Mitochondrial DNA Copy Number (mtDNA-CN) Can Influence Mortality and Cardiovascular Disease via Methylation of Nuclear DNA CpGs

Christina A. Castellani[1], Ryan J. Longchamps[1], Jason A. Sumpter[1], Charles E. Newcomb[1], John A. Lane[2], Megan L. Grove[3], Jan Bressler[3], Jennifer A. Brody[4], James S. Floyd[4], Traci M. Bartz[4,5], Kent D. Taylor[6], Penglong Wang[7], Adrienne Tin[8], Josef Coresh[8], James S. Pankow[9], Myriam Fornage[3,10], Eliseo Guallar[8], Brian O'Rourke[11], Nathan Pankratz[2], Chunyu Liu[12], Daniel Levy[7], Nona Sotoodehnia[4], Eric Boerwinkle[3], Dan E. Arking[1,11,*]

1) McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, MD; 2) Department of Laboratory Medicine and Pathology, University of Minnesota School of Medicine, Minneapolis, MN; 3) Human Genetics Center, School of Public Health, University of Texas Health Science Center at Houston, Houston, TX; 4) Cardiovascular Health Research Unit, Department of Medicine, University of Washington, Seattle, WA; 5) Department of Biostatistics, University of Washington, Seattle, WA; 6) Institute for Translational Genomics and Population Sciences, Los Angeles BioMedical Research Institute at Harbor-UCLA Medical Center, Torrance, CA; 7) Framingham Heart Study, Framingham, MA, USA, Population Sciences Branch, National Heart, Lung, and Blood Institute, National Institutes of Health, Bethesda, MD; 8) Department of Epidemiology and the Welch Center for Prevention, Epidemiology and Clinical Research, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD; 9) Division of Epidemiology & Community Health, School of Public Health, University of Minnesota, Minneapolis, MN; 10) Brown Foundation Institute of Molecular Medicine, McGovern Medical School, University of Texas Health Science Center at Houston, Houston, TX; 11) Division of Cardiology, Department of Medicine, Johns Hopkins University School of Medicine, Baltimore, MD.; 12) Department of Biostatistics, Boston University School of Public Health.

*corresponding author, arking@jhmi.edu
Johns Hopkins University School of Medicine, 733 N. Broadway Miller Research Building, Room 459, Baltimore, MD, 21205; 410-502-4867 (Phone), 410-614-8600 (Fax).

ABSTRACT

Mitochondrial DNA copy number (mtDNA-CN) has been associated with a variety of aging-related diseases, including all-cause mortality. However, the mechanism by which mtDNA-CN influences disease is not currently understood. One such mechanism may be through regulation of nuclear gene expression via the modification of nuclear DNA (nDNA) methylation. To investigate this hypothesis, we assessed the relationship between mtDNA-CN and nDNA methylation in 2,507 African American (AA) and European Americans (EA) participants from the Atherosclerosis Risk in Communities (ARIC) study using the Infinium Human Methylation 450K Beadchip (485,764 CpGs). Thirty-four independent CpGs were associated with mtDNA-CN at genome-wide significance ($P<5x10^{-8}$). To validate our findings we assayed an additional 2,528 participants from the Cardiovascular Health Study (CHS) (N=533) and Framingham Heart Study (FHS) (N=1995). Meta-analysis across all cohorts identified 6 mtDNA-CN associated CpGs to be validated across cohorts at genome-wide significance ($P<5x10^{-8}$). Additionally, over half of these CpGs were associated with phenotypes known to be associated with mtDNA-CN, including CHD, CVD, and mortality. Experimental modification of mtDNA-CN through knockout via CRISPR-Cas9 of *TFAM*, a regulator of mtDNA replication, demonstrated that modulation of mtDNA-CN directly drives changes in nDNA methylation and gene expression of specific CpGs and nearby transcripts. Strikingly, the 'neuroactive ligand receptor interaction' KEGG pathway was found to be highly overrepresented in the ARIC cohort ($P= 5.24x10^{-12}$), as well as the *TFAM* knockout methylation ($P=4.41x10^{-4}$) and expression ($P=4.30x10^{-4}$) studies. These results demonstrate that changes in mtDNA-CN influence nDNA methylation at specific loci and result in differential gene expression of specific genes, including those acting in the 'neuroactive ligand receptor interaction' pathway that may impact human health and disease via altered cell signaling.

INTRODUCTION

Mitochondria are cytoplasmic organelles primarily responsible for cellular metabolism, and have pivotal roles in many cellular processes, including aging, apoptosis and oxidative phosphorylation[1]. Dysfunction of the mitochondria has been associated with complex disease presentation including susceptibility to disease and severity of disease[2]. Mitochondrial DNA copy number (mtDNA-CN), a measure of mtDNA levels per cell, while not a direct measure of mtDNA damage, is associated with mitochondrial enzyme activity and adenosine triphosphate production. mtDNA-CN is regulated in a tissue-specific manner and in contrast to the nuclear genome, is present in multiple copies per cell, with the number being highly dependent on cell type[3]. Further, levels of mtDNA-CN correlate with mitochondrial function[4]. mtDNA-CN is therefore a relatively easily attainable biomarker of mitochondrial function. Cells with reduced mtDNA-CN show reduced expression of vital complex proteins, altered cellular morphology, and lower respiratory enzyme activity[5]. Variation in mtDNA-CN has been associated with numerous diseases and traits, including cardiovascular disease[6–8], chronic kidney disease[9], diabetes[10,11], and liver disease[12,13]. Lower mtDNA-CN has also been found to be associated with frailty and all-cause mortality[10].

Communication between the mitochondria and the nucleus is bi-directional and it has long been known that cross-talk between nDNA and mtDNA is required for proper cellular functioning and homeostasis[14,15]. Specifically, bi-directional cross-talk is essential for the maintenance and integrity of cells[16,17], and interactions between mtDNA and nuclear DNA (nDNA) contribute to a number of pathologies[18,19]. However, the precise relationship between mtDNA and the nuclear epigenome has not been well defined despite a number of reports which have identified a relationship between mitochondria and the nuclear epigenome. For example, mtDNA polymorphisms have been previously demonstrated to alter nDNA methylation patterns[20] and

hyper- and hypo- methylation of nuclear sites has been observed in mitochondria-depleted cancer cell lines[21]. Additionally, differential DNA methylation in brain tissue and corresponding differential gene expression were observed between strains of mice having identical nDNA, but different mtDNA[18] and reduced mtDNA-CN has been associated with inducing cancer progression via hypermethylation of nuclear DNA promoters[22]. Further, mtDNA-CN has been previously associated with changes in nuclear gene expression[23].

Thus, gene expression changes identified as a result of mitochondrial variation may be mediated, at least in part, by nDNA methylation. Further, given that it has been well-established that mtDNA-CN influences a number of human diseases we propose that one mechanism by which mtDNA-CN influences disease may be through regulation of nuclear gene expression via the modification of nDNA methylation.

To this end, we report the results of cross-sectional analysis of this association between mtDNA-CN and nDNA methylation in 5,035 individuals from the ARIC, CHS and FHS cohorts. Further, to determine the causal direction of the association between mtDNA-CN and nDNA methylation, we present results from experimental modification of mtDNA-CN followed by assessment of nDNA methylation and gene expression profiles in mtDNA-CN depleted cell lines.

RESULTS

mtDNA-CN is associated with nuclear DNA methylation at independent genome-wide loci in cross-sectional analysis

We performed an epigenome-wide association study (EWAS) in 2,507 individuals from the Atherosclerosis Risk in Communities (ARIC) study, comprised of 1,567 African American (AA)

and 940 European American (EA) subjects (Figure 1, Table 1, Table S1). 34 independent CpGs were significantly associated with mtDNA-CN ($P<5x10^{-8}$) in a meta-analysis combining the race groups (Figure 2, Figure S1, Table 2, Table S2A). This conservative *P*-value cutoff was confirmed by permutation testing. In stratified analysis of ARIC AA and EA participants, we identified 23 and 15 independent CpGs at epigenome-wide significance, respectively (Figure S2, Table S2B,C). Two CpGs were shared by both race groups (cg26094004 and cg21051031). ARIC AA and EA effect sizes for significant results were strongly correlated (R^2=0.49) (Figure S3). Further, 16/23 (70%) of AA cohort-identified CpGs showed the same direction of effect in EA participants *(P*=0.06) and 12/14 (86%) of EA cohort-identified CpGs displayed the same direction of effect in AA participants *(P*=0.008). Given these observations, we have focused on the ARIC results from combining both races (N=2,507) in further analyses.

Additionally, an association was observed between increased mtDNA-CN and global hypermethylation ($P<2.2x10^{-16}$, ß=0.1487) in ARIC AA, however no such association was seen in ARIC EA (*P*<0.77, ß=0.013) (Figure S4).

Pathway and biological process analysis displays associations with cell signaling functions and the 'Neuroactive ligand-receptor interaction' pathway

To assess the potential mechanism underlying the identified associations we performed GO and KEGG pathway analysis. mtDNA-CN associated CpGs were annotated with their nearest gene. KEGG analysis identified the *neuroactive ligand-receptor interaction pathway (path:hsa04080)* to be the top overrepresented pathway ($P= 5.24x10^{-12}$, Permuted $P=3.84x10^{-5}$) (Table 3a). Further, GO analyses identified a number of biological processes related to cell signaling and ligand interactions including *Cell-cell signaling* ($P=1.42x10^{-3}$)*, Trans-synaptic signaling*

(*P*=1.88x10^{-3}) *and Synaptic signaling* (*P*=1.88x10^{-3}), among others (Table 3b). These results were confirmed by both permutation testing and through their robustness to ten different associated-CpG cutoffs (cutoff used for final analysis: 300 CpGs).

Validation of CpG associations in independent cohorts

We performed a validation study to replicate findings from the ARIC discovery population in 239 AA and 294 EA participants from the Cardiovascular Health Study (CHS) as well as 1,995 EA participants from the Framingham Heart Study (FHS), for a total of 2,528 individuals (Table 1). 7/34 CpGs identified in the discovery cohort were nominally significant (*P*<0.05 and displaying the same direction of effect as the ARIC cohort results) (Table 2) and the effect sizes from the ARIC results and the validation meta-analysis were largely correlated (R^2=0.36) (Figure 3). Overall, the results were consistent across individual cohorts (Figure S5, Table S2) and analysis of the results from the 34 CpGs across all 3 cohorts (ARIC, CHS and FHS, N=5,035), identify 6 CpGs as validated mtDNA-CN associated CpGs (*P*<5x10^{-8}) (Table 2, Figure S6).

Establishing causality via TFAM knockout

mtDNA-CN is causative of changes in nuclear DNA methylation at loci of interest

To assess if modification of mtDNA-CN drives changes to nuclear DNA methylation we used CRISPR-Cas9 to knock out the *TFAM* gene, which encodes a regulator of mtDNA replication and has been shown to reduce mtDNA-CN[24]. Heterozygous knockout of the *TFAM* gene in HEK293T cells resulted in a 5-fold reduction in the expression of *TFAM*, negligible protein production, and an 18-fold reduction in mtDNA-CN across three biological replicates (Figure 4). We then assayed methylation of the validated mtDNA-CN associated CpGs using the Illumina Infinium Methylation EPIC Beadchip (Table S3). Specifically, the direct assessment of

methylation levels for 4 of the 6 validated mtDNA-CN associated CpGs and one surrogate CpG, as 2 CpGs were not present on the EPIC array and for one missing CpG a reasonable surrogate was not available (see Methods). Reduction of mtDNA-CN in *TFAM* knockout cell lines led to subsequent site-specific changes to DNA methylation for 3 of the 5 EWAS-identified CpGs (nominally significant, $P<0.05$), two of which were in the expected direction of effect (reduced mtDNA-CN led to an increase in methylation) (Table 4, Figure S7). Further, two of the validated mtDNA-CN associated CpGs were differentially methylated even after Bonferroni correction ($P<0.01$) (Table 4). Pyrosequencing was also performed for 3 of the 6 sites (the other sites did not pass pyrosequencing quality control) which confirmed methylation changes at all assayed sites ($P<0.05$) in the expected direction of effect (Table S4).

