# CELLULAR AND CELL-FREE MITOCHONDRIAL DNA HETEROPLASMY AND ITS MEDICAL SIGNIFICANCE

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Ruoyu Zhang

May 2018

CELLULAR AND CELL-FREE MITOCHONDRIAL DNA HETEROPLASMY AND ITS MEDICAL SIGNIFICANCE

Ruoyu Zhang, Ph. D.

Cornell University 2018

Since the discovery of pathogenic mitochondrial DNA (mtDNA) mutations in the 1980's, these mutations have been shown to be involved in a number of human diseases and have caught increasing attention. The emergence of next generation sequencing technology has allowed us to do more comprehensive studies of these mutations in the mitochondrial genome, especially for low frequency heteroplasmic mutations. In my research projects, I combined experimental and computational approaches to investigate mtDNA heteroplasmies in different medical disciplines.

I first conducted a literature review to summarize recent findings on the implications of mtDNA mutations in human diseases, with a focus on complex diseases such as cancer, neurodegenerative diseases and aging. My first project was to investigate mtDNA heteroplasmy and copy number variations in a general healthy population. This population-based study indicated that both mtDNA quality and quantity decreased with age. I further found that mtDNA copy number was associated with serum bicarbonate level and white blood cell counts, while aggregate heteroplasmy load was associated with blood apolipoprotein B level. These results suggested that maintaining optimal mtDNA copy number and preventing the expansion of heteroplasmy could promote healthy aging.

In my second project, I first identified heteroplasmies in 466 pairs of DNA and RNA sequencing data. I verified that most of the heteroplasmies were transcribed to RNA regardless of their pathogenic potentials. I then experimentally tested whether the heteroplasmy frequencies could change over time. My test showed that a heteroplasmy with ~50% frequency could decrease to ~1% in only 28 days. Moreover, these observed heteroplasmy dynamics could significantly affect certain gene expression levels.

My third project focused on mtDNA fragments circulating in blood stream in a cell-free format. I developed an experiment protocol tailored for sequencing short DNA fragments from plasma samples. After analyzing the sequencing data, I found that the fragment length of mtDNA in plasma is much shorter than that of nuclear DNA. We also demonstrated that mtDNA heteroplasmy was detectable in the plasma sample, suggesting its potential to serve as a biomarker in different clinical applications.

## BIOGRAPHICAL SKETCH

Ruoyu Zhang grew up in Kunming, the so-called “Spring City” in China. She almost never saw snow before she moved to Beijing for her college, whereas she may already saw most snows of her entire life during her PhD’s study in Ithaca.

Ruoyu was first intrigued by the DNA double helix when she was in elementary school by reading a short story about Human Genome Project from an encyclopaedia. In her senior year at high school. She won a first-place prize in the National Olympic Competition of Chemistry, which brought her to Tsinghua University to study Chemical and Biological Engineering for her undergraduate study. Afterwards, with the MEXT scholarship, she completed a Master Program in Dr. Susumu Kajiwara and Dr. Takahiro Oura’s lab at Tokyo Institute of Technology, where she received solid experiment training in molecular biology. During her trip to the US for a scientific conference to present her master’s study, she met Dr. Zhenglong Gu from Cornell University. After a lively conversation with Dr. Gu, she decided to resume her childhood dream, to analyze the human DNA sequences.

Before joining Dr. Gu’s lab at Cornell University, Ruoyu was trained as a bioinformatics analyst at Beijing Genome Institute, Shenzhen. In the past five years Ruoyu has been working with Dr. Zhenglong Gu at Cornell University to investigate the mitochondrial DNA mutations and their implications in diseases.

Ruoyu is also an amateur soccer player for more than 20 years. She won the third-place prize in Beijing University League with Tsinghua University women’s soccer team, she was a registered soccer player in Japan and a four-time champion of Cornell IM league.

*To my parents*

*And to Gavin and his devoted mom*

ACKNOWLEDGMENTS

When I was five years old, I asked my mom what I can do after I finish my college. My mom told me: “You can try to pursue a Master”, “What next?” I asked again. “You can become a Doctor”. Therefore, I would first thank my mom for directing me to this very tough but worthwhile road.

On this very challenge journey, I felt very fortunate to have met my respectful mentors. Foremost, I would like to thank my advisor, Dr. **Zhenglong Gu**. He has always been very supportive, even before I came to Cornell to join his lab. I appreciate all his contributions of time, energy, effort and ideas to support my research. He is a contagious, passionate, very creative and also very careful scientist. Under his guidance, I gradually became an independent researcher. Most important, he taught me that always being curious is such a precious characteristic for a scientist. I also want to thank my committee members, Drs. Patrick Sullivan, Iwijn De Vlaminck and Tom Brenna to provide guidance from different aspects to help me complete my PhD’s study. I also want to thank Dr. Kiichi Nakahira for providing the collaboration opportunities from medical school, as well as many valuable discussions to push forward my projects.

I would like to thank all my fellow labmates in Dr. Gu’s lab. Especially, I want to thank Dr. Kaixiong Ye and Mr. Yiqin Wang for countless discussions about data analysis. I want to thank Dr. Xiaoxian Guo, he is always willing to help with all the experiments without any reservation. I want to thank Dr. Yuan Si and Mr. Yiping Wang, my same year PhD labmates, to share all the joy and pain together during our

study.

I would like to thank Drs. Jonathan Toung and Li Liu to welcome me to Illumina for the impressive internship, exposing me to the most advanced research group in the next generation sequencing industry. I would also like to thank Dr. Sherry Xian to host me at iCarbonX for another unique internship experience.

I would like to thank my dear friends Feng Yang, Xi Zhan, Yiyang Dai, Xinzi Yu and Lingxia Sun to listen to my words and always encourage me. I would like to thank Qingyang Chen for running the social media account about women's soccer together with me, to keep my spare time very joyful. I would like to thank Xuan Qi for watching AC Milan's games together with me to make sure I am not pissed off by the team during these years. I would like to thank Stephanie French to organize the wonderful Ithaca Women's Soccer League, made my time in Ithaca much more interesting, and kept me doing enough exercise to stay healthy to complete my study.

I finish with my parents, where the most of my life energy resides, to support me in all my pursuits.

Table of Contents

Chapter 1 –mtDNA and human disease ..... 1

1.1 Abstract ..... 1
1.2 Introduction ..... 2
1.3 mtDNA genetics, mtDNA heteroplasmy and diseases ..... 2
1.3.1 Mitochondrial genetics ..... 3
1.3.2 mtDNA heteroplasmy ..... 3
1.3.3 Heteroplasmy sources and changes during lifetime ..... 7
1.4 mtDNA heteroplasmy implications in diseases ..... 9
1.4.1 Overview ..... 9
1.4.2 Mitochondrial diseases ..... 10
1.4.3 Neurodegenerative diseases ..... 11
1.4.4 Cancer ..... 12
1.4.5 Aging ..... 14
1.5 Acknowledgements ..... 16
1.6 References ..... 17

Chapter 2 – Impacts of Aging on Mitochondrial DNA Quantity and Quality in Humans ..... 23

2.1 Abstract ..... 23
2.2 Introduction ..... 25
2.3 Methods ..... 28
2.3.1 Data Access Permission ..... 28
2.3.2 mtDNA variation identification and haplogroup assignment ..... 28
2.3.3 Potential cross-sample contamination inspection ..... 29
2.3.4 Annotation of mtDNA variants ..... 30
2.3.5 mtDNA copy number estimation ..... 30
2.3.6 Association testing for mtDNA copy number and heteroplasmy ..... 31
2.3.7 Phenotypic associations of mtDNA copy number and heteroplasmy ..... 31
2.3.8 Mitochondrial heteroplasmy load and SKAT test ..... 32
2.4 Results ..... 33

2.4.1 Mitochondrial heteroplasmy is prevalent in UK10K TwinsUK cohort....33
2.4.2 Mitochondrial heteroplasmy has high pathogenic potential....37
2.4.3 Mitochondrial heteroplasmy burden increases with age....39
2.4.4 Age has effects on mtDNA heteroplasmy and copy number....43
2.4.5 Mitochondrial DNA copy number is associated with number of heteroplasmies...48
2.4.6 Phenotypic associations of mtDNA copy number and heteroplasmy load....51
2.5 Discussion....53
2.6 Conclusions....60
2.7 Acknowledgement....61
2.8 Reference....62
Chapter 3 – Heteroplasmy concordance between mitochondrial DNA and RNA....67
3.1 Abstract....67
3.2 Introduction....69
3.3 Methods....70
3.3.1 Data Retrieval and Pre-processing....70
3.3.2 Heteroplasmy identification and annotation....71
3.3.3 Cell line culture and point heteroplasmy sequencing....72
3.3.4 Genome wide association study to locate SNPs associated with editing events....72
3.4 Result....72
3.4.1 Identification of heteroplasmies using DNA and RNA sequencing data....73
3.4.2 Compare heteroplasmy frequencies between DNA and RNA....75
3.4.3 RDDs could be caused by heteroplasmy dynamics....77
3.4.4 Potential modification/editing events in mitochondrial RNA....80
3.5 Discussion....87
3.6 Acknowledgement....89
3.7 Reference....90
Chapter 4 - Mitochondrial DNA heteroplasmy dynamics causes global gene expression changes. ....92
4.1 Introduction....92
4.2 Material and Methods....93
4.2.1 Cell culture and single heteroplasmy site sequencing....93

4.2.2 RNA sequencing and bioinformatics analysis....93
4.2.3 Simulation of mtDNA segregation....94
4.3 Results....94
4.3.1 Heteroplasmy frequency can change over time....95
4.3.2 Heteroplasmy frequency changes can affect gene expression profiles....95
4.3.3 In-silico simulation of heteroplasmy changes....102
4.4 Discussion....106
4.5 Acknowledgement....107
4.6 Reference....108
Chapter 5 – Very Short Mitochondrial DNA Fragments and Heteroplasmy in Human Plasma....109
5.1 Abstract....109
5.2 Introduction....110
5.3 Materials and Methods....112
5.3.1 Clinical sample collection and plasma processing....112
5.3.2 Library preparation for plasma DNA....114
5.3.3 Library preparation for white blood cell DNA....114
5.3.4 Analysis workflow for Next-Generation Sequencing data....115
5.3.5 Heteroplasmy identification for white blood cell and plasma....115
5.3.6 Haplotype Analysis....117
5.3.7 Data Access....117
5.4 Results....117
5.4.1 Plasma mtDNA has a distinct size distribution compared to nDNA....117
5.4.2 cf-mtDNA recovery rate is improved by new method....122
5.4.3 Size selection further improves cell free mitochondrial DNA recover rate....124
5.4.4 mtDNA heteroplasmy in plasma....126
5.5 Discussion....132
5.6 Acknowledgement....135
5.7 Reference....136
Chapter 6 AFTERWORD....139
APPENDIX A: Publication Inclusion Authorizations....143

LIST OF FIGURES

Figure 1-1. Schematic diagram of the human mitochondrial genome. ..... 4

Figure 1-2. mtDNA heteroplasmy and the threshold effects..... 6

Figure 1-3. mtDNA heteroplasmy variations caused by mitochondrial bottleneck. ..... 8

Figure 2-1 mtDNA copy number comparison between cell lines and PBMCs DNA. ..... 35

Figure 2-2. Distribution of heteroplasmy in UK10K TwinsUK cohort. ..... 36

Figure 2-3. Pathogenic potential for nonsynonymous heteroplasmies. ..... 38

Figure 2-4 Association between mtDNA heteroplasmy number and age. ..... 41

Figure 2-5. Distribution of mtDNA copy number in the UK10K Twins cohort and its association with age..... 45

Figure 2-6. Association between mtDNA heteroplasmy number and copy number..... 49

Figure 2-7. Manhattan plot of associations between homoplasmic variants and mtDNA copy number. 50

Figure 2-8. mtDNA copy number association with phenotypic traits. ..... 52

Figure 3-1 mtDNA sequencing statistics in DNA and RNA seq data. ..... 74

Figure 3-2 Heteroplasmy frequency difference between DNA and RNA. ..... 76

Figure 3-3 CADD pathogenic scores of HDLR and non-HDLR heteroplasmies. ..... 78

Figure 3-4 RDDs could be caused by heteroplasmy dynamics ..... 79

Figure 3-5 Heteroplasmy frequency would change over time ..... 82

Figure 3-6 Edited allele frequency distribution in 3 previously reported mtRNA editing sites ..... 85

Figure 4-1 Heteroplasmy frequency changes during 28 days in the study cell line..... 97

Figure 4-2 The expression profile of the 13 mtDNA protein encoding genes in 28 growing days. .....101

Figure 4-3 The expression profile of the TCA related genes in 28 growing days. .....103

Figure 4-4 The expression profile of the glycolysis related genes in 28 growing days. .....104

Figure 4-5 Heteroplasmy frequency changes from simulation results.....105

Figure 5-1. Comparison of the standard method and optimized method. .....119

Figure 5-2. Relative recovery rate of DNA fragments with different lengths.....120

Figure 5-3. Plasma DNA size distribution. .....121

Figure 5-4. mtDNA fractional concentration in different size windows.....125

LIST OF TABLES

Table 2-1. Heteroplasmy number in different age groups ..... 42

Table 2-2. Regional distribution of heteroplasmy in different age groups ..... 44

Table 2-3. Correlation of age with mtDNA heteroplasmy number and copy number ..... 47

Table 3-1. examples of heteroplasmy with noticeable frequency difference between DNA and RNA .. 81

Table 3-2 Three previous reported mtRNA editing sites in this dataset ..... 84

Table 3-3 mtRNA editing events number identified in the study population ..... 86

Table 4-1. Heteroplasmy frequency changes during 28 days in cell line GM12282 in two replicate experiments ..... 96

Table 4-2. Upregulated genes GO enrichment ..... 99

Table 4-3. Downregulated genes GO enrichment ..... 100

Table 5-1. mtDNA fractional concentration by different approaches ..... 123

Table 5-2. mtDNA heteroplasmy in patient 1, 93 ..... 128

Table 5-3. mtDNA heteroplasmy present in both WBC and plasma in patient 42 ..... 129

Table 5-4. mtDNA heteroplasmy present only in WBC in patient 42 ..... 130

Table 5-5. mtDNA heteroplasmy present only in plasma in patient 42 ..... 131

## *Chapter 1 –mtDNA and human disease*[1]

### 1.1 Abstract

Mitochondria, more than just being the powerhouses of the cell, are involved in a wide range of cellular and metabolic processes. Besides classic mitochondrial diseases, the role of mitochondrial dysfunction has been increasingly recognized in many common and complex human diseases, such as metabolic disorders, neurodegenerative diseases, and cancers. Mitochondrial DNA mutations are common causes of mitochondrial dysfunction. Investigating the pathogenic roles of mitochondrial DNA mutations has been challenging because there are usually hundreds to thousands of DNA copies in a single cell, with mutant DNA coexisting with wild-type copies. The recent advance in high-throughput sequencing technologies provides new approaches to detect low-frequency mitochondrial DNA mutations and to examine their implications in the onset and progression of complex diseases. The accumulating knowledge of the normal mitochondrial biology will assist our development of new health management strategies by maintaining proper mitochondrial function.

[1] Plan to submit to *Advances in Nutrition*. Ruoyu Zhang and Yuan Si contribute equally to this work.

## 1.2 Introduction

Mitochondria are double membrane organelles, presenting in almost all eukaryotic cells. Different from other organelles, mitochondria host their own genome: mitochondrial DNA (mtDNA). The primary function of mitochondria is the production of ATP, supplying over 90% of cellular energy[1]. Mitochondria are also responsible for a series of cellular processes, including calcium signaling, iron homeostasis, and cell apoptosis [2, 3]. In addition to the well-established mitochondrial diseases, such as Leber hereditary optic neuropathy (LHON), mitochondrial encephalomyopathy, and lactic acidosis and stroke-like episodes (MELAS), a growing spectrum of human diseases have been found to be associated with mitochondrial dysfunction, including cancer, neurodegenerative diseases, and metabolic disorders [3-7]. Multiple molecular mechanisms have been proposed to explain the precise roles of mitochondria in these pathological processes, such as the excessive generation of reactive oxygen species (ROS), the accumulation of mtDNA mutations, and the mitochondria-mediated apoptosis. However, there is still no consensus on these issues. On the other hand, understanding the contributions of mitochondrial dysfunction to these common diseases may suggest a new way of intervention just by preserving mitochondrial functions.

In this chapter, we will do a brief overview of the mtDNA genetics and the implications of mitochondrial dysfunction in human diseases with a focus on mtDNA mutations.

## 1.3 mtDNA genetics, mtDNA heteroplasmy and diseases

### 1.3.1 Mitochondrial genetics

Compelling evidence suggested that mitochondria were once primitive bacterial cells and were acquired by the host through endosymbiotic event. Along the way of endosymbiosis, the bacteria became double membrane organelles and gradually transferred genes to their symbiotic cell nucleus, with only a few genetic materials retained as mtDNA [3, 8, 9]. Human mtDNA is a circular double-stranded molecule comprising 16569 base pairs. The two strands are distinguished by their molecular weight: a guanine-rich heavy strand and a cytosine-rich light strand. mtDNA encodes 13 peptides, serving as core subunits of the five enzyme complexes in the oxidative phosphorylation (OXPHOS) system. mtDNA also encodes 2 rRNA and 22 tRNA, which are essential for intra-mitochondrial protein synthesis (Figure 1-1).
Unlike human nuclear DNA (nDNA), mtDNA is condensed with genes. About 93% of its entirety encodes genes, which also lack intronic regions. The 13 protein-coding genes are separated by tRNA or 1~2 non-coding bases. The non-coding region is mainly located in the displacement loop (D-loop), which hosts the mtDNA replication initiation site and two H-strand transcription promoters. Because of this functionally dense organization, nucleotide substitutions in mtDNA are more likely to cause functional outcomes than nDNA mutations.

### 1.3.2 mtDNA heteroplasmy

In comparison to only two copies of nDNA within a cell, there are hundreds to thousands of copies of mtDNA. As a result, the mutation could be present in all copies of mtDNA (homoplasmy) or only a proportion of them (heteroplasmy), as illustrated

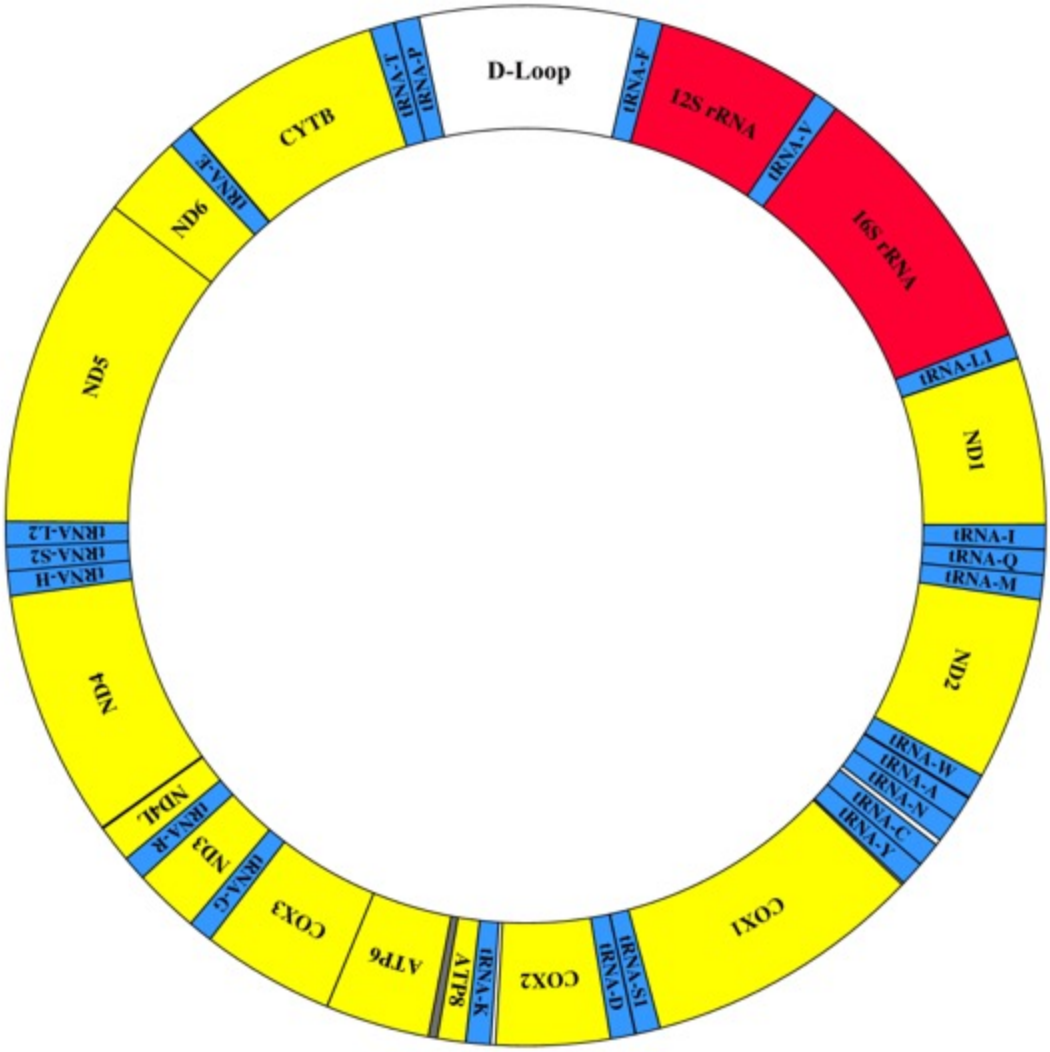


Figure 1-1. Schematic diagram of the human mitochondrial genome.

Human mtDNA is a 16569 bp double-stranded circular DNA molecule. The 13 protein encoding genes are shown in yellow blocks, the 22 transfer RNA genes are shown in blue blocks and the 2 ribosomal RNA genes are shown in red blocks. Figure was generated with mtviz (http://pacosy.informatik.uni-leipzig.de/mtviz)

in Figure 1-2. The proportion of mtDNA mutant copies is referred as heteroplasmy frequency. This frequency critically determines the phenotypic effect of a specific mutation. It has been suggested that there is a "phenotypic threshold effect". At low heteroplasmy frequencies, the deleterious effect of mutant mtDNA is masked by coexisting wild type copies, and once exceeding a threshold value (typically 60%-80%), mutant mtDNA will result in an altered phenotype (Figure 1-2) [10-12]. This frequency threshold varies across mutations and tissues [3]. The heteroplasmy frequency of a specific mtDNA mutation can vary across individuals. Take mutation 3243A>G as an example, which is the most common pathogenic heteroplasmic mutation and can cause multiple mitochondrial diseases, including MELAS, chronic progressive external ophthalmoplegia (CPEO), and Kearns–Sayre syndrome (KSS), Rajasimha *et al.* examined the frequencies of this mutation in 275 individuals and found the frequencies range from a few percent to higher than 80% [13]. It has also been reported that heteroplasmy frequency has tissue-specific variations within the same individual [14]. Heteroplasmy 72T>C showed high frequencies in the liver and kidney, a moderate frequency at skeletal muscle and low frequencies in other tissues [14]. While variations across individuals and tissues are well-established, the heteroplasmy frequency variation at the single cell level is still controversial. Jayaprakash *et al.* examined mtDNA heteroplasmy in colonies derived from single cells and found that heteroplasmy frequency was stably maintained in individual daughter cells [15]. On the other hand, Neupane *et al.* reported that in mouse embryonic stem cells, heteroplasmy frequency could vary up to 61% between individual cells. This wide variability in stem cells may explain the existence of

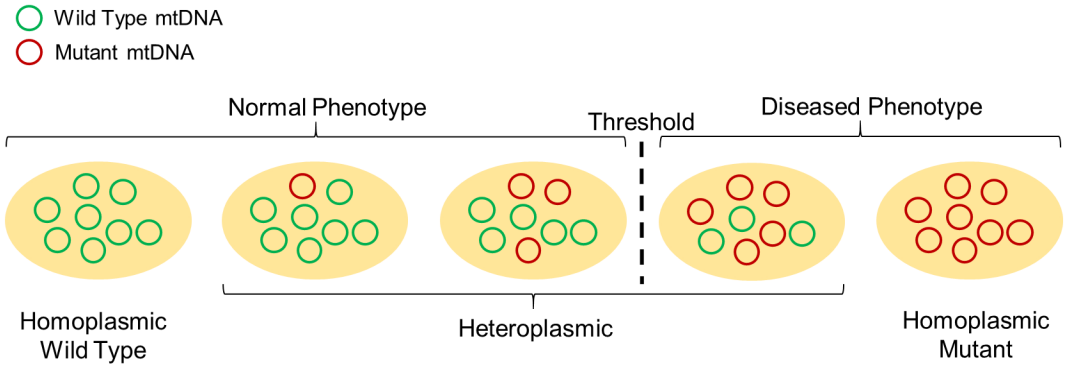


Figure 1-2. mtDNA heteroplasmy and the threshold effects.

The mtDNA within a same cell can be all identical (homoplasmic) or a mixture of wild type and mutant (heteroplasmic). The cells can harbor different proportions of mutated and wild type mtDNA (referred as heteroplasmy frequency). The heteroplasmy frequency is critical for the mutations' pathogenicity. If the heteroplasmy frequency is below a certain threshold, the cell can maintain normal phenotypes. However, once the frequency exceeds the threshold, the cell will show signs of mitochondrial dysfunctions.

tissues with different heteroplasmy loads in an individual [16].

### 1.3.3 Heteroplasmy sources and changes during lifetime

Taking advantage of the recent advance in next-generation sequencing (NGS) technology, several studies have demonstrated that most of the individuals, if not all, have heteroplasmy in their mitochondrial genomes [17, 18]. Heteroplasmy mutations may be inherited mutations from maternal mtDNA, or *de novo* mutations arising during embryonic development. Unlike nuclear genome which is transmitted by sexual reproduction, the human mitochondrial genome is strictly maternal transmitted. Although inherited from a single parent, extensive differences of mtDNA heteroplasmy frequency between mothers and offspring, and among siblings have been observed [19, 20]. For example, Li et al. found that the average difference in heteroplasmy frequency between mothers and offspring was 10.8%, with a maximal of 78.7% in a Netherlands cohort [21]. These variations could be a result of mitochondrial bottleneck effects: During oocytes development, only a small number of mtDNA are sampled from primordial germ cells and transmitted to primary oocytes, which leads to the heteroplasmy frequencies vary drastically among mature oocytes (Figure 1-3). Because of the bottleneck, low frequency mutant mtDNA has a relative lower probability to be transmitted to the next generation. But if they were selected during the bottleneck, the frequency of the mutant mtDNA could dramatically increase. In extreme cases, this process may also allow the frequency of a disease mutation to reach the phenotypic threshold within a single generation, resulting in

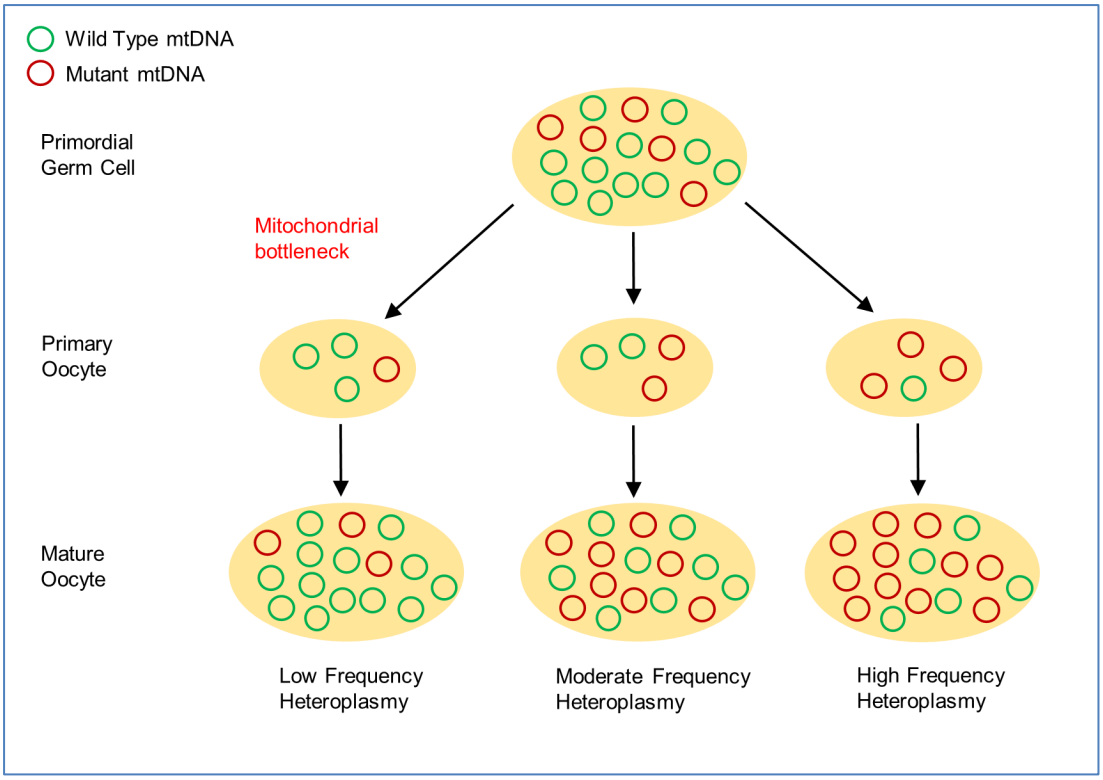


Figure 1-3. mtDNA heteroplasmy variations caused by mitochondrial bottleneck.

The transmission of mtDNA heteroplasmy from mother to offspring is affected by mitochondrial bottleneck effects during egg development. There is a significant reduction of mtDNA copy number from primordial germ cells to primary oocytes, meaning that only a small proportion of mtDNA will be sampled and transmitted to primary oocytes, leading to extensive variation of heteroplasmy frequency in the mature oocytes.

childhood-onset mitochondrial diseases or even complex diseases, such as autism spectrum disorder (ASD). Additionally, recent evidence has shown that there is a strong purifying selection of mtDNA in the maternal germline development and embryogenic development [22, 23], which also help prevent or reduce the transmission of deleterious mtDNA mutations.

mtDNA heteroplasmy can also be a result of somatic mutations. The triggers of the mtDNA somatic mutation are still highly debated. According to the original theory, it is the immensely oxidative microenvironment of mitochondria that cause mtDNA mutations. However, an emerging hypothesis argues that most mtDNA mutations originate from replication errors because the mtDNA mutation signatures are dominated by transition changes [7, 24-26]. mtDNA is replicated constantly through the lifetime and is independent of the cell cycle. Moreover, mtDNA replication and repairing system are less accurate than nDNA [27, 28]. Therefore, both dividing cells and post-mitotic cells can accumulate mtDNA mutations, especially heteroplasmic mutations over time. A newly introduced mutation in a single mtDNA molecule is possible to clonally expand to a higher frequency and even to reach the phenotypic threshold. Computational models are suggesting that mtDNA mutations arose early in life have sufficient time to reach the phenotypic threshold and to cause mitochondrial dysfunctions [29].

## 1.4 mtDNA heteroplasmy implications in diseases

### 1.4.1 Overview

Mitochondrial dysfunction is implicated in a broad spectrum of human diseases. Apart

from the classic mitochondrial diseases, such as MELAS and LHON [30-34], emerging evidence associates mtDNA mutations with common human diseases. In early studies, technology limitations drastically affect the sensitivity and resolution to identify mtDNA heteroplasmy, thus the implication of heteroplasmy in common and complex diseases might be overlooked. The advent of NGS technologies offers a new opportunity to revisit the connection between heteroplasmy and different diseases. In the following section, we will provide only a brief recapitulation of the roles of mtDNA mutations in classic mitochondrial diseases and focus most of our discussion on the recent insights into their implications in common and complex human diseases.

### 1.4.2 Mitochondrial diseases

"Classic" or "primary" mitochondrial disease refers to a group of diseases caused by defects in OXPHOS, which are results of mutations in nDNA-or mtDNA-encoded mitochondrial genes. Some mtDNA mutations can contribute to several different mitochondrial DNA diseases. The most common disease-causing mutation 3243A>G is associated with CPEO, MELAS, maternally inherited diabetes and deafness (MIDD) [35]. On the other side, a specific mitochondrial disease can be caused by a set of mutations. To date, mutations located in > 75 genes (in both mitochondrial and nuclear) have been identified to be involved in Leigh syndrome [36]. mtDNA mutation 3460G>A, 11778G>A and 14484T>C were found in both homoplasmic and heteroplasmic status in LHON families [37]. Advances in NGS technology helped the elucidation of the genetic basis of mitochondrial diseases and their diagnosis, but the treatment of these diseases remains a challenge, the fast development of DNA editing

techniques could be a promising direction which has the potential to fix the ETC defects at DNA level.

