Theoretical and Statistical Approaches to Understand Human Mitochondrial DNA Heteroplasmy Inheritance

Passorn Wonnapinij

Dissertation submitted to the faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
In
Genetics, Bioinformatics, and Computational Biology

David C. Samuels (Committee Chair)
Edward J. Smith
Ina Hoeschele
Ronald M. Lewis
Allan W. Dickerman

April 9, 2010
Blacksburg, VA

Keywords: mitochondrial genetic bottleneck, Kimura distribution, sampling error, mutation level variance, offspring gender bias

Copyright © 2010 Passorn Wonnapinij
unless otherwise stated
Theoretical and Statistical Approaches to Understand Human Mitochondrial DNA Heteroplasmy Inheritance

Passorn Wonnapeinij

ABSTRACT

Mitochondrial DNA (mtDNA) mutations have been widely observed to cause a variety of human diseases, especially late-onset neurodegenerative disorders. The prevalence of mitochondrial diseases caused by mtDNA mutation is approximately 1 in 5,000 of the population. There is no effective way to treat patients carrying pathogenic mtDNA mutation; therefore preventing transmission of mutant mtDNA became an important strategy. However, transmission of human mtDNA mutation is complicated by a large intergenerational random shift in heteroplasmy level causing uncertainty for genetic counseling. The aim of this dissertation is to gain insight into how human mtDNA heteroplasmy is inherited.

By working closely with our experimental collaborators, the computational simulation of mouse embryogenesis has been developed in our lab using their measurements of mouse mtDNA copy number. This experimental-computational interplay shows that the variation of offspring heteroplasmy level has been largely generated by random partition of mtDNA molecules during pre- and early post-implantation development.

By adapting a set of probability functions developed to describe the segregation of allele frequencies under a pure random drift process, we now can model mtDNA heteroplasmy distribution using parameters estimated from experimental data.

The absence of an estimate of sampling error of mtDNA heteroplasmy variance may largely affect the biological interpretation drawn from this high-order statistic, thereby we have developed three different methods to estimate sampling error values for mtDNA heteroplasmy variance. Applying this error estimation to the comparison of mouse to human mtDNA heteroplasmy variance reveals the difference of the mitochondrial genetic bottleneck between these organisms.

In humans, the mothers who carry a high proportion of m.3243A>G mutation tend to have fewer daughters than sons. This offspring gender bias has been revealed by applying basic statistical tests on the human clinical pedigrees carrying this mtDNA mutation. This gender bias may partially determine the mtDNA mutation level among female family members.

In conclusion, the application of population genetic theory, statistical analysis, and computational simulation help us gain understanding of human mtDNA heteroplasmy inheritance. The results of these studies would be of benefit to both scientific research and clinical application.
Dedication

To my parents,
Pan-Apinya Wonnapinij,
and brother
Prapass Wonnapinij
Acknowledgements

All studies presented in this dissertation would not be possible without help and support from many people. First, I would like to thank my adviser, Dr. Davild C. Samuels, for his invaluable advice, and scientific temperament. Being his student is not only a wonderful experience but also giving me a nice opportunity to meet and work with other scientists and clinicians in the field of mitochondrial genetics. Working with him is a wonderful experience at possibly the best ever.

I would like to thank my experimental and clinical collaborator, Dr. Patrick F. Chinnery whose biological and clinical point of views always drives the study to the better one. I also would like to thank all other experimental collaborators as well as the scientists who came across continent to visit our research group. Having a chance to work and discuss with them help me gain scientific knowledge about mitochondrial biology and genetics.

I would like to show my gratitude to all of my committee members for their suggestions and supports both for my academic study and scientific research.

I am thankful to all of my colleagues in Dr. Samuels’ research group: Zhuo Song, Vishal Gandhi and Mark Lee, for their kindness and warm support, especially Song. I also thank all my former lab members: Dr. Jonghoon Kang, Dr. Hasha Karur Rajasimha, and Katherine Wendelsdorf, for their helps and suggestions. Special thanks to Dr. Harsha, his advice was not only useful for my research project but also for my academic life.