Global methylation patterns did not show differences between negative control and *TFAM* knockout cell lines suggesting that these differences are site-specific (Figure S8).

mtDNA-CN is causative of changes in nuclear gene expression at nearby genes of interest

The same *TFAM* knockout and negative control cell lines were analyzed for differential gene expression nearby the methylated mtDNA-CN associated CpGs using RNA-seq (Table S5). RNA-seq resulted in expected clustering of knockout and control lines (Figure S9). All nominally differentially expressed genes ($P<0.05$) within 1Mb of the *TFAM* knockout differentially methylated CpGs were identified (Table S6). Five genes nearby the three differentially methylated CpGs were differentially expressed after Bonferroni correction for the number of genes within 1Mb of each CpG ($P<6.41\times10^{-4}$) (Table 5). The five differentially expressed genes were: *IFI35* ($P=3.76\times10^{-5}$) and *RAMP2* ($P=5.51\times10^{-4}$) near cg26094004; *RPIA* near cg26563141 ($P=5.04\times10^{-6}$); and *HLA-DRB5* ($P=6.50\times10^{-7}$) and *MSH5* ($P=2.50\times10^{-4}$) near cg08899667.

These results demonstrate that modulation of mtDNA-CN drives changes in nDNA methylation and gene expression of specific CpGs and transcripts in a cell culture model.

Pathway and biological process analysis of TFAM KO methylation and expression results independently identify pathways identified in cross-sectional analysis

We sought to independently assess the underlying pathways and biological processes that were overrepresented following *TFAM* knockout in our cell-culture model. Specifically, we analyzed the most over-represented terms resulting from GO and KEGG analysis of our full list of differentially methylated CpGs and differentially expressed genes as well as a list of integrated methylation and expression results (Cutoffs used: TFAM Methylation - top 300 differentially methylated CpGs, TFAM Expression – differentially expressed genes (169 genes), TFAM Integrated Methylation/Expression – top 188 genes). The independent results confirmed the findings from our ARIC cross-sectional analysis. Specifically, KEGG analysis of *TFAM* knockout results identified the *neuroactive ligand-receptor interaction pathway (path:hsa04080)* to be the second most overrepresented pathway in the *TFAM* knockout methylation analysis (P=4.41x10^{-4}) and the top overrepresented pathway in the *TFAM* knockout RNA sequencing analysis (P=4.30x10^{-4}) (Table 3a). Accordingly, integration of results from *TFAM* knockout methylation and expression also resulted in strong association with this pathway (P=8.77x10^{-6}). Further, combining of P-values (Fisher's method) across ARIC meta-analysis, *TFAM* knockout methylation and *TFAM* knockout expression analyses yielded a combined P-value of 8.96x10^{-16} for this pathway which was also the top pathway identified in integrated analysis (Table 3a).

The specific genes identified by each analysis to be part of the *neuroactive ligand receptor interaction pathway* were unique to each study (Table S7), with only one gene (*GABRG3*) in common between ARIC analyses and *TFAM* knockout methylation analysis and only one gene (*GABRB1*) in common between *TFAM* knockout methylation and expression analyses (Table S7).

GO analyses of *TFAM* knockout cell lines also confirmed the finding from cross-sectional analysis that biological processes related to cell signaling and ligand interactions including *Cell-cell signaling* (combined $P=7.63\text{x}10^{-8}$), *Trans-synaptic signaling* (combined $P=2.89\text{x}10^{-7}$) *and Synaptic signaling* (combined $P=2.97\text{x}10^{-7}$) were over-represented, among others (Table 3b). These results suggest that mtDNA-CN drives changes to nDNA methylation at sites nearby genes relating to cell signaling processes which in turn may cause gene expression changes to these genes and contribute to disease.

Establishing causality via Mendelian Randomization (MR): Nuclear DNA methylation does not appear to be causative of changes in mtDNA-CN at identified CpGs

Mendelian randomization, a form of instrument variable analysis, was used to further test the direction of causality between mtDNA-CN and nuclear methylation by exploring the relationship between methylation quantitative trait loci (meQTLs) and mtDNA-CN (Table S8). Specifically, if nDNA methylation at our sites of interest is causative of changes in mtDNA-CN, then meQTL SNPs for these CpGs of interest would be expected to also be associated with mtDNA-CN. Alternatively, if mtDNA-CN is not associated with meQTL SNPs, then it would follow that changes to nDNA methylation likely do not drive changes to mtDNA-CN at these CpGs.

We identified 4 independent *cis* meQTLs in the ARIC EA cohort (Permuted $P=7.84\text{x}10^{-4}$) and 6 independent *cis* meQTLs in the ARIC AA cohort (Permuted $P=9.12\text{x}10^{-4}$) across 5 mtDNA-CN associated CpGs for use as an instrument variable for MR (Table S8A). We further identified 2 independent meta-analysis derived meQTLs by combining results from ARIC EA and AA cohorts (Permuted $P=3.97\text{x}10^{-5}$, fixed effects (FE) model) (Table S8B).

We then assessed the relationship between meQTL SNPs and mtDNA-CN. The results of the MR were null for each independent meQTL (Bonferroni $P=0.005$) (Table S8). While our power for a single meQTLvaried depending on the specific meQTL assessed, with power to detect an individual association ranging from 0.18 to 0.99 across the 12 meQTLs, overall power was >99% to detect at least 1 associated meQTL. These results support the experimentaly established direction of causality by suggesting that modification of nDNA methylation at CpG sites of interest does not directly drive alterations in mtDNA-CN.

Association of CpG methylation with mtDNA-CN associated phenotypes

Since decreased mtDNA-CN has been associated with a number of aging-related diseases, and given our hypothesis that mtDNA-CN leads to nDNA methylation changes which influence disease outcomes, associated CpGs should also be associated with mtDNA-CN related phenotypes. To test these associations, we performed linear regression and survival analysis for prevalent and incident diseases, respectively, for each of the 6 validated CpGs as they relate to CHD, CVD, and mortality in the ARIC, FHS and CHS cohorts (Table 6, Table S9). Results from each cohort were meta-analyzed to derive an overall association for each validated CpG with each outcome of interest.

We identify nominally significant phenotype associations with at least one of the mtDNA-CN associated traits of interest for 4 of the 6 validated mtDNA-CN associated CpGs (P<0.05). Specifically, results in the expected direction of effect for prevalent CHD and prevalent CVD were identified for two mtDNA-CN associated CpGs (cg26094004 and cg08899667). Similarly, results in the expected direction of effect were identified for the association between all-cause mortality and cg26563141 and cg08899667. Thus, we found cg08899667 to be associated with three of the five mtDNA-CN associated phenotypes, including all-cause mortality (Table 6).

DISCUSSION

We report evidence that changes in mtDNA-CN influence nDNA methylation at specific, validated loci and lead to changes in gene expression of nearby genes, including those acting in the 'neuroactive ligand receptor interaction' pathway which may impact human health and disease via altered cell signaling. A number of these associations were validated across three independent cohorts and identified both cross-sectionally and experimentally. Interestingly, these associations were found to be site-specific in nature. It is important to note that the methods used to estimate mtDNA-CN differed between the three cohorts with a qPCR based approach used for CHS, a whole-genome sequencing approach for FHS and microarray analysis for ARIC. This may reflect the robustness of results across mtDNA-CN estimation methods and also explain why some but not all CpGs replicated in our validation analysis[25]. We also report that our experimental approach using cell lines replicated some but not all of the cohort validated CpGs. These findings likely reflect both the intrinsic differences between cell line data and cross-sectional data as well as the inherent complexity of mitochondrial-to-nuclear signaling which would be expected to vary across cell-types, developmental timepoints and environmental conditions.

DNA methylation as the link between mtDNA-CN and changes in nuclear gene expression

A symbiotic relationship between the nuclear and mitochondrial genomes has developed in eukaryotes. This relationship strongly implicates communication between the mitochondrial and nuclear genomes as vital for proper cell functioning. Epigenetic mechanisms allow for control of gene expression beyond DNA sequence and have the capacity to be influenced by environmental stimuli. Given the function of the mitochondria in meeting cellular energy demands, mitochondria may play an important role in translating environmental stimuli into epigenetic changes. In addition, mtDNA-CN levels are sensitive to a number of chemicals[26], highlighting the role of mtDNA as an environmental biosensor. Also supporting the notion that bioenergetics are involved in modulating the epigenetic status of the cell is the observation that clinical phenotypes of mitochondrial diseases are strikingly similar to those found in a number of epigenetic diseases such as Angelmans, Rett and Fragile X syndromes[27]. Further, epigenetic changes in nuclear DNA correlate with reduced cancer survival and low mtDNA-CN correlates with poor survival across a number of cancer types[28,29]. Thus, retrograde signals from the mitochondria to the nucleus may be crucial in sensing homeostasis and translating extracellular signals into altered gene expression[18].

Our results implicate the *neuroactive ligand receptor interaction* pathway and in general additional processes involved in cellular signaling. The results also show that although the same pathways are implicated across our independent datasets, the specific genes affected differ between conditions. Interestingly, the *neuroactive ligand receptor interaction* pathway has been identified as having the second highest number of atherosclerosis candidate genes of any KEGG pathway, harboring 53 atherosclerosis candidate genes (272 total genes in the

pathway)[30] . This is an interesting finding given the association of mtDNA-CN with cardiovascular disease[6–8]. Perhaps unsurprisingly, this pathway also belongs to the class of KEGG pathways that are responsible for environmental information processing.

Proposed mechanisms for the methylation of nDNA as a result of changes in mtDNA

The precise identity of the signal(s) coming from the mitochondria that might be responsible for modifying nDNA methylation has not yet been identified and warrants further experimentation. It is likely that metabolite intermediates, non-coding RNA, and/or histones, may play a role in this signaling process. For example, mitochondria-to-nucleus retrograde signaling has been shown to regulate histone acetylation and alter nuclear gene expression through the heterogenous ribonucleoprotein A2 (hnRNAP2)[23]. In fact, histone modifications co-vary with mitochondrial content and are linked with chromatin activation, namely H4K16, H3L4me3 and H3K36me2[31].

The differentially expressed genes identified from the experimental knockout may provide some evidence with regards to the mechanism behind these findings. For example, *IFI35*, a gene involved in Interferon response, is associated with mtDNA-CN through the antiviral innate immune response[32]. Futher, the differentially expressed genes, *RAMP2* and *MSH5*, are known to be related to oxidative phosphorylation protein expression and genome stability, respectively[33,34].

Uncovering the precise nature of this signaling from mitochondria to the nucleus would be expected to expose essential clues that will integrate epigenetic regulation, mitochondrial and genomic polymorphisms, and complex phenotypes. Further assessment of the functional mechanisms underlying the crosstalk between mtDNA-CN, methylation and disease will be

required to fully appreciate the diagnostic and therapeutic utility of the interaction between mtDNA and nDNA as identified in this study.

Influence of findings on complex disease etiology

The observation that differential methylation occurred at specific-sites throughout the nuclear genome as a result of changes to mtDNA-CN, provides an explanation for how mtDNA could alter normal homeostasis as well as susceptibility and/or severity of diseases. It is particularly interesting to note that these changes appear to be site-specific rather than global in nature. The association of mtDNA-CN associated CpGs with mtDNA-CN related disease states lends further support to the hypothesis that modulation of mtDNA-CN not only modifies the nuclear epigenome, and the expression of nearby genes, but does so at locations which may be relevant to disease outcomes, including cardiovascular disease and all-cause mortality. In particular, these observations may explain how mitochondrial-to-nuclear signaling could influence polygenic traits with complex etiology and in particular those for which environmental insults play a role. Together, mitochondrial signaling, and subsequent nDNA methylation, may have an important role in modifying gene expression which may in turn lead to disease outcomes or influence the severity of disease manifestation. Thus, the mechanism(s) by which mtDNA-CN influences disease status may be, at least in part, through modification of nDNA methylation and subsequent modification and/or regulation of nuclear gene expression.

Further, these findings have direct implications for the recent emergence of mitochondrial donation in humans as they suggest that mitochondrial replacement into recipient oocytes may lead to unexpected changes to the nuclear epigenome. Thus, with the recent development of

mitochondrial replacement therapy, unravelling the complex interplay of the mitochondria and nucleus is also critical to properly informing medical decision makers.

This study design had a number of strengths and limitations. A possible limitation of the cross-sectional analysis is the potential for some common factor we have not been able to account for to influence both mtDNA-CN and nDNA methylation. In experimental analysis, we used HEK293T cells for our knockdown studies and we note that the use of a blood cell line may be more relevant to direct interpretation of the results. Further, prevalent disease is subject to reverse causality and therefore the results on prevalent phentoypes should be interpreted with caution. Strengths of this study include the well phenotyped and carefully collected incident disease data, the robustness of the findings across multiple cohorts and ethnic groups, as well as the carefuly quality control employed. Further, our results stood up to rigorous permutation testing which increases the reliability of these observations.

