Mitochondrial DNA variants modulate N-formylmethionine, proteostasis and risk of late-onset human diseases

Author list:

Na Cai[1,2,§,†], Aurora Gomez-Duran[3,4,£,†], Ekaterina Yonova-Doing[5,^], Kousik Kundu[1], Annette I. Burgess[6,$], Zoe J. Golder[3,4,$], Claudia Calabrese[3,4], Marc J. Bonder[2,#], Marta Camacho[3], Rachael A. Lawson[7], Lixin Li[6], Caroline H Williams-Gray[3], ICICLE-PD Study Group[±], Emanuele Di Angelantonio[5,6,9,10], David J. Roberts[9,11,12], Nick A. Watkins[13], Willem H. Ouwehand[1,6,13,14], Adam S. Butterworth[5,6,9,10], Isobel D. Stewart[15], Maik Pietzner[15], Nick J. Wareham[15], Claudia Langenberg[15], John Danesh[1,5,6,9,10], Klaudia Walter[1], Peter M. Rothwell[6], Joanna M. M. Howson[5,^,*], Oliver Stegle[2,16,#,*], Patrick F. Chinnery[3,4,*], Nicole Soranzo[1,14,*]

Affiliations:

1. Human Genetics Department, Wellcome Sanger Institute (WT), Hinxton, UK
2. European Bioinformatics Institute (EMBL-EBI), Hinxton, UK
3. Department of Clinical Neurosciences, School of Clinical Medicine, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK
4. Medical Research Council Mitochondrial Biology Unit, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK
5. British Heart Foundation Cardiovascular Epidemiology Unit, Department of Primary Public Health and Primary care, University of Cambridge, Cambridge, UK
6. Stroke Prevention Research Unit, Nuffield Department of Clinical Neurosciences, John Radcliffe Hospital, Oxford, UK
7. Translational and Clinical Research Institute, Newcastle University, Newcastle, UK
8. British Heart Foundation Centre of Research Excellence, University of Cambridge, Cambridge, UK
9. National Institute for Health Research Blood and Transplant Research Unit in Donor Health and Genomics, University of Cambridge, Cambridge, UK
10. Health Data Research UK Cambridge, Wellcome Genome Campus and University of Cambridge, Cambridge, UK
11. NHS Blood and Transplant-Oxford Centre, Level 2, John Radcliffe Hospital, Oxford, UK
12. Radcliffe Department of Medicine, University of Oxford, Oxford, UK
13. NHS Blood and Transplant, Cambridge Biomedical Campus, Long Road, Cambridge, UK
14. Department of Haematology, University of Cambridge, Cambridge, UK
15. MRC Epidemiology Unit, University of Cambridge, Cambridge, UK
16. European Molecular Biology Laboratory, Meyerhofstraße 1, Heidelberg, Germany

[§] Current address: Helmholtz Pioneer Campus, Helmholtz Zentrum München, Ingolstädter Landstraße 1, Neuherberg, Germany

[£] Current address: Centro de Investigaciones Biológicas "Margarita Salas", Consejo Superior de Investigaciones Científicas (CIB-CSIC), Madrid, Spain.

[^] Current address: Novo Nordisk Research Centre Oxford, Innovation Building, Old Road Campus, Oxford, UK

[#] Current address: Deutsches Krebsforschungszentrum, Im Neuenheimer Feld 280, Heidelberg, Germany

† These authors contributed equally
* Jointly supervised the work
± A list of authors and their affiliations appears at the end of the paper

Correspondence should be addressed to O.S. (o.stegle@dkfz-heidelberg.de), P.F.C. (pfc25@cam.ac.uk) and N. S. (ns6@sanger.ac.uk)

Abstract:

Mitochondrial DNA (mtDNA) variants influence the risk of late-onset human diseases, but the reasons are poorly understood. Undertaking an hypothesis-free analysis of 5,689 blood-derived biomarkers with mtDNA variants in 16,220 healthy donors, here we show that variants defining mtDNA haplogroups Uk and H4 modulate the level of circulating N-formylmethionine (fMet), which initiates mitochondrial protein translation. In human cybrid lines, fMet modulated both mitochondrial and cytosolic proteins on multiple levels - through transcription, post-translational modification, and proteolysis by an N-degron pathway - abolishing known differences between mtDNA haplogroups. In a further 11,966 individuals, fMet levels contributed to all-cause mortality and the disease risk of several common cardiovascular disorders. Together these findings indicate that fMet plays a key role in common age-related disease through pleiotropic effects on cell proteostasis.

Introduction

The 16.5kb human mitochondrial genome (mtDNA) encodes 13 proteins of the electron transport chain (ETC) and the tRNA and rRNA machinery necessary for their transcription and translation *in situ*[1]. Genetic diversity in the maternally inherited[2] mtDNA with high mutation rates[3] has been effectively used for studies of human evolution and phylogenies[4–7]. Because of the ubiquitous and essential roles of mtDNA encoded proteins in cellular metabolism[8], mtDNA sequence variations have been examined extensively for their effects on cellular metabolism[9–12] and human health and diseases.

A compendium of rare mutations in mtDNA genes encoding ETC subunits have been identified to cause severe multisystemic diseases[13,14], commonly due to the primary biochemical consequences of the mutations on oxidative phosphorylation (OXPHOS) and the synthesis of adenosine triphosphate (ATP)[15]. mtDNA variations and somatic mutations have also been shown to be important in the pathology of cancers[16–18] and inducing the Warburg effect[19]. Common mtDNA variants with less severe phenotypes have been shown to affect the risk of complex late-onset human diseases[20], including neurodegenerative diseases like Alzheimer's disease [21] and Parkinson's disease[22], cardiovascular diseases like ischemic stroke[23], myocardial infarction[24] and coronary artery disease[25], and metabolic diseases like type 2 diabetes[26,27]. In some cases, mtDNA haplogroups have pleiotropic effects on multiple diseases, while in others they have opposite effects between diseases [28,29].

The relevance of mtDNA variations to health and disease is apparent[30], but understanding the molecular underpinnings of genetic associations on the mtDNA is not straightforward, especially when they do not directly implicate OXPHOS[31], implicating hitherto unknown mechanisms. Studies have identified mitophagy due to accumulation of reactive oxygen species (ROS)[32,33] and impairment of intra-mitochondrial protein synthesis[34,35] as potential mechanisms behind mtDNA associations with diseases, in addition to those that directly impact ETC complex functions and OXPHOS efficiency[32,36]. Most of the mtDNA associations examined, however, are rare variations with large effects on rare and severe diseases. Functional analysis of mtDNA variations altering risks to common diseases are relatively underexplored, and more difficult due to their smaller effect sizes.

To bridge this gap and discover new ways that mtDNA variants contribute to physiology and complex diseases, we took a phenome-driven and unbiased approach to survey the impact of mtDNA polymorphisms on a wide set of 5,689 molecular and metabolic traits beyond ATP synthesis (**Figure 1A**). Using mtDNA variations identified from whole-genome and whole-exome sequencing in 16,220 healthy individuals, we found novel associations between mtDNA variations in Haplogroup Uk with the metabolite N-formylmethionine (fMet). We followed up on this finding with analysis of mtDNA effects on gene expression in 44 tissues from the GTEx Consortium[37], and dissected the molecular consequences of the associations using human cytoplasmic hybrid (cybrids) cell lines, including the biogenesis of mitochondrial complexes, efficiency of OXPHOS, and both cytoplasmic and intra-mitochondrial protein synthesis and degradation. Finally, we examined the relevance of our findings in health and disease by verifying the effects of fMet and mtDNA Haplogroup Uk in disease and longitudinal cohorts. Our findings explain how common mtDNA variants can impact cellular proteostasis, providing a hitherto unknown genetic checkpoint and an easily measured circulating biomarker for late-onset diseases.

Results

mtDNA genotyping in 16,220 individuals

To identify mtDNA polymorphic sites for association analysis, we obtained short-read whole-genome sequencing (WGS, mean coverage = 26.8x, SD = 3.1x, in N=12,111 participants) and whole-exome sequencing (WES, mean coverage = 48.0x, SD = 8.6x, in N=4,470 participants) data in a total of 16,220 unrelated European descent participants in the INTERVAL study[38] (**Supplementary Figure 1**). We recovered a mean coverage of 2022.6x (SD = 566.5x) and 30.6x (SD = 13.5x) on the mtDNA from WGS and WES respectively, and identified 5,247 homoplasmic variants from WGS and WES (of which 5,161 are single nucleotide polymorphisms [SNPs]) using GATK HaplotypeCaller v4 [39] (**Supplementary Figure 1**).

We took rigorous steps to avoid spurious variant calls, leveraging reference mtDNA genotyping results from Affymetrix array data to recalibrate genotype quality filters, and high coverage on mtDNA from the WGS for identifying loci with high frequency of heteroplasmic variants that may be misidentified as homoplasmic (**Supplementary Figure 2**). For subsequent analyses we considered the 396 high-confidence SNPs that could be accurately identified from both WES and WGS; of these, 184 were common with minor population allele frequency (MAF) greater than 0.01 (**Supplementary Figure 2, Supplementary Table 1**). We use these 184 SNPs, consistently in all following analyses on the INTERVAL cohort. Principal component analysis using common mtDNA SNPs demonstrated the expected

clustering of individuals by their mitochondrial haplogroup (**Figure 1B**), distinct from population structure identified using SNPs from the nDNA (**Figure 1 C,D**).

Common mtDNA variants are associated with blood N-formylmethionine

We considered 5,689 distinct quantitative phenotypes representing a broad spectrum of biological processes and pathways for association with mtDNA variations. These included a total of 36 haematological traits, 1,344 small-molecule metabolites and 4,309 proteins measured in whole blood or plasma using nine high-throughput phenotyping platforms (**Figure 1A, Supplementary Table 2**). Many of these measures have been previously reported to be affected by genetic variants in the nDNA[40,41]. To specifically investigate the effect of mtDNA variants on these biomarkers, we tested the association of each of the 184 common mtDNA SNPs with each of the phenotypes using linear mixed models (LMM, **Online Methods**)[42].

One metabolite, N-formylmethionine (fMet), was significantly associated with mtDNA variants **(**5% false discovery rate [FDR], adjusted for SNPs and phenotypes; equivalent to $P= 1.65x10^{-7}$**)**, driven by eight mtDNA SNPs (**Figure 2A)**. These represent three independent associations **(Figure 2B,C)**, two with increased fMet levels (top SNP mt.1811A>G in *MT-RNR2*, MAF= 0.129, $P= 3.03x10^{-10}$, Beta[SE]= 0.12[0.02], and mt.1189T>C in *MT-RNR1*, MAF= 0.073, $P= 7.57x10^{-8}$, Beta[SE]= 0.13[0.03]), and one with decreased fMet (mt.3992C>T, T229M in *MT-ND1*, MAF= 0.02, $P= 7.86x10^{-6}$, Beta[SE]= -0.19[0.04]). Twenty-eight more mtDNA SNPs reached suggestive levels of association with one or more of the proteins or metabolites ($P\leq 5x10^{-5}$, **Supplementary Table 3**). To test the robustness of the association with fMet, we first sought to assess the contribution of potential biological pathways or bacterial exposure that may increase exogenous fMet. Using the Covariates in Multi-phenotype Studies (CMS)[43], we selected informative covariates from 896 metabolites and 36 blood cell measures (which can act as proxies for immune reactions against infections) and we assessed the association between mtDNA SNPs and fMet accounting for their contributions. Accounting for these additional covariates increased the statistical significance of the associations between mtDNA SNPs and fMet levels (e.g. at mt.1811A>G, CMS $P= 4.63x10^{-18}$, Beta[SE]= 0.06[0.01]) and identified seven additional SNPs, yielding a total of 15 fMet-associated variants (**Figure 2B**). We verified that these findings are not likely due to mtDNA copy number differences between haplogroups (**Online Methods, Supplementary Figure 3**), or artefacts introduced during our processing of metabolite data (**Online Methods, Supplementary Figure 4**).

We found that 14 of the 15 mtDNA SNPs associated with fMet levels (**Figure 2B**) defined a single phylogenetic lineage, super-haplogroup Uk (containing branches Uk1 and Uk2, **Figure 2C,D**). This includes two out of the three independent SNPs found to be associated with fMet, mt.1811A>G, which resides on super-haplogroup Uk, and mt.1189T>C, which resides on haplogroup Uk1. The remaining SNP mt.3992C>T resides on the H4 branch of Haplogroup H4'9 [44]. In a LMM association analysis between all 896 metabolites and haplogroups (instead of mtDNA SNPs), we found that only Haplogroup Uk was significantly associated with fMet ($P= 4.41x10^{-8}$, Beta[SE]= 0.25[0.05], at $P< 0.0029$ = 0.05/17 haplogroups tested). We then reassessed all 15 fMet-associated mtDNA SNPs conditioning on Haplogroup Uk, and confirmed that associations of SNPs mt.1811A>G (conditional $P= 1.37x10^{-3}$, Beta[SE]= 0.10[0.03] and mt.3992C>T (conditional $P= 2.57x10^{-5}$, Beta[SE]= -0.18[0.04]) were independent from haplogroup Uk (**Supplementary Table 4**).

Finally, we sought replication in plasma samples from 11,538 participants from the EPIC-Norfolk study[45]. Of the 15 variants associated with fMet levels in INTERVAL, eight were available in EPIC-Norfolk, with at least one SNP representing each of the three independent signals (). All eight SNPs replicated the fMet association (P<0.0063 = 0.05/8, Bonferroni adjusted for eight SNPs tested, **Figure 2B**).

Relative contribution of mtDNA and nDNA to fMet levels

A first question is whether the three mtDNA variants (mt.1811A>G on Uk, mt.1189T>C on Uk1 and mt.3992C>T on H4) exerted effects on fMet independently of the nuclear genome. We considered 5,577,007 array-imputed nuclear DNA (nDNA) SNPs with MAF≥ 0.05, and used a LMM model to test for association with fMet (**Online Methods**). This analysis yielded an association at intronic variants (top SNP rs550045, chr9:130477160, MAF= 0.488, P= $1.14x10^{-26}$, Beta[SE]= 0.19[0.02], **Figure 3A)** in *PTRH1*, encoding a human homolog of the yeast peptidyl-tRNA hydrolase 1 gene, with unverified hydrolase function in humans. Conditional analyses showed that all mtDNA effects on fMet were independent of the nDNA association at *PTRH1* (LMM P< $1.86x10^{-3}$; **Supplementary Table 4**). Further, no significant interaction effects were found between the top nDNA SNP rs550045 and the top mtDNA SNP mt.1811A>G (interaction P= 0.95, **Figure 3B**). Overall, mtDNA SNPs collectively contribute to 5.85% (SE= 1.20%) of variance in fMet levels, as compared to 14.00% (SE= 3.32%) explained by nDNA SNPs and 59.03% (SE= 1.75%) by unknown factors captured through all plasma metabolites and blood cell counts (**Figure 3C, Supplementary Discussion**).

fMet modulates mitochondrial function throughout coordination of mitochondrial transcription-translation

fMet is the initiation amino acid intra-mitochondrial protein translation[46,47]. Thus, we hypothesized that variations in fMet levels would be accompanied by changes in mtDNA gene expression and proteostasis. We first sought to assess whether fMet-associated mitochondrial variants impact on intra-mitochondrial gene expression, and assessed mtDNA SNP effects on mtDNA-encoded transcript levels in 41 primary tissues from a total of 456 unrelated donors of European descent in the GTEx Consortium v7 [37] (**Figure 1D,E Supplementary Table 5**, **Supplementary Table 6, Supplementary Figure 5**). Five transcripts were associated with mtDNA SNPs (*MT-ND1, MT-ND3*, *MT-ND4, MT-CO3, MT-CYB*) at 5% study-wide FDR, in 29 out of the 41 tissues tested (**Figure 3D**, **Supplementary Tables 7,8**). Strikingly, the 14 fMet-associated mtDNA SNPs on Haplogroup Uk accounted for 93.75% of top eQTLs on mtDNA encoded genes at 5% study-wide FDR (30 out of 32, **Supplementary Table 7**), with associations with *MT-ND3* accounting for 87.5% (28 out of 32, **Figure 3E**) of the top eQTLs, and associations with genes encoding subunits in Complex I accounting for 93.3% (28 out of 30).

We next assessed the effects of fMet levels *in vitro* using trans-mitochondrial cytoplasmic hybrids (cybrids). Cybrids were generated by fusing a single donor cell line depleted of mtDNA [48] with cytoplasts from donors with mtDNA of haplogroups Uk and H respectively [10] (**Figure 4A, Online Methods**). We used four transmitochondrial cybrid cell lines generated from four different healthy donors for each mtDNA haplogroup. As all cybrid lines contain the same nuclear DNA so that any functional differences between them is due to differences in the mtDNA sequence. **(Supplementary Table 9)**. We found higher fMet levels in Uk cybrids than H cybrids (P = 0.02, two-tailed t-test), in line with our

population-level findings (**Figure 4B**). As shown in **Figure 4C and Extended Figure 1A**, fMet is synthesized by products of the one-carbon metabolism in mitochondria.

To determine the source of the different fMet levels between the different mtDNA haplogroups we first looked for evidence of increased synthesis. This included an analysis of one carbon metabolism proteins which supplement the metabolites necessary for fMet synthesis (**Extended Data Figure 1A-B**); methionyl-tRNA formyltransferase (MTFMT) (**Extended Data Figure 1A-D**); and mitochondrial ribosomal proteins (for example, MRPL19). We found no differences between haplogroup Uk and H cybrids (**Extended Data Figure 1C**). Next, we looked for differences in the degradation of fMet by studying levels of the peptide deformylase (PDF) (**Extended Data Figure 1C**) and its deformylation bioproduct formate (**Extended Data Figure 1D**). Again, we saw no differences between haplogroup Uk and H cybrids. Next, we explored whether the increased fMet in haplogroup Uk was caused by the accumulation of fMet due to decreased protein synthesis, by studying serum samples from patients with a rare genetic mitochondrial disorder caused by a mutation in the mtDNA-encoded gene for tRNA Leucine (m.3243A>G). Despite the known profound defect of intra-mitochondrial translation[49,50], we did not observe any difference in fMet levels between the patient serums and controls (**Extended Data Figure 1E**), indicating that blocking protein synthesis itself does not lead to an increase in fMet levels. This raises the possibility of other sources not known at present, including the release of fMet from unstable supercomplexes[51.]

Our findings raised the possibility that previously reported differences in intra-mitochondrial protein synthesis between the haplogroups H and Uk[10] could be accounted for by the observed differences in fMet. In previous work, increasing fMet in fibroblasts by over-expressing MTFMT also decreased intra-mitochondrial protein synthesis[52]. In keeping with this hypothesis, treating the cybrid lines with exogenous fMet at a similar concentration seen in plasma significantly increased the intracellular fMet levels ~1.19-fold (**Extended Data Figure 2A)**. This was similar to the fold-differences observed in INTERVAL between the individuals carrying mtDNA SNPs alleles from haplogroup Uk as compared to those carrying mtDNA SNP alleles from haplogroup H (fold change = 1.25, SE=0.05, P= $1.7x10^{-7}$, **Extended Data Figure 2B)**. This increase significantly decreased intra-mitochondrial protein synthesis in the cells of both haplogroups (P = 0.04, 2-way-ANOVA; **Figure 4D**).

This was accompanied by increased levels of *MT-CO3* transcripts (**Figure 4E)** (consistent with eQTL analysis in GTEx as shown in **Extended Data Figure 3**), which is the characteristic compensatory response to reduced intra-mitochondrial translation[49]. We therefore asked if altering fMet levels through downregulation of mitochondrial methionyl-tRNA formyltransferase (MTFMT) using siRNA (**Extended Data Figure 4A**) would result in a similar compensatory response. Downregulation of MTFMT reduces fMet and the synthesis of fMet-dependent proteins in complex I and IV [46,47], resulted in lower levels of the fMet-dependent protein MT-CO1 (**Extended Data Figure 4A**) and a parallel increase in transcript levels for the fMet-independent *MT-CO3* (**Extended Data Figure 4B**). Taken together, these findings indicate that different fMet levels associated with common mtDNA polymorphisms modulate intra-mitochondrial protein synthesis under homeostatic conditions (**Figure 4D**).

