Mitochondrial Genome Variation in Healthy Aging

by

Daniel John Fornika

Bachelor of Science, Simon Fraser University, 2006

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Science

in

THE FACULTY OF GRADUATE STUDIES
(Medical Genetics)

The University Of British Columbia
(Vancouver)

August 2012

Abstract

Mitochondria are thought to play a role in the aging process through their production of reactive oxygen species (ROS), and their regulation of cell fate via senescence and apoptosis. We hypothesize that genetic variation in the mitochondrial genome may explain a portion of the phenotypic variance in the development of long-term good health. To test this hypothesis, we have performed genetic association tests on a set of common mitochondrial polymorphisms, in a study of 419 exceptionally healthy seniors (cases) and 415 population-based mid-life individuals (controls).

Variant discovery was performed using Sanger sequencing of 834 individuals for the 1.1 kb non-coding mitochondrial control region, and identified 277 SNPs present in at least one individual. A set of 92 mitochondrial coding-region SNPs were chosen via pooled high-throughput sequencing, combined with a previously-published set of European-specific mitochondrial tag SNPs.

After filtering for minor-allele frequency of $> 10\%$, a set of nine control-region SNPs and seven coding-region SNPs were tested for association with healthy aging. None showed a statistically-significant association signal. Additionally, one control-region variant that had shown association in an Italian centenarian population was tested in our sample set, but the association was not replicated.

Preface

Funding for these experiments was obtained from the Canadian Institute for Health Research (CIHR) by Angela Brooks-Wilson, in collaboration with the Marco Marra, Joseph Connors, Stephen Jones, Nhu Le and Graydon Meneilly. The experiments described in this thesis are a part of the Genomics, Genetics and Gerontology (G^3) Study of Healthy Aging.

All experiments were concieved of and designed by Angela Brooks-Wilson. This thesis is broadly composed of three sub-experiments. They are: pooled next-generation sequencing of mitochondrial DNA, Sanger sequencing of individual mitochondrial control-regions, and determination of mitochondrial genotypes by Sequenom genotyping.

Dan Fornika performed the Polymerase Chain Reaction (PCR) reactions necessary to provide template mitochondrial DNA for both pooled mitochondrial DNA sequencing and control-region Sanger sequencing. Mitochondrial DNA pools were constructed by Dan Fornika.

Sanger sequencing was prepared by Julius Halaschek-Wiener with assistance from Dan Fornika. Sanger sequencing was performed by the sequencing group of the British Columbia Genome Sciences Centre (BCGSC).

Library construction and Illumina Genome Analyzer (GA) sequencing was performed by the sequencing group of the BCGSC.

Sequenom genotyping was performed by the McGill University/Genome Québec Innovation Centre.

All bioinformatics and statistical analyses were performed by Dan Fornika, except for quality control of Sequenom genotype data, which was performed by Denise Daley.

Table of Contents

List of Tables

List of Figures

Glossary

BCGSC British Columbia Genome Sciences Centre

CIHR the Canadian Institute for Health Research

GA Genome Analyzer

kb kilobase

FMNH Flavin Mononucleotide

MAF Minor Allele Frequency

MELAS Mitochondrial Encephalopathy Lactic Acidosis and Stroke-like episodes

MERRF Myoclonic Epilepsy and Ragged-Red Fibres

mtDNA Mitochondrial DNA

NADH Nicotinamide Adenine Dinucleotide

PCR Polymerase Chain Reaction

rCRS revised Cambridge Reference Sequence

ROS Reactive Oxygen Species

rRNA Ribosomal RNA

tRNA Transfer RNA

OR Odds Ratio

SNP Single Nucleotide Polymorphism

Chapter 1

Introduction

1.1 Aging as a Genetic Disease

1.1.1 The “Healthy Aging” Phenotype

Our goal is to study biological mechanisms of aging by identifying genetic variants that are associated with healthy aging. This study focuses on individuals who have reached the upper end of the normal human lifespan in good health, as opposed to other longevity-based studies that focus on centenarians who may not be exceptionally healthy[1, 2, 3].

This project has been carried out using samples and phenotype data from the Genomics, Genetics and Gerontology (G^3) Study of Healthy Aging. In this study, cases are defined as having a “healthy aging” phenotype if they reached the age of 85 years without being diagnosed with cancer, (excluding non-melanoma skin cancer) cardiovascular disease, major pulmonary disease (excluding asthma), Alzheimer disease or diabetes. They have been further characterized by means of the Mini Mental State Examination for determination of moderate to severe cognitive impairment[4], the Timed Up and Go test of basic mobility skills[5] , the Geriatric Depression Scale[6] and the Instrumental Activities of Daily Living Scale[7].

Controls are between the ages of 40 and 54 years, and were not recruited with respect to health status. As such, they are representative of the general population with respect to their probability of reaching the age of 85 years without acquiring one of the five common age-related diseases listed above. Ideally, our controls would be a random sample of the population at the time that our cases were in mid-life, and we believe that these controls are a good proxy for that ideal sample. Specifically, we believe that the allele frequencies of our control sample should be a good approximation of the allele frequencies of the (now largely deceased) population that our cases originated from.

1.2 Mitochondria and Aging

1.2.1 Mitochondrial Genome Structure and Regulation

Human mitochondria have a 16.5 kb circular genome, which encodes 13 protein-coding genes, (See table 1.2) 22 Transfer RNA (tRNA)s, and two Ribosomal RNA (rRNA)s (see figure 1.1). The protein-coding genes encode subunits of the mitochondrial electron transport chain complexes I, III and IV, and two subunits of ATP synthase. In contrast with the nuclear genome, there is very little non-coding sequence in the human mitochondrial genome. The majority of the non-coding sequence is contained within the 1.1 kb control region, where three known promoters coordinate expression of the entire mitochondrial chromosome. Outside of the control region, the mitochondrial genes are tightly spaced, with clusters of tRNA genes located between protein-coding genes.

Table 1.1: Selected Mitochondrial Diseases

OMIM ID	Name	rCRS Positions Mutated	Symptoms
535000	LHON	11,778, 3,460, 14484	Blindness
540000	MELAS	3,243,	Myopathy, Lactic acidosis
220110	Complex IV Deficiency	(various mutations in MT-CO1-3)	Myopathy
256000	Leigh Syndrome	4,681	CNS Lesions
545000	MERRF	8,344	Seizures, myopathy
530000	Kearns-Sayre Syndrome	(various deletions)	Blindness, cardiomyopathy
157640	CPEO	(various deletions)	Eye turn, hypogonadism

Hundreds of additional mitochondrial proteins are encoded by the nuclear genome, and coordinated control of the two genomes is required for normal mitochondrial function[8, 9]. Mitochondrial gene expression is controlled by transcription factors (TFAM, TFB1M, TFB2M) and an RNA polymerase (POLRMT) that are

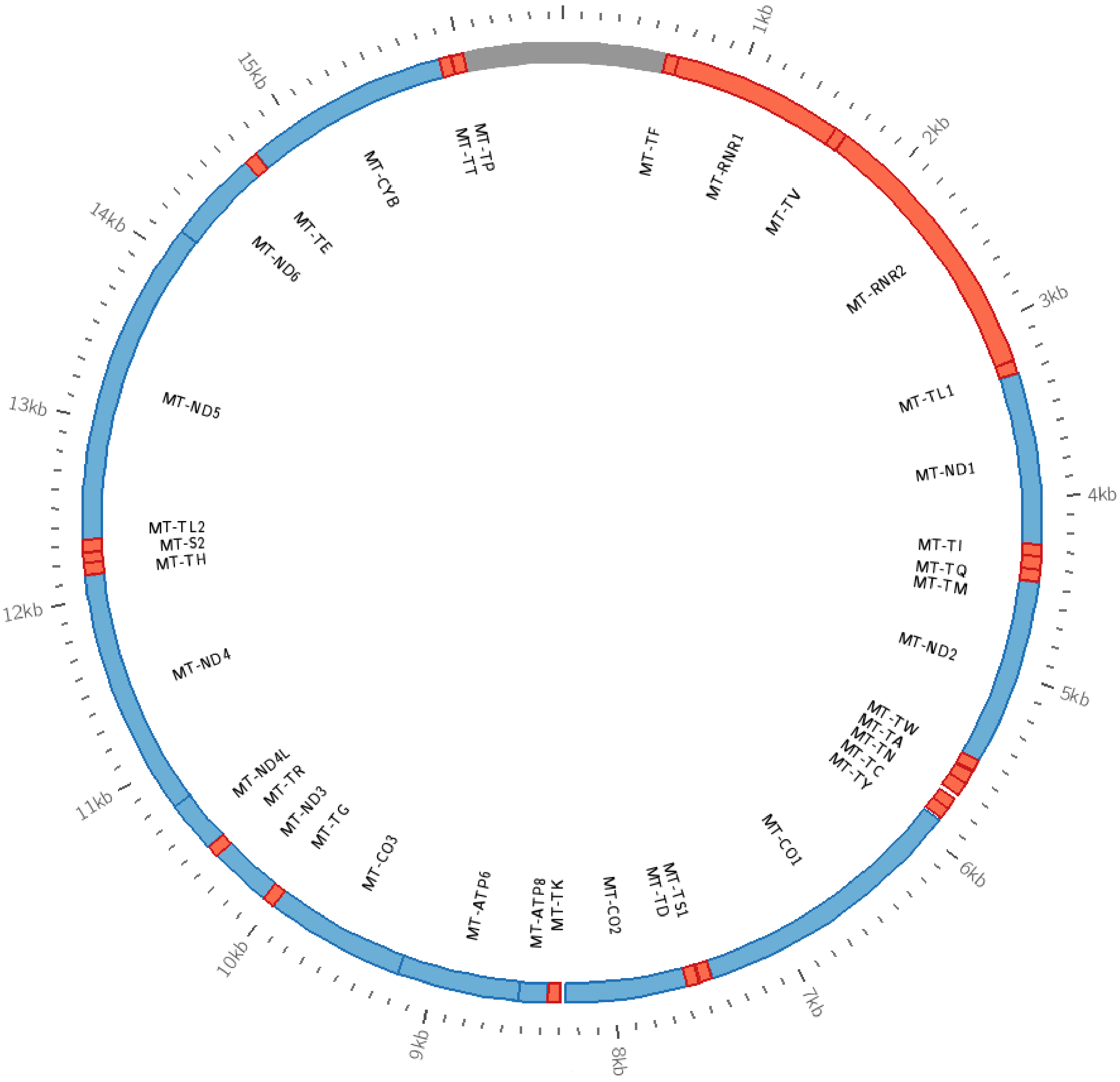


Figure 1.1: Map of the Human Mitochondrial Genome. Non-coding control region (position 16,024-576) is shown in grey. Protein-coding genes are shown in blue, while RNA-coding genes are shown in red. All gene labels are from the HUGO Gene Nomenclature Committee (www.genenames.org)

encoded on the nuclear genome. A transcriptional regulatory network links a master regulator, PGC-1α (also known as PPARGC1A) to the mitochondrial genome.

Table 1.2: Mitochondrial Protein Genes

Gene	Uniprot Acession	ETC Complex	rCRS Position
MT-ND1	P03886	Complex I	3,307–4,262
MT-ND2	P03891	Complex I	4,470–5,511
MT-COX1	P00395	Complex IV	5,904–7,445
MT-COX2	P00403	Complex IV	7,586–8,269
MT-ATP8	P03928	Complex V	8,366–8,572
MT-ATP6	P00846	Complex V	8,527–9,207
MT-COX3	P00414	Complex IV	9,207–9,990
MT-ND3	P03897	Complex I	10,059–10,404
MT-ND4L	P03901	Complex I	10,470–10,766
MT-ND4	P03905	Complex I	10,760–12,137
MT-ND5	P03915	Complex I	12,337–14,148
MT-ND6	P03923	Complex I	14,149–14,673
MT-CYB	P00156	Complex III	14,747–15,887

The mitochondrial genome also contains a 1.1 kb control region (position 16024-576 on GenBank NC_012920) which includes promoters for both the heavy and light strands, and the heavy strand origin of replication. The control region also contains numerous transcription factor binding sites. There are three hyper-variable sequences (HVS1, HVS2 and HVS3) within the control region that contain a relatively high density of polymorphisms, in comparison to the rest of the mitochondrial genome[10].

The human mitochondrial genome is inherited exclusively from the mother. Paternal mitochondria are selectively degraded after fertilization, by ubiquitin-mediated proteasomal degradation[11, 12]. There is no conclusive evidence for recombination in human Mitochondrial DNA (mtDNA)[13]. An extensive map of the geographic distribution of mitochondrial haplogroups in human populations has been recorded. Together with geographic and genotype data from the non-recombining portion of the Y chromosome, this information has helped to trace early human migration out of Africa and across the globe[14].

The mitochondrial genome is also highly polymorphic in all human populations. A previous study of European mitochondrial genome diversity identified 144 single nucleotide polymorphisms present in $> 1\%$ of a sample of 928 publicly available European mitochondrial genome sequences [15].

Numerous mitochondrial genetic diseases have been identified[16, 17]. Several of these diseases, with their characteristic mutations are listed in table 1.1. Symptoms vary widely, and include blindness, deafness, diabetes and ataxia.

1.2.2 Reactive Oxygen Species

Mitochondria are thought to contribute to the aging process through the production of Reactive Oxygen Species (ROS), as a byproduct of oxidative phosphorylation[18, 19]. Prolonged exposure to intracellular ROS can cause damage to protein and lipids, and can cause somatic mutations in both the nuclear and mitochondrial genomes.

During oxidative phosphorylation, electrons are passed from reduced Nicotinamide Adenine Dinucleotide (NADH) and Flavin Mononucleotide (FMNH) to a group of mitochondrial inner membrane-bound enzymes that comprise the electron transport chain. Electrons are passed down the chain in a series of redox reactions, releasing energy that is used to pump protons into the intermembrane space. These reactions maintain the mitochondrial elechemical gradient that drives the production of ATP. The majority of electrons passing through the electron transport chain will finally be combined with H^+ and $\frac{1}{2}O_2$ to form H_2O, but a small percentage will form side-reactions that result in the production of highly unstable superoxide radicals, $O_2^{-\cdot}$. Superoxide quickly reacts with H_2O to form hydrogen peroxide, (H_2O_2) itself a strong oxidizing agent. Although small amounts of ROS are a normal byproduct of cellular metabolism, the accumulated effects of these reactions can degrade tissue, cause somatic mutations and lead to cellular senescence[20, 21].

1.2.3 The Role of Mitochondria in Apoptosis

Mitochondria integrate several intracellular signals including DNA damage response and pro-survival signals, as well as metabolic signals such as the ADP/ATP ratio and intracellular Ca^{2+} concentrations. Under high cellular stress conditions, these signals can initiate cell death via the intrinsic apoptotic pathway. The pro-apoptotic proteins BAX and BAK are recruited to the mitochondrial membrane, resulting in increased membrane permeability and release of Cytochrome-c and SMAC/DIABLO from the mitochondrial intermembrane space into the cytosol. The release of Cytochrome-c and SMAC/DIABLO leads to the activation of effector caspases that initiate the process of apoptosis.

Apoptosis is a key protective mechanism against cancer. When a cell acquires mutations or DNA damage that may lead to escape from the cell cycle and uncontrolled cell division, the apoptotic pathway can be activated to prevent the development of a malignancy. Model organisms such as p53 knockout mice fail to activate the intrinsic apoptotic pathway in response to DNA damage and develop malignancies at a much higher rate than wild-type mice[22]. There is also evidence that variation in the mitochondrial genome itself can alter the probability that a cell will undergo apoptosis. Studies of a lymphoblastoid cell line showed that a A4263G mutation in the mitochondrial isoleucine tRNA could alter mitochondrial membrane potential and lead to an increased rate of apoptosis[23]. Some have argued that many of the phenotypic hallmarks of aging (muscle loss, wrinkled skin, functional decline of internal organs) are due to the accumulated effects of apoptosis and senescence[24]. They hypothesize that successful aging, (defined as reaching the age of 85 without being diagnosed with cancer, cardiovascular disease, diabetes, major pulmonary disease, or Alzheimer disease.)[25] requires a fine balance between cancer surveillance by apoptosis and a maintenance of healthy pre-senescent tissue[26].

1.2.4 The Role of Mitochondria in Cellular Senescence

Several lines of evidence indicate that mitochondria play a role in induction of cellular senescence. Senescent cells are characterized by growth arrest in the G1 phase of the cell cycle, accumulation of H2A.X foci and increased p53 activity indicative of DNA damage, and decreased telomere length[27]. The telomerase reverse transcriptase hTERT is translocated to mitochondria in response to oxidative stress, where it increases the rate of mtDNA damage and promotes apoptosis[28]. This relationship between telomere maintenance and mtDNA maintenance is a recent discovery, and is not yet completely understood[29]. Cells grown in high oxygen concentrations become senescent at an increased rate, and senescence can be delayed by addition of antioxidants or mild uncoupling agents to the growth medium[30].

