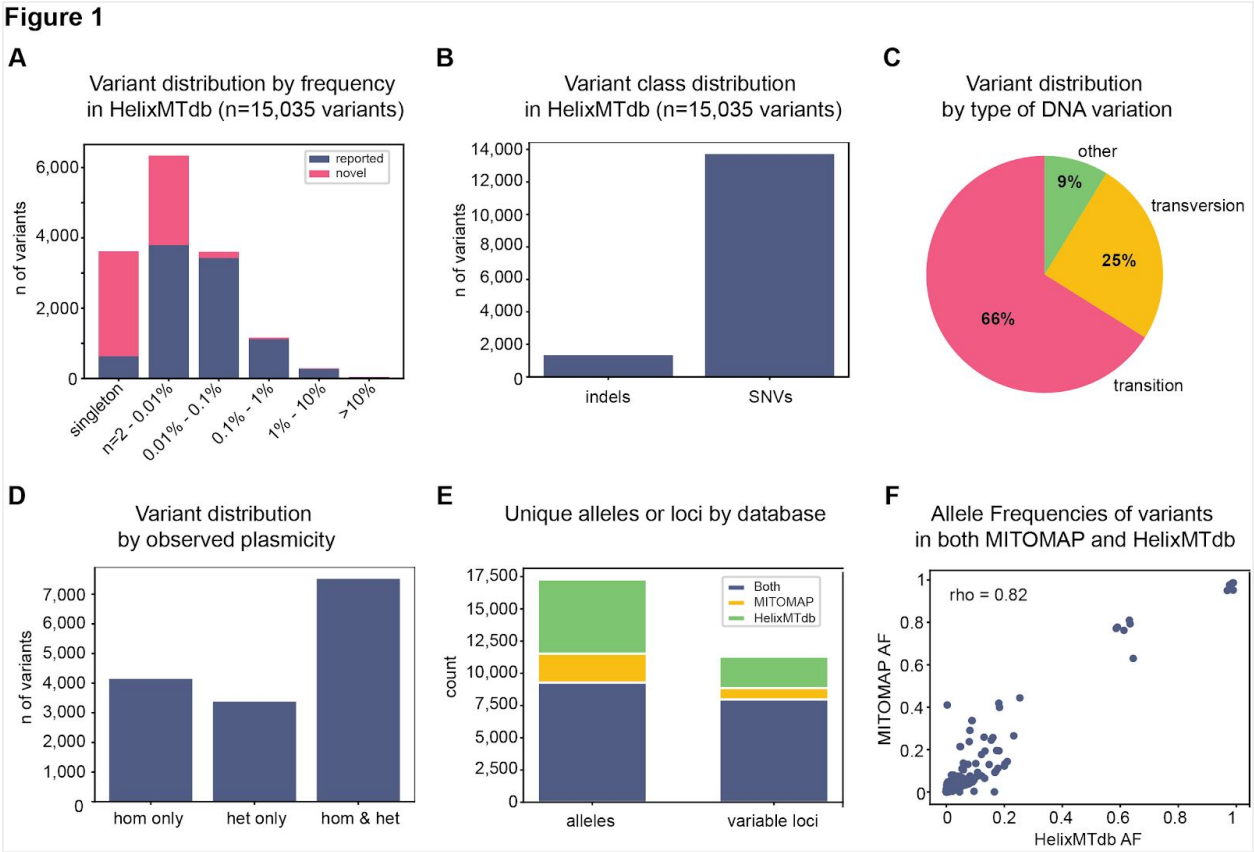

Figure 1. Overview of all mtDNA variants identified in 196,554 individuals

(A) Variants were grouped by their frequency in this cohort. Variants identified in HelixMTdb and also reported in MITOMAP are indicated in blue. Novel variants were defined as not being reported in MITOMAP, and are indicated in pink. **(B)** Counts by variant type are indicated. Indel includes insertions (n=999), deletions (n=333) and one tri-allelic variant with one alternate allele longer than the reference and one alternate allele shorter than the reference; SNVs includes single nucleotide substitutions. **(C)** Proportion of transition, transversion, or other (indels) variants. **(D)** Distribution of variants that are seen in HelixMTdb only at homoplasmic levels (hom only), only at heteroplasmic levels (het only), or both (hom & het). Present in both means that there is at least one occurrence of the variant as homoplasmic and one occurrence as heteroplasmic for the given variant. **(E)** Composition of unique alleles or position/loci by database where it is found: in both MITOMAP and HelixMTdb (blue), only in MITOMAP (yellow), or only in HelixMTdb (green). **(F)** Comparison of allele frequencies (AF) for variants present in both HelixMTdb and MITOMAP.

bioRxiv preprint doi: https://doi.org/10.1101/798264; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