Next, we studied the downstream consequences of fMet on mitochondrial oxidative phosphorylation. In keeping with the observed decrease in mtDNA-encoded protein MT-CO1 in the cell lines from the haplogroup H (*P*= 0.016, 2-way-ANOVA; **Extended Data Figure 4D**), fMet supplementation reduced the abundance of the N-formylation dependent[53]

complexes I and IV (P = 0.02, 2-way-ANOVA; **Figure 4F, Extended Data Figure 4E&F**). This reduction was accompanied by lowered enzyme activities [54] ($P\leq$ 0.0005; **Figure 4G**), abolishing the differences between the haplogroups. We, however, did not observe this decrease in the mitochondrial complexes III and V, which are less dependent on fMet [51] (**Extended Data Figure 4G-H**). As expected, the effect of fMet supplementation on ETC complexes I and IV abundance was associated with a decrease in oxygen consumption in both haplogroup H and Uk cybrids ($P\leq$ 0.03, 2-way-ANOVA; **Figure 4H**), and consequent increase in both glycolytic ATP (P= 0.02; **Figure 4I**) and cytoplasmic ROS **(**$P\leq$ 0.0005; **Figure 4J**) as previously observed in mouse fibroblasts lacking MTFMT[51], the main enzyme required for fMet synthesis. fMet had no significant effect on mitochondrial ATP levels ($P>$ 0.05, **Extended Data Figure 4J**), mitochondrial mass, membrane potential or intra-mitochondrial reactive oxygen species (ROS) ($P>$ 0.05, **Extended Data Figure 4K**), reflecting the relative sparing of complex III and V[51]. Thus, the differences in intra-mitochondrial protein synthesis modulated by fMet have downstream effects on mitochondrial respiratory chain function, and transcription mediated through the N-formylation of specific intra-mitochondrial proteins. Our findings suggest that fMet levels are physiologically balanced in the haplogroup H and Uk cybrids due to additional factors that are poorly understood at present. Adding more fMet disrupted this equilibrium resulting in the down-stream consequences that we observed.

fMet regulates cellular stress and cytosolic proteostasis

Although N-formylation of methionine is important for intra-mitochondrial protein synthesis[47], its effects are less well documented in the cytosol, particularly in mammals. Interestingly, we observed that fMet supplementation at levels seen in plasma globally suppressed the *de novo* cytosolic translation (P= 0.0001, 3-way-ANOVA; **Figure 5A**). In accordance with this, we observed activation of the eukaryotic translation initiation factor 2A (EIF2A) the downstream kinase of the highly conserved integrated stress response (ISR) pathway that leads to global repression of protein translation[55] (**Figure 5B, Extended Data Figure 5A**). Consistently with the activation of EIF2A we also saw increased mRNA expression of its downstream target activation factor 4 (ATF4)[56] and the CCAAT-enhancer-binding protein homologous protein (CHOP)[57], enough to abolish the basal differences between the haplogroups ($P\leq$0.03, 2-way-ANOVA, **Figure 5C**). These observations were independent of the ATF5-mitochondrial chaperone dependent mitochondrial unfolded response[58] (**Extended Data Figure 5B**). Exogenous fMet also abolished the difference between haplogroups Uk and H seen under basal conditions, implicating fMet in modulating whole cell differences in protein homeostasis associated with the different mtDNA polymorphisms.

The EIF2A/ATF4 pathway mediates repression of protein translation through inhibition of the mammalian target of rapamycin (mTORC1) and thus growth[59]. However, we did not find any differences in mTORC1 activation (**Extended Data Figure 5A, C**) or growth (**Extended Data Figure 5D-E**), implying an alternative mechanism for fMet in cytosolic proteostasis. In *Saccharomyces cerevisiae*, the formylation of methionine in the cytoplasm, has been proposed as a new protein degradation mechanism (N-degron) under stress conditions, in a process dependent on the amino acid sensor GNC2 (general control nonderepressible 2)[60] and presumably its downstream target EIF2A[61]. In line with these observations, supplementation with fMet significantly increased the levels of ubiquitinated proteins in both H and Uk cybrids ($P\leq$ 0.0002, 2-way-ANOVA; **Figure 5D, Extended Data Figure 5F**). These are in keeping with fMet also being an N-degron in humans. Thus, fMet

modulates cytosolic protein homeostasis at multiple levels, including translation (**Figure 5A**) and degradation (**Figure 5D, Extended Data Figure 5F**).

Given our previous findings that fMet modulates mitochondrial and cytosolic protein homeostasis, we determined the potential downstream consequences of the mitochondrial background H and Uk by comparing the transcriptome of individuals with haplogroup Uk against those with haplogroup H across 49 tissues using again data from the GTEx Consortium[37]. This revealed 1 to 619 differentially expressed genes (total 4,244 genes) in 47 out of 49 tissues using a quasi-likelihood F test in edgeR [62,63] at 5% tissue-wide FDR (**Supplementary Tables 10,11**), with the majority (94.5%) of differentially expressed genes being more highly expressed in haplogroup Uk across all tissues (**Supplementary Table 10,11**). Pathway (**Figure 5E)** and gene set enrichment analysis (GSEA) (**Supplementary Tables 12,13**) of all the differentially expressed genes showed an enrichment for pathways involved in metabolism and immunity in addition to protein homeostasis pathways and ribosomal translation initiation (**Figure 5F**). In particular, we found three significant differential expression signals (at 5% tissue-wide FDR) among nuclear DNA encoded mitochondrial ribosomal genes and genes involved in the processing of mitochondrial rRNAs: *RMRP* (logFC=2.58, FDR=3.93x10-8) in coronary artery, and *MRPS6* (logFC=0.82, FDR=2.24x10-2) and MRPL14 (logFC=0.68, FDR=3.07x10-2) in the prostate (**Supplementary Figure 6**). Taken together, these independent observations validate our earlier *in vitro* findings, and indicate that mtDNA haplogroup-associated differences in fMet have multiple potential downstream consequences for cellular function beyond bioenergetics, oxidative phosphorylation, and mitochondrial ATP synthesis (**Extended Data Figure 4J-K)**.

fMet levels mediate late-onset disease risk

Haplogroup Uk has been previously associated with reduced risk of developing late-onset diseases including ischemic stroke (IS)[23] and Parkinson's disease (PD)[22], but the reasons for this are not known. To explore these potential mechanisms, we measured blood fMet levels in an IS cohort (282 cases, 181 controls, **Supplementary Table 14**) where we previously described a protective effect of haplogroup Uk[23]. In order to remove haplogroup effects on IS when testing for fMet associations with the disease, we enriched for individuals of haplogroup Uk in both cases and controls (32% haplogroup Uk in cases, 23% haplogroup Uk in controls). We then asked whether fMet is associated with IS, and if its effects can be due to haplogroup differences. We found a marginal and negative association between fMet and IS (OR= 0.83, SE= 0.08, logistic regression *P*= 0.06). This association is present in non-Uk individuals (OR= 0.77, SE= 0.08, logistic *P*= 0.02), but not in those with haplogroup Uk (OR= 2.24, SE= 1.06, logistic *P*= 0.09) (**Figure 6A**). Our results are consistent with a potential involvement of fMet in IS etiology, only part of which is due to mtDNA haplogroups. We also considered a PD cohort without enrichment for haplogroup Uk (120 cases, 43 controls, **Supplementary Table 14**), which however did not reveal any associations, possibly due to low sample sizes.

Next, we investigated whether fMet levels may be associated with the risk of other ageing-related diseases (cardiometabolic and common cancers, **Supplementary Table 15**), and if those associations were mtDNA haplogroup-dependent. We used Cox-proportional hazards models to test associations between fMet and incident risk of 24 non-communicable diseases and all-cause mortality in 11,966 individuals from the EPIC-Norfolk study covering more than 20-years of follow-up (**Online Methods**). We observed significant (P< 0.002; 0.05/23 tests) positive associations between fMet levels and incident renal disease, heart failure, coronary artery disease, abdominal aortic aneurysms, peripheral artery disease and

chronic obstructive pulmonary disease (COPD) as well as mortality (**Figure 6B**, **Supplementary Table 16**). Hazard ratios ranged between 1.10 and 1.29 per 1 standard deviation (SD) increase in log-transformed fMet levels. As fMet was correlated with age (Pearson correlation of age and fMet r= 0.33, P= 2.48x10^{-307}, **Figure 6C**), all Cox models accounted for age. To ensure the age association does not violate the proportional hazards assumption, we investigated Schoenefeld residuals and age-interaction terms, and none showed evidence of violation (**Supplementary Table 16**). Associations did not differ significantly between the Uk (N= 895) and other haplogroups (N= 9,887), however confidence intervals were wide in the smaller Uk haplogroup for some outcomes (**Supplementary Table 16, Figure 6D**). While this highlights the low power of the current study to detect heterogeneity between groups, the potential role of fMet as a marker of ageing-related diseases in different haplogroups warrants further investigations.

Discussion

In this study, we profiled more than 5,000 molecular traits in a healthy population-based cohort, and found novel associations between three mtDNA variants in Haplogroups Uk and H4 and the metabolite fMet. Whilst it is possible that the differences in fMet levels and mitochondrial transcription and translation are due to independent effects of haplogroup-specific variants rather than through a common causal pathway, it is not clear how this would occur, and results from our experiments are consistent with the latter. In fact, two of the variants associated with fMet (**Figure 2B**) affect the non-coding D-loop which is involved in regulation mtDNA transcription, and two also involve the rRNA genes directly involved in protein synthesis. Four variants affect the amino acid sequence of critical respiratory chain proteins, potentially influencing their function[10,64], assembly or stability[65], with the non-synonymous variants being associated indirectly through co-inheritance on the same mtDNA haplogroup. This complex scenario highlights the need of future experiments dissecting the effect of each independent variant.

fMet is the initiation amino acid for intra-mitochondrial translation[66]. Previous studies have shown that fMet is not necessary for initiation of translation or stability of newly synthesized polypeptides. However, a lack of fMet decreases synthesis of mtDNA encoded proteins and their integration into OXPHOS complexes and supercomplexes: MTFMT knockouts or mutants[47] display inefficient OXPHOS and increased risk of disease[47,51]. On the other hand, increasing fMet by MTFMT overexpression[52] and our experiments with fMet supplementation also compromised mitochondrial protein synthesis, OXPHOS complex levels and respiratory supercomplex function[51]. This implies that under physiological conditions, fMet is maintained within a narrow window, and increasing or decreasing fMet can have detrimental effects. Our findings demonstrate a role for fMet-associated mtDNA variants regulating levels of intra-mitochondrial gene expression, and modulating intra-mitochondrial protein synthesis and OXPHOS complex formation under homeostatic conditions through fMet. This may have tissue-specific and age-cumulative effects on metabolism and disease risk.

One of our most intriguing findings is that of a mechanism of cellular proteostasis that is modulated by mtDNA. Within mitochondria, mtDNA haplogroup-dependent fMet levels directly and specifically affect the abundance of mtDNA encoded, N-formylated ETC subunits and complexes. In the cytosol, fMet exerts indirect influence on global intra-mitochondrial protein synthesis through transcription and proteolysis without effects on cell growth (**Extended Data Figure 5D,E**) and/or mitochondrial membrane stability (**Extended**

Figure 4J,K); higher fMet levels in individuals from haplogroup Uk increase the ubiquitin-targeted N-degron mediated proteolysis (**Figure 5G**), and thus decrease the formation of protein aggregates[67] and the regulation of apoptosis[68]. This can explain the previously found protective effects of haplogroup Uk on late-onset neurodegenerative disorders[69]. In addition to the elimination of mis-folded proteins, N-end rule pathways also play a role in controlling subunit stoichiometries[70] in protein complexes such as the ETC, and the elimination of proteins mis-localised from their primary cellular compartment[71]. Furthermore, the profound reduction in cytosolic protein synthesis is likely to have multiple downstream effects on cell function. This emphasizes the importance of maintaining fMet levels within a narrow physiological range and the need of future studies dissecting its fluctuations in different tissues and disease models.

It is therefore plausible that fMet is involved in 'matching' protein synthesis with the mitochondrial and cytosolic compartments[72] in response to cellular bioenergetic needs in a tissue-specific manner. Subtle differences in fMet, partly attributable to mtDNA haplogroup effects, could have a cumulative effect on proteostasis and degradation throughout life, and thereby modify the risk of developing several late-onset diseases. fMet levels were significantly associated with late-onset diseases in the EPIC Norfolk cohort independent of age. Further experimental work is needed to definitively prove a causal role for fMet, however, given that it can be readily measured in serum as a circulating biomarker of cellular proteostasis, fMet is likely to be valuable for monitoring new treatments across a wide range of common human disorders.

In conclusion, the use of deep molecular phenotyping based on high-throughput metabolomics, transcriptomics, and proteomics is proving effective in identifying molecular hypotheses underpinning genetic associations with health and disease endpoints[40,41,73,74]. Our findings open up possibilities for further investigations into mtDNA control over metabolism and cellular physiology, and its implications on human health and disease. fMet may not be the only metabolite mtDNA variants regulated; most plasma biomarkers included in our study were only assayed in around 3,000 individuals, limiting statistical power. Continued investigation of mtDNA effects on biomarkers may lead to further elucidation of mitochondria's role in cellular physiology and function.

References

1. Anderson, S. *et al.* Sequence and organization of the human mitochondrial genome. *Nature* **290**, 457–465 (1981).
2. Giles, R. E., Blanc, H., Cann, H. M. & Wallace, D. C. Maternal inheritance of human mitochondrial DNA. *Proc. Natl. Acad. Sci. U. S. A.* **77**, 6715–6719 (1980).
3. Howell, N. Mutational analysis of the human mitochondrial genome branches into the realm of bacterial genetics. *American journal of human genetics* vol. 59 749–755 (1996).
4. Lippold, S. *et al.* Human paternal and maternal demographic histories: insights from high-resolution Y chromosome and mtDNA sequences. *Investig. Genet.* **5**, 13 (2014).
5. Stone, A. C. & Stoneking, M. mtDNA analysis of a prehistoric Oneota population: implications for the peopling of the New World. *Am. J. Hum. Genet.* **62**, 1153–1170 (1998).
6. Vigilant, L., Stoneking, M., Harpending, H., Hawkes, K. & Wilson, A. C. African populations and the evolution of human mitochondrial DNA. *Science* **253**, 1503–1507 (1991).
7. Ruiz-Pesini, E., Mishmar, D., Brandon, M., Procaccio, V. & Wallace, D. C. Effects of purifying and adaptive selection on regional variation in human mtDNA. *Science* **303**, 223–226 (2004).
8. Lane, N. & Martin, W. The energetics of genome complexity. *Nature* **467**, 929–934 (2010).
9. Ji, F. *et al.* Mitochondrial DNA variant associated with Leber hereditary optic neuropathy and high-altitude Tibetans. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 7391–7396 (2012).
10. Gómez-Durán, A. *et al.* Unmasking the causes of multifactorial disorders: OXPHOS differences between mitochondrial haplogroups. *Hum. Mol. Genet.* **19**, 3343–3353 (2010).
11. Suissa, S. *et al.* Ancient mtDNA genetic variants modulate mtDNA transcription and replication. *PLoS Genet.* **5**, e1000474 (2009).
12. Tranah, G. J. *et al.* Mitochondrial DNA variation in human metabolic rate and energy expenditure. *Mitochondrion* **11**, 855–861 (2011).
13. Gorman, G. S. *et al.* Prevalence of nuclear and mitochondrial DNA mutations related to adult mitochondrial disease. *Ann. Neurol.* **77**, 753–759 (2015).
14. Ruiz-Pesini, E. *et al.* An enhanced MITOMAP with a global mtDNA mutational phylogeny. *Nucleic Acids Res.* **35**, D823–8 (2007).
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. Zong, W.-X., Rabinowitz, J. D. & White, E. Mitochondria and Cancer. *Mol. Cell* **61**, 667–676 (2016).
17. Larman, T. C. *et al.* Spectrum of somatic mitochondrial mutations in five cancers. *Proc. Natl. Acad. Sci. U. S. A.* **109**, 14087–14091 (2012).
18. Yuan, Y. *et al.* Comprehensive molecular characterization of mitochondrial genomes in human cancers. *Nat. Genet.* **52**, 342–352 (2020).
19. Vander Heiden, M. G., Cantley, L. C. & Thompson, C. B. Understanding the Warburg effect: the metabolic requirements of cell proliferation. *Science* **324**, 1029–1033 (2009).
20. Marom, S., Friger, M. & Mishmar, D. MtDNA meta-analysis reveals both phenotype specificity and allele heterogeneity: a model for differential association. *Sci. Rep.* **7**, 43449 (2017).
21. Bi, R. *et al.* Mitochondrial DNA haplogroup B5 confers genetic susceptibility to Alzheimer's disease in Han Chinese. *Neurobiol. Aging* **36**, 1604.e7–16 (2015).

22. Hudson, G. *et al.* Two-stage association study and meta-analysis of mitochondrial DNA variants in Parkinson disease. *Neurology* **80**, 2042–2048 (2013).
23. Chinnery, P. F., Elliott, H. R., Syed, A., Rothwell, P. M. & Oxford Vascular Study. Mitochondrial DNA haplogroups and risk of transient ischaemic attack and ischaemic stroke: a genetic association study. *Lancet Neurol.* **9**, 498–503 (2010).
24. Nishigaki, Y. *et al.* Mitochondrial haplogroup N9b is protective against myocardial infarction in Japanese males. *Hum. Genet.* **120**, 827–836 (2007).
25. Kofler, B. *et al.* Mitochondrial DNA haplogroup T is associated with coronary artery disease and diabetic retinopathy: a case control study. *BMC Medical Genetics* vol. 10 (2009).
26. Chinnery, P. F. *et al.* Mitochondrial DNA haplogroups and type 2 diabetes: a study of 897 cases and 1010 controls. *J. Med. Genet.* **44**, e80 (2007).
27. Poulton, J. *et al.* Type 2 diabetes is associated with a common mitochondrial variant: evidence from a population-based case-control study. *Hum. Mol. Genet.* **11**, 1581–1583 (2002).
28. Hudson, G., Gomez-Duran, A., Wilson, I. J. & Chinnery, P. F. Recent mitochondrial DNA mutations increase the risk of developing common late-onset human diseases. *PLoS Genet.* **10**, e1004369 (2014).
29. Marom, S., Friger, M. & Mishmar, D. MtDNA meta-analysis reveals both phenotype specificity and allele heterogeneity: a model for differential association. *Sci. Rep.* **7**, 43449 (2017).
30. Wallace, D. C. A mitochondrial paradigm of metabolic and degenerative diseases, aging, and cancer: a dawn for evolutionary medicine. *Annu. Rev. Genet.* **39**, 359–407 (2005).
31. Chinnery, P. F. & Gomez-Duran, A. Oldies but Goldies mtDNA Population Variants and Neurodegenerative Diseases. *Frontiers in Neuroscience* vol. 12 (2018).
32. Trounce, I., Neill, S. & Wallace, D. C. Cytoplasmic transfer of the mtDNA nt 8993 T-->G (ATP6) point mutation associated with Leigh syndrome into mtDNA-less cells demonstrates cosegregation with a decrease in state III respiration and ADP/O ratio. *Proceedings of the National Academy of Sciences* vol. 91 8334–8338 (1994).
33. Mattiazzi, M. The mtDNA T8993G (NARP) mutation results in an impairment of oxidative phosphorylation that can be improved by antioxidants. *Human Molecular Genetics* vol. 13 869–879 (2004).
34. Goto, Y., Nonaka, I. & Horai, S. A mutation in the tRNA(Leu)(UUR) gene associated with the MELAS subgroup of mitochondrial encephalomyopathies. *Nature* **348**, 651–653 (1990).
35. van den Ouweland, J. M. *et al.* Maternally inherited diabetes and deafness is a distinct subtype of diabetes and associates with a single point mutation in the mitochondrial tRNA(Leu(UUR)) gene. *Diabetes* **43**, 746–751 (1994).
36. Brown, M. D., Trounce, I. A., Jun, A. S., Allen, J. C. & Wallace, D. C. Functional analysis of lymphoblast and cybrid mitochondria containing the 3460, 11778, or 14484 Leber's hereditary optic neuropathy mitochondrial DNA mutation. *J. Biol. Chem.* **275**, 39831–39836 (2000).
37. GTEx Consortium *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
38. Moore, C. *et al.* The INTERVAL trial to determine whether intervals between blood donations can be safely and acceptably decreased to optimise blood supply: study protocol for a randomised controlled trial. *Trials* **15**, 363 (2014).
39. Poplin, R. *et al.* Scaling accurate genetic variant discovery to tens of thousands of samples. doi:10.1101/201178.
40. Sun, B. B. *et al.* Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79