CONCLUSION

Cross-sectionally we have shown that variation in mtDNA-CN is associated with nuclear epigenetic modifications at specific CpGs across multiple independent cohorts. Specifically, six mtDNA-CN associated CpGs were robustly identified across three independent cohorts, three of which were confirmed in experimental analysis. Second, we found meQTL SNPs to not be associated with mtDNA-CN, suggesting that nuclear methylation at these CpGs does not directly cause altered mtDNA-CN. Third, functional results show that modulation of mtDNA-CN causes site-specific changes to nuclear DNA methylation and RNA expression near genes relating to cell signaling processes including those in the *neuroactive-ligand-receptor interaction* pathway. Further, mtDNA-CN associated CpGs display association with mtDNA-CN related phenotypes, namely cardiovascular disease and all-cause mortality. These findings demonstrate that the mechanism(s) by which mtDNA-CN influences disease is at least in part

via regulation of nuclear gene expression through modification of nDNA methylation. Specifically, the data presented here support the model that modification of mtDNA-CN leads to changes to nDNA methylation which in turn influence nuclear DNA expression of nearby genes which contribute to disease pathology. These results have implications for understanding the mechanisms behind mitochondrial and nuclear communication as it relates to complex disease etiology as well as the consequences of mitochondrial replacement therapeutic strategies. Taken together, the results confirm that in elucidating the underpinnings of complex disease, knowledge of only nuclear DNA dynamics is not sufficient to fully elucidating disease etiology.

ONLINE METHODS

A flow chart of general methods can be found in Figure 1.

Ethics

The Atherosclerosis Risk in Communities (ARIC) study, Cardiovascular Health Study (CHS) and Framingham Heart Study (FHS) have been approved by the Institutional Review Board (IRB) at each participating institution. All participants provided written informed consent.

The ARIC study design and methods were approved by four different IRBs at each of the collaborating medical institutions: University of Mississippi Medical Center Institutional Review Board (Jackson Field Center); Wake Forest University Health Sciences Institutional Review Board (Forsyth County Field Center); University of Minnesota Institutional Review Board (Minnesota Field Center); and Johns Hopkins University School of Public Health Institutional Review Board (Washington County Field Center).

FHS is approved by the IRB at Boston University Medical Center. CHS recruited participants from Medicare lists at 4 sites and IRBs at each site were involved in human subjects approval.

Discovery Study Analysis

The Atherosclerosis Risk in Communities Cohort (ARIC)

The ARIC study is a prospective cohort intended for the study of cardiovascular disease in subjects from four communities across the USA: Forsyth County, NC, northwest suburbs of Minneapolis, MN, Jackson, MS, and Washington County, MD[35]. Sample characteristics are available in Table 1. Following quality control, 1,567 African Americans (AA) and 940 European Americans (EA) were used as a discovery cohort. Participants for ARIC EA were derived from two existing projects, Brain MRI (81.7%) and OMICS (18.3%). DNA was extracted from peripheral blood leukocyte samples from visit 2 or 3 using the Gentra Puregene Blood Kit (Qiagen; Valencia, CA, USA) according to the manufacturer's instructions (www.qiagen.com) and hybridized to the Illumina Infinium Human Methylation 450K BeadChip and the Genome-Wide Human SNP Array 6.0.

Estimation of mtDNA-CN from Affymetrix Human SNP 6.0 Arrays

The Affymetrix Genome-Wide Human SNP 6.0 Array was used to estimate mtDNA-CN for each participant as previously described[36]. Briefly, mtDNA copy number (mtDNA-CN) was determined utilizing the Genvisis software package (http://www.genvisis.org). Initially, a list of high-quality mitochondrial SNPs were hand-curated by employing BLAST to remove SNPs without a perfect match to the annotated mitochondrial location and SNPs with off-target matches longer than 20 bp. The probe intensities of the 25 remaining mitochondrial SNPs was determined using

quantile sketch normalization (apt-probeset-summarize) as implemented in the Affymetrix Power Tools software. To correct for DNA quality, DNA quantity, hybridization efficiency and other technical artifacts, surrogate variable analysis was applied to the BLAST filtered, GC corrected LRR of 43,316 autosomal SNPs. These autosomal SNPs were selected based on the following quality filters: call rate >98%, HWE *P*-value >0.00001, PLINK mishap for non-random missingness *P*-value >0.0001, association with sex *P*-value 0.00001, linkage disequilibrium pruning (r^2 <0.30), maximal autosomal spacing of 41.7 kb. The median of the normalized intensity, log R ratio (LRR) for all homozygous calls was GC corrected and used as initial estimates of mtDNA-CN for each sample. The final measure of mtDNA-CN is represented as the standardized residuals from a race-stratified linear regression adjusting the initial estimate of mtDNA-CN for 15 surrogate variables (SVs), age, sex, sample collection site, and white blood cell count. Technical covariates such as DNA quality, DNA quantity, and hybridization efficiency were captured via surrogate variable analysis (SVA) as previously described[7,37].

Illumina Infinium Human Methylation 450K Beadchip Analysis

The Infinium Human Methylation 450K BeadChip was used to determine DNA methylation profiles from blood for >450,000 CpGs across the human genome.

Bisulfite Conversion

Bisulfite conversion of 1 ug genomic DNA was performed using the EZ-96 DNA Methylation Kit (Deep Well Format) (Zymo Research; Irvine, CA, USA) according to the manufacturer's instructions (www.zymoresearch.com). Bisulfite conversion efficiency was determined by PCR amplification of the converted DNA before proceeding with methylation analyses on the Illumina

platform using Zymo Research's Universal Methylated Human DNA Standard and Control Primers.

Normalization and Quality Control

Probes included on the list of cross-reactive 450K probes as reported by Chen *et al* were removed prior to analysis[38]. The cross-reactive target had to match a minimum of 47 bases to be considered cross-reactive. This led to the removal of ~28,000 probes.

Genome studio background correction and BMIQ normalization were performed[39] and the wateRmelon R package was used to conduct QC filtering[40].

Samples were removed for the following reasons: 1. Failed bisulfite conversion, 2. Call rate <95%, 3. Sex mismatch using minfi, 4. Weak correlation between available genotypes and genotypes on 450K array, 5. Weak clustering according to sex in MDS plot, 6. PCA analysis identified them as an outlier (≥4SD from mean), 7. Failed sex check, 8. Sample pass rate <99%, 9. Only sample to pass on a chip. These filtering settings led to the removal of 68 samples in the AA group and 24 samples in the EA group. If samples were run in duplicate, the sample with the lowest missing rate was retained.

Surrogate Variable Analysis (SVA)

SVAs were generated using the package SVA in R and protecting mtDNA-CN[37].

Control Probe Principal Components in ARIC European Americans

The control probe principal components are based on 42 measures, which are transformed from control probes and out-of-band probes in the 450K data[41].

Statistical Analysis

All statistical analyses were performed using R (version 3.3.3).

Linear Mixed Model – Association between mtDNA-CN and nuclear DNA methylation

Linear-mixed-effects regression analysis was performed to determine the association between mtDNA-CN and nuclear DNA methylation at specific CpGs (Table S1).

ARIC AA: Methylation ~ MtDNA-CN + Age + Sex + Site + Visit + Chip Position + Plate + CD8 Count + CD4 Count + B-Cell Count + Monocyte Count + Granulocyte Count + Smoking Status + First 10 Surrogate Variables + Chip (as random effect).

ARIC EA: Same model as ARIC AA but further inclusion of Project (Brain MRI or Omics) as well as the first 10 PCs derived from methylation microarray control probes and the composition of natural killer (NK) cells.

Cell types were imputed using the method of Houseman *et al.*[42]. All correlations were performed using the Pearson method.

Global methylation distributions were assessed by a chi-square test to compare observed to expected site-specific methylation.

Meta-Analysis

A meta-analysis was performed to combine the results from the individual ARIC AA and EA analyses (Table S1). This analysis was done using the standard error scheme implemented in Metal[43]. CpGs had to have a *P*-value cutoff of $P<0.05$ in ARIC AA and EA analyses to be included in the meta-analysis. Associations that met genome-wide significance were included in

subsequent analyses (P=5.0x10^{-8}). 100 meta-analysis permutations were also performed (Permuted P=3.94x10^{-8}).

Residual Bootstrapping

Residual bootstrapping was used to determine the most appropriate genome-wide significance cutoff in ARIC EA and AA cohorts (AA: P<6.22x10^{-8}, EA: P<3.03x10^{-7}). The steps taken were as follows: 1) Residuals were derived from the full model, 2) Fitted values were derived from the null model (model without mtDNA-CN as independent factor), 3) The residuals from Step 1 were resampled and added to the fits from Step 2, 4) Each resulting matrix from Step 3 was run as pseudonull input in the formula lme(pseudonull~CN+covariates) to refit the full model and obtain null statistics, 4a) The most extreme P-value was pulled from each iteration, 4b) The resulting 100 most extreme P-values were ranked from least to most significant and the 95th value was chosen to be the 'genome-wide significance level' for the corresponding cohort. Additionally, the qq plots show minimal inflation in ARIC AA, EA and meta-analysis (Figure S1).

Significant CpGs with high correlation (R^2≥0.6) were identified as non-independent and the CpGs with the more significant P-value was retained. Highly correlated CpGs were consistent between AA and EA results, specifically these CpGs were cg21051031 and cg03964851 (R^2: AA=0.62, EA=0.63) and cg06809544 and cg13393978 (R^2: AA=0.65, EA=0.70).

Validation Cohorts

The Cardiovascular Health Study (CHS)

The CHS is a population-based cohort study of risk factors for coronary heart disease and stroke in adults ≥65 years conducted across four field centers[44]. The original predominantly European ancestry cohort of 5,201 persons was recruited in 1989-1990 from random samples of the Medicare eligibility lists; subsequently, an additional predominantly African-American cohort of 687 persons was enrolled in 1992-1993 for a total sample of 5,888. The validation cohort includes 239 AA participants and 294 EA participants from CHS with mtDNA-CN and 450K methylation derived from the same visit (Table 1).

mtDNA-CN Estimation using Quantitative PCR

mtDNA copy number (mtDNA-CN) was determined utilizing a multiplexed real time quantitative polymerase chain reaction (qPCR) assay with ABI TaqMan chemistry (Applied Biosystems) as previously described[7]. Briefly, each well consisted of a VIC-labeled, primer limited assay specific to a mitochondrial target (ND1), and a FAM-labeled assay specific to a region of the nuclear genome selected for being non-repetitive (RPPH1). Each sample was run in triplicate on a 384 well plate in a 10μL reaction containing 20ng of DNA. The cycle threshold (Ct) value was determined from the amplification curve for each target by the ABI Viia7 software. A ΔCT value was computed for each well as the difference between the Ct for the RPPH1 target and the Ct for the ND1 target, as a measure of mtDNA copy number relative to nuclear DNA copy number. For samples with a standard deviation of ΔCT for the three replicates >0.5, an outlier replicate was identified and excluded. If the ΔCT standard deviation remained >0.5 after exclusion, the sample was completely excluded from future analyses. Replicates with Ct values for ND1 > 28, Ct values for RPPH1 > 5 standard deviations from the mean, or ΔCT values >3 standard deviations from the mean of the plate were removed. Additionally, due to an observed linear increase in ΔCT value by the order in which the replicate was pipetted onto the plate, a linear regression was used to correct for pipetting order. Plate effects are controlled for by performing

a linear regression whereby the plate a sample is run on is treated as a random effect. The final measure of mtDNA-CN is represented as the standardized residuals from a race-stratified mixed linear regression adjusting for age, sex, and sample collection site.

Methylation Analysis

Methylation measurements were performed at the Institute for Translational Genomics and Population Sciences at the Harbor-UCLA Medical Center Institute for Translational Genomics and Population Sciences (Los Angeles, CA). DNA was extracted from Buffy coat fractions and subsequently underwent bisulfite conversion using the EZ DNA Methylation kit (Zymo Research, Irvine, CA). Methylation was then assayed using the Infinium HumanMethylation450 BeadChip (Illumina Inc, San Diego, CA).

Quality control was performed in in the minfi R package[45] (version 1.12.0, http://www.bioconductor.org/packages/release/bioc/html/minfi.html). Samples with low median intensities of below 10.5 (log2) across the methylated and unmethylated channels, samples with a proportion of probes falling detection of greater than 0.5%, samples with QC probes falling greater than 3 standard deviation from the mean, sex-check mismatches, failed concordance with prior genotyping or > 0.5% of probes with a detection *P*-value > 0.01 were removed. Probes with >1% of values below detection were removed. In total, 11 samples were removed for sample QC resulting in a sample of 323 European-ancestry and 326 African-American samples. Methylation values were normalized using the SWAN quantile normalization method[46]. Since white blood cell proportions were not directly measured in CHS they were estimated from the methylation data using the Houseman method[42].