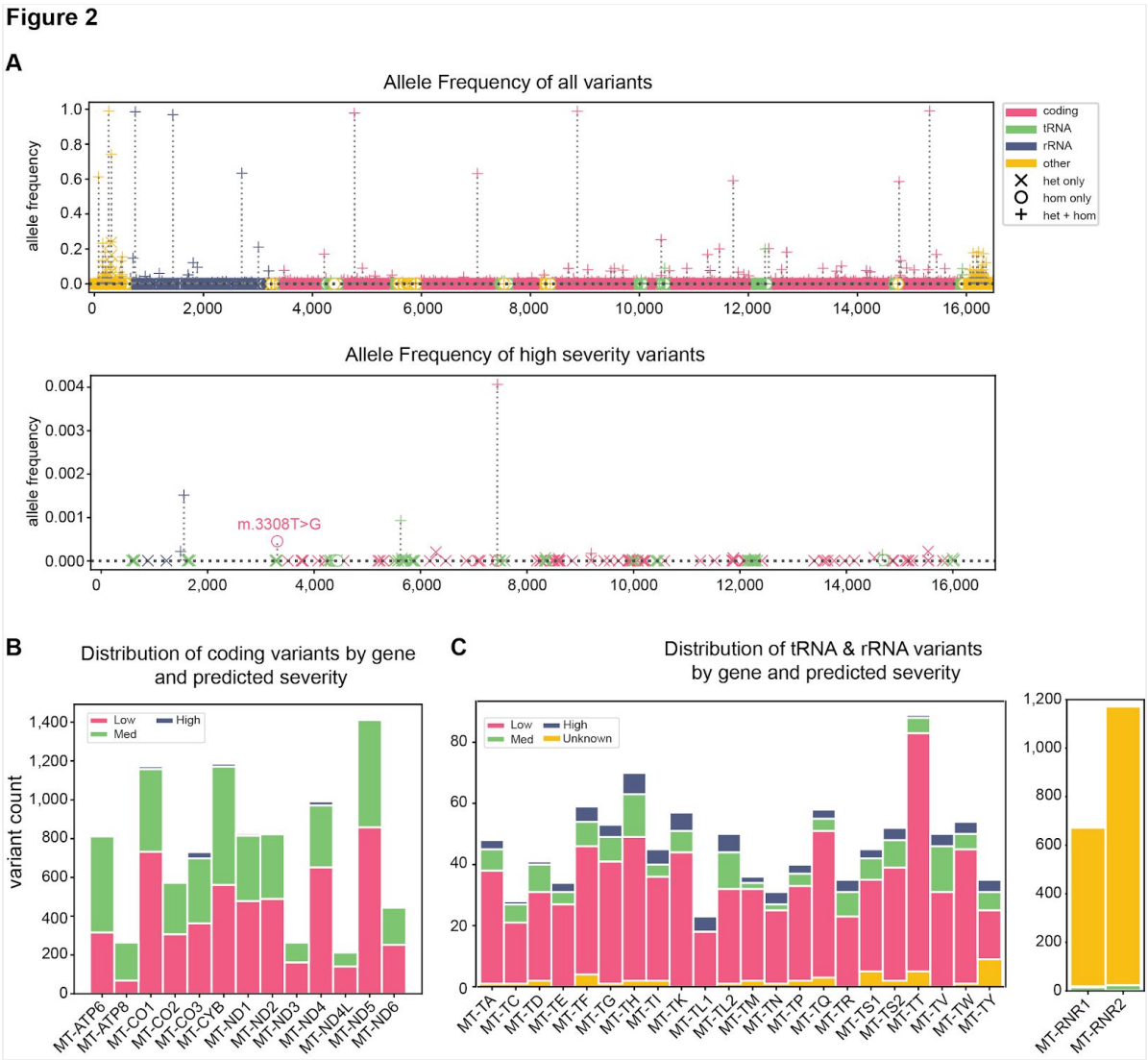

Figure 2. Distribution of variants in protein-coding and RNA genes

(A) Linearized view of mitochondrial genome. pink: protein-coding genes; green: tRNA genes; blue: rRNA genes; yellow: noncoding. Lollipops above genomic features indicate variants observed at heteroplasmic levels only (x), at homoplasmic levels only (o), and at both heteroplasmic and homoplasmic levels (+) of plasmicity. Top panel shows all variants; bottom panel only shows "high" severity variants. **(B-C)** Summary counts of variants per gene, colored based on predicted severity. Pink: low; green: med; blue: high; yellow: unknown. **(B)** Severity annotated using VEP most_severe_consequence and grouped as follows: high (stop gained, frameshift, stop lost), medium (nonsynonymous, inframe indel, coding_sequence_variant, protein_altering_variant), low (synonymous, incomplete_terminal_codon_variant, non_coding_transcript_exon_variant). For clarity, High-severity variants are observed in every protein-coding gene (range: 1 variant in *MT-ND4L* to 32 variants in *MT-CO3*). **(C)** Severity was calculated using MitoTip for tRNA, and manually determined from previous publications for rRNA.

bioRxiv preprint doi: https://doi.org/10.1101/798264; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

