scientific reports

OPEN

Modeling of mitochondrial genetic polymorphisms reveals induction of heteroplasmy by pleiotropic disease locus 10398A>G

Molly Smullen[1,2,3,4], Meagan N. Olson[1,2,3,4], Liam F. Murray[1,2,3,4], Madhusoodhanan Suresh[1,2,3,4], Guang Yan[1,2,3,4], Pepper Dawes[1,2,3,4], Nathaniel J. Barton[1,2,3,4], Jivanna N. Mason[1,2,3,4], Yucheng Zhang[1,2,3,4], Aria A. Fernandez-Fontaine[1,2,3,4], George M. Church[5,6], Diego Mastroeni[7], Qi Wang[7], Elaine T. Lim[1,2,3,4], Yingleong Chan[1,2,3,8]✉ & Benjamin Readhead[7,8]✉

Mitochondrial (MT) dysfunction has been associated with several neurodegenerative diseases including Alzheimer's disease (AD). While MT-copy number differences have been implicated in AD, the effect of MT heteroplasmy on AD has not been well characterized. Here, we analyzed over 1800 whole genome sequencing data from four AD cohorts in seven different tissue types to determine the extent of MT heteroplasmy present. While MT heteroplasmy was present throughout the entire MT genome for blood samples, we detected MT heteroplasmy only within the MT control region for brain samples. We observed that an MT variant 10398A>G (rs2853826) was significantly associated with overall MT heteroplasmy in brain tissue while also being linked with the largest number of distinct disease phenotypes of all annotated MT variants in *MitoMap*. Using gene-expression data from our brain samples, our modeling discovered several gene networks involved in mitochondrial respiratory chain and Complex I function associated with 10398A>G. The variant was also found to be an expression quantitative trait loci (eQTL) for the gene MT-ND3. We further characterized the effect of 10398A>G by phenotyping a population of lymphoblastoid cell-lines (LCLs) with and without the variant allele. Examination of RNA sequence data from these LCLs reveal that 10398A>G was an eQTL for MT-ND4. We also observed in LCLs that 10398A>G was significantly associated with overall MT heteroplasmy within the MT control region, confirming the initial findings observed in post-mortem brain tissue. These results provide novel evidence linking MT SNPs with MT heteroplasmy and open novel avenues for the investigation of pathomechanisms that are driven by this pleiotropic disease associated loci.

Mutations in the MT genome have been reported to cause a wide variety of human diseases[1]. In particular, the oxidative phosphorylation (OXPHOS) function of the mitochondria has been shown to be disrupted by these mutations[2], and defects in OXPHOS are linked with an increased risk of cancer, including prostate and breast cancer[3,4]. Besides mutations, dysregulation of mitochondria leading to metabolic defects have also been associated with Alzheimer's disease (AD)[5,6]. By analyzing large metabolic data sets, researchers have shown that at later stages of AD, there is a change of energy utilization from fatty acids to amino acids and glucose, indicating an AD-associated switch in energy substrate utilization[7]. Great effort has been made to test the association of AD pathogenesis on a number of phenotypes related to mitochondrial dysfunction, such as overexpression of reactive oxygen species (ROS), calcium imbalance, and other defects of mitochondrial function and dynamics[8]. There is

[1]Department of Neurology, University of Massachusetts Chan Medical School, Worcester, MA 01605, USA. [2]Program in Bioinformatics and Integrative Biology, University of Massachusetts Chan Medical School, Worcester, MA 01605, USA. [3]NeuroNexus Institute, University of Massachusetts Chan Medical School, Worcester, MA 01605, USA. [4]Department of Molecular, Cell and Cancer Biology, University of Massachusetts Chan Medical School, Worcester, MA 01605, USA. [5]Department of Genetics, Blavatnik Institute, Harvard Medical School, Boston, MA 02115, USA. [6]Wyss Institute for Biologically Inspired Engineering, Harvard University, Boston, MA 02115, USA. [7]ASU-Banner Neurodegenerative Disease Research Center, Arizona State University, Tempe, AZ 85281, USA. [8]These authors contributed equally: Yingleong Chan and Benjamin Readhead. ✉email: Rigel.Chan@umassmed.edu; Ben.Readhead@asu.edu

also evidence of significantly lower MT-DNA copy number in the temporal cortex[9] and dorsolateral prefrontal cortex[10] of AD patient brains compared to control subjects. Lastly, APOE ε4 is the common DNA variant that confers the highest increase in risk and age of onset of sporadic AD[11] and was associated with impaired MT structure and function from proteins measured within postmortem human brain tissues of AD patients[12]. As such, studying the role of MT function in AD pathogenesis is currently an active area of research[13,14].

Here, we primarily focus on MT heteroplasmy as a potential mechanism affecting MT function and thus impacting on risk of disease. MT heteroplasmy refers to the presence of MT mutations that occur in only a fraction of mitochondrial DNA within a given sample[15]. We analyzed data from 1801 whole-genome sequenced (WGS) post-mortem tissue samples generated by the Accelerating Medicines Partnership in Alzheimer's Disease (AMP-AD) for the presence of MT heteroplasmy. After discovering a significant association between the mitochondrial SNP 10398A>G (rs2853826) and MT heteroplasmy, we characterized the gene networks and gene expression changes associated with this single-nucleotide polymorphism (SNP). 10398A>G corresponds to the Thr114Ala missense mutation of the MT-ND3 gene coded by the MT, and is a common variant found within most populations with a reported allele frequency of 41.8% from the gnomAD database v3.1.1[16]. The 10398A>G allele has been collectively associated with an expansive set of phenotypically diverse diseases including AD[17], Parkinson's disease[18–24], breast cancer[25–34], gastric cancer, type 2 diabetes and several other human diseases and phenotypes[35–39].

To explore the effect of 10398A>G on transcriptomic and MT variables, we performed experiments using lymphoblastoid cell lines (LCLs) from the Harvard Personal Genome Project (PGP)[40,41]. The PGP consists of many participants that have high coverage WGS data as well as self-reported phenotypic data that are publicly available for research[42,43]. We selected 60 participant LCLs that vary by the 10398 genotype (major allele A: n = 30, minor allele G: n = 30) and performed bulk RNA sequencing, as well as quantifying MT copy number and heteroplasmy, and report our findings herein.

Results

Significant heteroplasmy detected within the mitochondria control region from brain samples.

We analyzed WGS data generated from 1801 post-mortem tissue samples collected as part of the Religious Order Study (ROS), Memory and Aging Project (MAP), Mount Sinai Brain Bank (MSBB) and Mayo Temporal Cortex (MAYO) studies. Across these four studies, WGS data was available from a total of seven different tissues (Table 1). We analyzed data from these samples to perform MT variant calling and to estimate MT heteroplasmy at each MT genomic site (MT heteroplasmy was classified as present if at least 5% of reads that were reliably mapped to either the major or minor allele were discordant with the remaining reads).

While MT heteroplasmy was robustly detected in many samples, there was much higher overall heteroplasmy within the blood compared to the brain samples (Fig. 1A). When the analysis was restricted to only brain samples, we detected significantly more heteroplasmy in the cortically derived samples than the cerebellum derived samples (Fig. 1B).

We observed that MT heteroplasmy detected within the blood derived samples demonstrated a distribution across the entire MT genome, whereas in the brain derived samples, heteroplasmy was entirely restricted to a set of several hundred bases corresponding to the MT control region (Fig. 2), a highly polymorphic[44], non-coding region of the MT genome which contains the origins of both transcription and replication[45].

Mitochondrial heteroplasmy in the dorsolateral prefrontal cortex is not associated with Alzheimer's disease.

We used a logistic regression approach to examine whether there are any systematic differences in mitochondrial heteroplasmy between samples from subjects with AD compared with aged con-

Cohort	WGS tissue	#Samples
ROS	Blood	227
	Dorsolateral prefrontal cortex	211
	Posterior cingulate cortex	38
	Cerebellum	91
MAP	Blood	147
	Dorsolateral prefrontal cortex	240
	Posterior cingulate cortex	29
	Cerebellum	155
MSBB	BM-10 (anterior prefrontal cortex)	80
	BM-22 (superior temporal gyrus)	246
MAYO	Temporal cortex	337
Total		1801

Table 1. The full set of cohorts and tissue groups from which samples were used for analyzing whole-genome sequencing data for mitochondrial heteroplasmy. Samples were sequenced from the ROS (Religious Order Study), MAP (Memory and Aging Project), MSBB (Mount Sinai Brain Bank) and MAYO (Mayo Clinic) samples.

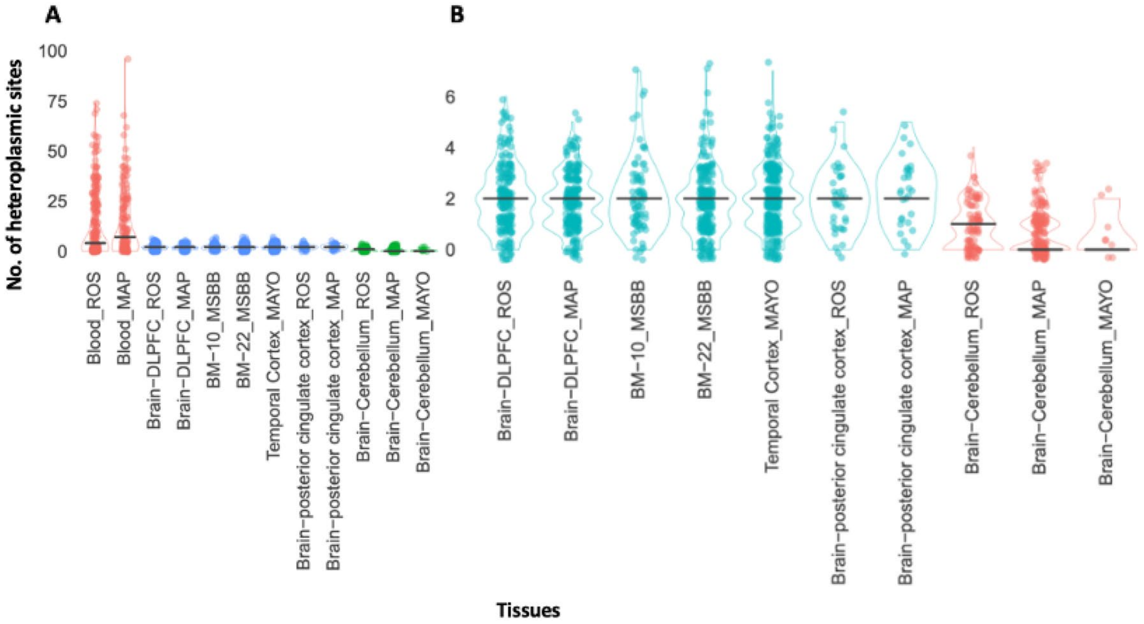

Figure 1. The number of mitochondrial heteroplasmic sites detected within each sample across the 12 tissues and cohort combinations. (**A**) Comparing the number of mitochondrial heteroplasmic sites detected across all blood and brain sample groups. (**B**) Comparing the number of mitochondrial heteroplasmic sites detected across different brain region sample groups.