### 1.4.3 Neurodegenerative diseases

The central nervous system is highly energy-demanding and thus heavily depends on mitochondrial energy production. mtDNA mutations and associated OXPHOS activity reduction have been observed in several complex neurodegenerative diseases, such as Parkinson's Disease (PD), Alzheimer's Disease (AD) and ASD. Early work revealed that mitochondrial haplogroups are associated with different disease risks. Researchers use the certain mtDNA homoplasmic variations to define the mitochondrial haplogroups, with different letters representing different subpopulations. For example, haplogroup J represents a subpopulation originated from Eurasia and is reported to have reduced risk of PD in a meta-analysis study [38], while haplogroup H, U, K, T, I, W and X are associated with increased risk of AD [39-43]. However, it is still unclear about the role of heteroplasmic mutations in these diseases. Here, we will summarize the recent findings of mtDNA heteroplasmy in complex neurodegenerative diseases. Lin *et al.* first found that mtDNA heteroplasmic mutations are significantly elevated in the substantia nigra of early PD patients [44]. As revealed by Coxhead *et al.*, PD patient brains are more vulnerable to mtDNA mutation by analyzing different regions in brain. Heteroplasmic nonsynonymous variants in several mitochondrial genes are also overrepresented in PD patients [45]. Considering AD, it has been reported that mtDNA heteroplasmic mutation is overrepresented in the hippocampus of early stage patients, and the mutation signature is consistent with replication errors rather than

oxidative damages [46]. These reports support that the accumulation of mtDNA mutations may contribute to the neurodegeneration in both PD and AD. Unlike PD and AD, Huntington's disease (HD) has elucidated genetic causes, the autosomal dominant repeat expansions in the Huntington gene. However, mitochondria still play an important role in HD progression. While mtDNA depletion and deletion are reported in several studies [47-50], there are very few studies about heteroplasmy. Filosto *et al.* reported the 5613T>C heteroplasmy could cause chronic progressive external ophthalmoplegia in HD patients [51]. Our lab is conducting a large-scale study of mtDNA sequencing in HD patients, which will systemically investigate the involvement of mtDNA heteroplasmy in this disease and assist the development of novel strategies to postpone the onset of the disease or even to counteract its progression. In addition to these age-related neurodegenerative diseases, ASD, which usually affects pre-pubertal children, is also associated with mitochondrial dysfunction. Our lab has analyzed 903 ASD children along with their unaffected siblings and mothers. We found that nonsynonymous and predicted pathogenic heteroplasmic mutations are enriched in autistic probands. Moreover, a large fraction of these mutations is inherited [52] , indicating that evaluating mtDNA heteroplasmic mutations in high-risk families may help early diagnosis and treatment of ASD.

### 1.4.4 Cancer

Reprogrammed energy metabolism is a hallmark of cancer [53]. Otto Warburg first observed the rewiring of cancer cell energy metabolism: mitochondrial respiration and OXPHOS are suppressed, substituted by a strong enhancement of glycolysis even in

the presence of oxygen. This phenomenon was termed as “aerobic glycolysis” [54-56]. The depression of mitochondrial activity in cancer cells may have various reasons. Recently, disruption of mitochondrial respiration complexes due to detrimental mtDNA mutations has been suggested to be one neglected reason.

Several studies have reported that mtDNA somatic mutations are frequently found in tumor tissues [57-59]. With precise quantification achieved by NGS, in 10 matched tumor-normal colorectal tissues, He *et al.* found 90% of the cancer tissues had at least one cancer-specific mtDNA point mutation (present in cancer tissue, but not present in the matched normal tissue), and most of these cancer-specific mutations are heteroplasmic rather than homoplasmic [60]. Davis *et al.* also found loss-of-function heteroplasmic mutations in NADH dehydrogenase subunit by sequencing mtDNA of 66 chromophobe renal cell carcinomas [61]. A more comprehensive study was conducted later by comparing mtDNA sequences from 1657 cancer and matched normal tissues in 31 cancer types. This study identified 1907 cancer specific somatic substitutions, most of which are heteroplasmic [24]. Notably, this study also suggested that the majority of mutations are generated from mtDNA replication errors rather than external mutagens such as ROS, cigarette smoking, and UV light. However, the understanding of how mtDNA mutations affect mitochondrial function and cellular metabolism is still limited. Hardie *et al.* investigated the metabolic genotype-phenotype relationships between mtDNA mutations and pancreatic cancer, discovering that heterogeneous genomic landscapes of cancer can converge towards common metabolic phenotypes, including reduced oxygen consumption and increased glycolysis [62].

It is apparent that mtDNA mutations, including heteroplasmic mutations, are widespread in cancer cells. However, the precise role of mtDNA mutations in oncogenesis is currently unresolved. To answer these questions, future investigation should put effort on elucidating the causal mechanisms of mtDNA in cancer.

### 1.4.5 Aging

Aging is a degenerative process with the gradually impaired physiological function that eventually leads to deterioration of cellular function, disease, and death [7]. During last several decades, multiple lines of evidence have shown that impaired mitochondrial function is implicated in aging and age-associated disease [7, 63-65]. Accumulation of mutations in mtDNA during time can lead to severe impairment of cellular energy production and mitochondrial dysfunction [64]. In humans, the accumulation of mtDNA mutations over time has been observed in both dividing cells and non-dividing (post-mitotic) cells, such as brain, muscle, and colon [66-69]. The first experimental evidence for the causative link between accumulation of mtDNA mutations and aging was from the mutator mice model. These mutator mice had proofreading-deficient mtDNA polymerase (Polγ), leading to accumulation of extensive mtDNA mutations. These mice had reduced lifespan and premature onset of aging-related phenotypes such as weight loss, hair loss, reduced fertility, *etc.* [65, 70]. It has been proposed for decades that ROS generated during cellular metabolism can damage mtDNA, while the resulted mtDNA mutations would further lead to disruption of the electron transport chain (ETC), which produces more ROS, creating a vicious cycle. Recent studies have suggested that the age-associated mtDNA

mutation accumulation is not from ROS damage, but rather from the spontaneous errors during the mtDNA replication. These replication errors will arise as low-frequency heteroplasmy and the potential subsequent clonal expansion of these heteroplasmic mutations would disturb the mitochondrial function [3]. Therefore, managing the expansion of the mtDNA mutations could be critical for aging.

## 1.5 Acknowledgements

I would like to thank Drs. Kaixiong Ye, Yuan Si, Xiaoxian Guo, Lingfeng Tang and Mr. Yiqin Wang for their valuable comments.

## 1.6 References


1. Wallace, D.C., *Mitochondrial DNA in aging and disease*. Scientific American, 1997. **277**(2): p. 40-59.
2. Schon, E.A., S. DiMauro, and M. Hirano, *Human mitochondrial DNA: roles of inherited and somatic mutations*. Nat Rev Genet, 2012. **13**(12): p. 878-890.
3. Stewart, J.B. and P.F. Chinnery, *The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease*. Nat Rev Genet, 2015. **16**(9): p. 530-542.
4. Picard, M., D.C. Wallace, and Y. Burelle, *The rise of mitochondria in medicine*. Mitochondrion, 2016. **30**: p. 105-116.
5. Garcia-Heredia, J.M. and A. Carnero, *Decoding Warburg's hypothesis: tumor-related mutations in the mitochondrial respiratory chain*. Oncotarget, 2015. **6**(39): p. 41582-99.
6. Chrysostomou, A. and D.M. Turnbull, *Mitochondria, the Synapse, and Neurodegeneration*, in *Mitochondrial Dysfunction in Neurodegenerative Disorders*, A.K. Reeve, et al., Editors. 2016, Springer International Publishing: Cham. p. 219-239.
7. Kauppila, T.E.S., J.H.K. Kauppila, and N.-G. Larsson, *Mammalian Mitochondria and Aging: An Update*. Cell Metabolism, 2017. **25**(1): p. 57-71.
8. Wallace, D.C. and D. Chalkia, *Mitochondrial DNA genetics and the heteroplasmy conundrum in evolution and disease*. Cold Spring Harb Perspect Biol, 2013. **5**(11): p. a021220.
9. Marlow, F.L., *Mitochondrial matters: Mitochondrial bottlenecks, self-assembling structures, and entrapment in the female germline*. Stem Cell Research, 2017. **21**: p. 178-186.
10. Rossignol, R., et al., *Mitochondrial threshold effects*. Biochem J, 2003. **370**(Pt 3): p. 751-62.
11. Boulet, L., G. Karpati, and E. Shoubridge, *Distribution and threshold expression of the tRNA (Lys) mutation in skeletal muscle of patients with myoclonic epilepsy and ragged-red fibers (MERRF)*. American journal of human genetics, 1992. **51**(6): p. 1187.
12. Larsson, N., et al., *Segregation and manifestations of the mtDNA tRNA (Lys) A--> G (8344) mutation of myoclonus epilepsy and ragged-red fibers (MERRF) syndrome*. American journal of human genetics, 1992. **51**(6): p. 1201.
13. Rajasimha, H.K., P.F. Chinnery, and D.C. Samuels, *Selection against Pathogenic mtDNA Mutations in a Stem Cell Population Leads to the Loss of the 3243A→G Mutation in Blood*. American Journal of Human Genetics, 2008. **82**(2): p. 333-343.
14. Li, M., et al., *Extensive tissue-related and allele-related mtDNA heteroplasmy suggests positive selection for somatic mutations*. Proc Natl Acad Sci U S A, 2015. **112**(8): p. 2491-6.
15. Jayaprakash, A.D., et al., *Stable heteroplasmy at the single-cell level is*

*facilitated by intercellular exchange of mtDNA*. Nucleic Acids Res, 2015. **43**(4): p. 2177-87.
16. Neupane, J., et al., *Cellular Heterogeneity in the Level of mtDNA Heteroplasmy in Mouse Embryonic Stem Cells*. Cell Reports, 2015. **13**(7): p. 1304-1309.
17. Ye, K., et al., *Extensive pathogenicity of mitochondrial heteroplasmy in healthy human individuals*. Proceedings of the National Academy of Sciences, 2014. **111**(29): p. 10654-10659.
18. Ding, J., et al., *Assessing Mitochondrial DNA Variation and Copy Number in Lymphocytes of ~2,000 Sardinians Using Tailored Sequencing Analysis Tools*. PLoS Genet, 2015. **11**(7): p. e1005306.
19. Ghosh, S.S., et al., *Longitudinal study of a heteroplasmic 3460 Leber hereditary optic neuropathy family by multiplexed primer-extension analysis and nucleotide sequencing*. Am J Hum Genet, 1996. **58**(2): p. 325-34.
20. Larsson, N.G., et al., *Segregation and manifestations of the mtDNA tRNA(Lys) A-->G(8344) mutation of myoclonus epilepsy and ragged-red fibers (MERRF) syndrome*. Am J Hum Genet, 1992. **51**(6): p. 1201-12.
21. Li, M., et al., *Transmission of human mtDNA heteroplasmy in the Genome of the Netherlands families: support for a variable-size bottleneck*. Genome Res, 2016. **26**(4): p. 417-26.
22. Samuels, D.C., et al., *Recurrent Tissue-Specific mtDNA Mutations Are Common in Humans*. PLOS Genetics, 2013. **9**(11): p. e1003929.
23. Liau, W.S., et al., *A persistent mitochondrial deletion reduces fitness and sperm performance in heteroplasmic populations of C. elegans*. BMC Genet, 2007. **8**: p. 8.
24. Ju, Y.S., et al., *Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer*. eLife, 2014. **3**: p. e02935.
25. Payne, B.A.I. and P.F. Chinnery, *Mitochondrial dysfunction in aging: Much progress but many unresolved questions*. Biochimica et Biophysica Acta, 2015. **1847**(11): p. 1347-1353.
26. Sevini, F., et al., *mtDNA mutations in human aging and longevity: Controversies and new perspectives opened by high-throughput technologies*. Experimental Gerontology, 2014. **56**: p. 234-244.
27. Holt, I.J. and A. Reyes, *Human mitochondrial DNA replication*. Cold Spring Harb Perspect Biol, 2012. **4**(12).
28. Alexeyev, M., et al., *The Maintenance of Mitochondrial DNA Integrity—Critical Analysis and Update*. Cold Spring Harb Perspect Biol, 2013. **5**(5).
29. Elson, J.L., et al., *Random Intracellular Drift Explains the Clonal Expansion of Mitochondrial DNA Mutations with Age*. The American Journal of Human Genetics, 2001. **68**(3): p. 802-806.
30. Wallace, D.C., *Mitochondrial DNA mutations in disease and aging*. Environmental and molecular mutagenesis, 2010. **51**(5): p. 440-450.
31. Wallace, D.C. and D. Chalkia, *Mitochondrial DNA genetics and the heteroplasmy conundrum in evolution and disease*. Cold Spring Harbor perspectives in biology, 2013. **5**(11): p. a021220.

32. Wallace, D.C., *A mitochondrial bioenergetic etiology of disease*. The Journal of clinical investigation, 2013. **123**(4): p. 1405-1412.
33. Gorman, G.S., et al., *Mitochondrial diseases*. Nature Reviews Disease Primers, 2016. **2**: p. 16080.
34. Alston, C.L., et al., *The genetics and pathology of mitochondrial disease*. The Journal of Pathology, 2017. **241**(2): p. 236-250.
35. Nesbitt, V., et al., *The UK MRC Mitochondrial Disease Patient Cohort Study: clinical phenotypes associated with the m.3243A>G mutation--implications for diagnosis and management*. J Neurol Neurosurg Psychiatry, 2013. **84**(8): p. 936-8.
36. Lake, N.J., et al., *Leigh syndrome: One disorder, more than 75 monogenic causes*. Ann Neurol, 2016. **79**(2): p. 190-203.
37. Carelli, V., et al., *Leber's hereditary optic neuropathy: biochemical effect of 11778/ND4 and 3460/ND1 mutations and correlation with the mitochondrial genotype*. Neurology, 1997. **48**(6): p. 1623-32.
38. Hudson, G., et al., *Two-stage association study and meta-analysis of mitochondrial DNA variants in Parkinson disease*. Neurology, 2013. **80**(22): p. 2042-2048.
39. Krüger, J., et al., *Mitochondrial DNA haplogroups in early-onset Alzheimer's disease and frontotemporal lobar degeneration*. Molecular neurodegeneration, 2010. **5**(1): p. 8.
40. Lakatos, A., et al., *Association between mitochondrial DNA variations and Alzheimer's disease in the ADNI cohort*. Neurobiology of aging, 2010. **31**(8): p. 1355-1363.
41. Maruszak, A., et al., *Mitochondrial haplogroup H and Alzheimer's disease—is there a connection?* Neurobiology of aging, 2009. **30**(11): p. 1749-1755.
42. Santoro, A., et al., *Evidence for sub-haplogroup h5 of mitochondrial DNA as a risk factor for late onset Alzheimer's disease*. PLoS One, 2010. **5**(8): p. e12037.
43. Van Der Walt, J.M., et al., *Analysis of European mitochondrial haplogroups with Alzheimer disease risk*. Neuroscience letters, 2004. **365**(1): p. 28-32.
44. Lin, M.T., et al., *Somatic mitochondrial DNA mutations in early Parkinson's and incidental Lewy body disease*. Ann Neurol, 2012. **71**(6): p. 850-4.
45. Coxhead, J., et al., *Somatic mtDNA variation is an important component of Parkinson's disease*. Neurobiology of Aging, 2016. **38**: p. 217.e1-217.e6.
46. Hoekstra, J.G., et al., *Mitochondrial DNA mutations increase in early stage Alzheimer disease and are inconsistent with oxidative damage*. Annals of Neurology, 2016. **80**(2): p. 301-306.
47. Hering, T., et al., *Selective striatal mtDNA depletion in end-stage Huntington's disease R6/2 mice*. Experimental neurology, 2015. **266**: p. 22-29.
48. Siddiqui, A., et al., *Mitochondrial DNA damage is associated with reduced mitochondrial bioenergetics in Huntington's disease*. Free Radical Biology and Medicine, 2012. **53**(7): p. 1478-1488.
49. Banoei, M.M., et al., *Huntington's disease and mitochondrial DNA deletions: event or regular mechanism for mutant huntingtin protein and CAG repeats expansion?!* Cellular and molecular neurobiology, 2007. **27**(7): p. 867.

50. Chen, C.-M., et al., *Increased oxidative damage and mitochondrial abnormalities in the peripheral blood of Huntington's disease patients*. Biochemical and biophysical research communications, 2007. **359**(2): p. 335-340.
51. Filosto, M., et al., *A novel mitochondrial tRNAAla gene variant causes chronic progressive external ophthalmoplegia in a patient with Huntington disease*. Molecular Genetics and Metabolism Reports, 2016. **6**: p. 70-73.
52. Wang, Y., M. Picard, and Z. Gu, *Genetic Evidence for Elevated Pathogenicity of Mitochondrial DNA Heteroplasmy in Autism Spectrum Disorder*. PLOS Genetics, 2016. **12**(10): p. e1006391.
53. Hanahan, D. and Robert A. Weinberg, *Hallmarks of Cancer: The Next Generation*. Cell. **144**(5): p. 646-674.
54. Warburg, O., *On the origin of cancer cells*. Science, 1956. **123**(3191): p. 309-314.
55. Warburg, O., *On respiratory impairment in cancer cells*. Science, 1956. **124**(3215): p. 269-270.
56. Warburg, O., *The Metabolism of Tumours: Investigations from the Kaiser Wilhelm Institute for Biology, translated by Frank Dickens*. Constable & Co Ltd, 1930.
57. Wong, L.J., et al., *Detection of mitochondrial DNA mutations in the tumor and cerebrospinal fluid of medulloblastoma patients*. Cancer Res, 2003. **63**(14): p. 3866-71.
58. Jones, J.B., et al., *Detection of mitochondrial DNA mutations in pancreatic cancer offers a "mass"-ive advantage over detection of nuclear DNA mutations*. Cancer Res, 2001. **61**(4): p. 1299-304.
59. Chatterjee, A., E. Mambo, and D. Sidransky, *Mitochondrial DNA mutations in human cancer*. Oncogene, 2006. **25**(34): p. 4663-74.
60. He, Y., et al., *Heteroplasmic mitochondrial DNA mutations in normal and tumour cells*. Nature, 2010. **464**(7288): p. 610-614.
61. Davis, Caleb F., et al., *The Somatic Genomic Landscape of Chromophobe Renal Cell Carcinoma*. Cancer Cell, 2014. **26**(3): p. 319-330.
62. Hardie, R.-A., et al., *Mitochondrial mutations and metabolic adaptation in pancreatic cancer*. Cancer & Metabolism, 2017. **5**(1): p. 2.
63. Bratic, A., et al., *The role of mitochondria in aging*. The Journal of Clinical Investigation, 2013. **123**(3): p. 951-957.
64. López-Otín, C., et al., *The Hallmarks of Aging*. Cell, 2013. **153**(6): p. 1194-1217.
65. Kujoth, G., et al., *Mitochondrial DNA mutations, oxidative stress, and apoptosis in mammalian aging*. Science, 2005. **309**(5733): p. 481-484.
66. Ross, J.M., et al., *Germline mitochondrial DNA mutations aggravate ageing and can impair brain development*. Nature, 2013. **501**(7467): p. 412-415.
67. Bua, E., et al., *Mitochondrial DNA–Deletion Mutations Accumulate Intracellularly to Detrimental Levels in Aged Human Skeletal Muscle Fibers*. The American Journal of Human Genetics, 2006. **79**(3): p. 469-480.
68. Greaves, L.C., et al., *Comparison of Mitochondrial Mutation Spectra in*

*Ageing Human Colonic Epithelium and Disease: Absence of Evidence for Purifying Selection in Somatic Mitochondrial DNA Point Mutations*. PLOS Genetics, 2012. **8**(11): p. e1003082.
69. Greaves, L.C., et al., *Clonal Expansion of Early to Mid-Life Mitochondrial DNA Point Mutations Drives Mitochondrial Dysfunction during Human Ageing*. PLOS Genetics, 2014. **10**(9): p. e1004620.
70. Trifunovic, A., et al., *Premature ageing in mice expressing defective mitochondrial DNA polymerase*. Nature, 2004. **429**(6990): p. 417-423.

## *Chapter 2 – Impacts of Aging on Mitochondrial DNA Quantity and Quality in Humans* [2]

### 2.1 Abstract

The accumulation of mitochondrial DNA (mtDNA) mutations, and the reduction of mtDNA copy number, both disrupt mitochondrial energetics, and may contribute to aging and age-associated phenotypes. However, there are few genetic and epidemiological studies on the spectra of blood mtDNA heteroplasmies, and the distribution of mtDNA copy numbers in different age groups and their impact on age-related phenotypes. In this work, we used whole-genome sequencing data of isolated peripheral blood mononuclear cells (PBMCs) from the UK10K project to investigate in parallel mtDNA heteroplasmy and copy number in 1,511 women, between 17-85 years old, recruited in the TwinsUK cohorts. We report a high prevalence of pathogenic mtDNA heteroplasmies in this population. We also find an increase in mtDNA heteroplasmies with age ($\beta$ = 0.011, $P$ = 5.77e-6), and showed that, on average, individuals aged 70-years or older had 58.5% more mtDNA heteroplasmies than those under 40-years old. Conversely, mtDNA copy number decreased by an average of 0.4 copies per year ($\beta$ = -0.395, $P$ = 0.0097). Finally, mtDNA copy number was positively associated with serum bicarbonate level ($P$ = 4.46e-5), and inversely correlated with white blood cell count ($P$ = 0.0006). Moreover, the aggregated heteroplasmy load was associated with blood apolipoprotein B level ($P$ = 1.33e-5),

[2] Published on *BMC Genomics*.

linking the accumulation of mtDNA mutations to age-related physiological markers. Our population-based study indicates that both mtDNA quality and quantity are influenced by age. An open question for the future is whether interventions that would contribute to maintain optimal mtDNA copy number and prevent the expansion of heteroplasmy could promote healthy aging.

## 2.2 Introduction

Mitochondria play a central role in cellular energy metabolism, as well as in a range of other cellular activities, such as calcium signaling, iron homeostasis, hormone synthesis, and programmed cell death [1-3]. Mitochondria differ from all other organelles in animals in having their own DNA (mitochondrial DNA, mtDNA), which in humans encodes 37 genes: 22 tRNAs, 2 rRNAs and 13 protein subunits of the electron transport chain and Complex V/ATP synthase. Although they contribute only ~1% of the mitochondrial proteome, the 13 mtDNA-encoded proteins are nevertheless essential for mitochondrial oxidative phosphorylation and cellular energetics [4]. A single mammalian cell hosts hundreds to thousands of copies of mtDNA, which are thought to have played a critical role in the evolution of mammalian genomic complexity [5]. Because of its multi-copy nature, spontaneous mtDNA mutations often affect only a small proportion of the cell's mtDNA, a state termed heteroplasmy. In contrast, if all mtDNA molecules harbor a specific mutation, it is said to be in a state of homoplasmy. mtDNA heteroplasmy is implicated in several human diseases, in which the ratio of mutated to wild-type mtDNA is critical in determining whether a specific mutation is deleterious [2, 6, 7]. In previous studies, we demonstrated that even in healthy adults, low-frequency heteroplasmies with high pathogenic potential were common [8].

In addition to mtDNA mutation burden, the number of mtDNA molecules per cell, or "mtDNA copy number", is also strictly regulated, ensuring that mitochondria can generate appropriate levels of energy and intracellular signals to maintain normal cellular functions. Altered mtDNA copy number has been shown to be involved in

age-related diseases, including cancer, neurodegeneration disorders, and diabetes [9-11]. In the general population, mtDNA copy number measured in peripheral blood has also been shown to be associated with a variety of physiological phenotypes, and to be linked with aging and mortality [12, 13]. For example, higher mtDNA copy number was linked with better physical and mental health status in aged populations [12]. There has been speculation that both mtDNA heteroplasmy and copy number may contribute to the aging process, but the effects of mtDNA heteroplasmy and copy number were only discussed separately, thus remaining inconclusive in humans [14]. Aging is commonly characterized as a time-dependent progressive loss of physiological integrity, leading to impaired function and increased vulnerability to death [14]. One important factor in aging is the accumulation of DNA damage over time [15]. mtDNA has been considered a major target of aging-associated mutation accumulation, possibly because it experiences higher oxidative damage, more turnover, and has lower replication fidelity compared to nuclear DNA (nDNA) [16-18]. Mice carrying elevated mtDNA mutation burden present premature signs of aging including hair loss, kyphosis, and premature death (lifespan shortened by up to 50%) [19, 20]. In human studies, mtDNA heteroplasmy incidence increases with age [21-23], while lower mtDNA copy number has been reported in aged populations [12, 24]. Ding et.al reported an trend of increased heteroplasmies and decreased mtDNA copy number with age in their study population [25]. However, previous studies were limited in one or more ways: i) limited power in detecting low-to-medium frequency heteroplasmies in blood due to low sequencing depth; ii) relatively small sample sizes, limiting statistical power; iii) small age range; iv) whole blood as the source of DNA,

which contains several sources of contaminants for mtDNA analysis; and/or v) assessing either mtDNA mutation or copy number, but not both in the same biological samples.

Whole genome sequencing (WGS) data allows us to study mtDNA heteroplasmy and copy number simultaneously. Previous large-scale studies of mtDNA heteroplasmy or copy number mostly used sequencing data of total genomic DNA extracted from transformed cell lines or whole blood. It is possible that during cell line transformation, both mtDNA heteroplasmy and copy number could undergo marked changes [26]. Moreover, estimating mtDNA copy number from WGS data relies on the ratio of sequencing reads for the mitochondrial and nuclear genomes extracted from the biological samples. There are numerous factors in the whole blood that can bias the estimation of mtDNA copy number. For example, platelets have high mtDNA content, but lack nuclear DNA; mtDNA from platelets therefore artificially raises the estimated mtDNA copy number from the whole blood [27, 28]. In the current study, we focused our analysis on WGS data of isolated platelet-free peripheral blood mononuclear cells (PBMCs) DNA obtained from the UK10K project TwinsUK cohort, which includes individuals ranging from 17-85 years of age [29]. The TwinsUK is a cohort with WGS data for more than 1,500 generally healthy female individuals of European ancestry with phenotypic data. The resulting mtDNA-phenotypic dataset is one of the largest available in a general human population for analysis on the relationship between age and mtDNA heteroplasmy and copy number. Our analyses reveal that these two mtDNA properties are significantly correlated with age. Our results further indicate that mtDNA copy number and heteroplasmy load

were significantly associated with age-related physiological parameters in this population, suggesting potential pathways by which age-related mtDNA alterations may impact the aging process.

## 2.3 Methods

### 2.3.1 Data Access Permission

Data used in this study was obtained from UK10K project, "UK10K Data Access Agreement" was approved by UK10K Data Access Officer.

### 2.3.2 mtDNA variation identification and haplogroup assignment

Whole genome sequencing and subsequent read mapping of the TwinsUK cohorts were accomplished by the UK10K project [29]. Briefly, DNA (1-3 μg) extracted from PBMCs was sheared to 100-1000 bp, and sheared DNA was sent to Illumina paired-end DNA library preparation. After size selection (300-500 bp insert size), the DNA library was sequenced by the Illumina HiSeq platform with paired-end read lengths of 100 bp. Sequencing reads mapping to the mitochondrial genome were extracted from indexed bam files to identify heteroplasmy in each individual. Retrieved reads were re-mapped to the combined human genome, hg19 for the nuclear genome and the revised Cambridge Reference Sequence (rCRS) for the mitochondrial genome, using bowtie2 [30]. Read pairs with proper orientation and less than 5 mismatches were retained from the mapping results. The nuclear genome contains some regions with high similarity to part of the mtDNA (nuclear mitochondrial DNA, abbreviated as NUMTs). To minimize the effect of NUMTs for heteroplasmy calling, we further

required the retained reads to be uniquely mapped to the mitochondrial genome. Filtered reads were further processed following the GATK best practice workflow, including Mark duplicates (duplicated reads were removed), Indel realignment, and Base quality score recalibration steps. Homoplasmies were identified using GATK HaplotypeCaller and GenotypeGVCFs [31]. Haplogroups were assigned using homoplasmic variants identified from each sample by HaploGrep2 [32]. To identify heteroplasmy, sequencing data for each position of the mitochondrial genome was extracted by Samtools mpileup [33], and bases were further filtered by sequencing quality (>=20). Heteroplasmy was then identified with the following criteria: 1) Sequencing coverage > 200. 2) Minor allele frequency >= 2%. 3) Minor allele must be observed at least twice from each strand.

### 2.3.3 Potential cross-sample contamination inspection

Potential cross-sample contamination was assessed in the UK10K project's original data processing by VerifyBamID [34] and "fraction skewed hets" [29]. Potential contaminated samples were already removed from the dataset. However, to be more conservative, we further tested contamination using mtDNA sequencing data, which has better sensitivity than nuclear DNA. We evaluated potential contamination by 2 criteria 1) if a sample had extremely high heteroplasmy number (Q3 + 1.5IQR rule); by this criterion, samples having more than 8 heteroplasmies were suspected to be contaminated. 2) We constructed two consensus mtDNA sequences for each sample, one covering the major alleles at heteroplasmic sites, the other covering minor alleles. A sample was suspected to be contaminated if these two consensus sequences

belonged to different haplogroups. If a sample met both criteria, we would recognize it as contamination and remove it from further analysis.

### 2.3.4 Annotation of mtDNA variants

Heteroplasmy and homoplasmy were annotated by customized scripts. Pathogenic potential of variants was predicted using Combined Annotation-Dependent Depletion (CADD) score (version 1.3) [35]. The CADD score integrated many diverse annotations of variants, including functionality, pathogenicity, experimentally measured effects etc., into a single score, which has been shown to have better performance than other predictive methods such as Grantham, SIFT and PolyPhen. As recommended, a scaled CADD score of 15 was used to define the pathogenic mutations. To avoid the bias of an arbitrary cutoff, a series of cutoffs from 12 to 22 were also applied to evaluate the variants' pathogenicity. The disease associated mtDNA mutations were obtained from the MITOMAP database [36].

### 2.3.5 mtDNA copy number estimation

Whole genome sequencing data of the study population were retrieved from the UK10K project [29]. To estimate mtDNA copy number, we further filtered mapped reads by the following criteria: 1) Mapping quality > 20. 2) Reads were not PCR duplicates. 3) Mismatches < 5. We proceeded with qualified reads for subsequent calculation. Sequencing coverage of each site in the reference genome was calculated by the Samtools mpileup function [33]. The average sequencing coverage was then calculated for each autosomal DNA and mtDNA locus. mtDNA copy number of each

individual was further estimated based on **Eq1** and **Eq2**. It has been shown that NUMTs have a negligible impact on mtDNA copy number estimation by this method [37].

$$\frac{mtDNA\ average\ coverage}{autosomal\ DNA\ average\ coverage} = \frac{mtDNA\ copies}{autosomal\ DNA\ copies} \quad (1)$$

$$mtDNA\ copy\ number = \frac{2 * mtDNA\ coverage}{\frac{1}{22}\sum_{i=1}^{22} autosomal\ coverage} \quad (2)$$

2.3.6 Association testing for mtDNA copy number and heteroplasmy

Linear regression was carried out to test the association of mtDNA heteroplasmy with age, as well as copy number with age separately. Down sampling was carried out by randomly sampling 0.06 million mtDNA reads from each individual and identifying heteroplasmy following the same criteria. The association of mtDNA copy number with mtDNA heteroplasmy number was assessed by linear regression, with age and mean nuclear coverage included as covariates. The influence of mtDNA variants (heteroplasmy, homoplasmy, haplogroup) on mtDNA copy number was tested by linear regression, with age and mean nuclear coverage included as covariates. For homoplasmy, variants present in more than 1% individuals were tested. The significance level was adjusted for multiple testing. Homoplasmic variants with *P* value < 2.69e-4 were considered to be significant.

2.3.7 Phenotypic associations of mtDNA copy number and heteroplasmy

We assessed the associations of 32 phenotypes directly measured from blood samples with mtDNA copy number and heteroplasmy number. The details of phenotypic data

measurements can be found from the UK10K project [29]. Linear regression models were applied to test for associations, and age was included as a covariate to adjust for its effects on these phenotypes. Significance levels were adjusted by Bonferroni correction. Effects were considered as significant if *P* value < 0.0016.