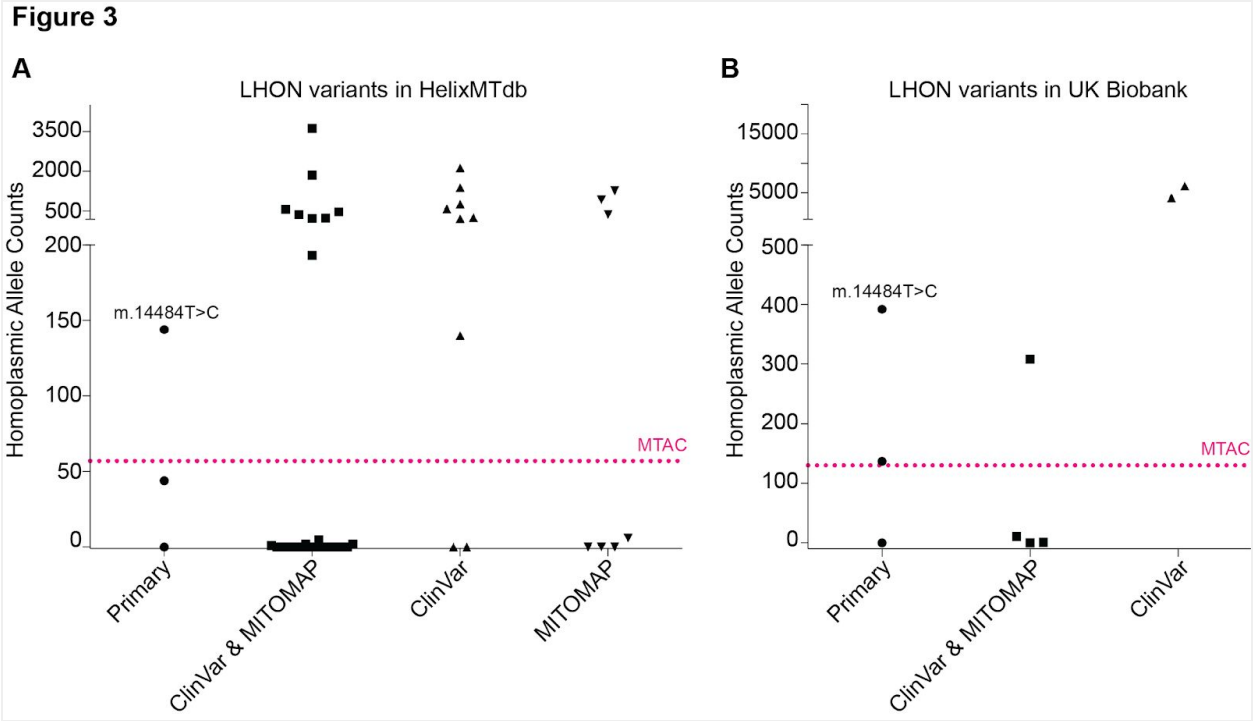

Figure 3. Counts of LHON variants in HelixMTdb and UK Biobank

(A) Allele counts for reported LHON variants in HelixMTdb. Each circle, square or triangle represents a unique mitochondrial DNA variant. The 3 LHON primary mutations are represented by circles. LHON variants reported as pathogenic in ClinVar and present in the LHON MITOMAP page are represented by squares. Triangles represent LHON variants described as pathogenic on ClinVar or in the MITOMAP LHON page, but not by both. The pink dotted line represents the maximum tolerated allele count (MTAC), which is 57 for HelixMTdb. **(B)** Allele counts for reported LHON variants in UK Biobank. MTAC is 130 for the UK Biobank.

bioRxiv preprint doi: https://doi.org/10.1101/798264; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Tables

Table 1: Summary of invariant positions and constrained regions in the mitochondrial genome by genomic features

			homoplasmic variants		homoplasmic + heteroplasmic with max RAF >= 0.5		all homoplasmic and heteroplasmic	
	feature count*	number of bases*	% of invariant bases (# of bp)	number of regions in top 1% of constrained regions**	% of invariant bases (# of bp)	number of regions in top 1% of constrained regions**	% of invariant bases (# of bp)	number of regions in top 1% of constrained regions**
protein-coding genes	13	11,341	45 (5,058)	1	43 (4,929)	1	35 (4,018)	1
tRNA genes	22	1,504	61 (922)	14	59 (888)	11	45 (676)	8
rRNA genes	2	2,513	66 (1,659)	27	64 (1,619)	26	45 (1,140)	18
non-coding region(s)	13	1,211	25 (297)	1	23 (276)	2	21 (258)	4
full mitochondrial genome	50	16,569	48 (7,936)	43	47 (7,712)	40	37 (6,092)	31

* see Table S8

** if an interval overlapped 2 features, we assigned to the one with the biggest overlap. If equal, it was attributed to the first feature.

bioRxiv preprint doi: https://doi.org/10.1101/798264; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Table 2: List of the most constrained regions in the mitochondrial genome (note: this table spreads on 2 pages)