I thank all faculty, staff and student at Virginia Tech, especially GBCB students. I thank Dr. Surot Thangjitham for all his help. Special thanks to Dr. David Bevan, his attitude helps me moving to Nashville with a peaceful mind. Many thanks to Ms. Dennie Munson, she has been very helpful in every senses from the beginning. Her helps for all administrative and graduate school paperwork only make my life easier as a foreign PhD student.

I thank all of faculty, staff, and student both of Virginia Bioinformatics Institute (VBI) at Virginia Tech and of Center of Human Genetic Genetics Research (CHGR) at Vanderbilt University for their warm welcome and assistance whenever I needed.

I am gratified by an opportunity to study abroad at Virginia Tech which would not be possible without funding from the Commission on Higher Education, Royal Thai Government. It would not be too much to say that my financial support comes from all Thais; thereby I would like to thank them all.

I thank all Thai friends, brothers and sisters at Virginia Tech, special thank to Ms. Kanokwan Nontapot, Ms. Boonta Chutvirasakul, and Ms. Monrudee Liangruksa, for their support both for my academic and personal life. I also thank my roommates and friends in Nashville, Mr. Lertop Supadhiloke, Mr. Steven Chen, Ms. Younhee Rye, and Ms. Pinkwan Jaru-ampornpan, for their warm welcome and help, while I live in Nashville.
Last but not least, I would like to show my deeply gratitude to all of my family members in Thailand. I thank my elder brother, Mr. Prapass Wonnapanij, for his practical advice and support. Many thanks to my parents, Pan and Apinya Wonnapanij, I am gratified by being born from these incredible persons. I cannot achieve any goal of my life without their unconditional love, invaluable advice, and warm support. My gratitude to them is ways beyond my words.
<table>
<thead>
<tr>
<th>Chapter</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4. Chapter 4: Previous Estimation of Mitochondrial DNA Mutation Level</td>
<td>45</td>
</tr>
<tr>
<td>Variance Did not Account for Sampling Error: Comparing the MtDNA Genetics Bottleneck in Mice and Humans</td>
<td></td>
</tr>
<tr>
<td>Abstract</td>
<td>46</td>
</tr>
<tr>
<td>Introduction</td>
<td>46</td>
</tr>
<tr>
<td>Material and methods</td>
<td>47</td>
</tr>
<tr>
<td>Results</td>
<td>47</td>
</tr>
<tr>
<td>Discussion</td>
<td>54</td>
</tr>
<tr>
<td>References</td>
<td>55</td>
</tr>
<tr>
<td>5. Chapter 5: The Mother Carrying a High Pathogenic A3243G Mutation Tends to Have Fewer Daughters than Sons</td>
<td>57</td>
</tr>
<tr>
<td>Abstract</td>
<td>58</td>
</tr>
<tr>
<td>Introduction and discussion</td>
<td>59</td>
</tr>
<tr>
<td>Method summary</td>
<td>63</td>
</tr>
<tr>
<td>References</td>
<td>65</td>
</tr>
<tr>
<td>Figures legends</td>
<td>66</td>
</tr>
<tr>
<td>Online method</td>
<td>67</td>
</tr>
<tr>
<td>References</td>
<td>69</td>
</tr>
<tr>
<td>Figures</td>
<td>72</td>
</tr>
<tr>
<td>Supplementary information</td>
<td>72</td>
</tr>
<tr>
<td>6. Chapter 6: Conclusion and Future Directions</td>
<td>79</td>
</tr>
<tr>
<td>Section 6.1 Mitochondrial genetic bottleneck</td>
<td>79</td>
</tr>
<tr>
<td>Section 6.2: Distribution pattern of mtDNA heteroplasmy level</td>
<td>79</td>
</tr>
</tbody>
</table>
List of Figures

1 Chapter 1: Introduction

Figure 1: The simple structure of mitochondrion

Figure 2: Schematic diagram of human mitochondrial DNA showing all 37 genes encoded in this organelle genome including some common pathogenic mtDNA mutations

Figure 3: Transmission of mtDNA heteroplasmy level during prenatal development of female germ line. After fertilization, zygote undergoes multiple cell divisions without mtDNA replication to generate blastocyst causing a small copy numbers of mtDNA molecules per blastomere. At implantation, some blastomeres are recruited to be primordial germ cells (PGCs). These PGCs experience serial cell divisions to generate a population of primary oocytes. The mtDNA replication is resumed at an early post-implantation development. The random partition of mtDNA molecules during this prenatal development of female germ lines and stochastic replication of mtDNA molecules during post-implantation development generate variation of heteroplasy level in the population of primary oocytes.