2.3.8 Mitochondrial heteroplasmy load and SKAT test

The Sequence Kernel Association Test (SKAT) has been shown to have high statistical power under a variety of conditions [38]. We used it to test the association of mtDNA heteroplasmic mutation load with different phenotypes. To do the association test, we first constructed a genotype matrix containing mtDNA heteroplasmy information. Assuming that we have mtDNA sequences for **n** individuals, and there are, in total, **m** unique heteroplasmic variants among those individuals, then the genotype matrix could be constructed an **n** x **m** matrix **X**, where the entry $\mathbf{a_{ij}}$ in **X** represents the minor allele frequency of individual i at heteroplasmic site j. To calculate $\mathbf{a_{ij}}$, the number of all possible bases (A, T, G, C) with sequencing quality > 20 were counted for individual i at site j. If the minor allele count exceeded 5, minor allele frequency would be calculated by dividing the minor allele count by total coverage at the given site, otherwise $\mathbf{a_{ij}}$ would be set to 0. We also constructed another genotype matrix **X'**, whose entries were either 0 or 1. This matrix will only consider whether a given site is heteroplasmic (as 1) or not (as 0), regardless of the minor allele frequency. Notably, the genotype matrices only contained the non-polymorphic heteroplasmies. After constructing the genotype matrix, a linear regression model can be considered (**Eq5**):

$$Y = \alpha_0 + C\alpha + X\beta \quad (5)$$

Where C is the covariates matrix, in which we included age, mtDNA copy number, and the top 2 PCs from principle component analysis of population structure. Phenotypes were log transformed to achieve normal distributed residuals. The CADD score of each heteroplasmic variant was used as the weights. This weighting scheme could upweight heteroplasmic variates which are predicted to be more deleterious. The tests were performed using the R package SKAT [38].

## 2.4 Results

### 2.4.1 Mitochondrial heteroplasmy is prevalent in UK10K TwinsUK cohort

The UK10K-cohorts arm [29] provided WGS data for healthy individuals from two British cohorts of European ancestry, namely the Avon Longitudinal Study of Parents and Children (ALSPAC) [39] and TwinsUK [40]. However, the genomic DNA used in these two cohorts was different. In ALSPAC, DNA was extracted from lymphoblastoid cell lines established *in vitro*, while DNA in TwinsUK was extracted from isolated PBMCs. We compared the mtDNA copy number distribution between the two cohorts and observed a dramatically higher mtDNA copy number in cell line DNA (Figure 2-1). The higher mtDNA copy number in the cell line compared to PBMCs is likely attributable to the difference in biological material, rather than a genuine cohort difference. Therefore, in the current study, we only focused on individuals from the TwinsUK cohort for which PBMC DNA was available. In the UK10K project's original study design, in order to increase the genetic diversity and decrease the sequencing costs, only one individual from each of the twin pairs

recruited was sequenced at the whole-genome level. After excluding 73 individuals with potential cross-sample contamination, we retained 1,511 individuals for further analysis. The average age of these studied individuals was 55.5 years (SD = 12.8 years), ranging from 17.3 years to 84.5 years.

The average sequencing coverage of the mitochondrial genome in these individuals was ~568X, allowing us to reliably identify mtDNA heteroplasmy at 2% minor allele frequency (MAF) cutoff. After applying a series of criteria to filter out low-quality heteroplasmies, we identified 1,348 mtDNA heteroplasmies in 1,511 individuals. 794 (52.5%) individuals harbored at least one heteroplasmy in the mitochondrial genome (Figure 2-2**A**). Most heteroplasmies presented at low-to-medium frequency (62.7% of heteroplasmies have MAF < 5%, Figure 2-2**B**). The gene-length normalized distribution of heteroplasmy frequency among mtDNA loci is shown in Figure 2-2**C**. The distribution of homoplasmy frequency was also plotted. Heteroplasmies were

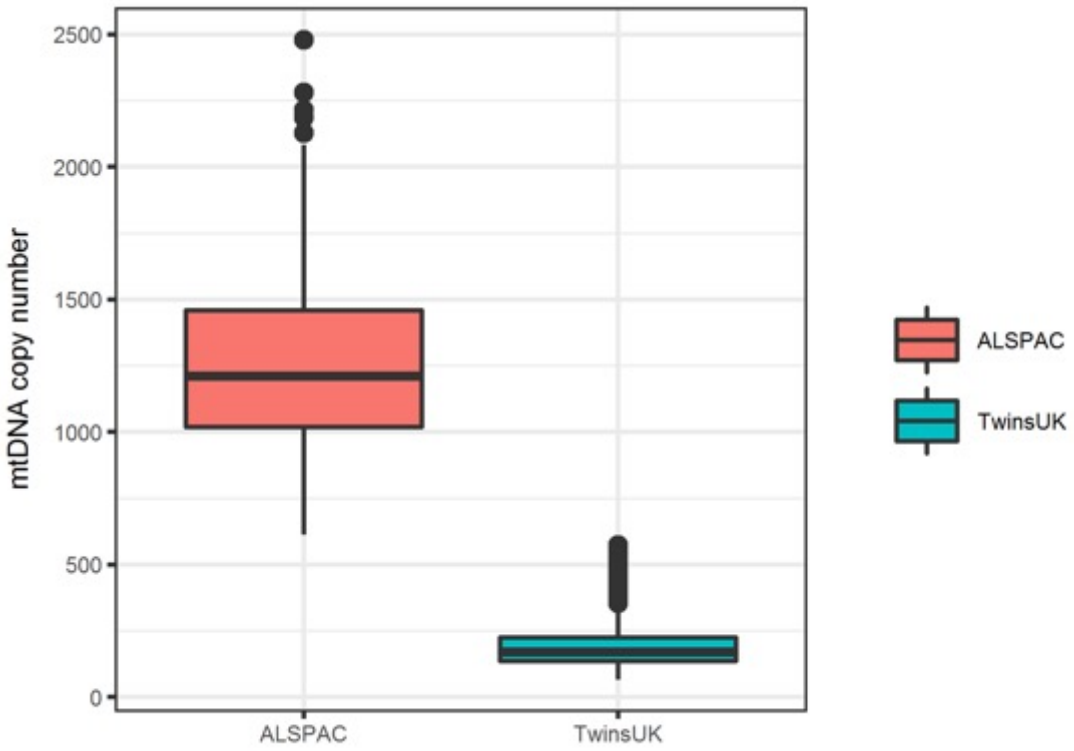


Figure 2-1 mtDNA copy number comparison between cell lines and PBMCs DNA.

mtDNA copy number was estimated in two UK10K cohorts, ALSPAC (DNA extracted from cell line) and TwinsUK (DNA extracted from PBMCs). On average, mtDNA copy number in cell lines was 5~10-fold higher than in PBMCs.

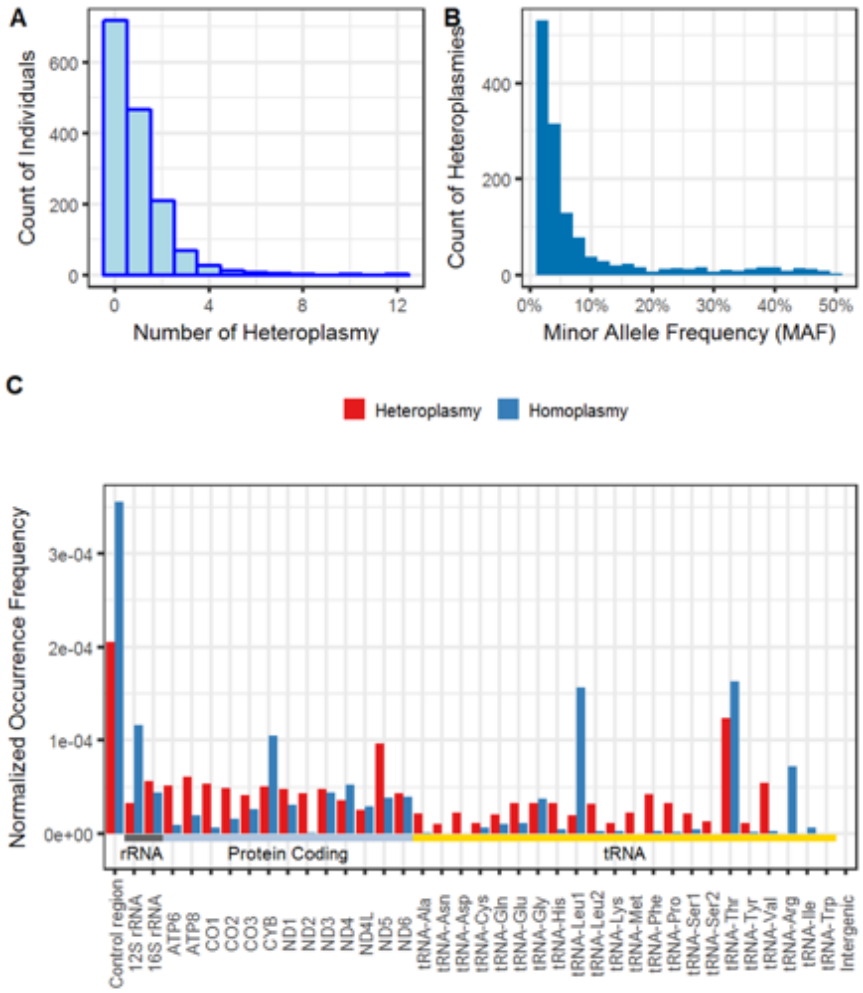


Figure 2-2. Distribution of heteroplasmy in UK10K TwinsUK cohort.

(A) Counts of individuals harboring a specific number of heteroplasmies (0-12 heteroplasmies at MAF > 2% cutoff). More than half of individuals (52.5%) carried at least one heteroplasmy in their genome. (B) Histogram for MAF of all heteroplasmies. 62.7% of heteroplasmies had MAF < 5% and 20.1% had MAF > 10%. (C) Normalized occurrence frequency distribution of heteroplasmies and homoplasmies. The frequency was normalized by the length of the mitochondrial loci. Dark gray, light blue and yellow bars indicate the genes in three different functional categories: rRNA, Protein coding and tRNA, respectively. The distribution of variants was relatively homogeneous among coding regions, except for some regions, such as higher frequency in ND5 (heteroplasmy) and tRNA Thr (heteroplasmy and homoplasmy).

observed over the entire mitochondrial genome. One exception was the control region, also known as the "hypervariable region" [41], which harbored the highest occurrence (normalized by region length) of both homoplasmic and heteroplasmic variants. Other regions were relatively homogenous with a few exceptions: tRNA-Thr, which is positioned immediately upstream from the control region, had significantly higher frequencies than other tRNA genes in both heteroplasmy and homoplasmy ($P$ = 0.00015 and 0.00297, respectively, Chi-squared outlier test). Interestingly, we observed that ND5 had a moderate occurrence frequency in homoplasmy, but significant higher frequency in heteroplasmy than other protein coding genes ($P$ = 0.00494, Chi-squared outlier test).

### 2.4.2 Mitochondrial heteroplasmy has high pathogenic potential

Among the 1,348 heteroplasmies, 192 (14.2%) were previously reported to be associated with diseases. 11.4% of individuals harboring these diseases-associated heteroplasmies. To further investigate the pathogenicity of the heteroplasmies, combined annotation dependent depletion (CADD) scores [35] were used to predict the potential pathogenicity of nonsynonymous mutations in heteroplasmy and homoplasmy. We also annotated CADD scores for disease-associated mutations (retrieved from MITOMAP [36]) as a comparison. Disease-causing nonsynonymous mutations had a mean pathogenicity score of 17.43; in comparison, the mean CADD score of the 294 unique nonsynonymous heteroplasmic mutations was 14.32, which is significant higher than the 359 unique nonsynonymous homoplasmic mutations (10.84. $P$ = 3.967e-7, Welch two sample t-test, Figure 2-3**A**). As suggested by the

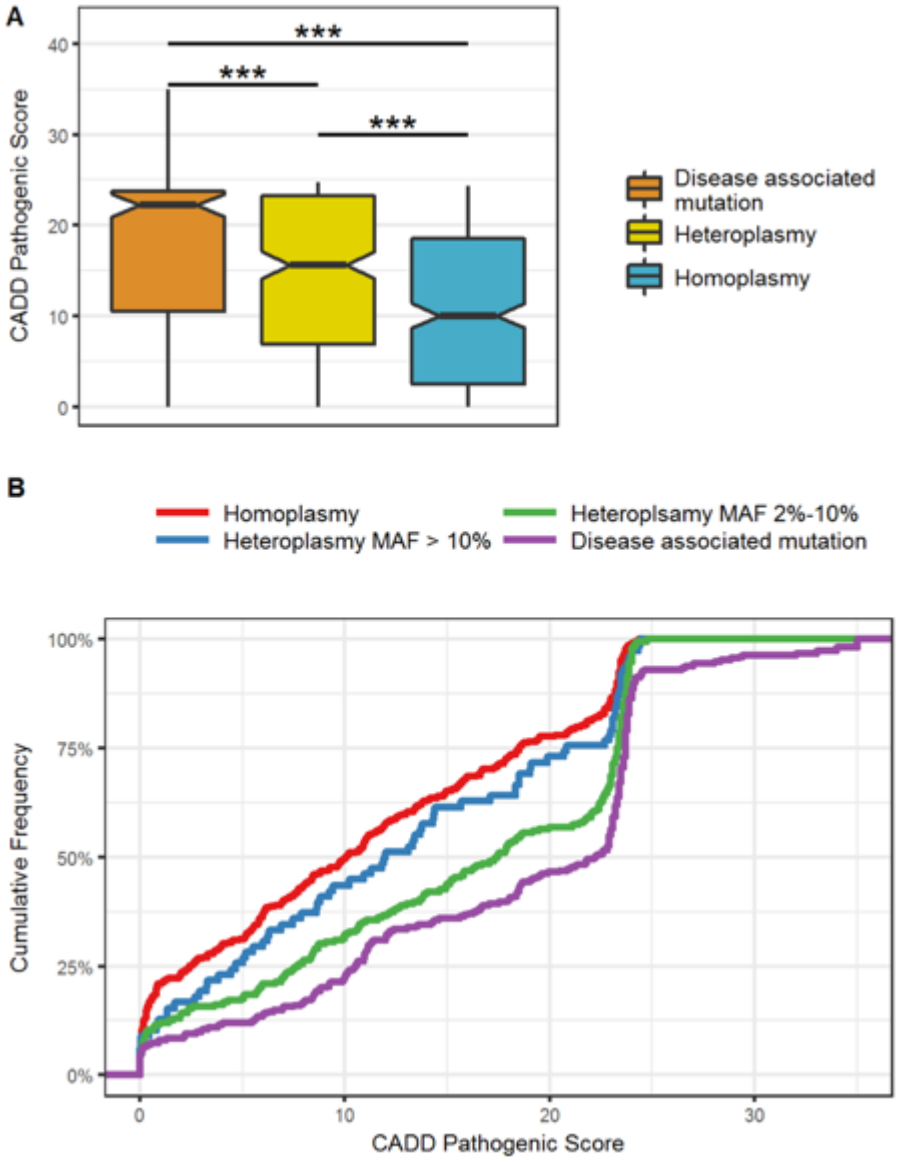


Figure 2-3. Pathogenic potential for nonsynonymous heteroplasmies.

(A) The box plot of CADD pathogenic score for disease associated mutations, nonsynonymous heteroplasmies and nonsynonymous homoplasmies (heteroplasmy and homoplasmy occurring in multiple individuals were counted only once). Heteroplasmies had significant higher pathogenic scores than homoplasmies ($P$ = 3.967e-7) although still lower than disease associated mutations ($P$ < 2.2e-16). (B) The cumulative distribution of CADD pathogenic scores of disease associated mutation, homoplasmy, low frequency heteroplasmy (MAF 2%-10%) and high frequency heteroplasmy (MAF > 10%). The distribution of low frequency heteroplasmy was close to disease associated mutations, indicating higher pathogenic potential.

CADD score database, CADD score > 15 can be used as a cutoff to define mutations with high possibility to be pathogenic [35]. With this cutoff, the proportion of high pathogenic potential mutations in heteroplasmy was significantly higher than that in homoplasmy ($P = 0.00825$, Chi-square test). With the same criterion, 51.4% of heteroplasmies were high pathogenic potential while only 34.8% of homoplasmic mutations were high pathogenic potential. To avoid the potential bias of the arbitrary cutoff, we also applied a series of CADD score cutoffs from 12 to 22, and heteroplasmy was 1.42 to 1.94 times more likely to be high pathogenic potential than homoplasmy under different cutoffs, consistent with the notion that more pathogenic mutations are more likely to be eliminated through purifying selection than less deleterious ones [8].

To further investigate this hypothesis, we separated nonsynonymous heteroplasmy into low frequency and high frequency groups using 10% MAF as a cutoff. The low frequency heteroplasmy group had significantly higher CADD scores than the high frequency group ($P = 0.019$, Welch two sample t test). Again, to avoid the potential bias of arbitrary cutoffs, we applied several MAF frequency cutoffs to separate low and high frequency groups (from 5% to 9%), and the difference remained significant until the cutoff was as low as 6%. To visualize this difference, we plotted the cumulative distribution of CADD scores for each group. The distribution of CADD scores for low frequency heteroplasmies approached that of disease-associated mutations, whereas the distribution of high frequency heteroplasmies moved towards that of homoplasmic mutations (Figure 2-3**B**).

### 2.4.3 Mitochondrial heteroplasmy burden increases with age

The mtDNA haplogroup of each individual was identified using Haplogrep2 [32]. In this population, 48.5% (733) of the individuals belonged to H haplogroup, and there were 4 other haplogroups having more than 100 individuals: U (220), K (143), J (135) and T (125). This distribution is typical for a population of predominantly European descent. The haplogroups did not significantly contribute to the heteroplasmy variance ($P > 0.05$ for each haplogroup), and thus were not considered in subsequent analysis.

mtDNA mutations have been thought to play an important role in aging. To investigate changes of heteroplasmies with age, we first applied linear regression and found that heteroplasmy number increased with age ($\beta = 0.011$, $P = 5.77e\text{-}6$, linear regression. Figure 2-4). To better describe the changing heteroplasmy trend during aging, we further divided the 1,511 individuals into five age groups. We observed a gradual and consistent increase of heteroplasmy number from the youngest group aged under 40-years to the oldest group aged over 70-years (Table 2-1). On average, individuals over 70-years old had 1.11 heteroplasmies, significantly higher than individuals under 40-years old (0.70 heteroplasmy, $P = 0.001593$, Welch two sample t test). We also separated heteroplasmy into low-to-medium MAF (2%-5%) and medium-to-high MAF (>5%) intervals and found that this increasing trend was consistent for heteroplasmy in different MAF intervals. Individuals under 40-years old had 0.41 heteroplasmy with MAF 2%-5% and 0.29 heteroplasmy with MAF >5%, while individuals over 70-years old had 0.68 and 0.43, respectively.

Next, we evaluated the spectra of heteroplasmy in the five age groups. In all groups, heteroplasmy was predominantly present in protein coding regions, which was not surprising since protein coding sequences account for > 67% of mtDNA. However,

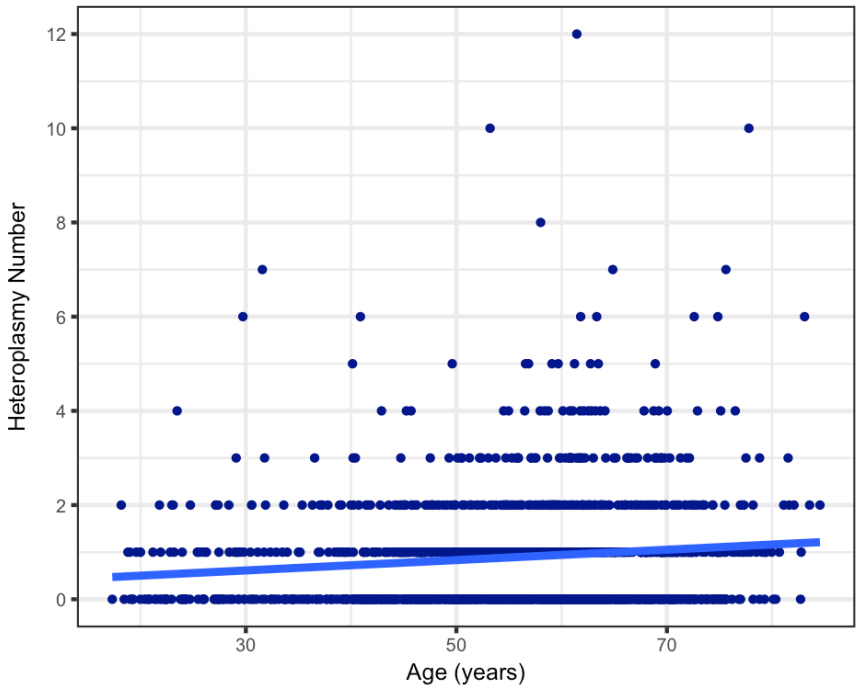


Figure 2-4 Association between mtDNA heteroplasmy number and age.

mtDNA heteroplasmy number was significantly associated with age ($\beta$ = 0.011, $P$ = 5.77e-6,). The blue line represents a linear regression line.

Table 2-1. Heteroplasmy number in different age groups

| Age Group | < 40 | 40-50 | 50-60 | 60-70 | > 70 |
|---|---|---|---|---|---|
| **Age Mean** | 30.12 | 45.24 | 55.27 | 64.28 | 74.07 |
| **Age SD** | 6.41 | 2.72 | 2.92 | 2.94 | 3.35 |
| **Individual count** | 166 | 267 | 464 | 447 | 167 |
| **Heteroplasmy count**[a] | 0.70; (44.6%) | 0.72; (47.2%) | 0.89; (55.0%) | 0.98; (53.5%) | 1.11; (60.0%) |
| **Heteroplasmy count (MAF 2-5 %)** | 0.41; (28.3%) | 0.48; (32.2%) | 0.56; (36.9%) | 0.62; (38.7%) | 0.68; (39.5%) |
| **Heteroplasmy count (MAF > 5%)** | 0.29; (22.9%) | 0.24; (20.6%) | 0.33; (28.4%) | 0.37; (27.7%) | 0.43; (32.3%) |

[a.] numbers in parentheses indicate the proportion of individuals harboring heteroplasmy with specified MAF cutoffs. An individual can have heteroplasmies in both MAF groups.

there was a tendency for the proportion of nonsynonymous heteroplasmies to increase with age. 25.9% of heteroplasmies were nonsynonymous in the under 40-years group, which increased to 28.6% in the over 70-years group, while the proportion of synonymous heteroplasmies did not significantly change (Table 2-2). Since nonsynonymous mutations are more likely to cause functional consequences than synonymous ones, this increased nonsynonymous proportion, together with the increased absolute heteroplasmy number in older individuals, could suggest that mtDNA integrity, or "quality" deteriorates with age.

### 2.4.4 Age has effects on mtDNA heteroplasmy and copy number

Because mtDNA heteroplasmy level can be affected by copies of mtDNA in blood, the age-related increase of heteroplasmy may reflect the consequences of decreased mtDNA copy number in older individuals [37]. In normal human cells, there are two fixed copies of the nuclear genome, and therefore the ratio of average WGS sequencing coverage for mitochondrial and nuclear genomes can be used to estimate mtDNA copy number. Assuming that autosomal and mtDNA are processed and sequenced with no significant difference, average sequencing coverage should be proportional to DNA copy number for autosomal and mtDNA (**Eq1**), thus mtDNA copy number can be estimated using **Eq2**. By this method, we observed a broad range of mtDNA copy number among these individuals (Figure 2-5**A**), from 65 to 573, with mean 169 and median 188. The distribution of mtDNA

Table 2-2. Regional distribution of heteroplasmy in different age groups

| | Control region | Intergenic region | rRNA | tRNA | Nonsynonymous | Synonymous |
|---|---|---|---|---|---|---|
| **<40** | 18.1% | 0.9% | 12.9% | 5.2% | 25.9% | 37.1% |
| **40-50** | 24.9% | 0.5% | 13.0% | 6.2% | 25.9% | 29.5% |
| **50-60** | 27.3% | 0.5% | 9.9% | 3.4% | 26.8% | 32.1% |
| **60-70** | 22.0% | 0.2% | 12.0% | 2.5% | 28.2% | 35.0% |
| **>70** | 16.8% | 0.0% | 13.5% | 4.9% | 28.6% | 36.2% |

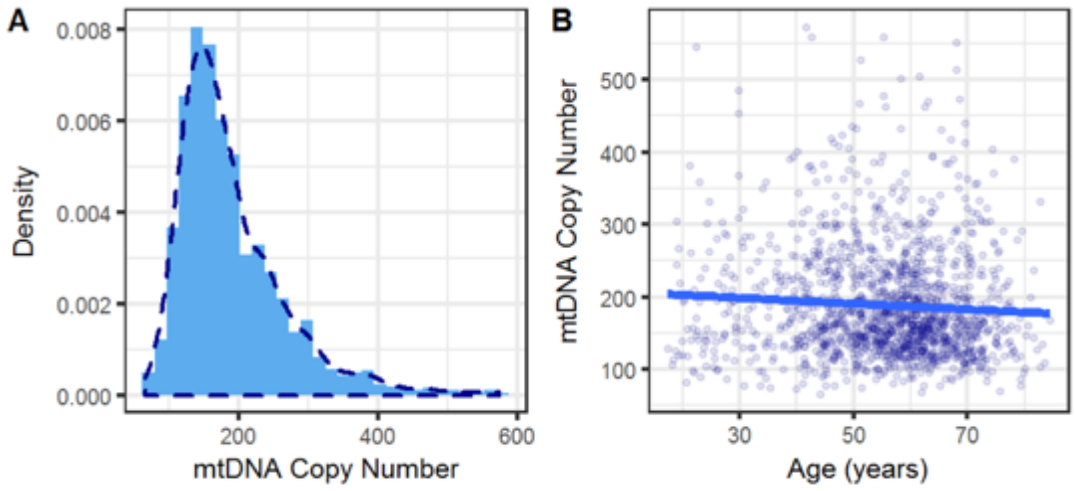


Figure 2-5. Distribution of mtDNA copy number in the UK10K Twins cohort and its association with age.
(A) mtDNA copy number was estimated using WGS data by comparing the mean sequencing coverage of mtDNA and nDNA. The distribution of mtDNA copy number was positively skewed, and most individuals had moderate numbers of mtDNA (mean 169 and median 188). (B) mtDNA copy number was negatively correlated with age ($\beta$ = -0.395, $P$ = 0.00972). Blue line represents the linear regression line. For every 10 years, mtDNA copy number decreases about 4 copies.

copy number was positively skewed ($P$ < 2.2e-16, D'Agostino's test), with a coefficient of skewness of 1.55. Our results showed that mtDNA copy number and age were negatively correlated ($\beta$ = -0.395, $P$ = 0.00972, linear regression, Figure 2-5**B**). For every 10 years, mtDNA copy number decreases about 4 copies. Similar to mtDNA heteroplasmy, mtDNA copy number was also not significantly affected by haplogroups.

Table 2-3. Correlation of age with mtDNA heteroplasmy number and copy number

| Parameter | Parameter Estimate | SE | P Value |
|---|---|---|---|
| **mtDNA heteroplasmy number** | 1.185 | 0.271 | 1.27e-5 *** |
| **mtDNA copy number** | -0.010 | 0.004 | 0.0228 * |

Significance level (* $P < 0.05$, ** $P < 0.01$ and *** $P < 0.001$)

### 2.4.5 Mitochondrial DNA copy number is associated with number of heteroplasmies

We next tested the correlation between mtDNA copy number and heteroplasmy. Since most heteroplasmies were unique to only one individual, especially those with high pathogenic potentials, instead of testing each single mtDNA heteroplasmy, our analysis was restricted to test the association between mtDNA copy number and the total number of heteroplasmies within an individual. With increasing heteroplasmy number, mtDNA copy number significantly decreased (Figure 2-6. $\beta$ = -4.34, $P$ = 0.007, linear regression, adjusted for age and average nuclear DNA sequencing coverage).

We also tested whether single mtDNA homoplasmic variants are associated with copy number. We identified 186 unique homoplasmic single nucleotide variants, each presented in > 1% of individuals in this study population. The associations between mtDNA copy number and these variants were tested using a linear model including age and mean nuclear DNA sequence coverage as covariates. After Bonferroni correction, none of these homoplasmic variants were significantly associated with mtDNA copy number (Figure 2-7). Ridge et al. previously reported that 3 mtDNA variants (A9667G, T5277C and C6489A), belonging to haplogroups T2 and U5A1, were significantly associated with higher mtDNA copy number [42]. There were 26 individuals in our dataset harboring A9667G, but this variant was not associated with mtDNA copy number in our test ($P$ = 0.5669). The other two variants were missing or found at a rare frequency (7 individuals) in our study, and thus were excluded from

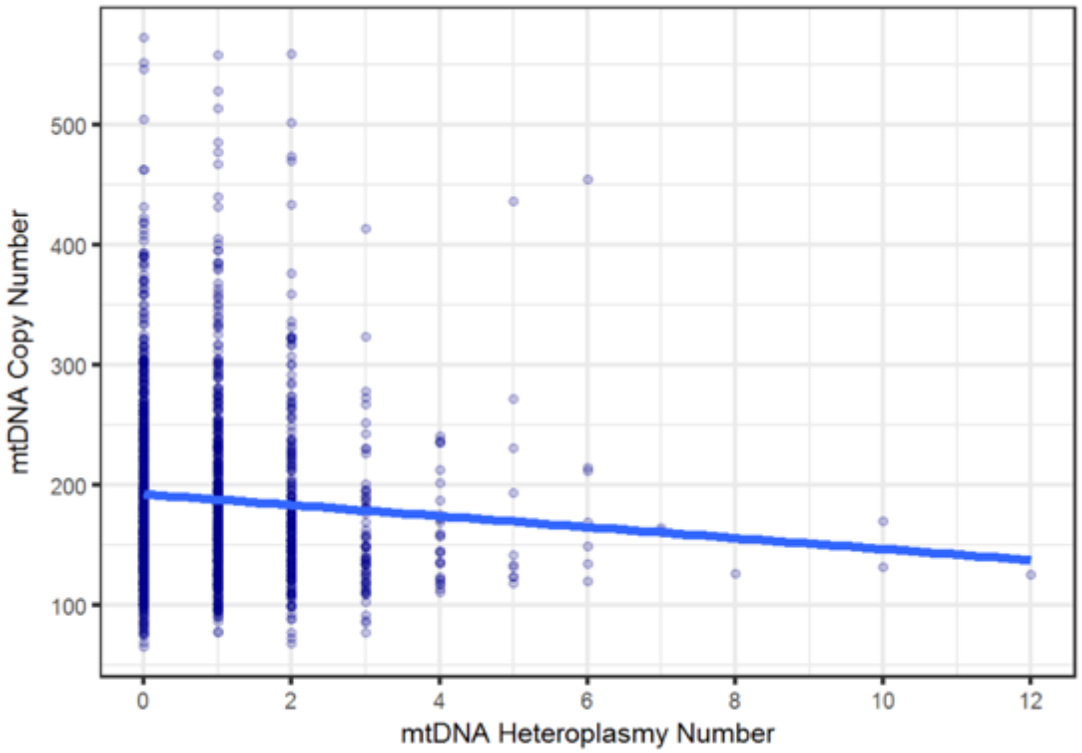


Figure 2-6. Association between mtDNA heteroplasmy number and copy number.

mtDNA copy number was significantly associated with the total heteroplasmy number within an individual, adjusting for age and mean nuclear sequencing coverage ($\beta$ = -4.34, $P$ = 0.007). Individuals harboring higher numbers of heteroplasmies were more likely to have low mtDNA copy number.

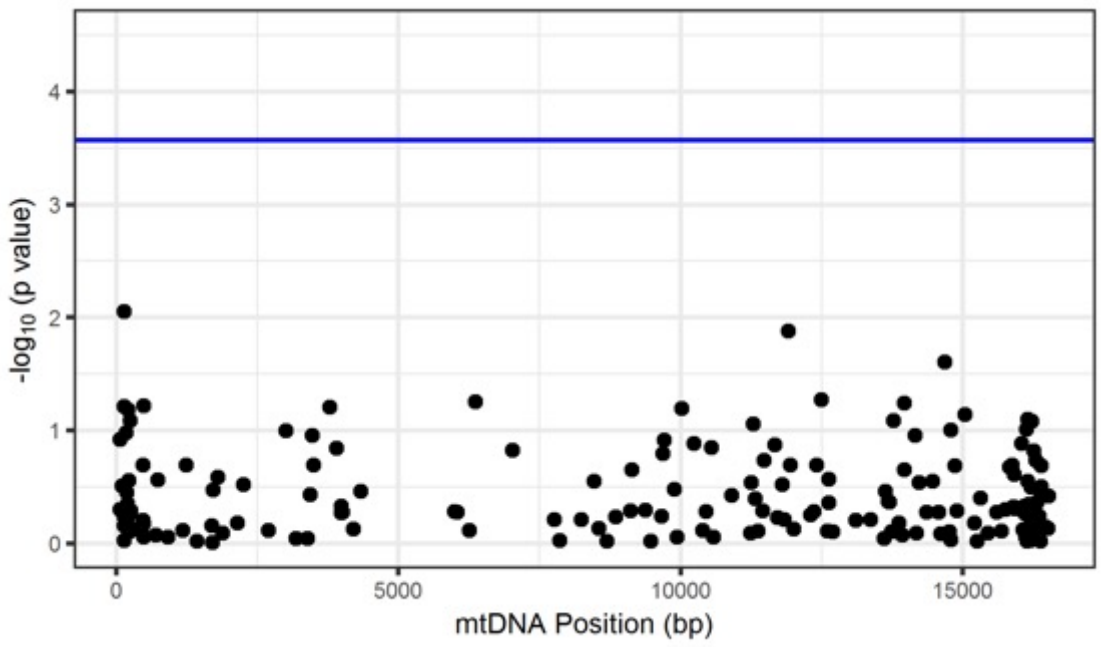


Figure 2-7. Manhattan plot of associations between homoplasmic variants and mtDNA copy number.
The Blue horizontal line indicates the mtDNA genome wide significance threshold (*P* = 2.69e-4). No significantly associated homoplasmy was detected in this analysis.

further association analysis.