Regression Analysis

CHS was analysed using linear regression with methylation beta values as the dependent variable and mtDNA-CN as the independent variable. Analyses were adjusted for age, sex, batch, measured white blood cell count and estimated cell type counts.

The Framingham Heart Study (FHS)

FHS is a prospective study of individuals from Framingham, Massachusetts[47]. The validation cohort includes 1,995 EA participants from FHS with mtDNA-CN and 450K methylation derived from the same visit (Table 1).

mtDNA-CN Estimation from Whole Genome Sequencing

Cohort-specific mtDNA-CN residuals were obtained by regressing mtDNA-CN on age, sex, and WBC counts. Mitochondrial DNA copy number was estimated by applying the fastMitoCalc software[48] to harmonized build 37 mappings of TOPMed deep whole genome sequencing data (freeze 5). The estimated mitochondrial copy number is twice the ratio of average mitochondrial sequencing depth to average autosomal sequencing depth. We applied inverse normal transformation to mtDNA-CN residuals.

Methylation Analysis

DNA extraction, methylation quantification (450k-BeadChip), and QC were detailed previously[49]. We obtained lab-specific and cohort-specific DNA methylation residuals by regressing methylation beta values on age, sex, batch effects (plate, col, row), and WBC counts. We applied inverse normal transformation to DNA methylation residuals.

Regression Analysis

A linear mixed model was applied with inverse normal transformed DNA methylation residuals as the dependent variable and inverse normal transformed mtDNA-CN residuals as the independent variable, accounting for family structure.

Validation and all-cohort meta-analyses

A meta-analysis was performed of all validation cohorts (FHS EA, CHS EA, CHS AA). We also performed an all-cohort meta-analysis (ARIC AA, ARIC EA, FHS EA, CHS EA, CHS AA). Both meta-analyses were performed using the standard error scheme implemented in Metal[43].

Mendelian Randomization

meQTL Analysis

meQTLs were identified using MatrixEQTL[50]. Imputed genotypes which were previously derived from ARIC for the relevant participants as well as normalized residuals from our 450K methylation dataset were used in regression analysis. Haplotype phasing was performed using ShapeIt[51] and imputation was performed using IMPUTE2[52]. SNPs were filtered for allele frequency >0.05, and imputation quality >0.4. Genotypes were imputed to the 1000G reference panel (Phase I, version 3). The same covariates used for the ARIC EWAS analysis were used to call meQTLs as well as the addition of genotyping PCs (4 for EA, 10 for AA). Only meQTLs which had an individual cohort *P* value >0.05 were included in the meta-analysis.

A linear model was used for MatrixEQTL and a cis meQTL was defined as having a distance less than 100 kb. Only cis meQTLs derived from the 6 CpGs of interest and which met a cohort-specific permuted *P*-value cutoff (Permuted *P:* EA=7.84×10^{-4}, AA=9.12×10^{-4}) or a permuted

meta-analysis *P*-value cutoff (Permuted *P*, fixed effects (FE) model: 3.97×10^{-5}) were retained for use in Mendelian randomization. Metasoft[53] was used for meta-anlaysis; in addition to the fixed effects (FE) model, a random effects (RE) and Han and Eskin's Random Effects model (RE2) were also used and yielded very similar results (Table S8).

Mendelian Randomization Methods

Independent meQTLs were used for MR. Independence was defined by including SNPs in the same linear model. MR with mtDNA-CN as the outcome and methylation as the exposure was undertaken. meQTLs served as the known relationship of genotype on exposure (methylation) and the results of the linear model, lm(mtDNA~meQTL SNP) were calculated. Power for the MR was calculated using the YZ association function in mRnd[54].

Phenotype Analysis

We compared methylation at the 6 validated CpGs to phenotypes that are known to be associated with mtDNA-CN. Phenotypes included prevalent diseases (CHD, CVD) as well as incident diseases (CHD, CVD, Mortality). The analysis was performed as follows for each cohort:

A) Prevalent diseases (CHD, CVD): glm(PRVCVD ~ resids(methyl) + AGE + SEX + CENTER + RACE, family=binomial(logit))
B) Incident diseases (CHD, CVD, Mortality): coxph(Surv(STime,dead) ~ resids(methyl) + AGE + SEX + CENTER + RACE))

Where resids(methyl) represents methylation adjusted for all relevant covariates from the EWAS. The event adjudication process in ARIC, CHS and FHS consisted of expert committee

review of hospital records, telephone interviews, and death certificates. In addition, adjudicated events between visit 1 and the baseline visit for this study were considered prevalent events.

Analyses of prevalent and incident events in CHS were adjusted for age, sex, clinic site and batch.

In ARIC, prevalent coronary heart disease (CHD) was defined as history of myocardial infarction (MI) or cardiac procedures (heart or arterial surgery, coronary bypass, or angioplasty). Cardiovascular disease (CVD) was defined as either CHD or stroke. Prevalent stroke was defined as stroke at baseline. For all phenotypes, prevalent disease was a combination of self-report at visit 1 plus adjudicated events between visit 1 and the baseline visit. Incident CHD was defined as the first incident MI or death owing to CHD. Incident stroke was defined as the first nonfatal stroke or death owing to stroke. In ARIC, the mean follow-up time was 20.6 years in the EA cohort and 18.1 years in the AA cohort. Follow-up for incident events was administratively censored at December 31, 2016.

CHS and FHS followed similar phenotype definitions as ARIC. For FHS, the mean follow-up time was 6.0 years and individuals were removed if follow-up years equaled 0, FHS events were adjudicated through 12/2016. In CHS, prevalent CVD/CHD was excluded during sampling and events were adjudicated through June 30, 2015. The follow up time for incident events from the time of methylation measurement was 23 years.

Results from each of the 5 individual cohorts were meta-analyzed across cohorts using an inverse weighted standard error method[43] to derive an overall phenotype association for each CpG of interest.

CRISPR-Cas9 Knockout of TFAM

Generation of *TFAM* Knockout

The stable *TFAM* CRISPR-Cas9 knockout was generated in HEK293T cells using the Origene *TFAM* – Human Gene Knockout Kit via CRISPR (catalog number: KN215488) following the manufacturer's protocol. The following sgRNA guide sequence was used to generate the stable *TFAM* knockout lines: GCGTTTCTCCGAAGCATGTG. Lipofection was conducted using Turbofectin 8.0 (catalog number: TF81001). Puromycin was used for selection at a concentration of 1.5 μg/mL. Fluorescence-activated cell sorting (FACS) was used for single cell sorting and clonal expansion. HEK293T cells were grown in DMEM containing 10% FBS and 1% penicillin-streptomycin at 37°C and 5% CO2. Sequencing primers used to confirm the *TFAM* knockout and proper insertion of the Donor plasmid are as follows: *TFAM*_Left_Forward_Primer_2: AGCGACTGTGGACAACTAGC, GFP_Reverse_Primer_2: TCATCTTGTTGGTCATGCGG, Puro-Forward_Primer_1: CACAACCTCCCCTTCTACGAG, *TFAM*_Right_Reverse_Primer_1: CCCCAAACTCCTTACCTGGG.

DNA Isolation

DNA extraction was performed on harvested HEK293T *TFAM* knockout cells using the AllPrep DNA/RNA Mini Kit (Qiagen #80204) following the manufacturer's protocol. DNA was eluted in 100 μL ultrapure water. DNA was quantified using a Nanodrop 1000. Low purity samples were subjected to ethanol precipitation.

RNA Isolation

Total RNA was extracted from confluent T75 culture flasks of *TFAM* CRISPR Negative Control and KO cell lines (p32) using the AllPrep DNA/RNA/Protein Kit (Qiagen #80004). RNA was extracted using the provided kit manual/instructions for RNA extraction, except all microcentrifuge spins were performed at 10,000 x g. RNA was eluted twice in 50 uL molecular biology grade water and stored in a -80C freezer.

mtDNA-CN Estimation on TFAM Knockout Cell Lines

qPCR was used to measure mtDNA-CN as described above for CHS in section "mtDNA-CN Estimation using Quantitative PCR".

TFAM Expression Assay

cDNA synthesis was performed with the SuperScript III First-Strand Synthesis System for RT-PCR (ThermoFisher #18080-051) following the manufacturer's protocol. 1.5 μg of total RNA from each cell line was used as input and primed with 50 ng random hexamers using the appropriate incubation conditions from the manufacture's protocol. Following completed cDNA synthesis, samples were quantified using the Qubit ssDNA assay kit (Invitrogen #Q10212) and Qubit 2.0 Fluorometer. Synthesized cDNA was then diluted to 10 ng/μL using ultrapure water and stored in -20^{o}C.

qPCR to determine TFAM gene expression for *TFAM* KO

20 ng of synthesized cDNA from each cell line was used as input for a 10 μL volume reaction. *TFAM* cDNA were amplified using TaqMan probe Hs00273372_s1 (20x, FAM-labeled, Applied Biosystems #4331182). GAPDH cDNA served as a housekeeping reference control and was amplified with probe Hs03929097_g1 (20x, VIC-labeled, Applied Biosystems #4448489). Both probes were multiplexed together and all qPCR reactions were conducted at 50^{o} C for 2 min,

95°C for 10 min, and then 40 cycles of 95°C for 15 sec and 60°C for 1 min. Expression fold change was determined using double delta cycle threshold using GAPDH as the housekeeping reference control.

Total Protein Extraction

Total protein lysates from HEK293T *TFAM* CRISPR knockout cell lines were extracted using ice-cold radioimmunoprecipitation assay buffer (RIPA) buffer supplemented with Halt Protease and Phosphatase Inhibitor Cocktail (Thermo Scientific #78440). Protein concentrations were quantified using the Pierce BCA Protein Assay Kit (Thermo Scientific #23227) and lysates were stored at -80°C.

Western Blotting

Equal amounts of each lysate were diluted 1:1 with 2x Laemmli Sample Buffer (Bio-Rad #161-0737) supplemented with 5% β-mercaptoethanol. Samples were then heated at 95°C for 5 minutes to denature the proteins. 30 μg of each protein lysate was separated on a 12% polyacrylamide Mini-PROTEAN TGX Gel (Bio-Rad #456-1044) and then transferred to a PVDF membrane (Bio-Rad #1704156) using the Trans-Blot Turbo Transfer System. The membrane was blocked overnight at 4° C in Tris-Buffered Saline and Tween 20 (TBST) containing 5% nonfat milk with gentle shaking. After blocking, the membrane was incubated with rabbit anti-*TFAM* primary antibody diluted 1:2000 in 5% milk (Abcam #ab131607) and rabbit anti- β-Tubulin primary antibody diluted 1:3000 in 5% milk (Invitrogen #PA5-27552) for 1 hour at room temperature with gentle shaking. The membrane was washed 5-times with TBST after primary antibody incubation, then incubated with goat anti-rabbit secondary antibody conjugated with horseradish peroxidase (1:20,000 dilution, Abcam #ab97080) in the dark for 1 hour at room temperature with shaking. Signals were visualized by enhanced chemiluminescent substrate

(SuperSignal West Pico PLUS, Thermo Scientific #34577) and photographed digitally using the ChemiDoc-It^2 Imager.

Methylation Analysis of *TFAM* Knockout Lines

TFAM KO cell lines were hybridized to the Illumina Infinium EPIC BeadChip at The University of Texas Health Science Center at Houston (UTHealth). Bisulfite conversion efficiency was reviewed in the laboratory using the Bead Array Controls Reporter (BACR) tool, and Illumina chemistry (sample independent controls) performed within acceptable specifications. All samples passed with detected CpG (0.01) >97%.

EPIC BeadChip analysis was performed using the minfi package[55]. Data was normalized using Functional Normalization[41] and differential methylation was calculated using the dmpFinder function in minfi (Table S3).

In the cases where the CpG from the 450k array was not represented on the EPIC array a CpG surrogate was chosen if there was a nearby CpG within 1000 bp upstream or downstream of the original CpG that was highly correlated with the original CpG ($R^2 \geq 0.6$) and associated with mtDNA-CN in the ARIC analysis ($P<5x10^{-8}$).