1.2.5 Somatic Mitochondrial DNA Mutations and Aging

Mutations in the mitochondrial genome accumulate with age in somatic tissues. Mitochondrial DNA mutations have been observed to correlate with age in tissues such as heart muscle,[31] brain,[32] and skeletal muscle[33]. In addition to point mutations, accumulation of ROS-damaged deoxyguanosine in the form of 8-Hydroxy-deoxyguanosine has been observed.

1.2.6 Mitochondrial Heteroplasmy and Tissue Heterogeneity

The number of mitochondria per cell varies from zero in red blood cells to several hundred in skeletal muscle cells, and each mitochondrion contains several copies of the mitochondrial genome. Mutations can arise in somatic cells because of oxidative damage or replication errors by DNA polymerase-γ, and can be propagated to daughter cells after division. Since each cell contains many copies of the mitochondrial genome, there may be a combination of mtDNA alleles in a particular cell or tissue[34, 35]. This phenomenon is known as heteroplasmy. Several mitochondrial diseases, such as Myoclonic Epilepsy and Ragged-Red Fibres (MERRF) or Mitochondrial Encephalopathy Lactic Acidosis and Stroke-like episodes (MELAS), do not present physiological symptoms unless the causative mutation accumulates beyond a certain threshold level, sufficient to disrupt normal mitochondrial function[36].

Although the accumulation of somatic mtDNA mutations is suspected to play a role in the aging process, our study is designed to detect heritable genetic factors that influence long-term good health. Mutations that arise in skeletal muscle, epithelium, neurons and other somatic tissues are not passed on in the germline. Only mutations that arise in the ova (or pre-oval germ cell lineage) can be passed on to the next generation.

1.2.7 Reported Associations of Mitochondrial Genome Variants with Longevity

Several longevity-associated mtDNA variants have been reported in populations around the world. A control region polymorphism at position 150 was associated with longevity in the Italian population and has been hypothesized to cause a re-organization of an origin of replication on the mtDNA[3]. The comparison of 52 centenarians (age range 99-106 years) and 117 controls (age range 18-98 years) showed a statistically significant difference (Odds Ratio (OR) $= 5.09$, $P = 0.0035$, Fisher's exact test) in the frequency of homoplasmic C150T transition in leukocytes. Furthermore, the researchers noted that the abundance of heteroplasmic C150T mutation in fibroblasts was correlated with age.

The association signal at control region position 150 has been replicated in both the Japanese and Finnish populations[37]. In Finns, a comparison of 46 seniors (age 90 or 91 years) and 57 middle-aged controls showed a significant association (OR$= 1.50$, $P = 0.037$, χ^2 test) of the 150T allele with longevity. A similar result was found in a smaller Japanese sample set of 19 seniors and 9 controls (OR$= 1.41$, $P = 0.032$, χ^2 test).

A polymorphism in the MT-ND2 gene at position 5,178 of the coding region was found to be associated with longevity in the Japanese population[38]. The study investigated the relative frequencies of the 5178A and 5178C alleles, and found that the 5178A allele in 9 of 11 centenarians, versus 12 of 43 controls. This same polymorphism was also associated with glucose tolerance in Japanese men, and may contribute to resistance to type II diabetes. The MT-ND2 gene encodes a subunit of NADH dehydrogenase, complex I of the mitochondrial electron transport chain.

1.3 Hypothesis and Specific Aims

We hypothesize that healthy aging is influenced by sequence variation in the mitochondrial genome. Therefore, one or more common mitochondrial alleles will be associated with healthy aging.

The specific aims of this study are as follows:

1. Survey the mitochondrial genomes of cases and controls for sequence variants.

2. Determine whether common variation in the mitochondrial genome is associated with healthy aging in our study population.

Chapter 2

Variant Detection

2.1 Introduction

The mitochondrial genome is amenable to targeted resequencing by second-generation technologies. Its relatively small size (16.5 kilobase (kb)) and low repetitive sequence content make it much more likely that a unique sequence alignment can be determined for the short reads generated by current second-generation sequencing systems.

In order to identify the extent of mitochondrial genome variation in our sample set, we used two sequencing technologies for two distinct segments of the mitochondrial genome. For the non-coding control region (position 16,024-576), we performed bi-directional Sanger sequencing on all 419 cases and 415 controls. A pooled Illumina Genome Analyzer (GA) sequencing strategy was used to identify variants in the entire mitochondrial genome. See figure 2.1 for an outline of the variant detection strategy.

2.2 Methods

This study was approved by the joint Clinical Research Ethics Board of the British Columbia Cancer Agency and the University of British Columbia. All subjects gave written informed consent.

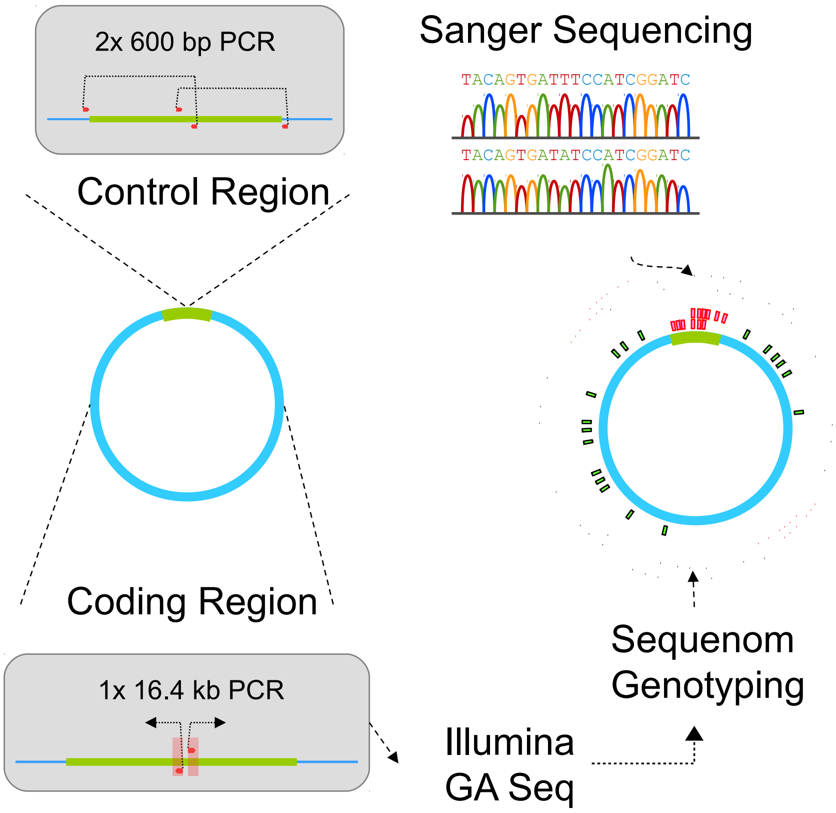


Figure 2.1: Experimental Design for Variant Detection. Two sequencing technologies were used to identify mitochondrial genome variation. The control-region was PCR-amplified in two segments and sequenced by Sanger sequencing in individuals. The entire mitochondrial genome was PCR-amplified, pooled, and sequenced on the Illumina Genome Analyzer. Coding region variants were carried forward to Sequenom genotyping in individuals.

2.2.1 Subjects and Samples

The subjects of this study were 419 healthy elderly individuals (cases) and 415 mid-life controls. Cases were $>$ 85 years old at the time of recruitment, and had not been diagnosed with cancer, cardiovascular disease, Alzheimer disease or diabetes. Controls were 40-50 years old at recruitment, and were ascertained without regard to health status. All participants are of European descent, based on subject-reported ethnicity of their four grandparents. Total DNA was extracted from peripheral blood leukocytes using the Gentra Puregene Blood Kit (Qiagen), according to the manufacturer's protocol.

2.2.2 Control Region PCR and Sanger Sequencing

PCR primers were designed not to overlap with common polymorphic loci. In-silico PCR was performed using web service based at Kyushu University, to ensure that no nuclear DNA segments would be co-amplified[39]. The mitochondrial control region was PCR-amplified with Platinum Pfx polymerase (Invitrogen). PCR reactions were performed in 20 μL total volume containing: 20 ng template genomic DNA, 10 μM each of forward primer (MAP001_F or MAP002.1_F) and reverse primer (MAP001_R or MAP002.1_R) (Table 2.1), 0.4 U Platinum Pfx enzyme, 10 mM each dNTPs, and 1x Phusion Buffer GC. Forward and reverse primers incorporated the -21M13F (TGTAAAACGACGGCCAGT) and M13R (CAGGAAACAGCTATGAC) extensions, respectively, at their 5' ends. Sequencing reactions were carried out as described previously [40].

2.2.3 Sanger Sequence Assembly

Sanger sequence traces were aligned to the revised Cambridge Reference Sequence (rCRS) reference sequence (GenBank accession NC_012920) with the Phred/Phrap/-Consed suite, version 20.0 [41, 42, 43]. Polymorphisms were first detected automatically using Polyphred version 6.18. To minimize false-positives all non-reference alleles were manually confirmed by visual inspection of chromatograms by two people.

2.2.4 Long PCR

The mitochondrial genome was amplified using long-PCR with Phusion polymerase (Finnzymes). PCR reactions were performed in 20 μL total volume containing: 20 ng template genomic DNA (2 ng/μL), 10 μM each of forward primer MAP011.1_F and reverse primer MAP011.1_R (Table 2.1), 0.4 U Phusion enzyme, 10mM each dNTPs, and 1x Phusion Buffer GC. The thermocycler program was: 1.) initial melt at 98°C for 30 seconds, 2.) melt at 98°C for 10 seconds, 3.) anneal/extend at 72°C for 8 minutes, 15 seconds 4.) repeat steps 2 and 3, 29 times 5.) final extension at 72°C for 10 minutes.

Table 2.1: PCR primers used for Sanger sequencing and long-PCR.

Primer ID	T_m (°C)	Sequence	rCRS Position
MAP011.1-F	66.3	GGGAGCTCTCCATGCATTTGG	34-54
MAP011.1-R	64.7	AGACCTGTGATCCATCGTGATGTC	16,558-12
MAP001-F	57.1	(-21M13-Fwd[a])GAAAAAGTCTTTAACTCCACCATT	15,961-15,984
MAP001-R	58.9	(M13-Rev[b])TACTGCGACATAGGGTGCTC	107-126
MAP002.1-F	59.3	(-21M13-Fwd)GAGCTCTCCATGCATTTGG	36-54
MAP002.1-R	57.3	(M13-Rev)AGGGTGAACTCACTGGAACG	707-726

[a] '-21M13-Fwd' = TGTAAAACGACGGCCAGT

[b] 'M13-Rev' = CAGGAAACAGCTATGAC

2.2.5 Construction of DNA Pools

DNA products from long-PCR were quantitated with Quant-iT™PicoGreen® reagent (Invitrogen). Two DNA pools were constructed. One pool consisted of 10 ng mtDNA from each of 419 case samples, and the other consisted of 10 ng mtDNA from each of 415 control samples. DNA was concentrated by speed-vac.

2.2.6 Library Construction and Sequencing

Library construction and DNA sequencing was carried out by the sequencing platform of the BC Genome Sciences Centre. Pooled mtDNA was sheared using sonication and size-separated using electrophoresis. The ~ 300-bp fraction was isolated for library construction using the Illumina Genome Analyzer single-end library protocol (Illumina). Sequencing was performed on an Illumina GA using two lanes of a flow cell per pool, generating 36-bp reads.

Table 2.2: Summary of Illumina Sequence Mapping (Untrimmed Reads)

Chr.	Length	Reads Mapped, Cases		Reads Mapped, Controls	
1	249,250,621	512,872	(4.30%)	539,464	(4.23%)
2	243,199,373	81,462	(0.68%)	85,478	(0.67%)
3	198,022,430	74,357	(0.62%)	77,525	(0.61%)
4	191,154,276	48,754	(0.41%)	53,598	(0.42%)
5	180,915,260	206,336	(1.73%)	225,336	(1.77%)
6	171,115,067	32,858	(0.28%)	35,324	(0.28%)
7	159,138,663	170,779	(1.43%)	227,925	(1.79%)
8	146,364,022	35,068	(0.29%)	38,300	(0.30%)
9	141,213,431	27,073	(0.23%)	26,942	(0.21%)
10	135,534,747	25,878	(0.22%)	25,809	(0.20%)
11	135,006,516	97,174	(0.81%)	98,442	(0.77%)
12	133,851,895	28,901	(0.24%)	30,747	(0.24%)
13	115,169,878	35,384	(0.30%)	36,679	(0.29%)
14	107,349,540	23,239	(0.19%)	24,482	(0.19%)
15	102,531,392	13,155	(0.11%)	14,121	(0.11%)
16	90,354,753	13,580	(0.11%)	13,829	(0.11%)
17	81,195,210	311,656	(2.61%)	299,374	(2.35%)
18	78,077,248	19,447	(0.16%)	20,532	(0.16%)
19	59,128,983	7,709	(0.06%)	7,346	(0.06%)
20	63,025,520	11,142	(0.09%)	11,389	(0.09%)
21	48,129,895	13,807	(0.12%)	14,062	(0.11%)
22	51,304,566	5,967	(0.05%)	5,941	(0.05%)
X	155,270,560	54,631	(0.46%)	62,053	(0.49%)
Y	59,373,566	11,447	(0.10%)	11,022	(0.09%)
MT	16,569	4,286,809	(35.94%)	4,681,659	(36.68%)
other	6,110,758	5,068	(0.04%)	5,087	(0.04%)
total mapped	-	6,154,553	(51.59%)	6,672,466	(52.27%)
total unmapped	-	5,774,653	(48.41%)	6,092,354	(47.73%)
grand total	-	11,929,206	(100.00%)	12,764,820	(100.00%)

2.2.7 Statistical Analysis

In order to assess the effect of read trimming on mapping, alignments were done with both full 36-base reads and trimmed reads. For trimmed reads, the BWA read-trimming parameter (q=25) was used. Short sequence reads were aligned to the GRCh37 (hg19) reference using the BWA sequence alignment program, version 0.6.1-r104[44]. Aside from the read-trimming parameter, all reads were mapped using default BWA parameters.

Per-base quality scores for both untrimmed and trimmed reads were calculated with FastQC software[45] (See figures 2.4, 2.5)

SNPs were detected by analyzing BWA ‘pileup’ output files with a custom perl

Table 2.3: Summary of Illumina Sequence Mapping (Trimmed Reads)

Chr.	Length	Reads Mapped, Cases		Reads Mapped, Controls	
1	249,250,621	496,262	(6.92%)	508,839	(6.92%)
2	243,199,373	95,787	(1.34%)	94,431	(1.28%)
3	198,022,430	86,431	(1.20%)	82,319	(1.12%)
4	191,154,276	64,637	(0.90%)	65,229	(0.89%)
5	180,915,260	263,349	(3.67%)	270,725	(3.68%)
6	171,115,067	38,931	(0.54%)	37,264	(0.51%)
7	159,138,663	112,712	(1.57%)	107,757	(1.47%)
8	146,364,022	43,377	(0.60%)	42,340	(0.58%)
9	141,213,431	35,488	(0.49%)	32,654	(0.44%)
10	135,534,747	32,305	(0.45%)	29,541	(0.40%)
11	135,006,516	101,180	(1.41%)	96,423	(1.31%)
12	133,851,895	35,422	(0.49%)	35,238	(0.48%)
13	115,169,878	36,647	(0.51%)	36,389	(0.50%)
14	107,349,540	29,971	(0.42%)	28,723	(0.39%)
15	102,531,392	16,489	(0.23%)	15,896	(0.22%)
16	90,354,753	17,579	(0.25%)	16,484	(0.22%)
17	81,195,210	392,551	(5.47%)	340,912	(4.64%)
18	78,077,248	22,205	(0.31%)	21,735	(0.30%)
19	59,128,983	8,908	(0.12%)	7,365	(0.10%)
20	63,025,520	14,805	(0.21%)	13,690	(0.19%)
21	48,129,895	16,454	(0.23%)	14,409	(0.20%)
22	51,304,566	7,768	(0.11%)	6,995	(0.10%)
X	155,270,560	66,354	(0.93%)	64,348	(0.88%)
Y	59,373,566	11,477	(0.16%)	10,403	(0.14%)
MT	16,569	3,750,715	(52.29%)	4,015,230	(54.64%)
other	6,110,758	5,121	(0.07%)	4,722	(0.06%)
total mapped	-	5,802,925	(80.90%)	6,000,061	(81.64%)
total unmapped	-	1,370,272	(19.10%)	1,348,969	(18.36%)
grand total	-	7,173,197	(100.00%)	7,349,030	(100.00%)

script. At each position, the numbers of reference and non-reference bases were counted. Only those bases with phred-scaled quality scores of 40 were included for SNP detection.