### 2.4.6 Phenotypic associations of mtDNA copy number and heteroplasmy load

Age is the most significant risk factor for several diseases. It is possible that the effects of age on mtDNA copy number and heteroplasmy could mediate these effects via their effects on physiological variables known to be perturbed in disease states and with aging. We examined the associations between 32 phenotypic traits provided by TwinsUK cohort and mtDNA copy number / heteroplasmy load. After correcting for multiple testing, mtDNA copy number was significantly associated with serum bicarbonate level ($P$ = 4.46e-5, Figure 2-8**A**) and WBC count ($P$ = 0.0006, Figure 2-8**B**).

Bicarbonate is an essential component of the pH buffering system and is indirectly related to mitochondrial oxidative reactions. In our analysis, there was a positive correlation between mtDNA copy number and serum bicarbonate level, such that, for each increase of 1 SD in mtDNA copy number (75.8 copies), serum bicarbonate level increased by 0.102 SD (0.27 mmol/L), indicating a potential interplay between the buffering system and mitochondrial activity. Conversely, WBC showed a significant negative correlation with mtDNA copy number. With each increase of 1 SD in mtDNA copy number, WBC count decreased by 0.116 SD ($0.2*10^9$ cell/L). WBC count is related to inflammation and immune senescence, so this observation indicated that mtDNA copy number could be associated with immune function.

Since the majority of the heteroplasmies were present in < 1% of individuals in the samples, our ability to test their phenotypic associations were limited. Instead of

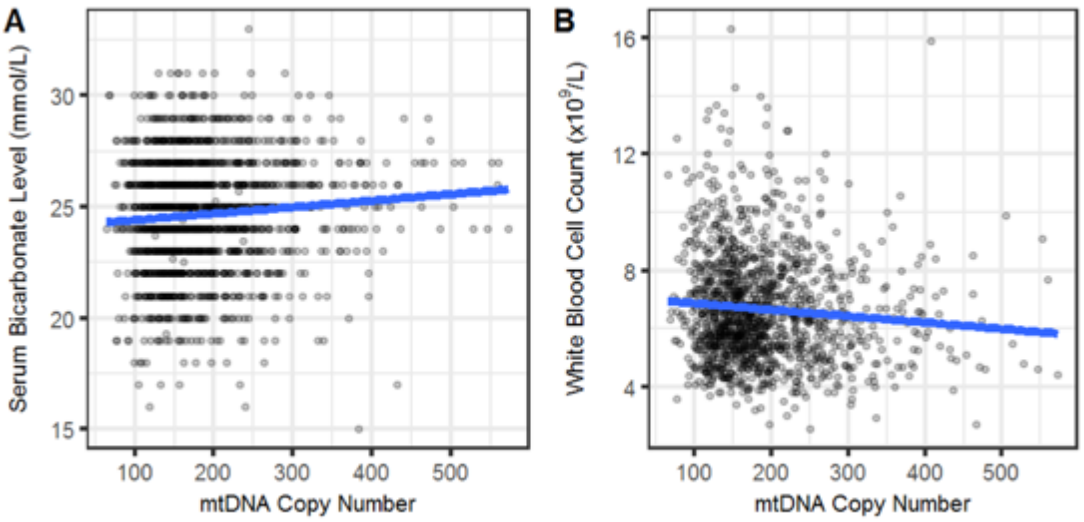


Figure 2-8. mtDNA copy number association with phenotypic traits.

(A) mtDNA copy number was positively associated with serum bicarbonate level ($P$ = 4.46e-5). The reference range for bicarbonate level is 22-29 mmol/L (B) mtDNA copy number was negatively associated with WBC count ($P$ = 0.0006). The reference range for WBC count is 4.0-11.0 (x$10^9$/L). The blue lines represent linear regression lines in each case.

performing analysis on each single heteroplasmy, we aggregated the heteroplasmic mutation information across the entire mitochondrial genome for each individual, and attempted to test the overall cumulative effects of heteroplasmy on different traits. We used the Sequence Kernel Association Test (SKAT) algorithm, which has been shown to have high statistical power under a variety of conditions [38]. Under SKAT default settings, the population frequencies of the variants were used as testing weights, since rare mutations were more susceptible to being deleterious. Here, because most heteroplasmies were only found in one person, we used the predicted pathogenicity of heteroplasmy (CADD scores) as weights. We tried two different genotype matrices, one taking the heteroplasmy MAF into account, the other only considering whether a site was a heteroplasmy or not, regardless of the MAF. In both cases, after multiple test correction, we observed that mtDNA heteroplasmy load was significantly associated with blood apolipoprotein B (ApoB) level ($P$ = 1.33e-5 and 3.73e-6, respectively), but not with other phenotypes. Because ApoB is a component of the lipid transport system linked to cardiovascular disease risk [43-46], this suggested a potential link between mtDNA integrity and physiological lipid regulation.

## 2.5 Discussion

In this study, we first identified mtDNA heteroplasmies in 1,511 generally healthy women using PBMC whole genome sequencing data from the UK10K project TwinsUK cohort, with an age range from 17-85 years of age. With 2% MAF cutoff, we demonstrated that more than half of the individuals (52.5%) harbor at least one

heteroplasmy in their mitochondrial genome, and on average each individual had ~0.9 heteroplasmy. Both the proportion of individuals harboring heteroplasmy and the average heteroplasmy number per person (using the same MAF 2% cutoff) were lower compared to our previous heteroplasmy study using sequencing data from the 1000 genome project, which utilized lymphoblastoid cell lines as source of DNA [8]. This difference could be caused by the difference between the sources of biological material: PMBCs versus cell lines. In cell line transformation, only a small proportion of original cells are induced, hence the heteroplasmy identified in a cell line only represents the heteroplasmy pattern for a few cells instead of the whole cell populations [26]. Because the cell line and PBMC samples show big differences in mtDNA characteristics (Figure 2-1), our study reinforces the notion that using DNA directly extracted from human samples is ideal for studying the impact of aging on mtDNA heteroplasmy and copy number. It is also important to point out that our analysis is limited to PBMCs, which are a mixture of different cell types. Further investigation of mtDNA characteristics in these different cell types and their relationship with aging will enable a better understanding on how mtDNA changes during aging.

We observed that heteroplasmy was not distributed uniformly across the mitochondrial genome, and several regions had enriched heteroplasmy: 1. The control region had the highest length normalized occurrence frequency; 2. Mutations located in tRNA-Thr had high occurrence frequency in both heteroplasmy and homoplasmy; 3. Notably, the normalized occurrence frequency of heteroplasmy in the ND5 gene was significantly higher than other protein coding genes, while the frequency of

homoplasmy in ND5 was comparable to other genes. mtDNA mutations in the tRNA-Thr and ND5 regions have been reported to be implicated in diseases including Leigh syndrome, mitochondrial myopathy, Parkinson's disease and thyroid cancer [36]. The high occurrence of mtDNA heteroplasmy in those regions may be a potential source of future diseases or could reflect an underlying prodromal disease state that independently promotes the accumulation of mtDNA defects.

Using CADD score as a measure of pathogenicity, we observed higher pathogenic potential of mtDNA heteroplasmy compared to homoplasmy, although still lower than disease-associated mutations. We further grouped heteroplasmies by their MAF, and found that heteroplasmies with lower MAF were more pathogenic than the ones with higher MAF, implying that the selective pressure on highly pathogenic heteroplasmies could be stronger, which could occur during germline selection, and hence reduce those heteroplasmies to low frequency. Due to mitochondrial threshold effects [47], highly-pathogenic heteroplasmies can persist in healthy individuals at low frequency; however, once they reach high frequency, they could potentially contribute to mitochondrial dysfunction and further lead to the onset and/or progression of various age-related diseases, as previously suggested [48-50].

It has been proposed that patients with mitochondrial diseases experience a monoclonal expansion of a single deleterious mtDNA mutation (for example, 3243A>G), whereas aging is associated with a mosaic of multiple low-level mtDNA mutations accumulated during a lifetime [51]. mtDNA is replicated throughout the lifetime of an individual, independent of cell cycle. Both inherited and de novo mutations that emerged early in life could clonally expand to increase the

heteroplasmy burden over time in a sub-population of cells. Several studies have reported that the amount of mtDNA mutation increases with age in several human tissues, including muscle, colon, putamen and heart [52-55]. Consistent with these reports, we observed that the heteroplasmy burden was elevated in older individuals. In the current sample, the absolute heteroplasmy number increased by 58.5% in individuals over 70-years (mean age 74.04) compared to individuals under 40-years (mean age 30.12). Meanwhile there was a trend for an increasing proportion of nonsynonymous heteroplasmy in older individuals. Given that the individuals involved in this study were generally healthy, it is possible that in aged individuals with diseases, more pronounced increases of heteroplasmy burden and pathogenicity would be observed. Future large-scale prospective studies should investigate the relationship between mtDNA heteroplasmy, disease status, and mortality.

Besides mtDNA quality, mtDNA quantity has also been suspected to be influenced by age. We estimated mtDNA copy number using WGS data, and found that mtDNA copy number was negatively correlated with age. These age-related mtDNA copy number changes were also reported in other studies. Wachsmuth *et al* suggested that mtDNA copy number decreased with age in human muscle tissue [37] and Sahin *et al* found a similar decrease in mice and rats in myocardial, hepatic, and hematopoietic cells [56, 57]. In measurements from whole blood, which are potentially confounded by several factors, Mengel-From et al. also reported a decline of 5.4 copies per decade of life in individuals above 48 years old [12].

However, no studies have investigated the effects of age on the two mitochondrial characteristics simultaneously, as it is possible that age can affect mtDNA copy

number through age-related heteroplasmy changes or vice versa. In this study, we demonstrated that age could affect both mtDNA copy number and heteroplasmy. Furthermore, compared to previous studies, we also included WBC count and platelet count as covariates in the regression model to adjust for potential bias caused by blood cell contaminations. Mitochondrial biogenesis has been proposed as a marker of many age-related health outcomes or even the aging process itself [58]. Our results suggested that both mtDNA heteroplasmy and copy number should be included to establish this relationship. Mitochondrial mutations that occur early in life can clonally expand to cause mitochondrial dysfunction and further contribute to aging through a number of potential mechanisms including decreased oxidative capacity and energy production capacity, but also nuclear signaling and transcriptional dysregulation [59-63]. In addition, decreased mtDNA copy number may also lead to decreased energy production and/or decreased mitochondrial gene expression [57, 64]. Maintaining both mtDNA quality and quantity together may help to counteract or slow down the aging process.

Our data are also consistent with the idea that mtDNA copy number and heteroplasmy can influence each other. We observed a negative correlation between mtDNA copy number and total number of heteroplasmies in an individual. In mitochondrial disease, a compensatory increase in mtDNA copy number via mitochondrial biogenesis may effectively compensate for heteroplasmic mtDNA mutations and mitochondrial dysfunction [65-67]. Thus, the observed age-related copy number decrease may result in a weaker copy number buffering effect during aging. In contrast, our results suggest that mtDNA haplogroups and homoplasmic variants were not strongly associated with

mtDNA copy number. Although haplogroup T2 has been reported to be associated with higher mtDNA copy number [42], we did not observe this association in our dataset. This may be caused by different sample sizes for this specific haplogroup. Our data had 177 individuals belonging to T2 while the conclusions in the previous study [42] only included 12 individuals. Another study suggested that haplogroup J had higher copy number compared to haplogroup H [68]. However, neither Wachsmuth et al [37] nor our data found this difference. It should be noted that our dataset only included UK females of European descent. To identify potential haplogroup-related effects on mtDNA copy number, further studies are needed to include a more ethnically-diverse range of populations, with both men and women, and larger sample sizes.

These age-related mitochondrial changes, combined with the fact that age is the main risk factor of many diseases in the population [69], further directed us to investigate mitochondrial associations with human physiological traits. After controlling for age, we found that serum bicarbonate level and WBC count were significantly associated with mtDNA copy number. The bicarbonate-carbon dioxide buffer system in blood can influence the pH gradient across the inner membrane of mitochondria, and thus may provide a link between systemic acid-base balance and regulation of mitochondrial metabolism [70, 71]. It has also been reported that reducing muscle hydrogen ion accumulation by sodium bicarbonate during running training was associated with greater improvements in both mitochondrial mass and mitochondrial respiration in rat models [72]. Our result is consistent with these reports and suggests a potential interplay between the bicarbonate buffer system and mitochondrial

biogenesis with aging.

WBC count is a well-established marker for inflammation [73, 74]. Its negative correlation with mtDNA copy number indicated a potential change in mitochondrial biogenesis during the immune response. Decreased peripheral blood mtDNA copy number is observed in various diseases accompanied with inflammation, for example, COPD was associated with decreased leukocyte mtDNA copy number [75] . Decreased mtDNA copy number was also observed to be significantly associated with adverse clinical outcomes in peritoneal dialysis patients [76]. Mitochondria play an important role in inflammatory signaling; conversely, inflammation may also damage mtDNA, promoting a vicious inflammatory cycle [77]. However, because WBCs are a mixture of different immune cells, a change in the composition of different immune cells, or all immune cell types undergo similar age-related changes in mtDNA copy number, may both contribute to the decrease of mtDNA copy number detected here. Further studies are needed to elucidate this observation. The link to specific pro- and anti-inflammatory biomarkers will also be important to resolve.

Most heteroplasmic variants had very low frequency in the population, which limited our ability to test for associations. Inspired by studies on nuclear DNA rare variants [78, 79], instead of evaluating single variants, we aggregated heteroplasmic mutations across the entire mitochondrial genome as a “heteroplasmy mutation load”, and tested the association between this mutation load and different healthy traits. By applying the SKAT algorithm, we found that mtDNA heteroplasmy load was significantly associated with blood ApoB level independent of age. Mitochondria play a critical role in fatty acid metabolism (eg, β-oxidation). Furthermore, ApoB is the main structural

surface protein found on all beta-lipoproteins, which is important for lipid transportation. The ApoB level is predictive for atherosclerosis [80], and the onset of obesity is usually accompanied by overproduction of ApoB [81]. Our result suggests a potential interaction between mitochondrial function and ApoB metabolism. It has been reported that the suppression of the PPARα signaling pathway would result in disrupted mitochondrial integrity and upregulated hepatic *apoB* gene expression at both the transcriptional and translational level in liver [82], providing a potential mechanism for how mitochondrial dysfunction is connected with ApoB metabolism. Nonetheless, further studies are needed to elucidate this connection.

One limitation of our study is that all participates were female. Given sex differences in mtDNA copy number measured in whole blood [24, 25, 83], our findings may not be representative for both men and women. In a study of whole blood, mtDNA copy number was previously reported to be associated with waist circumference and waist-hip ratio, suggesting an association between mtDNA copy number and fat distribution and lipid metabolism [25]. In our study, we did not observe these associations, which could possibly be caused by sex differences, or by other confounding factors (platelets, cell-free DNA, or other) in previous studies compared to purified leukocytes in this study.

## 2.6 Conclusions

In conclusion, using WGS data from the UK10K project TwinsUK cohort, we conducted, age has effects on both mtDNA heteroplasmy and copy number. Our analyses reveal that mtDNA copy number is inversely correlated with heteroplasmy

number, and associated with serum bicarbonate level and WBC count. Moreover, heteroplasmy load is associated with blood ApoB level, suggesting future avenues for research aimed at understanding the role of mitochondrial dysfunction in human aging. Mitochondria play a central role in cellular energy metabolism and regulate a broad range of cellular activities, and alterations of mtDNA sequence integrity and copy number have been implicated in human disease. Therefore, it remains promising to further investigate whether approaches to maintain mtDNA copy number and manage the expansion of mtDNA heteroplasmic mutations could help improve health status, especially in the elderly.

## 2.7 Acknowledgement

We thank Mr. Yiping Wang, Drs. Xiaoxian Guo and Yudong Li for their discussion and comments on the manuscript. This study makes use of data generated by the UK10K Consortium, derived from samples from EGAD00001000741, EGAD00001000790, EGAD00001000740. A full list of the investigators who contributed to the generation of the data is available from www.UK10K.org. Funding for UK10K was provided by the Wellcome Trust under award WT091310.

2.8 Reference


1. Schon, E.A., S. DiMauro, and M. Hirano, *Human mitochondrial DNA: roles of inherited and somatic mutations*. Nat Rev Genet, 2012. **13**(12): p. 878-890.
2. Stewart, J.B. and P.F. Chinnery, *The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease*. Nat Rev Genet, 2015. **16**(9): p. 530-542.
3. Picard, M., D.C. Wallace, and Y. Burelle, *The rise of mitochondria in medicine*. Mitochondrion, 2016. **30**: p. 105-116.
4. Calvo, S.E., K.R. Clauser, and V.K. Mootha, *MitoCarta2.0: an updated inventory of mammalian mitochondrial proteins*. Nucleic Acids Research, 2016. **44**(D1): p. D1251-D1257.
5. Lane, N. and W. Martin, *The energetics of genome complexity*. Nature, 2010. **467**(7318): p. 929-934.
6. Lightowlers, R.N., et al., *Mammalian mitochondrial genetics: heredity, heteroplasmy and disease*. Trends in Genetics, 1997. **13**(11): p. 450-455.
7. Russell, O. and D. Turnbull, *Mitochondrial DNA disease—molecular insights and potential routes to a cure*. Experimental Cell Research, 2014. **325**(1): p. 38-43.
8. Ye, K., et al., *Extensive pathogenicity of mitochondrial heteroplasmy in healthy human individuals*. Proceedings of the National Academy of Sciences, 2014. **111**(29): p. 10654-10659.
9. Reznik, E., et al., *Mitochondrial DNA copy number variation across human cancers*. eLife, 2016. **5**: p. e10769.
10. Schon, E.A. and G. Manfredi, *Neuronal degeneration and mitochondrial dysfunction*. Journal of Clinical Investigation, 2003. **111**(3): p. 303-312.
11. Kwak, S.H., et al., *Mitochondrial metabolism and diabetes*. Journal of Diabetes Investigation, 2010. **1**(5): p. 161-169.
12. Mengel-From, J., et al., *Mitochondrial DNA copy number in peripheral blood cells declines with age and is associated with general health among elderly*. Human Genetics, 2014. **133**(9): p. 1149-1159.
13. Lee, J.W., et al., *Mitochondrial DNA copy number in peripheral blood is associated with cognitive function in apparently healthy elderly women*. Clin Chim Acta, 2010. **411**(7-8): p. 592-6.
14. López-Otín, C., et al., *The Hallmarks of Aging*. Cell, 2013. **153**(6): p. 1194-1217.
15. Moskalev, A.A., et al., *The role of DNA damage and repair in aging through the prism of Koch-like criteria*. Ageing research reviews, 2013. **12**(2): p. 661-684.
16. Itsara, L.S., et al., *Oxidative stress is not a major contributor to somatic mitochondrial DNA mutations*. PLoS Genet, 2014. **10**(2): p. e1003974.
17. Ballard, J.W.O. and M.C. Whitlock, *The incomplete natural history of mitochondria*. Molecular ecology, 2004. **13**(4): p. 729-744.
18. Lynch, M. and B. Walsh, *The origins of genome architecture*. Vol. 98. 2007:

Sinauer Associates Sunderland.
19. Ross, J.M., et al., *Maternally transmitted mitochondrial DNA mutations can reduce lifespan*. Scientific reports, 2014. **4**: p. 6569.
20. Ross, J.M., et al., *Germline mitochondrial DNA mutations aggravate ageing and can impair brain development*. Nature, 2013. **501**(7467): p. 412-415.
21. Sondheimer, N., et al., *Neutral mitochondrial heteroplasmy and the influence of aging*. Human Molecular Genetics, 2011. **20**(8): p. 1653-1659.
22. Li, M., et al., *Extensive tissue-related and allele-related mtDNA heteroplasmy suggests positive selection for somatic mutations*. Proc Natl Acad Sci U S A, 2015. **112**(8): p. 2491-6.
23. Li, M., et al., *Transmission of human mtDNA heteroplasmy in the Genome of the Netherlands families: support for a variable-size bottleneck*. Genome Res, 2016. **26**(4): p. 417-26.
24. Knez, J., et al., *Correlates of Peripheral Blood Mitochondrial DNA Content in a General Population*. Am J Epidemiol, 2016. **183**(2): p. 138-46.
25. Ding, J., et al., *Assessing Mitochondrial DNA Variation and Copy Number in Lymphocytes of ~2,000 Sardinians Using Tailored Sequencing Analysis Tools*. PLoS Genet, 2015. **11**(7): p. e1005306.
26. Kang, E., et al., *Age-related accumulation of somatic mitochondrial DNA mutations in adult-derived human iPSCs*. Cell Stem Cell, 2016. **18**(5): p. 625-636.
27. Urata, M., et al., *Platelet contamination causes large variation as well as overestimation of mitochondrial DNA content of peripheral blood mononuclear cells*. Annals of clinical biochemistry, 2008. **45**(5): p. 513-514.
28. Hurtado-Roca, Y., et al., *Adjusting MtDNA Quantification in Whole Blood for Peripheral Blood Platelet and Leukocyte Counts*. PloS one, 2016. **11**(10): p. e0163770.
29. UKKC, *The UK10K project identifies rare variants in health and disease*. Nature, 2015. **526**(7571): p. 82-90.
30. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2*. Nat Meth, 2012. **9**(4): p. 357-359.
31. Van der Auwera, G.A., et al., *From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline*. Curr Protoc Bioinformatics, 2013. **43**: p. 11.10.1-33.
32. Weissensteiner, H., et al., *HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing*. Nucleic Acids Res, 2016. **44**(W1): p. W58-63.
33. Li, H., et al., *The Sequence Alignment/Map format and SAMtools*. Bioinformatics, 2009. **25**(16): p. 2078-9.
34. Jun, G., et al., *Detecting and Estimating Contamination of Human DNA Samples in Sequencing and Array-Based Genotype Data*. The American Journal of Human Genetics. **91**(5): p. 839-848.
35. Kircher, M., et al., *A general framework for estimating the relative pathogenicity of human genetic variants*. 2014. **46**(3): p. 310-5.
36. Ruiz-Pesini, E., et al., *An enhanced MITOMAP with a global mtDNA*

*mutational phylogeny*. Nucleic acids research, 2007. **35**(suppl 1): p. D823-D828.
37. Wachsmuth, M., et al., *Age-Related and Heteroplasmy-Related Variation in Human mtDNA Copy Number*. PLoS Genet, 2016. **12**(3): p. e1005939.
38. Ionita-Laza, I., et al., *Sequence Kernel Association Tests for the Combined Effect of Rare and Common Variants*. The American Journal of Human Genetics, 2013. **92**(6): p. 841-853.
39. Boyd, A., et al., *Cohort Profile: the 'children of the 90s'--the index offspring of the Avon Longitudinal Study of Parents and Children*. Int J Epidemiol, 2013. **42**(1): p. 111-27.
40. Moayyeri, A., et al., *The UK Adult Twin Registry (TwinsUK Resource)*. Twin Res Hum Genet, 2013. **16**(1): p. 144-9.
41. Stoneking, M., *Hypervariable sites in the mtDNA control region are mutational hotspots*. The American Journal of Human Genetics, 2000. **67**(4): p. 1029-1032.
42. Ridge, P.G., et al., *Mitochondrial genomic variation associated with higher mitochondrial copy number: the Cache County Study on Memory Health and Aging*. BMC Bioinformatics, 2014. **15**(7): p. S6.
43. Mahley, R.W., et al., *Plasma lipoproteins: apolipoprotein structure and function*. Journal of lipid research, 1984. **25**(12): p. 1277-1294.
44. Andrikoula, M. and I.F.W. McDowell, *The contribution of ApoB and ApoA1 measurements to cardiovascular risk assessment*. Diabetes, Obesity and Metabolism, 2008. **10**(4): p. 271-278.
45. Benn, M., *Apolipoprotein B levels, APOB alleles, and risk of ischemic cardiovascular disease in the general population, a review*. Atherosclerosis, 2009. **206**(1): p. 17-30.
46. Boekholdt, S.M., et al., *Association of LDL cholesterol, non–HDL cholesterol, and apolipoprotein B levels with risk of cardiovascular events among patients treated with statins: a meta-analysis*. Jama, 2012. **307**(12): p. 1302-1309.
47. Rossignol, R., et al., *Mitochondrial threshold effects*. Biochem J, 2003. **370**(Pt 3): p. 751-62.
48. Bender, A., et al., *High levels of mitochondrial DNA deletions in substantia nigra neurons in aging and Parkinson disease*. Nature genetics, 2006. **38**(5): p. 515-517.
49. Corral-Debrinski, M., et al., *Mitochondrial DNA deletions in human brain: regional variability and increase with advanced age*. Nature genetics, 1992. **2**(4): p. 324-329.
50. Wallace, D.C., *A mitochondrial paradigm of metabolic and degenerative diseases, aging, and cancer: a dawn for evolutionary medicine*. Annu. Rev. Genet., 2005. **39**: p. 359-407.
51. Park, C.B. and N.-G. Larsson, *Mitochondrial DNA mutations in disease and aging*. The Journal of Cell Biology, 2011. **193**(5): p. 809-818.
52. Bua, E., et al., *Mitochondrial DNA–deletion mutations accumulate intracellularly to detrimental levels in aged human skeletal muscle fibers*. The American Journal of Human Genetics, 2006. **79**(3): p. 469-480.

53. Greaves, L.C., et al., *Clonal Expansion of Early to Mid-Life Mitochondrial DNA Point Mutations Drives Mitochondrial Dysfunction during Human Ageing*. PLOS Genetics, 2014. **10**(9): p. e1004620.
54. Williams, S.L., et al., *Somatic mtDNA Mutation Spectra in the Aging Human Putamen*. PLOS Genetics, 2013. **9**(12): p. e1003990.
55. Cortopassi, G.A. and N. Arnheim, *Detection of a specific mitochondrial DNA deletion in tissues of older humans*. Nucleic acids research, 1990. **18**(23): p. 6927-6933.
56. Sahin, E., et al., *Telomere dysfunction induces metabolic and mitochondrial compromise*. Nature, 2011. **470**(7334): p. 359-365.
57. Barazzoni, R., K.R. Short, and K.S. Nair, *Effects of aging on mitochondrial DNA copy number and cytochromec oxidase gene expression in rat skeletal muscle, liver, and heart*. Journal of Biological Chemistry, 2000. **275**(5): p. 3343-3347.
58. Carré, J.E., et al., *Survival in critical illness is associated with early activation of mitochondrial biogenesis*. American journal of respiratory and critical care medicine, 2010. **182**(6): p. 745-751.
59. Raffaello, A. and R. Rizzuto, *Mitochondrial longevity pathways*. Biochim Biophys Acta, 2011. **1813**(1): p. 260-8.
60. Kroemer, G., L. Galluzzi, and C. Brenner, *Mitochondrial membrane permeabilization in cell death*. Physiol Rev, 2007. **87**(1): p. 99-163.
61. Green, D.R., L. Galluzzi, and G. Kroemer, *Mitochondria and the autophagy-inflammation-cell death axis in organismal aging*. Science, 2011. **333**(6046): p. 1109-12.
62. Bratic, A. and N.-G. Larsson, *The role of mitochondria in aging*. The Journal of clinical investigation, 2013. **123**(3): p. 951-957.
63. Picard, M., et al., *Progressive increase in mtDNA 3243A> G heteroplasmy causes abrupt transcriptional reprogramming*. Proceedings of the National Academy of Sciences, 2014. **111**(38): p. E4033-E4042.
64. Clay Montier, L.L., J.J. Deng, and Y. Bai, *Number matters: control of mammalian mitochondrial DNA copy number*. J Genet Genomics, 2009. **36**(3): p. 125-31.
65. Kauppila, T.E.S., J.H.K. Kauppila, and N.-G. Larsson, *Mammalian Mitochondria and Aging: An Update*. Cell Metabolism, 2017. **25**(1): p. 57-71.
66. Giordano, C., et al., *Efficient mitochondrial biogenesis drives incomplete penetrance in Leber's hereditary optic neuropathy*. Brain, 2014. **137**(2): p. 335-353.
67. Yu-Wai-Man, P., et al., *OPA1 mutations cause cytochrome c oxidase deficiency due to loss of wild-type mtDNA molecules*. Human molecular genetics, 2010: p. ddq209.
68. Suissa, S., et al., *Ancient mtDNA Genetic Variants Modulate mtDNA Transcription and Replication*. PLOS Genetics, 2009. **5**(5): p. e1000474.
69. Niccoli, T. and L. Partridge, *Ageing as a Risk Factor for Disease*. Current Biology, 2012. **22**(17): p. R741-R752.
70. Simpson, D.P. and S.R. Hager, *Bicarbonate-carbon dioxide buffer system: a*

*determinant of the mitochondrial pH gradient*. Am J Physiol, 1984. **247**(3 Pt 2): p. F440-6.
71. Durand, T., et al., *Role of Intracellular Buffering Power on the Mitochondria-Cytosol pH Gradient in the Rat Liver Perfused at 4°C*. Cryobiology, 1999. **38**(1): p. 68-80.
72. Bishop, D.J., et al., *Sodium bicarbonate ingestion prior to training improves mitochondrial adaptations in rats*. American Journal of Physiology - Endocrinology And Metabolism, 2010. **299**(2): p. E225-E233.
73. Pearson, T.A., et al., *Markers of inflammation and cardiovascular disease*. Circulation, 2003. **107**(3): p. 499-511.
74. Barati, M., et al., *Comparison of WBC, ESR, CRP and PCT serum levels in septic and non-septic burn cases*. Burns, 2008. **34**(6): p. 770-774.
75. Liu, S.-F., et al., *Leukocyte Mitochondrial DNA Copy Number Is Associated with Chronic Obstructive Pulmonary Disease*. PLOS ONE, 2015. **10**(9): p. e0138716.
76. Yoon, C.-Y., et al., *Low Mitochondrial DNA Copy Number is Associated With Adverse Clinical Outcomes in Peritoneal Dialysis Patients*. Medicine, 2016. **95**(7): p. e2717.
77. López-Armada, M.J., et al., *Mitochondrial dysfunction and the inflammatory response*. Mitochondrion, 2013. **13**(2): p. 106-118.
78. Arnedo, J., et al., *Uncovering the hidden risk architecture of the schizophrenias: confirmation in three independent genome-wide association studies*. Am J Psychiatry, 2015. **172**(2): p. 139-53.
79. Lohmueller, Kirk E., et al., *Whole-Exome Sequencing of 2,000 Danish Individuals and the Role of Rare Coding Variants in Type 2 Diabetes*. The American Journal of Human Genetics, 2013. **93**(6): p. 1072-1086.
80. Olofsson, S.O. and J. Boren, *Apolipoprotein B: a clinically important apolipoprotein which assembles atherogenic lipoproteins and promotes the development of atherosclerosis*. J Intern Med, 2005. **258**(5): p. 395-410.
81. Choi, S.H. and H.N. Ginsberg, *Increased very low density lipoprotein (VLDL) secretion, hepatic steatosis, and insulin resistance*. Trends in Endocrinology & Metabolism, 2011. **22**(9): p. 353-363.
82. Su, Q., et al., *Hepatic mitochondrial and ER stress induced by defective PPARα signaling in the pathogenesis of hepatic steatosis*. American Journal of Physiology-Endocrinology and Metabolism, 2014. **306**(11): p. E1264-E1273.
83. Reiling, E., et al., *The Association of Mitochondrial Content with Prevalent and Incident Type 2 Diabetes*. The Journal of Clinical Endocrinology & Metabolism, 2010. **95**(4): p. 1909-1915.