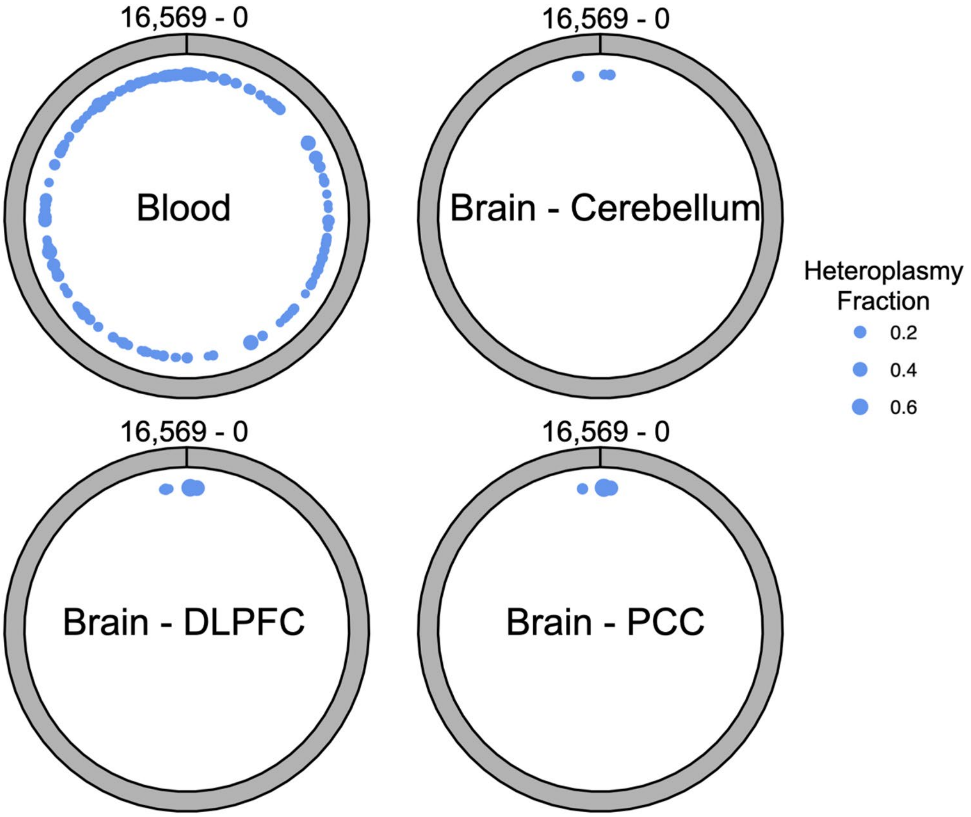

Figure 2. The distribution of detected heteroplasmic sites across the mitochondrial genome, within sample groups from the main tissue types represented within this study. Dot size area is proportional to the fraction of samples in which heteroplasmy was detected at that site.

trols, while adjusting for Study, Sex, Age at death, Years of education, Post-Mortem Interval, and Ancestry. We restricted this analysis to the brain region with the largest subset of WGS samples, the dorsolateral prefrontal cortex (DLPFC) from the ROS/MAP studies (n = 451). Analyses were restricted to nine MT genomic sites with detectable heteroplasmy in at least five samples across this cohort. We did not identify any individual heteroplasmy sites with a significant (FDR < 0.05) association to AD status (Table S1). A complementary analysis performed on 'total mitochondrial heteroplasmy' estimated by summarizing each sample according to the number of mitochondrial sites at which heteroplasmy was detected also did not reveal an association with AD (*P* = 0.67).

10398A>G is significantly associated with mitochondrial heteroplasmy. Based on the analysis of mitochondrial heteroplasmy within brain samples, we performed a cis-heteroplasmy-QTL across the mitochondrial to determine if any SNPs were associated with heteroplasmy in the DLPFC samples from the ROS/MAP studies. While a number of SNPs were found to be significantly associated with heteroplasmy, the 10398A>G variant had the most abundant signal, being significantly associated with heteroplasmy at five separate loci (FDR < 0.05) (Table 2). We further annotated each variant for its associated disease phenotypes using MITOMAP[46] and observed that 10398A>G was associated with the largest number of distinct phenotypes of any reported MT variant (n = 10). 10398A>G has been linked with AD[17], Parkinson's disease[18–24], breast cancer[25–34], type 2 diabetes and many other diseases and conditions[35–39].

The 10398A>G variant is significantly associated with MT-ND3 expression in human prefrontal cortex. Given the association of 10398A>G with mitochondrial heteroplasmy and diverse disease phenotypes, we analyzed its association with gene expression changes. By integrating RNA sequence data available on ROS/MAP cohort DLPFC samples, we performed a cis-eQTL analysis and found that 10398A>G was associated with MT-ND3 (Mitochondrially Encoded NADH Dehydrogenase 3) expression, where the minor G allele is associated with decreased expression (Fig. 3A). MT-ND3 is one of the subunits of the mitochondrial respiratory chain complex I that enables NADH dehydrogenase (ubiquinone) activity. We thus constructed a targeted gene regulatory network aimed at identifying genes that are downstream of MT-ND3, conditioned on the relationship with 10398A>G, using a causal inference testing approach[47] and identified a number of significantly associated downstream genes (Fig. 3B). We then performed a gene set enrichment analysis on the MT-ND3 subnetwork using *Enrichr*[48–50] focused on enrichments in Biological pathways and Gene Ontology components (see "Materials and methods" for gene set libraries) and observed several enrichments for NADH dehydrogenase activity and Mitochondrial respiratory chain Complex I components (Table 3).

The 10398A>G variant is significantly associated with MT-ND4 expression in LCLs. After performing RNA extraction and whole transcriptome RNA sequencing on all 60 LCL samples, we tested our data

MT hetQTL locus	# Heteroplasmy sites	MT heteroplasmy sites	Estimates (range)	FDR (range)	# MITOMAP diseases	MITOMAP diseases
10398_A/G	5	16129_G/A, 189_A/G, 185_G/A, 65_TG/T, 16093_T/C	1.27–2.33	4.69e−02 to 1.35e−07	10	PD protective factor/longevity/altered cell pH/metabolic syndrome/breast cancer risk/Leigh Syndrome risk/ADHD/cognitive decline/SCA2 age of onset/Fuchs endothelial corneal dystrophy
11467_A/G	4	65_TG/T, 189_A/G, 16093_T/C, 16188_CT/C	1.15–3.22	4.72e−02 to 1.29e−06	2	Altered brain pH/sCJD patients
12308_A/G	4	65_TG/T, 189_A/G, 16093_T/C, 16188_CT/C	1.15–3.22	4.72e−02 to 1.29e−06	5	CPEO/Stroke/CM/Breast & Renal & Prostate Cancer Risk/Altered brain pH /sCJD
12372_G/A	4	65_TG/T, 189_A/G, 16093_T/C, 16188_CT/C	1.15–3.22	4.72e−02 to 1.29e−06	2	Altered brain pH/sCJD patients
14798_T/C	4	185_G/A, 189_A/G, 65_TG/T, 16093_T/C	1.22–2.79	1.45e−03 to 1.80e−06		
11251_A/G	3	185_G/A, 189_A/G, 65_TG/T	1.31–2.61	1.07e−04 to 7E−06	1	Reduced risk of PD
11719_G/A	3	189_A/G, 65_TG/T, 16129_G/A	1.05–2.07	4.26e−02 to 1.08e−14		
12612_A/G	3	185_G/A, 189_A/G, 65_TG/T	1.18–3.58	2.62e−02 to 2.89e−08		
13617_T/C	3	65_TG/T, 16188_CT/C, 185_G/A	1.29–2.80	3.50e−02 to 8.59e−03		
13708_G/A	3	185_G/A, 189_A/G, 65_TG/T	0.93–3.30	4.94e−02 to 1.58e−07	3	LHON/Increased MS risk/higher freq in PD-ADS

Table 2. Inherited heteroplasmy quantitative trait loci (hetQTL) alleles rank ordered according to number of significantly associated mitochondrial heteroplasmy sites (FDR < 0.05) within the 411 dorsolateral prefrontal cortex samples from the ROS and MAP cohorts. Disease and phenotype annotations for listed hetQTL from MITOMAP.

