Population-level expression variability of mitochondrial DNA-encoded genes in humans

Gang Wang1, Ence Yang1, Ishita Mandhan2, Candice L Brinkmeyer-Langford1 and James J Cai*,1,3

Human mitochondria contain multiple copies of a circular genome made up of double-stranded DNA (mtDNA) that encodes proteins involved in cellular respiration. Transcript abundance of mtDNA-encoded genes varies between human individuals, yet the level of variation in the general population has not been systematically assessed. In the present study, we revisited large-scale RNA sequencing data generated from lymphoblastoid cell lines of HapMap samples of European and African ancestry to estimate transcript abundance and quantify expression variation for mtDNA-encoded genes. In both populations, we detected up to over 100-fold difference in mtDNA gene expression between individuals. The marked variation was not due to differences in mtDNA copy number between individuals, but was shaped by the transcription of hundreds of nuclear genes. Many of these nuclear genes were co-expressed with one another, resulting in a module-enriched co-expression network. Significant correlations in expression between genes of the mtDNA and nuclear genomes were used to identify factors involved with the regulation of mitochondrial functions. In conclusion, we determined the baseline amount of variability in mtDNA gene expression in general human populations and cataloged a complete set of nuclear genes whose expression levels are correlated with those of mtDNA-encoded genes. Our findings will enable the integration of information from both mtDNA and nuclear genetic systems, and facilitate the discovery of novel regulatory pathways involving mitochondrial functions.

Keywords: mitochondrial DNA-encoded genes; nuclear genes; gene expression variability; human populations; mitochondrial DNA copy number

The mitochondrion is the primary energy-generating organelle, harboring critical components of the electron transport chain for ATP synthesis through oxidative phosphorylation. In almost all eukaryotic cells, mitochondria have central roles in biosynthesis, homeostasis, and programmed cell death.1–3 Dysfunctional mitochondria have pleiotropic negative effects, giving rise to a large spectrum of defects that primarily affect tissues with high energy requirements such as the brain, heart, liver, skeletal muscles, kidney, and the endocrine and respiratory systems.4–7 Despite the fundamental role of mitochondria in eukaryotic cell functions and human health, the transcriptional pattern of mtDNA has been overlooked by the majority of genomic investigations. Most transcriptome profiling studies do not take into account the transcripts of the mtDNA-encoded genes. Current procedures used for the analysis of RNA sequencing (RNA-seq) data typically filter out and discard short reads mapped onto mtDNA. As a result, none of the large-scale RNA-seq studies in humans have been focused on characterizing mtDNA transcription. Comprehensive transcriptomic data sets from large-scale RNA-seq studies have revealed significant expression variation for nuclear genes,8,9 but to date virtually no progress has been made toward investigating expression variation in mtDNA genes. The lack of a precedent for evaluating the expression variability of mtDNA genes will ultimately hinder the understanding of mitochondrial biology as well as the advance of mitochondrial medicine.

To simultaneously estimate the expression of mtDNA and nuclear genes, in the present study we revisited published RNA-seq data8,9 that included a large number of polyadenylated10–13 mitochondrial transcripts. We quantified between-individual variation in the expression of mtDNA genes in HapMap samples of European (CEU) and African (YRI) ancestry to evaluate this variation at the population level. We detected co-varying expression between mtDNA and nuclear genes, and explored various potential mechanisms that may underlie the association.

MATERIALS AND METHODS

RNA sequencing data

Short sequence data produced for two RNA-seq studies8,9 were obtained from GEO using accessions GSE19480 and GSE25530. The SRA (sequence read archives) files were downloaded and subsequently converted into FASTQ files using the NCBI SRA toolkit program, fastq-dump (v 2.1.16). Mitochondria and mitoplast RNA-seq data produced by Mercer et al14 were obtained from GEO using accession GSE30772.