## *Chapter 3 – Heteroplasmy concordance between mitochondrial DNA and RNA*

### 3.1 Abstract

Mitochondrial DNA (mtDNA) heteroplasmies are associated with various diseases. Recent studies suggested that heteroplasmic mutations have high prevalence even in healthy individuals. However, the transmission of heteroplasmic variations from mtDNA to mitochondrial RNA (mtRNA) are rarely studied. In this study, we compared RNA sequences from 446 human lymphoblastoid cell lines to their corresponding DNA sequence, where RNA sequencing data was retrieve from Geuvadis RNA-seq project DNA sequencing data was from 1000 genome project. We identified 2786 heteroplasmies presenting in both DNA and RNA at 1% minor allele frequency (MAF) cutoff. The heteroplasmy frequencies were highly consistent between DNA and RNA: in 93.7% of these heteroplasmies, the frequency difference was less than 5%. We did not observe obvious negative selections for heteroplasmy during the transcription, indicating that the potential pathogenic heteroplasmies in DNA can be transcribed into RNA and may further cause deleterious consequences. In addition, owing to the heteroplasmy frequencies consistency between DNA and RNA, for future studies, RNA-seq data could be a potential source to infer mitochondrial heteroplasmic variations. We also investigated RNA editing/modification events in mitochondrial genome by investigating the heteroplasmies only observed in RNA. We confirmed the previously reported RNA editing sites in our data set and further identified novel editing/modification sites. Notably, we found RNA editing events are more frequent in African population than European populations, suggesting a potential

mitochondrial transcriptome regulatory transition during human out of Africa process.

## 3.2 Introduction

The genetic information from DNA need to be transcribed to RNA and eventually proteins to achieve biological functions. Therefore, the fidelity of the sequences transmission from DNA to RNA is critical, since the inconsistency introduced during the transcription process may contribute to genetic variations, further affect protein synthesis and/or the gene expression level. In addition, post-transcription RNA processing, such as RNA editing and modification, adds another layer of RNA sequence diversity. Studies has shown that there is widespread RNA and DNA differences (RDDs) in human nuclear genome [1, 2]. However, the knowledge of mitochondrial RDDs is still limited.

Different from only two copies of nuclear DNA (nDNA), there can be hundreds to thousands copies of mitochondrial DNA (mtDNA) existing within a single eukaryotic cell. The nature of multiple copies of mtDNA and their highly variable sequences make it possible that mutated mtDNA can co-exist with wild type mtDNA, which is termed as heteroplasmy [3]. Recently, there are many studies to show that mitochondrial heteroplasmic mutations are associated with broad spectrum of diseases (Chapter 1) and the allele frequency of a deleterious mutation would be critical for its pathogenicity [3-11]. Nevertheless, there are few studies to investigate the transmission of mtDNA heteroplasmic variations to mitochondrial RNA (mtRNA). Therefore, whether the mutated mtDNA and wild type mtDNA are transcribed proportionally is largely unknown. Moreover, RNA editing/modification also play a role of genetic regulation in many organisms [1]. However, these posttranscriptional processing of mtRNA is still subject to more comprehensive investigations.

Advances in sequencing technologies enable us to study heteroplasmy at single nucleotide resolution. By comparing the sequence of mitochondrial RNA and DNA from the same individual, we are able assess the RDDs in mitochondrial genome. In this study, we comprehensively analyzed 446 pairs of RNA and DNA sequences from human lymphoblastoid cell lines, which were a subset of individuals from 1000 genome project [12, 13]. We investigated whether the extensively existing heteroplasmic variations in mtDNA were also observable in mtRNA, moreover, whether the heteroplasmy frequency presented in RNA would keep consistent with that in DNA. Generally, we found that most heteroplasmy frequency difference between DNA and RNA were less than 5%. We also identified several potential mtRNA editing sites in these populations and found that genes related to RNA processing may contribute the individual variations of mtRNA modification/editing.

## 3.3 Methods

### 3.3.1 Data Retrieval and Pre-processing

DNA and RNA raw sequencing data was retrieved from 1000 Genome Project (http://www.internationalgenome.org/data) and Geuvadis RNA Sequencing Project (http://www.geuvadis.org/web/geuvadis/rnaseq-project), respectively. Totally, there were 446 pairs of DNA and RNA sequences. To retrieve mtDNA candidates reads, raw sequencing was first mapped to the mitochondrial genome. The mapped were then re-mapped to the combined human genome, hg19 for the nuclear genome and the revised Cambridge Reference Sequence (rCRS) for the mitochondrial genome with bowtie2 [14]. To remove possible contaminations from nuclear mitochondrial

sequence (NUMTs), we only retained reads which could be uniquely mapped to mitochondrial genome. Retained reads were further processed with the GATK best practice workflow, including Mark duplicates, Indel realignment, and Base quality score recalibration steps [15]. The sequencing information for each mtDNA sites were compiled with Samtools mpileup function [16].

### 3.3.2 Heteroplasmy identification and annotation

We first applied relative conservative criteria to identify heteroplasmies in DNA and RNA sequencing data separately at 1% minor allele frequency cutoff. The criteria were as followings: 1) Sequencing coverage > 200. 2) Minor allele frequency >= 1%. 3) Minor allele must be observed at least twice from each strand. We then integrate heteroplasmy information from both sides to evaluate heteroplasmy concordance between DNA and RNA. To avoid possible artifacts caused sequencing coverage, we only consider sites with depth > 200 in both DNA and RNA data. The heteroplasmies were grouped into the following categories: A heteroplasmy was grouped into "Observed in both DNA and RNA" (BDR) if it met any of the followings: 1. Identified as a heteroplasmy at 1% frequency cutoff in both DNA and RNA-seq data of the same individual. 2. If identified as a heteroplasmy only in DNA-seq data, the minor allele of this heteroplasmy needs had at least 3 sequencing reads to support in the RNA-seq data from the same individual. 3. If identified as a heteroplasmy only in RNA-seq data, the minor allele of this heteroplasmy needs had at least 3 sequencing reads to support in the DNA-seq data from the same individual. Otherwise, a heteroplasmy would be grouped into "Observed Only in DNA" (OD) and "Observed

Only in RNA" (OR) groups if it was identified in either DNA or RNA data. Heteroplasmies were annotated by customized scripts. Pathogenic potential of variants was predicted using Combined Annotation-Dependent Depletion (CADD) score (version 1.3) [17]. The disease associated mtDNA mutations were obtained from the MITOMAP database [18].

### 3.3.3 Cell line culture and point heteroplasmy sequencing

Cell lines were purchased from Coriell Institute (https://www.coriell.org/). Cells were grown at 37°C, 5% $CO_2$, in Roswell Park Memorial Institute Medium 1640 with 2mM L-glutamine and 15% fetal bovine serum. Cells were split every 2-3 days and fresh medium were replaced. Cells were collected at Day 0, 7, 14, 21 and 28, respectively. DNA and RNA were extracted from collected cells simultaneously with RNeasy Plus Mini Kit (Qiagen, catalog no.74134). RNA was reverse transcribed to cDNA with SuperScript™ III First-Strand Synthesis System (Invitrogen, catalog number: 18080051). Heteroplasmy sites of interest were then PCR amplified and sequenced using Illumina Miseq sequencing platform with pair-end 250 bp reads.

### 3.3.4 Genome wide association study to locate SNPs associated with editing events

SNPs from a list of 108 human RNA-binding and transcription-associated mitochondrial genes [19] were included in the association study. Association testing was performed with plink 1.9 [20] and the significance level was set at 7.8e-6 due to multiple testing.

## 3.4 Result

### 3.4.1 Identification of heteroplasmies using DNA and RNA sequencing data

Geuvadis RNA-seq Project [13] provided RNA-seq data for 446 individuals from a variety of human populations (One African population, and four European descent population). These individuals were a subset of 1000 Genome Project, thus the DNA-seq data from the same individual could be retrieved from 1000 Genome Project [12]. To investigate mitochondrial heteroplasmic variants, the sequencing reads were first aligned to human reference genome hg19, and after a series of quality control steps, we retained high quality mitochondrial DNA sequencing reads for subsequent mitochondrial heteroplasmy analysis(**Methods**). We examined two statistics of mitochondrial genome alignment results: median sequencing coverage and average cover rate of mitochondrial genome (Figure 3-1 A and B). The median sequencing coverage of mtDNA and mtRNA were 2077 and 4557 for these samples, respectively. The sequencing coverage of DNA was lower than that of RNA, but the distribution is more uniform across the entire mitochondrial genome (Figure 3-1 C). It's not surprising that mtRNA has fluctuating coverage distribution, since the coverage was highly affected by the gene expression level of each mitochondrial gene. The mitochondrial genome cover rate was calculated as the percentage of mitochondrial positions with sequencing coverage > 200. The average cover rates by DNA and RNA-seq data were 99.76% and 97.09%, respectively. These sequencing coverage and mitochondrial genome cover rate were sufficient to systemically investigate heteroplasmy at 1% frequency cutoff.

The details of heteroplasmy identification procedure can be found in Methods part. We first applied relative conservative criteria to call heteroplasmy in either DNA or

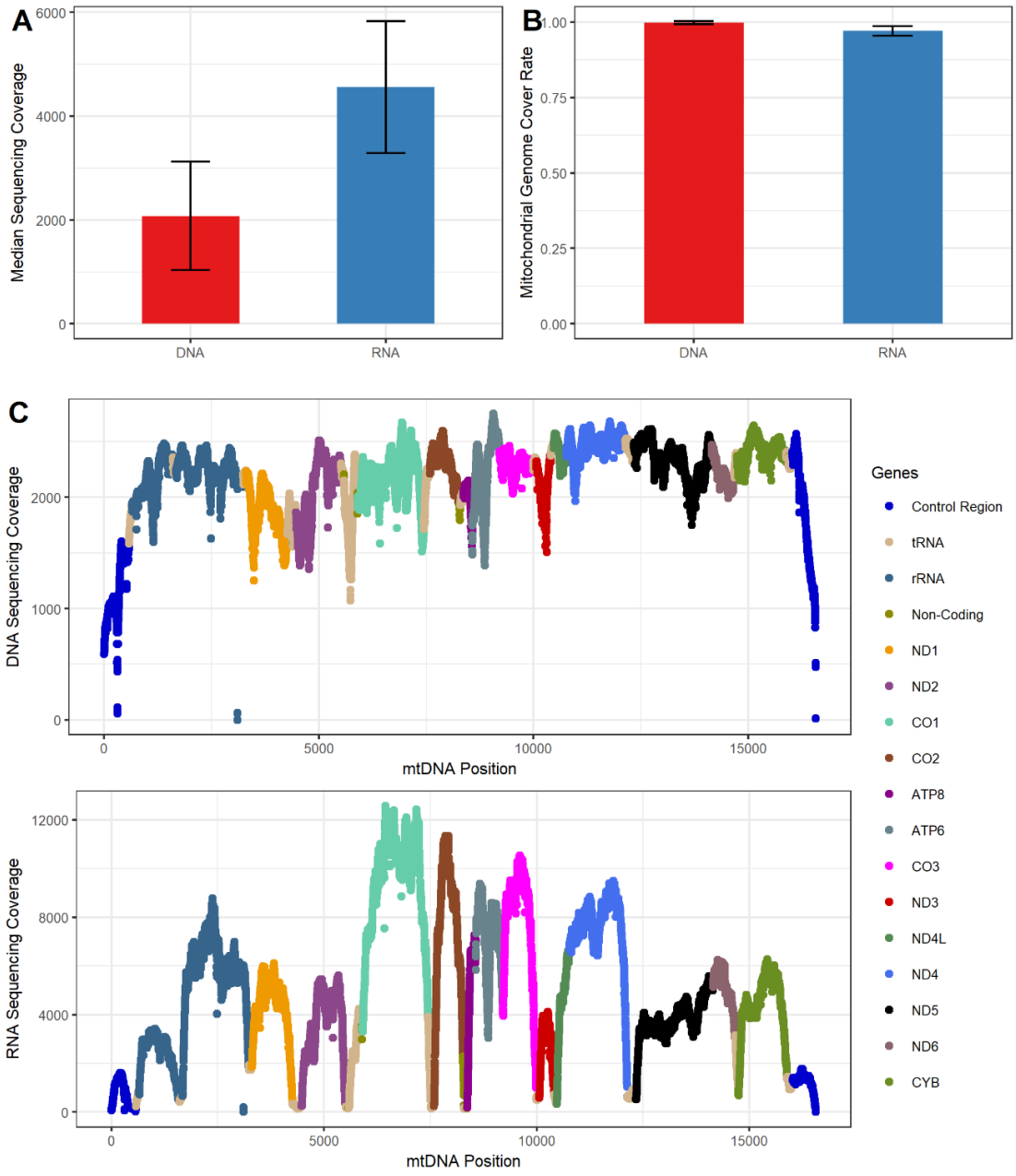


Figure 3-1 mtDNA sequencing statistics in DNA and RNA seq data.

(A) The median sequencing coverage of mtDNA and mtRNA among the 446 samples. (B) The mitochondrial genome cover rate by DNA and RNA seq data. (C) The sequencing coverage across the entire mitochondrial genome in DNA and RNA, respectively.

RNA sequencing data at 1% minor allele frequency cutoff. We then integrated the results from both DNA and RNA sides to separate heteroplasmies into "Observed in both DNA and RNA" (BDR), 'Observed Only in DNA" (OD) and "Observed Only in RNA" (OR) groups (**Methods**). Totally, we identified 2786 heteroplasmies presenting in both DNA and RNA, 219 only found in DNA and 682 only in RNA, respectively. Consistent with previous studies [21], most of the heteroplasmies has low frequencies. In DNA-seq data, 77.5% heteroplasmy has frequency < 5%, and 86.8% has frequency < 10%.

### 3.4.2 Compare heteroplasmy frequencies between DNA and RNA

According to the heteroplasmy threshold effects [22], heteroplasmy frequencies could be determinant for the phenotypic variations. Therefore, whether the heteroplasmic mutations in mtDNA can be transcribed into RNA proportionally could be important to investigate. In these datasets, we compared the heteroplasmy frequency differences between DNA and RNA. For each heteroplasmy in BDR group (**Methods**), we computed its MAF in DNA, then computed the frequency of the same allele in RNA. The frequency differences Δ were calculated as RNA-frequency minus DNA-frequency (RNA-DNA). In general, the heteroplasmy frequencies were highly consistent between DNA and RNA, 93.7% heteroplasmies had frequency difference < 5%, and 96.6% had frequency difference < 10% (Figure 3-2**)**, indicating that most of the heteroplasmy information can be transmitted faithfully from DNA to RNA. We also observed that some heteroplasmies had noticeable frequency difference from DNA to RNA. One possible explanation for the heteroplasmies with high frequencies

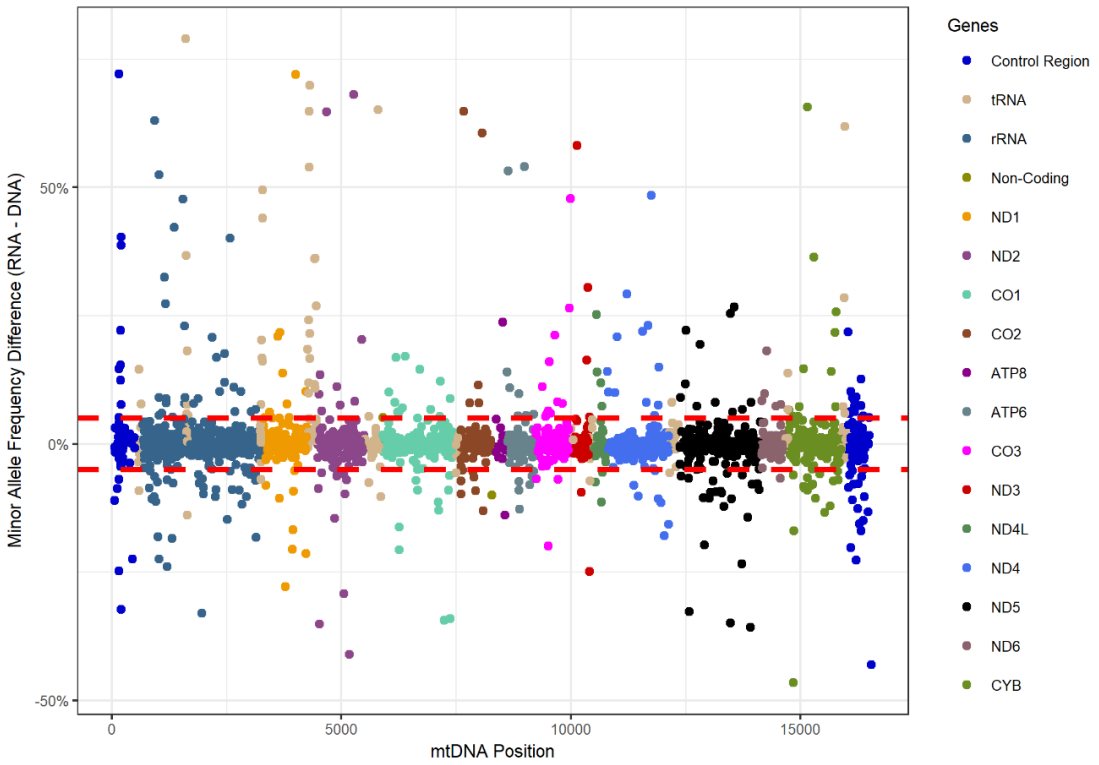


Figure 3-2 Heteroplasmy frequency difference between DNA and RNA.

Each point is a heteroplasmy sites. Positive number indicates frequency in RNA is higher than that in DNA, and negative number frequency in DNA is higher than that in RNA. Red dashed line indicates 5% frequency difference.

in DNA but low frequency in RNA (HDLR heteroplasmy), is that these mutations were deleterious thus the mutation transmission from DNA to RNA are suppressed. To test this hypothesis, we annotated the heteroplasmy with Combined Annotation Dependent Depletion (CADD) scores [17] as a measurement of their pathogenic potentials. We compared the CADD scores of HDLR heteroplasmies (DNA frequency at least 5% greater than RNA frequency) to the rest of heteroplasmies. However, the CADD scores of the HDLR heteroplasmies were not significantly higher (one-side t-test, ***P value*** = 0.2276, Figure 3-3). We next tested whether the HDLR heteroplasmies were more like to be diseases associated by chi-squared test, but the result was also not significant (***P value*** = 0.6877). In addition, we also identified 219 heteroplasmies in OD group. We then tested whether the CADD scores of these heteroplasmies were significantly higher than those of BDR group. The result was not significant (one-side t-test, ***P value*** = 0.1). Thus, we didn't find the evidence for negative selections of heteroplasmy transmission from DNA to RNA.

### 3.4.3 RDDs could be caused by heteroplasmy dynamics

In this study, DNA and RNA mtDNA sequencing data were retrieved from two different large-scale sequencing projects (1000 Genome Project and Geuvadis RNA-seq Project), and the cell lines for extracting DNA or RNA were grown independently. If the heteroplasmy frequencies could change during the cell growth and DNA and RNA were extract at different time points, it was not surprising that we would observe quite different heteroplasmy frequencies between DNA-seq and RNA-seq data (Figure 3-4). To test this hypothesis, we picked up five heteroplasmy sites with DNA-RNA

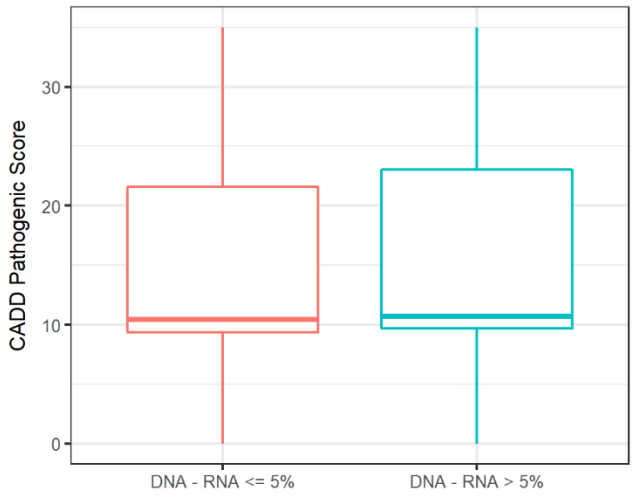


Figure 3-3 CADD pathogenic scores of HDLR and non-HDLR heteroplasmies.

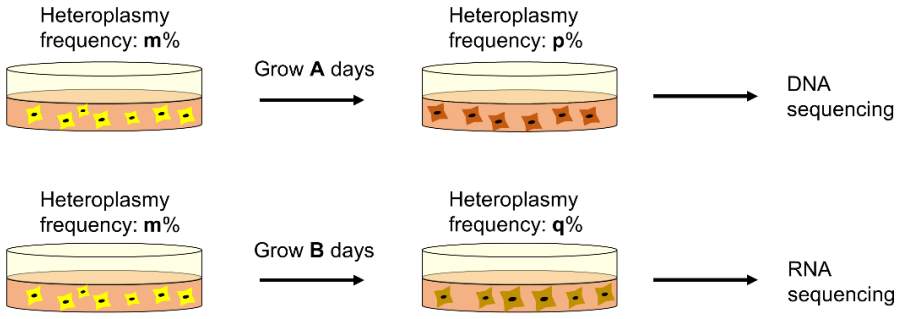


Figure 3-4 RDDs could be caused by heteroplasmy dynamics

Cells used for DNA and RNA sequencing were cultured independently. At day 0, the heteroplasmy frequencies of both cell cultures were m%. Cells for DNA sequencing were grown for A days and the heteroplasmy frequency was changed to p% while cells for RNA sequencing were grown for B days and heteroplasmy frequency was changed to q%. In the sequencing data, the RNA-DNA heteroplasmy frequency difference would be around q%-p%.

frequency differences > 45% from the above analysis (Table 3-1). These heteroplasmies belonged to two cell lines: GM12282 and GM18934. We grew these two cell lines for four consecutive weeks and extracted DNA and RNA simultaneously from exact same cells every week to track the heteroplasmy frequency dynamics of the five heteroplasmy sites (**Methods**). We observed striking frequency changes in three of the five heteroplasmic sites, while the rest two also had mild frequency alterations (Figure 3-5). For example, at heteroplasmy site 15153G>A, the frequencies of A allele were 54.0% and 46.9% in DNA and RNA, respectively, at Day 0, and gradually decreased in the later time points. In Day 28, they decreased to 2.1% (DNA) and 0.8% (RNA).

These heteroplasmy sites were selected since they showed noticeable frequency difference between DNA and RNA in the public available dataset (1000 Genome and Geuvadis). However, in our experiment data, the heteroplasmy frequencies were highly consistent between DNA and RNA at each time point (Figure 3-5). For example, the heteroplasmy difference of 929A>G was 63.1% (DNA 79.8%, RNA 16.7%) in the public dataset. In contrast, the difference was only 1.1% (DNA 21.4%, RNA 20.3%) at day 0 and 0.2% (0.5% DNA 0.3% RNA) at Day 28 in our experiment. Taking these results together, the pronounced heteroplasmy frequency differences observed from the public dataset might be artifacts caused by the dynamics of heteroplasmy rather than real biological difference.

### 3.4.4 Potential modification/editing events in mitochondrial RNA

For the same individual, heteroplasmies observed in RNA but not DNA were

Table 3-1. examples of heteroplasmy with noticeable frequency difference between DNA and RNA

| Cell line | mtDNA Position | Heteroplasmy frequency Difference between DNA and RNA | Ref | Alternative frequency DNA | Alternative frequency RNA | Annotation |
|---|---|---|---|---|---|---|
| **NA12282** | 929 | 63.1% | A | G 79.8% | G 16.7% | 12S RNA |
| **NA12282** | 5794 | 65.1% | T | C 78.9% | C 13.8% | tRNA-Cys |
| **NA12282** | 15153 | 65.7% | G | A 80.3% | A 14.6% | CYB (NS) |
| **NA18934** | 5262 | 68.2% | G | C 20.0% | C 88.2% | ND2 (NS) |
| **NA18934** | 9984 | 47.8% | G | A 12.6% | A 60.4% | CO3 (SY) |

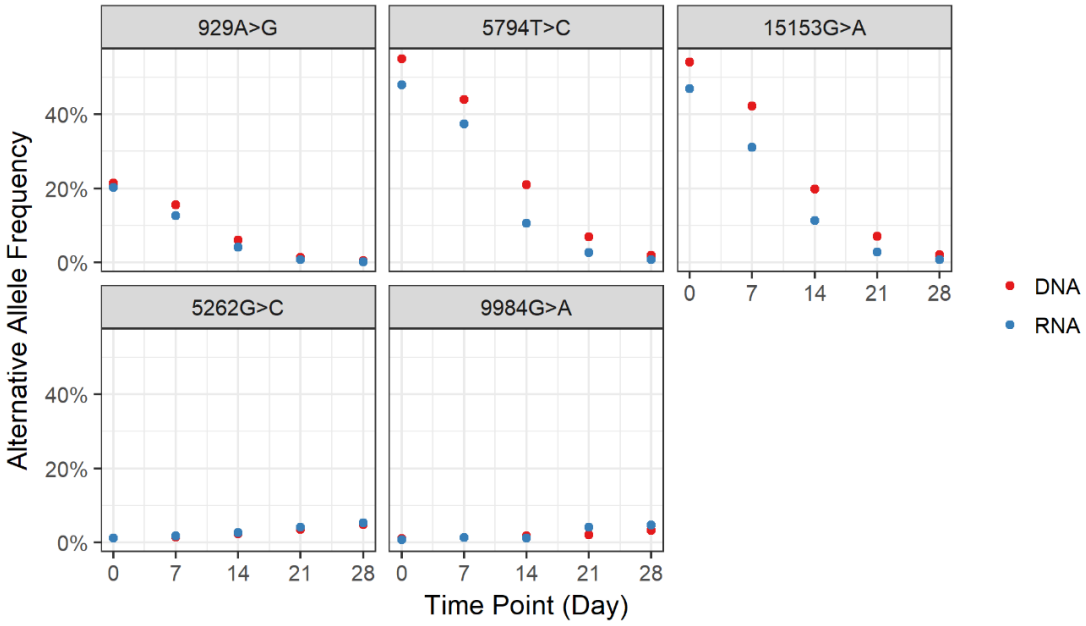


Figure 3-5 Heteroplasmy frequency would change over time

Heteroplasmy frequency changes were tracked in 5 different sites in 2 cell lines. The frequencies were gradually changing from day 0 to day 28, and keep consistent between DNA and RNA.

potentially caused by RNA posttranscriptional modification/editing. There are three mitochondrial sites (295, 2617 and 13710) were found to be RDD sites in most individuals [23, 24]. We first evaluated these three sites in our study dataset. To avoid noise, for each site, we only included the individuals with sequencing coverage > 400 in both DNA and RNA data, and major allele is same as reference allele. Most of the individuals has modification/editing at these three sites (Table 3-2), and the edited allele frequency were varied from individual to individual (Figure 3-6).

Besides these three sites, there were another 472 heteroplasmies at 200 unique mitochondrial sites were only observed in RNA (both DNA and RNA had sequencing coverage > 400). Among these 472 heteroplasmies, 216 (45.7%) were observed in YRI population, significant higher than expected (***P value*** < 2.2e-16, Chi-square test, Table 3-3). We also identified 6 sites have potential editing in more than 10 individuals. Among these sites, major of individuals were from YRI population.

If we divide all the individuals into with editing and without editing groups (183 individuals had 0 editing while 263 individuals have >=1 editing). If we compared the gene expression level between these two groups, 1540 genes had p value < 0.05, and 372 genes had p value <0.01 (t-test). We did a GO term analysis for the genes with p value < 0.01, and 174 genes can be mapped to GO biological processes. The top 1 GO term hit was "regulation of mRNA stability" (***P value*** = 4.39e-6). This functional pathway may be associated with RNA editing process. Next, we performed a genome wide association study (GWAS) for the two groups to see whether we could locate any nuclear variations that may associated with the mitochondrial RNA modification/editing. Because this dataset only had 446 individuals, to avoid the

Table 3-2 Three previous reported mtRNA editing sites in this dataset

| | # of individuals passed filter | # ofindividuals has heteroplasmy at given site | # of individuals has edited allele frequency > 1% | Reference allele | Editing allele |
|---|---|---|---|---|---|
| **295** | 319 | 2 | 306 | C | T |
| **2617** | 445 | 0 | 445 | A | T, G, C |
| **13710** | 445 | 0 | 326 | A | T, G |

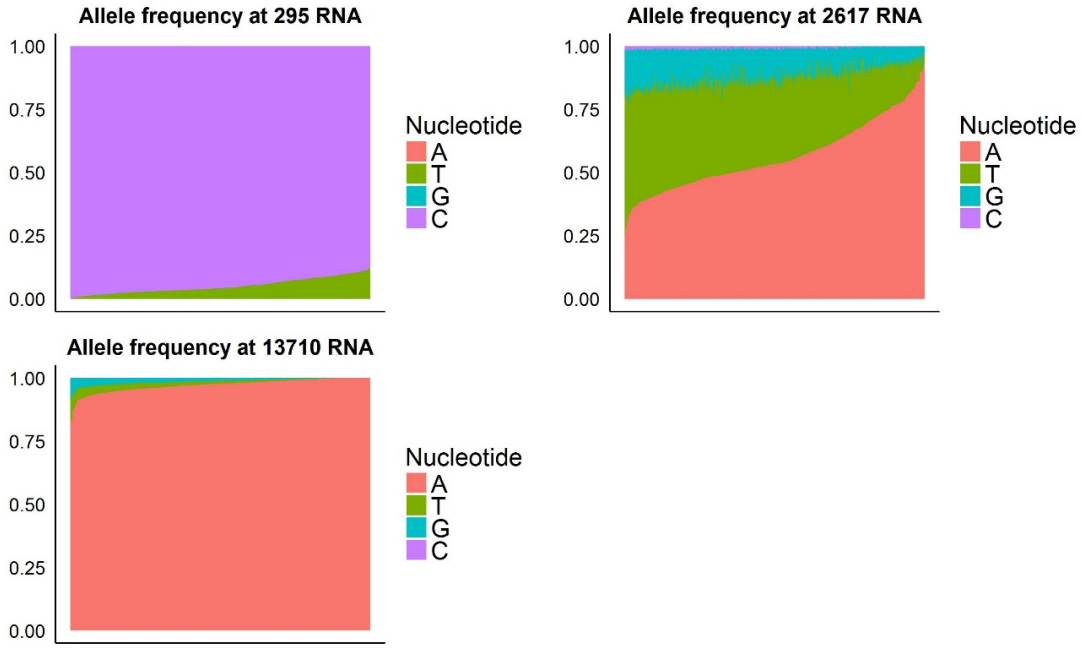


Figure 3-6 Edited allele frequency distribution in 3 previously reported mtRNA editing sites

Table 3-3 mtRNA editing events number identified in the study population

| Population | CEU | FIN | GBR | TSI | YRI |
|---|---|---|---|---|---|
| **# of individuals** | 89 | 92 | 87 | 91 | 87 |
| **Total editing events** | 40 | 79 | 51 | 86 | **216** |
| **Editing at 2129** | 6 | 16 | 15 | 12 | **28** |
| **Editing at 5746** | 4 | 7 | 3 | 4 | **26** |
| **Editing at 12146** | 7 | 9 | 6 | 10 | 7 |
| **Editing at 6691** | 3 | 6 | 5 | 4 | **13** |
| **Editing at 4104** | 0 | 0 | 0 | 0 | **19** |
| **Editing at 296** | 1 | 1 | 0 | 1 | **10** |

statistical power issue, we only screened SNPs on a list of 108 human RNA-binding and transcription-associated mitochondrial genes [19]. Totally, 6405 variants were tested, and the significance level was set to 7.8e-6. We located a significant SNP on homo sapiens transcription factor B2 gene, which may contribute to mitochondrial RNA posttranscription process.

## 3.5 Discussion

The sequence information transmission from DNA to RNA is critical for cell's function. There are several studies reported widespread DNA/RNA sequence disconcordance in human nuclear genome. Li et al. uncovered that more 10,000 exonic sites were inconsistent between DNA and RNA in human B cells resulted from RNA editing [1]. Park et al. further studied that the A-to-I RNA editing sites may play a mechanistic role in linking genetic variation to complex traits and diseases [2]. RNA-DNA differences were also identified in mitochondrial. The first reported mitochondrial RDD sites were 295 (C>U), 13710 (A>U and A>G) and 2617 (A>U and A>G) [23]. Hodgkinson et al. found that mitochondrial tRNA posttranscriptional modification may affect cellular energy production [24]. They later reported that variation of mitochondrial RNA processing between normal and tumor tissues may impact cancer patient survival outcomes [25]. However, these studies only focused on the RNA editing caused RDDs. The information about heteroplasmy transcription fidelity were still overlooked. In this study, by comparing the genomic and transcriptomic data from 446 human lymphoblastoid cell lines, we identified 2786 heteroplasmies presented in both DNA and RNA, 472 only observed in RNA (except

the 3 common RNA editing sites) and 219 only observed in DNA. Therefore, most of heteroplasmy observed at DNA level can be transcribed into RNA. Moreover, the frequency of heteroplasmies could keep consistent from DNA to RNA. For the sites with noticeable heteroplasmy frequency differences, we experimentally verified that these were artifacts caused by heteroplasmy dynamics (heteroplasmy dynamics would be discussed in more detail in next chapter). Our result suggested that there might be limited negative selection in the mitochondrial transcription step, thus the pathogenic mutations were likely to be transmitted to RNA and further lead to deleterious consequence. Due to the sequencing cost, public available RNA sequencing dataset were much richer than whole genome sequencing data. Our study also provided the import guidance that RNA sequencing data could be a potential source for mitochondrial variation data mining.