2.3 Results

2.3.1 Sequencing of the Mitochondrial Conrol Region

The highly polymorphic mitochondrial control region rCRS (position 16024-576) was sequenced using bi-directional Sanger sequencing. We discovered 277 SNPs in the control region that were present in at least one sample.

2.3.2 Next-Generation Sequencing of Pooled mtDNA

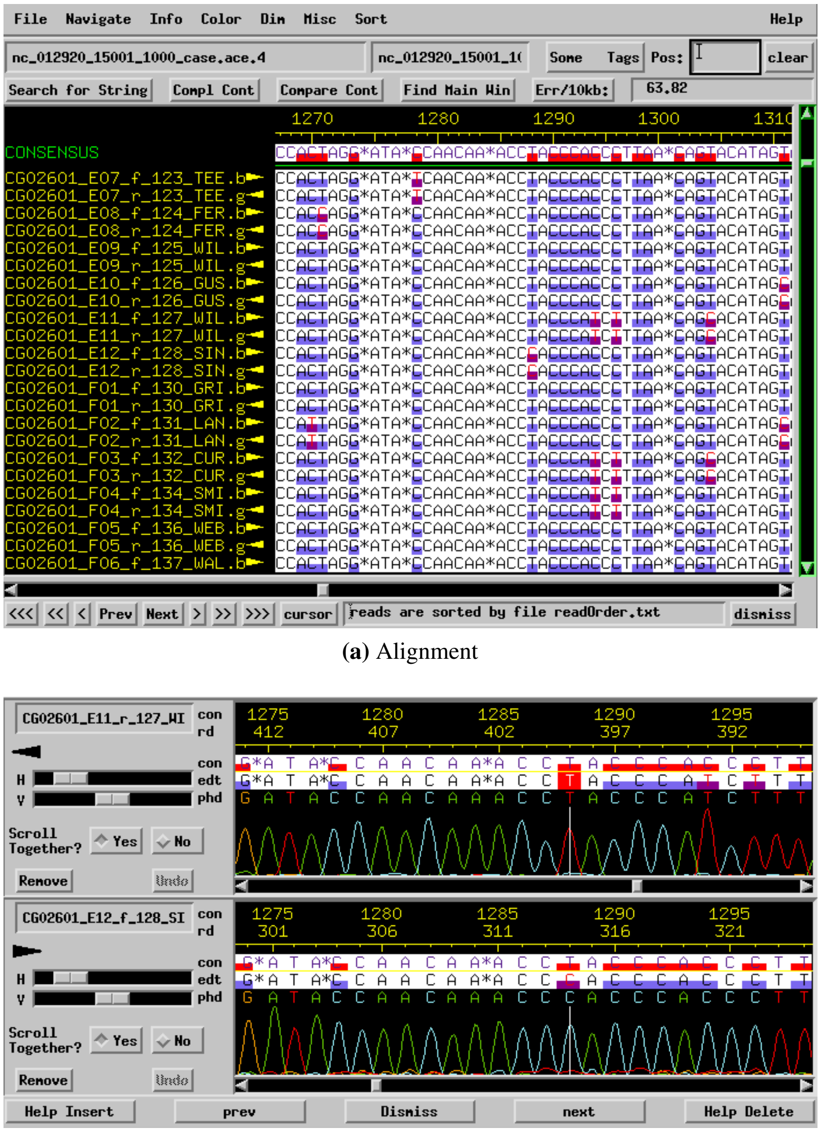
The median depth of sequence coverage was 31,134 reads for the case pool and 12,683 reads for the control pool. This represents approximately 30x coverage for each sample that is included in the pool. To reduce the number of false-positive variant calls that are due to sequencing errors, we only considered high-quality bases (base-quality score > 35 and mapping-qualty score > 20) for SNP-calling. We identified 90 SNPs in the case pool and 113 SNPs in the control pool with MAF $> 1\%$. 84 of these SNPs are common to both pools, with 6 SNPs only being observed in the case pool and 29 SNPs only being observed in the control pool. (see figs 2.8 and 2.9).

Comparison of minor allele frequencies for control region SNPs in Sanger and Illumina GA datasets is shown in Figures 2.10 and 2.11. We found close correlation (Spearman's $r = 0.88$ in cases, Spearman's $r = 0.88$ in controls, $N = 277$). We also observed that pooled Illumina GA sequencing produced consistently lower MAF estimates than Sanger sequencing of individual samples in this region.

Table 2.4: Functional Consequences for Non-synonymous SNPs

ID	Position	MAF (Seniors)	MAF(Controls)	Gene	Amino Acid Change	PolyPhen
rs28357980	4,917	0.073	0.060	MT-ND2	N [Asn] ⇒ D [Asp]	0.129 (benign)
rs28358886	8,697	0.078	0.053	MT-ATP6	M [Met] ⇒ I [Ile]	0.890 (possibly damaging)
rs9645429	9,055	0.070	0.051	MT-ATP6	A [Ala] ⇒ T [Thr]	0.845 (possibly damaging)
rs2853826	10,398	0.055	0.059	MT-ND3	T [Thr] ⇒ A [Ala]	0.000 (benign)
rs28359178	13,708	0.041	0.028	MT-ND5	A [Ala] ⇒ T [Thr]	0.000 (benign)
rs3135031	14,766	0.084	0.072	MT-CYTB	T [Thr] ⇒ I [Ile]	0.000 (benign)
rs28357681	14,798	0.109	0.077	MT-CYTB	F [Phe] ⇒ L [Leu]	0.000 (benign)
rs2853508	15,326	0.245	0.218	MT-CYTB	T [Thr] ⇒ A [Ala]	0.000 (benign)
rs3088309	15,452	0.134	0.118	MT-CYTB	L [Leu] ⇒ I [Ile]	0.029 (benign)

For each gene in the mitochondrial genome, the number of variants observed at $\geq 1\%$ frequency were tabulated (Table 2.5). The most variable protein-coding gene is MT-ND3, with 20.2 variants per kb in cases, and 17.3 variants per kb in controls. The most variable RNA gene is MT-TT, with 45.5 variants per kb in cases and 75.5 variants per kb in controls.



```
                          1270       1280       1290       1300       1310
CONSENSUS                 CCACTAGG*ATA*CCAACAA*ACCTACCCACCCTTAA*CAGTACATAGT
CG02601_E07_f_123_TEE.b   CCACTAGG*ATA*TCAACAA*ACCTACCCACCCTTAA*CAGTACATAGT
CG02601_E07_r_123_TEE.g   CCACTAGG*ATA*TCAACAA*ACCTACCCACCCTTAA*CAGTACATAGT
CG02601_E08_f_124_FER.b   CCACCAGG*ATA*CCAACAA*ACCTACCCACCCTTAA*CAGTACATAGT
CG02601_E08_r_124_FER.g   CCACCAGG*ATA*CCAACAA*ACCTACCCACCCTTAA*CAGTACATAGT
CG02601_E09_f_125_WIL.b   CCACTAGG*ATA*CCAACAA*ACCTACCCACCCTTAA*CAGTACATAGT
CG02601_E09_r_125_WIL.g   CCACTAGG*ATA*CCAACAA*ACCTACCCACCCTTAA*CAGTACATAGT
CG02601_E10_f_126_GUS.b   CCACTAGG*ATA*CCAACAA*ACCTACCCACCCTTAA*CAGTACATAGC
CG02601_E10_r_126_GUS.g   CCACTAGG*ATA*CCAACAA*ACCTACCCACCCTTAA*CAGTACATAGC
CG02601_E11_f_127_WIL.b   CCACTAGG*ATA*CCAACAA*ACCTACCCATCTTTAA*CAGCACATAGT
CG02601_E11_r_127_WIL.g   CCACTAGG*ATA*CCAACAA*ACCTACCCATCTTTAA*CAGCACATAGT
CG02601_E12_f_128_SIN.b   CCACTAGG*ATA*CCAACAA*ACCCACCCACCCTTAA*CAGTACATAGT
CG02601_E12_r_128_SIN.g   CCACTAGG*ATA*CCAACAA*ACCCACCCACCCTTAA*CAGTACATAGT
CG02601_F01_f_130_GRI.b   CCACTAGG*ATA*CCAACAA*ACCTACCCACCCTTAA*CAGTACATAGT
CG02601_F01_r_130_GRI.g   CCACTAGG*ATA*CCAACAA*ACCTACCCACCCTTAA*CAGTACATAGT
CG02601_F02_f_131_LAN.b   CCATTAGG*ATA*CCAACAA*ACCTACCCACCCTTAA*CAGTACATAGC
CG02601_F02_r_131_LAN.g   CCATTAGG*ATA*CCAACAA*ACCTACCCACCCTTAA*CAGTACATAGC
CG02601_F03_f_132_CUR.b   CCACTAGG*ATA*CCAACAA*ACCTACCCATCTTTAA*CAGCACATAGT
CG02601_F03_r_132_CUR.g   CCACTAGG*ATA*CCAACAA*ACCTACCCATCTTTAA*CAGCACATAGT
CG02601_F04_f_134_SMI.b   CCACTAGG*ATA*CCAACAA*ACCTACCCATCTTTAA*CAGTACATAGT
CG02601_F04_r_134_SMI.g   CCACTAGG*ATA*CCAACAA*ACCTACCCATCTTTAA*CAGTACATAGT
CG02601_F05_f_136_WEB.b   CCACTAGG*ATA*CCAACAA*ACCTACCCACCCTTAA*CAGTACATAGT
CG02601_F05_r_136_WEB.g   CCACTAGG*ATA*CCAACAA*ACCTACCCACCCTTAA*CAGTACATAGT
CG02601_F06_f_137_WAL.b   CCACTAGG*ATA*CCAACAA*ACCTACCCACCCTTAA*CAGTACATAGT
```

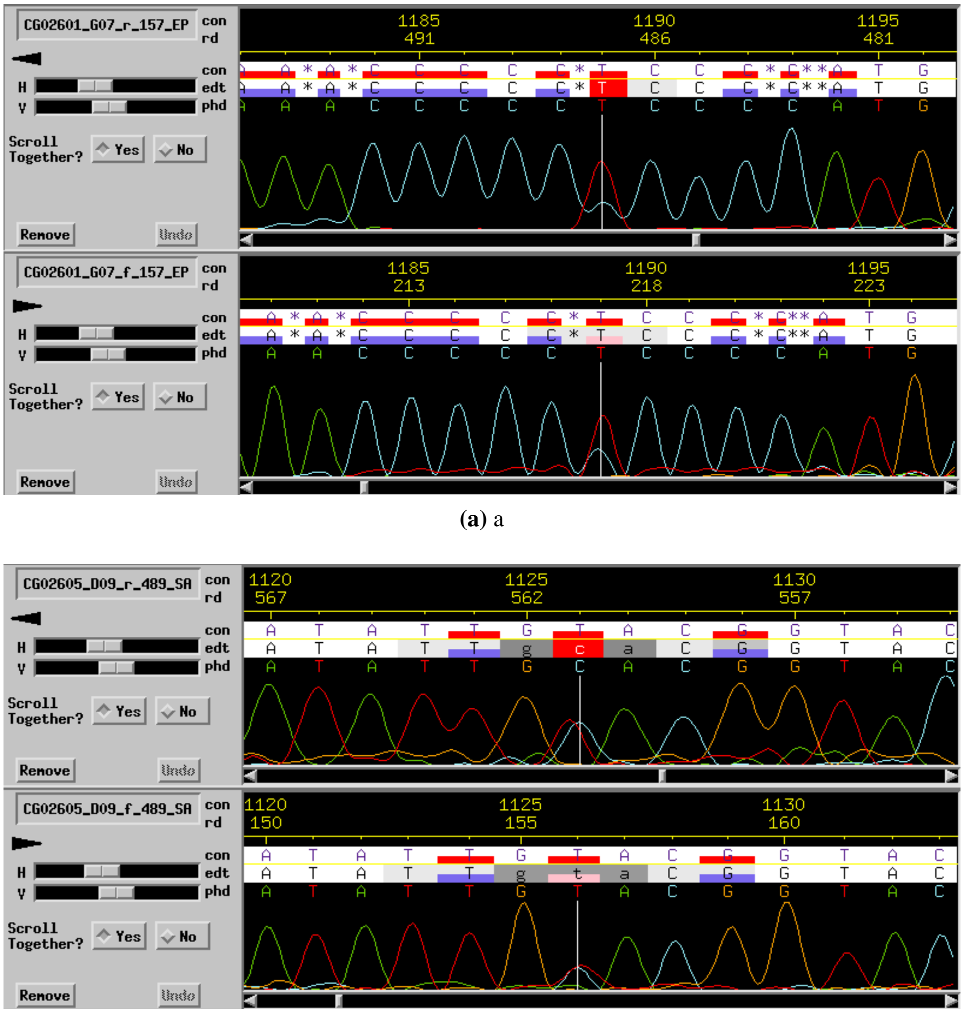
(**a**) Alignment

```
CG02601_E11_r_127_WI
con  1275      1280      1285      1290      1295
rd   412       407       402       397       392
con  G*A T A*C C A A C A A*A C C T A C C C A C C C T T
edt  G*A T A*C C A A C A A*A C C T A C C C A T C T T T
phd  G A T A C C A A C A A A C C T A C C C A T C T T T

CG02601_E12_f_128_SI
con  1275      1280      1285      1290      1295
rd   301       306       311       316       321
con  G*A T A*C C A A C A A*A C C T A C C C A C C C T T
edt  G*A T A*C C A A C A A*A C C C A C C C A C C C T T
phd  G A T A C C A A C A A A C C C A C C C A C C C T T
```

(**b**) Traces

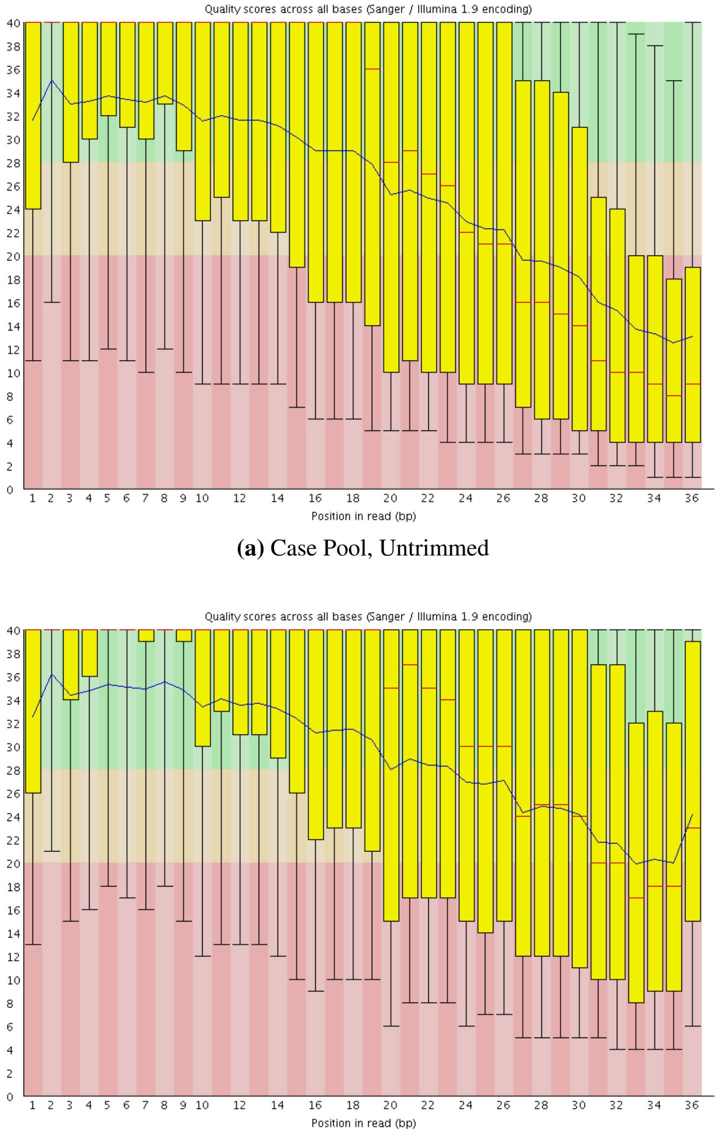
Figure 2.2: SNP Calling by Phred/Phrap/Consed + PolyPhred. (**a**) Reads were aligned to the revised Cambridge Reference Sequence (rCRS). Each sample was sequenced in both forward and reverse directions. Only a subset of samples are shown Sample IDs are at left in yellow type. (**b**) Variants were identified automatically using PolyPhred, and confirmed manually by visual inspection of sequence traces. Two samples (127_WIL and 128_SIN) with differing alleles at contig position 1,288 (rCRS position 16,288) are shown.