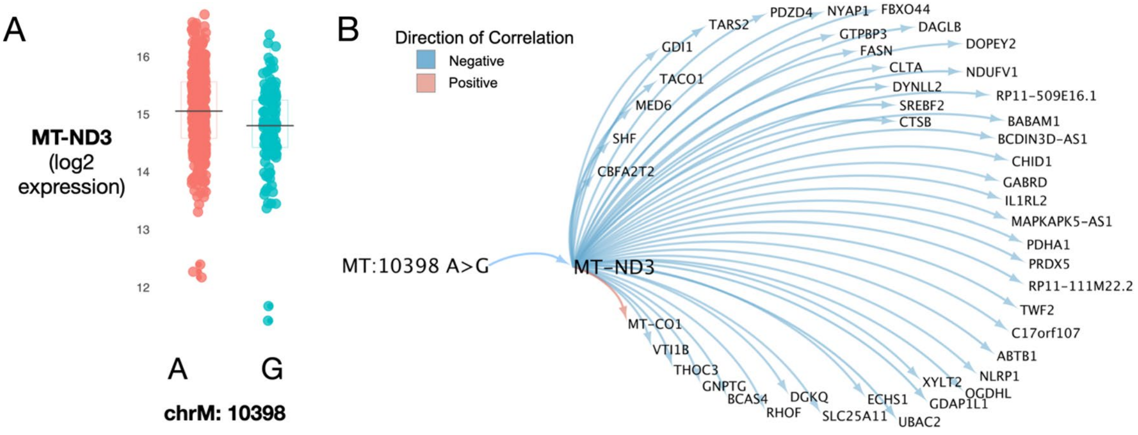

Figure 3. (**A**) Gene expression levels of MT-ND3 between dorsolateral prefrontal cortex samples with A or G allele for 10398A>G. (**B**) Gene regulatory network constructed from neighbors downstream of MT-ND3, conditioned on MT:10398 A>G genotype using a Causal Inference Testing approach. Top 45 strongest connections shown.

Library	Gene set	FDR	Odds ratio
GO_Molecular_Function_2021	Oxidoreduction-driven active transmembrane transporter activity (GO:0015453)	2.74e−05	6.08
GO_Molecular_Function_2021	NADH dehydrogenase (quinone) activity (GO:0050136)	2.74e−05	8.65
GO_Molecular_Function_2021	NADH dehydrogenase (ubiquinone) activity (GO:0008137)	2.74e−05	8.65
GO_Cellular_Component_2021	Mitochondrial respiratory chain complex I (GO:0005747)	3.16e−05	7.32
GO_Cellular_Component_2021	Respiratory chain complex I (GO:0045271)	3.16e−05	7.32
Reactome_2022	Respiratory Electron Transport R-HSA-611105	5.31e−05	4.76
GO_Biological_Process_2021	Mitochondrial electron transport, NADH to ubiquinone (GO:0006120)	2.15e−04	8.20
Reactome_2022	Complex I Biogenesis R-HSA-6799198	2.87e−04	6.10
GO_Biological_Process_2021	Aerobic electron transport chain (GO:0019646)	4.23e−04	5.08
GO_Biological_Process_2021	Mitochondrial ATP synthesis coupled electron transport (GO:0042775)	4.23e−04	4.98

Table 3. Selection of top Gene Ontology and Biological Pathways that are over-represented among the gene regulatory network constructed from neighbors downstream of MT-ND3, conditioned on MT:10398 A>G genotype. Enrichments were calculated using Enrichr. Gene sets with a false discovery rate (FDR) < 0.05 were classified as significantly over-represented among the MT-ND3 gene regulatory network.

against previously discovered eQTLs in LCLs to determine if they were replicated within our samples. To do this, we integrated cis-eQTL data obtained from the GTEx portal for LCLs mapped in European-American subjects[51]. After filtering and pruning, we retained 764 SNPs that were also annotated as cis-eQTLs in GTEx and with available association statistics within our LCL data. We found an enrichment of GTEx cis-eQTL from our data in the same direction of effect (Figure S1, Table S2). Of the 764 SNPs, 98 of them were nominally significant from our data using a threshold of $P < 0.01$, whereas one would expect only 7.64 to be significant under the null hypothesis ($P = 7.4 \times 10^{-74}$). This result suggests that our LCL gene expression data is broadly concordant with the LCL data obtained from GTEx.

We then analyzed whether 10398A>G is associated with expression levels of any MT encoded genes. Of the 37 MT genes we examined for analysis, only 28 demonstrated robust expression in our data. We observed that 10398A>G is significantly associated with lower expression of MT-ND4 ($P = 5.96 \times 10^{-9}$, $FDR = 1.67 \times 10^{-7}$) (Fig. 4A, Table S3). MT-ND4 (NADH Dehydrogenase 4), which is a different subunit from MT-ND3, is also part of the mitochondrial respiratory chain complex I. The G allele is also significantly associated with lower expression of MT-ATP8, MT-ND2, MT-ND4L, MT-ATP6 and MT-CYB ($FDR < 0.005$) (Table S3). Notably, there was no significant difference in expression detected for the MT-ND3 gene (as we had observed in the ROS/MAP DLPFC samples) despite the G allele causing a Thr114Ala missense mutation ($P = 0.853$, $FDR = 0.884$) (Fig. 4B, Table S3).

Testing for association of 10398A>G with mitochondrial DNA copy number in the LCLs. Given that the effect of 10398A>G on MT gene expression within the LCL samples was to uniformly decrease MT gene expression (Table S3), we hypothesized that this may reflect an effect on reducing overall MT copy number, as has been described for 10398A>G in a previous disease context[52]. We thus tested whether 10398A>G had any effect on MT copy number in the LCLs. We obtained genomic and mitochondrial DNA for each of the 60 PGP LCLs and performed multiplex PCR of a mitochondrial fragment, a fragment on the X-chromosome, and a

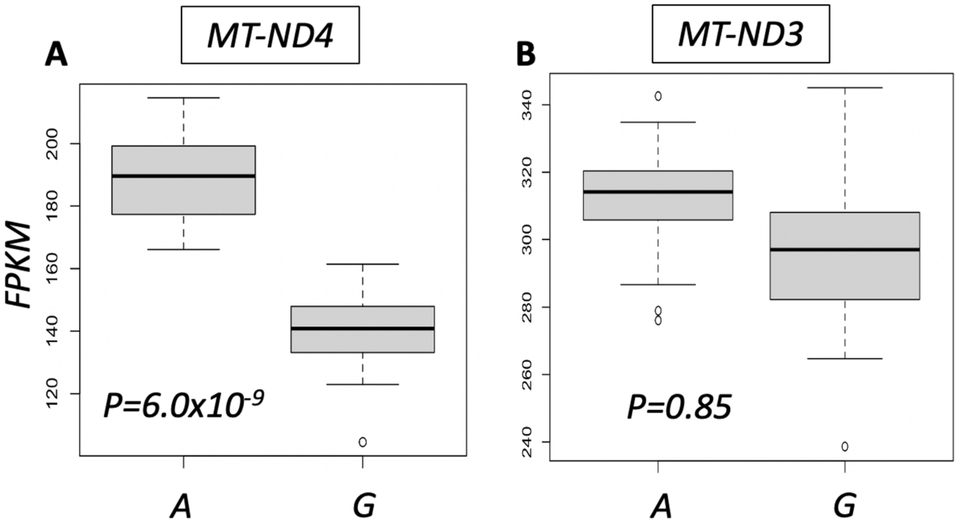

Figure 4. Comparing gene expression differences between LCL samples with the A allele versus the G allele of 10398A>G and representing them as standard boxplots. (**A**) Gene expression difference of MT-ND4. (**B**) Gene expression difference of MT-ND3. The P values for (**A**,**B**) are depicted within each of the individual plots. P values for differential expression were calculated using edgeR.

fragment on chromosome 22 as a genomic control. The X-chromosome fragment serves as a positive control given that female samples should have significantly more X-chromosome DNA than male samples. Consistent with our expectation, the multiplex PCR approach demonstrated a significantly increased X-chromosome DNA copy number in our female samples compared to males, with clear separation between female and male samples ($P = 1.28 \times 10^{-7}$) (Fig. 5A). We performed the equivalent analysis to determine if 10398A>G is associated with

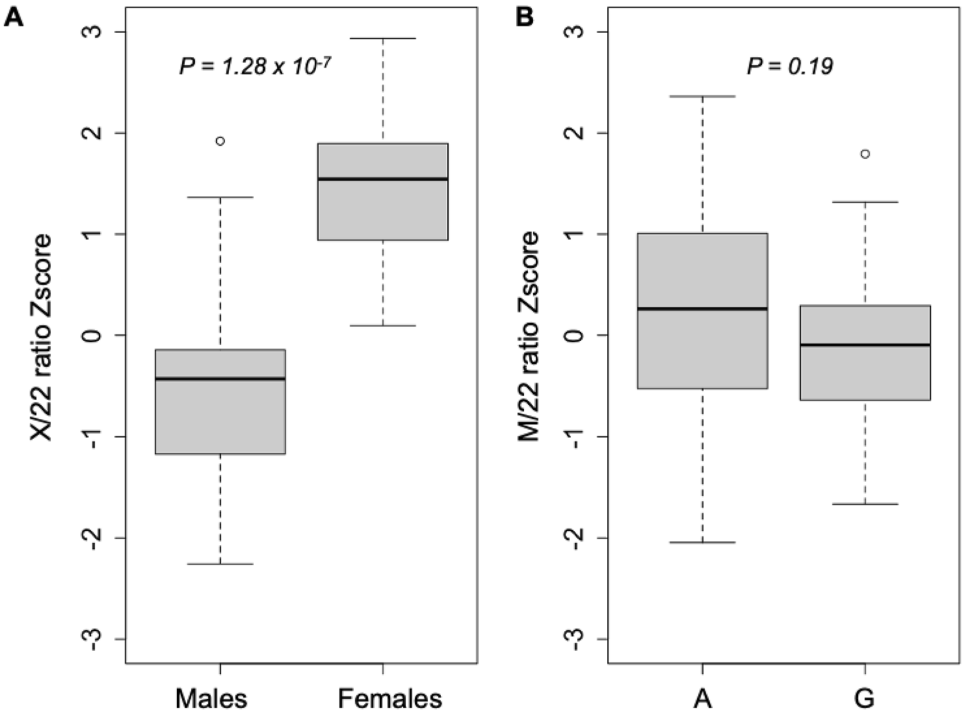

Figure 5. DNA copy number analysis using multiplex PCR represented as standard boxplots. (**A**) Analysis of the X-chromosome normalized against chromosome 22 between male and female samples. The vertical axis represents the X versus chromosome 22 ratio while the horizontal axis represents the different sexes. (**B**) Mitochondrial DNA copy number analysis normalized against chromosome 22 between the A and G allele at position 10398 of the mitochondrial DNA. The vertical axis represents the Mitochondria versus chromosome 22 ratio while the horizontal axis represents the allele status of position 10398 of the mitochondria.

altered MT copy number. The analysis did not detect any significant differences, suggesting that 10398A>G is not associated with altered MT copy number ($P = 0.19$) (Fig. 5B). In addition, we also performed MT copy number analysis using quantitative PCR (qPCR) on DNA from 52 LCL samples. Using a genomic DNA fragment on chromosome 17 as a baseline control, we did not observe any significant MT copy number difference with 10398A>G ($P = 0.33$) (Table S4).