(2018).
41. Shin, S.-Y. *et al.* An atlas of genetic influences on human blood metabolites. *Nat. Genet.* **46**, 543–550 (2014).
42. Speed, D. *et al.* Reevaluation of SNP heritability in complex human traits. *Nat. Genet.* **49**, 986–992 (2017).
43. Aschard, H. *et al.* Covariate selection for association screening in multiphenotype genetic studies. *Nat. Genet.* **49**, 1789–1795 (2017).
44. Derenko, M. *et al.* Western Eurasian ancestry in modern Siberians based on mitogenomic data. *BMC Evol. Biol.* **14**, 217 (2014).
45. Day, N. *et al.* EPIC-Norfolk: study design and characteristics of the cohort. European Prospective Investigation of Cancer. *Br. J. Cancer* **80 Suppl 1**, 95–103 (1999).
46. Rajbhandary, U. L. & Ming Chow, C. Initiator tRNAs and Initiation of Protein Synthesis. *tRNA* 511–528 doi:10.1128/9781555818333.ch25.
47. Tucker, E. J. *et al.* Mutations in MTFMT Underlie a Human Disorder of Formylation Causing Impaired Mitochondrial Translation. *Cell Metabolism* vol. 14 428–434 (2011).
48. King, M. P. & Attardi, G. Human cells lacking mtDNA: repopulation with exogenous mitochondria by complementation. *Science* **246**, 500–503 (1989).
49. Picard, M. *et al.* Progressive increase in mtDNA 3243A>G heteroplasmy causes abrupt transcriptional reprogramming. *Proc. Natl. Acad. Sci. U. S. A.* **111**, E4033–42 (2014).
50. King, M. P., Koga, Y., Davidson, M. & Schon, E. A. Defects in mitochondrial protein synthesis and respiratory chain activity segregate with the tRNA(Leu(UUR)) mutation associated with mitochondrial myopathy, encephalopathy, lactic acidosis, and strokelike episodes. *Mol. Cell. Biol.* **12**, 480–490 (1992).
51. Arguello, T., Köhrer, C., RajBhandary, U. L. & Moraes, C. T. Mitochondrial methionyl - formylation affects steady-state levels of oxidative phosphorylation complexes and their organization into supercomplexes. *J. Biol. Chem.* **293**, 15021–15032 (2018).
52. Hinttala, R. *et al.* An N-terminal formyl methionine on COX 1 is required for the assembly of cytochrome c oxidase. *Hum. Mol. Genet.* **24**, 4103–4113 (2015).
53. Neeve, V. C. M. *et al.* Clinical and functional characterisation of the combined respiratory chain defect in two sisters due to autosomal recessive mutations in MTFMT. *Mitochondrion* **13**, 743–748 (2013).
54. Nijtmans, L. G. J., Henderson, N. S. & Holt, I. J. Blue Native electrophoresis to study mitochondrial and other protein complexes. *Methods* **26**, 327–334 (2002).
55. Wek, R. C., Jiang, H.-Y. & Anthony, T. G. Coping with stress: eIF2 kinases and translational control. *Biochem. Soc. Trans.* **34**, 7–11 (2006).
56. Ameri, K. & Harris, A. L. Activating transcription factor 4. *Int. J. Biochem. Cell Biol.* **40**, 14–21 (2008).
57. Su, N. & Kilberg, M. S. C/EBP homology protein (CHOP) interacts with activating transcription factor 4 (ATF4) and negatively regulates the stress-dependent induction of the asparagine synthetase gene. *J. Biol. Chem.* **283**, 35106–35117 (2008).
58. Quirós, P. M. *et al.* Multi-omics analysis identifies ATF4 as a key regulator of the mitochondrial stress response in mammals. *J. Cell Biol.* **216**, 2027–2045 (2017).
59. Walter, P. & Ron, D. The unfolded protein response: from stress pathway to homeostatic regulation. *Science* **334**, 1081–1086 (2011).
60. Kim, J.-M. *et al.* Formyl-methionine as an N-degron of a eukaryotic N-end rule pathway. *Science* vol. 362 eaat0174 (2018).
61. Eldeeb, M. A., Fahlman, R. P., Esmaili, M. & Fon, E. A. Formylation of Eukaryotic Cytoplasmic Proteins: Linking Stress to Degradation. *Trends in biochemical sciences* vol. 44 181–183 (2019).
62. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for

differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
63. McCarthy, D. J., Chen, Y. & Smyth, G. K. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Res.* **40**, 4288–4297 (2012).
64. Pacheu-Grau, D., Gómez-Durán, A., López-Pérez, M. J., Montoya, J. & Ruiz-Pesini, E. Mitochondrial pharmacogenomics: barcode for antibiotic therapy. *Drug Discov. Today* **15**, 33–39 (2010).
65. Pello, R. *et al.* Mitochondrial DNA background modulates the assembly kinetics of OXPHOS complexes in a cellular model of mitochondrial disease. *Hum. Mol. Genet.* **17**, 4001–4011 (2008).
66. Bianchetti, R., Lucchini, G., Crosti, P. & Tortora, P. Dependence of mitochondrial protein synthesis initiation on formylation of the initiator methionyl-tRNAf. *J. Biol. Chem.* **252**, 2519–2523 (1977).
67. Sorrentino, V. *et al.* Enhancing mitochondrial proteostasis reduces amyloid-β proteotoxicity. *Nature* **552**, 187–193 (2017).
68. Eldeeb, M. A., Fahlman, R. P., Esmaili, M. & Ragheb, M. A. Regulating Apoptosis by Degradation: The N-End Rule-Mediated Regulation of Apoptotic Proteolytic Fragments in Mammalian Cells. *Int. J. Mol. Sci.* **19**, (2018).
69. Hudson, G., Gomez-Duran, A., Wilson, I. J. & Chinnery, P. F. Recent mitochondrial DNA mutations increase the risk of developing common late-onset human diseases. *PLoS Genet.* **10**, e1004369 (2014).
70. Shemorry, A., Hwang, C.-S. & Varshavsky, A. Control of protein quality and stoichiometries by N-terminal acetylation and the N-end rule pathway. *Mol. Cell* **50**, 540–551 (2013).
71. Wrobel, L. *et al.* Mistargeted mitochondrial proteins activate a proteostatic response in the cytosol. *Nature* **524**, 485–488 (2015).
72. Couvillion, M. T., Soto, I. C., Shipkovenska, G. & Churchman, L. S. Synchronized mitochondrial and cytosolic translation programs. *Nature* **533**, 499–503 (2016).
73. Suhre, K. *et al.* Connecting genetic risk to disease end points through the human blood plasma proteome. *Nat. Commun.* **8**, 14357 (2017).
74. Yao, C. *et al.* Genome-wide mapping of plasma protein QTLs identifies putatively causal genes and pathways for cardiovascular disease. *Nat. Commun.* **9**, 3268 (2018).
75. Astle, W. J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* **167**, 1415–1429.e19 (2016).
76. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
77. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
78. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
79. Weissensteiner, H. *et al.* mtDNA-Server: next-generation sequencing data analysis of human mitochondrial DNA in the cloud. *Nucleic Acids Res.* **44**, W64–9 (2016).
80. Van der Auwera, G. A. *et al.* From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr. Protoc. Bioinformatics* **43**, 11.10.1–11.10.33 (2013).
81. 1000 Genomes Project Consortium *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
82. Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* **26**, 2867–2873 (2010).

83. Kiełbasa, S. M., Wan, R., Sato, K., Horton, P. & Frith, M. C. Adaptive seeds tame genomic sequence comparison. *Genome Res.* **21**, 487–493 (2011).
84. Stegle, O., Parts, L., Piipari, M., Winn, J. & Durbin, R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protoc.* **7**, 500–507 (2012).
85. Lippert, C., Casale, F. P., Rakitsch, B. & Stegle, O. LIMIX: genetic analysis of multiple traits. *Genetics* 833 (2014).
86. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 15545–15550 (2005).
87. Mootha, V. K. *et al.* PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nat. Genet.* **34**, 267–273 (2003).
88. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
89. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
90. Koboldt, D. C. *et al.* VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**, 2283–2285 (2009).
91. Koboldt, D. C. *et al.* VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Res.* **22**, 568–576 (2012).
92. Wilm, A. *et al.* LoFreq: a sequence-quality aware, ultra-sensitive variant caller for uncovering cell-population heterogeneity from high-throughput sequencing datasets. *Nucleic Acids Res.* **40**, 11189–11201 (2012).
93. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
94. Torroni, A. *et al.* Classification of European mtDNAs from an analysis of three European populations. *Genetics* **144**, 1835–1850 (1996).
95. van Oven, M. & Kayser, M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.* **30**, E386–94 (2009).
96. Weissensteiner, H. *et al.* HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Res.* **44**, W58–W63 (2016).
97. Pereira, L., Soares, P., Radivojac, P., Li, B. & Samuels, D. C. Comparing phylogeny and the predicted pathogenicity of protein variations reveals equal purifying selection across the global human mtDNA diversity. *Am. J. Hum. Genet.* **88**, 433–439 (2011).
98. Levin, L., Zhidkov, I., Gurman, Y., Hawlena, H. & Mishmar, D. Functional recurrent mutations in the human mitochondrial phylogeny: dual roles in evolution and disease. *Genome Biol. Evol.* **5**, 876–890 (2013).
99. Bustin, S. A. *et al.* The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin. Chem.* **55**, 611–622 (2009).
100. Bradford, M. M. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.* **72**, 248–254 (1976).
101. Wittig, I., Braun, H.-P. & Schägger, H. Blue native PAGE. *Nat. Protoc.* **1**, 418–428 (2006).
102. Yarnall, A. J. *et al.* Characterizing mild cognitive impairment in incident Parkinson disease: the ICICLE-PD study. *Neurology* **82**, 308–316 (2014).
103. Gibb, W. R. & Lees, A. J. The relevance of the Lewy body to the pathogenesis of idiopathic Parkinson's disease. *J. Neurol. Neurosurg. Psychiatry* **51**, 745–752 (1988).
104. Rothwell, P. M. *et al.* Population-based study of event-rate, incidence, case fatality, and

mortality for all acute vascular events in all arterial territories (Oxford Vascular Study). *Lancet* **366**, 1773–1783 (2005).
105. Floros, V. I. *et al.* Segregation of mitochondrial DNA heteroplasmy through a developmental genetic bottleneck in human embryos. *Nat. Cell Biol.* **20**, 144–151 (2018).
106. Calabrese, C. *et al.* MToolBox: a highly automated pipeline for heteroplasmy annotation and prioritization analysis of human mitochondrial variants in high-throughput sequencing. *Bioinformatics* **30**, 3115–3117 (2014).
107. Xu, H. *et al.* FastUniq: a fast de novo duplicates removal tool for paired short reads. *PLoS One* **7**, e52249 (2012).

Acknowledgments:

Participants in the INTERVAL randomised controlled trial were recruited with the active collaboration of NHS Blood and Transplant England (www.nhsbt.nhs.uk), which has supported field work and other elements of the trial. DNA extraction and genotyping was co-funded by the National Institute for Health Research (NIHR), the NIHR BioResource (http://bioresource.nihr.ac.uk) and the NIHR Cambridge Biomedical Research Centre (BRC-1215-20014)*.

The academic coordinating centre for INTERVAL was supported by core funding from: NIHR Blood and Transplant Research Unit in Donor Health and Genomics (NIHR BTRU-2014-10024), UK Medical Research Council (MR/L003120/1), British Heart Foundation (SP/09/002; RG/13/13/30194; RG/18/13/33946) and the NIHR Cambridge BRC (BRC-1215-20014)*. A complete list of the investigators and contributors to the INTERVAL trial is provided in Di Angelantonio E, et al. Lancet. 2017 Nov 25;390(10110):2360-2371. The academic coordinating centre would like to thank blood donor centre staff and blood donors for participating in the INTERVAL trial.

This work was supported by Health Data Research UK, which is funded by the UK Medical Research Council, Engineering and Physical Sciences Research Council, Economic and Social Research Council, Department of Health and Social Care (England), Chief Scientist Office of the Scottish Government Health and Social Care Directorates, Health and Social Care Research and Development Division (Welsh Government), Public Health Agency (Northern Ireland), British Heart Foundation and Wellcome.

NC is supported by the EBI-Sanger Postdoctoral Fellowship. AGD receives funding from the National Institute for Health Research (NIHR) Biomedical Research Centre based at Cambridge University Hospitals NHS Foundation Trust. EY-D was funded by the Isaac Newton Trust/Wellcome Trust ISSF / University of Cambridge Joint Research Grants Scheme. JMMH was funded by the NIHR Cambridge BRC (BRC-1215-20014)*. PFC is a Wellcome Trust Principal Research Fellow (212219/Z/18/Z), and a UK NIHR Senior Investigator, who receives support from the Medical Research Council Mitochondrial Biology Unit (MC_UU_00015/9), the Evelyn Trust, and the National Institute for Health Research (NIHR) Biomedical Research Centre based at Cambridge University Hospitals NHS Foundation Trust and the University of Cambridge. CHWG is supported by a RCUK/UKRI Research Innovation Fellowship awarded by the Medical Research Council (MR/R007446/1) and by the Cambridge Centre for Parkinson-Plus. NS is supported by the Wellcome Trust, the BHF and the NIHR. JD holds a British Heart Foundation Professorship and a NIHR Senior Investigator Award*. RAL is supported by grants from Parkinson's UK. Sequencing of the INTERVAL samples was supported by the Wellcome Trust grant number 206194. Metabolon Metabolomics assays were funded by the NIHR BioResource, the Wellcome Trust (grant number 206194) and the NIHR Cambridge BRC (BRC-1215-20014)*. Nightingale Health NMR assays were funded by the European Commission Framework Programme 7 (HEALTH-F2-2012-279233). SomaLogic assays were funded by Merck and the NIHR Cambridge BRC (BRC-1215-20014)*. ICICLE-PD was funded by Parkinson's UK (J-0802, G-1301, G-1507), the Lockhart Parkinson's Disease Research Fund, and supported by the National Institute for Health Research (NIHR) Newcastle Biomedical Research Unit and the National Institute for Health Research (NIHR) Cambridge Biomedical Research Centre (146281). The EPIC-Norfolk study (https://doi.org/10.22025/2019.10.105.00004) has received funding from the Medical Research Council (MR/N003284/1 and MC-UU_12015/1) and Cancer Research UK (C864/A14136). The genetics work in the EPIC-Norfolk study was funded by the Medical

Research Council (MC_PC_13048). Metabolite measurements in the EPIC-Norfolk study were supported by the MRC Cambridge Initiative in Metabolic Science (MR/L00002/1) and the Innovative Medicines Initiative Joint Undertaking under EMIF grant agreement no. 115372. We are grateful to all the participants who have been part of the project and to the many members of the study teams at the University of Cambridge who have enabled this research. The collection of samples and assay work done on the Oxford Vascular Study cohort was funded by the National Institute for Health Research (NIHR) Biomedical Research Centre, Oxford, and the Wellcome Trust.

*The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

The authors thank all members of the ICICLE-PD Consortium for their support of this work. A full list of members of the ICICLE-PD Consortium is included in the Supplementary Materials.

Author contributions: NC, AGD, OS, NS and PFC conceived the study; NC, AGD, OS, NS and PFC designed the analyses and experiments; NC performed the quantitative genetics analyses; AGD performed the wet lab experiments; EYD and JMMH obtained replication data; AGD, AIB, LL and ZG performed the wet lab experiments on cell lines and patient cohorts; NC and CC performed the analysis on the patient cohorts; IDS and MP performed the analysis on the prospective cohort; NC, AGD, OS, NS and PFC interpreted the results; NS, PFC, OS, JMMH, JD, PMR, AIB, MC, RAL, CWG, NJW and CL acquired the resources and datasets; NC, KK, MJB, and KW curated the data and wrote the bioinformatics methods; NS, PFC, JMMH and OS supervised the project; NS and PFC acquired the funding and administered the project; NC, AGD, OS, NS, and PFC wrote the paper with input from all authors, who approved the final version of the manuscript. NC and AGD contributed equally as first authors; EYD and KK contributed equally as second authors; JMMH, OS, NS and PFC jointly supervised the work.

Competing interests: JMMH and EKYD are full time employees of Novo Nordisk Ltd. ASB has received grants outside of this work from AstraZeneca, Biogen, BioMarin, Bioverativ, Merck and Sanofi and personal fees from Novartis outside of this work. JD sits on the International Cardiovascular and Metabolic Advisory Board for Novartis (since 2010); the Steering Committee of UK Biobank (since 2011); the MRC International Advisory Group (ING) member, London (since 2013); the MRC High Throughput Science 'Omics Panel Member, London (since 2013); the Scientific Advisory Committee for Sanofi (since 2013); the International Cardiovascular and Metabolism Research and Development Portfolio Committee for Novartis; and the AstraZeneca Genomics Advisory Board (2018).

Figure Legends

Figure 1: Overview of analysis and population structure on the nuclear and mitochondrial genomes

A. Overview of analyses. A flowchat overview of the analyses we performed in this study, summarizing the number of samples and phenotypes used in each cohort we analysed, and the number of mtDNA associations with metabolites and cis-eQTLs we found. **B.** A simplified mtDNA haplogroup tree with haplogroups present in INTERVAL and GTEx participants individually coloured. Haplogroups not present in INTERVAL and GTEx participants are coloured grey. Each haplogroup is consistently represented by the same colours throughout this manuscript. **C.** A plot of principal component (PC) 1 and 2 from a principal component analysis (PCA) performed using 187 mtDNA SNPs (MAF >=1%) in 16,220 participants in INTERVAL, coloured by haplogroups identified for each individual using Haplogrep v2. Haplogroups U, K (Uk), and H are labelled. **D.** A plot of PC1 and PC2 from a PCA performed using 5,511,276 nDNA SNPs (MAF >=5%) in the same participants in INTERVAL, coloured by their haplogroups, as previously described. **E.** A plot of PC1 and PC2 from a PCA performed using 215 common mtDNA SNPs (MAF >=1%) in 456 participants in GTEx, coloured by haplogroups identified for each individual using Haplogrep v2. Haplogroups U, K (Uk), and H are labelled. **F.** A plot of PC1 and PC2 from a PCA performed using 5,451,305 common nucDNA SNPs (MAF >=5%) in the same participants in GTEx, coloured by their haplogroups

Figure 2. Metabolites and their associations with mtDNA SNPs in INTERVAL.

A. Manhattan plot summarizing results of association between 183 common mtDNA variants (MAF≥ 5%) and 896 metabolites. Each dot corresponds to the association between a mtDNA variant and a metabolite. Its x coordinate represents its position along the mitochondrial genome and its y coordinate represents the -log10 (p value) for the association from a Wald Test between LMMs with and without genotypes at a mtDNA SNP as a predictor for each metabolite, implemented in LDAK v5. The red dotted horizontal line at P=1.04x10^{-6} represents the significance threshold upon correcting for 896 metabolites and an estimate of 53.56 independent mtDNA SNPs (**Supplementary Discussion**). Red dots are associations where fMet is the metabolite tested, labelled with their position and genotype. The x axis is annotated with ranges of positions in the mtDNA with function, including the D-LOOP, the mtDNA encoded rRNAs (MT-rRNA), the mtDNA encoded tRNAs (MT-tRNA) and the mtDNA encoded protein-coding genes. **B.** Table of association statistics between 15 mtDNA SNPs associated with fMet in the discovery cohort INTERVAL and replication cohort EPIC-Norfolk; for each mtDNA SNP we show its position (BP), the gene it is in (GENE), the allele whose effects we test (A1), the other allele (A0), the frequency of the tested allele (A1FREQ) and functional annotations of the variant (ANNOT), the standardised effect size (BETA) of its association with fMet, its standard error (SE) and P values (P). Association statistics from both Wald tests on LMM and linear regressions in CMS are shown for the discovery cohort, while association statistics from linear regression are shown for the replication cohort. **C.** Pearson correlation coefficient r^2 between the mtDNA SNPs significantly associated with fMet; the red, light blue and dark blue squares denote the most significantly associated variants at each of the three independent signals. **D.** The haplogroup lineage tree for super-haplogroup Uk (on the left) and H (on the right). SNPs that are part of this tree and significantly associated with fMet are in bold and coloured according to their haplogroups. All except one SNP (mt.3992C>T) associated with fMet are on the branch for the super-haplogroup Uk.

Figure 3. fMet-associated genes regulate mtDNA gene expression.

A. Manhattan plot of association between common nDNA SNPs (MAF≥ 5%) with fMet; the x coordinate represents positions for each nDNA SNP tested; the y coordinate represents the -log10(P) for the Wald Test associations for nDNA SNP effects on fMet levels; red horizontal line indicates the significance threshold of $P= 5x10^{-8}$; red dots represent SNPs with significant associations, orange SNPs represent SNPs with $P< 5x10^{-7}$. **B.** Boxplot of fMet levels in INTERVAL participants with each genotype at rs550045 and mt.1811. Centre of the boxplots show median fMet levels, upper and lower limits of boxplots show interquartile ranges, while the whiskers show values within 1.5 times the interquartile range. Outliers show values beyond 1.5 times the interquartile range. **C**. The variance decomposition model for quantifying the variance in fMet levels explained by variation in nDNA SNPs, mtDNA SNPs, other metabolites and blood cell counts, and those that cannot be accounted for by all the above. The pie chart shows the relative contribution from all four components. **D.** A Manhattan plot of association between 13 protein-coding mtDNA genes and their top eQTL on the mtDNA for 41 GTEx tissues. X axis represents the position of the SNP along the mtDNA and Y axis represents the -log10(P) from log-likelihood ratio (LRT) tests for mtDNA SNP effects on expression of mtDNA encoded genes. Dots are coloured by genes, and the x axis is annotated with ranges of positions in the mtDNA with function. **E.** This figure shows the standardized effect size (BETA) of the mtDNA SNPs on log10(TPM+1) expression levels of mtDNA encoded *MT-ND3* in 41 GTEx tissues. Colour of the boxes corresponds to the tissue type; centre of the boxplots show median BETA values from all mtDNA SNPs; upper and lower limits show the interquartile range, while the whiskers show values within 1.5 times the interquartile range. Outliers show values beyond 1.5 times the interquartile range. The upper panel shows the results for the 15 fMet-associated SNPs, while the bottom panel shows the results for the non-fMet associated mtDNA SNPs.