RNA sequencing of TFAM Knockout Lines

RNA Preparation

RNA quantification was performed using the Qubit RNA BR Assay (Invitrogen #Q10211) and Qubit 2.0 Fluorometer. The Agilent BioAnalyzer was used for quality control of the RNA prior to library creation, with a minimum RIN of 8.5. Samples were diluted to 300 ng/uL in 12 uL molecular

biology grade water, and then submitted to the Genetic Resources Core Facility for RNA sequencing.

Library Preparation and Sequencing

Illumina's TruSeq Stranded Total RNA kit protocol was used to generate libraries. Specifically, total RNA is converted to cDNA and size selected to 150 to 200 bp in length with 3' or 5' overhangs. End repair is performed where 3' to 5' exonuclease activity of enzymes removes 3' overhangs and the polymerase activity fills in the 5' overhangs. An 'A' base is then added to the 3' end of the blunt phosphorylated DNA fragments which prepares the DNA fragments for ligation to the sequencing adapters, which have a single 'T' base overhang at their 3' end. Ligated fragments are subsequently size selected through purification using SPRI beads and undergo PCR amplification techniques to prepare the 'libraries'. The BioAnalyzer is used for quality control of the libraries to ensure adequate concentration and appropriate fragment size. The resulting library insert size is 120-200 bp with a median size of 150 bp. Libraries were uniquely barcoded and pooled for sequencing. DNA sequencing was performed in duplicate on an Illumina® HiSeq 2500 instrument using standard protocols for paired end 150 bp sequencing. As per Illumina's recommendation, 3% PhiX was added to each lane as a control, and to assist the analysis software with any library diversity issues.

Primary Analysis

Illumina HiSeq reads were processed through Illumina's Real-Time Analysis (RTA) software generating base calls and corresponding base call quality scores. CIDRSeqSuite 7.1.0 was used to convert compressed bcl files into compressed fastq files.

Secondary Analysis

Each independent cell-line was sequenced twice. RNA sequencing fastq files were pseudoaligned to Genome Reference Consortium Human Build 37 (GRCh37) using Kallisto[56]. 100 bootstraps were performed using Kallisto. The R package Sleuth was used for RNA sequencing analysis[57] (Table S5). Lane was included as a covariate in the Sleuth model. Differentially expressed genes were defined as those with a $P<0.05$.

Integrated analysis of *TFAM* knockout methylation and expression

The linear-gwis method in FAST (genotype mode) was used to collapse *TFAM* KO methylation data into one gene level *P*-value per gene[58]. These gene-level methylation results were combined with gene-level gene expression results for the same gene using the Fisher *P*-value combination method to generate an integrated gene level Methylation/RNA sequencing *P*-value.

GO/KEGG Analysis

Each CpG was annotated with the nearest gene as defined by the closest gene which harbors the CpG within 1,500 bp of the transcriptional start site and extending to the polyA signal. A bias exists when performing gene set analysis for genome-wide methylation data that occurs due to the differing numbers of CpG sites profiled for each gene[59]. Due to this, we used gometh for GO and KEGG analysis since it is based off of the goseq method which accounts for this bias[60]. We analyzed our individual ARIC/*TFAM* datasets as well as our *TFAM* integrated (meth/expression) dataset. We also combined GO/KEGG results for ARIC, TFAM methylation and TFAM RNA sequencing using the Fisher *P*-value combination method to generate an overall combined *P*-value for each term. 10 stepwise cutoffs ranging from 75 CpGs to 300 CpGs were performed to ensure robustness of results. Final *P*-value cutoffs used for each analysis were as follows: ARIC Meta-Analysis (300 CpGs, $P=5.24\times10^{-12}$), TFAM Methylation (300 CpGs, $P=4.41\times10^{-4}$), TFAM

Expression (169 genes, P=4.30x10^{-4}), TFAM Integrated (Methylation/Expression) (188 genes, P=8.77x10^{-6}).

All statistical analyses were performed using R (version 3.3.3).

ACKNOWLEDGEMENTS

Infinium Methylation EPIC BeadChip array hybridization was performed at the UTHealth Human Genetics Center, The University of Texas, Houston, TX. Illumina sequencing was conducted at the Genetic Resources Core Facility, Johns Hopkins Institute of Genetic Medicine, Baltimore, MD. This research was supported by grant R01HL131573 from the US National Institutes of Health. Castellani was supported by a CIHR Postdoctoral Fellowship.

ARIC Acknowledgements

The Atherosclerosis Risk in Communities study has been funded in whole or in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, Department of Health and Human Services (contract numbers HHSN268201700001I, HHSN268201700002I, HHSN268201700003I, HHSN268201700004I and HHSN268201700005I). The authors thank the staff and participants of the ARIC study for their important contributions. Funding was also supported by 5RC2HL102419 and R01NS087541.

CHS Acknowledgements

Infrastructure for the CHARGE Consortium is supported in part by the National Heart, Lung, and Blood Institute grant R01HL105756. The CHS research was supported by NHLBI contracts HHSN268201200036C, HHSN268200800007C, HHSN268201800001C, N01HC55222, N01HC85079, N01HC85080, N01HC85081, N01HC85082, N01HC85083, N01HC85086; and

NHLBI grants U01HL080295, U01HL130114, K08HL116640, R01HL087652, R01HL092111, R01HL103612, R01HL103612, R01HL111089, R01HL116747 and R01HL120393 with additional contribution from the National Institute of Neurological Disorders and Stroke (NINDS). Additional support was provided through R01AG023629 from the National Institute on Aging (NIA), Merck Foundation / Society of Epidemiologic Research as well as Laughlin Family, Alpha Phi Foundation, and Locke Charitable Foundation. A full list of principal CHS investigators and institutions can be found at CHS-NHLBI.org. The provision of genotyping data was supported in part by the National Center for Advancing Translational Sciences, CTSI grant UL1TR000124, and the National Institute of Diabetes and Digestive and Kidney Disease Diabetes Research Center (DRC) grant DK063491 to the Southern California Diabetes Endocrinology Research Center. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

FHS Acknowledgements

Whole genome sequencing (WGS) for the Trans-Omics in Precision Medicine (TOPMed) program was supported by the National Heart, Lung and Blood Institute (NHLBI). WGS for "NHLBI TOPMed: Whole Genome Sequencing and Related Phenotypes in the Framingham Heart Study" (phs000974.v1.p1) was performed at the Broad Institute of MIT and Harvard (HHSN268201500014C). Centralized read mapping and genotype calling, along with variant quality metrics and filtering were provided by the TOPMed Informatics Research Center (3R01HL-117626-02S1). Phenotype harmonization, data management, sample-identity QC, and general study coordination, were provided by the TOPMed Data Coordinating Center (3R01HL-120393-02S1). We gratefully acknowledge the studies and participants who provided biological samples and data for TOPMed.

This work is supported by National Institutes of Health (NIH) contract N01-HC-25195 and HHSN268201500001I and grant R01 HL092577, also supported by intramural funding of Dan Levy, National Heart, Lung, and Blood Institute (NHLBI) (for DNA methylation profiling), and Trans-Omics for Precision Medicine (TOPMed) sponsored by NHLBI/NIH. The Framingham Heart Study thanks the study participants and the multitude of investigators who over its 70 year history continue to contribute so much to further our knowledge of heart, lung, blood and sleep disorders and associated traits.

The views expressed in this manuscript are those of the authors and do not necessarily represent the views of the National Heart, Lung, and Blood Institute; the National Institutes of Health; or the U.S. Department of Health and Human Services.

AUTHOR CONTRIBUTIONS

Concept and design: Castellani, Guallar, Pankratz, O'Rourke, Coresh, Arking.
Acquisition, analysis, or interpretation of data: Castellani, Longchamps, Newcomb, Sumpter, Lane, Brody, Bartz, Grove, Fornage, Floyd, Bressler, Pankow, Tin, O'Rourke, Guallar, Pankratz, Taylor, Wang, Liu, Boerwinkle, Arking.
Drafting of the manuscript: Castellani, Arking.
Critical revision of the manuscript for important intellectual content: Castellani, Longchamps, Floyd, Liu, Tin, Fornage, O'Rourke, Brody, Pankow, Bartz, Arking.
Statistical analysis: Castellani, Longchamps, Lane, Brody, Liu, Guallar, Pankratz, Arking.
Obtained funding: Coresh, Guallar, Boerwinkle, Arking.
Administrative, technical, or material support: Newcomb, Sumpter, Grove, Bressler.
Supervision: Sotoodehnia, Levy, Guallar, Arking.

COMPETING INTERESTS STATEMENT

The authors declare no competing interests.

REFERENCES

1. Clayton, D. A. Transcription and replication of animal mitochondrial DNAs. *Int. Rev. Cytol.* **141**, 217–232 (1992).
2. 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).
3. Wai, T. *et al.* The role of mitochondrial DNA copy number in mammalian fertility. *Biol. Reprod.* **83**, 52–62 (2010).
4. Guha, M. & Avadhani, N. G. Mitochondrial retrograde signaling at the crossroads of tumor bioenergetics, genetics and epigenetics. *Mitochondrion* **13**, 577–591 (2013).
5. Jeng, J.-Y. *et al.* Maintenance of mitochondrial DNA copy number and expression are essential for preservation of mitochondrial function and cell growth. *J. Cell. Biochem.* **103**, 347–357 (2008).
6. Dai, D.-F., Rabinovitch, P. S. & Ungvari, Z. Mitochondria and cardiovascular aging. *Circ. Res.* **110**, 1109–1124 (2012).
7. Ashar, F. N. *et al.* Association of Mitochondrial DNA Copy Number With Cardiovascular Disease. *JAMA Cardiol* **2**, 1247–1255 (2017).
8. Chen, S. *et al.* Association between leukocyte mitochondrial DNA content and risk of coronary heart disease: A case-control study. *Atherosclerosis* **237**, 220–226 (2014).
9. Tin, A. *et al.* Association between Mitochondrial DNA Copy Number in Peripheral Blood and Incident CKD in the Atherosclerosis Risk in Communities Study. *J Am Soc Nephrol* **27**, 2467–2473 (2016).

10. Ashar, F. N. *et al.* Association of mitochondrial DNA levels with frailty and all-cause mortality. *J. Mol. Med.* **93**, 177–186 (2015).

11. Crovetto, F. *et al.* A role for mitochondria in gestational diabetes mellitus? *Gynecol. Endocrinol.* **29**, 259–262 (2013).

12. Sookoian, S. *et al.* Epigenetic regulation of insulin resistance in nonalcoholic fatty liver disease: Impact of liver methylation of the peroxisome proliferator–activated receptor γ coactivator 1α promoter. *Hepatology* **52**, 1992–2000 (2010).

13. Pirola, C. J. *et al.* Epigenetic Modifications in the Biology of Nonalcoholic Fatty Liver Disease: The Role of DNA Hydroxymethylation and TET Proteins. *Medicine (Baltimore)* **94**, e1480 (2015).

14. Delsite, R., Kachhap, S., Anbazhagan, R., Gabrielson, E. & Singh, K. K. Nuclear genes involved in mitochondria-to-nucleus communication in breast cancer cells. *Mol. Cancer* **1**, 6 (2002).

15. Kitamura, E. *et al.* Analysis of tissue-specific differentially methylated regions (TDMs) in humans. *Genomics* **89**, 326–337 (2007).

16. Horan, M. P. & Cooper, D. N. The emergence of the mitochondrial genome as a partial regulator of nuclear function is providing new insights into the genetic mechanisms underlying age-related complex disease. *Hum. Genet.* **133**, 435–458 (2014).

17. Cagin, U. & Enriquez, J. A. The complex crosstalk between mitochondria and the nucleus: What goes in between? *Int. J. Biochem. Cell Biol.* **63**, 10–15 (2015).

18. Vivian, C. J. *et al.* Mitochondrial Genomic Backgrounds Affect Nuclear DNA Methylation and Gene Expression. *Cancer Res.* **77**, 6202–6214 (2017).

19. Latorre-Pellicer, A. *et al.* Mitochondrial and nuclear DNA matching shapes metabolism and healthy ageing. *Nature* **535**, 561–565 (2016).

20. Bellizzi, D., D'Aquila, P., Giordano, M., Montesanto, A. & Passarino, G. Global DNA methylation levels are modulated by mitochondrial DNA variants. *Epigenomics* **4**, 17–27 (2012).

21. Smiraglia, D. J., Kulawiec, M., Bistulfi, G. L., Gupta, S. G. & Singh, K. K. A novel role for mitochondria in regulating epigenetic modification in the nucleus. *Cancer Biol. Ther.* **7**, 1182–1190 (2008).

22. Xie, C. *et al.* Mitochondrial regulation of cancer associated nuclear DNA methylation. *Biochem. Biophys. Res. Commun.* **364**, 656–661 (2007).