(a) a

(b) b

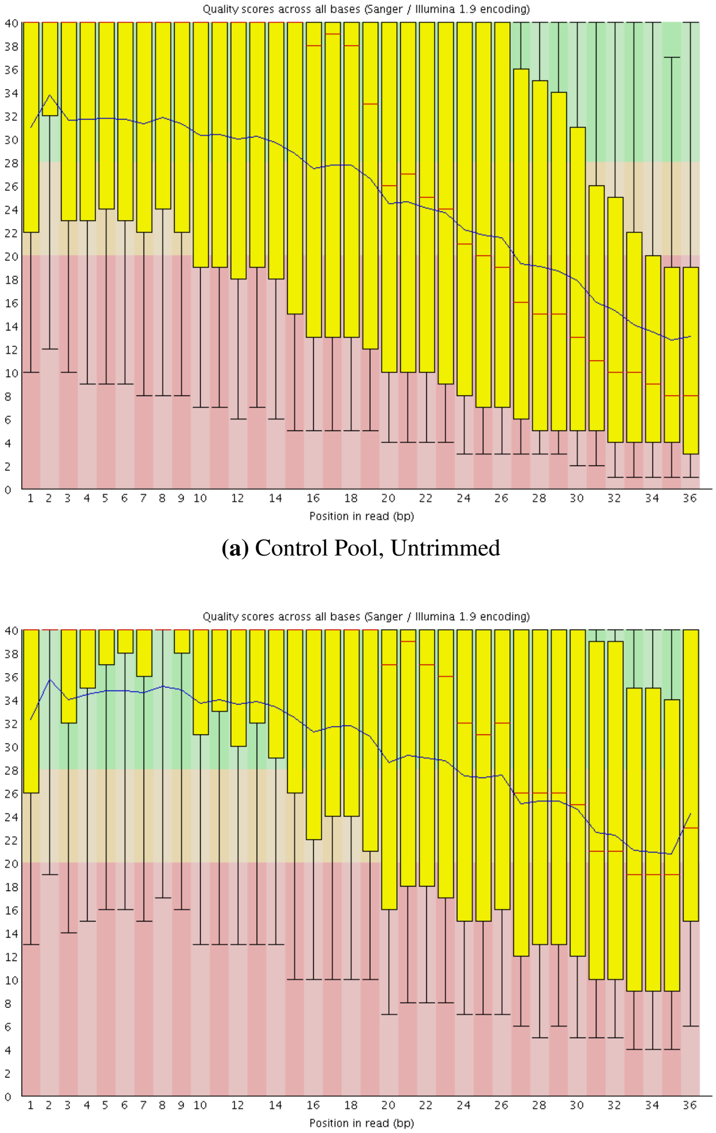
Figure 2.3: Putative Heteroplasmic Positions. Heteroplasmy was observed in some samples by identifying double-peaks in sequence traces. (**a**) Sample '157_EPP' shows putative heteroplasmy level of ∼ 25% at contig position 1,189 (rCRS position 16,189). (**b**) Sample '489_SAM' shows putative heteroplasmy level of ∼ 50% at contig position 1,126 (rCRS position 16,126). Note that the relative heights of the two peaks at the heteroplasmic positions are consistent in forward and reverse reads.



(a) Case Pool, Untrimmed

(b) Case Pool, Trimmed

Figure 2.4: Effect of Read-trimming on Per-base Quality Distributions (Case Pool) Average base quality score was calculated at each read position, across all reads. For each position, red line indicates median quality score, yellow box indicates interquartile range (25-75%), upper and lower whiskers represent 90% and 10% quantiles, respectively, and blue line represents mean quality score. The upwards shift in average quality for trimmed reads indicates that poor-qualty sequence near the 3’ end of reads has been removed in trimmed reads.



(a) Control Pool, Untrimmed

(b) Control Pool, Trimmed

Figure 2.5: Effect of Read-trimming on Per-base Quality Distributions (Control Pool) Average base quality score was calculated at each read position, across all reads. For each position, red line indicates median quality score, yellow box indicates interquartile range (25-75%), upper and lower whiskers represent 90% and 10% quantiles, respectively, and blue line represents mean quality score. The upwards shift in average quality for trimmed reads indicates that poor-qualty sequence near the 3' end of reads has been removed in trimmed reads.

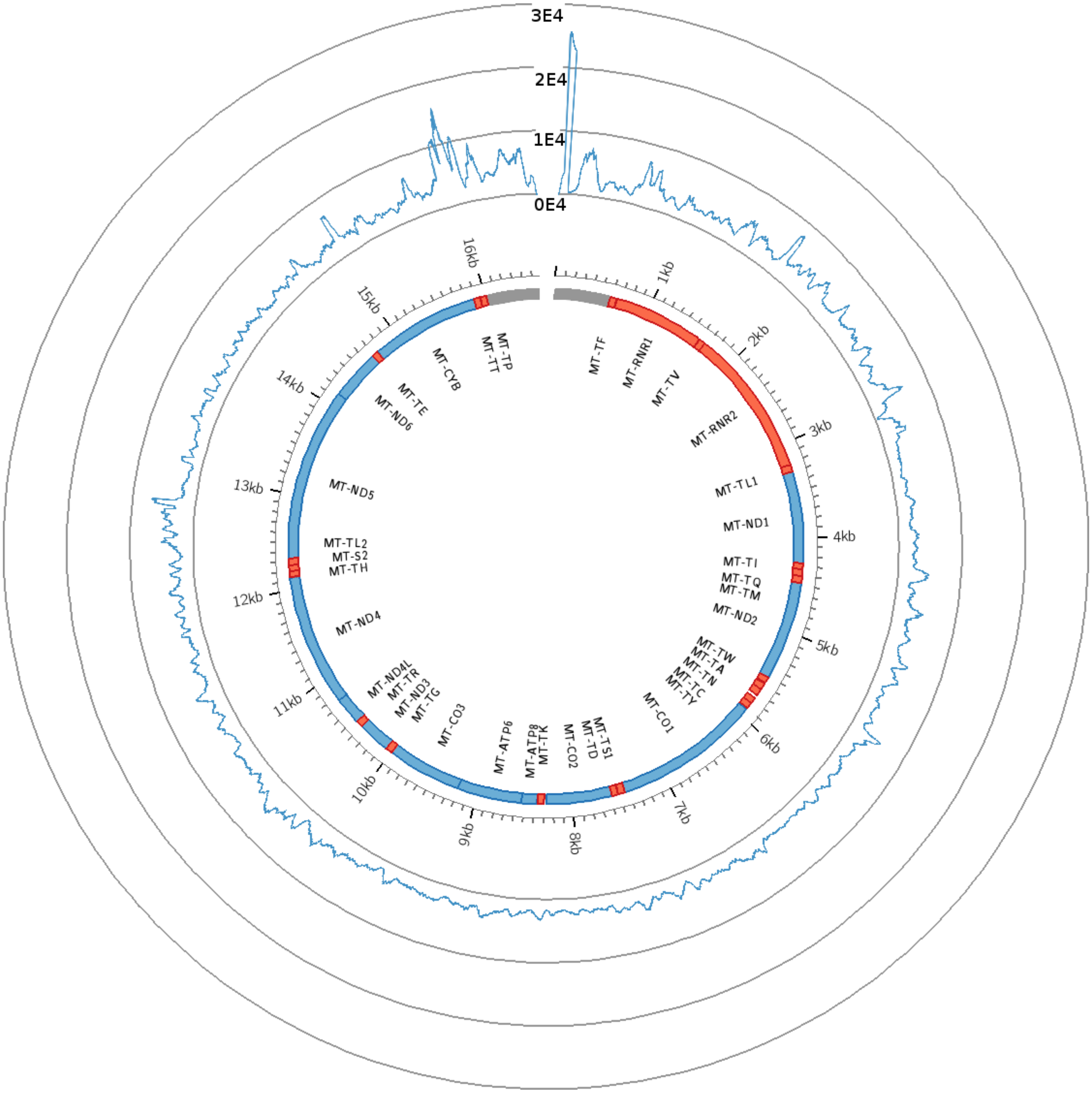


Figure 2.6: Sequence coverage across the mitochondrial genome (Case Pool). Blue line indicates high-quality (phred-scaled quality score = 40) sequence coverage. Graph lines every 10,000-fold depth.

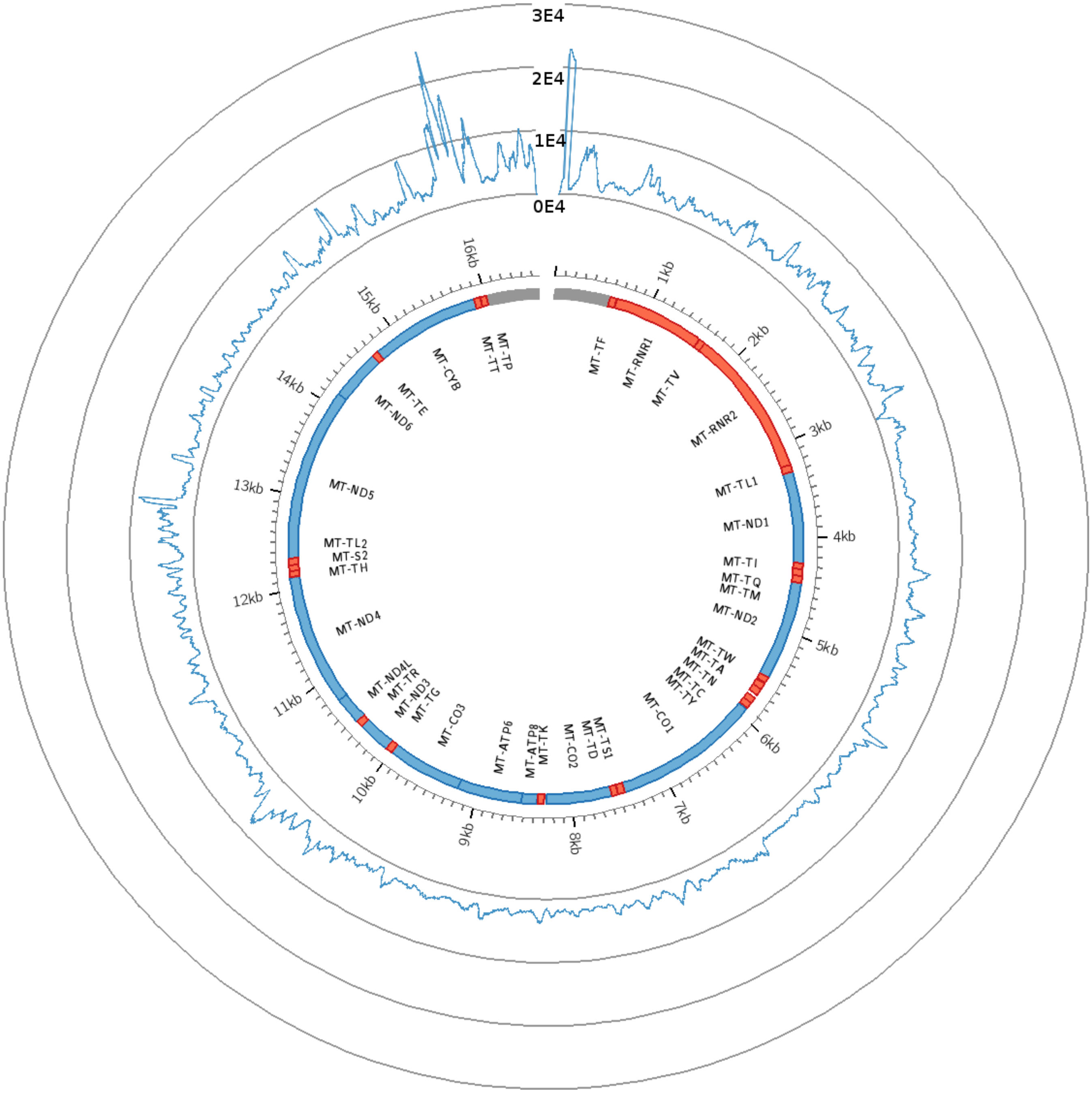


Figure 2.7: Sequence coverage across the mitochondrial genome (Control Pool). Blue line indicates high-quality (phred-scaled quality score = 40) sequence coverage. Graph lines every 10,000-fold depth.

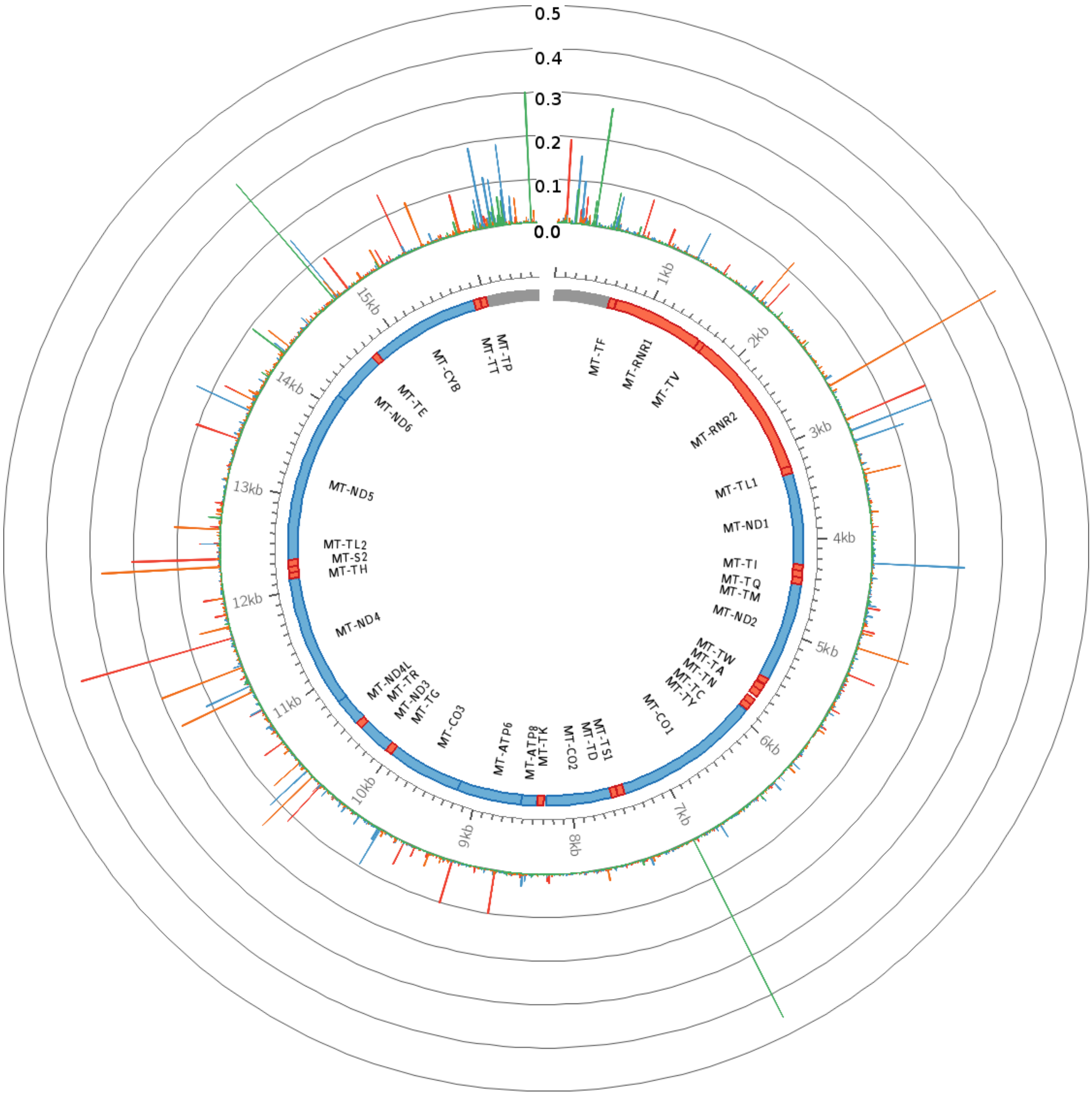


Figure 2.8: Minor allele frequencies (Case Pool). Locations and minor allele frequencies for all SNPs detected by Illumina GA sequencing. Base identities are indicated as follows: A = Red, C = Blue, G = Orange, T = Green. Heights of data bars indicate minor allele frequencies, scale bars every 10% allele frequency.