CEU and YRI reference genomes

The human reference genome (hg19) was re-engineered for CEU and YRI by replacing the mtDNA sequence to produce two population-specific reference genomes. For CEU, the Revised Cambridge Reference Sequence (NC_012920) was obtained from the MITOMAP website (http://www.mitomap.org/MITOMAP). For YRI, the sequence of NC_001807 in GenBank’s RefSeq database was used. CEU and YRI mtDNA sequences differ by 41 single-nucleotide variants and four indels. The GTF file of Gencode v11 with updated mtDNA annotation (for CEU and YRI, respectively) was used to guide the short-read mapping.
Estimation of transcript abundance

To estimate FPKM (fragments per kilobase of exon per million fragments), RNA-seq short reads were mapped to the corresponding population-specific reference genome using TopHat v2.0.1 and processed using Cufflinks v2.0.2. TopHat option – read-mismatches was set to 2 and 3 (ie, allowing up to 2 and 3 mismatches in final read alignments), for CEU and YRI, respectively, because the mtDNA nucleotide diversity in YRI is higher than that in CEU.15 By setting the TopHat option -g to 1, we deliberately allowed reads to be mapped on a single specific position on the genome. For each population, the expression values of \( \log_{2}(\text{FPKM} + 1) \) for all genes were quantile-normalized across individuals.16 The gene expression level per mtDNA copy was computed according to \( \log_{2}(\text{FPKM} + 1) \) with \( C_{\text{mt}} \) the mtDNA copy number. Thirteen representative housekeeping (HK) genes were randomly selected from the list of HK genes in a previous study.17

Weighted gene co-expression network analysis

The weighted gene co-expression network analysis (WGCNA)18 was used to identify nuclear gene modules in the co-expression network including mtDNA genes. For each module, we defined eigengene significance by measuring the correlation between the modules and the trait under consideration (ie, the average expression level of mtDNA genes). To ensure the results are robust to the WGCNA parameter settings,19 we allowed the key parameter, the thresholding power for network construction (or the power), to vary between 5, 6 (the optimal value determined by WGCNA), and 7, where the other parameters were kept fixed (ie, the smallest value of the scale independence = 0.9, the minimum module size = 30, and the maximum joining height = 10). Thus, instead of producing WGCNA modules using one single value for the power, we ran WGCNA three times with three different power values and produced three sets of WGCNA modules. We then iterated all pairs of genes, and identified gene pairs in which two genes appeared in the same module in all three sets of WGCNA modules. The identified pairs of genes were represented with 1 in an adjacency matrix, while the rest pairs of genes were represented with 0. We used the MCL algorithm implemented in SBEToolbox20 to identify clusters of genes within the adjacency matrix. In this way, we identified genes in the same clusters that were highly connected and had been consistently grouped in all three sets of WGCNA modules.

Gene ontology analysis

For certain sets of genes, we computed enrichment scores for the gene ontology (GO) biological process and molecular function terms using DAVID.21 The program compares the annotation composition in a list of genes to that of background genes. The full set of genes expressed in LCLs (with average FPKM > 1.0) was used as the background gene set.

Estimation of mtDNA copy number

The mtDNA copy number was estimated using the mtDNA/nDNA ratio.22–25 For a given chromosomal region, short reads mapped to the region were retrieved from the BAM file at the FTP site of the 1000 Genomes Project26 using samtools.27 This was done for the whole mtDNA region and the autosomal regions; the numbers of short reads in the two regions were used to compute mtDNA/nDNA. If the read coverage along the genome is even, then the number of reads in different genomic regions of the same length should not vary substantially. That is to say, the autosomal regions could be selected arbitrarily as long as the total length of these regions is long enough (eg, comparable with the length of mtDNA). In this study, we chose to use the genomic regions of the 13 HK genes. To confirm that mtDNA copy number estimation was indeed not affected by the selection of autosomal regions, we reestimated the mtDNA copy numbers for all samples 1000 times using randomly selected autosomal regions of the same total length. Each time, the result was compared with the result derived from the genomic regions of the 13 HK genes. The correlations between estimates were consistently high (average Spearman correlation coefficient \( [SCC] = 0.981 \); Supplementary Figure 1).

Analysis of eQTL

HapMap SNPs28 with minor allele frequency (MAF) of 5% were selected from CEU and YRI populations (~2.2 million per population). We tested the eQTL associations between each SNP and gene expression with a linear regression model using Matrix-eQTL.29 To establish the null distribution of \( P \)-values, randomly shuffled expression data were used to perform the regression analysis.

Distribution of source code and data

The associated source code and data are available at http://www.github.com/jamesjcai/mtRNA-seq.