interval	size of interval in bp	genomic feature(s) mapping to the interval
[chrM:3053-3082)	29	MT-RNR2
[chrM:1251-1275)	24	MT-RNR1
[chrM:1127-1148)	21	MT-RNR1
[chrM:2502-2523)	21	MT-RNR2
[chrM:2979-3000)	21	MT-RNR2
[chrM:3119-3140)	21	MT-RNR2
[chrM:1474-1494)	20	MT-RNR1
[chrM:2463-2483)	20	MT-RNR2
[chrM:1064-1081)	17	MT-RNR1
[chrM:5687-5704)	17	MT-TN
[chrM:16019-16036)	17	MT-TP, MT-CR
[chrM:4419-4435)	16	MT-TM
[chrM:2446-2461)	15	MT-RNR2
[chrM:2714-2729)	15	MT-RNR2
[chrM:12312-12327)	15	MT-TL2
[chrM:577-591)	14	MT-TF
[chrM:907-921)	14	MT-RNR1
[chrM:1557-1571)	14	MT-RNR1
[chrM:2012-2026)	14	MT-RNR2
[chrM:3291-3305)	14	MT-TL1
[chrM:5754-5768)	14	MT-OLR, MT-TC
[chrM:14728-14742)	14	MT-TE
[chrM:1355-1368)	13	MT-RNR1
[chrM:1572-1585)	13	MT-RNR1
[chrM:2488-2501)	13	MT-RNR2
[chrM:2673-2686)	13	MT-RNR2
[chrM:2932-2945)	13	MT-RNR2
[chrM:4264-4277)	13	MT-TI
[chrM:966-978)	12	MT-RNR1
[chrM:1082-1094)	12	MT-RNR1
[chrM:3040-3052)	12	MT-RNR2
[chrM:5877-5889)	12	MT-TY
[chrM:12253-12265)	12	MT-TS2
[chrM:691-702)	11	MT-RNR1

bioRxiv preprint doi: https://doi.org/10.1101/798264; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

[chrM:1586-1597)	11	MT-RNR1
[chrM:1915-1926)	11	MT-RNR2
[chrM:2564-2575)	11	MT-RNR2
[chrM:3279-3290)	11	MT-TL1
[chrM:4563-4574)	11	MT-ND2
[chrM:7522-7533)	11	MT-TD

bioRxiv preprint doi: https://doi.org/10.1101/798264; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Table 3: Phenotype of individuals carrying the m.14484T>C variant

	HelixMTdb all individuals	HelixMTdb homoplasmic m.14484T>C	UK Biobank all individuals	UK Biobank homoplasmic m.14484T>C
n samples	196,554	144	486,428	392
n samples with EHR available (% of samples)	18,503 (9%)	18	413,647	318
Mean number of records in EHR when available (range)	54 (1 - 823)	39 (1 - 156)	21 (1 - 5012)	17 (1 - 209)
Median number of records in EHR when available	35	34	8	8
Number of individuals with ICD10 code H47.2 (% of samples)	12 (0.06%)	0 (0%)	97 (0.02%)	0 (0%)

bioRxiv preprint doi: https://doi.org/10.1101/798264; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Supplementary Information

There are 1 supplementary figure and 9 supplementary tables.

Figure S1: Comparison between HelixMTdb and MITOMAP

Table S1: Distribution of mitochondrial haplogroups in HelixMTdb

Table S2: List of all constrained mitochondrial regions inferred from homoplasmic calls
Regions / intervals of 1bp were not included in this analysis.

Table S3: List of all constrained mitochondrial regions inferred from homoplasmic calls and heteroplasmic calls with a RAF >=0.5
Regions / intervals of 1bp were not included in this analysis.

Table S4: List of all constrained mitochondrial regions inferred from all homoplasmic calls and all heteroplasmic calls
Regions / intervals of 1bp were not included in this analysis.

Table S5: Variant attributes by genomic features
Mean allele frequency = ratio of n_non_ref / n_samples. So the % of individuals either with a homoplasmic or heteroplasmic variant vs number of total samples.
% conservation = hl.agg.mean(phastcons100v)

Table S6: Details of all putative loss-of-function variants in HelixMTdb

Table S7: Counts of all LHON variants reported in HelixMTdb and the UK Biobank

Table S8: Location of genomic features in the mitochondrial genome

Table S9: ICD10 codes in Healthy Nevada Project and UK Biobank

bioRxiv preprint doi: https://doi.org/10.1101/798264; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