10398A>G is significantly associated with increased mitochondrial heteroplasmy. Finally, we quantified MT DNA heteroplasmy within the DNA obtained from our 60 PGP LCL samples by next-generation sequencing of the MT control region, where we had observed all MT heteroplasmy within brain derived samples. We designed primers to sequence neighboring regions chrM:16043–16238 and chrM:16469–262 of the MT DNA and calculated a heteroplasmy score for each site (see "Materials and methods"). We used simple linear regression to look for association between age, sex, and MT copy number with individual heteroplasmy scores. There was no significant association between total heteroplasmy score and age ($R^2 = -0.01609$, $P = 0.79838$), sex ($R^2 = -0.01567$, $P = 0.75984$), or copy number mitochondrial Z-score ($R^2 = -1.5 \times 10^{-3}$, $P = 0.344$). We then performed an association test of 10398A>G with the heteroplasmy scores and found that the minor G allele of 10398 was associated with significantly increased heteroplasmy at chrM:16469–262 ($P = 0.011$) (Fig. 6B). While there was a slight increase in heteroplasmy detected for chrM:16043–16238, the association statistic was not significant ($P = 0.29$) (Fig. 6A). Probing further into the individual sites, we found certain individual sites were contributing disproportionately to the overall increase in heteroplasmy (Fig. 7).

We then performed an association test between all MT SNPs with overall heteroplasmy. We identified eight mitochondrial SNPs significantly associated ($P < 0.05$) with total heteroplasmy score (Table 4). 10398A>G was significantly associated with higher heteroplasmy score ($P = 0.021$). Furthermore, we computed the squared R^2 correlation coefficient between each SNP identified as significantly associated with heteroplasmy score with 10398A>G and found that some of them were significantly correlated with 10398A>G (Table 4).

Discussion

We analyzed large cohorts of DNA sequencing data to determine the extent of detectable mitochondrial heteroplasmy in various tissue types for different donor samples. We discovered that the 10398A>G (rs2853826) allele is significantly associated with mitochondrial heteroplasmy in the control region of brain samples. Given the reported associations between 10398A>G and a multitude of human diseases, we explored the functional consequence of 10398A>G with data available to us and discovered that it is associated with MT-ND3 expression levels in brain samples. We then characterized lymphoblastoid cell lines (LCLs) from the Harvard Personal Genome Project to determine the effect of 10398A>G in LCLs and discovered that it is associated with MT-ND4 expression levels and the MT-ND3 expression levels were not significantly altered in the LCLs. This suggests that 10398A>G could be affecting mitochondrial gene expression in a tissue or cell-type specific manner. Finally, from the sequencing data obtained from the LCLs, we found that 10398A>G is also associated with mitochondrial heteroplasmy, suggesting that heteroplasmy may be a general consequence of 10398A>G within multiple tissue types and potentially a mechanism by which 10398A>G confers an effect upon diverse human diseases.

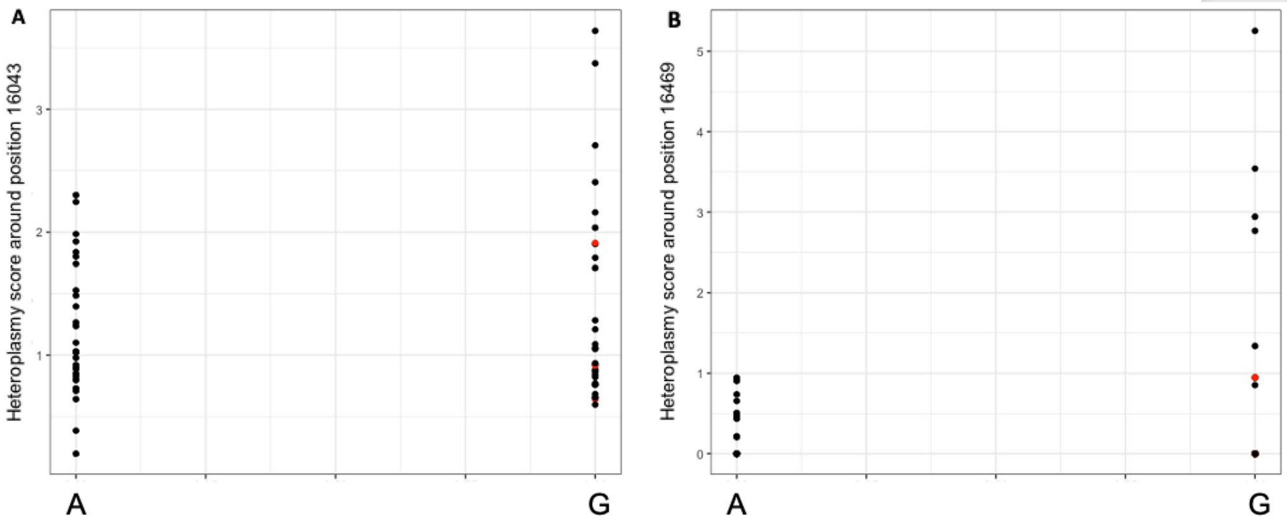

Figure 6. Association of A and G allele at position 10398 with (**A**) overall heteroplasmy score calculated for sites in the mitochondria between positions 16043 and 16238 (chrM:16043–16238) and (**B**) overall heteroplasmy score calculated for sites in the mitochondria between positions 16469 and 262 (chrM:16469–262). Each dot represents a particular individual and red dots indicate individuals of non-European ancestry stratified during PCA (see "Materials and methods"). The vertical axis represents the overall heteroplasmy score while the horizontal axis represents the allele status at position 10398 of the mitochondria.

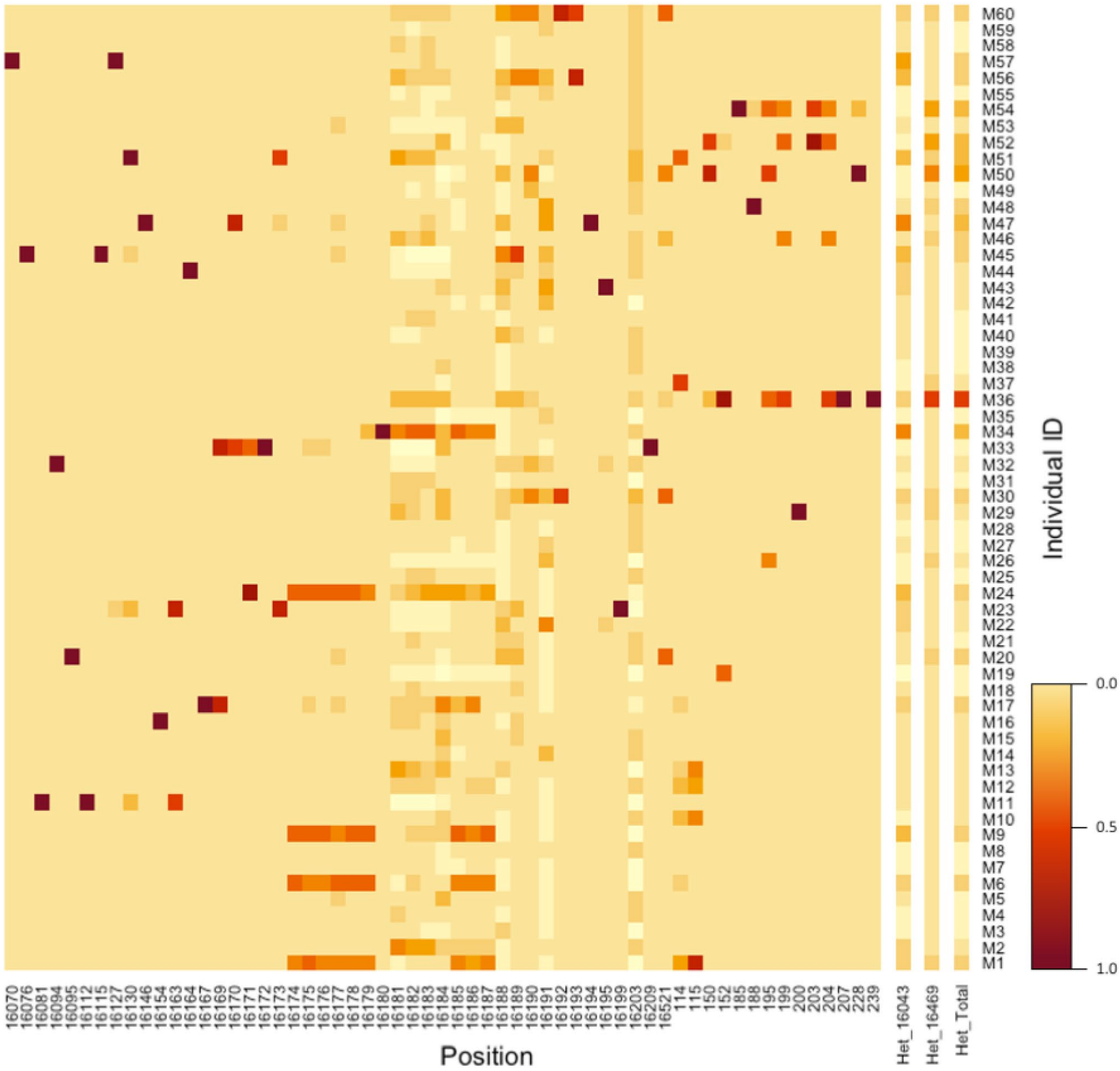

Figure 7. Individual heteroplasmic sites detected. Each row shows the heteroplasmy detected for each individual donor (M1–M60). Darker regions show a higher heteroplasmy fraction for that site. The number on the x-axis shows selected positions on the mitochondrial DNA where heteroplasmy was detected. Het_16043 depicts the cumulative heteroplasmy for sites within position 16043-of the mitochondria and Het_16469 depicts the cumulative heteroplasmy for sites within position 16469–262. Het_Total depicts the combined heteroplasmy for both fragments.