RESULTS

Expression levels of mtDNA genes

We used RNA-seq data sets generated from LCLs of 60 CEU20 and 69 YRI21 individuals to estimate expression levels for both mtDNA and nuclear genes. RNA-seq short reads were mapped onto population-specific reference genomes. Multiple-hit reads were discarded to avoid the influence of mapping artifacts caused by the improper assignment of mtDNA reads to nuclear genome sequences of mitochondrial origin (NUMTs), which include >500 sequences covering ~627 kb or 0.021% of the human genome.20,31 Our results showed that, overall, mtDNA genes were expressed significantly more abundantly than other genes (Kolmogorov-Smirnov (K–S) test: \( P = 1.3e-12 \); Supplementary Figure 2). This is consistent with our expectation and indicates that the RNA-seq data sets are replete with mitochondrial transcripts. Additional analyses showed that the accuracy of this estimation was not affected by: (1) the number of mapped reads, (2) the ages of individuals from whom blood samples were collected, (3) the batch effect of cell line processing in the RNA-seq experiments, or (4) EBV copy number and/or cell doubling time for the LCLs (see Supplementary Note 1: analyses showing that the accuracy of mtDNA gene expression estimation is not affected by confounding factors).

Expression variability of mtDNA genes

There was substantial variation in mtDNA gene expression among CEU (Figure 1) and YRI (Supplementary Figure 3) individuals. The variation was more pronounced in mtDNA genes (Figure 1a) than HK genes (Figure 1b). To quantify this, we computed the coefficient of variation (CV) of expression for all genes with an average FPKM > 2.0. Indeed, mtDNA genes tended to have a significantly larger CV than nuclear genes (K–S test: \( P = 2.7e-6 \); see Supplementary Figure 4). We hypothesized that the marked variation in mtDNA gene expression was due to differences in mtDNA copy number between individuals. To test this, we estimated the mtDNA copy number for each sample using the ratio between the number of short reads mapped in mtDNA versus nuclear DNA (Materials and Methods). Our estimates showed a significant, positive correlation with the estimates obtained by Maranville et al32 using a PCR-based method with 46 YRI samples (Pearson correlation test: \( r = 0.445, P = 0.002 \)). As before, we found considerable differences in mtDNA copy number among CEU and YRI individuals (Supplementary Figure 5), but this variation did not correlate with that of mtDNA gene expression (Spearman correlation test: \( \rho = 0.067 \) and 0.112, \( \rho = 0.61 \) and 0.42, for CEU and YRI, respectively). In fact, gene expression level per copy was found to be highly variable for mtDNA genes, ranging from 0.09 to 0.20 in CEU and 0.06 to 0.11 in YRI. These results suggest that differences in mtDNA copy number do not account for mtDNA gene expression variability.

Coordinated expression of mtDNA genes

It is known that functionally related genes, such as the 13 mtDNA protein-coding genes, are likely to be expressed coordinately.33

---

*European Journal of Human Genetics*

G. Wang et al
Indeed, we observed remarkably strong, positive correlations between expression levels of mtDNA genes: SCCs between possible pairs ranged from 0.55 to 0.97 with an average of 0.86 (Figure 1c and Supplementary Figure 6). In contrast, SCCs between the HK genes were much smaller (Supplementary Figure 7a, b): only 7 out of 78 HK gene pairs showed a SCC > 0.50 (Figure 1d). Additionally, we examined the SCCs between genes whose products form a single protein complex – the SNF2h/cohesion complex. For this particular example, the SCCs between genes of the same protein complex (Supplementary Figure 7c, d) were stronger than those between HK genes, but still much weaker than those between mtDNA genes.

Co-expression of mtDNA and nuclear genes
To examine co-expression of mtDNA and nuclear genes, we performed pairwise correlation tests and identified nuclear genes whose expressions were significantly correlated with one or more mtDNA genes (SCC with Bonferroni correction, adjusted P < 0.05; Supplementary Table 1). These included 496 positive and 434 negative correlations between nuclear and mtDNA gene expression in CEU, plus 203 positive and 184 negative correlations in YRI. A total of 63 of these genes (29 positive and 34 negative) were found in common between CEU and YRI (Figure 2), including 15 ribosomal protein genes, ACIN1 (apoptotic chromatin condensation inducer protein activated by caspase-3, Figure 3a), and ZSWIM1 (zinc-finger SWIM domain-containing protein indirectly interacting with MT-ND2 through MAPK14, Figure 3b). GO analyses indicated that positively-correlated nuclear genes were often involved in the regulation of transcription; those negatively-correlated were more likely to be involved in translation (for details, see Supplementary Table 2).