Figure 4: fMet regulates mitochondrial protein synthesis and oxidative phosphorylation function.

A. Schematic representation of transmitochondrial cybrids. Black, red and blue dots represent the absence of mtDNA, haplogroup H and haplogroup Uk respectively. **B.** Quantification of fMet levels in cybrids of different haplogroups. Statistical testing was performed by unpaired t-test. Normality was assessed using the Kolmogorov–Smirnov test. The average raw value of fMet is 3.06 pg/mg. **C.** Schematic representation of mitochondrial protein synthesis and fMet. **D.** Effect of fMet on mitochondrial translation. Electrophoretic patterns of the synthesized mitochondrial products and fragments of the gel stained with Coomassie (used as a loading control); molecular weight marker (left) and each mitochondrial protein (right) are shown. Quantification of the bands was corrected by the loading control in each cell line. **E.** Effect of fMet on the levels of the mitochondrial transcript MT-CO3. **F.** One-dimensional Blue Native Gel Electrophoresis (1D-BNGE) and Western blot quantification (see **Extended Data Figure 3F)**; values are corrected with relative levels between loading control (CII) in each cell line and untreated samples from haplogroup H. **G.** Complex I (left gel) and IV (right gel) *in gel* activity assays (IGA) after 1D-BNGE analysis of digitonin treated cybrid cell lines with and without fMet treatment. Super-complexes (SC) composition is indicated. **H.** Basal respiration. **I.** Glycolytic ATP levels. The average raw value of ATP of all the cybrids is 19367878,79 luminescence units /mg of protein. **J.** Cytoplasmic ROS levels. The average raw value of ROS of all the cybrids is 17434,4/20000 cells fluorescence units. In all the graphs (**B-I**) bars/lines represent the mean ± SD of the biological replicates (n=4) of – (Control) and + (fMet treated) cell lines of each haplogroup that were measured in 3-5 independent technical replicates each. Colors red and blue represent haplogroup H and Uk respectively. The values are represented as relative to the average of untreated samples from

haplogroup H, unless otherwise indicated. Statistical testing was performed by using a 2-way-ANOVA test followed by Holm-Sidak's multiple comparison test unless stated otherwise. *P*-values corrected for multiple comparisons are indicated. Unprocessed S35 Blots and loadings can be found in **Source Data Figure 4**.

Figure 5. fMet modulates cytosolic protein homeostasis.

A. Effect of fMet on cytosolic translation products. Electrophoretic patterns of the synthesized proteins and fragments of the gel stained with Coomassie (loading control) and molecular weight marker (left) are shown. Quantification of the bands was corrected by the loading control in each cell line. **B.** Effect of fMet and mitochondrial haplogroup on EIF2A activation (**Extended Data Figure 4A**). Quantification of the immune detected bands for p.EIF2A^{Ser51}and EIF2A corrected by loading control (B-actin) in each cell line. Activation of EIF2A is calculated as ratio p.EIF2A^{Ser51}/ EIF2A. Values are represented as relative to the average of untreated samples from haplogroup H. **C.** Effect of fMet and haplogroup on the expression of EIF2A downstream targets ATF4 and CHOP. Box plots represent minimum, maximum, sample median, and the first and third quartiles. All data points are plotted. **D**. Effect of fMet on protein ubiquitination. Immunoblot detection with anti-ubiquitin and anti-B-actin as a loading control for untreated (-/-), proteasome inhibition with MG132 (+, -) and proteasome inhibition plus fMet (+, +). The quantification of the bands for ubiquitin smear was corrected by loading control (B-actin). Statistical testing was performed with a 3-way-ANOVA followed by Holm-Sidak's multiple comparisons. **E.** Consensus Pathway Analysis of all the differentially expressed genes between haplogroup H and Uk in all tissues in GTEx. Grey represents Reactome pathways and green wikipathways. **F.** Gene Set Enrichment Analysis (GSEA) analysis of all the differentially expressed genes between haplogroup H and Uk in all tissues in GTEx. Enrichment score is shown. **G.** Schematic representations of the effects of fMet. In all the plots (**B-D**) bars/lines represent the mean ± SD of the biological replicates (n= 4) of – (Control) and + (fMet treated) cell lines that were performed in 3-5 independent technical replicates each. Colors red and blue represent haplogroup H and Uk respectively. The values are represented as relative to the average of untreated samples from haplogroup H, unless indicated. Statistical testing was performed with a 2-way-ANOVA followed by Holm-Sidak's multiple comparisons, unless otherwise stated. *P*-values corrected for multiple comparisons are shown. Unprocessed blots and loadings can be found in **Source Data Figure 5**.

Figure 6: fMet as a biomarker for IS and other late-onset disorders.

A. Rank normalised residuals levels of fMet in IS and controls, separated by their mtDNA haplogroups, after regressing out batch for quantification of fMet and site of data collection as covariates; *P*-values are from logistic regression of IS disease status with haplogroup. The centre of the boxplots show the median normalised fMet levels; upper and lower limits of boxplots show the interquartile range, while the whiskers show values within 1.5 times the interquartile range. Outliers beyond whiskers show values beyond 1.5 times the interquartile range. All data points are plotted. **B.** Hazard ratio of fMet levels measured at baseline for 24 health outcomes, including mortality, over a 20-year follow-up. Points and error bars shown in blue represent the point estimates and 95% confidence intervals of the hazard ratio in 11,966 EPIC-Norfolk participants. **C.** Relationship between Z scores of fMet measured at baseline of a 20-year longitudinal study with 11,966 individuals, and their ages at baseline (mean 60 years, SD 6 years). Significant Spearman correlations were found between fMet levels and age at measurement in participants of both mtDNA haplogroup Uk and other haplogroups (Spearman correlation $P = 3.51 \times 10^{-25}$ and 1.49×10^{-284} respectively). **D.** Hazard ratio of fMet levels measured at baseline for 24 health outcomes, including mortality, over a

946 20-year follow-up. Points and error bars shown in blue represent the point estimates and 95%
947 confidence intervals of the hazard ratio in participants of mtDNA haplogroup Uk, while those
948 in red represent those in participants of other haplogroups.
949
950

Online Methods

Sample filtering in INTERVAL cohort

In the INTERVAL dataset[38], 12,395 participants were sequenced across the whole genome (WGS, mean coverage = 26.8x, SD = 3.1x; mtDNA mean coverage = 2022.6x, SD = 566.5x, **Supplementary Figure 1A**), and 4,502 participants were sequenced across the whole exome (WES, mean coverage = 48.0x, SD = 8.6x; mtDNA mean coverage = 30.6x, SD = 13.5x, **Supplementary Figure 1B)**, including 60 participants on whom WGS was also performed. Of the 12,395 participants with WGS in the INTERVAL cohort, we identified 32 participants who were sequenced in duplicate. All 32 participants had their blood samples taken and sequenced at two time points, so we removed one sample per participant sequenced at the later of the two time points. Second, 12,112 out of the remaining 12,363 participants with WGS and 4,471 out of 4,502 participants with WES can be linked to both phenotypic data and genotypes assayed with the UK Biobank Affymetrix Axiom array[75,76]. We retain only these participants for further analysis. Third, we compared genotype calls from WES against those from WGS in 56 overlapping participants at 307 overlapping polymorphic mtDNA SNPs between WES and WGS. One individual showed a particularly high rate of discordance between WES and WGS genotype calls (14 sites discordant), and was hence removed from further analysis. In the remaining 55 samples with both WGS and WES, we obtained a mean per variant Pearson r2 of 0.994 (SD = 0.039) between genotypes at all 307 mtDNA SNPs called in both WGS and WES, and a mean per sample Pearson r2 of 0.980 (SD = 0.134). We retained all 55 participants for further evaluation of the genotype qualities of variants called from WGS and WES, giving us 12,111 participants with WGS and 4,470 participants with WES (55 of whom also have WGS).

Genotype quality control in INTERVAL cohort

For mtDNA, we extracted reads mapping to the rCRS mitochondrial reference genome (NC_012920) from WGS in 10,704 individuals and WES in 4,502 individuals in INTERVAL, and called mtDNA variants using GATK v4[39], obtaining 396 high quality variants, of which 187 have MAF >= 1%. We use the 187 variants for assigning Haplogroups to each individual in INTERVAL using Haplogrep v2 and all further analyses. For nDNA, we obtained imputed genotypes from array genotypes at 5,511,276 autosomal, biallelic SNPs in 43,059 unrelated individuals from European descent from the INTERVAL project[75], filtering raw imputation results with information score (INFO > 0.9), minor allele frequency (MAF > 5%), P-value of violation of Hardy Weinberg Equilibrium (HWE > 10-6), and missingness (< 0.1).

mtDNA variant calling from sequencing data

To utilize the maximum number of samples for association testing, we called mtDNA variations in both WGS and WES using GATK HaplotypeCaller v4. Using --ploidy 1 in GATK HaplotypeCaller v4[77,78], we called 4,696 variants from WGS (of which 4,602 are SNPs and 255 are present on the UK Biobank Affymetrix Axiom array, **Supplementary Figure 1C**), and 3,618 variants from WES (of which 3,546 are SNPs and 254 are present on the UK Biobank Affymetrix Axiom array, **Supplementary Figure 1D**), giving a union of 5,247 variants (of which 5,161 are SNPs). We then performed the following checks and filters for sample and mtDNA variants. As mtDNA coverage is extremely high on the WGS (mean coverage = 2015.2x, SD = 578.7x), homoplasmic variations in the mtDNA would be

supported by thousands of reads and are therefore of high confidence. This confidence however cannot be extended to WES, due to the low coverage (mean coverage = 30.6, SD = 13.5x) and off-target nature of mtDNA reads on the WES (no mtDNA probes are present on the Agilent SureSelect Human All Exon v.5 kit).

mtDNA heteroplasmy check with mtdna-server

We checked the heteroplasmy levels at all WGS variant calls using the local version of mtdna-server (v1.1.11), a specialized software for variant calling in mtDNA that is particularly optimized for identification of heteroplasmic mutations in the mtDNA [79], so as to ensure the post-VQSR variants we obtained from WGS are not likely due to misidentification of heteroplasmic mutations as inherited homoplasmic variants. We were unable to perform this check directly on the WES data because mtdna-server was able to call only variations at only coverage of 30x and above - as the average coverage on the mtDNA in WES is 30.6x, we were only able to call 2,341 variants from 540 out of 4,502 samples, much fewer than present in the WGS samples. We first checked for correlation between genotypes called with GATK HaplotypeCaller and mtdna-server at 4,544 SNPs called with both methods. While 4,407 SNPs showed high correlation between both methods (Pearson r2 >= 0.9), 137 SNPs showed lower correlation. We checked the level of heteroplasmy at all levels of correlation and at all minor allele frequencies (MAF) determined with the mtdna-server genotype calls, and found that levels of heteroplasmy are higher for SNPs with lower correlation between GATK HaplotypeCaller and mtdna-server at all levels of MAF (**Supplementary Figure 1E**), and SNPs with high correlation between the two methods have lower heteroplasmy (mean heteroplasmy = 5.54×10^{-5}, SD = 2.9×10^{-4}) than those with low correlation (mean heteroplasmy = 0.018, SD = 0.08, **Supplementary Figure 1F**). This suggests that GATK may have mis-called these heteroplasmic sites as homoplasmic variants. We removed these 137 SNPs from both WGS and WES variant calls from all further analyses.

Sequencing-based mtDNA quality

Using 12,111 and 4,470 samples with WGS and WES, we compared genotypes at biallelic, polymorphic, and non-strand-ambiguous mtDNA SNPs called in WGS (209 SNPs) and WES (206 SNPs) against genotypes called in the Affymetrix Axiom array (out of a total of 235 biallelic, non strand-ambiguous SNPs genotyped on the array). We find that WES showed similar correlation and concordance with array genotypes (mean Pearson r^2 = 0.964, SD = 0.131, mean concordance = 0.999, SD = 0.004, 5 sites with more than 1% participants discordant) as WGS (mean Pearson r^2 = 0.961, SD = 0.131, mean concordance = 0.999, SD = 0.004, with the same 5 sites with more than 1% participants discordant as WES). Both analyses showed that WES can produce high quality genotypes and variant calls in the mtDNA, despite having two orders of magnitude lower coverage on the mtDNA than the WGS. In addition, we found that the same sites were discordant between Affymetrix and both sequencing platforms: 13 sites had Pearson r^2 of lower than 0.9 between Affymetrix and both WGS and WES, inclusive of the 5 that had greater than 1% discordant rate between Affymetrix and both sequencing platforms. This indicates that these 13 sites (magenta points in **Supplementary Figure 2**) likely represent errors in the Affymetrix array rather than either sequencing platforms.

mtDNA variant quality recalibration

To assess the quality of variant and genotyping calling at sites that are polymorphic in all participants in the WES and WGS cohorts, we used Gaussian mixture models in Variant Quality Score Recalibration (VQSR) in GATK (version 4.0.3.0)[80] to cluster all SNP calls from WES with SNPs of high concordance between the three platforms based on their variant call metrics. We first identified 92 SNPs of high concordance between all three platforms (Pearson $r^2 > 0.9$), and designated them as the "training" set for training the Gaussian mixture model (orange points in **Supplementary Figure 2**). We then restricted our "known" set to a) the 92 SNPs in the training set, b) the 189 SNPs genotyped on the array that showed good correlation between WES and array (Pearson $r^2 > 0.9$), and c) the 231 common SNPs with high minor allele frequency (MAF > 1%) in WGS, WES or array calls. From c), we removed 10 SNPs with low Affymetrix genotype quality (Pearson r^2 with WGS and WES < 0.9), 9 SNPs with low WES quality (Pearson r^2 with WGS < 0.9), and 1 SNP with low WGS quality (Pearson r^2 with Affymetrix < 0.9). We took the union of the remaining 211 SNPs from c) with a) and b) to arrive at 314 SNPs to use as the "known" set. The correlation between platforms and MAF of all SNPs from the three platforms is shown in **Supplementary Figure 2**. We then applied VQSR separately on the WES callset with "--trust-all-polymorphic" and "--max-gaussians 2" in SNP mode using the following annotations: QD, FS, MQ, MQRankSum, ReadPosRankSum, BaseQRankSum, SOR, and MLEAF. We obtained VQSLOD scores for all variants based on the clustering of their annotations with "known" and "training" sets, and selected variants not in the above sets whose transition to transversion ratio (Ti/Tv) most closely matched that of "known" variants in WES (known Ti/Tv = 63.3). After filtering WES variant calls with VQSLOD scores of lower than 1.91 (TruthSensitivity = 50%, Ti/Tv = 59.0), we obtained 212 high-quality SNPs from WES. We took the union of this set with a) 92 polymorphic SNPs from the "training" set and b) 189 SNPs with high correlation between WES and array, to obtain a total of 403 high-quality SNPs from the WES. 7 SNPs were strand-ambiguous or multi-allelic (mt.373, mt.1766, mt.3308, mt.7960, mt.13816, mt.14605, mt.15625), and were hence removed, leaving 396 high-quality SNPs, all of which were called from WGS. Of this, 187 are common at MAF > 1%, and we used these SNPs for association analyses as well as quantification of the total mtDNA contribution to molecular phenotypes.

Metabolite data quality control

Metabolome profiling was performed in two batches on plasma samples extracted from the blood of 9000 INTERVAL participants using the Metabolon HD4 mass spectroscopy discovery platform. This platform quantifies plasma metabolites using the Ultrahigh Performance Liquid Chromatography-Tandem Mass Spectroscopy (UPLC-MS/MS) method, and produces ion-counts for specific fragments that identify specific metabolites. Raw data was extracted, peak-identified and QC processed using Metabolon's hardware and software. As such, the raw data from this platform corresponds to but is not a direct measurement of plasma metabolite concentrations. Where metabolite levels were below the lower limit of detection, they were set to "missing" rather than 0 or the lowest detectable value, in order to prevent skewing of the data. The following steps were carried out for the filtering of samples and metabolites for ensuring only high-quality metabolite quantification was used in all subsequent analyses. First, only 7,778 out of 9,000 INTERVAL participants who were previously found to be unrelated and of European ancestry and had not withdrawn from the INTERVAL study are included in this study. Second, 68 participants had metabolite data quantified from repeat blood samples (taken either at the same or different times). For these participants, metabolite data quantified from the first blood samples (baseline) were kept, where available, and repeats were dropped. Where samples were from the same survey time,

both samples were dropped. Third, metabolites which were measured in only one batch (n=22) were excluded from the analysis. Fourth, participants with missing data for at least 300 metabolites (n=9) were excluded (arbitrary cut-point decided based on histogram of missingness) from further analysis. Quantification measures of all remaining 995 metabolites were then transformed by taking the natural logarithm, and then winsorized where the value was 5 or more standard deviations away from the mean metabolite value. The transformed metabolites were then regressed (linear regression) against the following covariates: age, sex, batch, INTERVAL recruitment centre, plate number, appointment month, the lag time between the blood donation appointment and sample processing, and the first 5 ancestry principal components obtained from genome-wide genotyping data from the Affymetrix UK Biobank Axiom array [75]. Following regression, the residuals were generated and inverse rank normalized. As metabolite levels below detection limits set to "missing" did not factor into any of the above transformations and remained "missing", we finally removed all metabolites that were missing in more than 50% (4000) of the samples, leaving 896 metabolites for association analysis with mtDNA variants.

Partitioning of contributions to metabolite levels

We tested for single mtDNA or nDNA SNP association with levels of each metabolite in LDAK, which implements a linear mixed model: for each of the common mtDNA or nDNA variants x_j of MAF >=5% in INTERVAL, we tested for their effect β on metabolite level y_i in a linear mixed model, controlling for all autosomal genetic effects and population structure in the random effect term g mean metabolite level (using an intercept term) in the fixed effect term F_i: $y_i = \alpha F_i + \beta x_j + g + \varphi$ where $g \sim N(0, \sigma_g^2 K_g)$ and $\varphi \sim N(0, \sigma_e^2 I)$, g is the random effect of the LD-weighted relatedness matrix K_g constructed using LDAk v5[42] with 5,511,276 common (MAF >= 5%) autosomal, biallelic SNPs nuclear DNA variants using --power -0.25 as recommended, and φ is the residual variance assuming an independent identically distribution (i.i.d) matrix for noise and uncaptured environmental effects. For each mtDNA SNP-metabolite association, we asked if we could increase the power to detect true associations by controlling environmental factors indexed by levels of selected metabolites or blood cell counts using the CMS framework[43] (**Supplementary Discussion**). To quantify relative levels of contribution of autosomal, mtDNA, and known environmental contribution to each metabolite, we performed a variance decomposition analysis with LDAK v5, using combinations of the following variance components: i) LD-weighted and MAF-adjusted (with --power 0.25 as recommended) relatedness matrix from 5,511,276 common SNPs on the autosomes only, ii) non-weighted relatedness matrix from 187 common SNPs on the mtDNA only, and iii) non- relatedness matrix from levels of 896 metabolites and 36 blood cell types, and assessed significance of contribution from each of the three variance components using likelihood ratio tests.

GTEx sample selection and genotype quality control

We obtained 5,696,458 autosomal, biallelic SNPs in 635 GTEx samples from the variant call set from WGS data in version 7[37], filtering on minor allele frequency (MAF >= 5%), P-value of violation of Hardy Weinberg Equilibrium (HWE > 10-6), and missingness (< 0.1). We then extracted reads mapping to the rCRS mitochondrial reference genome (NC_012920), and called mtDNA variants using mtdna-server[79], obtaining 1,756 variants, all of which are SNPs, and of which 46 are multi-allelic. For multi-allelic SNPs, we retained the two alleles with highest frequencies for analysis, hence obtaining a total of 1,714 SNPs, 64 of which are common (MAF >= 5%) for use in association testing. To identify unrelated individuals, we

used KING9 to identify related samples among the 635 GTEx samples. Two pairs of individuals in GTEx are related up to third-degree (kinship >=0.04419), though only marginally (kinship among pairs = 0.0477 and 0.0657 respectively), so we did not remove them from analyses. To identify individuals of European ancestry, we selected 4,812,475 common SNPs (MAF > 5%, P value for HWE > 10-6) from the autosomes that overlap between GTEx and 1000 Genomes Project Phase 3 (1000G)[81], built an LD-weighted kinship matrix using LDAK v5 and performed PCA on 1000G samples, projected the GTEx samples onto PCs from 1000G samples, and selected 491 GTEx samples that cluster with 1000G samples from the EUR superpopulation. We then built a kinship matrix using 5,696,456 common SNPs (MAF >= 5%, missingness < 0.1, P value for HWE > 10-6) in 491 European samples in GTEx for use in testing for association between mtDNA encoded genes and mtDNA SNPs with a linear mixed model.