2 Chapter 2: A Reduction of Mitochondrial DNA Molecules During Embryogenesis Explains the Rapid Segregation of Genotype

Figure 1: The amount of mtDNA in mature mouse oocytes and pre-implantation mouse embryos. (a) Mature oocytes and whole embryos. Box and whisker diagram showing the median, ± 1 s.d., and the range from maximum to minimum values. (b) Single blastocyst cells (blastomeres) from pre-implantation embryos. Each data point corresponds to a single blastomere. (c) Primordial germ cells (PGCs) from post-implantation mouse embryos on a logarithmic plot. Each data point corresponds to a single PGC. Horizontal line indicates median value. Data from male and female embryos is shown separately for 14.5 d.p.c. Blue circles, males; red triangles, females.

Figure 2: Models of the mitochondrial genetic bottleneck. Schematic diagram showing a heteroplagomic fertilized oocyte (top), a model of the mitochondrial genetic bottleneck (middle) and subsequent primary oocytes (bottom). Blue circles, wild-type mtDNA; red circles, mutated mtDNA. Time scale shown on the left in d.p.c. (a) Single-sampling model. A single random sample of mtDNA molecules is assumed to repopulate each primary oocyte. (b) Multiple sampling model. Using an adaptation of the population genetic model of Sewall Wright, this approach assumes that an identical moderate genetic bottleneck is present over multiple cell divisions, G. In mice, G = 15, which is the number of divisions required to produce the 25,791 primary oocytes present by 13.5 d.p.c. from a single blastomere. (c) Complex biological model based upon the number of cell divisions leading up to the formation of 40 primordial germ cells (PGCs) at 7.25 d.p.c., followed by 9–10 cell divisions required to generate the 25,791 PGCs required to form a full complement of primary oocytes (for details see Supplementary Fig (1).
Supplementary information

Supplementary Figure 1: Comparison between the number of cells simulated and the actual number of cells observed experimentally. The rate of cell division in pre-implantation embryos was derived from Ref\(^4\), and for post-implantation embryos from Ref\(^2,3\), base upon published values for alkaline phosphatase stained primordial germ cells (PGCs). PGCs are first discernable at 7.25 days post coitus (dpc) during the mid primitive streak stage, just posterior to the primitive streak in the extra-embryonic mesoderm. Note the logarithmic Y-axis indicating exponential growth to a final figure of 25,791 primary oocytes at day 13.5\(^\circ\)………………. 27

Supplementary Figure 2: The total amount of mtDNA in the simulated mouse embryos to 7 dpc, followed by the total amount of mtDNA in the entire PGC population. Dpc = days post coitus……………………………….. 28

Supplementary Figure 3: Comparison of three real-time PCR assays targeting different regions of the mitochondrial genome: ND5 (nt12789 to nt12876), ND4 (nt1031 to nt11174), and ND1 (nt2751 to nt3709). Total genomic DNA was extracted from the tail of a C57Bl6 mouse and serially diluted. Each dilution was assayed in quadruplicate using the same methods employed in the manuscript to determine the absolute amount of mtDNA in each sample. The graphs shows two independent serial dilutions of a genomic DNA template serially diluted and measured by two different assays: (a) comparison of ND4 with ND5, (b) comparison of ND1 with ND5, (c) comparison of ND4 with ND1. Pearson’s correlation co-efficient showed a tight correlation in the measured copy number values using each assay (R\(^2\) > 0.99). The ND5 assay was used in the experiments described in the manuscript…………………………………. 29

Chapter 3: The Distribution of Mitochondrial DNA Heteroplasy Due to Random Genetic Drift……………………………………………………… 32

Figure 1: The heteroplasmy distribution of the A3243G mtDNA mutation in a sample of human primary oocytes is compared to the Kimura distribution (A) Frequency histogram of the heteroplasmy in both the data and the Kimura distribution fit to the data. Parameter values for the Kimura distribution are given in
Table 1 for all figures. (B) Cumulative probability distribution functions for the data and the Kimura distribution fit to the data. A KS test indicates that there is no significant difference between the measured and the theoretical probability distributions.  