23. Guha, M. *et al.* HnRNPA2 is a novel histone acetyltransferase that mediates mitochondrial stress-induced nuclear gene expression. *Cell Discov* **2**, 16045 (2016).

24. Ekstrand, M. I. *et al.* Mitochondrial transcription factor A regulates mtDNA copy number in mammals. *Hum. Mol. Genet.* **13**, 935–944 (2004).

25. Longchamps, R. J. *et al.* Evaluation of mitochondrial DNA copy number estimation techniques. *bioRxiv* 610238 (2019). doi:10.1101/610238

26. Roubicek, D. A. & Souza-Pinto, N. C. de. Mitochondria and mitochondrial DNA as relevant targets for environmental contaminants. *Toxicology* **391**, 100–108 (2017).

27. Wallace, D. C. A mitochondrial bioenergetic etiology of disease. *J Clin Invest* **123**, 1405–1412 (2013).

28. Munot, K. *et al.* Pattern of expression of genes linked to epigenetic silencing in human breast cancer. *Hum. Pathol.* **37**, 989–999 (2006).

29. Yu, M. *et al.* Reduced mitochondrial DNA copy number is correlated with tumor progression and prognosis in Chinese breast cancer patients. *IUBMB Life* **59**, 450–457 (2007).

30. Timinskas, A., Au, Z. & Ku, V. Atherosclerosis: alterations in cell communication. 6 (2007).

31. Guantes, R. *et al.* Global variability in gene expression and alternative splicing is modulated by mitochondrial content. *Genome Res.* **25**, 633–644 (2015).

32. West, A. P. *et al.* Mitochondrial DNA stress primes the antiviral innate immune response. *Nature* **520**, 553–557 (2015).

33. Shindo, T. *et al.* Regulation of cardiovascular development and homeostasis by the adrenomedullin-RAMP system. *Peptides* **111**, 55–61 (2019).

34. Bannwarth, S. *et al.* The human MSH5 (MutSHomolog 5) protein localizes to mitochondria and protects the mitochondrial genome from oxidative damage. *Mitochondrion* **12**, 654–665 (2012).

35. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators. *Am. J. Epidemiol.* **129**, 687–702 (1989).

36. MitoPipeline, G. MitoPipeline: Generating Mitochondrial copy number estimates from SNP array data in Genvisis. (2018). Available at: http://genvisis.org/MitoPipeline/. (Accessed: 27th November 2017)

37. Leek, J. T., Johnson, W. E., Parker, H. S., Jaffe, A. E. & Storey, J. D. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics* **28**, 882–883 (2012).

38. Chen, Y. *et al.* Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray. *Epigenetics* **8**, 203–209 (2013).

39. Teschendorff, A. E. *et al.* A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* **29**, 189–196 (2013).

40. Pidsley, R. *et al.* A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics* **14**, 293 (2013).

41. Fortin, J.-P. *et al.* Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol* **15**, (2014).

42. Houseman, E. A., Molitor, J. & Marsit, C. J. Reference-free cell mixture adjustments in analysis of DNA methylation data. *Bioinformatics* **30**, 1431–1439 (2014).

43. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).

44. Fried, L. P. *et al.* The Cardiovascular Health Study: design and rationale. *Ann Epidemiol* **1**, 263–276 (1991).

45. Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30**, 1363–1369 (2014).

46. Maksimovic, J., Gordon, L. & Oshlack, A. SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome Biol.* **13**, R44 (2012).

47. Dawber, T. R., Meadors, G. F. & Moore, F. E. Epidemiological approaches to heart disease: the Framingham Study. *Am J Public Health Nations Health* **41**, 279–281 (1951).

48. Ding, J. *et al.* Assessing Mitochondrial DNA Variation and Copy Number in Lymphocytes of ~2,000 Sardinians Using Tailored Sequencing Analysis Tools. *PLoS Genet.* **11**, e1005306 (2015).

49. Joehanes Roby *et al.* Epigenetic Signatures of Cigarette Smoking. *Circulation: Cardiovascular Genetics* **9**, 436–447 (2016).

50. Shabalin, A. A. Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics* **28**, 1353–1358 (2012).

51. Delaneau, O., Zagury, J.-F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods* **10**, 5–6 (2013).

52. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G. R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.* **44**, 955–959 (2012).

53. Han, B. & Eskin, E. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am. J. Hum. Genet.* **88**, 586–598 (2011).

54. Brion, M.-J. A., Shakhbazov, K. & Visscher, P. M. Calculating statistical power in Mendelian randomization studies. *Int J Epidemiol* **42**, 1497–1501 (2013).

55. Fortin, J.-P., Triche, T. J. & Hansen, K. D. Preprocessing, normalization and integration of the Illumina HumanMethylationEPIC array with minfi. *Bioinformatics* **33**, 558–560 (2017).

56. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).

57. Pimentel, H., Bray, N. L., Puente, S., Melsted, P. & Pachter, L. Differential analysis of RNA-seq incorporating quantification uncertainty. *Nat. Methods* **14**, 687–690 (2017).

58. Huang, H., Chanda, P., Alonso, A., Bader, J. S. & Arking, D. E. Gene-Based Tests of Association. *PLOS Genetics* **7**, e1002177 (2011).

59. Geeleher, P. *et al.* Gene-set analysis is severely biased when applied to genome-wide methylation data. *Bioinformatics* **29**, 1851–1857 (2013).

60. Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol.* **11**, R14 (2010).

Figure 1. Flow chart of methods.

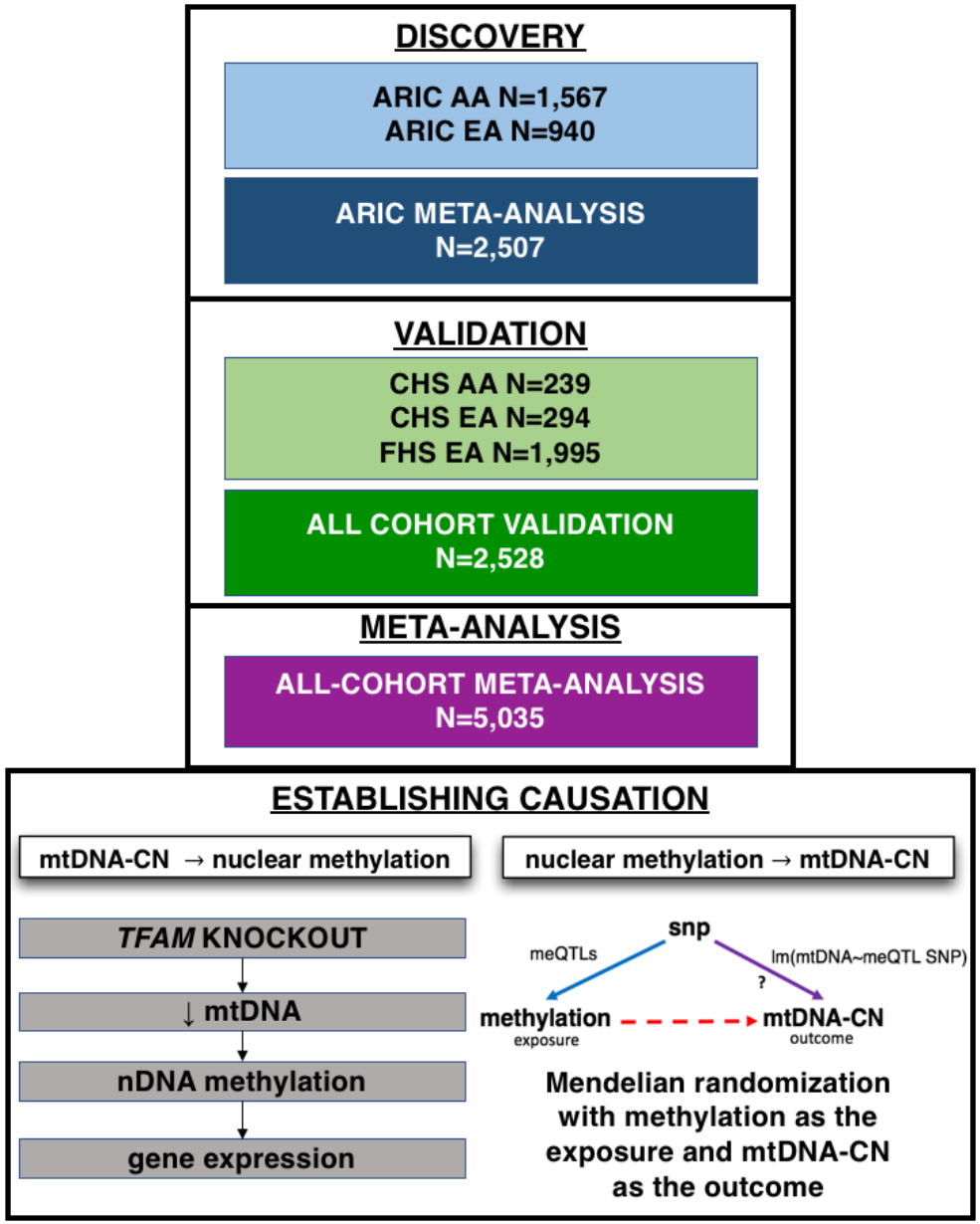

Figure 2. ARIC Meta-Analysis (AA and EA) Results. 34 Independent genome-wide significant CpGs were identified in ARIC meta-analysis to be associated with mtDNA-CN (red dots). Blue dotted line represents genome-wide significance cutoff (P=5x10^{-8}). CpGs had to be independent and nominally significant in both cohorts (P<0.05), as well as meet the meta-analysis significance cutoff (P=5x10^{-8}) to be considered significant.

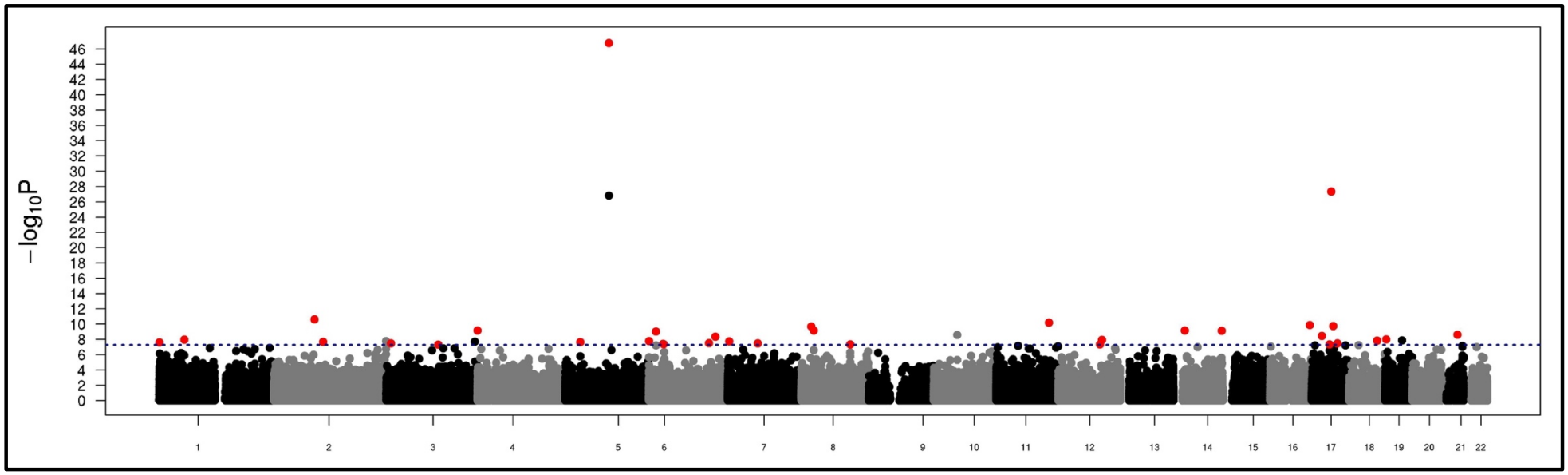

Figure 3. Validation of Meta-analysis identified CpGs in CHS and FHS combined cohorts (N=2,528, R^2=0.36).