References

1. Schaefer, A. M. *et al.* Prevalence of mitochondrial DNA disease in adults. *Ann. Neurol.* **63**, 35–39 (2008).
2. Gorman, G. S. *et al.* Prevalence of nuclear and mitochondrial DNA mutations related to adult mitochondrial disease. *Ann. Neurol.* **77**, 753–759 (2015).
3. Anderson, S. *et al.* Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457–465 (1981).
4. Wei, W. *et al.* Germline selection shapes human mitochondrial DNA diversity. *Science* **364**, (2019).
5. Chinnery, P. F. & Samuels, D. C. Relaxed replication of mtDNA: A model with implications for the expression of disease. *Am. J. Hum. Genet.* **64**, 1158–1165 (1999).
6. Chinnery, P. F., Samuels, D. C., Elson, J. & Turnbull, D. M. Accumulation of mitochondrial DNA mutations in ageing, cancer, and mitochondrial disease: is there a common mechanism? *Lancet* **360**, 1323–1325 (2002).
7. Havrilla, J. M., Pedersen, B. S., Layer, R. M. & Quinlan, A. R. A map of constrained coding regions in the human genome. *Nat. Genet.* **51**, 88–95 (2019).
8. Whiffin, N. *et al.* Using high-resolution variant frequencies to empower clinical genome interpretation. *Genet. Med.* **19**, 1151–1158 (2017).
9. University of Michigan and NHLBI. The NHLBI Trans-Omics for Precision Medicine

bioRxiv preprint doi: https://doi.org/10.1101/798264; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

(TOPMed) Whole Genome Sequencing Program. *BRAVO variant browser.* (2018). Available at: https://bravo.sph.umich.edu/freeze5/hg38/.

10. Karczewski, K. J. *et al.* Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. *bioRxiv* 531210 (2019). doi:10.1101/531210
11. Lott, M. T. *et al.* mtDNA Variation and Analysis Using Mitomap and Mitomaster. *Curr. Protoc. Bioinformatics* **44**, 1.23.1–26 (2013).
12. Preste, R., Vitale, O., Clima, R., Gasparre, G. & Attimonelli, M. HmtVar: a new resource for human mitochondrial variations and pathogenicity data. *Nucleic Acids Res.* **47**, D1202–D1210 (2019).
13. Connor, T. M. *et al.* Mutations in mitochondrial DNA causing tubulointerstitial kidney disease. *PLoS Genet.* **13**, e1006620 (2017).
14. Wallace, D. C. Mitochondrial genetic medicine. *Nat. Genet.* **50**, 1642–1649 (2018).
15. Wallace, D. C. *et al.* Mitochondrial DNA mutation associated with Leber's hereditary optic neuropathy. *Science* **242**, 1427–1430 (1988).
16. GBFreqInfo < MITOMAP < Foswiki. Available at: https://www.mitomap.org/foswiki/bin/view/MITOMAP/GBFreqInfo. (Accessed: 2nd October 2019)
17. Andrews, R. M. *et al.* Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat. Genet.* **23**, 147 (1999).
18. Sonney, S. *et al.* Predicting the pathogenicity of novel variants in mitochondrial tRNA with MitoTIP. *PLoS Comput. Biol.* **13**, e1005867 (2017).
19. Smith, P. M. *et al.* The role of the mitochondrial ribosome in human disease: searching for mutations in 12S mitochondrial rRNA with high disruptive potential. *Hum. Mol. Genet.* **23**,

bioRxiv preprint doi: https://doi.org/10.1101/798264; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

949–967 (2014).