Population structure on nuclear DNA and mtDNA in GTEx

We obtained 5,696,456 autosomal, biallelic SNPs in 635 GTEx participants from the variant call set from WGS data in version 7 [37], filtering on minor allele frequency (MAF > 5%), P-value of violation of Hardy Weinberg Equilibrium (HWE > 10^{-6}), and missingness (< 0.1). To identify unrelated participants, we used KING [82] to identify related participants among the 635 GTEx participants. Two pairs of participants in GTEx are related up to third-degree (kinship >=0.04419), though only marginally (kinship among pairs = 0.0477 and 0.0657 respectively), so we did not remove them from our analyses. To identify participants of European ancestry, we selected 4,812,475 common SNPs (MAF > 5%, P value for HWE > 10^{-6}) from the autosomes that overlap between GTEx and 1000 Genomes Project Phase 3 (1000G) [81] and built an LD-weighted kinship matrix using LDAK v5 [42], and projected the GTEx participants onto PCs obtained from 1000G samples using the same SNPs (**Supplementary Figure 5A**). We selected 491 GTEx participants that cluster with 1000G participants from the EUR superpopulation (**Supplementary Figure 5B**). Of these 491 participants, we were able to obtain WGS reads on the mtDNA of 456 participants. We then extracted reads mapping to the rCRS mitochondrial reference genome (NC_012920) from the 456 participants, and called mtDNA variants using both GATK HaplotypeCaller, using the same settings as we did in INTERVAL, as well as mtdna-server (v1.1.11) [79] to check for heteroplasmy. We obtained a total of 1,180 SNPs. Of the 1,180 SNPs, 38 are likely mis-identified as inherited homoplasmic SNPs due to heteroplasmy (mean heteroplasmy = 0.23, SD = 0.35, **Supplementary Figure 5C,D**). Of these, 12 are also found to be potential heteroplasmic sites in INTERVAL, indicating that these sites are consistently mis-identified as homoplasmic SNPs from WGS data due to heteroplasmy across studies. We removed all 38 potentially heteroplasmic sites, leaving us with a total of 1,142 SNPs (mean heteroplasmy = 2.48×10^{-4}, SD = 9.90×10^{-4}, **Supplementary Figure 5D**). Of these, we use the 56 SNPs that are common (MAF > 5%), for use in association testing. To assess the diversity on nuclear DNA and mtDNA in GTEx in these 456 participants, we built a LD-weighted kinship matrix using LDAK v5 [42] with 5,451,305 common SNPs (MAF > 5%, missingness < 0.1, P value for HWE > 10^{-6} in just the EUR participants) to obtain PCs specifically in these participants to compare against PCs obtained from the mtDNA (**Figure 1E,F**).

Obtaining PEER factors as covariates for eQTL analysis

Inaccuracies in quantification of expression levels of some nDNA genes attributable to mis-mapping of 100bp RNAseq reads originating from mtDNA encoded genes to the nuclear

mitochondrial sequence (NUMT) regions of the nucDNA, sequences on the nDNA that are highly similar to the mtDNA. As such, we aligned the sequence of each nuclear gene (57,820 genes) in GENCODE v19 on the rCRS mtDNA reference sequence NC012920 using lastal (version 744) [83] and found 651 genes with extensive sequence similarity (>=100bp) where the total fraction of genes with such alignment >= 5% (**Supplementary Table 6**). Of these, 84.0% are pseudogenes, 5.68% are lincRNAs, 2.30% are antisense RNAs, and 7.93% are protein coding genes, of which 0.1% (of total) are in introns. We excluded all of them in the calculation of PEER factors[84] for capturing unknown confounding factors in the RNAseq data and the association analysis. The numbers of PEER factors we used for correction of gene expression levels in each tissue increased with the number of participants with gene expression data in each tissue, following suggestions from GTEx release 7 [37], and are shown in **Supplementary Table 5**.

GTEx mtDNA eQTLs and multiple testing correction

We tested for single mtDNA SNP association with nDNA gene expression levels for every gene i (mt-eQTL) using LIMIX[85], a linear mixed model package in python. For each of the common mtDNA variants x_j of MAF > 5% in the GTEx participants, we tested for its effect β on PEER-factor[84] corrected gene expression y_i in a linear mixed model, controlling for all autosomal genetic effects and population structure in the random effect term g, and age, sex and mean gene expression level (using an intercept term) in the fixed effect term F_i: $y_i = \alpha F_i + \beta x_j + g + \varphi$, where $g \sim N(0, \sigma_g^2 K_g)$ and $\varphi \sim N(0, \sigma_e^2 I)$, g is the random effect of the LD-weighted relatedness matrix K_g constructed using LDAK with 5,696,456 common (MAF >= 5%) autosomal, biallelic SNPs nDNA variants, and φ is the residual variance assuming an independent identically distributed (i.i.d) matrix for noise and uncaptured environmental effects. We obtained nominal p-values for each variant-gene pair by testing the alternative hypothesis that the β between genotype and expression deviates from 0. We then calculated Beta distribution-adjusted (using the beta distribution model of the minimum P value distribution) empirical P values for the top cis-eQTL per gene using P values generated from 100 permutations of SNPs. These empirical P values were used to calculate Q values, the false discovery rate (FDR), using the "qvalue" package in R. A FDR threshold of 0.05 was applied to identify genes with a significant eQTL ("eGenes"), and the maximum empirical P value with FDR smaller than or equal to 0.05 was the gene-level threshold for identifying significant eQTLs (equivalent to empirical P value < 0.004).

Differential expression and pathway enrichment

We downloaded counts of reads mapping onto each gene quantified in each tissue in GTEx from the GTEx portal (version 2016-01-15_v7_RNASeQCv1.1.8), and performed normalization of the read counts by effective library size using calcNormFactors with the default trimmed-mean of M values (TMM) method using R package "edgeR"[62,63]. We then estimated the common, trended and tagwise dispersions over all genes using estimateDisp in edgeR, before computing the log(fold change), P values and false discovery rate (FDR) for differential expression in each gene between donors with mtDNA haplogroup H and haplogroup Uk using the exactTest function in edgeR. Pathway and GSEA analysis was performed using the WEB-based GEne SeT AnaLysis Toolkit[86,87] following their instructions online. Results are shown in **Supplementary Data 12,13**.

mtDNA sequencing in cytoplasmic hybrid cell lines

We verified the mtDNA sequence from the cybrid cell lines. mtDNA from cybrid cell lines was enriched using long-range PCR. To eliminate the potential for error and nDNA contamination (nuclear-mitochondrial sequences, NUMTs), amplicons were polymerised using PrimeSTAR GXL DNA polymerase (error rate = 0.00108 %, Takara Bio, Saint-Germain-en-Laye, France) in two overlapping fragments, using primer set-1: CCC TCT CTC CTA CTC CTG-F (m.6222-6239) and CAG GTG GTC AAG TAT TTA TGG–R (m.16133-16153), and set-2: CAT CTT GCC CTT CAT TAT TGC–F (m.15295-15315) and GGC AGG ATA GTT CAG ACG-R (7773-7791). Primer efficiency and specificity was assessed as successful after no amplification of DNA from rho^0 cell lines, minimising the unintended amplification of nuclear pseudogenes. Amplified products were assessed by gel electrophoresis, against DNA+ve and DNA-ve controls, and quantified using a Qubit 2.0 fluorimeter (Life Technologies, Paisley, UK). Each amplicon was individually purified using Agencourt AMPure XP beads (Beckman-Coulter, USA), pooled in equimolar concentrations and re-quantified. For the mtDNA sequencing pooled amplicons were 'tagmented', amplified, cleaned, normalised and pooled into 48 sample multiplexes using the Illumina Nextera XT DNA sample preparation kit (Illumina, CA, USA). Multiplex pools were sequenced using MiSeq Reagent Kit v3.0 (Illumina, CA, USA) in paired-end, 250 bp reads. Post run data, limited to reads with QV >= 30, were exported for analysis. Post-run FASTQ files were analysed using an in-house developed bioinformatic pipeline. Reads were aligned to the rCRS (NC_012920) using BWA v0.7.10, invoking –mem[88]. Aligned reads were sorted and indexed using Samtools v0.1.18 [89], duplicate åreads were removed using Picard v1.85 (http://broadinstitute.github.io/picard/). Variant calling (including somatic calling) was performed in tandem using VarScan v2.3.8 [90,91](minimum depth = 1,500, supporting reads = 10, base-quality (BQ) => 30, mapping quality (MQ) => 20 and variant threshold = 1.0 %) and LoFreq v0.6.1 [92]. Concordance calling between VarScan and LoFreq was > 99.5%. Concordant variants were annotated using ANNOVAR v529 [93]. In-house Perl scripts were used to extract base/read quality data and coverage data. The mtDNA haplogroup was determined through in-house algorithms based upon existing phylogenetic data [94,95] and through Haplogrep [96]. The pathogenicity Score was given to each of the SNPs as previously described [97,98].

Cell lines and formyl-methionine treatment

Cell lines were grown in Dulbecco's modified eagle medium (DMEM) containing glucose (4.5 g/l), pyruvate (0.11 g/l) and fetal bovine serum (FBS) (5 %) without supplemented fMet and/or antibiotics at 37^{o}C and 5 % CO2 conditions. 8 cell lines H (4 lines), Uk (4 lines) from 8 independent healthy control subjects were used. All the cybrids were obtained from cybrid pools after the selection process[10]. All the experiments were performed in at least 3 cell lines derived from different donors per haplogroup. The mtDNA sequences of all the cell lines can be found on GenBank and their mtDNA accession numbers[10] are included in **Supplementary Table 9.** In fMet supplementation experiments, cells were incubated with 1ng /ml of N-Formyl-L-methionine (F3377, SigmaAldrich) during 2 days prior to the experiments.

Cell growth and doubling time experiments

For growth experiments 30000 cells were seeded in a 6-well plate and counted every 4h using an Incucyte® Live-Cell Analysis system. Three to five growth curves were performed for every cell line, and each time point was counted in triplicate. Time 0h was used for correction of each well. For the doubling time analysis, the data was analyzed on an exponential curve and only those curves with $R2 \geq 0.9$ were considered.

Formyl-methionine quantification

fMet was quantified using the Formyl-methionine ELISA Kit from Elabscience (E-EL-0063) following manufacturer's conditions.

Real-time PCR quantification of transcripts.

Total RNA was isolated from cells exponentially growing using an RNA isolation kit from Qiagen® according to the manufacturer's protocol. Quantification of mRNA by real-time PCR (RT-PCR) was carried out using the High capacity cDNA reverse transcription kit (Applied Biosystems) following the manufacturer's conditions. The mRNA levels were determined using probes from Applied Biosystems and following MQIE guidelines[99]. The expression levels were normalized with *GADPH* and *B-ACTIN* as housekeeping genes. The codes of each of the probes are included in **Supplementary Table 17**. The comparative Cq method was used for relative quantification of gene expression. Differences in the Cq values (dCq) of the transcript of interest and the reference gene were used to determine the relative expression of the gene in each sample. The dCq method was used to calculate the number of copies.

Mitochondrial and Cytoplasmic Translation Assay

Protein translation was assessed by labeling with 35S-methionine/35S-cysteine (EXPRE35S35S Protein Labeling Mix; Perkin Elmer Life Sciences) in cells seeded at 80% of confluence. For cytoplasmic translation assessment, cells were washed twice with Met and Cys-free DMEM (21013024, ThermoFisher) followed by an incubation on the same media for one hour. Then, the cells were incubated with a labelling media containing Met and Cys-free DMEM, 2 mM Glutamine (25030081, ThermoFisher), 1mM Sodium Pyruvate, 96ug/ml Cystein (DOC0122, ForMedium) and 5% dialyzed Fetal Bovine Serum (30067334, ThermoFisher) for 10 minutes at 37 ºC, followed by the addition of 100uCi 35S L-Methionine and incubation for 30 minutes at 37 ºC. Mitochondrial translation assay was performed similarly than in the cytoplasmic assay with some adaptations. Cells were incubated during 20 minutes at 37 ºC with Labeling Medium including 100ug/ml emetine (E2375, Sigma) followed by the addition of 100μCi 35S L-Methionine and incubation for 60 min at 37 ºC. In both cases, cells were trypsinized and collected with PBS (and washed twice) and pellets kept at -80ºC. Proteins extraction was performed using a buffer containing 0.1% DDM (D4641, SigmaAldrich), 1% Sarkosyl (L9150, SigmaAldrich) and 50 units of Benzonase (Novagen 70664 25U/ul) vortexed vigorously and left on ice for 30 minutes. Protein quantities were assessed by DC assay (Biorad 500-0113) following manufacturer's instructions. Next, 15ug of protein was loaded onto 15-well-12% Tris-Glycine gels (Invitrogen NP0343BOX) using MES buffer and ran for 3 hours at 70 volts. Total protein levels were assessed by Coomassie blue staining (0.1% Coomassie blue in 7% acetic acid and 40% methanol and de-stained with (20%methanol, 7% Acetic acid) solution for 2-3h. Images of the gel were collected using a Scanner. Gels were dried at 80ºC for 2 hours. Dried gels were then exposed for several days and imaged using a phosphor imaging screen on an Amersham™ Typhoon™ Biomolecular Imager. The bands were quantified, aligned and cropped using the Fiji program and the OD was used as a value for statistical purposes.

Mitochondrial bioenergetics characterization

Oxygen consumption modifications. Briefly, 20 × 104 cells/well were seeded 8-12 hours before the measurement basal respiration, leaking respiration (LR), maximal respiratory capacity (MRC) and not mitochondrial respiration (NMR) were determined by adding 1 μM oligomycin (LR), 0.75 and 1.5 μM of carbonyl cyanide-p-trifluoromethoxyphenylhydrazone (FCCP) (MRC) and 1 μM rotenone/antimycin (NMR), respectively. Data were corrected by the NMR and expressed as pmol of oxygen/min/mg of protein. The quantity of protein in each well was measured by Bradford method [100].

Determination of MIMP and cytoplasmic and mitochondrial ROS,

The determination of mitochondrial inner membrane potential (MIMP) was carried using Tetramethylrhodamine, methyl ester (TMRM) at 20nM (DMSO) in parallel to the mitochondrial mass detection using Mito-Traker Green (20nM in DMSO). Mitochondrial superoxide content was measured using MitoSOX Red at 5μM in DMSO. Cytosolic ROS were measured using 2′,7′-dichlorofluorescin-diacetate at 9μM in DMSO. All the reagents were purchased in Invitrogen®. Fluorescence activated detection was carried using a BD LSRFortessa™ cell analyzer from BD. 20000 events were recorded and doublet Discrimination was carried using the FCS-Height and Area FlowJo Software. An example of the gating strategy is shown in **Supplementary Figure 7**. The data is expressed as intensity of fluorescence.

Determination of ATP levels

ATP levels were measured four times in three independent experiments using the CellTiter-Glo® Luminiscent Cell Viability Assay (Promega) according to the manufacturer's instructions. Briefly, 10,000 cells/well were seeded and the media was changed 48h before the measurement. After that time cells were lysed, and lysates were incubated with the luciferin/luciferase reagents. Samples were measured using a NovoStar MBG Labtech microplate luminometer, and the results referred to the protein quantity measured in a parallel plate.

Electrophoresis and Western blot analysis.

Samples for blue-native gel electrophoresis (BNGE) and *in gel* activities were prepared as previously described [54,101]. Native samples were run through precast NativePAGE 3–12% Bis–Tris gels during 6-9 hours. Total protein extracts were prepared according to each protein's solubilities. Mitochondrial proteins were prepared using 2% dodecyl-maltoside in PBS including protease inhibitors. Protein extracted for kinase phosphorylation analysis was extracted using PathScan® Sandwich ELISA Lysis Buffer from Cell signaling. In any case protein extracts were loaded on NuPAGE® Bis-Tris Precast Midi Protein Gels with MES (Invitrogen®) with 20 or 26 wells depending on the experiment. Electrophoresis was carried out following the manufacturer's conditions. SeeBlue® Plus2 Pre-stained Protein Standard from Invitrogen® was used in each electrophoresis as protein size markers. The separated proteins were transferred to polyvinylidene fluoride membranes using the iBLOT system (Invitrogen®) or Mini Trans-Blot® transfer system from Biorad®. The resulting blots were probed overnight at 4 ºC with primary antibodies with the appropriate concentration following manufacturer's condition with small adaptations (Antibodies, and concentrations are attached in **Supplementary Table 18**). After the primary antibody, blots were incubated for 1 h with secondary antibodies conjugated with horseradish peroxidase (HRP) and immuno-detected using an Amersham Imager 600. The bands for each antibody were

quantified, aligned and cropped using the Fiji program and the OD was used as a value for statistical purposes. In order to avoid inter-blot variation one cell line was used as an internal control and the values of the OD corrected by β-Actin were relative to it in each case.

Reproducibility of the experiments and statistical analysis

All of the experiments present in this work were performed in 3 independent biological replicates (unless noted otherwise) and statistical analyses were derived from these data (Prism 8.0.1). Normal distributions were validated by the Kolmogorov–Smirnov test. One-way ANOVA followed by the Holm-Sidak test for multiple comparisons Kruskal-Wallis was applied for group comparison tests.

Mitochondrial disease patients harboring the m.3243A>G variant

Serum samples from patients carrying the m.3243A>G variant were obtained from Prof Chinnery's neurogenetic/mitochondrial clinic at through the study: Genotype and Phenotype in Inherited Neurodegenerative Diseases (REC ID: 13/YH/0310, IRAS ID: 136697) Cambridge University Hospitals NHS Trust. Age and gender-matched controls were obtained in the NIHR BioResource and the Blood and Stem Cell Biobank (Cambridge, UK) Ethics ID: 13/YH/0310.

Oxford Vascular Study

OXVASC is a longitudinal population-based incidence cohort of all acute vascular events in a defined population of 92,728 people, covered by around 100 primary care physicians in nine primary care practices in Oxfordshire, UK. An estimated 97% of the true study residential population is registered with a primary care practice; most unregistered people are young students. The study area contains a mix of urban and rural populations. The OXVASC population is 94% Caucasian, 3% Asian, 2% Chinese, and 1% Afro-Caribbean. Written informed consent or assent from relatives is obtained in all participants for study, interview and follow-up, including ongoing review of primary care and hospital records and death certificate data. OXVASC was approved by the Oxfordshire research ethics committee (OREC A: 05/Q1604/70). Multiple overlapping methods are used for ascertainment of all participants with TIA and stroke, approaching 100% of events reaching medical attention. These include the following: (a) a daily, rapid access clinic to which participating general practitioners and the local emergency department refer participants with suspected TIA or minor stroke; (b) daily searches of admissions to the medical, stroke, neurology, and other relevant wards; (c) daily searches of the local emergency department attendance register; (d) daily searches of in-hospital death records via the Bereavement Office; (e) monthly searches of all death certificates and coroner's reports for out-of-hospital deaths; (f) monthly searches of general practitioner diagnostic coding and hospital discharge codes; and (g) monthly searches of all brain and vascular imaging referrals. Demographic data and stroke risk factors are collected from face-to-face interviews by study physicians as soon as possible after referral or hospital admission and cross-referenced with primary care records. Detailed clinical history was recorded in all patients and assessments were made for stroke severity using the National Institute of Health Stroke Scale (NIHSS) as recorded on assessment. Cause of ischaemic events was classified according to the Trial of Org 10172 in Acute Stroke Treatment (TOAST) criteria. Stroke and TIA were defined according to WHO criteria (acute onset of neurological deficit, persisting for >24 hours in case of a stroke, or for <24 hours in

case of a TIA), with review of all cases as soon as possible after presentation by the same senior neurologist throughout the study. Non-fasting blood samples were taken as soon as possible after the event, usually within one day. These included serum, 3.2% buffered tri-sodium citrate plasma and lithium heparin plasma (Vacutainer tubes; Becton Dickinson, United Kingdom). Samples were centrifuged at 3000 g for 10 minutes, and aliquots of serum and plasma were stored at -80°C before analysis when they were thawed for use at 37°C. All times from sampling to freezing were documented, typically within 4 hours of taking.