Our initial motivation for exploring mitochondrial heteroplasmy was to determine its role in Alzheimer's disease pathology. Early evidence indicated that post-mortem AD brains were enriched for heteroplasmic mitochondrial control region mutations[53]. Furthermore, recent evidence suggests mitochondrial DNA mutations can lead to mitochondrial dysfunction and increased Aβ deposition[54]. We did not observe any significant difference in overall, or locus-specific MT heteroplasmy within the DLPFC between AD and aged controls. Our subsequent eQTL findings lead us to the mitochondrial variant 10398A>G, which we found to be a strong predictor of heteroplasmic mutations in the mitochondrial control region. In an early study, 10398A>G was shown to be associated with AD risk[17]. However, a later study of a cohort of patients from Tuscany reported no mitochondrial haplotype group association with AD[55]. A later study in a cohort of Japanese individuals suggests little to no contribution of 10398A>G to AD[56]. Recently, there was a report of other mitochondrial variants associated with AD in a 254 individuals Tunisian cohort[57]. While that study did not find 10398A>G to be associated with AD, they identified other variants such as 5656A>G and 13759A>G to be marginally associated with AD. These studies suggest that the effect of 10398A>G on AD, if true, will be moderate and may not be observed in small cohort studies of only a few hundred patients. Nevertheless, it remains to be determined if the underlying cause

Position	rsID	Ref	Alt	MAF	Effect size	SE	P value	R^2 w. 10398
73	rs3087742	G	A	0.3333	− 0.671972	0.314507	0.0370148	0.486486
10398	rs2853826	A	G	0.4737	0.720784	0.304438	0.0215864	1
12705	rs2854122	C	T	0.2083	1.2496	0.394272	0.00247615	0.0362319
14766	rs3135031	T	C	0.3667	− 0.675267	0.307857	0.0324439	0.565714
15043	rs28357684	G	A	0.1167	1.94662	0.491097	0.000211222	0.0838574
15924	rs2853510	A	G	0.15	1.15994	0.39716	0.00502691	0.208333
16129	rs45134744	G	A	0.1583	1.27893	0.396306	0.0020912	0.00226131
16223	rs2853513	C	T	0.2308	1.32282	0.415834	0.0025717	0.0737281

Table 4. List of mitochondrial SNPs that were associated with overall heteroplasmy. Position depicts the specific base-pair position in the mitochondria. Ref. and Alt. stands for reference and alternate allele, respectively. MAF stands for minor allele frequency. Effect size, standard error (SE) and P value are the respective statistics for associating these SNPs with overall heteroplasmy. R^2 w. 10398 shows Pearson's correlation coefficient values for each SNP with the SNP at position 10398 on the mitochondria.

of the association of 10398A>G to AD and other human diseases could be due to the effect of mitochondrial heteroplasmy on cellular function.

Besides AD, the role of mitochondrial heteroplasmy has also been studied for other neurological diseases. In autism spectrum disorders (ASD), researchers have discovered pathogenic heteroplasmic mitochondrial mutation being associated with ASD[58,59]. Perhaps the most common and well-studied mutation linked with mitochondrial heteroplasmy is 3243A>G, which has been associated with mitochondrial disorders and dysfunctions, including Mitochondrial Encephalopathy, Lactic Acidosis, Stroke-like episodes (MELAS), and Leigh syndrome[60]. As such, our finding of an inherited polymorphism 10398A>G being linked to mitochondrial heteroplasmy may inspire future research to investigate the role of mitochondrial heteroplasmy as a causal component of disease pathology.

Our study has several important limitations. In the current paradigm, we are unable to definitively establish the directionality of causal relationships between MT-ND3/MT-ND4 expression and mitochondrial heteroplasmy (both of which are presumably downstream in their association with inherited variant 10398A>G). Whether heteroplasmy is a consequence of altered MT-ND3/MT-ND4, induced directly by 10398A>G or even indicates a pleiotropic locus that is independently driving regulatory effects and heteroplasmy, awaits further experimentation to resolve. An additional limitation pertains to our focus on mitochondrially encoded transcripts, rather than the broader set of mitochondrial genes which are also encoded within the nuclear genome. This limitation was motivated by our usage of a cis-eQTL framework, which is limited to regulatory relationships occurring within 20 kilobases, though this may well overlook important regulatory relationships that occur between genomes.

In addition, we have not presented any functional mitochondrial data within this study, which may represent a fruitful direction for future studies. While we report an intriguing eQTL relationship, the actual fold-change difference, while strongly significant, is not large in magnitude. Given that 10398A>G is a common genetic variant, this eQTL relationship may indicate a physiological association that is non-pathogenic under most tissues and conditions, but which may become pathogenic under more extreme 'stress' conditions, such as altered pH, temperature or conditions of inflammation. This could be consistent with previous reports of 10398A>G associating with altered mitochondrial matrix pH and calcium dynamics[61]. It is also possible that the observed relationship between heteroplasmy and 10398A>G in the DLPFC are confounded with other disease comorbidities that themselves are linked with 10398A>G, which are present in the ROS/MAP cohort, though not characterized within the study and are thus not accounted for in our analyses. Despite this limitation, the replication of 10398A>G as a heteroplasmy QTL in our LCL experiments is supportive of this effect not being solely mediated by unobserved comorbidities. Additional avenues for further exploration could be to look at additional populations, larger sample sizes and additional tissues beyond DLFPC.

While we found that the 10398A>G was associated with decreased gene expression in MT-ND3/MT-ND4 and increased mitochondrial heteroplasmy, we should also be cognizant that the cause of this association could be due to other genetic variants linked with 10398A>G. 10398A>G is an established marker for the I, J and K European mitochondrial haplogroup and it is possible that other variants within these haplogroups account for the heteroplasmy association. The frequency of the G allele is also significantly higher in non-European (e.g. African and Asian) populations and future follow up studies from data generated from these populations may allow us to answer this question.

Given the association of mitochondrial heteroplasmy with such a wide range of human diseases, and the previously reported progression from mitochondrial heteroplasmy to mitochondrial dysregulation and disruption of the OXPHOS pathway[62], an exploration of the downstream effects of heteroplasmy and the 10398A>G variant on metabolic pathways may be warranted. Further evaluation of associations between the 10398A>G variant with other AD-associated phenotypes may prove to be fruitful. 10398A>G has previously been linked with Human Papillomavirus (HPV) positivity in the context of cervical cancer, indicating a potential modulation of host-virus interactions[52]. Given prior evidence of the role of Herpes simplex virus 1 (HSV-1) in AD pathogenesis in APOE-ε4 carriers[63–66], one avenue of experimentation might be to test the viability of viral infection (e.g. HSV-1) in cell lines from individuals with the A versus G allele at 10398. This may provide better context for the relationships between the 10398A>G variant, mitochondrial heteroplasmy and damage, and AD pathology.

In summary, we have demonstrated the use of large scale whole-genome sequence data for analyzing mitochondrial heteroplasmy and performing cis-heteroplasmy-QTL analysis. Doing so led us to discover the association between 10398A>G and mitochondrial heteroplasmy. We then further characterize the effect of 10398A>G with the use of personalized LCL models. This highlights the utility of such resources as an effective way to perform *in-vitro* modeling for exploring genetic drivers of mitochondrial function in diseases, and suggests novel avenues for illuminating the role of heteroplasmy as a candidate mediating mechanism that links 10398A>G with diverse human diseases.

Materials and methods

Whole genome sequences from the ROS, MAP, MSBB and Mayo cohorts.

Mitochondrial variables (germline variation, heteroplasmy and copy number) were estimated on tissue samples with available whole genome sequences (WGS) generated by the Accelerating Medicines Partnership in Alzheimer's Disease (AMP-AD). BAM files can be accessed at the AD Knowledge Portal (https://adknowledgeportal.org) for the MSBB (syn19987071), Mayo (syn19989379) and ROS/MAP (syn20068543) cohorts.

Mitochondrial variant calling from brain derived whole genome sequences.

Mitochondrial SNV and INDEL variants were called on existing WGS data using the gatk4 mitochondrial pipeline available here: https://github.com/gatk-workflows/gatk4-mitochondria-pipeline.

1. Available BAM files were subset to reads with a primary mapping to MT genome using the gatk PrintReads function:
   ```
   gatk PrintReads -R human_g1k_v37.fasta -L MT --read-filter
    MateOnSameContigOrNoMappedMateReadFilter --read-filter
    MateUnmappedAndUnmappedReadFilter -I Full.WGS.bam -O MT.bam
   ```
2. MT.bam files were reheadered to exchange chromosomal notation from MT to chrM:
   ```
   samtools view -H MT.bam | sed -e "s/SN:MT/SN:chrM/" | samtools reheader -MT.bam > chrM.bam
   ```
3. chrM.bam files were indexed:
   ```
   java -jar picard.jar BuildBamIndex I = chrM.bam
   ```
4. chrM.bam files were submitted as inputs to Cromwell workflows designed for the gatk4-mitochondrial-pipeline
   ```
   java -Dconfig.file = cromwell.singularity.conf -jar cromwell-47.jar run mitochondria-pipeline.remove.unpaired.wdl -inputs sample_config.json
   ```
5. The called variants have been filtered by FilterMutectCalls to label false positives with a list of failed filters and true positives with PASS and only PASS variants were kept
   ```
   gatk FilterMutectCalls -V output.vcf.gz -contamination-table contamination.table -O filtered.vcf.gz
   gatk SelectVariants -R reference.fasta -V filtered.vcf.gz -exclude-filtered true -O final.vcf.gz
   ```
6. Called variants are merged to a single file for each cohort, leaving uncalled sites missing:bcftools merge -l file.lst -o all.vcf

Mitochondrial heteroplasmy quantitative trait loci analysis.