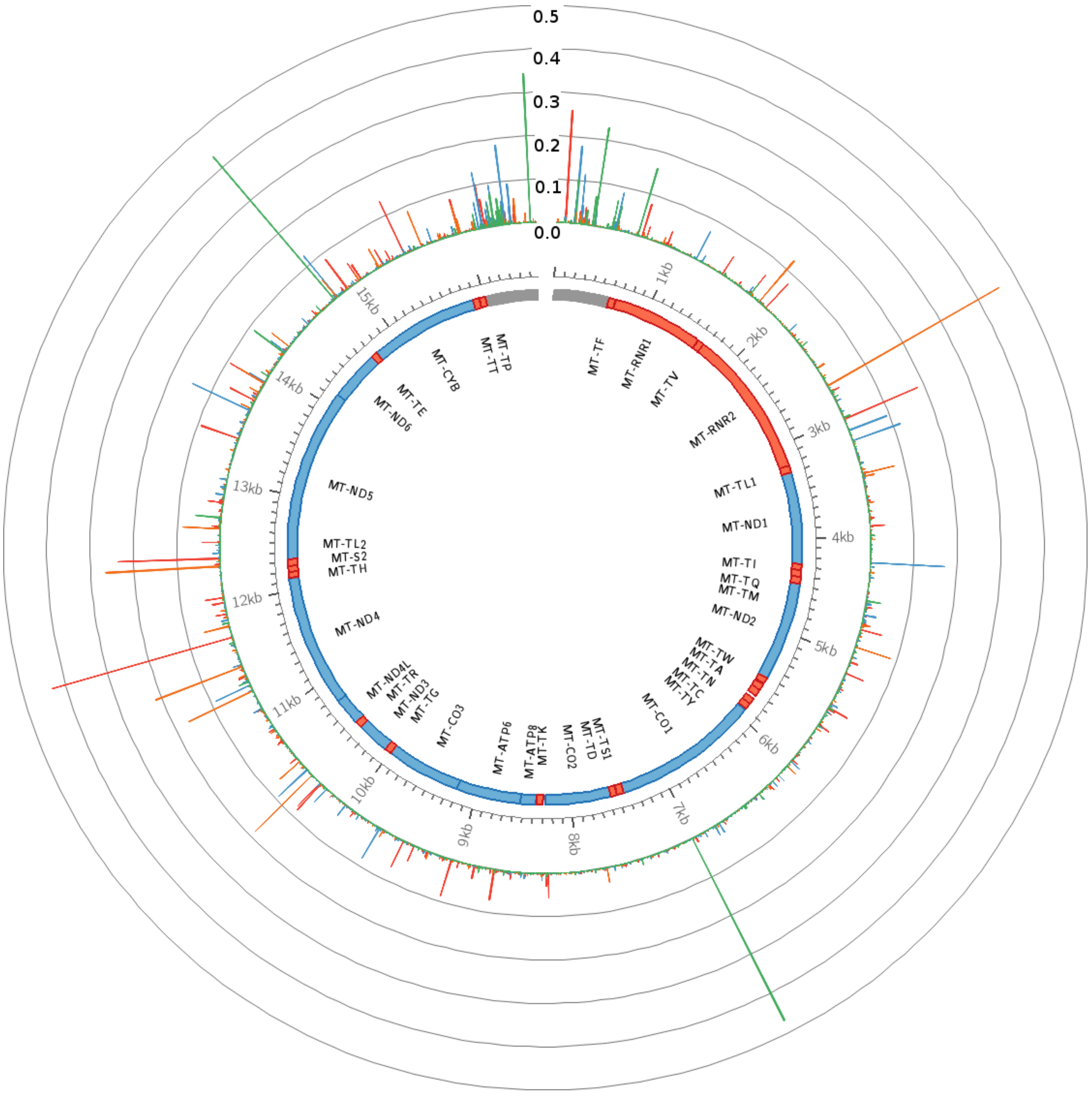


Figure 2.9: Minor allele frequencies (Control Pool). Locations and minor allele frequencies for all SNPs detected by Illumina GA sequencing. Base identities are indicated as follows: A = Red, C = Blue, G = Orange, T = Green. Heights of data bars indicate minor allele frequencies, scale bars every 10% allele frequency.

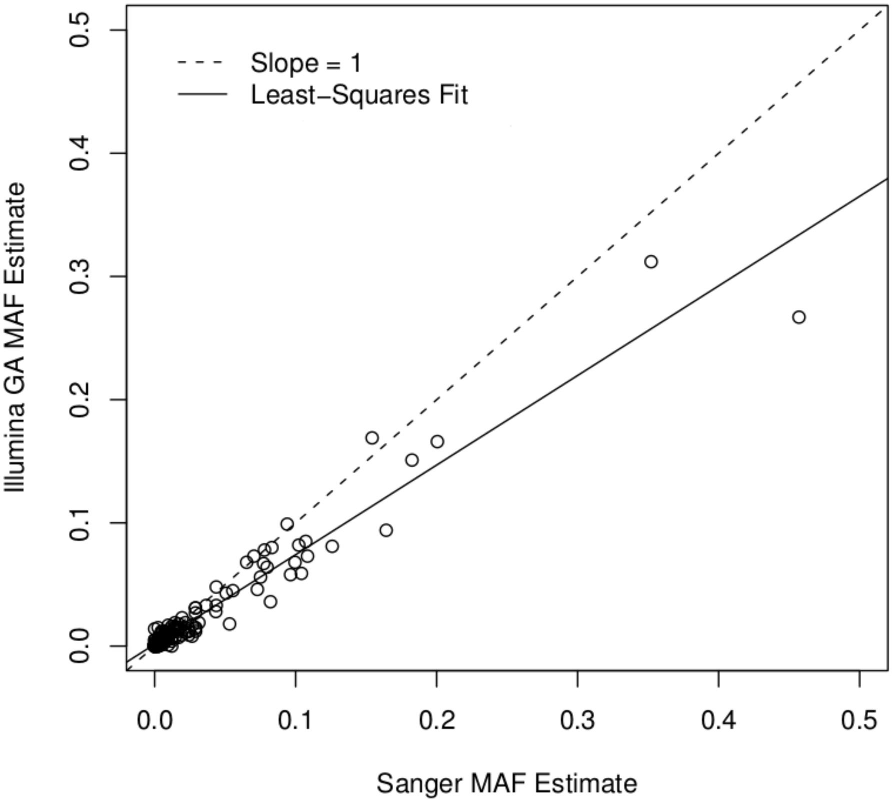


Figure 2.10: MAF comparison (Cases). Minor allele frequencies were determined by both Sanger sequencing and by pooled Illumina GA sequencing for 277 SNPs in the control region. The Spearman's rank correlation between the two estimates is 0.88 in the case sample set, and 0.91 in controls. Dashed line indicates slope = 1; the least-squares regression line is indicated by a solid line.

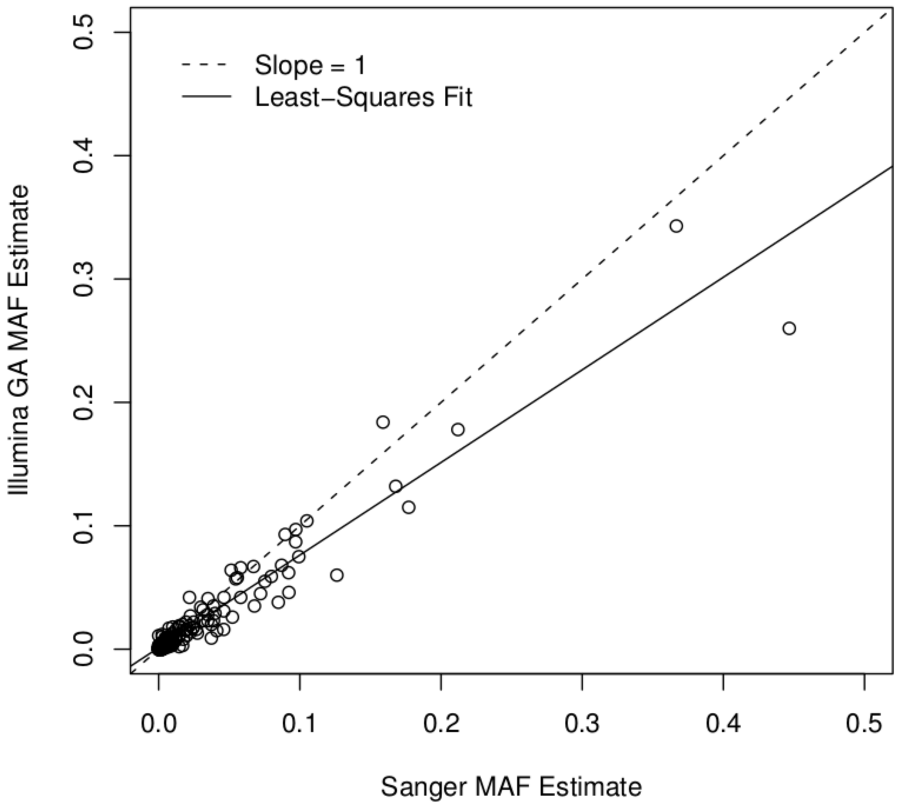


Figure 2.11: MAF comparison (Controls). Minor allele frequencies were determined by both Sanger sequencing and by pooled Illumina GA sequencing for 277 SNPs in the control region. The Spearman's rank correlation between the two estimates is 0.91 in control sample set. Dashed line indicates slope = 1; the least-squares regression line is indicated by a solid line.

Table 2.5: Number of Variants by Gene

		Total Variants		Variants per kb	
Gene	Size (bp)	Cases	Controls	Cases	Controls
MT-TF	71	0	0	0.0	0.0
MT-RNR1	954	7	10	7.3	10.5
MT-TV	69	0	0	0.0	0.0
MT-RNR2	1,559	17	16	10.9	10.3
MT-TL1	75	0	0	0.0	0.0
MT-ND1	956	8	7	8.4	7.3
MT-TI	69	0	0	0.0	0.0
MT-TQ	72	1	1	13.9	13.9
MT-TM	68	0	0	0.0	0.0
MT-ND2	1,042	12	19	11.5	18.2
MT-TW	68	0	0	0.0	0.0
MT-TA	69	1	2	14.5	29.0
MT-TN	73	0	0	0.0	0.0
MT-TC	66	0	1	0.0	15.2
MT-TY	66	0	0	0.0	0.0
MT-CO1	1,542	11	16	7.1	10.4
MT-TS1	69	1	1	14.5	14.5
MT-TD	68	0	0	0.0	0.0
MT-CO2	684	3	4	4.4	5.8
MT-TK	70	0	1	0.0	14.3
MT-ATP8	207	3	3	14.5	14.5
MT-ATP6	681	7	9	10.3	13.2
MT-C03	784	10	9	12.8	11.5
MT-TG	68	1	1	14.7	14.7
MT-ND3	346	7	6	20.2	17.3
MT-TR	65	1	1	15.4	15.4
MT-ND4L	297	2	3	6.7	10.1
MT-ND4	1,378	18	24	13.1	17.4
MT-TH	69	0	0	0.0	0.0
MT-TS2	59	0	0	0.0	0.0
MT-TL2	71	1	2	14.1	28.2
MT-ND5	1,812	28	27	15.5	14.9
MT-ND6	525	10	6	19.0	11.4
MT-TE	69	0	0	0.0	0.0
MT-CYTB	1,141	20	22	17.5	19.3
MT-TT	66	3	5	45.5	75.8
MT-TP	68	0	0	0.0	0.0
All Protein-coding	11,395	139	155	12.2	13.6
All RNA-coding	4,021	33	41	8.2	10.2

2.4 Discussion

We have shown here that it is possible to discover variants across the entire mitochondrial genome in over 400 samples in a single sequencing experiment. By combining long-PCR with second-generation sequencing technology, we were able to estimate the alllele frequencies of over 300 mitochondrial SNPs in our study population. This technique will be useful for rapidly surveying a large sample set for mitochondrial SNPs. Given its small size and high copy number per cell, mtDNA is a good candidate for pooled targeted resequencing efforts. The size of the mitochondrial chromosome (16.5 kb) makes it amenable to long PCR. The whole mtDNA genome can be amplified in one reaction, which simplifies the DNA pooling process. A similar variant detection has been employed by another group, using a pool size of 20 samples[46].

Figures 2.6 and 2.7 show that the entire mitochondrial genome was sufficiently covered by mapped reads to perform variant detection. There are strong peaks in coverage in both the case pool and control pool near position 200 within the control region. We attribute this peak to excess PCR primers that were carried through into the sequencing reaction.

A previous report showed accurate determination of allele frequencies of pooled genomic DNA on the ABI SOLiD, Roche 454 and Illumina GA II platforms [47]. Our estimation of MAF from Illumina sequencing of DNA pools correlates strongly with MAF calculated using genotypes determined using Sanger sequence data (Spearman's $r = 0.88$); this correlation is close to the value of $r^2 = 0.9637$ published by Druley et al[47]. The most likely source of discrepancy between these two datasets is due to small differences in the quantity of DNA that each sample contributes to the DNA pool.

In our analyses, MAF estimated from Illumina GA data is about 25% lower than our measurement from Sanger sequencing. We suggest that this discrepancy may represent a bias against mapping of reads containing non-reference bases. We suggest that a read that contains a real non-reference base in the form of a SNP is less likely to align than a read that contains no non-reference SNPs, and that this probem will be increased in low-quality sequence data. This phenomenon, referred to as 'reference bias,' has been observed in previous studies of next-generation sequence data[48].

The number of variants observed in at least 1% of samples varied from 0 (MT-TY, MT-TF for example) to 27 (MT-ND5) (see table 2.5). When normalized by the length of the gene, the most variable genes are MT-ND3 (20.2 variants/kb in cases, 17.3 variants/kb in controls) and MT-TT (45.5 variants/kb in cases, 75.8 variants/kb in controls). Note, however, that the short length of the tRNA genes ($\sim$ 70 bp) leads to a highly variable estimate of variants/kb. Overall the distribution of variants was similar in protien-coding and RNA-coding genes at roughly 10 variants/kb.

Although our study was not designed to investigate the role that heteroplasmic variants play in the aging process, we did detect a small number of putative heteroplasmic variants by Sanger sequencing. For low levels of heteroplasmy, (below $\sim$ 25%) it would be difficult to distinguish a true heteroplasmic variant from background noise in the sequence trace. The few instances of heteroplasmy that we were able to identify with some certainty appeared to be close to 50% heteroplasmic (See 2.3 for a representative example).

Chapter 3

A Case-Control Association Study for Mitochondrial Variants and Healthy Aging

In order to identify variants that are associated with the healthy-aging phenotype, case-control association tests were performed using PLINK software[49]. Each SNP is analyzed by comparing the major and minor allele frequencies in cases versus controls, by applying a Chi-squared (χ^2) test.

The power of a Chi-squared test to detect a genetic association is based on a comparison of a null $\chi^2_{(1-\alpha)}$ distribution to an alternative χ^2 distribution with non-centrality parameter λ, proportional to the effect size[50]. It is expressed as follows:

$$\text{Power} = \mathrm{P}(\chi^2(df,\lambda) \geq \chi^2_{1-\alpha}(df)), \tag{3.1}$$

where:

$$\lambda = \Delta^2 N = \left(\frac{(p-q)^2}{q}\right) N \tag{3.2}$$

and for a 2×2 contingency table, the number of degrees of freedom (df) are one.

Coding-region variants were nominated for genotyping based on three criteria.

Variants that showed a suggestive *P*-value (< 0.05) based on a comparison of the estimated Minor Allele Frequency (MAF) from pooled Illumina GA II sequencing were genotyped, as were a set of 64 tag Single Nucleotide Polymorphism (SNP)s that were designed to capture all common variants present at $> 1\%$ in the European population, with linkage disequilibrium of at least $r^2 = 0.8$[15]. Finally, any variant with an estimated MAF of > 0.05 based on pooled sequencing that did not fit the first two criteria was also included. In total, 92 SNPs were nominated for genotyping (see Supplemental Table B.1).

Due to our limited statistical power to detect moderate effects in low-frequency variants, a MAF cut-off of 10% was applied before association testing. This limited the number of SNPs that qualified for testing to nine control-region SNPs and seven coding-region SNPs (See tables 3.1 and 3.2, respectively).

One SNP in the control region was selected for testing based on previous reports of its association with longevity in Italian[3], Finnish and Japanese[37] populations.

3.1 Methods

3.1.1 Power Calculations

Statistical power was calculated with the PS power and sample size calculator[51]. Power curves were calculated for a sample of 419 cases and 415 controls, at minor allele frequencies of 0.01, 0.05, 0.10, 0.25 and 0.50, with a false-positive rate $\alpha = 0.05$.

3.1.2 Genotyping and Quality Control

Genotyping was performed on the Sequenom MassARRAY platform at the McGill University/Genome Québec Innovation Centre. A set of 92 SNPs were included in the first assay set. A set of 37 genotyping assays were repeated due to quality control failure.

Quality control was performed in collaboration with Dr. Denise Daley (University of British Columbia, St.Paul's Hospital). Assays with call rates below 95% were considered 'failed' and were re-designed. Genotype cluster plots were visually inspected for irregularities.