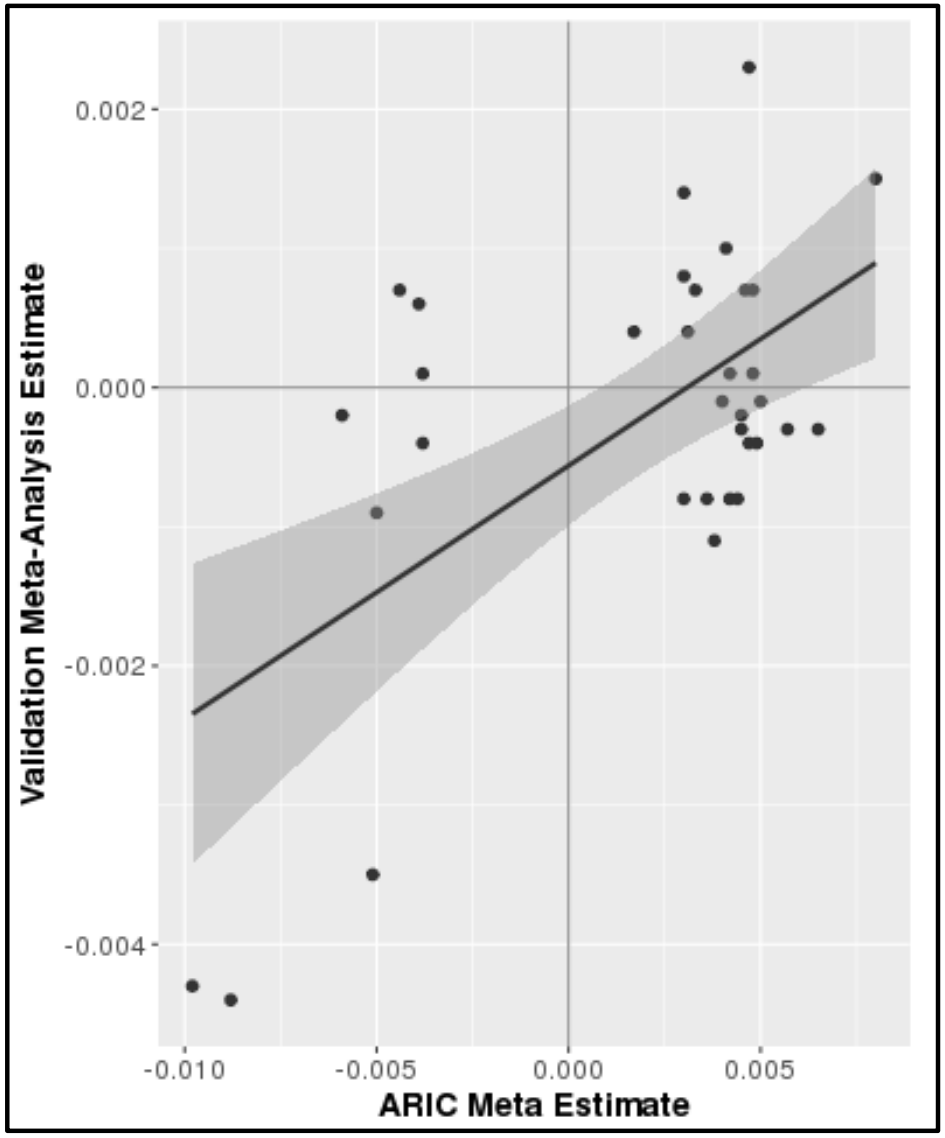

Figure 4. CRISPR-Cas9 induced heterozygous knockout of *TFAM* reduced RNA expression, mtDNA-CN and protein levels. **A.** RNA expression was reduced by over 80% relative to negative control (NC) expression (left) (passage 45). mtDNA-CN levels showed an ~18-fold reduction in *TFAM* knockout cell lines, (passage 32) (right). **B.** Western blot of CRISPR *TFAM* heterozygous knockout showed a significant reduction in *TFAM* protein (passage 35). NC=Negative Control lines. CRISPR=CRISPR *TFAM* knockout lines. Control is Tubulin.

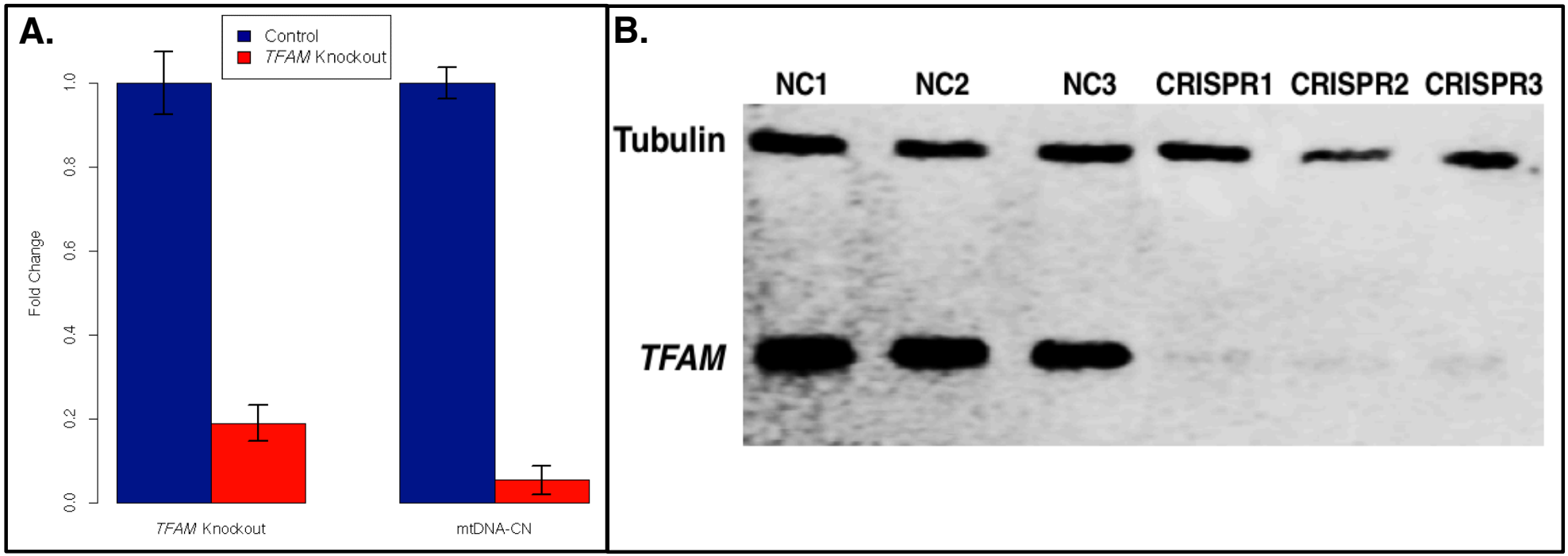

Table 1. Sample characteristics of discovery and validation cohorts.

	DISCOVERY COHORTS		VALIDATION COHORTS				
	ARIC AA (N=1567)	**ARIC EA (N=940)**	**CHS**	**CHS AA (N=239)**	**CHS EA (N=294)**	**FHS**	**FHS (N=1995)**
	mean (range)	*mean (range)*		*mean (range)*	*mean (range)*		*mean (range)*
Age	57.2 (47 - 71)	60.2 (47 - 72)		72.4 (65 - 92)	72.3 (65 - 95)		61.4 (20-91)
mtDNA-CN (SD units)	0.04 (-6.47 - 2.88)	0.04 (-4.41 - 2.87)		0.05 (-3.19 - 2.69)	0.05 (-2.28 - 2.76)		0.01 (-2.28 -8.11)
Sex	*N (percentage)*	*N (percentage)*		*N (percentage)*	*N (percentage)*		*N (percentage)*
Male	604 (38.5%)	381 (40.5%)		97 (41.8%)	112 (38.9%)		901 (45.16%)
Female	963 (61.5%)	559 (59.5%)		135 (58.2%)	176 (61.1%)		1094 (54.64%)
Collection Site			**Collection Site**			**Collection Site**	
Forsyth County, NC	188 (12%)	832 (88.5%)	Bowman Gray	74 (31.9%)	67 (23.3%)	Framingham Town	1995 (100%)
Suburbs of Minneapolis, MN	0 (0%)	86 (9.2%)	Davis	71 (30.6%)	71 (24.7%)		
Jackson, MS	1379 (88%)	0 (0%)	Hopkins	0 (0%)	83 (28.8%)		
Washington County, MD	0 (0%)	22 (2.3%)	Pittsburgh	87 (37.5%)	67 (23.3%)		
Smoking Status							
Current Smoker	655 (41.8%)	181 (19.3%)		32 (13.8%)	26 (9.0%)		205 (10.4%)
Former Smoker	484 (30.9%)	366 (38.9%)		89 (38.4%)	116 (40.3%)		947 (47.47%)
Never Smoker	428 (27.3%)	392 (41.7%)		85 (36.6%)	133 (46.2%)		838 (42.0%)
Unknown	0 (0.0%)	1 (0.1%)		26 (11.2%)	13 (4.5%)		5 (0.2%)
Phenotypes (# of cases)							
Mortality	605 (38.6%)	224 (23.8%)		194 (83.6%)	263 (91.3%)		217 (10.87%)
CVD Prevalent	154 (9.8%)	49 (5.2%)		N/A	N/A		94 (4.71%)
Incident	296 (18.9%)	108 (11.5%)		83 (35.8%)	99 (34.4%)		94 (4.71%)
CHD Prevalent	112 (7.1%)	40 (4.3%)		N/A	N/A		94 (4.71%)
Incident	193 (12.3%)	83 (8.8%)		48 (20.7%)	57 (19.8%)		68 (3.41%)
Cell Type Proportions	*mean (range)*	*mean (range)*		*mean (range)*	*mean (range)*		*mean (range)*
CD8	0.15 (0.00 - 0.48)	0.10 (0.00 - 0.27)		0.09 (0.00 - 0.38)	0.06 (0.00 - 0.22)		0.10 (0.00-0.36)
CD4	0.19 (0.00 - 0.52)	0.16 (0.00 - 0.44)		0.20 (0.00 - 0.48)	0.15 (0.00 - 0.52)		0.19 (0.02-0.44)
B-cell	0.07 (0.00 - 0.58)	0.06 (0.00 - 0.56)		0.08 (0.00 - 0.26)	0.06 (0.00 - 0.76)		0.04 (0.00-0.52)
Monocyte	0.13 (0.02 - 0.26)	0.09 (0.02 - 0.19)		0.10 (0.00 - 0.27)	0.09 (0.01 - 0.35)		0.12 (0.05-0.30)
Granulocyte	0.45 (0.15 - 0.98)	0.55 (0.16 - 0.93)		0.44 (0.11 - 0.75)	0.57 (0.03 - 0.92)		0.49 (0.02-0.85)
NK cells	N/A	0.07 (0.00 - 0.36)		0.12 (0.01 - 0.38)	0.09 (0.00 - 0.36)		0.02 (0.00-0.13)

Table 2. ARIC discovery meta-analysis identified 34 independent mtDNA-CN associated CpGs. Validation meta-analysis included CHS AA, CHS EA and FHS EA cohorts (*P*<0.05 and same direction, bolded cells). All cohort meta-analysis (ARIC AA, ARIC EA, CHS AA, CHS EA and FHS EA) identified 6 validated CpGs (*P*<5x10^{-8}, shaded cells).