Mitochondrial (MT) heteroplasmy quantitative trait loci (hetQTL) analysis was performed on whole genome sequences generated from dorsolateral prefrontal cortex (DLPFC) samples collected from 451 unique subjects within the Religious Orders Study (ROS) and Memory and Ageing Project (MAP). We applied a logistic regression approach, modelling the presence (or absence) of heteroplasmy at each MT site, as a function of each MT SNP genotype, while adjusting for AD status (NIA Reagan Score), Study (ROS or MAP), Sex, Age at death, Years of education, Post Mortem Interval, and Ancestry (estimated using the 10 largest population ancestry components precomputed from the WGS data). Within the ROS/MAP DLPFC samples, we identified 77 common MT variants (MAF ≥ 5%) and nine MT heteroplasmic sites that were detected in at least five subjects, which were carried forward into the hetQTL analysis. HetQTL associations were estimated using a generalized linear model using the R function "glm" and the family type "binomial". Associations between each MT variant and MT heteroplasmic site were adjusted using the Benjamini–Hochberg approach for controlling the false discovery rate (FDR). Associations with an FDR < 0.05 were classified as significant. Results were annotated with symbols from overlapping genes, and separately with symbols from genes within 20 kilobases of MT variants and heteroplasmy sites using annotations from the GRCh38 transcriptome assembly.

Mitochondrial expression quantitative trait loci analysis. Mitochondrial (MT) cis expression quantitative trait loci (cisQTL) analysis was performed across the set of samples with whole genome sequences (originating from any tissue) and RNA-sequences generated from dorsolateral prefrontal cortex (DLPFC) samples, comprising 573 unique subjects within the Religious Orders Study (ROS) and Memory and Ageing Project (MAP). Within the ROS/MAP DLPFC samples, we identified 212 common MT variants (MAF ≥ 5%). Transcriptomic data was subset to the 37 MT encoded genes. Genes were retained in the analysis if expression was detected in at least five samples, resulting in 23 included genes. Normalized abundance values were offset by a count of 1 and log2 transformed. Cis-eQTL analysis was performed using the R package, MatrixEQTL. Gene expression was modelled using a linear additive approach, including terms for MT variant genotype and while adjusting for AD status (NIA Reagan Score), Study (ROS or MAP), Sex, Age at death, Years of education, Post-Mortem Interval, RNA Integrity Number (RIN), RNA-seq batch, and Ancestry (estimated using the 10 largest population ancestry components precomputed from the WGS data). Cis-eQTL relationships were defined as a maximum distance of 1 MB between a MT variant/gene pair, which was definitionally true for all examined MT pairs.

Construction of MT-ND3 gene regulatory network. We performed causal inference testing[47], to build a causal gene regulatory network focused on MT-ND3 in post-mortem DLPFC brain tissue collected as part of the ROS/MAP studies. This approach requires paired gene expression and genotype data for a large number of samples to establish the direction of regulation between MT-ND3 and its correlated genes.

Causal inference testing (CIT) has been well described previously[47]. Briefly, it offers a hypothesis test for whether a molecule (in this case, the expression of MT-ND3) is potentially mediating a causal association between a DNA locus (10398A>G), and some other quantitative trait (such as the expression of genes correlated with MT-ND3 and 10398A>G). Causal relationships can be inferred from a chain of mathematic conditions, requiring that for a given trio of loci (L), a potential causal mediator i.e., MT-ND3 (G) and a quantitative trait (T), the following conditions must be satisfied to establish that G is a causal mediator of the association between L and T:

(a) L and G are associated
(b) L and T are associated
(c) L is associated with G, given T
(d) L is independent of T, given G

We used the R software package "cit"[47], to perform the causal inference test, calculating a false discovery rate using 1000 test permutations. Trios with a Qvalue < 0.05 were classified as significant, and the associated T genes were considered downstream of MT-ND3.

Gene set enrichment testing. We then submitted the 1293 the downstream neighbors of MT-ND3 to the Enrichr webtool using the enrichR R package(23586463, 27141961, 33780170), specifically querying the "Reactome_2022", "GO_Biological_Process_2021", "GO_Cellular_Component_2021 , "GO_Molecular_Function_2021", and "WikiPathway_2021_Human" gene set libraries. Enrichment results with an adjusted P value < 0.05 were classified as significant.

Personal genome project lymphoblastoid cell lines and whole genome sequencing data. The lymphoblastoid cell lines (LCL) from the Personal Genome Project (PGP) cohort were obtained from the Coriell Institute for Medical Research (https://www.coriell.org/). The whole genome sequencing data (WGS) for PGP samples were obtained from the PGP website (https://pgp.med.harvard.edu). There were a total of 123 unique donors with both LCLs and WGS data available. We performed data mining of the 123 donors and determined that 30 donors carry the rs2853826 "G" allele of which 3 individuals also have the "GCT" allele (Table S5). Most of these individuals were self-reported as having European ancestry except for 4 donors (2 Chinese, 1 African-American and 1 Hispanic). We then selected an additional 30 donor samples carrying the rs2853826 "A" allele as control samples bringing the total number of donors analyzed to 60. We also performed sanger sequencing of the rs2853826 locus for all 60 samples to confirm their rs2853826 genotype (Fig. S2).

LCL cell culture, DNA and RNA extraction and RNA sequencing. Lymphoblastoid cell lines (LCL) were cultured in RPMI medium supplemented with 10% Fetal Bovine Serum (FBS) and 1% penicillin–streptomycin. These cultured LCLs were maintained at 37 °C. DNA from the LCLs were extracted using the AccuPrep® Genomic DNA Extraction Kit (Bioneer). RNA was extracted from the LCLs using the PureLink™ RNA Mini Kit (Thermofisher). The extracted RNA for whole transcriptome sequencing was sent to Psomagen for sequencing. Preparation of samples for RNA-Seq analysis was performed using the TruSeq RNA Sample Preparation Kit (Illumina, San Diego, CA). Briefly, rRNA was depleted from total RNA using the Ribo-Zero rRNA Removal Kit (Human/Mouse/Rat) (Illumina, San Diego, CA) to enrich for coding RNA and long non-coding RNA, following the TruSeq Stranded Total RNA Sample Prep Guide, Part # 15031048 Rev. The Ribo-Zero libraries were sequenced on the Illumina NovaSeq 6000 System with 151 nucleotide paired end reads, according to the standard manufacturer's protocol (Illumina, San Diego, CA). The raw sequence reads were aligned to human genome hg38 (ensembl_GRCh38, GenBank Assembly ID GCA_000001405.15) with the star aligner (v2.5.2b) and gene level expression (read counts) were summarized by the "- quantMode GeneCounts" parameter in star.

Testing mitochondrial DNA copy number. We used a multiplex PCR approach to ascertain mitochondrial DNA copy number. We designed 3 primer pairs to target the mitochondrial DNA, a region on the X-chromosome and a region on chromosome 22. The primers for the mitochondrial DNA were TACACATGCAAGCATCCCCG for the forward primer and ATCACTGCTGTTTCCCGTGG for the reverse primer. These primers target chrM:692–826 which results in a 135 bp PCR fragment. The primers for the X-chromosome DNA were ATCCCCGTGTGGTAGTCTCC for the foward primer and AGTTGCCAGACGTCTTAAAGTCC for the reverse primer. These primers target chrX:13086693–13086892 which results in a 200 bp PCR fragment. The primers for chromosome 22 were CAGAGGCTCAGAGAGGTCATCT for the forward primer and CCTAAGGTTGAGTTTGGTCTCCC for the reverse primer. These primers target chr22:37103514–37103839 which results in a 326 bp PCR fragment. 50 ng of genomic and mitochondrial DNA was used as template DNA for the multiplex PCR reaction. All primer sequences are given in the 5′ to 3′ orientation. The genome coordinates were provided using the GRCh37 (hg19) assembly of the human genome. The resulting PCR product was sent for analysis using a 5200 Fragment Analyzer System (Agilent) and the quantification for each PCR fragment is given for each sample. The multiplex PCR was performed in 2 separate batches. Batch one had samples M1–15, M31–45 and batch 2 had samples M16–30, M46–60. We then calculated the X/22 and M/22 ratios for each sample and obtained a Zscore for both statistics by normalizing against each batch's sample mean and standard deviation (Table S6).

PCR amplification and sequencing of the mitochondrial control region. We PCR amplified 2 regions of the mitochondrial DNA for sequencing to determine mitochondrial heteroplasmy. For the first site, chrM:16043–16238, we used CTTTCCCTACACGACGCTCTTCCGATCTNNNNNNgatttgggtaccacccaagt as the forward primer and GGAGTTCAGACGTGTGCTCTTCCGATCTgttgaaggttgattgctgtact as the reverse primer. As for the second site, chrM:16469–262, which consist of a 365 bp fragment, we used CTTTCCCTACACGACGCTCTTCCGATCTNNNNNNcttgggggtagctaaagtga as the forward primer and GGAGTTCAGACGTGTGCTCTTCCGATCTggctgtgcagacattcaatt as the reverse primer. For these primer sequences, the lower-case bases are homologous to the target region while the upper-case bases are overhangs required for the second round PCR. The amplified fragments then undergo a second round of PCR where the illumina indexes and primer sequences are attached using i7 and i5 primers. The i7 primer sequence was CAAGCAGAAGACGGCATACGAGAT[i7]GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCT and the i5 primer sequence was AATGATACGGCGACCACCGAGATCTACAC[i5]ACACTCTTTCCCTACACGACGCTCTTCCGATCT. The i7 and i5 sequence consist of a unique 8 bp sequence to identify the dual-indexed illumina sequence library. All sequences are depicted in the 5′ to 3′ orientation. The primer sequences were synthesized using Custom DNA Oligos Synthesis Services (Thermofisher). We used a Q5 High-Fidelity DNA Polymerase (NEB) and ran the PCR on a ProFlex Thermal Cycler (Thermofisher).

The DNA libraries were then pooled and sent for whole-genome sequencing (Psomagen). We performed Illumina paired-end sequencing (read 150 bases) on the HiSeq X Ten system, with different index barcodes for each sample resulting in roughly one million paired-end reads per sample on average (Table S7).