ICICLE-PD Cohort

Plasma samples from PD patients and controls were obtained from the 'Incidence of Cognitive Impairment in Cohorts with Longitudinal Evaluation-PD' (ICICLE-PD) study, which includes newly-diagnosed PD cases and unrelated control subjects of a similar age recruited from the community and outpatient clinics in Newcastle and Cambridge, UK[102]. Idiopathic PD was diagnosed according to UKPDS Brain Bank criteria [103]. The study was approved by the Newcastle and North Tyneside Research Ethics Committee. All patients provided written informed consent. Venous blood samples were collected in EDTA tubes at baseline study visits (between 2009 and 2011), and centrifuged within 30 minutes at 2000rpm for 15 minutes. Plasma was removed and stored in 200µl aliquots at -80C until assays were performed.

mtDNA sequencing of patients cohorts

We selected 282 participants with ischaemic stroke (IS) and 181 age-matched controls from the Oxford Vascular Study (OXVASC) [104] for sequencing of mtDNA using the Illumina Hiseq 2000 using an amplicon-based paired-end library preparation; both groups are enriched for individuals with Haplogroup Uk (32% haplogroup Uk in cases, 23% haplogroup Uk in controls). We also sequenced 123 participants with Parkinson's disease (PD) and 40 age-matched controls from the Incidence of Cognitive Impairment in Cohorts with Longitudinal Evaluation–PD (ICICLE-PD) cohort [102] with the same platform; both groups represent population samples and are not enriched for any mtDNA haplogroups. The Fluidigm Access ArrayTM technology was used to generate tagged and indexed amplicons (on average 100 per sample of 150-200bp), with sample-specific barcodes and Illumina adaptor sequences. The resulting PCR products were checked for quality using the Agilent 2100 Bioanalyzer and then pooled together in equal volumes. The PCR product library was purified using AMPure XP beads and quantified with PicoGreen prior to loading for Illumina sequencing. 183 age-matched controls from the OXVASC cohort were sequenced with Illumina Miseq using a paired-end library preparation. Mitochondrial DNA MiSeq libraries were prepared by amplification of two overlapping fragments [105]. After individual purification and quantification, the amplicons from each sample were pooled in equal amounts. Libraries were prepared with NEBNext Ultra library prep reagents (New England BioLabs, MA) according to manufacturer's instructions and sequenced using a 2 × 250-cycle MiSeq Reagent kit v3.0 (Illumina, CA).

Mitochondrial variant calling and haplogroup prediction

Quality of raw sequencing fastq files was checked with FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) prior to mapping and eventually trimmed using TrimGalore!

(https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) to remove low-quality read ends (--q 20), remove adapters (--stringency 5) and Ns from either side of the read (–trim-n). Trimmed reads below 35bp were removed (–length 35). Read were mapped using the MToolBox pipeline (v.1.1) [106] which performs a two-step reads mapping, first on the rCRS mitochondrial reference genome and then simultaneously on the hg19 nuclear reference and rCRS reference, to remove possible nuclear-mitochondrial DNA sequences (NumtS) contaminations. PCR duplicates were removed with MarkDuplicates in the picard package (https://gatk.broadinstitute.org/hc/en-us/articles/360037052812-MarkDuplicates-Picard-) from sequencing generated with Illumina Miseq and with FastUniq [107] from sequencing generated with amplicon-based library preparations. The average coverage obtained (percentage of mtDNA molecules covered by at least one read) was 99.5% for IS samples, 98.7% for PD samples and 100% for controls. Average mitochondrial read depth was 1220X for IS samples, 1299X for PD samples and 2282X for controls. Mitochondrial variant calling was performed with the MToolBox pipeline, using the default options (minimum read depth per alternative allele ≥ 5 and minimum quality score per base ≥ 25). Haplogroup predictions were generated with the Haplogrep 2 software [96], using VCF files with homoplasmic and nearly homoplasmic (i.e. with heteroplasmic fraction ≥ 0.8) variants generated with MToolBox.

mtDNA haplogroup association in patient cohorts

We measured fMet levels 631 participants from both the IS and PD cohorts (Stroke N = 282, Parkinson's disease N = 124, Control N = 225) in three batches (**Supplementary Table 14**). Haplogroup predictions were available for 95% of the samples (N=601) whose mtDNA were deep-sequenced. In the IS cohort, 92 IS cases and 17 controls were of haplogroup Uk, while 190 IS cases and 139 controls were of other haplogroups. fMet levels measured in pg/ml from the samples were controlled for batch and data collection site (Cambridge or Newcastle) of fMet measurement using linear regression, and residuals were rank normalised for further analysis. Associations between normalized fMet levels with Uk haplogroup was tested using a logistic regression implemented with the R glm function (family = "binomial"). In the PD cohort, 2 PD cases and 4 controls were of haplogroup Uk, and 118 PD cases and 39 controls were of other haplogroups. fMet levels measured in pg/ml from the samples were controlled for batch of fMet measurement (all PD samples were collected at the same site) using linear regression, and residuals were rank normalised for further analysis in the same fashion as in the IS cohort.

EPIC Norfolk cohort

We obtained incidences of late-onset diseases from 11,966 men and women from the EPIC-Norfolk prospective cohort EPIC Norfolk cohort. Participants were identified as having experienced an event if the corresponding ICD-10 code was registered on the death certificate (as the underlying cause of death or as a contributing factor), or as the cause of hospitalization. Participants were on average 60 years (standard deviation: 6 years) old and 46.3% were men. All participants were flagged for mortality at the UK Office of National Statistics, and vital status was ascertained for the entire cohort. Death certificates were coded by trained nosologists according to the International Classification of Diseases (ICD), 10th revision. Hospitalization data were obtained using National Health Service numbers through linkage with the East Norfolk Health Authority (ENCORE) database, which contains information on all hospital contacts throughout England and Wales. Participants were identified as having experienced an event if the corresponding ICD-10 code was registered on

the death certificate (as the underlying cause of death or as a contributing factor), or as the cause of hospitalization, **Supplementary Data 11**). The current study is based on follow-up to 31st March 2016.

fMet measurement and mtDNA genotyping in EPIC-Norfolk

fMet was measured from plasma samples stored in liquid nitrogen since baseline in 1993-97 from a total of 11,966 men and women from the EPIC-Norfolk prospective cohort as part of an untargeted metabolomic profiling using Metabolon's DiscoveryHD4™ platform (Metabolon Inc., Morrisville, North Carolina, USA). Measurements were undertaken in two sub-cohorts of 5,989 and 5,977 participants, respectively, quasi-randomly selected from the full cohort. Prior to statistical analyses, fMet levels were transformed using the natural logarithm and values at the tail of the distribution, defined by mean ± 5 x standard deviation, were replaced by the respective lower/upper bound. They were then rescaled to a mean of zero and standard deviation of one. Processing steps were performed for each batch separately. mtDNA haplogroups in 10,782 participants were obtained with Haplogrep v2, using the --chip option with genotype data at 262 (2 of which were multi-allelic) mtDNA variants on the Affymetrix UK Biobank Axiom genotyping array (895 participants fall under haplogroup Uk, 9,887 individuals were of other haplogroups).

Cox-proportional hazards models

We used Cox-proportional hazards models to estimate hazard ratios for the association of fMet levels (log-transformed and standardized) with first incidences of 24 diseases and health outcomes during a 20-year follow up period, with age as the underlying time scale adjusting for sex. For each incident outcome, we excluded participants reporting an instance of the outcome at baseline. For cancer outcomes, we additionally excluded all participants with the onset of any cancer within six months after baseline. mtDNA haplogroups in 10,782 participants were obtained with Haplogrep v2, using the --chip option with genotype data at 262 (2 of which were multi-allelic) mtDNA variants on the Affymetrix UK Biobank Axiom genotyping array (895 participants fall under haplogroup Uk, 9,887 individuals were of other haplogroups). Cox-proportional hazard models for fMet effects on outcome incidence were calculated for all participants, as well as separately within each haplotype group on 24 outcomes, all of which had more than 10 incidences in both haplogroup Uk or otherwise, to obtain both cohort-based and haplotype-specific hazard ratios (**Supplementary Table 16**). As fMet levels are positively associated with age, we note that outcomes with significant fMet hazard ratios were late-onset diseases or outcomes. Further, we checked for potential violations of the proportional hazard assumption in Cox-proportional hazard models with Schoenfeld residual tests for each outcome, and found no violations except for cataracts, on which fMet does not have a significant effect (**Supplementary Table 16**). We further tested for interaction effects between mtDNA haplogroups and fMet levels to formally test for differences in effect estimates, and found no significant interaction effects (**Supplementary Table 16**).

Data availability:

The Genotype-Tissue Expression (GTEx) Project was supported by the Common Fund of the Office of the Director of the National Institutes of Health, and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. The data used for the analyses described in this manuscript were obtained from the GTEx Portal (GTEx_Analysis_2016-01-15_v7_RNASeQCv1.1.8) and

dbGaP accession number phs000424.v7.p2. All data is available in the main text or the supplementary materials, or available upon request to the authors.

Code Availability:

We conducted our analyses using the following published and publicly available software: 1) calling mtDNA variants: GATK v4.0.3.0 HaplotypeCaller (https://gatk.broadinstitute.org/hc/en-us/articles/360037225632-HaplotypeCaller) and mtdna-server local version (https://github.com/seppinho/mutserve), 2) for mtDNA association analysis using a linear mixed model and variance decomposition analysis: LDAK v5 (http://dougspeed.com/downloads2/); 3) for improving power of mtDNA association: CMS v1.0 (https://github.com/haschard/CMS); 4) for eQTL analyses: limix v3.0 (https://github.com/limix/limix); 5) for identifying pseudogenes in nuclear genome with high sequence similarity to the mtDNA: lastal 744 (http://last.cbrc.jp/doc/lastal.html); 6) for identifying PEER factors that capture unknown confounding in gene expression data: PEER v1.3 (https://github.com/PMBio/peer); 7) differential expression analysis: edgeR v3.11 (http://bioconductor.org/packages/release/bioc/html/edgeR.html); 8) gene set enrichment analysis: GSEA v4.1.0 (https://www.gsea-msigdb.org/gsea/index.jsp); 9) Flow cytometry analysis: FlowJo v10.2

Supplementary Materials:

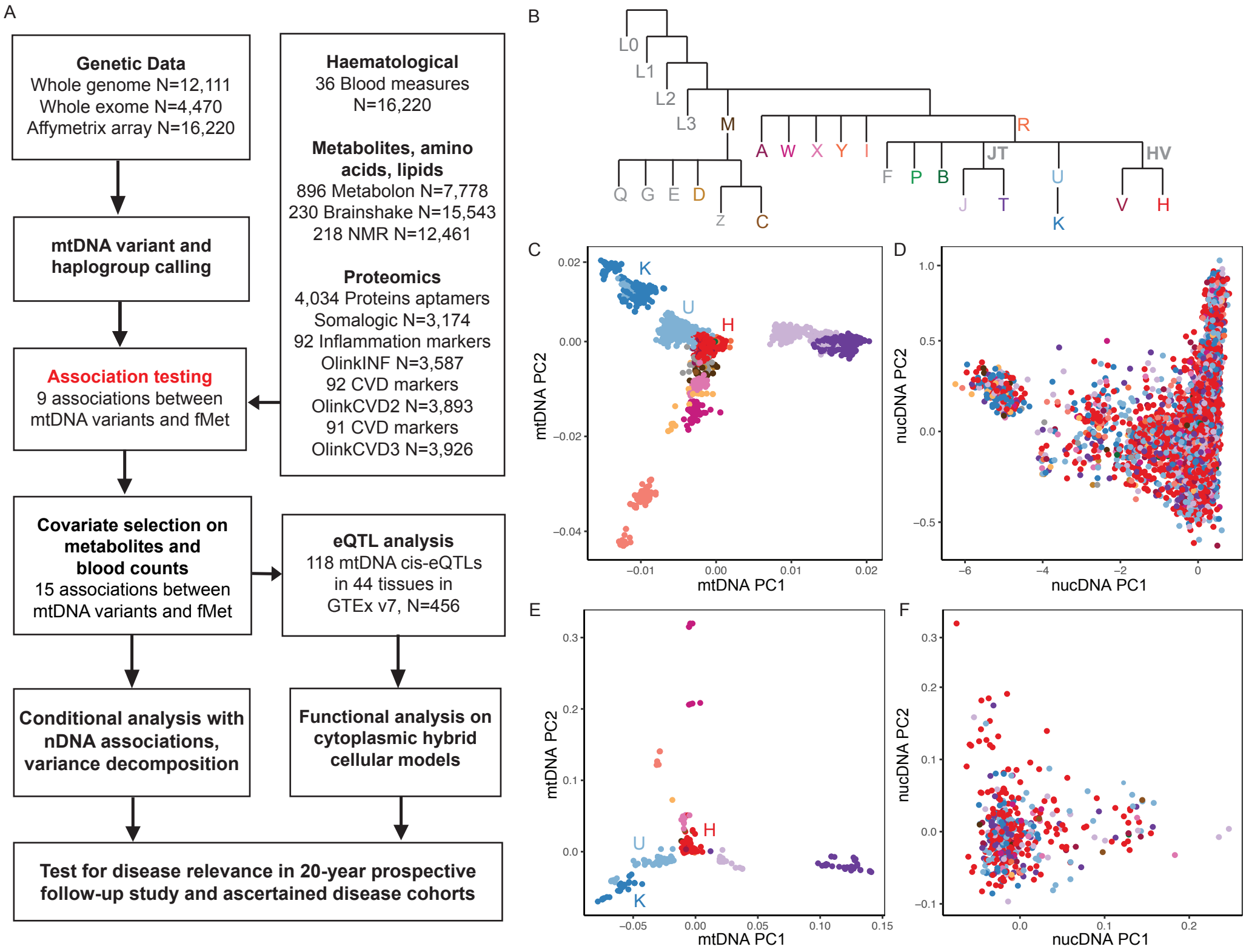
Members of the ICICLE-PD Consortium, Supplementary Discussion, Supplementary Figures 1-7, Supplementary Tables 1-18, References (*1-10*)

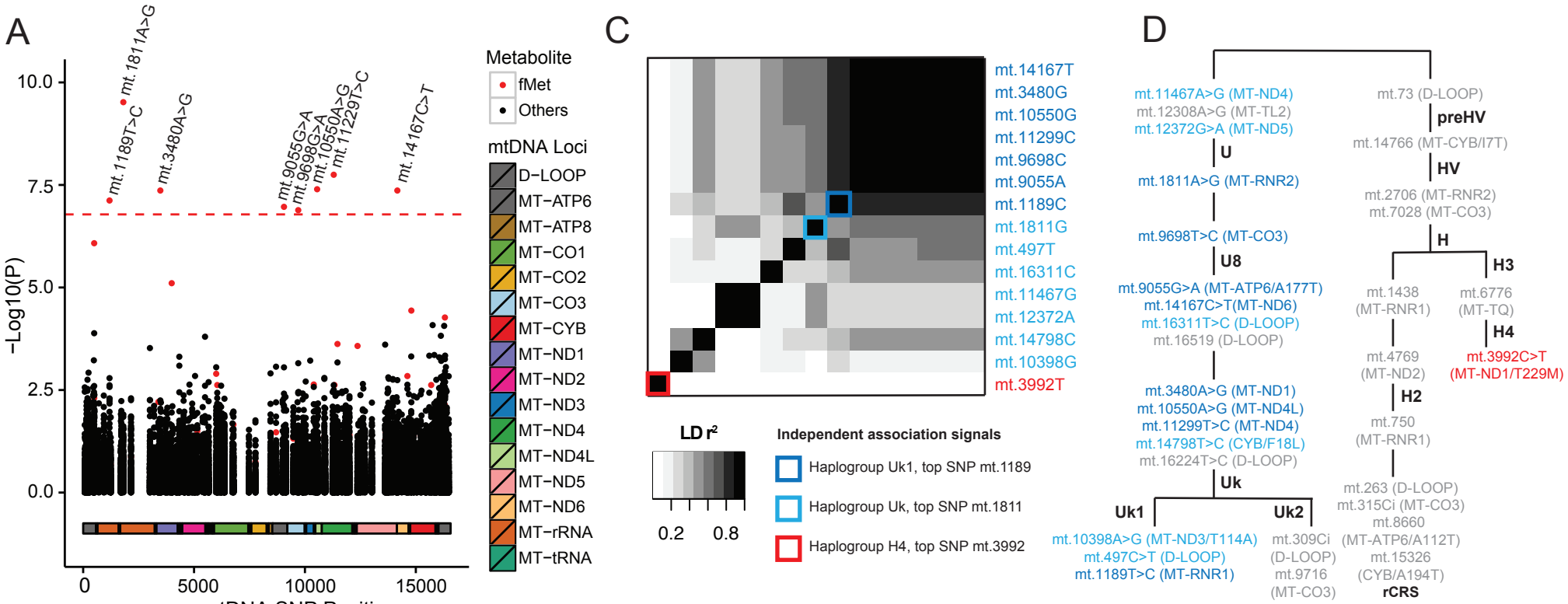
Members of the ICICLE-PD Consortium included in the author list:

Caroline H Williams-Gray[3]

3. Department of Clinical Neurosciences, School of Clinical Medicine, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK

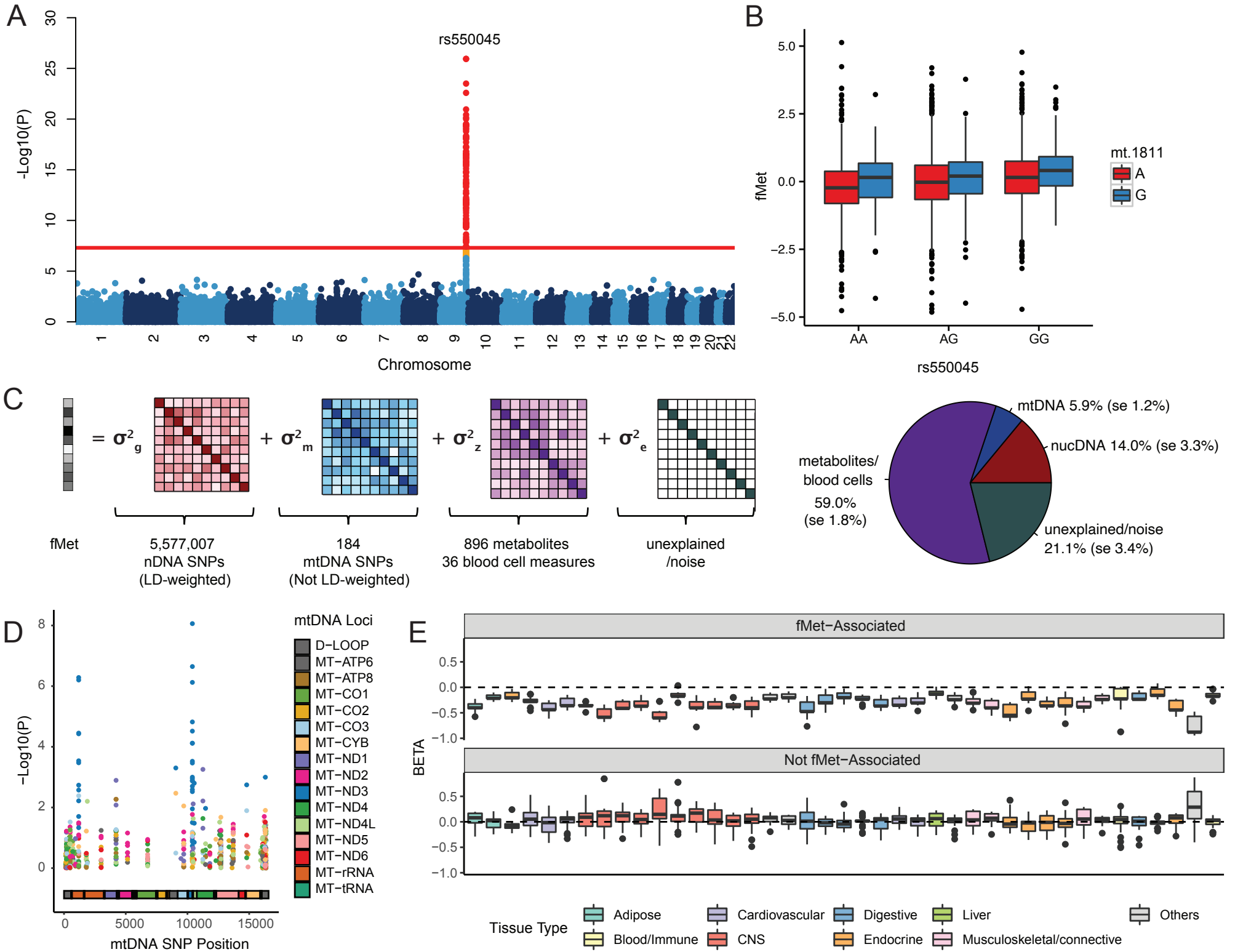
A full list of members and their affiliations appears in the Supplementary Information.