20. Elson, J. L. *et al.* The presence of highly disruptive 16S rRNA mutations in clinical samples indicates a wider role for mutations of the mitochondrial ribosome in human disease. *Mitochondrion* **25**, 17–27 (2015).
21. Yu-Wai-Man, P. & Chinnery, P. F. Leber Hereditary Optic Neuropathy. in *GeneReviews®* (eds. Adam, M. P. et al.) (University of Washington, Seattle, 2000).
22. MutationsLHON < MITOMAP < Foswiki. Available at: https://www.mitomap.org/foswiki/bin/view/MITOMAP/MutationsLHON. (Accessed: 7th October 2019)
23. Brown, M. D., Torroni, A., Reckord, C. L. & Wallace, D. C. Phylogenetic analysis of Leber's hereditary optic neuropathy mitochondrial DNA's indicates multiple independent occurrences of the common mutations. *Hum. Mutat.* **6**, 311–325 (1995).
24. Brown, M. D., Sun, F. & Wallace, D. C. Clustering of Caucasian Leber hereditary optic neuropathy patients containing the 11778 or 14484 mutations on an mtDNA lineage. *Am. J. Hum. Genet.* **60**, 381–387 (1997).
25. Torroni, A. *et al.* Haplotype and phylogenetic analyses suggest that one European-specific mtDNA background plays a role in the expression of Leber hereditary optic neuropathy by increasing the penetrance of the primary mutations 11778 and 14484. *Am. J. Hum. Genet.* **60**, 1107–1121 (1997).
26. Sigurðardóttir, S., Helgason, A., Gulcher, J. R., Stefansson, K. & Donnelly, P. The Mutation Rate in the Human mtDNA Control Region. *Am. J. Hum. Genet.* **66**, 1599–1609 (2000).
27. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
28. Hällberg, B. M. & Larsson, N.-G. Making proteins in the powerhouse. *Cell Metab.* **20**,

bioRxiv preprint doi: https://doi.org/10.1101/798264; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

226–240 (2014).

29. Howell, N., Herrnstadt, C., Shults, C. & Mackey, D. A. Low penetrance of the 14484 LHON mutation when it arises in a non-haplogroup J mtDNA background. *Am. J. Med. Genet. A* **119A**, 147–151 (2003).
30. Puomila, A. *et al.* Epidemiology and penetrance of Leber hereditary optic neuropathy in Finland. *Eur. J. Hum. Genet.* **15**, 1079–1089 (2007).
31. Carelli, V. *et al.* Haplogroup effects and recombination of mitochondrial DNA: novel clues from the analysis of Leber hereditary optic neuropathy pedigrees. *Am. J. Hum. Genet.* **78**, 564–574 (2006).
32. Richards, S. *et al.* Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet. Med.* **17**, 405–424 (2015).
33. Sudlow, C. *et al.* UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med.* **12**, e1001779 (2015).
34. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv [q-bio.GN]* (2013).
35. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
36. Freed, D., Aldana, R., Weber, J. A. & Edwards, J. S. The Sentieon Genomics Tools - A fast and accurate solution to variant calling from next-generation sequence data. *bioRxiv* 115717 (2017). doi:10.1101/115717
37. Weissensteiner, H. *et al.* HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res.* **44**, W58–63 (2016).
38. van Oven, M. & Kayser, M. Updated comprehensive phylogenetic tree of global human

bioRxiv preprint doi: https://doi.org/10.1101/798264; this version posted October 8, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

mitochondrial DNA variation. *Hum. Mutat.* **30**, E386–94 (2009).

39. Kendig, K. *et al.* Computational performance and accuracy of Sentieon DNASeq variant calling workflow. *bioRxiv* 396325 (2018). doi:10.1101/396325
40. Genetics — Hail. Available at: https://hail.is/docs/0.2/methods/genetics.html. (Accessed: 25th September 2019)
41. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
42. Yu-Wai-Man, P. *et al.* The epidemiology of Leber hereditary optic neuropathy in the North East of England. *Am. J. Hum. Genet.* **72**, 333–339 (2003).
43. Mackey, D. A. *et al.* Primary pathogenic mtDNA mutations in multigeneration pedigrees with Leber hereditary optic neuropathy. *Am. J. Hum. Genet.* **59**, 481–485 (1996).