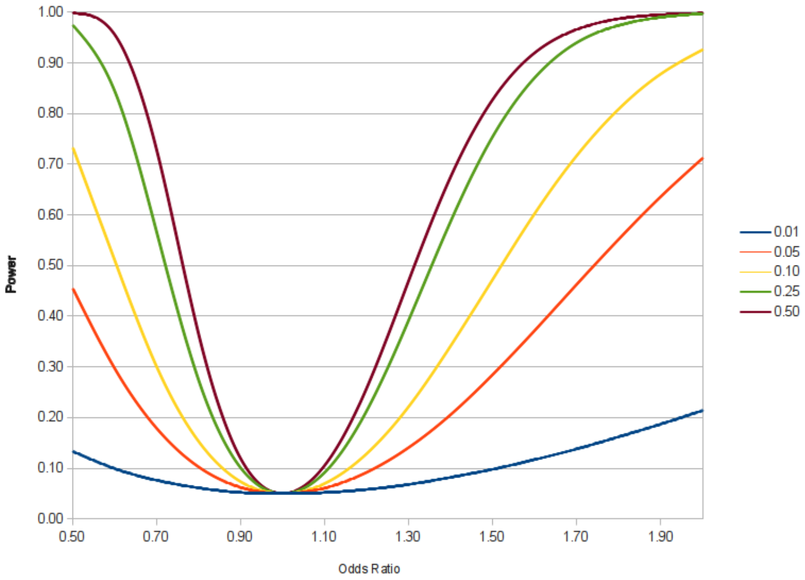


Figure 3.1: Power to Detect Association. Statistical power was calculated using PS software[51]. Curves are shown for the following control MAFs: 0.01 (blue), 0.05 (orange),0.10 (yellow), 0.25 (green) and 0.50 (purple). For all curves, $\alpha = 0.05$ number of cases = 419, number of controls = 415.

3.2 Results

This study is powered to detect an odds ratio of at least 1.75 (or 0.45) for a variant at minor allele frequency of 0.10 with a false-positive rate of 0.05 (see 3.1).

Of the 92 SNPs that were chosen for the initial round of Sequenom genotyping, 37 failed quality control (See B.3) due to low call rates. These assays were redesigned and repeated. Of the second set, only three assays failed quality controls (mt9947, rs41345446 and rs41347846).

After performing χ^2 tests for association between mtDNA alleles and healthy aging, no variants that were tested showed association with the healthy aging phenotype, at a p-value significance threshold of 0.05. The lowest p-value for control region SNPs was rs117135796 at position 152, with a p-value of 0.258 and odds ratio of 0.81. For coding region SNPs, the lowest p-value was 0.280, with odds ratio 1.11 for rs2853495 at position 11,719 within the MT-ND4 gene.

The rs62581312 variant at position 150 within the control region showed a *p*-value of 0.171 and odds ratio of 0.72.

Table 3.1: Mitochondrial Control Region (MAF > 0.10)

Chr	ID	Position	Minor Allele	Major Allele	$F_A{}^a$	$F_U{}^b$	χ^{2c}	P	Odds Ratio
M	rs3087742	73	A	G	0.456	0.442	0.163	0.726	1.06
M	rs117135796	152	C	T	0.185	0.218	1.405	0.258	0.81
M	rs2857291	195	C	T	0.170	0.181	0.173	0.714	0.93
M	rs28625645	489	C	T	0.102	0.098	0.031	0.908	1.04
M	mt16126	16,126	C	T	0.195	0.167	1.083	0.318	1.21
M	rs55749223	16,189	C	T	0.139	0.118	0.811	0.404	1.21
M	rs2857290	16,270	T	C	0.107	0.093	0.425	0.561	1.16
M	rs34799580	16,311	C	T	0.151	0.167	0.384	0.567	0.89
M	rs3937033	16,519	T	C	0.340	0.348	0.062	0.826	0.96

[a] Minor Allele Frequency in ‘Affecteds’ (seniors)
[b] Minor Allele Frequency in ‘Unaffecteds’ (controls)
[c] χ^2 test statistic

Table 3.2: Mitochondrial Coding Region (MAF > 0.10)

Chr	ID	Position	Minor Allele	Major Allele	$F_A{}^a$	$F_U{}^b$	χ^{2c}	P	Odds Ratio
M	rs2853517	709	G	A	0.144	0.128	0.942	0.332	1.14
M	rs3928306	3,010	C	T	0.264	0.246	0.760	0.383	1.10
M	rs2015062	7,028	A	G	0.445	0.436	0.155	0.694	1.04
M	rs2853825	9,477	G	A	0.104	0.097	0.214	0.644	1.08
M	rs2853495	11,719	A	G	0.493	0.468	1.167	0.280	1.11
M	rs2853499	12,372	C	T	0.242	0.241	0.002	0.961	1.01
M	rs28357681	14,798	A	G	0.158	0.145	0.516	0.473	1.10

[a] Minor Allele Frequency in ‘Affecteds’ (seniors)
[b] Minor Allele Frequency in ‘Unaffecteds’ (controls)
[c] χ^2 test statistic

Table 3.3: Replication of rs62581312 (C150T)

Chr	ID	Position	Minor Allele	Major Allele	$F_A{}^a$	$F_U{}^b$	χ^{2c}	P	Odds Ratio
M	rs62581312	150	T	C	0.082	0.110	1.88	0.171	0.72

[a] Minor Allele Frequency in ‘Affecteds’ (seniors)
[b] Minor Allele Frequency in ‘Unaffecteds’ (controls)
[c] χ^2 test statistic

3.3 Discussion

It is notable that this study did not replicate the previously-reported association at position 150 of the mitochondrial control region[3]. There are several potential explanations for this result. The study by Zhang et al. focused on a group of Italians aged 99-106 years, whereas our samples qualify at age 85 and are mainly of British ancestry. Although the association was replicated in both Finnish and Japanese populations,[37] there may be population-specific genetic or environmental factors that combine with the position 150 polymorphism to effect the aging phenotype.

Because the mitochondrial genome does not recombine, it is possible to identify sets of variants that are inherited together and form mitochondrial haplotypes. These haplotypes have been traced to geographic/ancestral lineages across the world[14]. Previous studies have identified haplogroups that are associated with longevity[52, 1, 53]. In our study, we elected to combine a previously-published set of common European mitochondrial tag SNPs[15] with additional variants that were discovered by pooled next-generation sequencing.

Chapter 4

Discussion

We have designed a cost-effective method of surveying the mitochondrial genomes of hundreds of samples for single-nucleotide polymorphisms. Our method combines long-range PCR with a high-processivity, low-error DNA polymerase with pooled next-generation sequencing on the Illumina Genome Analyzer platform. Our single-amplicon long-PCR mtDNA isolation method also eliminates complications due to co-amplification of of mtDNA-derived pseudogenes (NUMTs) in the nuclear genome.

While we have established that is is possible to isolate and sequence the whole mitochondrial genome via a single long-PCR reaction, our mtDNA isolation protocol was not designed to detect common deletions that have been observed in other studies[54]. Future studies may be able to take advantage of paired-end sequencing to detect relatively large-scale deletions such as the common 4.9 kb deletion that has been characterized between rCRS positions 8,470 and 13,446[55]. In a paired-end sequencing experiment, deletions can be detected when paired reads map further apart than the expected $\sim$ 300 bp insert size[56].

Previous reports have demonstrated accurate determination of allele frequencies of pooled genomic DNA on the ABI SOLiD, Roche 454 and Illumina GA IIx platforms[47, 57]. Our estimation of MAF from Illumina sequencing of DNA pools correlates strongly with MAF calculated using genotypes determined using Sanger sequence data (Spearman's $r = 0.88$); this correlation is close to the value of $r^2 = 0.9637$ published by Druley *et al*[47]. The most likely source of discrepancy between these two datasets is due to small differences in the quantity of DNA that each sample contributes to the DNA pool.

Our study was not designed for sensitive detection of heteroplasmic variants, though we did observe a small number of variants that were suggestive of heteroplasmy via analysis of Sanger sequence traces. These variants appear similar to heterozygous variants in diploid nuclear sequence data, with overlapping peaks of two different fluorophores (Fig 2.3), and suggest a roughly equal mixture of two alleles. For lower levels of heteroplasmy, it becomes difficult to distinguish true heteroplasmy from background noise in the Sanger sequence trace. The goal of this study was to investigate the role that common, heritable mitochondrial variants may play in the human aging process. Heteroplasmy can be inherited, and can also arise *de novo*, and can vary by tissue type[58, 59]. Recent studies have shown that next-generation sequencing can be a powerful tool to detect heteroplasmy[35]. In order to detect heteroplasmic variants in a pooled sequencing experiment, one would need a way to tie each read to a specific sample, rather than estimate the allele frequencies of the whole pool as was done in our experiment. New DNA barcoding methods (also called ‘indexed’ sequencing) have now made this possible[60]. It is well established that heteroplasmic variants accumulate with age[58, 61, 62, 63], so if we had used a sequencing technology that was sensitive to heteroplasmy then it is likely that we would have observed differences in levels in heteroplasmy between our cases ($>$ 85 years of age) and controls (40-54 years of age). It would remain unclear, however, if those somatic heteroplasmic variants would be passed down to future generations and also to what extent heteroplasmic variants are involved with healthy aging.

In our analyses, MAF estimated from Illumina GA data is about 25% lower than our measurement from Sanger sequencing. We suggest that this discrepancy may represent a bias against mapping of reads containing non-reference bases. We suggest that a read that contains a real non-reference base in the form of a SNP is less likely to align than a read that contains no non-reference SNPs, and that this problem will be increased in low-quality sequence data. This phenomenon is referred to as ‘reference bias,’ and has been observed in other next-generation sequencing experiments[48].

When conducting a case-control genetic association study, it is important to control for possible population stratification. If the case and control groups are composed of samples from different ethnic backgrounds, it is possible to observe false-positive associations due to differences in population-specific allele frequencies that play no functional role in the phenotype of interest. Another study that is also part of the G^3 Study of Healthy Aging, and used the same sample set has analyzed a set of ancestry-informative markers and found no evidence for population stratification[64].

Other studies have found evidence for gene-gene interactions in the etiology of type II diabetes mellitus. One study used a non-parametric machine learning method known as Multifactor Dimensionality Reduction (MDR) to study genetic association with the metabolic disease. Out of 23 loci on 15 candidate genes in the study, the researchers were able to identify a two-locus interaction between PPARγ and UCP2 that significantly reduced risk of T2DM in Koreans (odds ratio: 0.51, 95% CI: 0.34, 0.77, p=0.0016)[65]. Another study, using a more traditional logistic regression model, identified a three-locus interaction between variants in UCP2, PGC-1α and position 10,398 of the mitochondrial genome in the North Indian Population[66]. Although our study lacked the statistical power to detect these sorts of effects, this may be a fruitful direction for future studies of mitochondrial genetics in aging.

Bibliography

[1] Marta D. Costa et al. “Data from complete mtDNA sequencing of Tunisian centenarians: Testing haplogroup association and the ‘golden mean’ to longevity”. In: *Mechanisms of Ageing and Development* 130.4 (Apr. 2009), pp. 222–226 (cit. on pp. 1, 33).

[2] Paola Sebastiani et al. “Whole Genome Sequences of a Male and Female Supercentenarian, Ages Greater than 114 Years”. In: *Frontiers in Genetics* 2.January (2012), pp. 1–28. ISSN: 1664-8021 (cit. on p. 1).

[3] J. Zhang et al. “Strikingly higher frequency in centenarians and twins of mtDNA mutation causing remodeling of replication origin in leukocytes”. In: *Proceedings of the National Academy of Sciences* 100.3 (2003), pp. 1116–1121 (cit. on pp. 1, 7, 30, 33).

[4] M.F. Folstein, S.E. Folstein, P.R. McHugh, et al. “Mini-Mental State: a practical method for grading the cognitive state of patients for the clinician”. In: *J Psychiatr Res* 12.3 (1975), pp. 189–198 (cit. on p. 1).

[5] D. Podsiadlo and S. Richardson. “The timed” Up & Go”: a test of basic functional mobility for frail elderly persons.” In: *Journal of the American Geriatrics Society* 39.2 (1991), p. 142 (cit. on p. 1).

[6] JA Yesavage et al. “Geriatric Depression Scale (GDS)”. In: *Journal of Psychiatric Research* 17 (1983), pp. 37–49 (cit. on p. 1).

[7] S. Katz. “Assessing self-maintenance: activities of daily living, mobility, and instrumental activities of daily living”. In: *J Am Geriatr Soc* 31.12 (1983), pp. 721–27 (cit. on p. 1).

[8] M.B. Hock and A. Kralli. “Transcriptional Control of Mitochondrial Biogenesis and Function”. In: *Annual Review of Physiology* 71 (2009), pp. 177–203 (cit. on p. 2).

[9] M.T. Ryan and N.J. Hoogenraad. “Mitochondrial-nuclear communications”. In: *Annual Review of Biochemisrty* 76 (2007), pp. 701–722 (cit. on p. 2).

[10] M Stoneking. “Hypervariable sites in the mtDNA control region are mutational hotspots.” In: *American journal of human genetics* 67.4 (Oct. 2000), pp. 1029–32 (cit. on p. 4).

[11] P. Sutovsky et al. “Early degradation of paternal mitochondria in domestic pig (Sus scrofa) is prevented by selective proteasomal inhibitors lactacystin and MG132”. In: *Biology of reproduction* 68.5 (2003), p. 1793 (cit. on p. 4).

[12] W.E. Thompson, J. Ramalho-Santos, and P. Sutovsky. “Ubiquitination of prohibitin in mammalian sperm mitochondria: possible roles in the regulation of mitochondrial inheritance and sperm quality control”. In: *Biology of reproduction* 69.1 (2003), p. 254 (cit. on p. 4).

[13] A. Eyre-Walker and P. Awadalla. “Does human mtDNA recombine?” In: *Journal of Molecular Evolution* 53.4 (2001), pp. 430–435 (cit. on p. 4).

[14] D.M. Behar et al. “The Genographic Project public participation mitochondrial DNA database”. In: *PLoS Genetics* 3.6 (2007), e104 (cit. on pp. 4, 33).

[15] R. Saxena et al. “Comprehensive association testing of common mitochondrial DNA variation in metabolic disease”. In: *The American Journal of Human Genetics* 79.1 (2006), pp. 54–61 (cit. on pp. 4, 30, 33).

[16] PF Chinnery and DM Turnbull. “Mitochondrial DNA and disease”. In: *The Lancet* 354 (1999), S17–S21 (cit. on p. 5).

[17] A.H.V. Schapira. “Mitochondrial diseases”. In: *The Lancet* (2012) (cit. on p. 5).

[18] D. C. Wallace. “Mitochondrial Diseases in Man and Mouse”. In: *Science* 283.5407 (Mar. 1999), pp. 1482–1488 (cit. on p. 5).

[19] Raquel Moreno-Loshuertos et al. “Differences in reactive oxygen species production explain the phenotypes associated with common mouse mitochondrial DNA variants.” In: *Nature genetics* 38.11 (Nov. 2006), pp. 1261–8 (cit. on p. 5).

[20] S Toyokuni. “Reactive oxygen species-induced molecular damage and its application in pathology.” In: *Pathology international* 49.2 (Feb. 1999), pp. 91–102 (cit. on p. 5).

[21] João F Passos, Gabriele Saretzki, and Thomas von Zglinicki. “DNA damage in telomeres and mitochondria during cellular senescence: is there a connection?” In: *Nucleic acids research* 35.22 (Jan. 2007), pp. 7505–13 (cit. on p. 5).

[22] H. Symonds et al. “p53-dependent apoptosis suppresses tumor growth and progression in vivo.” In: *Cell* 78.4 (1994), p. 703 (cit. on p. 6).

[23] L. Yuqi et al. “Voltage-dependent anion channel(VDAC) is involved in apoptosis of cell lines carrying the mitochondrial DNA mutation”. In: *BMC Medical Genetics* 10.1 (2009), p. 114 (cit. on p. 6).

[24] J. Campisi. “Senescent cells, tumor suppression, and organismal aging: good citizens, bad neighbors”. In: *Cell* 120.4 (2005), pp. 513–522 (cit. on p. 6).

[25] J. Halaschek-Wiener et al. “Genetic variation in healthy oldest-old”. In: *PloS one* 4.8 (2009), e6641 (cit. on p. 6).

[26] F. Rodier, J. Campisi, and D. Bhaumik. “Two faces of p53: aging and tumor suppression”. In: *Nucleic Acids Research* (2007) (cit. on p. 6).

[27] J.F. Passos and T. von Zglinicki. “Mitochondria, telomeres and cell senescence”. In: *Experimental gerontology* 40.6 (2005), pp. 466–472 (cit. on p. 6).

[28] J.H. Santos et al. “Mitochondrial hTERT exacerbates free-radical-mediated mtDNA damage”. In: *Aging Cell* 3.6 (2004), pp. 399–411 (cit. on p. 6).

[29] J.F. Passos, G. Saretzki, and T. von Zglinicki. “DNA damage in telomeres and mitochondria during cellular senescence: is there a connection?” In: *Nucleic Acids Research* 35.22 (2007), p. 7505 (cit. on p. 6).

[30] J. Haendeler et al. “Antioxidants inhibit nuclear export of telomerase reverse transcriptase and delay replicative senescence of endothelial cells”. In: *Circulation research* 94.6 (2004), p. 768 (cit. on p. 6).