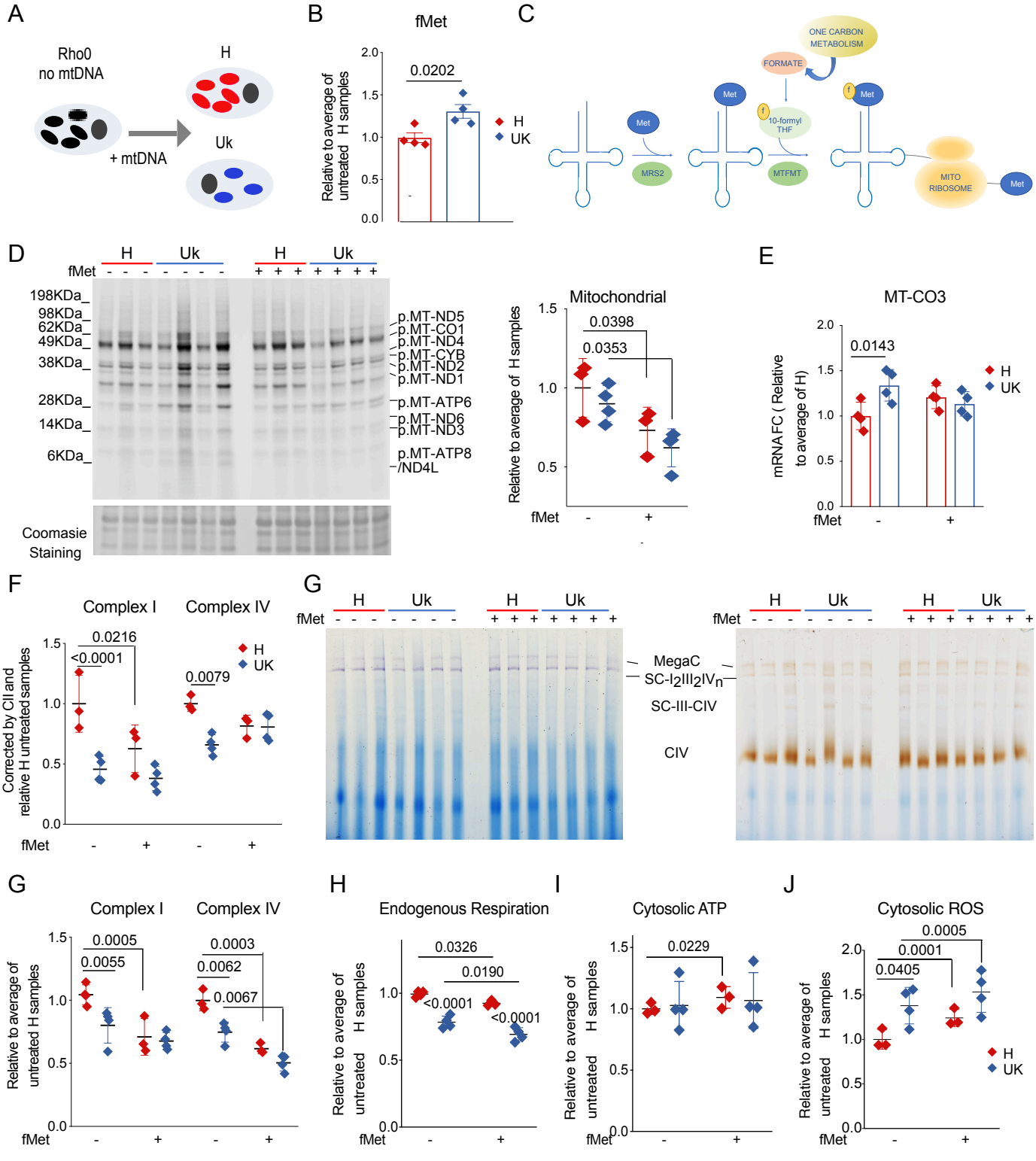

A
Genetic Data
Whole genome N=12,111
Whole exome N=4,470
Affymetrix array N=16,220
mtDNA variant and haplogroup calling
Association testing
9 associations between mtDNA variants and fMet
Haematological
36 Blood measures
N=16,220
Metabolites, amino acids, lipids
896 Metabolon N=7,778
230 Brainshake N=15,543
218 NMR N=12,461
Proteomics
4,034 Proteins aptamers
Somalogic N=3,174
92 Inflammation markers
OlinkINF N=3,587
92 CVD markers
OlinkCVD2 N=3,893
91 CVD markers
OlinkCVD3 N=3,926
Covariate selection on metabolites and blood counts
15 associations between mtDNA variants and fMet
eQTL analysis
118 mtDNA cis-eQTLs in 44 tissues in GTEx v7, N=456
Conditional analysis with nDNA associations, variance decomposition
Functional analysis on cytoplasmic hybrid cellular models
Test for disease relevance in 20-year prospective follow-up study and ascertained disease cohorts
B
L0
L1
L2
L3
M
R
A
W
X
Y
I
JT
HV
F
P
B
U
Q
G
E
D
J
T
V
H
Z
C
K
C
mtDNA PC2
mtDNA PC1
D
nucDNA PC2
nucDNA PC1
E
mtDNA PC2
mtDNA PC1
F
nucDNA PC2
nucDNA PC1

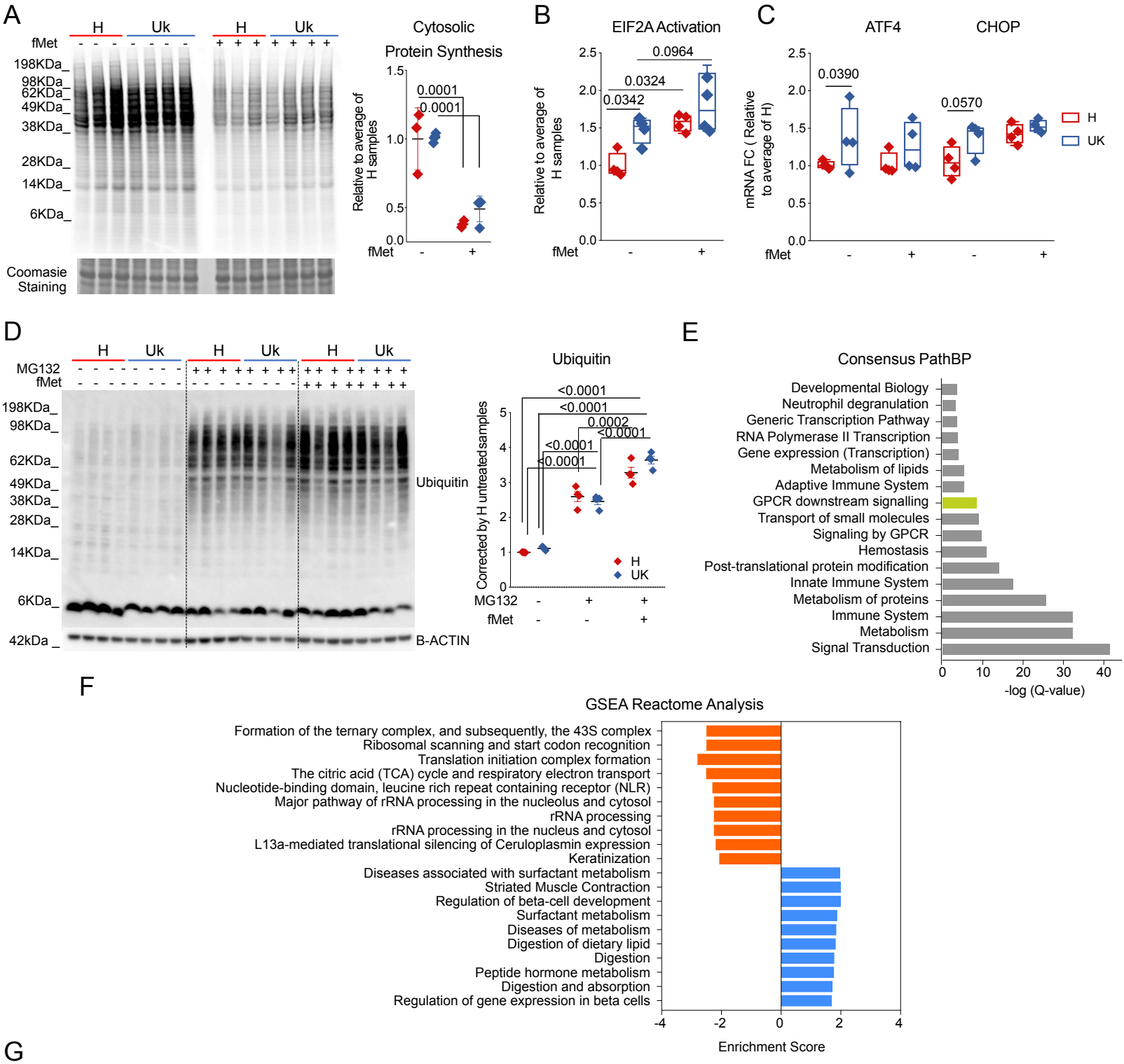


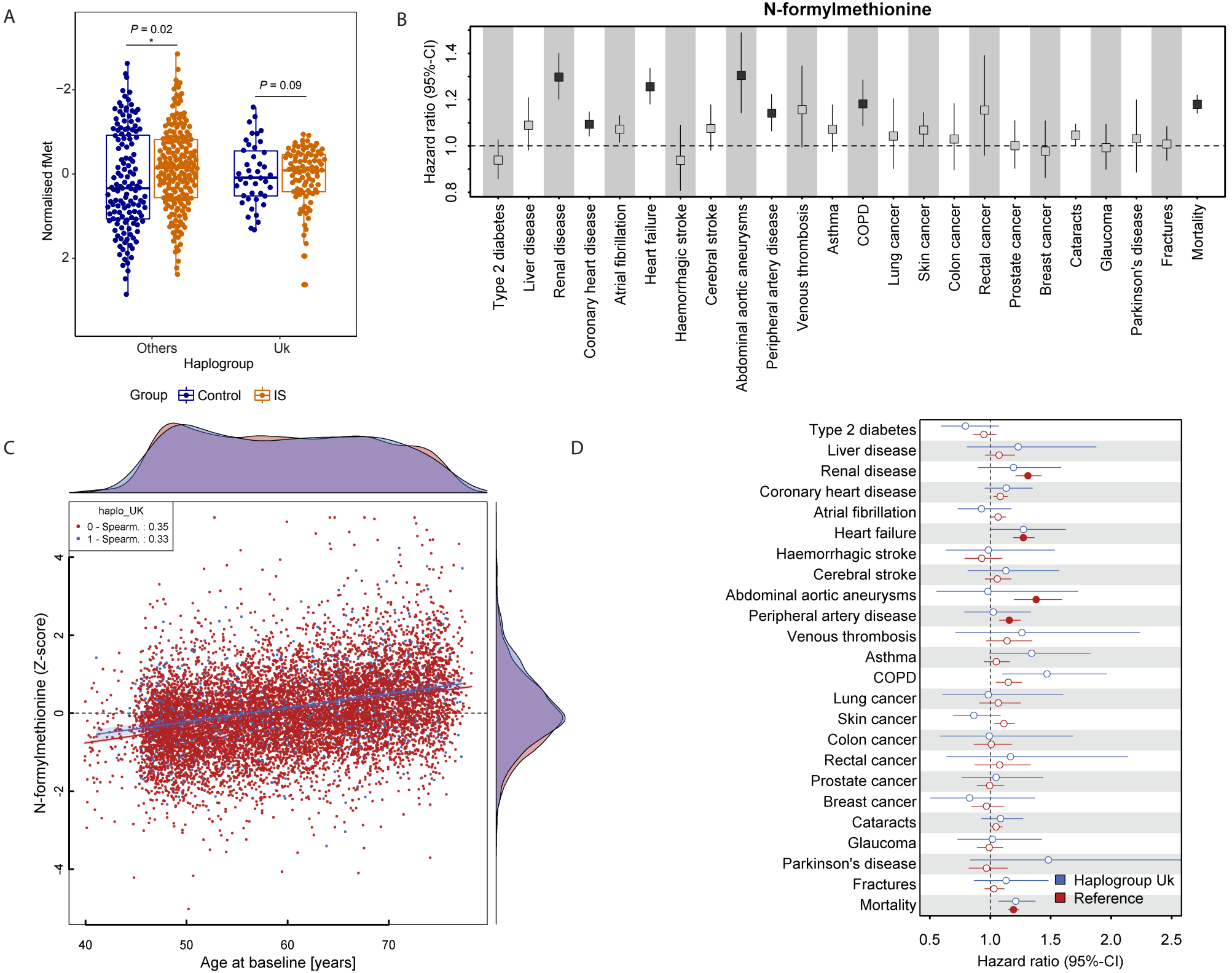
B

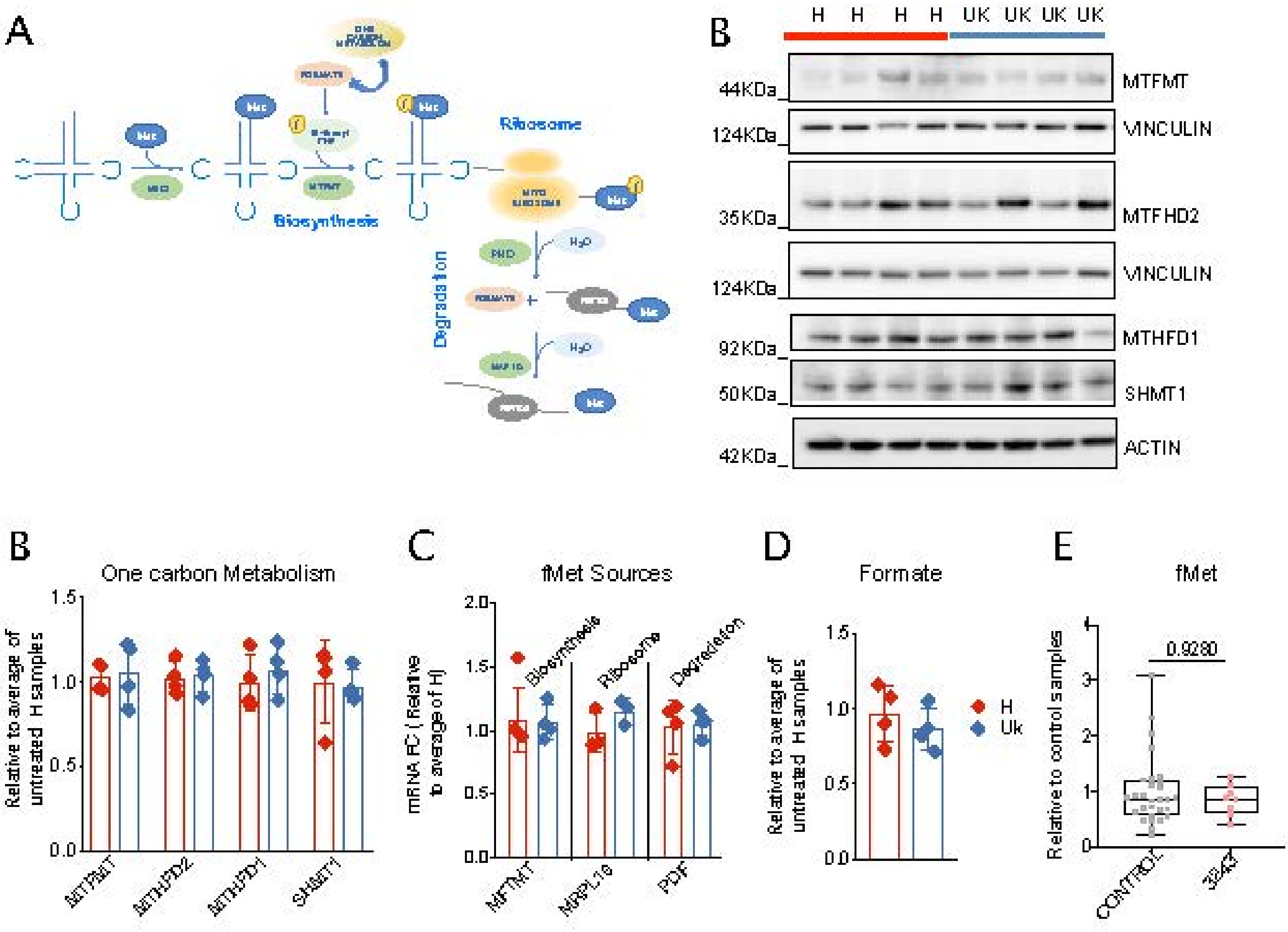
SNP						DISCOVERY (INTERVAL)						REPLICATION (EPIC)		
						LMM			CMS			LMM		
BP	GENE	A1	A0	A1FREQ	ANNOT	BETA	SD	P	BETA	SD	P	BETA	SD	P
497	D-LOOP	T	C	0.050	D-LOOP	0.144	0.029	8.33×10^{-7}	0.039	0.007	5.67×10^{-9}	0.100	0.022	4.74×10^{-6}
1189	MT-RNR1	C	T	0.072	12S rRNA	0.133	0.025	7.57×10^{-8}	0.057	0.007	7.36×10^{-18}	Not genotyped		
1811	MT-RNR2	G	A	0.129	16S rRNA	0.120	0.019	3.03×10^{-10}	0.057	0.007	4.63×10^{-18}	Not genotyped		
3480	MT-ND1	G	A	0.087	syn:K58K	0.125	0.023	4.32×10^{-8}	0.057	0.007	1.44×10^{-17}	0.078	0.017	3.33×10^{-6}
3992	MT-ND1	T	C	0.023	non-syn:T229M	-0.190	0.042	7.86×10^{-6}	-0.054	0.007	6.05×10^{-16}	-0.137	0.033	3.93×10^{-5}
9055	MT-ATP6	A	G	0.089	non-syn:A177T	0.119	0.022	1.08×10^{-7}	0.054	0.007	2.47×10^{-16}	0.070	0.016	2.13×10^{-5}
9698	MT-CO3	C	T	0.089	syn:L164L	0.119	0.023	1.30×10^{-7}	0.056	0.007	2.05×10^{-17}	0.079	0.017	3.23×10^{-6}
10398	MT-ND3	G	A	0.216	non-syn:T114S	0.048	0.016	2.32×10^{-3}	0.038	0.007	8.68×10^{-9}	Not genotyped		
10550	MT-ND4L	G	A	0.086	syn:M27M	0.126	0.023	4.02×10^{-8}	0.058	0.007	3.59×10^{-18}	0.079	0.017	2.82×10^{-6}
11229	MT-ND4	C	T	0.087	syn:T180T	0.128	0.023	1.78×10^{-8}	0.058	0.007	1.80×10^{-18}	Not genotyped		
11467	MT-ND4	G	A	0.229	syn:L236L	0.056	0.015	2.39×10^{-4}	0.040	0.007	1.50×10^{-9}	Not genotyped		
12372	MT-ND5	A	G	0.229	syn:L12L	0.056	0.015	2.67×10^{-4}	0.040	0.007	1.82×10^{-9}	Not genotyped		
14167	MT-ND6	T	C	0.086	syn:E169E	0.125	0.023	4.32×10^{-8}	0.057	0.007	1.44×10^{-17}	0.076	0.017	7.62×10^{-6}
14798	MT-CYB	C	T	0.166	non-syn:F18L	0.071	0.017	3.66×10^{-5}	0.039	0.007	3.92×10^{-9}	0.044	0.013	6.83×10^{-4}
16311	D-LOOP	C	T	0.088	D-LOOP	0.070	0.017	5.39×10^{-5}	0.050	0.007	3.60×10^{-12}	Not genotyped		


A
rs550045
-Log10(P)
Chromosome
B
fMet
mt.1811
A
G
AA
AG
GG
rs550045
C
fMet
5,577,007
nDNA SNPs
(LD-weighted)
184
mtDNA SNPs
(Not LD-weighted)
896 metabolites
36 blood cell measures
unexplained
/noise
mtDNA 5.9% (se 1.2%)
nucDNA 14.0% (se 3.3%)
metabolites/
blood cells
59.0%
(se 1.8%)
unexplained/noise
21.1% (se 3.4%)
D
−Log10(P)
mtDNA SNP Position
mtDNA Loci
D−LOOP
MT−ATP6
MT−ATP8
MT−CO1
MT−CO2
MT−CO3
MT−CYB
MT−ND1
MT−ND2
MT−ND3
MT−ND4
MT−ND4L
MT−ND5
MT−ND6
MT−rRNA
MT−tRNA
E
fMet−Associated
Not fMet−Associated
BETA
Tissue Type
Adipose
Blood/Immune
Cardiovascular
CNS
Digestive
Endocrine
Liver
Musculoskeletal/connective
Others


A
Rho0
no mtDNA
+ mtDNA
H
Uk
B
fMet
Relative to average of untreated H samples
0.0202
H
UK
C
ONE CARBON METABOLISM
FORMATE
Met
10-formyl THF
MTFMT
MRS2
MITO RIBOSOME
D
H
Uk
fMet
198KDa_
98KDa_
62KDa_
49KDa_
38KDa_
28KDa_
14KDa_
6KDa_
p.MT-ND5
p.MT-CO1
p.MT-ND4
p.MT-CYB
p.MT-ND2
p.MT-ND1
p.MT-ATP6
p.MT-ND6
p.MT-ND3
p.MT-ATP8
/ND4L
Coomasie Staining
Mitochondrial
Relative to average of H samples
0.0398
0.0353
E
MT-CO3
mRNA FC (Relative to average of H)
0.0143
F
Complex I
Complex IV
Corrected by CII and relative H untreated samples
0.0216
<0.0001
0.0079
G
MegaC
SC-I2III2IVn
SC-III-CIV
CIV
G
Complex I
Complex IV
Relative to average of untreated H samples
0.0005
0.0055
0.0003
0.0062
0.0067
H
Endogenous Respiration
0.0326
0.0190
<0.0001
<0.0001
I
Cytosolic ATP
0.0229
J
Cytosolic ROS
0.0005
0.0001
0.0405