Using these paired DNA and RNA sequencing data, we also confirmed the three common mtRNA editing sites were observable in most the individuals. Consistent with previous studies, we found that the modification/editing levels were high variable in the population. The editing level could be regulated by nuclear DNA. For example, Hodgkinson et al reported that the methylation level of mitochondrial tRNA p9 sites were significantly driven by a missense mutation in MRPP3 gene [24]. We also located a SNP in TFB2M gene to be associated with mtRNA modification/editing. TFB2M gene can dimethylate mitochondrial 12S rRNA [26]. TFB2M is also required by the transcription machinery of mitochondrial DNA, probably via its interaction with mitochondrial RNA polymerase POLRMT and TFAM [27, 28] and can stimulate transcription independently of the methyltransferase activity [29]. Notably, the

mtRNA modification/editing events were more common in African population than the other four European populations. The mtDNA transcript expression pattern were also found to be distinct in African population [30]. These unique features of mtRNA in African population suggested that a transition of regulatory of mtRNA may happen in ancient human population and populate later as human left Africa.

## 3.6 Acknowledgement

We would like to thank Drs. Kaixiong Ye, Yuan Si and Xiaoxian Guo for their valuable comments. We would like to thank Dr. Kiichi Nakahira for his suggestions of cell culture experiments. We would like to thank Dr. Patrick Sullivan for his comments on statistical tests.

3.7 Reference


1. Li, M., et al., *Widespread RNA and DNA Sequence Differences in the Human Transcriptome*. Science, 2011. **333**(6038): p. 53-58.
2. Park, E., et al., *Population and allelic variation of A-to-I RNA editing in human transcriptomes*. Genome Biology, 2017. **18**(1): p. 143.
3. Stewart, J.B. and P.F. Chinnery, *The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease*. Nat Rev Genet, 2015. **16**(9): p. 530-542.
4. Lightowlers, R.N., et al., *Mammalian mitochondrial genetics: heredity, heteroplasmy and disease*. Trends in Genetics, 1997. **13**(11): p. 450-455.
5. Wallace, D.C., *Mitochondrial DNA mutations in disease and aging*. Environmental and molecular mutagenesis, 2010. **51**(5): p. 440-450.
6. Wallace, D.C. and D. Chalkia, *Mitochondrial DNA genetics and the heteroplasmy conundrum in evolution and disease*. Cold Spring Harbor perspectives in biology, 2013. **5**(11): p. a021220.
7. Wallace, D.C., *A mitochondrial bioenergetic etiology of disease*. The Journal of clinical investigation, 2013. **123**(4): p. 1405-1412.
8. Gorman, G.S., et al., *Mitochondrial diseases*. Nature Reviews Disease Primers, 2016. **2**: p. 16080.
9. Alston, C.L., et al., *The genetics and pathology of mitochondrial disease*. The Journal of Pathology, 2017. **241**(2): p. 236-250.
10. Schon, E.A., S. DiMauro, and M. Hirano, *Human mitochondrial DNA: roles of inherited and somatic mutations*. Nat Rev Genet, 2012. **13**(12): p. 878-890.
11. Rossignol, R., et al., *Mitochondrial threshold effects*. Biochem J, 2003. **370**(Pt 3): p. 751-62.
12. The Genomes Project, C., *A global reference for human genetic variation*. Nature, 2015. **526**(7571): p. 68-74.
13. Lappalainen, T., et al., *Transcriptome and genome sequencing uncovers functional variation in humans*. Nature, 2013. **501**: p. 506.
14. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2*. Nat Meth, 2012. **9**(4): p. 357-359.
15. Van der Auwera, G.A., et al., *From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline*. Curr Protoc Bioinformatics, 2013. **43**: p. 11.10.1-33.
16. Li, H., et al., *The Sequence Alignment/Map format and SAMtools*. Bioinformatics, 2009. **25**(16): p. 2078-9.
17. Kircher, M., et al., *A general framework for estimating the relative pathogenicity of human genetic variants*. 2014. **46**(3): p. 310-5.
18. Ruiz-Pesini, E., et al., *An enhanced MITOMAP with a global mtDNA mutational phylogeny*. Nucleic acids research, 2007. **35**(suppl 1): p. D823-D828.
19. Wolf, Ashley R. and Vamsi K. Mootha, *Functional Genomic Analysis of Human Mitochondrial RNA Processing*. Cell Reports. **7**(3): p. 918-931.

20. Chang, C.C., et al., *Second-generation PLINK: rising to the challenge of larger and richer datasets*. GigaScience, 2015. **4**(1): p. 7.
21. Ye, K., et al., *Extensive pathogenicity of mitochondrial heteroplasmy in healthy human individuals*. Proceedings of the National Academy of Sciences, 2014. **111**(29): p. 10654-10659.
22. Rossignol, R., et al., *Mitochondrial threshold effects*. Biochemical Journal, 2003. **370**(3): p. 751-762.
23. Bar-Yaacov, D., et al., *RNA-DNA differences in human mitochondria restore ancestral form of 16S ribosomal RNA*. Genome Res, 2013. **23**(11): p. 1789-96.
24. Hodgkinson, A., et al., *High-Resolution Genomic Analysis of Human Mitochondrial RNA Sequence Variation*. Science, 2014. **344**(6182): p. 413-415.
25. Idaghdour, Y. and A. Hodgkinson, *Integrated genomic analysis of mitochondrial RNA processing in human cancers*. Genome Medicine, 2017. **9**(1): p. 36.
26. Metodiev, M.D., et al., *Methylation of 12S rRNA Is Necessary for In Vivo Stability of the Small Subunit of the Mammalian Mitochondrial Ribosome*. Cell Metabolism, 2009. **9**(4): p. 386-397.
27. Bestwick, M.L. and G.S. Shadel, *Accessorizing the human mitochondrial transcription machinery*. Trends in Biochemical Sciences. **38**(6): p. 283-291.
28. Litonin, D., et al., *Human Mitochondrial Transcription Revisited: ONLY TFAM AND TFB2M ARE REQUIRED FOR TRANSCRIPTION OF THE MITOCHONDRIAL GENES IN VITRO*. Journal of Biological Chemistry, 2010. **285**(24): p. 18129-18133.
29. McCulloch, V. and G.S. Shadel, *Human mitochondrial transcription factor B1 interacts with the C-terminal activation region of h-mtTFA and stimulates transcription independently of its RNA methyltransferase activity*. Mol Cell Biol, 2003. **23**(16): p. 5816-24.
30. Cohen, T., L. Levin, and D. Mishmar, *Ancient Out-of-Africa Mitochondrial DNA Variants Associate with Distinct Mitochondrial Gene Expression Patterns*. PLOS Genetics, 2016. **12**(11): p. e1006407.

## *Chapter 4 - Mitochondrial DNA heteroplasmy dynamics causes global gene expression changes.*

### 4.1 Introduction

The term *heteroplasmy* specify a condition that a mixture of mutant mitochondrial DNA (mtDNA) and wild type mtDNA presenting in the same cell due to the multiploid nature of mtDNA [1]. The ratio of mutated mtDNA (heteroplasmy frequency) can vary from 0% to 100%. For the pathogenic mutations, in many cases, the phenotypic manifestation occurs only when the heteroplasmy frequency reaches a certain level, which is the so-called "heteroplasmy phenotypic threshold effects" [2]. Therefore, the frequency of a given heteroplasmy could be very critical for its pathogenicity. Moreover, the heteroplasmy frequency can vary from tissue to tissue within a same individual and even vary from cell to cell in a same tissue, making its pathogenicity more complicated [3]. mtDNA is replicated constantly through the lifetime and is independent of the cell cycle, therefore, it is possible that the frequency of an existing heteroplasmic mutation could gradually change after multiple rounds of replications.

The dynamics of heteroplasmy has been reported for the most common pathogenic mtDNA mutation 3243A>G. Several clinical studies showed that 3243A>G heteroplasmy frequency would decrease slightly overtime [4, 5]. Rajasimha et al. collected clinical data with multiple heteroplasmy frequency measurements of same individuals to established a model to simulate the changes of heteroplasmy frequency [6]. Raap et al. used 3243A>G cybrid cells to study the segregation of the mutated

mtDNA, and provided possible explanations for the shifts of heteroplasmy frequency. In this study, we tracked the frequency changes of three heteroplasmies in a human B-lymphoblastoid cell line which all had high frequencies at experiment starting point. We observed that the frequencies of all the three heteroplasmies shifted dramatically in only 28 days. We then speculated that these noticeable heteroplasmy frequency changes could have some effects on the cellular functions. As a measurement, we performed RNA sequencing for the cells at 5 different time points (with different heteroplasmy frequencies). The RNA-seq results suggested that there was a global gene expression change with respect to different heteroplasmy frequencies.

## 4.2 Material and Methods

### 4.2.1 Cell culture and single heteroplasmy site sequencing

Cell lines were purchased from Coriell Institute (https://www.coriell.org/). Cells were grown at 37°C, 5% CO2, in Roswell Park Memorial Institute Medium 1640 with 2mM L-glutamine and 15% fetal bovine serum. Cells were split every 2-3 days and fresh medium were replaced. Cells were collected at Day 0, 7, 14, 21 and 28, respectively. RNA was extracted from collected cells with RNeasy Plus Mini Kit (Qiagen, catalog no.74134). RNA was reverse transcribed to cDNA with SuperScript™ III First-Strand Synthesis System (Invitrogen, catalog number: 18080051).

### 4.2.2 RNA sequencing and bioinformatics analysis

Total RNA was extracted from each time point and sequencing libraries were generated with KAPA stranded mRNA-Seq kit (catalog number: KK8420). Libraries

were sequenced by Illumina Hiseq 2500 platform with single end 50bp reads. Adapter were trimmed from raw data with Trimmomatic [7]. The trimmed reads were aligned to human reference genome hg19 with STAR [8]. The number of reads mapped to each gene were counted by HTSeq [9] and DESeq2 [10] were used to evaluate the differentially expressed genes. Heteroplasmies were identified from RNA-seq data as described in Chapter 3.

### 4.2.3 Simulation of mtDNA segregation

The model to simulate heteroplasmy dynamics were simplified with a few assumptions for mtDNA segregation: 1) the mtDNA copy number of each cell is fixed at 1000. 2) The heteroplasmy frequency (the fraction of mutated copy) at starting point was f. 3) For each generation, a new set of 1000 mtDNA copies was replicated, the mutated copy number followed a binomial distribution with p equals to f*fit, where fit was the selection coefficient of mutated copies, if there was no selection, fit = 1; if there was positive selection, fit > 1; if there was negative selection, fit < 1. 4) During segregation, the old 1000 mtDNA copies were mixed with newly replicated mtDNA copies, then they were randomly distributed into two daughter cells with same copy number. We also had several assumptions for cell growing: 1) The cell doubling time was 24h. 2) There were 200,000 cells at starting point. 3) For every 2 generations, a quarter of the cells were randomly sampled and used for the subsequent experiment. 4) The cells were grown for 30 days (equivalent to 30 generations).

## 4.3 Results

### 4.3.1 Heteroplasmy frequency can change over time

The dynamics of heteroplasmy frequency has been observed and discussed in several studies previously [4-6, 11, 12]. Unlike nDNA, mtDNA is continuously replicated and destructed even in post-mitotic cells. In a cell harboring heteroplasmy, the mutated molecule might be replicated more frequently or less frequently. Hence, after multiple replications, this can result in a recognizable heteroplasmy frequency change. In this study, we grew a B-lymphoblastoid cell line (GM12282) with three heteroplasmies for 28 days and extracted DNA and RNA simultaneously every 7 days to track the frequency changes of three heteroplasmies. The minor allele frequencies of the three heteroplasmies at the starting point (Day 0) were as followings: 929A>G (G frequency 15.8%), 5794T>C (C frequency 43.5%) and 15153G>A (A frequency 48.3%). The frequency would refer to the non-reference allele frequency in the following discussions. In the following 4 check points (Day 7, 14, 21 and 28), for all the three heteroplasmies, the frequencies decreased consistently. At Day 28, the frequencies were 0.2%, 0.6% and 0.8%, respectively. The frequency of minor alleles shifted to very low levels in the cells (Table 4-1, Figure 4-1).

### 4.3.2 Heteroplasmy frequency changes can affect gene expression profiles

There were three heteroplasmies identified in the studying cell line GM12282. One located at 12S ribosomal RNA (929A>G), one located at tRNA-Cys (5794T>C) and the third one was a nonsynonymous mutation in CYB gene (15153G>A). The frequencies of these three heteroplasmies were all gradually changing during the 28 growing days (Figure 4-1). To confirm that these frequency shifts were reproducible,

Table 4-1. Heteroplasmy frequency changes during 28 days in cell line GM12282 in two replicate experiments

| Heteroplasmy Site | Day 0 | Day 7 | Day 14 | Day 21 | Day 28 |
|---|---|---|---|---|---|
| 929 (G) rep 1 | 15.8% | 11.1% | 3.3% | 0.7% | 0.2% |
| 929 (G) rep 2 | 12.9% | 6.4% | 3.1% | 0% | 0.6% |
| 5794 (C) rep 1 | 43.5% | 38.3% | 11.4% | 2.9% | 0.8% |
| 5794 (C) rep 2 | 39.2% | 24.4% | 10.2% | 0% | 2.7% |
| 15153 (A) rep 1 | 48.3% | 32.6% | 11.3% | 2.5% | 0.8% |
| 15153 (A) rep 2 | 42.2% | 23.5% | 11.0% | 0% | 3.5% |

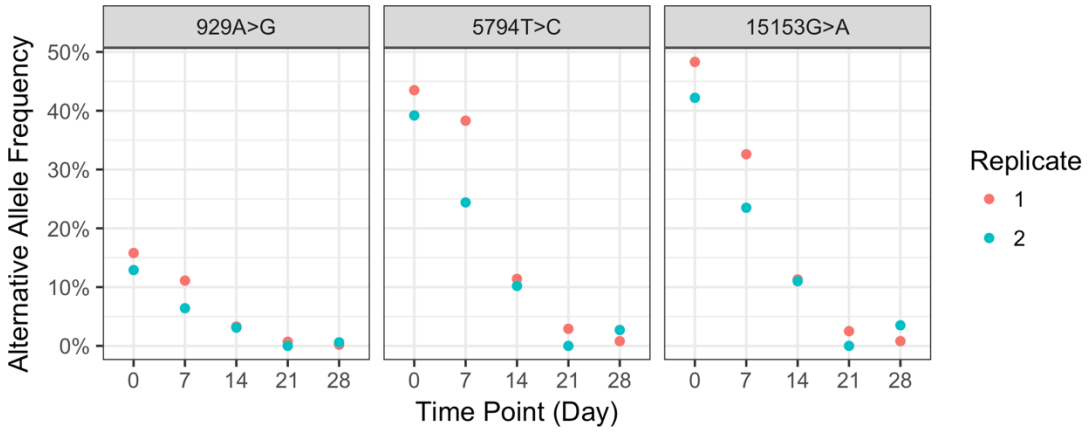


Figure 4-1 Heteroplasmy frequency changes during 28 days in the study cell line.

There were three heteroplasmies in GM12282 cell line. During the cell culture process, the heteroplasmy frequencies were gradually decreasing in all three heteroplasmies in two independent replicates experiments.

we repeated this growing experiment, and found that the general pattern of frequency changes of the three heteroplasmy sites were comparable to the first experiment (Table 4-1). According to the heteroplasmy threshold theory [13], the frequencies of a specific heteroplasmy may affect its functional impacts to the mitochondria and further to the cell function. To investigate the influence of these observed heteroplasmy dynamics, we analyzed the gene expression profiles of the study cell line with respect to different heteroplasmy frequencies at each time point. We first compared the gene expressions between high frequency and low frequency conditions. Repeat 1 at Day 0 and repeat 2 at Day 0 were used as high frequency replicates while repeat 1 at Day 28 and repeat 2 at Day 21 were used as low frequency replicates since they had relative similar frequency at each heteroplasmies site (Table 4-1). After multiple tests correction, there were 424 genes differentially expressed between these two heteroplasmy conditions (**Methods**). Among them, 242 genes were up-regulated while the other 182 genes were down-regulated. Gene ontology (GO) classified significant enrichment of genes in several catalogs (Table 4-2, Table 4-3). The up-regulated genes were enriched in immune response, and the down-regulated genes also showed some enrichment in immune related terms for instance, platelet activation. The top 3 down regulated genes were IGHV3-21, IGLV2-23 and IGKV4-1, and the top 3 up regulated genes were TMEM176A, TMEM176B and IL1R2. These genes were also closed related to cells immune functions.

We next evaluated the expression levels of genes related to cellular energy production. We first compare the expression level of the 13 mtDNA encoded genes at five time points (Figure 4-2). The expression levels were stable for these genes, and the

Table 4-2. Upregulated genes GO enrichment

| Go Term | Fold Enrichment | P Value | Adjusted P Value |
|---|---|---|---|
| **immune response** | 4.34655 | 1.81E-06 | 0.002072 |
| **Extracellular region** | 2.238271 | 1.05E-05 | 0.001927 |
| **Extracellular space** | 2.369543 | 1.40E-05 | 0.001277 |
| **Cytokine activity** | 5.957439 | 4.64E-05 | 0.013691 |
| **Desmosome** | 21.45009 | 7.68E-05 | 0.004675 |

Table 4-3. Downregulated genes GO enrichment

| Go Term | Fold Enrichment | P Value | Adjusted P Value |
|---|---|---|---|
| **Plasma membrane** | 2.0013 | 7.13E-09 | 1.76E-06 |
| **Focal adhesion** | 4.762932 | 7.47E-06 | 9.22E-04 |
| **Extracellular exosome** | 1.940198 | 2.52E-05 | 0.00207 |
| **Platelet activation** | 8.464776 | 4.34E-05 | 0.046459 |
| **Hepatocyte differentiation** | 48.67246 | 6.11E-05 | 0.03292 |

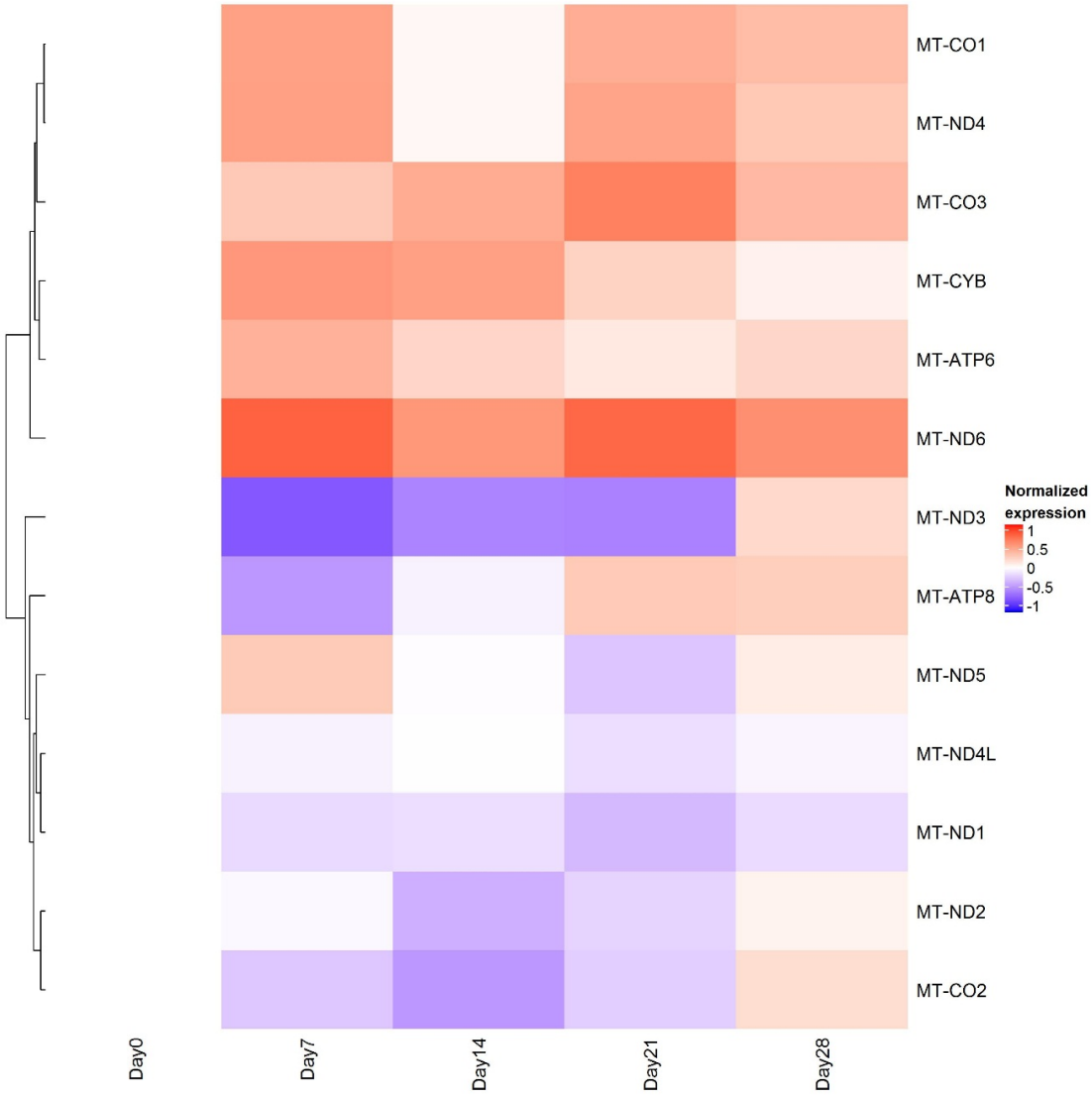


Figure 4-2 The expression profile of the 13 mtDNA protein encoding genes in 28 growing days.

fluctuation was with 2 folds. For gene CYB, where one of the heteroplasmies located, the expression level also kept consistent over the 5 time points. These observations indicated that the heteroplasmy frequency changes did not affect the expression mtDNA genes directly. We then did the similar comparisons for citric acid cycle (TCA cycle) and glycolysis related genes (Figure 4-3, Figure 4-4). While most of the genes had stable expression, we also found several genes with consistent and significant expression changes. For example, PCK1 gene expression gradually increased and the expression at day 28 was ~6.9 fold of that at day 0, suggesting that the heteroplasmy frequency could influence the cellular metabolisms.

To confirm that these transcriptional changes were not caused the cell growth process, we grew another cell line (GM12751) in parallel, and evaluated the gene expression changes between Day 0 and Day 28. For the differentially expressed gene identified from above analysis, we did not observe similar changes in this cell line, which suggested that these transcriptional alterations were caused by the heteroplasmy changes rather than the cell growing process.

### 4.3.3 In-silico simulation of heteroplasmy changes

To help explain the heteroplasmy frequency changes we observed from this study, we ran several simulations of mtDNA segregations during cell line growth to model the heteroplasmy dynamics. Using the simplified model (**Methods**), we simulated the heteroplasmy dynamics with different selection coefficient of mutation (Figure 4-5). From the simulation we could see that the complete neutral mtDNA heteroplasmies could keep a stable frequency over time, while slight negative selection or positive

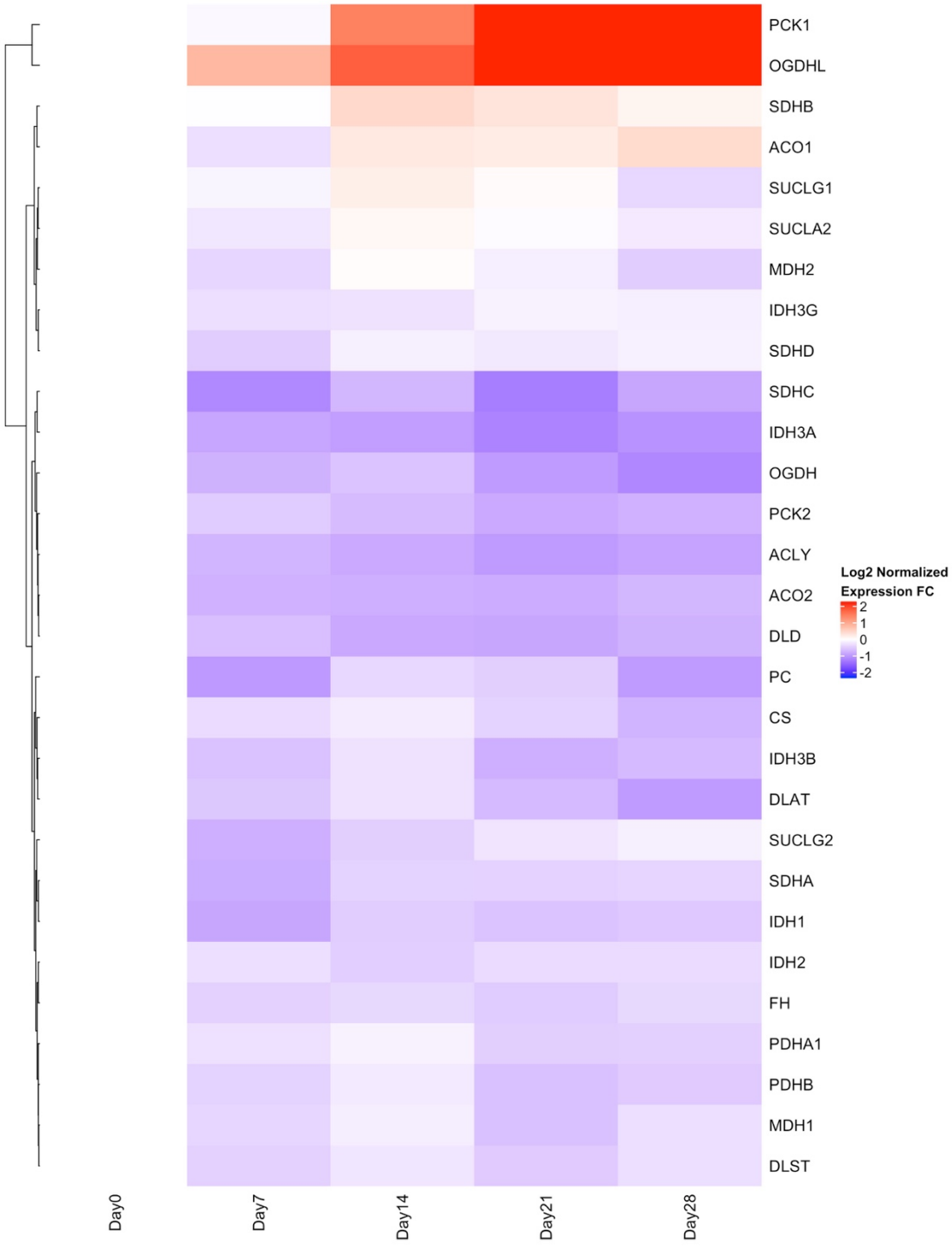


Figure 4-3 The expression profile of the TCA related genes in 28 growing days.

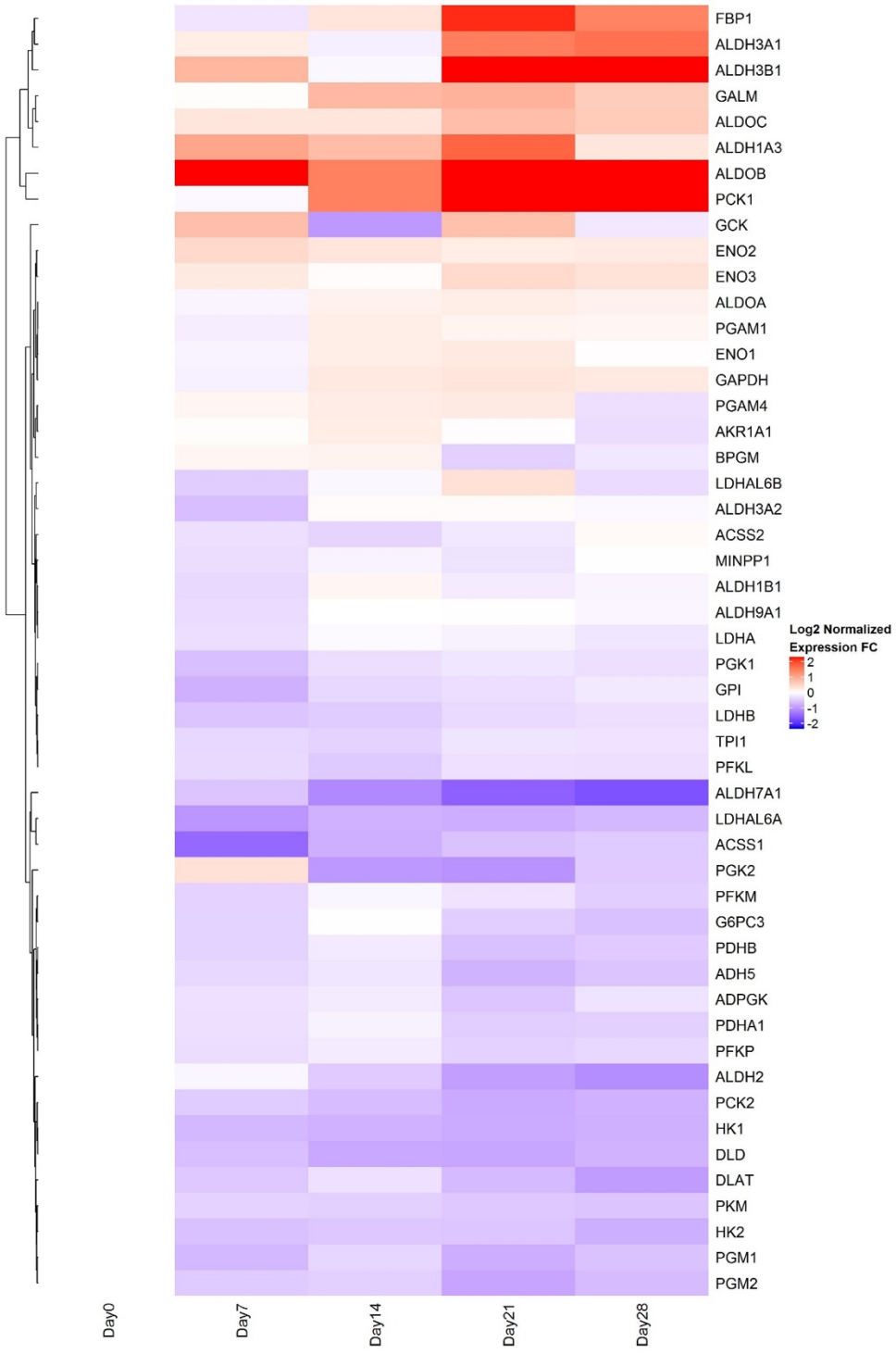


Figure 4-4 The expression profile of the glycolysis related genes in 28 growing days.