Figure 2: The measured heteroplasmy distribution from offspring and mature oocytes in the heteroplasmic mouse line 515 is compared to the Kimura distribution (A) The heteroplasmy frequency histogram from the offspring. (B) KS test comparing the offspring heteroplasmy data to the Kimura distribution fit to the data. (C) The heteroplasmy frequency histogram of the mature oocytes in the 515 line. (D) KS test for the line 515 mature oocyte data. There is a significant difference between the two distributions in the mature oocyte data.  

Figure 3: The measured heteroplasmy distribution from offspring and mature oocytes in the heteroplasmic mouse line 517 is compared to the Kimura distribution (A) The heteroplasmy frequency histogram from the offspring. (B) KS test comparing the offspring heteroplasmy data to the Kimura distribution fit to the data. (C) The heteroplasmy frequency histogram of the mature oocytes in the 517 line. (D) KS test for the data from line 517 mature oocytes.  

Figure 4: The measured heteroplasmy distribution from mature oocytes and primary oocytes in the heteroplasmic mouse line 603A is compared to the Kimura distribution (A) The heteroplasmy frequency histogram of the mature oocytes in the 603A line. (B) KS test for the data from line 603A mature oocytes. (C) The heteroplasmy frequency histogram of the primary oocytes in the 603A line. (D) KS test for the data from line 603A primary oocytes. There is a significant difference between the two distributions for the primary oocyte data.  

Figure 5: The measured heteroplasmy distribution from mature oocytes and primary oocytes in the heteroplasmic mouse line 603B is compared to the Kimura distribution (A) The heteroplasmy frequency histogram of the mature oocytes in the 603B line. (B) KS test for the data from line 603B mature oocytes. (C) The heteroplasmy frequency histogram of the primary oocytes in the 603B line. (D) KS test for the data from line 603B primary oocytes.  

Figure 6: The measured heteroplasmy distribution from unfertilized eggs in the heteroplasmic Drosophila mauritiana lines H1, G20-5, and G71-12 is compared to the Kimura distribution (A) The heteroplasmy frequency histogram of the Drosophila line H1 and the Kimura distribution fit to the mean and variance values from these data. (B) The KS test comparing the data with the Kimura distribution. There is a significant difference between the two distributions for line H1. (C) The heteroplasmy frequency histogram for the Drosophila line G20-5 is compared to the Kimura distribution. (D) KS test comparing the data for Drosophila line G20-5 to the Kimura distribution. (E) The heteroplasmy frequency histogram from the Drosophila line G71-12 is compared to the Kimura distribution. (F) KS test comparing the data for Drosophila line G71-12 to the Kimura distribution.
Figure 7: The measured heteroplasmy distribution from unfertilized eggs in the heteroplasmic *Drosophila mauritiana* lines H1-31M, H1-18D, and H1-12B is compared to the Kimura distribution (A) The heteroplasmy frequency histogram from the *Drosophila* line H1-31M and the Kimura distribution. (B) The KS test comparing the heteroplasmy data for *Drosophila* line H1-31M to the Kimura distribution. (C) The heteroplasmy frequency histogram from the *Drosophila* line H1-18D is compared to the Kimura distribution. (D) KS test comparing the *Drosophila* line H1-18D to the Kimura distribution. (E) The heteroplasmy frequency histogram from *Drosophila* line H1-12B is compared to the Kimura distribution. (F) KS test comparing data for *Drosophila* line H1-12B and the Kimura distribution………………….. 41

Figure 8: Comparison of measured heteroplasmy distribution from *Drosophila simulans* unfertilized eggs with the Kimura distribution (A) Heteroplasmy frequency histogram and the Kimura distribution fit to these data. (B) KS test comparing the data to the Kimura distribution……………………………….. 42

Figure 9: Comparison of a Kimura distribution, Normal distribution, and Binomial distribution with Mean = 0.1 and Variance = 0.01 (A) Kimura distribution. The probability density $\phi(x)$ is plotted. (B) Normal distribution. (C) Binomial distribution. The mean and variance values require a range of discrete states from zero to nine, giving discrete probability values of 0, 1/9, 2/9, etc……………. 42