Calculating heteroplasmy score. We performed alignment of the PCR-amplified sequencing reads for each sample to the hg19 assembly of the human genome using bwa[67]. We mapped the aligned sequences to a fasta file containing the primers and mitochondrial genome sequences using the mpileup function of samtools[68]. From the mpileup file, we constructed a statistic to measure the degree of heteroplasmy across our nine sites of interest and the surrounding bases using the following method.

For each of the 60 individuals in our cohort, we computed the fraction of non-reference alleles to the total number of reads at all PCR-amplified positions (each mtDNA site between bases 16043 to 16238 and bases 16469 to 262). We first identified any sites with measurable heteroplasmy (noted in the formula below with the indicator function) which are sites with allele fraction estimates between 0.05 and 0.95 (Table S8). Sites with non-reference allele fraction less than or equal to 0.05 or greater than or equal to 0.95 are assigned as stably inherited SNPs, and the heteroplasmy measure at these sites are set to zero. We then weighted each heteroplasmic site by the allele fraction and summed them up to generate heteroplasmy scores for each individual.

$$Het16043_i = \sum_{m=16043}^{16238} 1_{\{0.05<MAF_{i,m}<0.95\}} * \left[\frac{\#Alt.reads_{i,m}}{Total\#reads_{i,m}}\right]$$

$$Het16469_i = \sum_{m=1}^{262} 1_{\{0.05<MAF_{i,m}<0.95\}} * \left[\frac{\#Alt.reads_{i,m}}{Total\#reads_{i,m}}\right] + \sum_{m=16469}^{16571} 1_{\{0.05<MAF_{i,m}<0.95\}} * \left[\frac{\#Alt.reads_{i,m}}{Total\#reads_{i,m}}\right]$$

$$HetTotal_i = Het16043_i + Het16469_i$$

Het16043$_i$ is the heteroplasmy score of individual *i* generated from sites between bases 16043 and 16238. Het16469$_i$ is the heteroplasmy score of individual *i* generated from sites between bases 16469 and 262 and HetTotal$_i$ is the heteroplasmy score of individual *i* generated by taking the sum of Het16043$_i$ and Het16469$_i$.

Quality control and association analysis of MT SNPs with heteroplasmy. Plink 1.90b6.9[69] was used to perform quality control and association analysis. We ensured that no individuals had a genotyping rate less than 20%, discordant sex data, or were duplicated or shared significant identity-by-descent. Because of limited sample size, SNPs with minor allele frequency less than or equal to 0.09 as well as those not in Hardy-Weinburg equilibrium ($P < 1.0 \times 10^{-6}$) were removed. Of the 60 samples in our cohort of LCLs, we found that three individuals of non-European ancestry (2 Chinese, 1 African-American) were stratified in PCA (Fig. S3). We used Plink to implement a generalized linear regression model for the remaining 42 MT SNPs for the 60 LCL samples in our cohort, including the first two principal components from PCA as covariates, with HetTotal_i and calculated the overall association statistics (Table S9).

Ethical approval. The use of the human cell lines for research in this work was determined to not be human subjects research by the IRB of UMass Chan Medical School (H00021419).

Data availability

Mitochondrial data derived from whole genome sequences are available via the AD Knowledge Portal with data identifier syn25927578 at https://doi.org/10.7303/syn2580853.

Received: 23 February 2023; Accepted: 23 June 2023
Published online: 27 June 2023

References

1. Lightowlers, R. N., Taylor, R. W. & Turnbull, D. M. Mutations causing mitochondrial disease: What is new and what challenges remain?. *Science* **349**, 1494–1499 (2015).
2. Koopman, W. J. H., Distelmaier, F., Smeitink, J. A. & Willems, P. H. OXPHOS mutations and neurodegeneration. *EMBO J.* **32**, 9–29 (2013).
3. Yadav, N. & Chandra, D. Mitochondrial DNA mutations and breast tumorigenesis. *Biochim. Biophys. Acta* **1836**, 336–344 (2013).
4. Schöpf, B. *et al.* OXPHOS remodeling in high-grade prostate cancer involves mtDNA mutations and increased succinate oxidation. *Nat. Commun.* **11**, 1487 (2020).
5. PerezOrtiz, J. M. & Swerdlow, R. H. Mitochondrial dysfunction in Alzheimer's disease: Role in pathogenesis and novel therapeutic opportunities. *Br. J. Pharmacol.* **176**, 3489–3507 (2019).
6. Hoekstra, J. G., Hipp, M. J., Montine, T. J. & Kennedy, S. R. Mitochondrial DNA mutations increase in early stage Alzheimer disease and are inconsistent with oxidative damage: Mitochondrial Mutations in AD. *Ann. Neurol.* **80**, 301–306 (2016).
7. Toledo, J. B. *et al.* Metabolic network failures in Alzheimer's disease: A biochemical road map. *Alzheimers Dement. J. Alzheimers Assoc.* **13**, 965–984 (2017).
8. Misrani, A., Tabassum, S. & Yang, L. Mitochondrial dysfunction and oxidative stress in Alzheimer's disease. *Front. Aging Neurosci.* **13**, 617588 (2021).
9. Soltys, D. T. *et al.* Lower mitochondrial DNA content but not increased mutagenesis associates with decreased base excision repair activity in brains of AD subjects. *Neurobiol. Aging* **73**, 161–170 (2019).
10. Klein, H.-U. *et al.* Characterization of mitochondrial DNA quantity and quality in the human aged and Alzheimer's disease brain. *Mol. Neurodegener.* **16**, 75 (2021).
11. Reiman, E. M. *et al.* Exceptionally low likelihood of Alzheimer's dementia in APOE2 homozygotes from a 5000-person neuropathological study. *Nat. Commun.* **11**, 667 (2020).
12. Yin, J. *et al.* Effect of ApoE isoforms on mitochondria in Alzheimer disease. *Neurology* **94**, e2404–e2411 (2020).
13. Chakravorty, A., Jetto, C. T. & Manjithaya, R. Dysfunctional mitochondria and mitophagy as drivers of Alzheimer's disease pathogenesis. *Front. Aging Neurosci.* **11**, 311 (2019).
14. Wang, W., Zhao, F., Ma, X., Perry, G. & Zhu, X. Mitochondria dysfunction in the pathogenesis of Alzheimer's disease: Recent advances. *Mol. Neurodegener.* **15**, 30 (2020).
15. Stewart, J. B. & Chinnery, P. F. The dynamics of mitochondrial DNA heteroplasmy: Implications for human health and disease. *Nat. Rev. Genet.* **16**, 530–542 (2015).
16. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141456 humans. *Nature* **581**, 434–443 (2020).
17. van der Walt, J. M. *et al.* Analysis of European mitochondrial haplogroups with Alzheimer disease risk. *Neurosci. Lett.* **365**, 28–32 (2004).
18. van der Walt, J. M. *et al.* Mitochondrial polymorphisms significantly reduce the risk of Parkinson disease. *Am. J. Hum. Genet.* **72**, 804–811 (2003).
19. Otaegui, D. *et al.* Mitochondrial polymporphisms in Parkinson's disease. *Neurosci. Lett.* **370**, 171–174 (2004).
20. Pyle, A. *et al.* Mitochondrial DNA haplogroup cluster UKJT reduces the risk of PD. *Ann. Neurol.* **57**, 564–567 (2005).
21. Ghezzi, D. *et al.* Mitochondrial DNA haplogroup K is associated with a lower risk of Parkinson's disease in Italians. *Eur. J. Hum. Genet.* **13**, 748–752 (2005).
22. Huerta, C. *et al.* Mitochondrial DNA polymorphisms and risk of Parkinson's disease in Spanish population. *J. Neurol. Sci.* **236**, 49–54 (2005).
23. Latsoudis, H., Spanaki, C., Chlouverakis, G. & Plaitakis, A. Mitochondrial DNA polymorphisms and haplogroups in Parkinson's disease and control individuals with a similar genetic background. *J. Hum. Genet.* **53**, 349–356 (2008).
24. Huerta, C. *et al.* No association between Parkinson's disease and three polymorphisms in the eNOS, nNOS, and iNOS genes. *Neurosci. Lett.* **413**, 202–205 (2007).
25. Canter, J. A., Kallianpur, A. R., Parl, F. F. & Millikan, R. C. Mitochondrial DNA G10398A polymorphism and invasive breast cancer in African–American women. *Cancer Res.* **65**, 8028–8033 (2005).
26. Bai, R.-K., Leal, S. M., Covarrubias, D., Liu, A. & Wong, L.-J.C. Mitochondrial genetic background modifies breast cancer risk. *Cancer Res.* **67**, 4687–4694 (2007).
27. Darvishi, K., Sharma, S., Bhat, A. K., Rai, E. & Bamezai, R. N. K. Mitochondrial DNA G10398A polymorphism imparts maternal Haplogroup N a risk for breast and esophageal cancer. *Cancer Lett.* **249**, 249–255 (2007).
28. Setiawan, V. W. *et al.* Mitochondrial DNA G10398A variant is not associated with breast cancer in African–American women. *Cancer Genet. Cytogenet.* **181**, 16–19 (2008).
29. Covarrubias, D., Bai, R.-K., Wong, L.-J.C. & Leal, S. M. Mitochondrial DNA variant interactions modify breast cancer risk. *J. Hum. Genet.* **53**, 924–928 (2008).
30. Pezzotti, A. *et al.* The mitochondrial A10398G polymorphism, interaction with alcohol consumption, and breast cancer risk. *PLoS One* **4**, e5356 (2009).
31. Czarnecka, A. M. *et al.* Mitochondrial NADH-dehydrogenase polymorphisms as sporadic breast cancer risk factor. *Oncol. Rep.* **23**, 531–535 (2010).