Marker Name	Chr	Position	Gene‡	ARIC Meta-Analysis (N=2507)			Validation Meta-Analysis (N=2528)			All Cohort Meta-Analysis (N=5035)		
				Estimate	Standard Error	*P*-value	Estimate	Standard Error	*P-value*	Estimate	Standard Error	*P-value*
cg21051031	5	93,905,482	KIAA0825	0.0080	0.0005	1.66E-47	**0.0015**	**0.0007**	**2.45E-02**	0.0056	0.0004	1.48E-42
cg26094004	17	42,075,116	PYY	-0.0098	0.0009	4.54E-28	**-0.0043**	**0.0012**	**5.05E-04**	-0.0079	0.0007	4.13E-28
cg26563141	2	88,124,876	RGPD2; RGPD1	-0.0088	0.0013	2.42E-11	**-0.0044**	**0.0010**	**7.42E-06**	-0.0060	0.0008	2.20E-14
cg03597491	11	113,945,432	ZBTB16	0.0038	0.0006	6.20E-11	-0.0011	0.0004	6.94E-03	0.0006	0.0003	7.05E-02
cg04454285	16	86,016,317	IRF8	0.0033	0.0005	9.95E-11	0.0007	0.0004	1.25E-01	0.0018	0.0003	7.31E-08
cg01351315	17	46,667,737	LOC404266	0.0049	0.0008	1.84E-10	-0.0004	0.0005	4.26E-01	0.0013	0.0004	3.54E-03
cg21163717	8	21,769,903	DOK2	-0.0038	0.0006	2.13E-10	0.0001	0.0005	8.88E-01	-0.0014	0.0004	1.13E-04
cg13488078	8	27,469,338	CLU	0.0045	0.0007	6.72E-10	-0.0002	0.0005	7.13E-01	0.0013	0.0004	1.47E-03
cg01697902	14	25,046,117	CTSG	0.0041	0.0007	6.72E-10	**0.0010**	**0.0005**	**4.19E-02**	0.0020	0.0004	1.43E-07
cg26894523	4	107,725	ZNF718	0.0036	0.0006	6.90E-10	-0.0008	0.0004	8.68E-02	0.0007	0.0004	4.67E-02
cg10044470	14	104,866,284	C14orf144	-0.0059	0.0010	7.51E-10	-0.0002	0.0007	7.47E-01	-0.0021	0.0005	1.39E-04
cg20605134	6	15,400,462	JARID2	0.0030	0.0005	9.41E-10	0.0008	0.0005	8.75E-02	0.0018	0.0003	6.76E-08
cg17356733	21	34,774,627	IFNGR2	0.0044	0.0007	2.38E-09	-0.0008	0.0006	1.57E-01	0.0012	0.0005	7.50E-03
cg11212901	17	22,020,759	MTRNR2L1	0.0042	0.0007	3.60E-09	0.0001	0.0007	9.74E-01	0.0022	0.0005	2.33E-05
cg17586302	6	144,013,969	PHACTR2	0.0048	0.0008	4.24E-09	0.0007	0.0007	3.20E-01	0.0023	0.0005	7.39E-06
cg22068629	19	2,446,633	LMNB2	0.0050	0.0009	9.85E-09	-0.0001	0.0006	9.01E-01	0.0016	0.0005	1.29E-03
cg06358171	1	54,822,008	SSBP3	0.0057	0.0010	1.03E-08	-0.0003	0.0007	6.71E-01	0.0017	0.0006	2.30E-03
cg25006194	12	94,288,553	CRADD	0.0040	0.0007	1.15E-08	-0.0001	0.0005	8.82E-01	0.0012	0.0004	2.13E-03
cg13381110	18	60,646,614	PHLPP1	0.0065	0.0011	1.43E-08	-0.0003	0.0008	7.56E-01	0.0020	0.0007	2.55E-03
cg03910874	6	209,712	LOC285766	-0.0038	0.0007	1.64E-08	-0.0004	0.0006	4.67E-01	-0.0020	0.0005	1.28E-05
cg23304647	7	2,778,058	GNA12	0.0042	0.0007	1.84E-08	-0.0008	0.0005	1.65E-01	0.0010	0.0004	2.72E-02
cg00705730	2	106,438,120	NCK2	0.0046	0.0008	2.04E-08	0.0007	0.0006	2.52E-01	0.0020	0.0005	2.40E-05
cg00960906	5	31,769,846	PDZD2	-0.0050	0.0009	2.33E-08	-0.0009	0.0007	1.53E-01	-0.0024	0.0005	8.50E-06
cg14531564	1	1,154,853	SDF4	-0.0039	0.0007	2.34E-08	0.0006	0.0006	3.52E-01	-0.0013	0.0005	3.79E-03
cg14575356	6	130,013,903	ARHGAP18	0.0047	0.0008	3.14E-08	**0.0023**	**0.0007**	**1.10E-03**	0.0033	0.0005	1.22E-09
cg12430029	17	55,446,979	MSI2	-0.0044	0.0008	3.19E-08	0.0007	0.0006	2.83E-01	-0.0012	0.0005	1.10E-02

cg23513930	3	10,334,717	GHRLOS; GHRL	0.0030	0.0006	3.28E-08	**0.0014**	**0.0004**	**1.96E-03**	0.0020	0.0003	3.71E-09
cg01323964	7	65,219,171	SNORA22; CCT6P1	0.0048	0.0009	3.35E-08	0.0001	0.0007	9.95E-01	0.0017	0.0005	8.20E-04
cg03720100	6	30,720,263	IER3	0.0030	0.0005	3.73E-08	-0.0008	0.0005	8.46E-02	0.0008	0.0004	2.31E-02
cg08899667	6	31,761,055	VARS	-0.0051	0.0009	4.15E-08	**-0.0035**	**0.0007**	**3.11E-06**	-0.0041	0.0006	1.55E-12
cg16276850	17	38,498,914	RARA	0.0047	0.0008	4.35E-08	-0.0004	0.0005	4.38E-01	0.0010	0.0004	2.77E-02
cg17564205	12	89,992,940	ATP2B1	0.0017	0.0003	4.43E-08	0.0004	0.0004	2.66E-01	0.0012	0.0002	4.34E-07
cg12578100	8	106,330,170	ZFPM2	0.0045	0.0008	4.49E-08	-0.0003	0.0007	6.58E-01	0.0017	0.0005	1.33E-03
cg18548864	3	112,995,278	BOC	0.0031	0.0006	4.83E-08	0.0004	0.0005	4.46E-01	0.0016	0.0004	2.40E-05

‡Gene is defined as the closest gene(s) which harbor the CpG within 1500 bp of the transcriptional start site and extending to the polyA signal.

Table 3. Results of pathway and functional analysis in ARIC Meta-analysis and *TFAM* knockout methylation and expression datasets with combined *P-Value* and ARIC *P-Value* <0.05. *TFAM* integrated (INT) *P*-value represents combined methylation and expression results. Combined *P*-value represents combined ARIC and TFAM methylation and expression results. **A.** KEGG pathways sorted by combined p-value. **B.** GO pathways sorted by combined p-value.

A.

Pathway	Name	ARIC *P*-value	*TFAM* METH *P*-value	*TFAM* RNA *P*-value	*TFAM* INT *P*-value	Combined Pathway *P*-value (ARIC, *TFAM*-METH, *TFAM*-RNA)
path:hsa04080	Neuroactive ligand-receptor interaction	5.24E-12	4.41E-04	4.30E-04	8.77E-06	8.96E-16
path:hsa05033	Nicotine addiction	8.99E-04	6.30E-05	9.32E-04	7.72E-06	1.61E-08
path:hsa04024	cAMP signaling pathway	1.29E-05	2.32E-02	2.23E-01	3.25E-01	1.03E-05
path:hsa04614	Renin-angiotensin system	1.04E-05	5.11E-02	1.00E+00	1.00E+00	6.37E-05
path:hsa04723	Retrograde endocannabinoid signaling	1.89E-04	1.84E-02	1.56E-01	3.66E-03	6.48E-05
path:hsa05032	Morphine addiction	1.26E-02	4.38E-02	5.15E-02	1.40E-03	1.88E-03
path:hsa05031	Amphetamine addiction	3.54E-02	9.64E-02	4.23E-01	1.19E-01	4.18E-02
path:hsa04724	Glutamatergic synapse	1.80E-02	9.49E-02	1.00E+00	1.00E+00	4.73E-02

B.

Function	Name	ARIC *P*-value	*TFAM* METH *P*-value	*TFAM* RNA *P*-value	*TFAM* INT *P*-value	Combined Function *P*-value (ARIC, *TFAM*-METH, *TFAM*-RNA)
GO:0007267	Cell-cell signaling	1.42E-03	1.71E-05	1.19E-02	1.42E-02	7.63E-08
GO:0099537	Trans-synaptic signaling	1.88E-03	1.06E-05	6.27E-02	1.92E-02	2.89E-07
GO:0099536	Synaptic signaling	1.88E-03	1.08E-05	6.34E-02	1.92E-02	2.97E-07
GO:0007268	Chemical synaptic transmission	1.88E-03	2.22E-05	6.05E-02	1.87E-02	5.47E-07
GO:0098916	Anterograde trans-synaptic signaling	1.88E-03	2.22E-05	6.05E-02	1.87E-02	5.47E-07
GO:0099095	Ligand-gated anion channel activity	4.30E-02	8.02E-04	1.23E-04	3.74E-07	8.74E-07
GO:0045202	Synapse	1.98E-04	9.81E-05	2.74E-01	5.92E-03	1.07E-06
GO:0045211	Postsynaptic membrane	8.33E-03	1.03E-04	1.45E-01	6.19E-04	1.78E-05

Table 4. Methylation Status of Validated CpGs in *TFAM* KO cell lines (N=6). Bolded entries indicate differential expression $P<0.05$.

Marker Name	All Cohort Meta-Analysis				Average Methylation in Negative Control Lines	Average Methylation in *TFAM* Knockout Lines	*TFAM* Differential Expression	
	Mean Methylation	Estimate	Standard Error	*P*-value			Beta Estimate	*P*-Value
cg03964851 (surrogate for cg21051031)	0.83	0.0038	0.0004	7.34E-27	0.7685	0.7710	-0.0025	9.42E-01
cg26094004	0.55	-0.0079	0.0007	4.13E-28	0.6504	0.9001	-0.2497	2.91E-05
cg26563141	0.37	-0.0060	0.0008	2.20E-14	0.3071	0.4187	-0.1116	1.25E-02
cg14575356	0.55	0.0033	0.0005	1.22E-09	0.7906	0.7918	-0.0013	9.40E-01
cg23513930	0.35	0.0020	0.0003	3.71E-09	Not on EPIC array and no surrogate available			
cg08899667	0.58	-0.0041	0.0006	1.55E-12	0.7931	0.7014	0.0917	3.33E-03

***Note:** Mean methylation for cg21051031 = 0.85

Table 5. Differentially expressed genes ($P=6.41\times10^{-4}$) within 1Mb of differentially methylated CpGs.

EWAS CpG	Chr:Position	Number of Genes Within 1 Mb	*P*-Value for *TFAM* Methylation Difference	Gene	*P*-value for *TFAM* Differential Expression	Direction of Effect (following KO)*	Distance from CpG (Kb)	Description
cg26094004	**17:42,075,116**	**42**	2.91E-05	IFI35	3.76E-05	Increased	931.6	interferon induced protein 35 [Source:HGNC Symbol;Acc:HGNC:5399]
				RAMP2	5.51E-04	Increased	683.3	receptor activity modifying protein 2 [Source:HGNC Symbol;Acc:HGNC:9844]
cg26563141	**2:88,124,876**	**4**	1.25E-02	RPIA	5.04E-06	Decreased	566.8	ribose 5-phosphate isomerase A [Source:HGNC Symbol;Acc:HGNC:10297]
cg08899667	**6:31,761,055**	**32**	3.33E-03	HLA-DRB5	6.50E-07	Decreased	756.3	major histocompatibility complex: class II: DR beta 5 [Source:HGNC Symbol;Acc:HGNC:4953]
				MSH5	2.50E-04	Increased	1.8	mutS homolog 5 [Source:HGNC Symbol;Acc:HGNC:7328]

Table 6. Summary of phenotype associations from all-cohort meta-analysis for Validated CpGs. Bold entries highlight nominally significant associations (*P*<0.05).

All Cohorts	Meta Direction	Expected Direction of Phenotype Association	Prevalent CHD			Prevalent CVD			Incident CHD			Incident CVD			Mortality		
			Beta	Std. Error	*P*-Value	Beta	Std. Error	*P*-Value	Beta	Std. Error	*P*-Value	Beta	Std. Error	*P*-Value	Beta	Std. Error	*P*-Value
cg21051031	Positive	Negative	-0.56	1.95	7.74E-01	-0.67	1.88	7.23E-01	0.96	1.80	5.94E-01	-0.28	1.41	8.43E-01	-0.79	0.81	3.31E-01
cg26094004	Negative	Positive	2.91	1.48	**4.89E-02**	3.13	1.40	**2.55E-02**	0.89	0.93	3.39E-01	-0.13	0.77	8.63E-01	-0.78	0.49	1.16E-01
cg26563141	Negative	Positive	0.22	1.33	8.69E-01	-0.31	1.24	8.00E-01	0.13	0.94	8.88E-01	1.10	0.75	1.40E-01	0.98	0.45	**2.87E-02**
cg14575356	Positive	Negative	-1.80	2.04	3.79E-01	-2.19	1.88	2.45E-01	2.71	1.34	**4.36E-02**	1.75	1.08	1.04E-01	0.12	0.68	8.63E-01
cg23513930	Positive	Negative	1.99	3.37	5.54E-01	0.88	3.07	7.75E-01	2.24	1.84	2.24E-01	0.50	1.49	7.39E-01	1.07	0.94	2.53E-01
cg08899667	Negative	Positive	3.65	1.76	**3.81E-02**	3.35	1.61	**3.81E-02**	1.57	1.10	1.54E-01	0.06	0.93	9.44E-01	2.08	0.59	**3.93E-04**