4 Chapter 4: Previous Estimates of Mitochondrial DNA Mutation Level Variance Did Not Account for Sampling Error: Comparing the MtDNA Genetic Bottleneck in Mice and Humans………………………………………………….. 45

Figure 1: Measurements of the mean and variance from samples drawn from a Normal distribution. (A) The Normal distribution used with mean = 0.5 and variance = 0.01. (B) Mean values as a function of the sample size N ranging from 3 to 100. The error bars were set to twice the Standard Error of the Mean as calculated from Equation 1. (C) Values of variance as a function of the sample size N. The error bars were set to twice the Standard Error of the Variance for a Normal distribution as calculated from Equation 6. The 95% confidence intervals were determined from the mean and variance values from 10,000 samples of size N……………………………………………………………………….. 48

Figure 2: Measurements of the mean and variance from samples drawn from a Kimura distribution with moderate mean value. (A) The Kimura distribution $\phi(p)$ used with mean $p_0 = 0.5$ and $b = 0.9$. (B) Mean values as a function of the sample size N ranging from 3 to 100. The error bars were set to twice the Standard Error of the Mean as calculated from Equation 1. (C) Values of variance as a function of the sample size N. The error bars were set to twice the Standard Error of the Variance for a Kimura distribution as calculated from Equations 11 and 2. The 95% confidence intervals were determined from the mean and variance values from 10,000 independent samples of size N………………………………………………………………………………. 50
Figure 3: Measurements of the mean and variance from samples drawn from a Kimura distribution with an extreme mean value. (A) The Kimura distribution $\phi(p)$ used with mean $p_0 = 0.1$ and $b = 0.9$. (B) Mean values as a function of the sample size $N$ ranging from 3 to 100. The error bars were set to twice the Standard Error of the mean as calculated from Equation 1. (C) Values of variance as a function of the sample size $N$. The error bars were set to twice the Standard Error of the Variance for a Kimura distribution as calculated from Equations 11 and 2. The 95% confidence intervals were determined from the mean and variance values from 10,000 samples of size $N$.

Figure 4: Application of the Standard Error of Variance to data from human and mouse models. (A) Heteroplasmic mouse model data from Jenuth et al.\textsuperscript{12} (circles) at four stages of mtDNA inheritance; primordial germ cells (PGC), primary oocytes, mature oocytes and offspring. Human data (stars) from Brown et al.\textsuperscript{13} for primary oocytes and from numerous sources\textsuperscript{14-36} for offspring data are compared to the mouse data. (B) mtDNA mutation level variance with error bars measured in 21 mouse lineages. All error bars are twice the Standard Error calculated from a Kimura distribution. Variance values are normalized by dividing by $p_0(1-p_0)$.

Figure 5: mtDNA mutation level variance with error bars in a mouse model of the post-natal development of oocytes. The data are taken from Wai et al.\textsuperscript{10} and all error bars are twice the Standard Error of Variance calculated from a Kimura model. Variance values are normalized by dividing by $p_0(1-p_0)$.

Figure 6: Levene test $p$-values for comparisons of two data sets with different variances but equal mean mutation levels of 0.5. Both data sets were drawn randomly from a Kimura distribution. The distribution for the first data set was set to have $b=0.9$ while the value of $b$ for the second distribution was lowered according to Eq. 10 to give the stated variance difference. $P$-values were calculated using the standard Levene test. The horizontal line indicates a $p$-value of 0.05.

Figure 7: Levene test $p$-values for comparisons of two data sets with different variances but equal mean mutation levels of 0.1. Other details are the same as in Figure …

Figure 8: Dependence of the Standard Error of variance divided by the variance on the mean mtDNA mutation level, $p_0$ and on the sample size, $N$. (A) Dependence on the sample size $N$ for three values of the mean mutation level. (B) Dependence on the mean mutation level for a range of sample sizes. All Standard Error values are calculated for a Kimura distribution from Equations 11 and 2. All curves were calculated from Equations 11 and 2 with a value of $b = 0.9$. That $b$ value was chosen as a simple value that was close to the $b$ value calculated from the one human oocyte data set\textsuperscript{13,38}.