32. Salas, A., García-Magariños, M., Logan, I. & Bandelt, H.-J. The saga of the many studies wrongly associating mitochondrial DNA with breast cancer. *BMC Cancer* **14**, 659 (2014).
33. Grzybowska-Szatkowska, L. & Slaska, B. Mitochondrial NADH dehydrogenase polymorphisms are associated with breast cancer in Poland. *J. Appl. Genet.* **55**, 173–181 (2014).
34. Jahani, M. M., AzimiMeibody, A., Karimi, T., Banoei, M. M. & Houshmand, M. An A10398G mitochondrial DNA alteration is related to increased risk of breast cancer, and associates with Her2 positive receptor. *Mitochondrial DNA Part DNA Mapp. Seq. Anal.* **31**, 11–16 (2020).
35. Rai, E. *et al.* Interaction between the UCP2-866G/A, mtDNA 10398G/A and PGC1alpha p.Thr394Thr and p.Gly482Ser polymorphisms in type 2 diabetes susceptibility in North Indian population. *Hum. Genet.* **122**, 535–540 (2007).
36. Bhat, A. *et al.* The possible role of 10398A and 16189C mtDNA variants in providing susceptibility to T2DM in two North Indian populations: A replicative study. *Hum. Genet.* **120**, 821–826 (2007).
37. Liao, W.-Q. *et al.* Novel mutations of mitochondrial DNA associated with type 2 diabetes in Chinese Han population. *Tohoku J. Exp. Med.* **215**, 377–384 (2008).
38. Chen, S. *et al.* Mitochondrial NADH dehydrogenase subunit 3 polymorphism associated with an earlier age at onset in male Machado-Joseph disease patients. *CNS Neurosci. Ther.* **22**, 38–42 (2016).
39. Jin, E.-H., Sung, J. K., Lee, S.-I. & Hong, J. H. Mitochondrial NADH dehydrogenase subunit 3 (MTND3) polymorphisms are associated with gastric cancer susceptibility. *Int. J. Med. Sci.* **15**, 1329–1333 (2018).
40. Ball, M. P. *et al.* A public resource facilitating clinical use of genomes. *Proc. Natl. Acad. Sci. USA* **109**, 11920–11927 (2012).
41. Ball, M. P. *et al.* Harvard Personal Genome Project: Lessons from participatory public research. *Genome Med.* **6**, 10 (2014).
42. Mao, Q. *et al.* The whole genome sequences and experimentally phased haplotypes of over 100 personal genomes. *GigaScience* **5**, 42 (2016).
43. Chan, Y. *et al.* An unbiased index to quantify participant's phenotypic contribution to an open-access cohort. *Sci. Rep.* **7**, 46148 (2017).
44. Stoneking, M., Hedgecock, D., Higuchi, R. G., Vigilant, L. & Erlich, H. A. Population variation of human mtDNA control region sequences detected by enzymatic amplification and sequence-specific oligonucleotide probes. *Am. J. Hum. Genet.* **48**, 370–382 (1991).
45. Anderson, S. *et al.* Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457–465 (1981).
46. Lott, M. T. *et al.* mtDNA variation and analysis using mitomap and mitomaster. *Curr. Protoc. Bioinform.* **44**, 123126 (2013).
47. Millstein, J., Zhang, B., Zhu, J. & Schadt, E. E. Disentangling molecular relationships with a causal inference test. *BMC Genet.* **10**, 23 (2009).
48. Chen, E. Y. *et al.* Enrichr: Interactive and collaborative HTML5 gene list enrichment analysis tool. *BMC Bioinform.* **14**, 128 (2013).
49. Kuleshov, M. V. *et al.* Enrichr: A comprehensive gene set enrichment analysis web server 2016 update. *Nucleic Acids Res.* **44**, W90-97 (2016).
50. Xie, Z. *et al.* Gene set knowledge discovery with Enrichr. *Curr. Protoc.* **1**, e90 (2021).
51. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
52. Feng, D. *et al.* An association analysis between mitochondrial DNA content, G10398A polymorphism, HPV infection, and the prognosis of cervical cancer in the Chinese Han population. *Tumour Biol. J. Int. Soc. Oncodevelopmental Biol. Med.* **37**, 5599–5607 (2016).
53. Coskun, P. E., Beal, M. F. & Wallace, D. C. Alzheimer's brains harbor somatic mtDNA control-region mutations that suppress mitochondrial transcription and replication. *Proc. Natl. Acad. Sci. USA* **101**, 10726–10731 (2004).
54. Lee, Y. *et al.* Mitochondrial genome mutations and neuronal dysfunction of induced pluripotent stem cells derived from patients with Alzheimer's disease. *Cell Prolif.* **55**, e13274 (2022).
55. Mancuso, M. *et al.* Lack of association between mtDNA haplogroups and Alzheimer's disease in Tuscany. *Neurol. Sci.* **28**, 142–147 (2007).
56. Tanaka, N. *et al.* Mitochondrial DNA variants in a Japanese population of patients with Alzheimer's disease. *Mitochondrion* **10**, 32–37 (2010).
57. Ben Salem, N. *et al.* Mitochondrial DNA and Alzheimer's disease: A first case–control study of the Tunisian population. *Mol. Biol. Rep.* **49**, 1687–1700 (2022).
58. Wang, Y., Picard, M. & Gu, Z. Genetic evidence for elevated pathogenicity of mitochondrial DNA heteroplasmy in autism spectrum disorder. *PLoS Genet.* **12**, e1006391 (2016).
59. Wang, Y. *et al.* Association of mitochondrial DNA content, heteroplasmies and inter-generational transmission with autism. *Nat. Commun.* **13**, 3790 (2022).
60. Mancuso, M. *et al.* The m.3243A>G mitochondrial DNA mutation and related phenotypes. A matter of gender?. *J. Neurol.* **261**, 504–510 (2014).
61. Kazuno, A. *et al.* Identification of mitochondrial DNA polymorphisms that alter mitochondrial matrix pH and intracellular calcium dynamics. *PLoS Genet.* **2**, e128 (2006).
62. Elorza, A. A. & Soffia, J. P. mtDNA heteroplasmy at the core of aging-associated heart failure. An integrative view of OXPHOS and mitochondrial life cycle in cardiac mitochondrial physiology. *Front. Cell Dev. Biol.* **9**, 625020 (2021).
63. Readhead, B. *et al.* Multiscale analysis of independent Alzheimer's cohorts finds disruption of molecular, genetic, and clinical networks by human herpesvirus. *Neuron* **99**, 64-82.e7 (2018).
64. Eimer, W. A. *et al.* Alzheimer's disease-associated β-amyloid is rapidly seeded by herpesviridae to protect against brain infection. *Neuron* **99**, 56-63.e3 (2018).
65. Linard, M. *et al.* Interaction between *APOE4* and herpes simplex virus type 1 in Alzheimer's disease. *Alzheimers Dement.* **16**, 200–208 (2020).
66. Itzhaki, R. F. Overwhelming evidence for a major role for herpes simplex virus type 1 (HSV1) in Alzheimer's disease (AD); underwhelming evidence against. *Vaccines* **9**, 679 (2021).
67. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinf. Oxf. Engl.* **25**, 1754–1760 (2009).
68. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinf. Oxf. Engl.* **25**, 2078–2079 (2009).
69. Chang, C. C. *et al.* Second-generation PLINK: Rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).

Acknowledgements

The results published here are in whole or in part based on data obtained from the AD Knowledge Portal (https://adknowledgeportal.org). Data pertaining to the ROS and MAP studies were provided by the Rush Alzheimer's Disease Center, Rush University Medical Center, Chicago. Data collection was supported through funding by NIA Grants P30AG10161 (ROS), R01AG15819 (ROSMAP; genomics and RNAseq), R01AG17917 (MAP), R01AG30146, R01AG36042 (5hC methylation, ATACseq), RC2AG036547 (H3K9Ac), R01AG36836 (RNAseq), R01AG48015 (monocyte RNAseq) RF1AG57473 (single nucleus RNAseq), U01AG32984 (genomic and whole exome sequencing), U01AG46152 (ROSMAP AMP-AD, targeted proteomics), U01AG46161(TMT proteomics), U01AG61356 (whole genome sequencing, targeted proteomics, ROSMAP AMP-AD), the Illinois Department of Public Health (ROSMAP), and the Translational Genomics Research Institute (genomic). Additional phenotypic data can be requested at www.radc.rush.edu. Data pertaining to the MSBB study was generated from postmortem brain tissue collected through the Mount Sinai VA Medical Center Brain Bank and was provided by Dr. Eric Schadt from Mount Sinai School of Medicine. The Mayo study data was led by Dr. Nilüfer Ertekin-Taner, Mayo Clinic, Jacksonville, FL as part of the multi-PI U01 AG046139 (MPIs Golde, Ertekin-Taner, Younkin, Price). Samples were provided from the following sources: The Mayo Clinic Brain Bank. Data collection was supported through funding by NIA Grants P50 AG016574, R01 AG032990, U01 AG046139, R01 AG018023, U01 AG006576, U01 AG006786, R01 AG025711, R01 AG017216, R01 AG003949, NINDS Grant R01 NS080820, CurePSP Foundation, and support from Mayo Foundation. Study data includes samples collected through the Sun Health Research Institute Brain and Body Donation Program of Sun City, Arizona. The Brain and Body Donation Program is supported by the National Institute of Neurological Disorders and Stroke (U24 NS072026 National Brain and Tissue Resource for Parkinsons Disease and Related Disorders), the National Institute on Aging (P30 AG19610 Arizona Alzheimers Disease Core Center), the Arizona Department of Health Services (contract 211002, Arizona Alzheimers Research Center), the Arizona Biomedical Research Commission (contracts 4001, 0011, 05-901 and 1001 to the Arizona Parkinson's Disease Consortium) and the Michael J. Fox Foundation for Parkinsons Research. BR is supported in part by NIA Grants U01AG061835 and R21AG063068. QW is supported in part by NIA Grants U01AG061835. ETL and YC are supported in part by NIA Grant U01AG061835. BR and DM are grateful for funding support received from the Arizona Alzheimer's Consortium.

Author contributions

B.R and Y.C planned the study. M.S. P.D. and B.R performed bioinformatics analyses. M.S, B.R, Y.C wrote the main manuscript text and prepared figures. D.M performed experiments. All authors reviewed the manuscript.

Competing interests

GMC holds leadership positions in many companies related to DNA sequencing technologies. A full list of these companies is available at http://arep.med.harvard.edu/gmc/tech.html. The remaining authors declare that they have no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-37541-y.

Correspondence and requests for materials should be addressed to Y.C. or B.R.