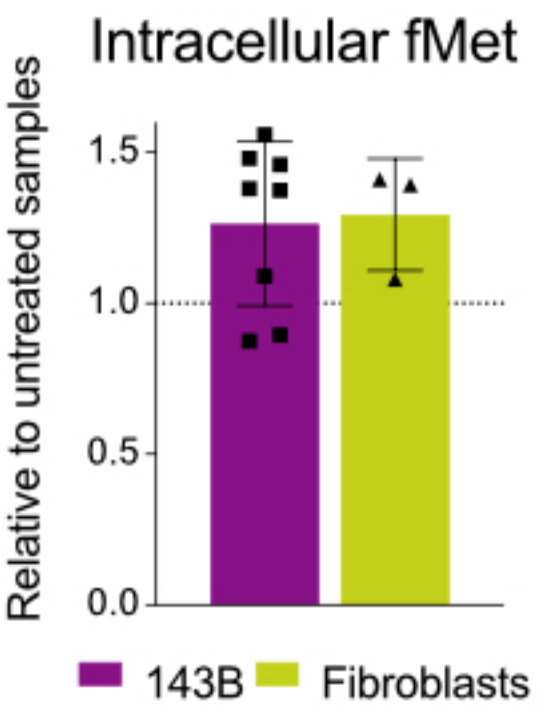

A
H
Uk
fMet
198KDa_
98KDa_
62KDa_
49KDa_
38KDa_
28KDa_
14KDa_
6KDa_
Coomasie Staining
Cytosolic Protein Synthesis
Relative to average of H samples
0.0001
0.0001
fMet
B
EIF2A Activation
0.0964
0.0324
0.0342
C
ATF4
CHOP
mRNA FC (Relative to average of H)
0.0390
0.0570
H
UK
D
MG132
Ubiquitin
B-ACTIN
42kDa _
Corrected by H untreated samples
<0.0001
<0.0001
0.0002
<0.0001
<0.0001
<0.0001
E
Consensus PathBP
Developmental Biology
Neutrophil degranulation
Generic Transcription Pathway
RNA Polymerase II Transcription
Gene expression (Transcription)
Metabolism of lipids
Adaptive Immune System
GPCR downstream signalling
Transport of small molecules
Signaling by GPCR
Hemostasis
Post-translational protein modification
Innate Immune System
Metabolism of proteins
Immune System
Metabolism
Signal Transduction
-log (Q-value)
F
GSEA Reactome Analysis
Formation of the ternary complex, and subsequently, the 43S complex
Ribosomal scanning and start codon recognition
Translation initiation complex formation
The citric acid (TCA) cycle and respiratory electron transport
Nucleotide-binding domain, leucine rich repeat containing receptor (NLR)
Major pathway of rRNA processing in the nucleolus and cytosol
rRNA processing
rRNA processing in the nucleus and cytosol
L13a-mediated translational silencing of Ceruloplasmin expression
Keratinization
Diseases associated with surfactant metabolism
Striated Muscle Contraction
Regulation of beta-cell development
Surfactant metabolism
Diseases of metabolism
Digestion of dietary lipid
Digestion
Peptide hormone metabolism
Digestion and absorption
Regulation of gene expression in beta cells
Enrichment Score
G
Model
Met
MRS2
tRNA
10-formyl-THF
MTFMT
mito-ribosome
peptide
Efficient synthesis/ fMet
Complex I and IV Oxygen consumption
mtStress
mtProtein synthesis
Mitochondria
Cytoplasm
MRS
Ribosome
Protein synthesis
Integrated Stress Response
Ub
Ubiquitination/ Proteasomal degradation


A
P = 0.02
*
P = 0.09
Normalised fMet
−2
0
2
Others
Uk
Haplogroup
Group
Control
IS
B
N-formylmethionine
Hazard ratio (95%-CI)
0.8
1.0
1.2
1.4
Type 2 diabetes
Liver disease
Renal disease
Coronary heart disease
Atrial fibrillation
Heart failure
Haemorrhagic stroke
Cerebral stroke
Abdominal aortic aneurysms
Peripheral artery disease
Venous thrombosis
Asthma
COPD
Lung cancer
Skin cancer
Colon cancer
Rectal cancer
Prostate cancer
Breast cancer
Cataracts
Glaucoma
Parkinson's disease
Fractures
Mortality
C
haplo_UK
0 - Spearm. : 0.35
1 - Spearm. : 0.33
N-formylmethionine (Z-score)
4
2
0
-2
-4
40
50
60
70
Age at baseline [years]
D
Type 2 diabetes
Liver disease
Renal disease
Coronary heart disease
Atrial fibrillation
Heart failure
Haemorrhagic stroke
Cerebral stroke
Abdominal aortic aneurysms
Peripheral artery disease
Venous thrombosis
Asthma
COPD
Lung cancer
Skin cancer
Colon cancer
Rectal cancer
Prostate cancer
Breast cancer
Cataracts
Glaucoma
Parkinson's disease
Fractures
Mortality
Haplogroup Uk
Reference
0.5
1.0
1.5
2.0
2.5
Hazard ratio (95%-CI)

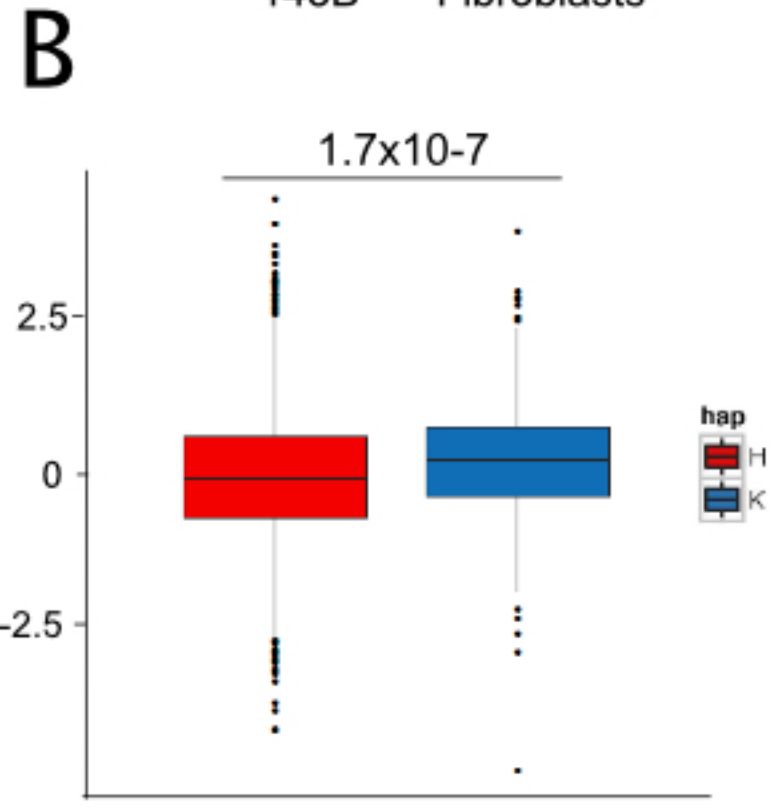


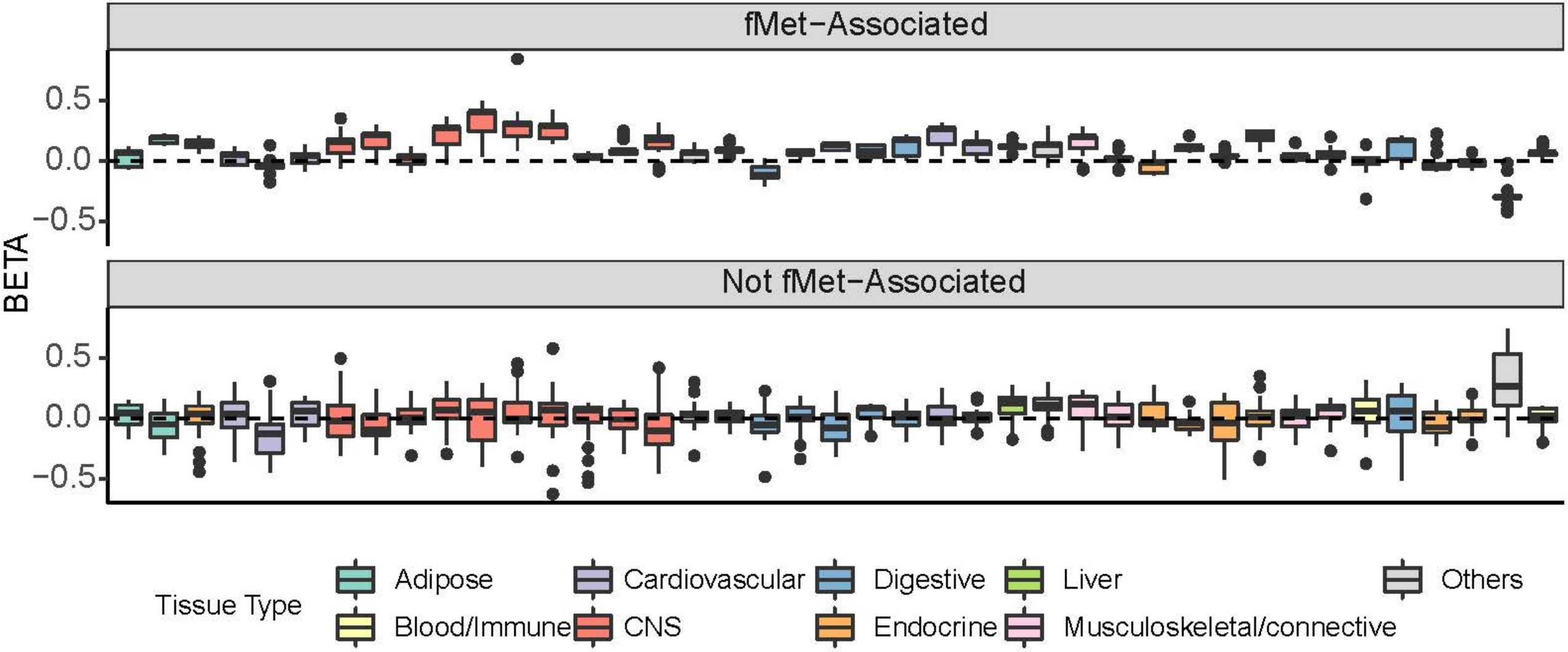
A
Ribosome
Biosynthesis
Degradation
B
H
UK
MTFMT
VINCULIN
MTFHD2
MTHFD1
SHMT1
ACTIN
44KDa
124KDa
35KDa
92KDa
50KDa
42KDa
One carbon Metabolism
Relative to average of untreated H samples
C
fMet Sources
mRNA FC (Relative to average of H)
Biosynthesis
Ribosome
Degradation
MTFMT
MRPL16
PDF
D
Formate
H
Uk
E
fMet
Relative to control samples
0.9280
CONTROL
3243

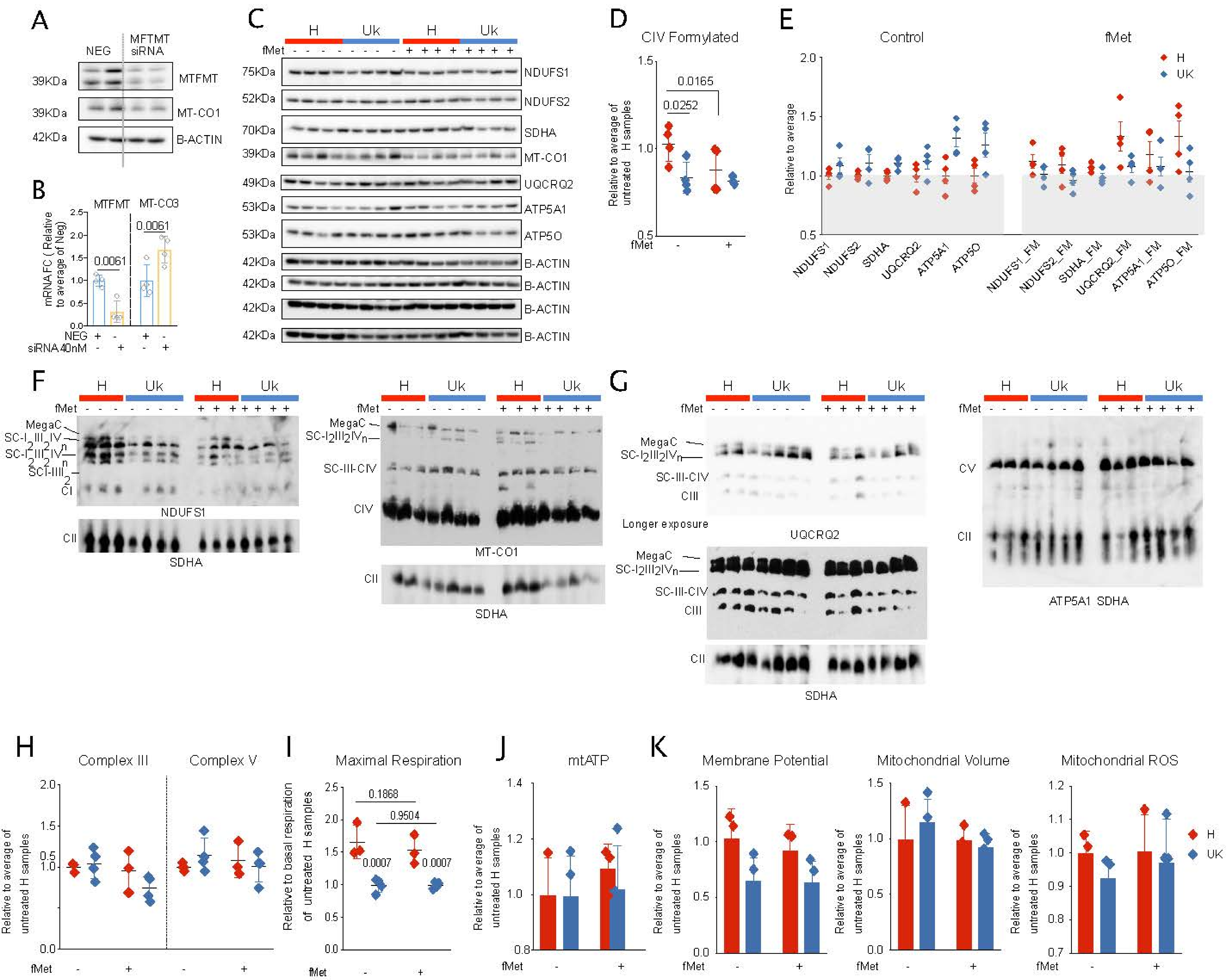
A

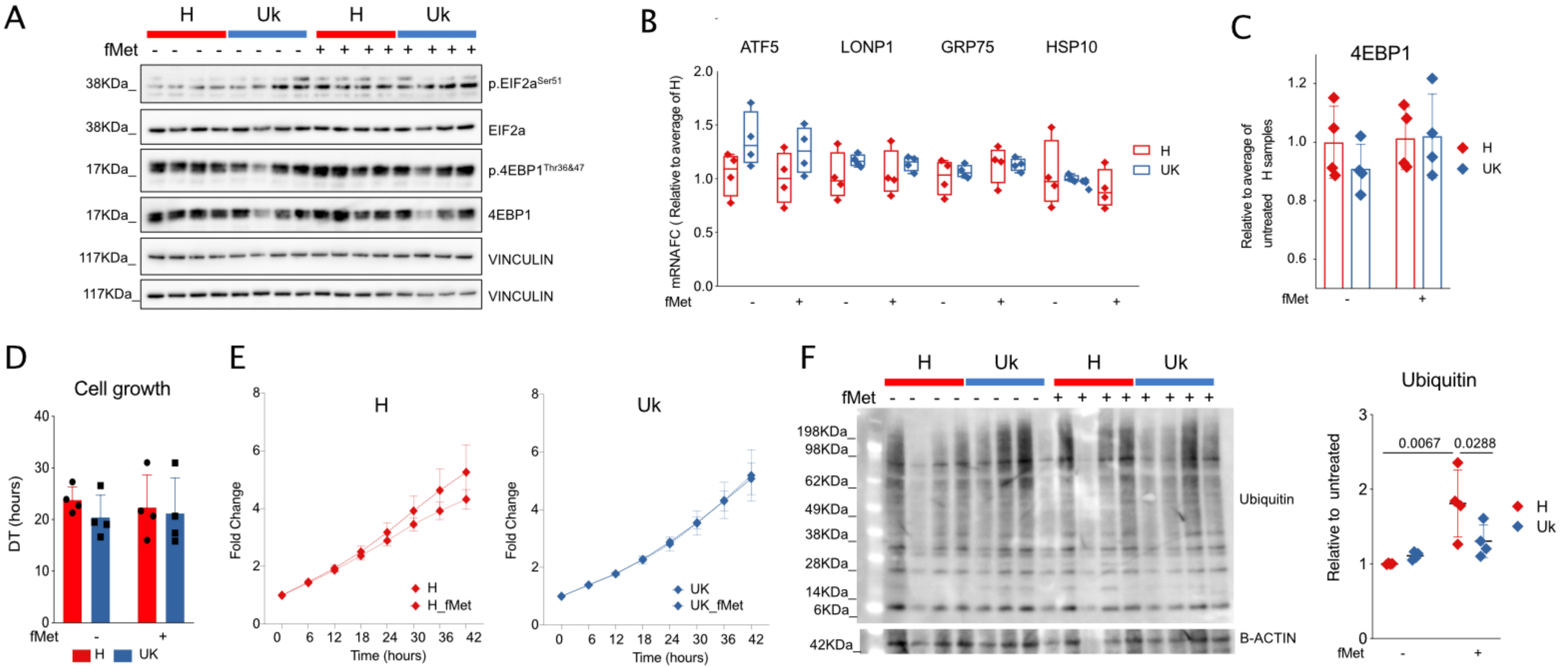

Intracellular fMet
Relative to untreated samples
1.5
1.0
0.5
0.0
143B
Fibroblasts

B


1.7x10-7
2.5
0
-2.5
hap
H
K


fMet−Associated
Not fMet−Associated
BETA
0.5
0.0
−0.5
Tissue Type
Adipose
Blood/Immune
Cardiovascular
CNS
Digestive
Endocrine
Liver
Musculoskeletal/connective
Others


A
NEG
MFTMT siRNA
39KDa
MTFMT
39KDa
MT-CO1
42KDa
B-ACTIN
B
MTFMT
MT-CO3
mRNA FC (Relative to average of Neg)
0.0061
0.0061
NEG
siRNA 40nM
C
H
Uk
fMet
75KDa
NDUFS1
52KDa
NDUFS2
70KDa
SDHA
39KDa
MT-CO1
49KDa
UQCRQ2
53KDa
ATP5A1
53KDa
ATP5O
42KDa
B-ACTIN
D
CIV Formylated
Relative to average of untreated H samples
0.0165
0.0252
fMet
E
Control
fMet
H
UK
Relative to average
NDUFS1
NDUFS2
SDHA
UQCRQ2
ATP5A1
ATP5O
NDUFS1_FM
NDUFS2_FM
SDHA_FM
UQCRQ2_FM
ATP5A1_FM
ATP5O_FM
F
MegaC
SC-I2III2IVn
SCI-III2
CI
NDUFS1
CII
SDHA
SC-III-CIV
CIV
MT-CO1
G
CIII
Longer exposure
UQCRQ2
CV
ATP5A1 SDHA
H
Complex III
Complex V
Relative to average of untreated H samples
I
Maximal Respiration
Relative to basal respiration of untreated H samples
0.1868
0.9504
0.0007
0.0007
J
mtATP
K
Membrane Potential
Mitochondrial Volume
Mitochondrial ROS


A
H
Uk
H
Uk
fMet
38KDa_
p.EIF2a Ser51
38KDa_
EIF2a
17KDa_
p.4EBP1 Thr36&47
17KDa_
4EBP1
117KDa_
VINCULIN
117KDa_
VINCULIN
B
ATF5
LONP1
GRP75
HSP10
mRNAFC (Relative to average of H)
H
UK
fMet
C
4EBP1
Relative to average of untreated H samples
H
UK
fMet
D
Cell growth
DT (hours)
fMet
H
UK
E
H
Uk
Fold Change
Time (hours)
H
H_fMet
UK
UK_fMet
F
H
Uk
H
Uk
fMet
198KDa_
98KDa_
62KDa_
49KDa_
38KDa_
28KDa_
14KDa_
6KDa_
Ubiquitin
42KDa_
B-ACTIN
Ubiquitin
0.0067
0.0288
Relative to untreated
H
Uk
fMet