Chapter 5: The Mother Carrying a High Pathogenic A3243G Mutation Tends to Have Fewer Daughters than Sons.
Figure 1 The intergenerational transmission of A3243G mutation level (a) The average mutation level per generation among females was compared across generations from the grandmother’s to the most recent generation. The average mutation level was calculated either from blood or age-corrected blood mutation level (b) The average mutation level per generation among males was compared across generations from the grandmother’s to the most recent generation. The average mutation level was calculated either from blood or age-corrected blood mutation level (c) The average age-corrected blood O-M analysis..................... 72

Figure 2: The comparison of the average number of daughters to the average number of sons within the group of mothers carrying low mutation level (maternal age-corrected blood mutation level ≤ 50%) and the group of mothers carrying high mutation level (maternal age-corrected blood mutation level > 50%) ............ 73
List of Tables

1 Chapter 2: A Reduction of Mitochondrial DNA Molecules During Embryogenesis Explains the Rapid Segregation of Genotypes………………………….. 17

Table 1: Amount of mtDNA in mature oocytes and pre-implantation mouse embryos………………………………………………………………………. 20

Table 2: Amount of mtDNA in primordial germ cells at different embryonic stages in mice………………………………………………………………………... 21

Supplementary information

Supplementary Table 1: Variation in the level of heteroplasm between the offspring of heteroplasmic female mice………………………………………. 25

Supplementary Table 2: Oligonucleotide primer sequences used to quantify NZB/C57BL.6J mtDNA heteroplasm………………………………………………….. 26

3 Chapter 3: The Distribution of Mitochondrial DNA Heteroplasmy Due to Random Genetic Drift……………………………………………………………. 32

Table 1: The parameters estimated from experimental data: The mean heteroplasmy, p0, and the b parameter calculated from the variance and the p value calculated from the KS test………………………………………………….. 35

5 Chapter 5: The Mother Carrying a High Pathogenic A3243G Mutation Tends to Have Fewer Daughters than Sons............................................................... 57

Supplementary information

Supplementary Table 1: The human pedigree data used in the analysis of A3243G heteroplasm transmission and the effect on female reproduction…………… 74
Preface

The main purpose of this dissertation is to gain insight into how human mtDNA heteroplasmy level is maternally transmitted. Several approaches have been applied to achieve this goal in which each project has been done in a collaborative manner, especially for the project presented in Chapter 2. My chairman, Dr. David C. Samuels, has played an important role in every single project presented in this dissertation. His invaluable suggestions have directed each study to its goal. Each manuscript would not gain recognition without timely discussion with Dr. Patrick F. Chinnery. He has played a role as my experimental and clinical collaborator. He provided the human primary oocyte mtDNA heteroplasmy measurements used in the study presented in Chapter 3.

The thorough study of mouse mtDNA heteroplasmy transmission during its embryogenesis presented in Chapter 2 would not be successfully done without a broad collaboration with several research groups. Dr. Patrick F. Chinnery and Dr. Lynsey M. Cree from Newcastle University helped with the designing of the overall experimental study. Dr. Lynsey M. Cree measured mtDNA copy number during mouse embryogenesis from the GFP-Stella mice. The GFP-Stella mouse model was developed by Dr. Susana Chuva de Sousa Lopes from University of Cambridge. The mouse embryogenesis simulation model was designed by Dr. David C. Samuels and programmed by Dr. Hasha Karur Rajasimha from our research group. Dr. David C. Samuels and I carried out this computational simulation model to study how the germ line heteroplasmy variation has been generated during mouse embryogenesis process. We calculated the mtDNA heteroplasmy variance of each stage of mouse embryogenesis simulation and plot the value against the developmental stages (dpc) as shown in Figure 3C in Chapter 2. We also calculated the 95% confident interval of the offspring heteroplasmy level from the simulation. This 95% confident interval was compared to the actual heteroplasmic mouse measurements. These heteroplasmic mice were developed by Dr. Hans-Henrik M. Dahl and Dr. Jeffrey R. Mann from the University of Melbourne.
Chapter 1: Introduction

Chapter 1 aims to give a brief introduction related to human mitochondrial DNA (mtDNA) heteroplasmy inheritance. In particular, this chapter introduces a general knowledge about mitochondrial biology and genetics, mitochondrial diseases caused by pathogenic mtDNA mutation, the prevalence of mitochondrial diseases and pathogenic mtDNA mutation in the human population, and the transmission of human mtDNA heteroplasmy level. It also introduces a basic concept in population genetic theory and how to apply it to understand the segregation pattern of human mtDNA heteroplasmy level. The application of statistical analysis and computational simulation to understand the transmission of human mtDNA heteroplasmy level is also briefly mentioned in this chapter. The last section of this chapter, research motivations, discusses why we do all the studies presented in this dissertation.