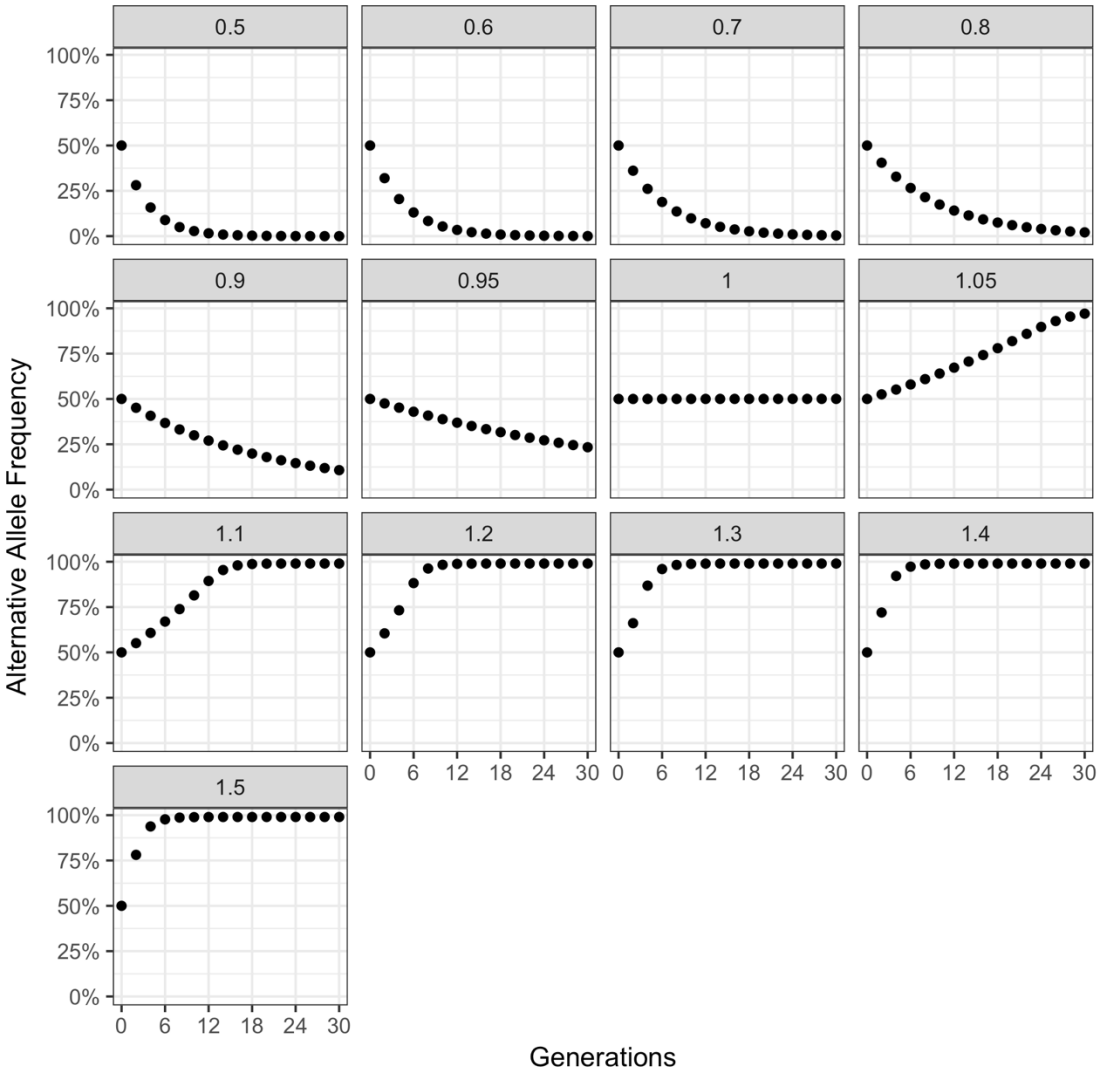


Figure 4-5 Heteroplasmy frequency changes from simulation results.

The number in the title of each block represents the selection coefficient of the mutation (fit). When there is not any selection (fit = 1), the heteroplasmy frequency will stay stable; when there is positive selection (fit > 1), the heteroplasmy frequency gradually increase over time; when there is negative selection (fit < 1), the heteroplasmy frequency gradually decreases.

selection could efficiently purify or enrich the mutations.

## 4.4 Discussion

mtDNA heteroplasmy had considerable prevalence even in general healthy population and most of the heteroplasmies stayed at low frequency status [14, 15]. The deleterious heteroplasmies among them, once reached high frequency, could be a potential source of mitochondrial dysfunction. Therefore, it could be important to investigate the dynamics of heteroplasmy frequency. The changes of heteroplasmy frequencies were reported in previous studies [4, 5]. In this study, in two independent growing experiments, we observed that heteroplasmy frequency can change as much as ~48% in only 28 days. This observation providing the experiment evidence that low frequency heteroplasmies can eventually reach a high frequency and may further make impacts for the cell functions.

We then examined the transcriptome changes with respect to different heteroplasmy frequencies. We detected a group of genes with significant expression changes at different heteroplasmy frequencies. Through GO analyzed, we found these genes were enriched in immune related terms. Considering that the cells we used in this study was B-lymphoblastoid cells, this result suggested that heteroplasmy frequency changes could affect the specified cell functions. Human cell line is widely used for many biological studies. Our results suggested mtDNA heteroplasmy frequency changes could introduce potential noise to the cell line studies, which is overlooked previously. In the future studies, it could be important to adjust the confounding effects caused by heteroplasmy dynamics.

We also did in-silico simulation of mtDNA segregation to model the heteroplasmy frequency dynamics during the cell line growth. We found that, without any selection pressure, the heteroplasmy frequency would keep stable during the cell growth, while very little disturbance can result in substantial changes in heteroplasmy frequencies. These in-silico simulation results, together with the experiment observations, indicated that the mtDNA mutations were subjected to the negative selections during cell line growing process.

## 4.5 Acknowledgement

We would like to thank Drs. Kaixiong Ye, Yuan Si and Xiaoxian Guo and Mr. Yiqin Wang for their valuable comments. We would like to thank Dr. Kiichi Nakahira for his suggestions of cell culture experiments and mitochondrial function measurement experiments. We would like to thank Dr. Iwijn De Vlaminck for his comments on simulation study.

4.6 Reference


1. Lightowlers, R.N., et al., *Mammalian mitochondrial genetics: heredity, heteroplasmy and disease*. Trends Genet, 1997. **13**(11): p. 450-5.
2. Rossignol, R., et al., *Mitochondrial threshold effects*. Biochem J, 2003. **370**(Pt 3): p. 751-62.
3. Li, M., et al., *Extensive tissue-related and allele-related mtDNA heteroplasmy suggests positive selection for somatic mutations*. Proc Natl Acad Sci U S A, 2015. **112**(8): p. 2491-6.
4. Rahman, S., et al., *Decrease of 3243 A-->G mtDNA mutation from blood in MELAS syndrome: a longitudinal study*. Am J Hum Genet, 2001. **68**(1): p. 238-40.
5. Pyle, A., et al., *Depletion of mitochondrial DNA in leucocytes harbouring the 3243A->G mtDNA mutation*. J Med Genet, 2007. **44**(1): p. 69-74.
6. Rajasimha, H.K., P.F. Chinnery, and D.C. Samuels, *Selection against Pathogenic mtDNA Mutations in a Stem Cell Population Leads to the Loss of the 3243A→G Mutation in Blood*. American Journal of Human Genetics, 2008. **82**(2): p. 333-343.
7. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data*. Bioinformatics, 2014. **30**(15): p. 2114-20.
8. Dobin, A., et al., *STAR: ultrafast universal RNA-seq aligner*. Bioinformatics, 2013. **29**(1): p. 15-21.
9. Anders, S., P.T. Pyl, and W. Huber, *HTSeq—a Python framework to work with high-throughput sequencing data*. Bioinformatics, 2015. **31**(2): p. 166-9.
10. Love, M.I., W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2*. Genome Biology, 2014. **15**(12): p. 550.
11. Stewart, J.B. and P.F. Chinnery, *The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease*. Nat Rev Genet, 2015. **16**(9): p. 530-542.
12. Raap, A.K., et al., *Non-Random mtDNA Segregation Patterns Indicate a Metastable Heteroplasmic Segregation Unit in m.3243A>G Cybrid Cells*. PLOS ONE, 2012. **7**(12): p. e52080.
13. Rossignol, R., et al., *Mitochondrial threshold effects*. Biochemical Journal, 2003. **370**(3): p. 751-762.
14. Zhang, R., et al., *Independent impacts of aging on mitochondrial DNA quantity and quality in humans*. BMC Genomics, 2017. **18**(1): p. 890.
15. Ye, K., et al., *Extensive pathogenicity of mitochondrial heteroplasmy in healthy human individuals*. Proceedings of the National Academy of Sciences, 2014. **111**(29): p. 10654-10659.
16. Alston, C.L., et al., *The genetics and pathology of mitochondrial disease*. The Journal of Pathology, 2017. **241**(2): p. 236-250.

## *Chapter 5 – Very Short Mitochondrial DNA Fragments and Heteroplasmy in Human Plasma* [3]

### 5.1 Abstract

Cell free DNA (cfDNA) has received increasing attention and has been studied in a broad range of clinical conditions implicating inflammation, cancer, and aging. However, few studies have focused on mitochondrial DNA (mtDNA) in the cell free form. We optimized DNA isolation and next-generation sequencing library preparation protocols to better retain short DNA fragments from plasma, and applied these optimized methods to plasma samples from patients with sepsis. Our methods can retain substantially shorter DNA fragments than the standard method, resulting in an average of 11.5-fold increase in short DNA fragments yield (DNA < 100bp). We report that cf-mtDNA in plasma is highly enriched in short-size cfDNA (30 ~ 60 bp). Motivated by this unique size distribution, we size-selected short cfDNA fragments from the sequencing library, which further increased the mtDNA recovery rate by an average of 10.4-fold. We then detected mixtures of different mtDNA sequences, termed heteroplasmy, in plasma from 3 patients. In one patient who previously received bone marrow transplantation, different minor allele frequencies were observed between plasma and leukocyte at heteroplasmic mtDNA sites, consistent with mixed-tissue origin for plasma DNA. This study is the first report of genome-wide identification of mtDNA heteroplasmy in human plasma.

---

[3] Published on Scientific Reports. Ruoyu Zhang and Kiichi Nakahira contribute equally to this work.

## 5.2 Introduction

Circulating cell free (cfDNA) has been proposed as a universal diagnostic and monitoring biomarker for many clinical applications, including cancer monitoring, prenatal diagnosis, and transplantation allograft rejection [1-3]. Although a number of the current studies investigating cfDNA have focused on cell free nuclear DNA (nDNA) in plasma, emerging evidence suggests that cell free mtDNA (cf-mtDNA) is also involved in disease progression. For instance, elevated cf-mtDNA concentrations have been observed in various diseases such as breast cancer, stroke, and myocardial infarction [4-6]. Furthermore, clinical reports have shown that the release of mtDNA into plasma is involved in immune responses [7], and increase with aging [8], suggesting that cf-mtDNA may serve as a biomarker to monitor disease onset and/or progression.

Although the origin of cf-mtDNA remains unclear, it has been suggested that mtDNA is released from apoptotic cells or necrotic cells [9, 10]. Interestingly cf-mtDNA levels are not always correlated with cf-nDNA levels in certain pathological conditions such as cancer [11], implying that cf-mtDNA may provide its unique patho-physiological information distinct from nDNA. It has been well reported that the size distribution of cf nDNA peaks at around 167 bp, suggesting cf nDNA may bind to histones and circulate as intact nucleosomes in blood [12]. Unlike nDNA, mtDNA lacks the protection of histones, making it more vulnerable to degradation [13], and possibly causing cf-mtDNA fragments to be shorter than cf nDNA. Ellinger *et. al*. [14], demonstrated that the levels of circulating mtDNA fragments (79 bp and 220 bp) were higher in patients with testicular germ cell cancer, compared to the control subjects.

Importantly the diagnostic information (*e.g.*, receiver operator characteristic (ROC) curve analysis) of 79 bp mtDNA fragments was relatively higher than that of 220 bp mtDNA fragments (REF), implying the importance of short fragment cf-mtDNA in human diseases. However, until now cf-mtDNA has been not fully characterized yet. Next generation sequencing makes it possible to measure the length of individual plasma DNA fragments at single nucleotide resolution. Using this sequence technique, Lo et al [13] showed that the size peak of plasma cf-mtDNA is ~140 bp, shorter than that of the predominant nDNA molecules in plasma (~167 bp). However, their strategy for cf-mtDNA size profiling was still technically limited by standard DNA library preparation methods [15], which contain several purification steps with poor recovery rates for short DNA fragment (<100 bp). In addition to measuring the length distribution of cf-mtDNA, deep sequencing also allows us to assess cf-mtDNA variants, including heteroplasmy. Heteroplasmy is defined as the coexistence of different mtDNA sequences within an individual, and has been reported to be implicated in various human diseases [16-18]. While heteroplasmy patterns have been well studied in different organs or tissues [19-21], their prevalence in cell free format remain unknown.

In this study, we optimized plasma DNA isolation and library preparation protocols in order to retain short DNA fragments. By our optimized protocol, we recovered substantially shorter DNA fragments than reported in previous studies, yielding a more comprehensive size profile of plasma DNA. Our results indicate that the average length of cf-mtDNA is much shorter than previously reported[13]. We further improved the recovery rate of cf-mtDNA by size selecting short DNA fragments in the

sequencing library. Using this strategy in combination with massive parallel sequencing, we investigated heteroplasmy patterns in cf-mtDNA and detected heteroplasmy in 3 individuals' cf-mtDNA. Interestingly, further investigation indicated that one individual who had bone marrow transplantation (BMT) previously was likely to have both donor and recipient-specific DNA in plasma. We observed very different heteroplasmy patterns between white blood cell (WBC) and plasma in this individual, which may indicate that cf-mtDNA originates from a mixture of different organs. Thus, heteroplasmy in cf-mtDNA may have the potential to provide information on patients' disease status.

## 5.3 Materials and Methods

### 5.3.1 Clinical sample collection and plasma processing

Blood samples were collected from sepsis patients registered in The New York Presbyterian-Weill Cornell Medicine Hospital Research Registry and Human Sample Repository for the study of the Biology of Critical Illness (Weill Cornell Medicine Biobank of Critical Illness (WCM BoCI)). WCMC BoCI is an ongoing registry that collects demographic and clinical information, and blood specimens from patients admitted into the medical intensive care unit (MICU). WCM BoCI is approved by Weill Cornell Medicine Institutional Review Board (IRB) and is carried out in "accordance" with IRB protocol #1405015116. All adults (age 18 and older) admitted to the MICU are considered for enrollment. The presence of any of the following excludes a patient from study enrollment: 1) Subjects with mental handicaps. 2) Subjects who are unable to provide consent directly and for whom an appropriate legal

representative cannot be found to provide consent. 3) Subjects who have previously indicated that they do not wish to be enrolled in this study, (e.g. during a prior admission to the MICU). 4) Subjects admitted to the MICU purely to facilitate comfort care and weaning of medical intervention at end of life. 5) Subjects who are Jehovah's witnesses or are otherwise unable or unwilling to receive blood transfusions during hospitalization. 6) Subjects with a hemoglobin level of less than 7 g/dL upon admission to the MICU or subjects with rare blood groups, or other antigens that might require minimization of blood draws. 7) Subjects with active bleeding at the time of MICU admission with hemoglobin levels less than 8 g/dL and subjects suffering from acute myocardial infarction with hemoglobin levels less than 8 g/dL. Written informed consent consistent with the research purposes in this proposal is obtained from all of the subjects prior to study procedures. Sepsis was identified according to the 2001 SCCM/ESICM/ACCP/ATS/SIS International Sepsis Definitions Conference guidelines [22]. Blood samples were drawn and transferred into EDTA-coated blood collection tubes within 24 h of study inclusion as previously described [23-25] processed within 4 h after venipuncture, and was and stored at −80°C. 50 µl of plasma was then mixed with 170 µl of sterile PBS, followed by brief vortex. The diluted plasma was centrifuged at 700g at 4°C for 5 min, and the supernatant (210 µl) was carefully saved by avoiding touching any pellets and the bottom of the tubes with pipette tips. The obtained supernatant was further centrifuged at 18,000g at 4°C for 15 min, and the resulting supernatant (200 µl) was carefully saved. Contamination of cells, cell debris, or pellets into supernatant might lead to a significant change of the results [24].

### 5.3.2 Library preparation for plasma DNA

The obtained plasma samples were processed for DNA isolation using an optimization of a previously published standard method [24]. In the standard method: DNA was extracted by DNeasy Blood and Tissue Kit (No. 69504; Qiagen), and indexed DNA libraries were prepared as described in [15], using the following steps: i) End repair. ii) Adenine tailing. iii) Adaptor ligation. iv) Library amplification. DNA was purified by 1.5X SPRI beads after each step. For our optimized method: DNA was extracted by QIAamp DSP DNA Blood Mini Kit (No. 61104; Qiagen). DNA libraries were prepared by KAPA Hyper Prep Kits (No. KK8502; Kapa), using the following steps: i) End repair and adenine tailing combined in a single step. ii) Proceeded directly to adaptor ligation. iii) Adaptor ligated DNA was purified by 1.5X SPRI beads and amplified by indexed primers. For size selection to enrich mtDNA fragments in the library, the library was size selected from 150 to 190 bp by Pippin electrophoresis (Sage Sciences Blue Pippen). Libraries were sequenced by the Illumina Hiseq platform with paired-end read lengths of 150 bp.

### 5.3.3 Library preparation for white blood cell DNA

DNA from white blood cell was isolated by QIAamp DSP DNA Blood Mini Kit (No. 61104; Qiagen). . MtDNA was amplified as two 9 kb long amplicons (primer sequence: Pair1 Forward: 5'-GATATCATAGCTCAGACCATACC-3'; Reverse: 5'-CCACATCACTCGAGACGTAAAT-3'. Pair2: 5'-CTGCTGGCATCACTATACTACTA-3'; Reverse: 5'-GATGTGTAGGAAGAGGCAGATAAA-3'.) and the PCR products were mixed with

equimolar ratio. Indexed sequencing libraries were prepared by Nextera XT DNA library preparation kit (Illumina, FC-131-1024). Libraries were sequenced by the Illumina Hiseq and Miseq platform with paired-end reads.

### 5.3.4 Analysis workflow for Next-Generation Sequencing data

Raw sequencing reads were trimmed by Trimmomatic [26] to remove adaptor sequence. Cleaned reads were then mapped to hg19 human reference sequence with rCRS mitochondrial reference genome by bowtie2 [27]. PCR duplicates were removed with Picard (http://picard.sourceforge.net.). Read pairs with proper orientation, mapping quality > 20, and mismatches less than 5% of trimmed read length were retained for downstream analysis. DNA fragment length was inferred from the coordinates of the nucleotides at the end of each pair of reads.

Nuclear mitochondrial DNA segments (NUMTs) in nuclear genome might be mismapped to mitochondrial genome and counted as mtDNA reads. To minimize the effect of NUMTs, reads mapped to mitochondrial genome were remapped to the hg19 reference genome, and we only retained reads that uniquely mapped to the mitochondrial genome for downstream analysis.

### 5.3.5 Heteroplasmy identification for white blood cell and plasma

Sequencing data for each position of mtDNA was extracted by Samtools mpileup [28], bases were further filtered by sequencing quality (>=20). Heteroplasmy in WBC was defined with following criteria: i) Sequencing coverage > 400. ii) Minor allele frequency >= 1%. iii) Minor allele frequency is no less than 0.6% for both strands, and

minor allele frequency at both strands is not significantly different (chi-square test). Heteroplasmy criteria are loosed for plasma DNA due to relative low coverage: i) Sequencing coverage > 50. ii) Minor allele counts >=4. iii) Observed at least once in both strands.

To minimize the false positive heteroplasimes introduced by sequencing error, we applied a maximum likelihood based algorithm to take the sequencing quality at each base in each sequence read into account [20, 21]. At an interested heteroplasimic site, if there were l bases with major alleles and k bases with minor alleles, and the probability of sequencing error corresponding to the sequencing quality of each base was εj, the likelihood function of the major allele frequency f can be derived as:

$$L(f)= \prod_{j=1}^{l}((1-f)\varepsilon_j + f(1-\varepsilon_j)) \prod_{j=1}^{k}((1-f)(1-\varepsilon_j) + f\varepsilon_j)$$

f can be estimated by heteroplasmic model ($f_{het}$) and homoplasmic model ($f_{homo}$) respectively, and log likelihood ratio of these two models can be calculated as $LLR = \log(L(\hat{f}_{\text{het}}) / L(\hat{f}_{\text{homo}}))$. $LLR > 5$ indicates a high confidence heteroplasmy (false positive rate < 10-5). We confirmed that heteroplasmy identified from previous step all had LLR scores > 5.

The strength of a heteroplasmy signal at an mtDNA site may be different between WBC and plasma, due to different mapping criteria. In order to compare heteroplasmy at same sites between WBC and plasma, we defined “heteroplasmy in both WBC and plasma” by the following criteria: i) LLR score > 5 in either WBC or plasma. ii) Major and minor alleles need present in both WBC and plasma. iii) Minor allele count >= 2.

iv) Minor allele count >= 1 on both strands. Otherwise, the heteroplasmy would be considered as only in WBC or only in plasma.

### 5.3.6 Haplotype Analysis

For both WBC and plasma, we constructed two consensus mtDNA sequences, one covering the major alleles at heteroplasmic sites, the other covering minor alleles. We then sent two sequences to HaploGrep [29] to classify haplogroups. The resulting haplogroups were denoted as major allele haplogroup and minor allele haplogroup respectively.

### 5.3.7 Data Access

Sequencing data have been archived in the National Center for Biotechnology Information Gene Expression Omnibus under accession number GSE81178.

## 5.4 Results

### 5.4.1 Plasma mtDNA has a distinct size distribution compared to nDNA

While most of recent plasma DNA extraction methods are column-based, in part due to the need for processing a large number of human samples, short DNA fragment recovery rates are limited. In addition, current standard library preparation protocols include several purification steps with either SPRI beads or columns, which also has poor short DNA fragment recovery rate [30]. Therefore, although these methods are widely used in a range of applications, they are unlikely to capture the complete cfDNA size profile. To circumvent these issues, we optimized these steps in order to

better preserve short fragments (Figure 5-1): First, we used QIAamp DSP DNA Blood Mini Kit for DNA extraction from plasma, which we verified was able to retain DNA fragments as short as 50 bp (Figure 5-2). Second, to avoid any further purification steps before sequencing adaptor ligation, we combined end-repair and dA tailing steps in a single reaction, and then proceeded directly to the adaptor ligation reaction. We also avoided the size selection step which is used in current protocols to remove adapter-dimers, since this may introduce some arbitrary elimination of DNA fragments. To validate the improvement of our optimized methods, we also extract DNA by DNeasy Blood and Tissue Kit and performed standard sequencing library preparation on same plasma sample as a comparison **(**Figure 5-1**).**

Pair-end sequencing was used to accurately measure the length of each individual DNA fragment. Adaptor sequences were trimmed by Trimmomatic [26] prior to alignment. Trimmed reads were aligned against the human reference genome hg19, with mitochondrial reference rCRS, using bowtie2 [27]. DNA fragment lengths were determined by the paired-end read alignment coordinates. Seven samples were processed by both optimized and standard methods. Compared to the standard method, our optimized method preserved many shorter DNA fragments; the proportion of DNA fragment <100 bp was 19.05% by the optimized method, but only 1.77% by the standard method. Figure 5-3 shows the fragment length distribution of mtDNA and nDNA in both methods. Although by the standard method, cf-mtDNA (Figure 5-3B blue line) already shows a left-shifted peak (around 90 bp), our optimized protocol revealed that cf-mtDNA (red line) has a length peak at around 42 bp, much shorter than previously reported [13]. nDNA has a peak at 167 bp (Figure 5-3C), consistent

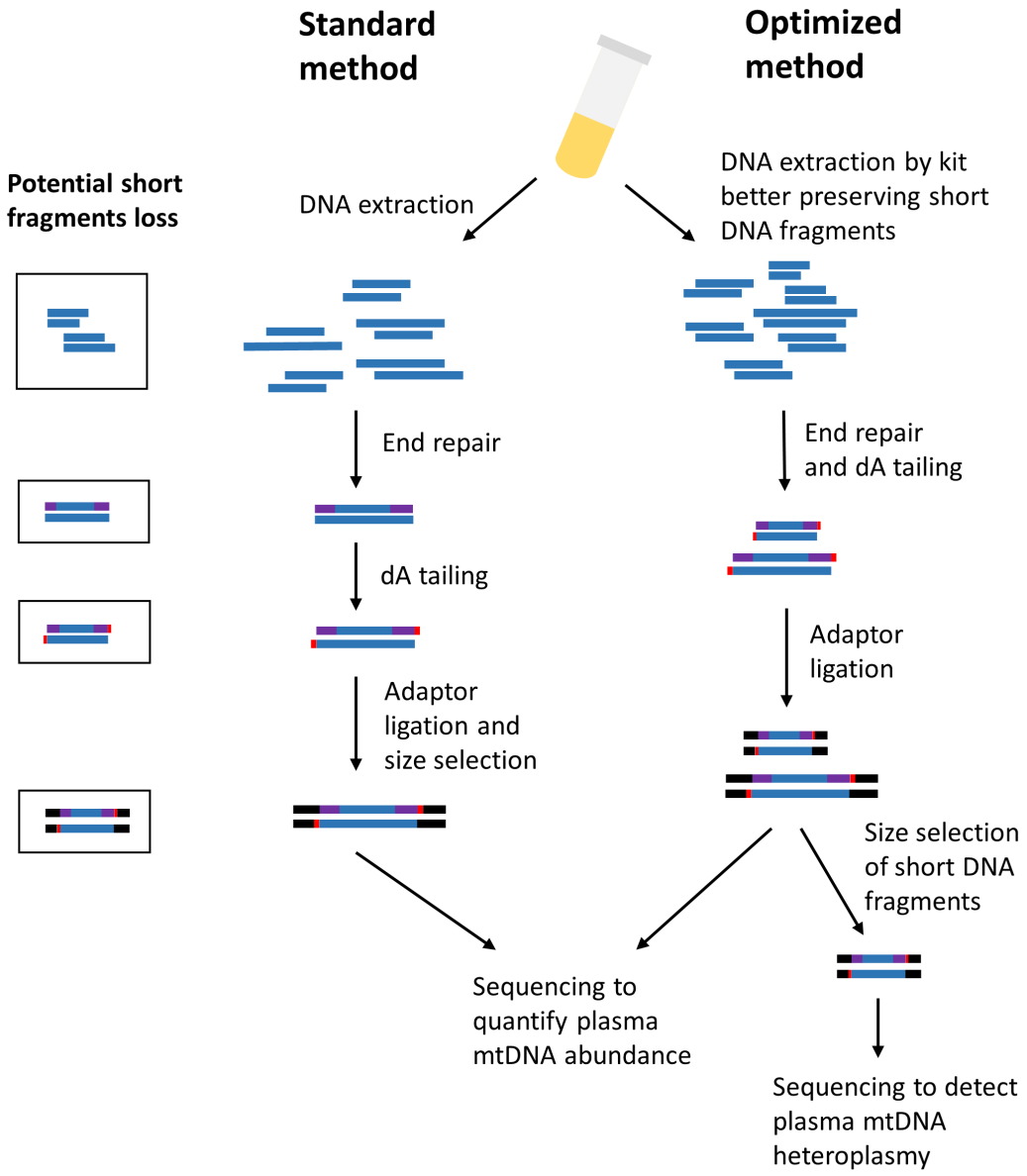


Figure 5-1. Comparison of the standard method and optimized method.

The optimized method has the following improvements: 1. using a DNA isolation kit which can better preserve short DNA fragments 2. combined end repair and dA tailing in a single step, avoiding purification before sequencing adaptor ligation. Black box indicates potential short DNA fragment loss during each step in the standard method. The optimized method produces a 2.41 to 17.88-fold increased in mtDNA concentration. The optimized method can further involve size selection of the ligated library product to enrich mtDNA.

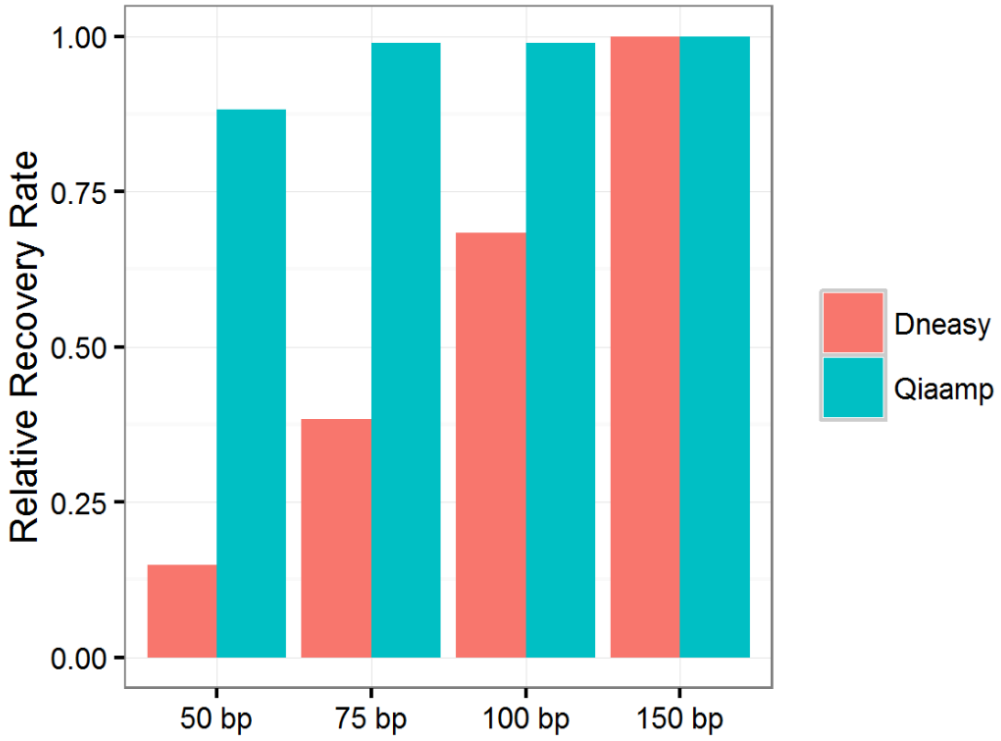


Figure 5-2. Relative recovery rate of DNA fragments with different lengths

Relative recovery rate of DNA fragments with different lengths (recovery rate of 150 bp DNA set as 1). QIAamp has better performance than DNeasy for short DNA fragments (<50 bp).

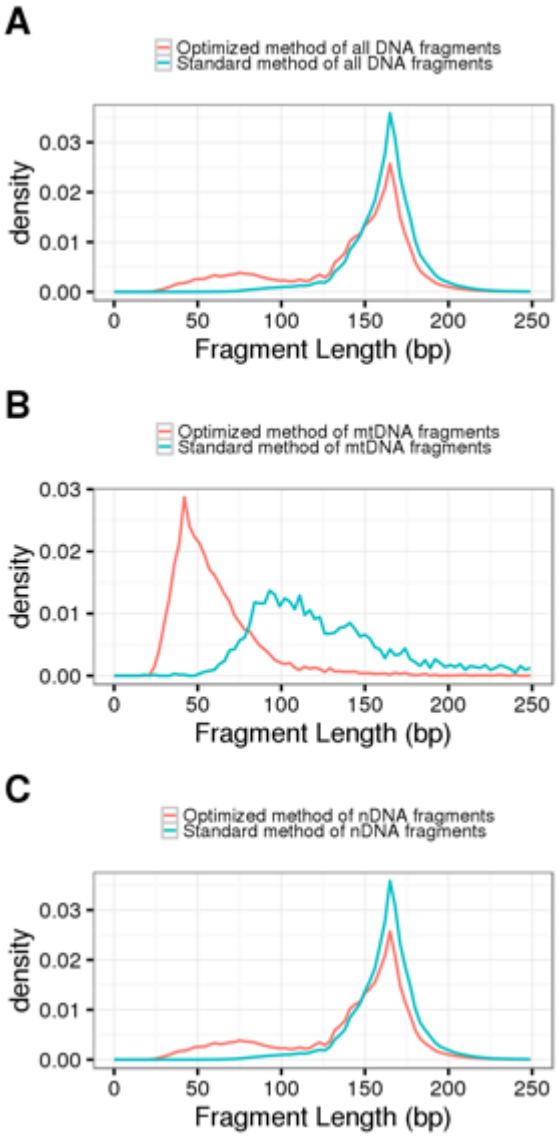


Figure 5-3. Plasma DNA size distribution.

(A) The size distribution of all DNA fragments in plasma. Optimized method and standard method were indicated in red line and blue line respectively. Both methods showed a sharp peak at ~ 166-167 bp; in optimized method, 23.2% DNA fragments have length shorter than 100 bp. while in standard method, there was only 1.7% DNA fragments <100 bp. (B) The size distribution of plasma mtDNA. The length of mtDNA peaks at 42 bp in our optimized method, and only small fraction of mtDNA had length greater than 100 bp. In standard method, mtDNA length was also shorter than nDNA, but the size peak was ~90 bp, much longer than our optimized method. (C) The size distribution of the nuclear DNA in plasma. The size peak was at ~ 166-167 bp, which was consistent with previous reports [12].

with previous studies [12, 13]. In addition, we noticed that by our optimized method, nDNA fragments also had a flat peak for shorter fragments around 77 bp (Figure 5-3C red line), which has not been found in previous studies (Figure 5-3C blue line). These results suggest that some nDNA also lacks histone protection, resulting in their degradation and shorter length distribution.