[31] M. Hayakawa et al. “Age-associated oxygen damage and mutations in mitochondrial DNA in human hearts.” In: *Biochemical and biophysical research communications* 189.2 (1992), p. 979 (cit. on p. 7).

[32] N.W. Soong et al. “Mosaicism for a specific somatic mitochondrial DNA mutation in adult human brain”. In: *Nature genetics* 2.4 (1992), pp. 318–323 (cit. on p. 7).

[33] S. Melov et al. “Marked increase in the number and variety of mitochondrial DNA rearrangements in aging human skeletal muscle”. In: *Nucleic acids research* 23.20 (1995), pp. 4122–4126 (cit. on p. 7).

[34] Ulf Gyllensten. “MtDNA substitution rate and segregation of heteroplasmy in coding and noncoding regions”. In: *Human Genetics* 107 (2000), pp. 45–50 (cit. on p. 7).

[35] "Detecting Heteroplasmy from High-Throughput Sequencing of Complete Human Mitochondrial DNA Genomes." In: *American journal of human genetics* 87.2 (Aug. 2010), pp. 237–249 (cit. on pp. 7, 35).

[36] Rodrigue Rossignol et al. "Mitochondrial threshold effects". In: *Biochemical Journal* 370 (2003), pp. 751–762 (cit. on p. 7).

[37] A.K. Niemi et al. "A combination of three common inherited mitochondrial DNA polymorphisms promotes longevity in Finnish and Japanese subjects". In: *European Journal of Human Genetics* 13.2 (2005), pp. 166–170 (cit. on pp. 8, 30, 33).

[38] Masashi Tanaka et al. "Mitochondrial genotype associated with longevity". In: *The Lancet* 351 (1998), pp. 185–186 (cit. on p. 8).

[39] K. Higasa. *Kyushu-U In Silico PCR*. 2006. URL: http://qsnp.gen.kyushu-u.ac.jp/genome/InSilicoPCR.html (cit. on p. 11).

[40] AR Brooks-Wilson et al. "Germline E-cadherin mutations in hereditary diffuse gastric cancer: assessment of 42 new families and review of genetic screening criteria". In: *British Medical Journal* 41.7 (2004), p. 508 (cit. on p. 11).

[41] D. Gordon, C. Abajian, and P. Green. "Consed: a graphical tool for sequence finishing". In: *Genome research* 8.3 (1998), pp. 195–202 (cit. on p. 11).

[42] B. Ewing et al. "Base-calling of automated sequencer traces usingPhred. I. Accuracy assessment". In: *Genome research* 8.3 (1998), pp. 175–185 (cit. on p. 11).

[43] B. Ewing and P. Green. "Base-calling of automated sequencer traces usingPhred. II. error probabilities". In: *Genome research* 8.3 (1998), pp. 186–194 (cit. on p. 11).

[44] H. Li and R. Durbin. "Fast and accurate short read alignment with Burrows–Wheeler transform". In: *Bioinformatics* 25.14 (2009), pp. 1754–1760 (cit. on p. 13).

[45] S. Andrews. *FASTQC. A quality control tool for high throughput sequence data*. 2010. URL: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (cit. on p. 13).

[46] T. Wang et al. "Estimating allele frequency from next-generation sequencing of pooled mitochondrial DNA samples". In: *Frontiers in genetics* 2 (2011) (cit. on p. 27).

[47] T.E. Druley et al. “Quantification of rare allelic variants from pooled genomic DNA”. In: *Nature Methods* 6.4 (2009), pp. 263–265 (cit. on pp. 27, 34).

[48] J.F. Degner et al. “Effect of read-mapping biases on detecting allele-specific expression from RNA-sequencing data”. In: *Bioinformatics* 25.24 (2009), pp. 3207–3212 (cit. on pp. 28, 35).

[49] Shaun Purcell et al. “PLINK: a tool set for whole-genome association and population-based linkage analyses.” In: *American journal of human genetics* 81.3 (Sept. 2007), pp. 559–75 (cit. on p. 29).

[50] Paul I W de Bakker et al. “Efficiency and power in genetic association studies.” In: *Nature genetics* 37.11 (2005), pp. 1217–23 (cit. on p. 29).

[51] W D Dupont and W D Plummer. “Power and sample size calculations. A review and computer program.” In: *Controlled clinical trials* 11.2 (Apr. 1990), pp. 116–28 (cit. on pp. 30, 31).

[52] S. Dato et al. “Association of the mitochondrial DNA haplogroup J with longevity is population specific”. In: *European journal of human genetics* 12.12 (2004), pp. 1080–1082 (cit. on p. 33).

[53] G. De Benedictis et al. “Mitochondrial DNA inherited variants are associated with successful aging and longevity in humans”. In: *The FASEB journal* 13.12 (1999), pp. 1532–1536 (cit. on p. 33).

[54] GA Cortopassi et al. “A pattern of accumulation of a somatic deletion of mitochondrial DNA in aging human tissues”. In: *Proceedings of the National Academy of Sciences* 89.16 (1992), p. 7370 (cit. on p. 34).

[55] C. Meissner et al. “The 4977bp deletion of mitochondrial DNA in human skeletal muscle, heart and different areas of the brain: A useful biomarker or more?” In: *Experimental gerontology* 43.7 (2008), pp. 645–652 (cit. on p. 34).

[56] I. Hajirasouliha et al. “Detection and characterization of novel sequence insertions using paired-end next-generation sequencing”. In: *Bioinformatics* 26.10 (2010), pp. 1277–1283 (cit. on p. 34).

[57] Zhi Wei et al. “SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data.” In: *Nucleic acids research* 39.19 (Oct. 2011), pp. 1–13 (cit. on p. 34).

[58] N. Sondheimer et al. “Neutral mitochondrial heteroplasmy and the influence of aging”. In: *Human molecular genetics* 20.8 (2011), pp. 1653–1659 (cit. on p. 35).

[59] H.A. Coller, N.D. Bodyak, and K. Khrapko. “Frequent intracellular clonal expansions of somatic mtDNA mutations”. In: *Annals of the New York Academy of Sciences* 959.1 (2002), pp. 434–447 (cit. on p. 35).

[60] S. Szelinger, A. Kurdoglu, D.W. Craig, et al. “Bar-coded, multiplexed sequencing of targeted DNA regions using the Illumina Genome Analyzer”. In: *Methods Mol Biol* 700 (2011), pp. 89–104 (cit. on p. 35).

[61] A. Bender et al. “High levels of mitochondrial DNA deletions in substantia nigra neurons in aging and Parkinson disease”. In: *Nature genetics* 38.5 (2006), pp. 515–517 (cit. on p. 35).

[62] Y. Michikawa et al. “Aging-dependent large accumulation of point mutations in the human mtDNA control region for replication”. In: *Science* 286.5440 (1999), p. 774 (cit. on p. 35).

[63] C.D. Calloway et al. “The frequency of heteroplasmy in the HVII region of mtDNA differs across tissue types and increases with age”. In: *The American Journal of Human Genetics* 66.4 (2000), pp. 1384–1397 (cit. on p. 35).

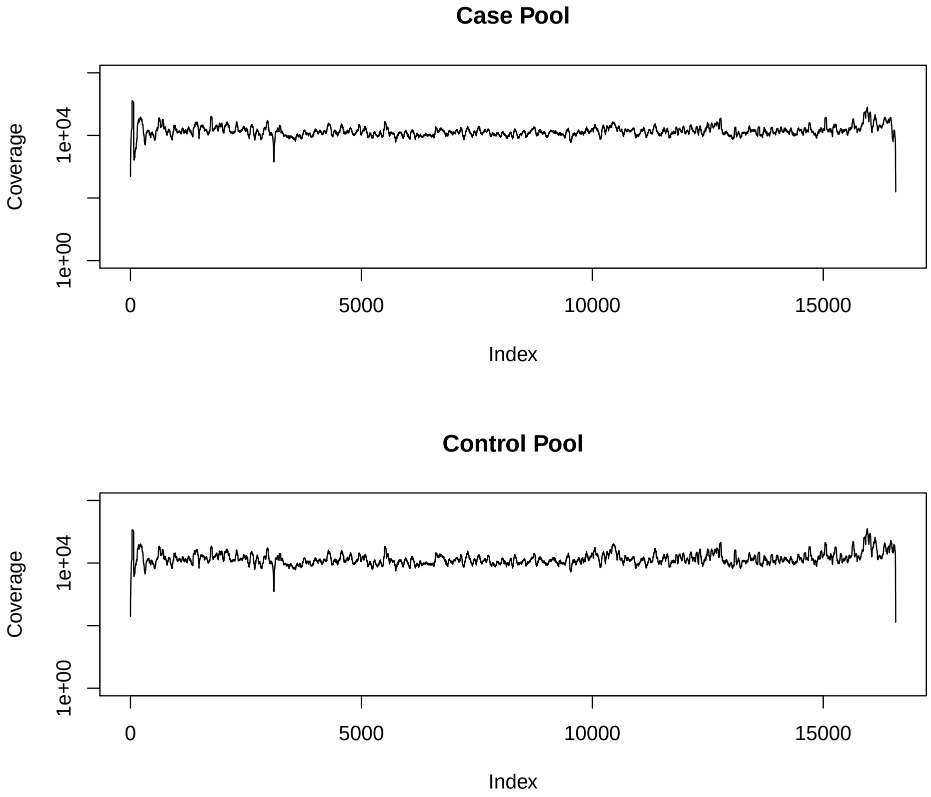
[64] J. Halaschek-Wiener et al. “Variants in BECN1 and MAPK14 are associated with healthy aging”. In Preparation (cit. on p. 36).

[65] YM Cho et al. “Multifactor-dimensionality reduction shows a two-locus interaction associated with Type 2 diabetes mellitus”. In: *Diabetologia* 47.3 (2004), pp. 549–554 (cit. on p. 36).

[66] A. Bhat et al. “PGC-1α Thr394Thr and Gly482Ser variants are significantly associated with T2DM in two North Indian populations: a replicate case-control study”. In: *Human genetics* 121.5 (2007), pp. 609–614 (cit. on p. 36).

Appendix A

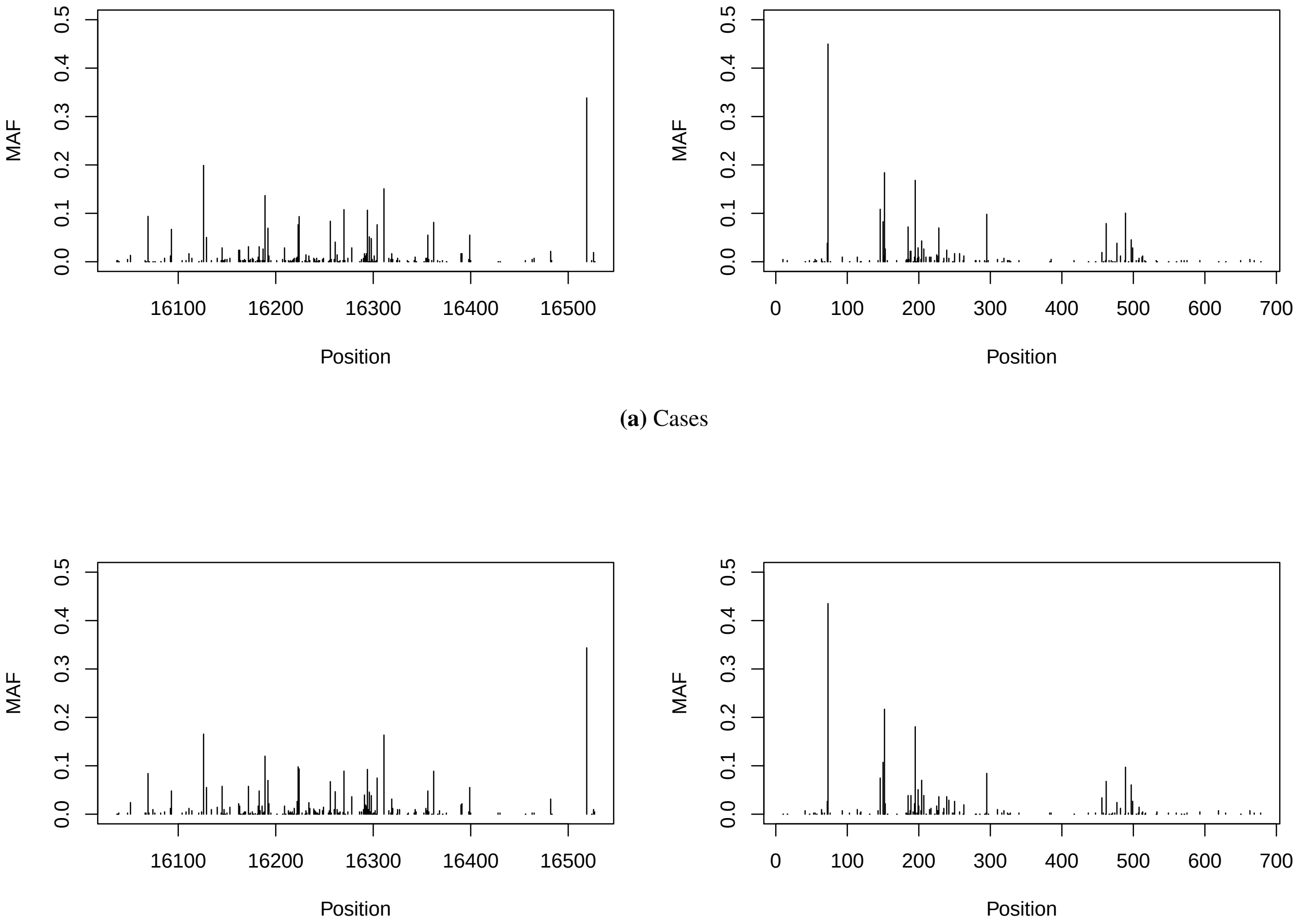
Supplemental Figures



Case Pool

Control Pool

Figure A.1: Sequence Coverage. Sequence reads were aligned to the rCRS (`NC_012920.1`) with MAQ. Median coverage was 13,134 reads (31.3 reads per sample) for the case pool, and 12,683 reads (30.6 reads per sample) for the control pool.



(a) Cases

(b) Controls

Figure A.2: Minor Allele Frequencies from Sanger Dataset. A total of 277 SNPs were identified by Sanger sequencing.

Appendix B

Mitochondrial Marker Data

Table B.1: Mitochondrial Marker Selection

Position	MAF case	MAF control	P value	rs Number	Reason for Inclusion
512	0.017	0.000	0.01524	NA	*P*-value < 0.050
675	0.165	0.162	0.92549	NA	MAF > 0.050
709	0.096	0.079	0.46286	rs2853517	Saxena *et al.* Tag SNP
750	0.014	0.028	0.16074	rs2853518	SAXENA_TAG
896	0.002	0.018	0.03740	NA	*P*-value < 0.050
930	0.048	0.035	0.49043	rs41352944	Saxena *et al.* Tag SNP
1,189	0.058	0.071	0.48061	rs28358571	Saxena *et al.* Tag SNP
3,010	0.223	0.184	0.16902	rs3928306	Saxena *et al.* Tag SNP
3,109	0.068	0.093	0.16225	NA	MAF > 0.050
3,348	0.001	0.004	0.24731	rs41423746	Saxena *et al.* Tag SNP
3,394	0.014	0.009	0.75245	rs41460449	Saxena *et al.* Tag SNP
3,505	0.004	0.026	0.01194	rs28358585	*P*-value < 0.050
3,849	0.001	0.011	0.03014	NA	*P*-value < 0.050
3,915	0.021	0.033	0.29868	rs41524046	Saxena *et al.* Tag SNP
4,336	0.016	0.012	0.77281	rs41456348	Saxena *et al.* Tag SNP
4,529	0.012	0.034	0.03835	NA	*P*-value < 0.050
4,769	0.035	0.052	0.24322	rs3021086	Saxena *et al.* Tag SNP
4,793	0.024	0.008	0.08982	NA	Saxena *et al.* Tag SNP
4,928	0.000	0.001	1.00000	rs41461545	Saxena *et al.* Tag SNP
5,426	0.010	0.017	0.38249	NA	Saxena *et al.* Tag SNP
5,465	0.000	0.000	1.00000	rs3902405	Saxena *et al.* Tag SNP
5,495	0.016	0.010	0.54615	rs3020602	Saxena *et al.* Tag SNP
5,656	0.017	0.012	0.77281	NA	Saxena *et al.* Tag SNP