Section 1.1: Mitochondrial biology

A mitochondrion is a small-dynamic organelle localized in the cytoplasm of a eukaryotic cell. This organelle is composed of two compartments, intermembrane space and matrix. The two main compartments are separated by two membranes, an outer and an inner membrane. The intermembrane space surrounds by the outer and the inner membrane. The matrix surrounds by the inner membrane. The inner membrane convolutes into the matrix to form the structure called cristae [1]. Mitochondrial DNA (mtDNA) is localized in the matrix. Figure 1 presents a diagram of the simple structure of the mitochondrion.

Figure1: The simple structure of the mitochondrion

Each compartment and membrane of mitochondrion contains a unique set of proteins functioning in different cellular mechanisms; for example all enzymes functioning in the citric acid cycle are localized in the matrix, and the transport proteins functioning in the oxidative phosphorylation process (OXPHOS) are embedded in the inner membrane. Although the schematic picture of the mitochondrion is generally presented as a static bean shape structure resembling bacteria as shown in Figure 1, this organelle is actually highly dynamic in that it moves in the cytoplasm using actin and myosin motors, changes its shape, and undergoes
frequent fission-fusion processes [2-6]. The dynamics of mitochondrial morphology are associated with its function [7-8].

Mitochondria play a key role in many cellular mechanisms such as amino acid metabolism, citric acid cycle, fatty acid catabolism, programmed cell death (apoptosis) and aging [9-12]. Mitochondrial dysfunction has also been observed in association with human diseases such as Parkinson’s disease [13-16], Alzheimer’s disease [17] and cancer [18-21]. The largest category of human diseases caused by mitochondrial gene mutations, located both in nuclear and mitochondrial genomes, is metabolic disease. There are also a large number of human diseases caused by mitochondrial gene mutations that affect multiple-tissue functions and cannot be classified [22]. Perhaps the most important function of mitochondria is as the center of Fe/S protein synthesis because this protein is essential for functional protein synthesis machinery [23].

The two critical roles of mitochondria involve OXPHOS, to produce cellular energy, and to regulate the apoptosis process, programmed cell death. The OXPHOS mechanism produces energy by transferring electrons generated from dietary calories through complex proteins located in the mitochondrial inner membrane. There are four complexes functioning in the electron transport chain: complex I (NADH dehydrogenase; EC 1.6.5.3), complex II (succinate dehydrogenase; EC 1.3.5.1) complex III (cytochrome bc1 complex; EC 1.10.2.2), and complex IV (cytochrome c oxidase; EC 1.9.3.1) [1, 9]. An electron is transported from complex I to complex IV creating a proton gradient across the mitochondrial inner membrane. This proton motive force drives complex V (ATP synthase) to produce cellular energy or adenosine triphosphate (ATP) from adenosine diphosphate (ADP), coupled with proton flow into the mitochondrial matrix [1, 9]. The OXPHOS process also produces toxic by-products called reactive oxygen species (ROS) which are harmful to both nuclear DNA (nDNA) and mitochondrial DNA (mtDNA) in the sense that they can mutate the DNA molecule. ROS is also hypothesized to be a major factor inducing apoptosis, causing progression of age-related diseases, and generating cancer [24].

**Section 1.2: Mitochondrial genetics**

Mitochondria have their own genetic material separated from the nuclear DNA (nDNA), called mitochondrial DNA (mtDNA). The size and the number of genes encoded in the mitochondrial genome are highly varied among different eukaryotes [25]. Eukaryotic mtDNA is mostly observed as a circular molecule; however a linear form has been observed as well [1]. In most organisms, mtDNA is maternally inherited; however the paternal leakage of mtDNA also exists in a variety of organisms [26-29].