We used several SCC cutoffs ranging from 0.5 to 0.9 to produce co-expression networks of different sizes. The size of the resulting co-expression networks decreased with the increase of the SCC cutoff value (Supplementary Table 3). The overlap between co-expression networks for CEU and YRI decreased rapidly with the increment of
SCC cutoff. This low level of overlap might be due to that the overall expression distributions were different between CEU and YRI (Supplementary Figure 8) that may be attributed to the technical difference in RNA-seq procedures (ie, paired-end RNA-seq for CEU and single-end RNA-seq for YRI) or differences in the number of passage from immortalization of LCLs between CEU and YRI samples.

Next, to examine the large-scale organization of co-expression networks, we used the WGCNA method to identify nuclear gene modules significantly correlated with mtDNA genes. WGCNA starts from the level of thousands of genes, with modules represented by their centroids, to assess relationships between modules and the trait under consideration. A conservative procedure that reduces the influence of WGCNA parameter choices (Materials and Methods), we identified a total of 98 gene clusters, each containing at least 30 genes. These clusters are available for download (Materials and Methods). All pairs of genes in one cluster had been consistently grouped into the same module, regardless of different values of the power used in WGCNA analysis.

A major concern when evaluating relationships between genes based on their expression is that transcriptional co-regulation among many genes can give rise to indirect interaction effects in expression data, and regular correlation networks cannot distinguish direct from indirect relationships. To control for the indirect effects, we employed the approach developed by Schafer and Strimmer based on the graphical Gaussian model (GGM) to reconstruct a GGM network. In this network, each link indicated a partial correlation between two genes that remained after removing the effects of other genes. Our results showed that many nuclear genes were partially correlated with mtDNA genes (Supplementary Table 4), and that partial correlation of mtDNA genes with RSPO1, PRRC2C, EIF4E2, and TMEM101 appeared in Figure 2 showed significance.

Effects of subcellular co-localization
Transcript localization by mRNA trafficking is an important cellular process that controls subcellular distribution of mRNAs and the subsequent distribution of proteins. Through this process, transcripts of a number of nuclear genes preferentially localize to the vicinity of mitochondria. We hypothesized that these nuclear genes are likely to be functionally related to mitochondria and that their expression is likely to be correlated with mtDNA genes. To test this, we obtained the RNA-seq data generated by Mercer et al. by using a subtractive approach. In that study, mRNAs were extracted and sequenced for a mitochondrial preparation as well as for a mitoplast.

Figure 2 Co-varying expression network between mtDNA and nuclear genes. The consensus network shared between CEU and YRI is plotted. Connections are placed between genes that are correlated at |Spearman’s $\rho| > 0.60$ in CEU and YRI. The width of connection lines is proportional to the average absolute value of SCCs for the two populations. Positive and negative correlation relationships are presented as red and blue lines, respectively. Scatter plots between expression of MT-ATP6 and RSPO1 for CEU and YRI are shown to depict the degree of correlations.
expression. We identified this kind of autosomal SNPs and tabulated these SNPs and corresponding nuclear genes whose expression was associated with mtDNA gene expression (Supplementary Table 5).

DISCUSSION

We systematically examined the between-individual variation in the expression level of mtDNA genes, a subject that has been neglected by the overwhelming majority of previous studies. Using existing data sets, we quantified the population-level expression variability of mtDNA genes for European and African populations. Up to > 100-fold between-individual difference in mtDNA gene expression was detected in both populations.

We further investigated whether the marked variation is due to differences in mtDNA copy numbers between-individual samples. Each mitochondrion contains between two and ten copies of mtDNA, cells have numerous mitochondria, and a cell may harbor several thousand mtDNA copies. It is generally believed that mtDNA gene expression is proportional to the number of mtDNA copies and the amount of mtDNA in a cell could provide a major regulatory point in mitochondrial activity. Our results based on the population-level expression variability, however, do not appear to favor the idea.