Position	MAF case	MAF control	P value	rs Number	Reason for Inclusion
5,785	0.000	0.011	0.03014	NA	*P*-value < 0.050
5,981	0.002	0.016	0.03740	NA	*P*-value < 0.050
6,182	0.000	0.011	0.03014	NA	*P*-value < 0.050
6,260	0.012	0.017	0.57664	NA	Saxena *et al.* Tag SNP
6,272	0.000	0.011	0.03014	NA	*P*-value < 0.050
6,365	0.017	0.019	0.80102	rs41464546	Saxena *et al.* Tag SNP
6,719	0.003	0.002	1.00000	rs28358872	Saxena *et al.* Tag SNP
6,776	0.029	0.024	0.82959	NA	Saxena *et al.* Tag SNP
7,028	0.498	0.471	0.40672	rs2015062	Saxena *et al.* Tag SNP
8,251	0.027	0.057	0.02499	rs3021089	*P*-value < 0.050
8,269	0.021	0.031	0.39659	rs8896	Saxena *et al.* Tag SNP
8,303	0.003	0.016	0.03740	NA	*P*-value < 0.050
8,697	0.117	0.072	0.03293	rs28358886	*P*-value < 0.050
8,705	0.012	0.007	0.72527	NA	Saxena *et al.* Tag SNP
8,869	0.000	0.000	1.00000	NA	Saxena *et al.* Tag SNP
9,123	0.011	0.010	1.00000	rs28358270	Saxena *et al.* Tag SNP
9,150	0.013	0.007	0.72527	NA	Saxena *et al.* Tag SNP
9,477	0.067	0.068	1.00000	rs2853825	Saxena *et al.* Tag SNP
9,548	0.016	0.000	0.01524	NA	*P*-value < 0.050
9,667	0.011	0.018	0.57664	rs41482146	Saxena *et al.* Tag SNP
9,716	0.027	0.012	0.20577	rs41502750	Saxena *et al.* Tag SNP
9,899	0.020	0.013	0.57807	rs41345446	Saxena *et al.* Tag SNP
9,947	0.003	0.016	0.03740	NA	*P*-value < 0.050
10,034	0.016	0.030	0.25570	rs41347846	Saxena *et al.* Tag SNP
10,084	0.012	0.006	0.45119	rs41487950	Saxena *et al.* Tag SNP
10,296	0.091	0.081	0.71196	NA	MAF > 0.050
10,314	0.033	0.074	0.00901	NA	*P*-value < 0.050
10,398	0.161	0.185	0.35980	rs2853826	MAF > 0.050
10,915	0.002	0.011	0.12210	rs2857285	Saxena *et al.* Tag SNP
11,377	0.018	0.019	1.00000	NA	Saxena *et al.* Tag SNP
11,485	0.019	0.021	0.81184	rs28529320	Saxena *et al.* Tag SNP
11,674	0.003	0.012	0.12210	rs28358286	Saxena *et al.* Tag SNP
11,719	0.428	0.433	0.88880	rs2853495	Saxena *et al.* Tag SNP
11,812	0.071	0.060	0.57741	rs3088053	Saxena *et al.* Tag SNP
11,914	0.034	0.027	0.68557	rs2853496	Saxena *et al.* Tag SNP
12,007	0.010	0.029	0.04604	rs2853497	Saxena *et al.* Tag SNP
12,372	0.213	0.236	0.45499	rs2853499	Saxena *et al.* Tag SNP
12,414	0.005	0.017	0.10603	NA	Saxena *et al.* Tag SNP
12,501	0.017	0.041	0.03954	rs28397767	*P*-value < 0.050

Position	MAF case	MAF control	P value	rs Number	Reason for Inclusion
12,633	0.025	0.020	0.81254	rs3926883	Saxena *et al.* Tag SNP
12,705	0.037	0.060	0.15203	rs2854122	Saxena *et al.* Tag SNP
13,020	0.010	0.017	0.38249	rs75577869	Saxena *et al.* Tag SNP
13,105	0.005	0.010	0.44948	rs2853501	Saxena *et al.* Tag SNP
13,637	0.007	0.026	0.03296	NA	P-value < 0.050
13,706	0.022	0.000	0.00374	NA	P-value < 0.050
13,708	0.090	0.065	0.19650	rs28359178	Saxena *et al.* Tag SNP
13,734	0.007	0.013	0.50388	rs41421644	Saxena *et al.* Tag SNP
13,869	0.030	0.091	0.00026	NA	P-value < 0.050
13,870	0.020	0.069	0.00034	NA	P-value < 0.050
13,879	0.012	0.042	0.00937	NA	Saxena *et al.* Tag SNP
13,934	0.023	0.020	0.81254	NA	Saxena *et al.* Tag SNP
13,965	0.006	0.005	1.00000	rs41509754	Saxena *et al.* Tag SNP
13,966	0.014	0.012	1.00000	rs41535848	Saxena *et al.* Tag SNP
14,182	0.032	0.042	0.46312	NA	Saxena *et al.* Tag SNP
14,470	0.020	0.020	1.00000	rs3135030	Saxena *et al.* Tag SNP
14,793	0.047	0.054	0.75375	rs2853504	Saxena *et al.* Tag SNP
14,798	0.167	0.123	0.07692	rs28357681	Saxena *et al.* Tag SNP
15,022	0.018	0.048	0.02131	NA	P-value < 0.050
15,043	0.025	0.040	0.17600	rs28357684	Saxena *et al.* Tag SNP
15,218	0.039	0.049	0.50023	rs2853506	Saxena *et al.* Tag SNP
15,257	0.041	0.042	1.00000	rs41518645	Saxena *et al.* Tag SNP
15,511	0.001	0.014	0.01491	NA	P-value < 0.050
15,758	0.007	0.015	0.33904	rs41337244	Saxena *et al.* Tag SNP
15,775	0.000	0.014	0.01491	NA	P-value < 0.050
15,784	0.005	0.004	1.00000	rs28357375	Saxena *et al.* Tag SNP
15,833	0.014	0.007	0.50548	rs41504845	Saxena *et al.* Tag SNP
15,884	0.007	0.021	0.08866	rs28617642	Saxena *et al.* Tag SNP
15,924	0.059	0.068	0.67229	rs2853510	Saxena *et al.* Tag SNP
15,937	0.012	0.034	0.03835	NA	P-value < 0.050

Table B.2: Mitochondrial Marker Set for Sequenom Genotyping

ID	Chr	Position	Gene	MAF	Call Rate	AA	BB	AB	U	Project State
mt512	M	512	Non-coding	0.000	0.994	0	971	0	6	good
mt675	M	675	RNR1	0.000	0.998	975	0	0	2	good
rs2853517	M	709	RNR1	0.141	0.999	138	838	0	1	good
rs2853518	M	750	RNR1	0.023	0.992	22	947	0	8	good
mt896	M	896	RNR1	0.000	0.000	0	0	0	977	failed
rs41352944	M	930	RNR1	0.046	1.000	932	45	0	0	good
rs28358571	M	1,189	RNR1	0.074	1.000	72	905	0	0	good
rs3928306	M	3,010	RNR2	0.248	0.998	733	242	0	2	good
mt3109	M	3,109	RNR2	0.000	0.999	0	976	0	1	good
rs41423746	M	3,348	ND1	0.002	0.988	963	2	0	12	good
rs41460449	M	3,394	ND1	0.013	0.996	960	13	0	4	good
rs28358585	M	3,505	ND1	0.018	1.000	959	18	0	0	good
mt3849	M	3,849	ND1	0.000	0.000	0	0	0	977	failed
rs41524046	M	3,915	ND1	0.027	0.996	26	947	0	4	good
rs41456348	M	4,336	TRNQ	0.017	0.959	16	921	0	40	good
mt4529	M	4,529	ND2	0.000	0.724	707	0	0	270	good
rs3021086	M	4,769	ND2	0.035	0.995	34	938	0	5	good
mt4793	M	4,793	ND2	0.000	0.000	0	0	0	977	failed
rs41461545	M	4,928	ND2	0.000	0.999	976	0	0	1	good
mt5426	M	5,426	ND2	0.000	0.000	0	0	0	977	failed
rs3902405	M	5,465	ND2	0.001	0.766	747	1	0	229	good
rs3020602	M	5,495	ND2	0.001	1.000	976	1	0	0	good
mt5656	M	5,656	Non-coding	0.000	0.000	0	0	0	977	failed
mt5785	M	5,785	TRNC	0.000	0.982	0	959	0	18	good
mt5981	M	5,981	COX1	0.000	0.000	0	0	0	977	failed
mt6182	M	6,182	COX1	0.000	0.000	0	0	0	977	failed
rs28623747	M	6,260	COX1	0.018	1.000	959	18	0	0	good
mt6272	M	6,272	COX1	0.000	0.898	877	0	0	100	good
rs41464546	M	6,365	COX1	0.016	0.997	16	958	0	3	good
rs28358872	M	6,719	COX1	0.000	0.000	0	0	0	977	failed
mt6776	M	6,776	COX1	0.000	0.000	0	0	0	977	failed
rs2015062	M	7,028	COX1	0.429	1.000	558	419	0	0	good
rs3021089	M	8,251	COX2	0.049	0.987	917	47	0	13	good
rs8896	M	8,269	ATP6	0.025	0.999	24	952	0	1	good
mt8303	M	8,303	ATP6	0.006	0.998	6	969	0	2	good
rs28358886	M	8,697	ATP6	0.090	0.987	87	877	0	13	good
mt8705	M	8,705	ATP6	0.000	0.000	0	0	0	977	failed
mt8869	M	8,869	ATP6	0.000	0.000	0	0	0	977	failed

ID	Chr	Position	Gene	MAF	Call Rate	AA	BB	AB	U	Project State
rs28358270	M	9,123	ATP6	0.009	0.997	965	9	0	3	good
mt9150	M	9,150	ATP6	0.000	0.000	0	0	0	977	failed
rs2853825	M	9,477	COX3	0.104	0.992	101	868	0	8	good
rs41482146	M	9,667	COX3	0.016	1.000	961	16	0	0	good
rs41502750	M	9,716	COX3	0.016	0.996	957	16	0	4	good
rs41345446	M	9,899	COX3	0.017	1.000	17	960	0	0	good
mt9947	M	9,947	COX3	0.000	0.000	0	0	0	977	failed
rs41347846	M	10,034	TRNG	0.000	0.000	0	0	0	977	failed
rs41487950	M	10,084	ND3	0.000	0.000	0	0	0	977	failed
mt10296	M	10,296	ND3	0.000	0.998	0	975	0	2	good
mt10314	M	10,314	ND3	0.000	0.991	968	0	0	9	good
rs2853826	M	10,398	ND3	0.000	0.981	958	0	0	19	good
rs2857285	M	10,915	ND4	0.005	0.989	5	961	0	11	good
rs41537746	M	11,377	ND4	0.015	1.000	962	15	0	0	good
rs28529320	M	11,485	ND4	0.023	1.000	955	22	0	0	good
rs28358286	M	11,674	ND4	0.018	1.000	18	959	0	0	good
rs2853495	M	11,719	ND4	0.467	0.999	520	456	0	1	good
rs3088053	M	11,812	ND4	0.073	0.995	901	71	0	5	good
rs2853496	M	11,914	ND4	0.000	0.000	0	0	0	977	failed
rs2853497	M	12,007	ND4	0.020	0.829	16	794	0	167	good
rs2853499	M	12,372	ND5	0.250	0.988	724	241	0	12	good
rs41520546	M	12,414	ND5	0.009	0.992	960	9	0	8	good
rs28397767	M	12,501	ND5	0.030	1.000	29	948	0	0	good
rs3926883	M	12,633	ND5	0.000	0.000	0	0	0	977	failed
rs2854122	M	12,705	ND5	0.000	0.000	0	0	0	977	failed
rs75577869	M	13,020	ND5	0.000	0.000	0	0	0	977	failed
rs2853501	M	13,105	ND5	0.000	0.000	0	0	0	977	failed
mt13637	M	13,637	ND5	0.000	0.000	0	0	0	977	failed
mt13706	M	13,706	ND5	0.000	0.869	849	0	0	128	good
rs28359178	M	13,708	ND5	0.000	1.000	977	0	0	0	good
rs41421644	M	13,734	ND5	0.000	1.000	977	0	0	0	good
mt13869	M	13,869	ND5	0.000	1.000	0	977	0	0	good
mt13870	M	13,870	ND5	0.000	0.000	0	0	0	977	failed
mt13879	M	13,879	ND5	0.000	0.000	0	0	0	977	failed
mt13934	M	13,934	ND5	0.000	0.000	0	0	0	977	failed
rs41509754	M	13,965	ND5	0.007	0.997	7	967	0	3	good
rs41535848	M	13,966	ND5	0.020	0.999	19	957	0	1	good
mt14182	M	14,182	ND6	0.000	0.000	0	0	0	977	failed
rs3135030	M	14,470	ND6	0.000	0.000	0	0	0	977	failed

ID	Chr	Position	Gene	MAF	Call Rate	AA	BB	AB	U	Project State
rs2853504	M	14,793	CYTB	0.064	0.999	62	914	0	1	good
rs28357681	M	14,798	CYTB	0.154	0.988	816	149	0	12	good
mt15022	M	15,022	CYTB	0.000	0.000	0	0	0	977	failed
rs28357684	M	15,043	CYTB	0.038	0.999	37	939	0	1	good
rs2853506	M	15,218	CYTB	0.049	0.999	48	928	0	1	good
rs41518645	M	15,257	CYTB	0.025	0.999	952	24	0	1	good
rs35070048	M	15,311	CYTB	0.000	0.972	950	0	0	27	good
rs41337244	M	15,758	CYTB	0.011	1.000	11	966	0	0	good
mt15775	M	15,775	CYTB	0.000	0.000	0	0	0	977	failed
rs28357375	M	15,784	CYTB	0.006	1.000	971	6	0	0	good
rs41504845	M	15,833	CYTB	0.000	0.000	0	0	0	977	failed
rs28617642	M	15,884	CYTB	0.003	0.964	3	939	0	35	good
rs2853510	M	15,924	TRNT	0.060	0.953	875	56	0	46	good
mt15937	M	15,937	TRNT	0.000	0.602	0	588	0	389	good
rs55749223	M	16,189	Non-coding	0.003	0.953	3	928	0	46	good

Table B.3: Markers Repeated Due to Quality Control Failure

ID	Chr	Position	Gene	MAF	Call Rate	AA	BB	AB	U	Project State
mt896	M	896	RNR1	0.005	0.998	973	5	0	2	good
mt3849	M	3,849	ND1	0.007	0.999	972	7	0	1	good
mt4529	M	4,529	ND2	0.022	1.000	22	958	0	0	good
mt4793	M	4,793	ND2	0.018	0.994	17	957	0	6	good
mt5426	M	5,426	ND2	0.007	0.988	2	957	9	12	good
rs3902405	M	5,465	ND2	0.001	1.000	979	1	0	0	good
mt5656	M	5,656	Non-coding	0.020	1.000	20	960	0	0	good
mt5981	M	5,981	COX1	0.000	1.000	980	0	0	0	good
mt6182	M	6,182	COX1	0.003	1.000	977	3	0	0	good
mt6272	M	6,272	COX1	0.000	1.000	980	0	0	0	good
mt6719	M	6,719	COX1	0.000	0.994	0	974	0	6	good
mt6776	M	6,776	COX1	0.044	0.994	931	43	0	6	good
rs2015062	M	7,028	COX1	0.434	0.999	553	424	2	1	good
mt8705	M	8,705	COX2	0.007	0.994	967	7	0	6	good
mt8869	M	8,869	ATP6	0.001	0.999	1	978	0	1	good
mt9150	M	9,150	ATP6	0.011	1.000	969	11	0	0	good
mt9947	M	9,947	COX3	0.000	0.000	0	0	0	980	failed
rs41345446	M	9,899	COX3	0.000	0.000	0	0	0	980	failed
rs41347846	M	10,034	TRNG	0.000	0.000	0	0	0	980	failed
rs2853495	M	11,719	ND4	0.471	0.999	518	461	0	1	good
rs2853496	M	11,914	ND4	0.029	0.990	28	941	1	10	good
rs2853497	M	12,007	ND4	0.020	0.998	958	20	0	2	good
rs3926883	M	12,633	ND5	0.014	0.947	13	915	0	52	good
rs2854122	M	12,705	ND5	0.061	0.833	50	766	0	164	good
mt13020	M	13,020	ND5	0.012	1.000	12	968	0	0	good
rs2853501	M	13,105	ND5	0.012	1.000	11	968	1	0	good
mt13637	M	13,637	ND5	0.022	0.993	952	21	0	7	good
mt13706	M	13,706	ND5	0.000	1.000	0	980	0	0	good
mt13870	M	13,870	ND5	0.000	1.000	980	0	0	0	good
mt13879	M	13,879	ND5	0.020	0.998	958	19	1	2	good
mt13934	M	13,934	ND5	0.016	0.998	0	947	31	2	good
mt14182	M	14,182	ND6	0.044	0.976	914	42	0	24	good
rs3135030	M	14,470	ND6	0.021	0.995	955	20	0	5	good
mt15022	M	15,022	CYTB	0.000	0.995	975	0	0	5	good
mt15775	M	15,775	CYTB	0.001	1.000	978	0	2	0	good
rs41504845	M	15,833	CYTB	0.003	0.891	870	3	0	107	good
mt15937	M	15,937	TRNT	0.002	0.957	2	936	0	42	good