5.4.2 cf-mtDNA recovery rate is improved by new method

We then compared the mtDNA concentration for the same individual resulting from the optimized and standard methods. We calculated the mtDNA concentration as the ratio of mtDNA reads to the total number of sequenced reads which can be mapped to the human genome.

$$\text{mtDNA fractional concentration} = \frac{\text{mtDNA reads number}}{\text{Total mapped reads number}}$$

The improvement of the mtDNA recovery rate was shown in Table 5-1. The fractional concentration of mtDNA in plasma increased by 2.41 to 17.88-fold among different individuals (Table 5-1). Previously, mtDNA was reported to make up about 0.003% in the plasma cfDNA in hepatocellular carcinoma patients[13]. We obtained a similar average mtDNA concentration of 0.0098% using the standard method. However, in our optimized method, the average mtDNA fractional concentration increased to 0.1428%. These results showed that our optimized method can recover more mtDNA, and thus estimate mtDNA fractional concentration in plasma much more accurately

Table 5-1. mtDNA fractional concentration by different approaches

| ID | mtDNA concentration (%) | | | Fold increase in mtDNA yield | | |
|---|---|---|---|---|---|---|
| | Standard Method | Optimized Method | Optimized Method with Size Selection | Optimized / Standard | (Optimized with Size Selection) / Optimized | (Optimized with Size Selection) / Standard |
| 1 | 0.00332 | 0.02114 | 0.17842 | 6.36 | 8.44 | 53.73 |
| 2 | 0.00307 | 0.01428 | 0.05366 | 4.65 | 3.76 | 17.49 |
| 17 | 0.00543 | 0.01310 | 0.15901 | 2.41 | 12.14 | 29.27 |
| 42 | 0.05152 | 0.92133 | 2.75686 | 17.88 | 2.99 | 53.51 |
| 69 | 0.00188 | 0.00688 | 0.10392 | 3.66 | 15.10 | 55.27 |
| 78 | 0.00114 | 0.01044 | 0.14689 | 9.16 | 14.07 | 128.93 |
| 93 | 0.00237 | 0.01256 | 0.20437 | 5.30 | 16.27 | 86.29 |

than previous approaches.

### 5.4.3 Size selection further improves cell free mitochondrial DNA recover rate

While each cell has two copies of nuclear genome, there are hundreds to thousands of mtDNA copies in a typical human cell. Mutations can be present in all mtDNA copies (homoplasmy) or only exist in a proportion of mtDNA (heteroplasmy). Few studies have investigated mtDNA heteroplasmy in plasma due to the extremely low concentration of plasma mtDNA.

Unlike total genomic DNA extracted from cells, plasma DNA is fragmented, therefore it's difficult to enrich plasma mtDNA by PCR. Since we have discovered that the size distributions for cf-mtDNA and nDNA are different (Figure 5-3), we then attempted to use this feature to recover more mtDNA reads. We calculated cf-mtDNA concentration in a series of size intervals by the sliding window strategy. Fig 3 showed that mtDNA fractional concentration varied dramatically in different size intervals, peaking at 43 bp, mtDNA with 0.8%. This number was 10 times higher than the mtDNA concentration across all size range (0.07%). We then size selected DNA molecules (with the ligated sequencing adaptors) between 150 to 190 bp including the ligated sequencing adaptors (corresponding to a DNA insert size of 30 to 60 bp) from the sequencing library in order to increase mtDNA concentration (Figure 5-4, blue dash lines).

Our optimized experiment protocols for library preparation improved the mtDNA recovery rate many-fold (Table 5-1). By size selection, we were able to further increase mtDNA concentration by another 2.99 to 16.27-fold from results without size

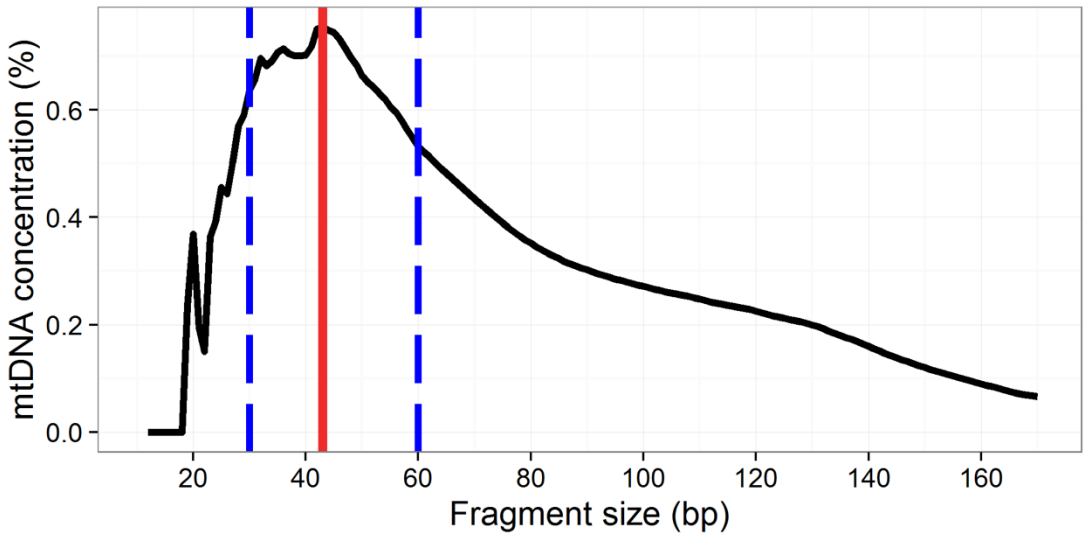


Figure 5-4. mtDNA fractional concentration in different size windows.

In plasma, most of mtDNA exists as short fragments, while the majority of nDNA has larger length. mtDNA thus has relatively high fractional concentration in short size ranges. Blue dashed lines indicate the size region between 30 bp to 60 bp, where mtDNA has highest concentration. Therefore, size selection for plasma DNA from this size region produces substantial enrichment of mtDNA.

selection. Compared to the standard method, the overall fold increase is from 17.49 to 128.93 fold for different individuals (Table 5-1). This improvement of mtDNA recovery rate made it feasible to assess cf-mtDNA heteroplsamy**.**

### 5.4.4 mtDNA heteroplasmy in plasma

Plasma cfDNA can be a mixture of DNA released from different origins origins (cell types, tissues or organs). In healthy individuals, plasma cfDNA can be derived from hematopoietic cell lineages [9, 10], but in disease contexts, multiple organs may be damaged and release cfDNA. Furthermore, since heteroplasmy frequencies can vary from cells to cells and tissues to tissues within the same individual [19], we suggest that the distribution mtDNA heteroplasmy in plasma can be informative to infer the tissue origins of cfDNA, and from this can inter patients' disease status.

However, as we indicated above, even with our improved methods, mtDNA only comprises a very small fraction of the total cfDNA in plasma, potentially making it very costly to retrieve enough sequencing reads to estimate heteroplasmy. Nevertheless, by using both our optimized method and size selection, we were able to evaluate heteroplasmy in 3 individuals, (ID: 1, 42, 93, average sequencing coverage 35.9, 503.8, 45.3 respectively). We also sequenced mtDNA in WBC and identified heteroplasmy in these three individuals for comparison (see Methods). Because the sequencing coverage for plasma was lower than WBC, we used less stringent criteria to identify heteroplasmy in plasma (total coverage >= 40, overall minor allele count >= 4, minor allele count >=1 on each strand).

In patient #1 and 93, we did not identify any heteroplasmy in their WBC using 1%

frequency cutoff (see Methods), while we found that patient #1 had heteroplasmy at mtDNA position 11836 and patient 93 had heteroplasmy at position 16111 in their plasma (Table 5-2). We then manually inspected sequencing details for these heteroplasmy positions in WBC. We found that the minor alleles were indeed presenting but were filtered out because they did not pass the 1% frequency threshold (0.46% and 0.26% respectively). Although the minor allele signals were weaker than those in plasma, we still considered these heteroplasmy were in both sides (see criteria in Methods). We then compared minor allele frequencies between plasma and WBC. The frequencies were 7.1% and 6.9% in plasma, but were lower than 0.5% in WBC (Table 5-2), indicating heteroplasmy pattern can be different between plasma and WBC.

In patient #42, we identified 12 heteroplasmy in plasma and WBC. Among them, 8 were commonly presented in both WBC and plasma, and 3 were presented in WBC and 1 in plasma (Table 5-3, Table 5-4, Table 5-5). Similar as patient #1 and 93, we also observed different heterolasmy patterns between WBC and plasma. For example, we only observed T (100%) at position 72 in WBC, but we observed high frequency of C alleles in plasma (15.2%). Interestingly, C allele has been shown to be very frequently observed in liver and kidney at this position [19]. Moreover, we noticed among these common 8 heteroplasmy sites, 7 of them had different major alleles between WBC and plasma. For instance, at position 207, major allele was G in plasma (98.9%, with 1.1% A), but in WBC, major allele was A (96.8% with 3.2%G). These differences may indicate that cf-mtDNA is released from multiple tissues and each tissue may contribute to different proportion of cfDNA in plasma. The heteroplasmy

Table 5-2. mtDNA heteroplasmy in patient 1, 93

| ID | Position | Type | Depth | Allele1 | Frequency of Allele1 | Allele2 | Frequency of Allele2 |
|---|---|---|---|---|---|---|---|
| 1 | 11836 | Plasma | 56 | A | 92.9% | G | 7.1% |
| | | Cell | 3714 | | 99.54% | | 0.46% |
| 93 | 16111 | Plasma | 58 | T | 93.10% | C | 6.90% |
| | | Cell | 2663 | | 99.74% | | 0.26% |

Table 5-3. mtDNA heteroplasmy present in both WBC and plasma in patient 42

| Position | Type | Depth | Allele1 | Frequency of Allele1 | Allele2 | Frequency of Allele2 |
|---|---|---|---|---|---|---|
| 186 | Plasma | 481 | T | 2.1% | G | 97.9% |
| | Cell | 2325 | | 0.13% | | 99.87% |
| 207 | Plasma | 335 | A | 1.8% | G | 98.2% |
| | Cell | 2040 | | 96.8% | | 3.2% |
| 8425 | Plasma | 46 | A | 19.6% | G | 80.4% |
| | Cell | 3352 | | 99.9% | | 0.1% |
| 12127 | Plasma | 193 | A | 92.7% | G | 7.3% |
| | Cell | 5602 | | 0.4% | | 99.6% |
| 13708 | Plasma | 709 | A | 0.3% | G | 99.7% |
| | Cell | 2015 | | 96.7% | | 3.2% |
| 14364 | Plasma | 870 | A | 0.7% | G | 99.2% |
| | Cell | 4353 | | 95.6% | | 4.4% |
| 16126 | Plasma | 264 | T | 17.0% | C | 83.0% |
| | Cell | 7077 | | 99.8% | | 0.2% |
| 16129 | Plasma | 157 | A | 72.0% | G | 27.4% |
| | Cell | 6878 | | 1.2% | | 98.7% |

Table 5-4. mtDNA heteroplasmy present only in WBC in patient 42

| Position | Type | Depth | Allele1 | Frequency of Allele1 | Allele2 | Frequency of Allele2 |
|---|---|---|---|---|---|---|
| 477 | Plasma | 452 | T | 100% | C | 0% |
| | Cell | 942 | | 96.1%% | | 3.9% |
| 3010 | Plasma | 865 | A | 0% | G | 100% |
| | Cell | 7087 | | 2.2%% | | 97.8% |
| 14350 | Plasma | 667 | T | 0% | C | 100% |
| | Cell | 4788 | | 2.5% | | 97.5% |

Table 5-5. mtDNA heteroplasmy present only in plasma in patient 42

| Position | Type | Depth | Allele1 | Frequency of Allele1 | Allele2 | Frequency of Allele2 |
|---|---|---|---|---|---|---|
| 72 | Plasma | 1199 | T | 84.8% | C | 15.1% |
| | Cell | 2325 | | 100% | | 0% |

difference between plasma and WBC could be caused by bone marrow transplantation.

Interestingly, the medical record indicated that patient #42 had allogeneic BMT due to hematological malignancy. Thus, plasma cfDNA in this patient could be derived from either the recipient's' or donor's cells/tissues. To analyze the mtDNA haplogroup of this patient, we constructed two consensus sequences, covering either the major and minor alleles at heteroplasmic sites (see Methods). The haplogroup analysis showed that the major allele haplogroups of WBC and plasma DNA were H18 and H2, respectively. One possible explanation for this difference is that WBC and plasma DNA were derived from different subjects (*i.e.*, either the BMT recipient or donor). In addition, the minor allele haplogroup of plasma mtDNA was H18, suggesting a proportion of plasma DNA may be released from WBC.

## 5.5 Discussion

cf-mtDNA has great potential to serve as a biomarker in various clinical situations. A number of studies have used real-time PCR to show that circulating cf-mtDNA is elevated in various disease conditions. However, our results indicate that these methods are not capable of detecting the majority of mtDNA molecules in plasma due to their ultra-short length. In this study, we optimized DNA isolation and library preparation protocols in order to preserve short DNA fragments in plasma. We verified that optimized method was able to capture short DNA fragments (<100 bp), and we found that mtDNA has a very short length in plasma, peaking at 42 bp, which is much shorter than previously reported [13]. Our optimized protocol can increase

mtDNA recovery rate by as much as 19 fold. Compared to real-time PCR based methods, our method can quantify mtDNA content in plasma much more accurately. The endosymbiont hypothesis suggested that the mitochondrion evolved from a bacterial progenitor [31], therefore mtDNA contains bacterial specific sequence motifs [32, 33]. Thus, the release of mtDNA into the circulation may cause severe immune consequences, especially when the release is enhanced in specific disease conditions. For example, mtDNA has been shown to increase in fluids in joints of patients with rheumatoid arthritis, and induce inflammation in vivo [34]. Furthermore, liver and kidney which may be responsible for eliminating circulating cf DNA are often damaged in critically illness such as sepsis due to systemic inflammation or infection [35, 36]. Such organ dysfunction can further lead to leveraged DNA releasing. Thus plasma mtDNA would be a better indicator of overall mtDNA status than WBC mtDNA. Nonetheless, few studies have been conducted to evaluate mtDNA heteroplasmy in the cell free form. One technical difficulty is the ultra-low concentration of mtDNA in plasma as well as its unique size distribution. By size selecting short fragments from the DNA sequencing library, we were be able to further enrich mtDNA by up to >100 fold compared to standard methods, enabling us to investigate cf-mtDNA heteroplasmy in three patients.

We observed different heteroplasmy patterns between WBC and plasma in all three patients. Most of the heteroplasmic positions have different allele frequencies between WBC and plasma. In patient #42 who had previously received BMT therapy, 7 out of 12 heteroplasmic positions even had flipped major alleles between WBC and plasma. In general, high doses of chemotherapy and/or radiation are given to patients who plan

to receive BMT in order to destroy cancer cells or the defective bone marrow (BM) of the patients. Therefore, after BMT the recipient's new BM is mostly replaced with the donor's, implying that DNA in newly generated WBC of the patients (the recipient) is likely to be the same as the donor's. However, in some cases the recipient's tumor cells can survive and remain in the BM even after radiation and chemotherapy, which may lead to co-existence of WBC derived from BM of both the recipient and donor. Taken this into consideration, it is not surprising that we observed different haplogroups for WBC and plasma cf-mtDNA. In addition, although the frequencies were relatively low, we could still detect WBC-derived allele in plasma mtDNA, which could be donor-derived mtDNA. mtDNA provides a valuable tool to identify DNA origins, since the high number of nucleotide polymorphisms in mtDNA can allow discrimination between the donor and recipient. cf-mtDNA has not been deeply studied in transplantation medicine yet, but our results suggest that mtDNA has great potential in monitoring allograft health.

Another possible explanation for this inconsistent heteroplasmy between plasma and WBC is that plasma DNA is a mixture of DNA released from different organs or cell types. For example, it has been reported that high levels of heteroplasmy are observed at position 72 in liver and kidney, moderate levels in skeletal muscle, but low levels in all other tissues [19]. In our analysis, the heteroplasmic C allele at position 72 showed a 15.4% frequency in patient #42, and we found that this patient had acute kidney injury when the plasma sample was collected. It is possible that more mtDNA may have been released from the damaged kidneys, contributing to the high level of the C allele in the plasma. This result suggests that patterns of heteroplasmy in plasma can

be used to infer cell death in specific tissues or organs, providing more information about patients' disease status.

Our study characterized certain properties of plasma mtDNA, which will give inform future plasma DNA studies. Using our optimized experimental protocols, the mtDNA concentration in plasma can be measured more accurately, which can be applied to study changes in plasma mtDNA concentrations in a wide range of diseases such as cancer, stroke and cardiovascular diseases. We also showed that plasma mtDNA can provide information on heteroplasmy that cannot be provided by a single cell type, which can be extended to infer the tissue origins of cfDNA under specific disease conditions, and provide more information about patients' disease status.

## 5.6 Acknowledgement

We thank Mr. Yiping Wang, Drs. Kaixiong Ye, Shu Hisata and Partin Picard for their discussion and comments on the manuscript. We also thank WCMC BioBank, Dr. Maria Angelica Pabon Porras and Dr. Eli Finkelsztein for blood specimen preparation.

5.7 Reference


1. Dawson, S.-J., et al., *Analysis of circulating tumor DNA to monitor metastatic breast cancer*. New England Journal of Medicine, 2013. **368**(13): p. 1199-1209.
2. Chiu, R.W., et al., *Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma*. Proceedings of the National Academy of Sciences, 2008. **105**(51): p. 20458-20463.
3. De Vlaminck, I., et al., *Circulating cell-free DNA enables noninvasive diagnosis of heart transplant rejection*. Science translational medicine, 2014. **6**(241): p. 241ra77-241ra77.
4. Kohler, C., et al., *Levels of plasma circulating cell free nuclear and mitochondrial DNA as potential biomarkers for breast tumors*. Mol Cancer, 2009. **8**: p. 105.
5. Rainer, T.H., et al., *Prognostic use of circulating plasma nucleic acid concentrations in patients with acute stroke*. Clinical Chemistry, 2003. **49**(4): p. 562-569.
6. Wang, L., et al., *Plasma nuclear and mitochondrial DNA levels in acute myocardial infarction patients*. Coron Artery Dis, 2015. **26**(4): p. 296-300.
7. Fang, C., X. Wei, and Y. Wei, *Mitochondrial DNA in the regulation of innate immune responses*. Protein & Cell, 2015. **7**(1): p. 11-16.
8. Pinti, M., et al., *Circulating mitochondrial DNA increases with age and is a familiar trait: Implications for "inflamm-aging"*. European Journal of Immunology, 2014. **44**(5): p. 1552-1562.
9. Stroun, M., et al., *The origin and mechanism of circulating DNA*. Annals of the New York Academy of Sciences, 2000. **906**(1): p. 161-168.
10. Jahr, S., et al., *DNA fragments in the blood plasma of cancer patients: quantitations and evidence for their origin from apoptotic and necrotic cells*. Cancer research, 2001. **61**(4): p. 1659-1665.
11. Mehra, N., et al., *Circulating mitochondrial nucleic acids have prognostic value for survival in patients with advanced prostate cancer*. Clin Cancer Res, 2007. **13**(2 Pt 1): p. 421-6.
12. Lo, Y.D., et al., *Maternal plasma DNA sequencing reveals the genome-wide genetic and mutational profile of the fetus*. Science translational medicine, 2010. **2**(61): p. 61ra91-61ra91.
13. Jiang, P., et al., *Lengthening and shortening of plasma DNA in hepatocellular carcinoma patients*. Proc Natl Acad Sci U S A, 2015. **112**(11): p. E1317-25.
14. Ellinger, J., et al., *Circulating mitochondrial DNA in the serum of patients with testicular germ cell cancer as a novel noninvasive diagnostic biomarker*. BJU Int, 2009. **104**(1): p. 48-52.
15. Wang, L., et al., *A low-cost library construction protocol and data analysis pipeline for Illumina-based strand-specific multiplex RNA-seq*. PLoS One, 2011. **6**(10): p. e26426.

16. Taylor, R.W. and D.M. Turnbull, *Mitochondrial DNA mutations in human disease*. Nature Reviews Genetics, 2005. **6**(5): p. 389-402.
17. Wallace, D.C. and D. Chalkia, *Mitochondrial DNA genetics and the heteroplasmy conundrum in evolution and disease*. Cold Spring Harb Perspect Biol, 2013. **5**(11): p. a021220.
18. Stewart, J.B. and P.F. Chinnery, *The dynamics of mitochondrial DNA heteroplasmy: implications for human health and disease*. Nat Rev Genet, 2015. **16**(9): p. 530-542.
19. Li, M., et al., *Extensive tissue-related and allele-related mtDNA heteroplasmy suggests positive selection for somatic mutations*. Proc Natl Acad Sci U S A, 2015. **112**(8): p. 2491-6.
20. Ye, K., et al., *Extensive pathogenicity of mitochondrial heteroplasmy in healthy human individuals*. Proceedings of the National Academy of Sciences, 2014. **111**(29): p. 10654-10659.
21. Ding, J., et al., *Assessing Mitochondrial DNA Variation and Copy Number in Lymphocytes of ~2,000 Sardinians Using Tailored Sequencing Analysis Tools*. PLoS Genet, 2015. **11**(7): p. e1005306.
22. Levy, M.M., et al., *2001 sccm/esicm/accp/ats/sis international sepsis definitions conference*. Intensive care medicine, 2003. **29**(4): p. 530-538.
23. Rogers, A.J., et al., *Metabolomic derangements are associated with mortality in critically ill adult patients*. PLoS One, 2014. **9**(1): p. e87538.
24. Nakahira, K., et al., *Circulating mitochondrial DNA in patients in the ICU as a marker of mortality: derivation and validation*, in *PLoS Med*. 2013. p. e1001577.
25. Dolinay, T., et al., *Inflammasome-regulated cytokines are critical mediators of acute lung injury*. Am J Respir Crit Care Med, 2012. **185**(11): p. 1225-34.
26. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data*. Bioinformatics, 2014. **30**(15): p. 2114-20.
27. Langmead, B. and S.L. Salzberg, *Fast gapped-read alignment with Bowtie 2*. Nature methods, 2012. **9**(4): p. 357-359.
28. Li, H., et al., *The Sequence Alignment/Map format and SAMtools*. Bioinformatics, 2009. **25**(16): p. 2078-9.
29. Kloss-Brandstätter, A., et al., *HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups*. Human Mutation, 2011. **32**(1): p. 25-32.
30. Bronner, I.F., et al., *Improved Protocols for Illumina Sequencing*. Current protocols in human genetics / editorial board, Jonathan L. Haines ... [et al.], 2009. **0 18**: p. 10.1002/0471142905.hg1802s62.
31. Sagan, L., *On the origin of mitosing cells*. J Theor Biol, 1967. **14**(3): p. 255-74.
32. Zhang, Q., et al., *Circulating mitochondrial DAMPs cause inflammatory responses to injury*. Nature, 2010. **464**(7285): p. 104-7.
33. Fang, C., X. Wei, and Y. Wei, *Mitochondrial DNA in the regulation of innate immune responses*. Protein & Cell, 2016. **7**(1): p. 11-16.
34. Collins, L.V., et al., *Endogenously oxidized mitochondrial DNA induces in vivo and in vitro inflammatory responses*. J Leukoc Biol, 2004. **75**(6): p. 995-

1000.
35. Lam, N.Y.L., et al., *Time Course of Early and Late Changes in Plasma DNA in Trauma Patients*. Clinical Chemistry, 2003. **49**(8): p. 1286-1291.
36. Tsumita, T. and M. Iwanaga, *Fate of injected deoxyribonucleic acid in mice*. Nature, 1963. **198**: p. 1088-9.

## *Chapter 6 AFTERWORD*

My graduate research focused on using next generation sequencing technology to identify mtDNA mutations, especially heteroplasmic mutations in human subjects, and to further investigate the implications of these variants in human diseases or/and their potential to serve as biomarkers in different medical conditions. My research provided some evidence for the association of mtDNA mutations with human diseases, and suggested that increasing knowledge of mitochondria and mtDNA can provide new opportunities for disease prevention and diagnosis. The potential applications and future follow-up studies of my research projects are discussed below.

**Project 1**: Independent impacts of aging on mitochondrial DNA quantity and quality in humans

This is a data mining project using the public available dataset from the UK10K project. It has been known that the dysfunction of mitochondria is a hallmark of aging. Inherited mtDNA mutations, or the depletion of mtDNA within cells, are major causes of human diseases, demonstrating that mtDNA integrity as well as its cellular content are critical for proper mitochondrial function. Evidence from small studies indicates that both mtDNA integrity and content may decrease with age, and that this may be a harbinger of the aging process in humans. In particular, spontaneous mtDNA mutations present at low levels, and which coexist with wild-type mtDNA copies (i.e., heteroplasmy) may alter cellular function in synergistic or independent ways. In this project, we analyzed whole genome sequencing data of 1,511 women with healthy phenotypic data from the UK10K project, and reported for the first time an independent effect of age on mtDNA heteroplasmy and copy number. Our results

demonstrate that aging is independently associated with high mtDNA mutation burden and lower copy number.

One limitation of this study is that we do not have data for both male and female subjects, so our findings may not be representative for both genders, given that there are some differences in the energy usage between males and females, and the primary function of mitochondria is in energy production. Follow-up studies including both genders are needed to generalize our findings.

Our results also suggested that maintenance of mtDNA copy number and managing the expansion of mtDNA heteroplasmic mutations could help improve health status, especially in the elderly. Future studies are required to identify genetic, behavioral and environmental factors that can prevent or accelerate age-related changes in mtDNA quality and quantity.

**Project 2**: mtDNA heteroplasmy concordance between DNA and RNA, and the effect of heteroplasmy dynamics on gene expression.

In this project, we first conducted data mining on paired DNA and RNA sequencing data of cell lines from the 1000 genomes project. Our results suggested that most heteroplasmies presenting in mtDNA can be transcribed to mtRNA, and their frequencies stayed consistent between DNA and RNA. This is one of the first studies to show the concordance of heteroplasmies, and it also showed that the RNA sequencing could be used as a data source to investigate mtDNA heteroplasmy. Meanwhile, we suspected that the heteroplasmies with big frequency difference between DNA and RNA could be artifacts caused by heteroplasmy dynamics. We then experimentally tested this hypothesis. Our data indicated that the heteroplasmy

frequencies could change very quickly during the cell culture process, and this may have substantial effects on cellular functions. Therefore, heteroplasmy dynamics could introduce some confounding factors in cell line based studies.

In this study, we used gene expression profiles to study the disturbance of heteroplasmy dynamics. We are now collaborating with Dr. Nakahira from Weill Cornell Medical School to replicate this experiment and perform more measurements to assess mitochondrial functions under different heteroplasmy status, such as reactive oxygen species (ROS) levels, mitochondrial membrane potential and cell viability. In addition, the heteroplasmies we investigated here were not reported in disease patients from previous studies. The dynamics of confirmed pathogenic mtDNA heteroplasmies should be investigated in future studies.

**Project 3**: Very Short Mitochondrial DNA Fragments and Heteroplasmy in Human Plasma

In this project, we developed a method to sequence mtDNA from plasma, integrating an improved experimental procedure and computational workflow. This study provided the first experimental evidence that mtDNA heteroplasmy was detectable in plasma cell-free DNA, and therefore had the potential to serve as a biomarker for disease diagnosis and monitoring, which could have promising applications in several fields such as cancer and transplantation medicine.

Although we successfully identified mtDNA heteroplasmies from the plasma sample, the sequencing cost for the current approach is still very high, making it infeasible for large scale studies. Follow-up optimization, including development of novel targeted sequencing methods, is necessary for future studies. Case control studies with large

sample sizes are also needed to locate the potential marker mtDNA mutations under different medical conditions.

## *APPENDIX A: Publication Inclusion Authorizations*

*Impacts of Aging on Mitochondrial DNA Quantity and Quality in Humans*

A version of this manuscript was accepted for publication by the journal *BMC Genomics* in November 2017 with following authorship: Ruoyu Zhang, Yiqin Wang, Kaixiong Ye, Martin Picard and Zhenglong Gu

The *BMC Genomics* Authors' Statement and Copyright Release Form authorizes the inclusion of the manuscript in this dissertation

E-mail correspondence with this journal shown below also confirms authorization of the inclusion of the manuscripts in this dissertation:

Gmail Ruoyu Zhang <zry0510@gmail.com>

**00861581 RE: Request written permission to include my publication...**

**"Joel Lagmay" <joel.lagmay@springernature.com>** <joel.lagmay@springernature.com> Fri, Apr 20, 2018 at 10:13 PM
To: "rz253@cornell.edu" <rz253@cornell.edu>

SPRINGER NATURE

Dear Ruoyu,

Thank you for contacting Springer Nature.

The open access articles published in BioMed Central's journals are made available under the Creative Commons Attribution (CC-BY) license, which means they are accessible online without any restrictions and can be re-used in any way, subject only to proper attribution (which, in an academic context, usually means citation).

The re-use rights enshrined in our license agreement (http://www.biomedcentral.com/about/policies/license-agreement) include the right for anyone to produce printed copies themselves, without formal permission or payment of permission fees. As a courtesy, however, anyone wishing to reproduce large quantities of an open access article (250+) should inform the copyright holder and we suggest a contribution in support of open access publication (see suggested contributions at http://www.biomedcentral.com/about/policies/reprints-and-permissions/suggested-contributions).

Please note that the following journals have published a small number of articles that, while freely accessible, are not open access as outlined above: Alzheimer's Research & Therapy, Arthritis Research & Therapy, Breast Cancer Research, Critical Care, Genome Biology, Genome Medicine, Stem Cell Research & Therapy.
You will be able to find details about these articles at http://www.biomedcentral.com/about/policies/reprints-and-permissions

If you have any questions, please do not hesitate to contact me.

With kind regards,
---
**Joel Lagmay**
Global Open Research Support Executive
Global Open Research Support

*Very Short Mitochondrial DNA Fragments and Heteroplasmy in Human Plasma*

A version of this manuscript was accepted for publication by the journal Scientific Reports in November 2016 with following authorship: Ruoyu Zhang, Kiichi Nakahira, Xiaoxian Guo Augustine M.K. Choi and Zhenglong Gu

The journal *Scientific Reports* authorizes the inclusion of the manuscript in this dissertation:


SPRINGER NATURE

**Title:** Very Short Mitochondrial DNA Fragments and Heteroplasmy in Human Plasma

**Author:** Ruoyu Zhang, Kiichi Nakahira, Xiaoxian Guo, Augustine M.K. Choi, Zhenglong Gu

**Publication:** Scientific Reports

**Publisher:** Springer Nature

**Date:** Nov 4, 2016



E-mail correspondence with this journal shown below also confirms authorization of the inclusion of the manuscripts in this dissertation:

Gmail **Ruoyu Zhang <zry0510@gmail.com>**

## Case #00496341 - Request written permission to include my publication in my PhD's thesis [ ref:_00D30oeGz._5000c1YEj6o:ref ]

**customercare@copyright.com** <customercare@copyright.com> Sun, Apr 22, 2018 at 8:12 PM
To: "rz253@cornell.edu" <rz253@cornell.edu>

Dear Ruoyu Zhang,

Good day! Welcome to the RightsLink service of Copyright Clearance Center.

Please be advised that "**Very Short Mitochondrial DNA Fragments and Heteroplasmy in Human Plasma**" is an Open Access article licensed under a **Creative Commons Attribution 4.0 International License**. This allows you to reuse the article without having to obtain a formal permission provided that the original source of publication is properly acknowledged. To know more about the terms of this license, please refer to the link below.

https://creativecommons.org/licenses/by/4.0/

If you have any further questions please don't hesitate to contact a Customer Account Specialist at 855-239-3415 Monday-Friday, 24 hours/day.

Best regards,

Rose Cabudoc
Customer Account Specialist
Copyright Clearance Center
222 Rosewood Drive
Danvers, MA 01923
www.copyright.com
+1.855.239.3415
Facebook - Twitter - LinkedIn
ref:_00D30oeGz._5000c1YEj6o:ref
---------- Original Message ----------
**From:** Ruoyu Zhang [rz253@cornell.edu]
**Sent:** 4/20/2018 2:41 PM
**To:** customercare@copyright.com
**Subject:** Request written permission to include my publication in my PhD's thesis

Dear editors,

I am a phd student from Cornell University. I am now writing my phd thesis. I previously published a manuscript with Scientific Reports
(Zhang, R., et al., Very Short Mitochondrial DNA Fragments and Heteroplasmy in Human Plasma. Scientific Reports, 2016. 6: p. 36097.
https://www.nature.com/articles/srep36097). Could I request the permission to include a version of this manuscript in my thesis?

Thanks!
Ruoyu

ref:_00D30oeGz._5000c1YEj6o:ref