We evaluated relationships between mtDNA and nuclear genes by constructing the co-expression network and using the WGCNA algorithm to detect the correlated genes and modules. Functional analyses of correlated genes and modules confirmed biological processes previously known to be associated with mitochondrial activities, such as apoptosis. These analyses also identified many genes involved in those biological processes previously unknown to be associated with mitochondrial activities. Furthermore, we evaluated whether functionally relevant pathways could be investigated by identifying associations between genetic variants (mainly, SNPs) and expression of mtDNA genes. However, we did not find evidence for the existence of any mitochondrial eQTL. We conclude that although genomic variants may make important contributions to the expression of mtDNA genes, the evidence to date is too weak to support such a conclusion. On the other hand, we know that evidence for mtDNA polymorphisms associated with susceptibility to complex disorders is also weak. Thus, establishing convincing relationships between phenotypes and mtDNA transcripts/variants remains highly challenging.

Several caveats and technical limitations are associated with our analysis. (1) Many factors that might alter the number of mtDNA copies could not be controlled. These factors include the stage of the cell cycle, the energetic requirements of the cells, the environmental effects on the redox balance of the cells, the stage of differentiation, and/or cell signaling mechanisms. An accurate determination of the number of mtDNA copies per cell is technically difficult to achieve because both the number of mitochondria per cell and the number of mtDNA copies per mitochondrion vary. Our analysis could only focus on the variability of mtDNA gene expression at the level of per sample, ignoring the details of expression variability at the level of per mtDNA copy. (2) Throughout the paper, the transcription of mtDNA genes was treated as occurring in a manner completely analogous to that of nuclear transcription. For mtDNA, its long polycistronic precursor transcripts are processed and released individually, which means that the transcription of mtDNA genes is treated as occurring in a manner completely analogous to that of nuclear transcription. For mtDNA, its long polycistronic precursor transcripts are processed and released individually, and stabilized with polyadenylation regulated by mitochondria-specific poly(A) polymerase and polynucleotide phosphorylase. The differences in these detailed transcriptional and post-transcriptional processes between mtDNA genes and nuclear genes were ignored in this study. (3) The two RNA-seq data sets used in this study were
derived from poly(A)-enriched RNA pools, and were not generated using strand-specific RNA-seq. These might influence the estimation of the levels of mtDNA gene expression because of the different polyadenylation statuses of mtDNA genes\textsuperscript{13,30,61} and the possible cross-mapping between L- and H-strand-derived mitochondrial transcripts.\textsuperscript{60} Nevertheless, because these technical limitations systematically influence all samples in the same manner on the same order of magnitude, our main results in connection with the between-individual expression variability should not be affected.

In summary, we have taken the first step toward characterizing the population-level expression variability of mtDNA genes in humans, which is done through exploiting widely accessible yet completely untapped RNA-seq reads originating from the transcribed mitochondrial genome. In doing this, we established an analytical framework for future analyses of the interplay between the two human genomes. This study demonstrates the utility of publically available data for answering interesting questions in studies of natural human variation. Using this data, we have confirmed that there is a substantial amount of variation in mtDNA gene expression across individuals, rejecting the hypothesis that the transcript abundance of mtDNA genes is determined by the number of mtDNA copies. Next, the expression of mtDNA genes may be either positively or negatively associated with the expression of many known and unknown network modules. These modules contain many genes, many of whose functions were hitherto not known to be linked with mitochondrial function, underscoring the need to further study the underlying mechanisms of these associations to increase our understanding of the genetic basis of the expression regulation of mtDNA genes.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

We thank all anonymous reviewers and section editor Peter Robinson for their valuable and constructive comments. We thank Hui Jiang, Yi Xing, Dongxiao Zhu, Stephen Montgomery, Maranville, and Di Rienzo for helpful discussions and/or sharing research data. EY is supported by CVM Postdoctoral Trainee Research Grant (02-14400-03504) and GW is supported by CVM Graduate and/or sharing research data. EY is supported by CVM Postdoctoral Trainee funding. EY is supported by CVM Postdoctoral Trainee funding.


Supplementary Information accompanies this paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)