Based on a serial endosymbiosis theory, the mitochondrion was a free-living bacterium that was engulfed by a nucleus containing eukaryote and developed an endosymbiotic relationship with the host cell. This serial theory has been challenged by the studies of some unicellular organisms. The results indicated the possibility that the mitochondrion may originate in a contemporary eukaryotic ancestor at approximately the same time as the nucleus [30]. The DNA sequence comparison had been applied to specify the bacterial ancestor of the mitochondrion. The protein sequence analysis result indicated that the mitochondrion is evolved from a purple photosynthetic bacterium in which its ability of photosynthesis had been previously lost [1]. During the course of evolution, many proto-mitochondrial genes were transferred to the nucleus of the eukaryotic cell. Different organisms have maintained different
genes in their mitochondrial genome; however the principle set of genes are maintained: some ribosomal RNA (rRNA) genes, some transfer RNA (tRNA) genes, some polypeptide-encoding genes, and a non-coding control region containing an origin of replication and promoters [31-32]. Therefore complex proteins functioning in a mitochondrion, especially the complex proteins associated with the OXPHOS, are encoded both in nuclear and mitochondrial genomes.

Human mtDNA is a 16,569 bp double-stranded circular molecule, almost exclusively transmitted through the maternal lineage with only a single case of paternal leakage reported [29]. The heavy strand (H strand) mtDNA is a guanine rich stand, while the light strand (L strand) is a cytosine rich strand. The human mitochondrial genome has a higher gene density than the nuclear genome because it codes for 37 genes with no introns in the coding region and only approximate 1 kb for non-coding region called D-loop. The D-loop contains essential components for mtDNA replication and transcription: H-strand origin of replication (O_H) and two promoters for both H and L strand mtDNA transcription. The 37 genes in the mitochondrial genome code for 13 polypeptides, 22 tRNAs and 2 rRNAs. All protein-coding genes are evenly distributed through the mitochondrial genome, interspersed by 22 tRNAs coding genes with 2 rRNAs coding genes that are sequentially located next to the D-loop as shown in Figure 2 [33]. All 13 polypeptides are essential subunits of transport proteins in OXPHOS process: 7 subunits for complex I, 1 subunit for complex III, 3 subunits for complex IV, and 2 subunits for complex V or ATP synthase [24]. All other OXPHOS protein subunits and other mitochondrial proteins are encoded in nuclear DNA (nDNA). Based on recent estimation, there are approximately 1,500 genes products contributing to the mitochondrial proteome, indicating that most of mitochondrial proteins are encoded in the nuclear genome [34]. The 22 tRNAs and 2 rRNAs are essential for intra-mitochondrial protein translation mechanism.

![Figure 2: Schematic diagram of human mitochondrial DNA showing all 37 genes encoded in this organelle genome including some common pathogenic mtDNA mutations](image)
Generally, different cell types contain different numbers of mitochondria. This number ranges from several hundred to several thousand mitochondria per cell, basically determined by energy requirement of the cell. Each mitochondrion contains multiple copies of mtDNAs ranging from 2 to 10 molecules [9]. The multiple copies of mtDNAs are not randomly distributed. They are rather distributed to several clusters in the matrix of each mitochondrion. Several mitochondrial DNAs in each cluster are packed together in a nucleoprotein complex structure called a nucleoid [35-37]. Each nucleoid is attached to the inner membrane of each mitochondrion [1]. Taking into account both the multiple number of mitochondria per cell and the multiple copies of mtDNAs per mitochondrion, a single eukaryotic cell contains a large copy number of mtDNAs. The variation of mtDNA copy numbers ranges from several hundred to several hundred thousand mtDNA molecules per cell; for example the mtDNA copy number in a human primary oocyte is approximately 100,000 molecules.

The state of a single cell containing multiple copies of mtDNAs is called polyploidy. In the sense of the content of mtDNAs in a single cell or a single tissue, if it contains only one type of mtDNA, this mitochondrial genetic condition is called homoplasy. On the contrary, if a single cell or tissue contains a mixture of two or more types of mtDNA, this mitochondrial genetic condition is called heteroplasy. Typically, the tissue sample of the patient diagnosed as carrying mitochondrial disease causing by the pathogenic mtDNA mutation contains both mutant and wild type mtDNAs, the heteroplasmic condition. The quantity of heteroplasmic is measured as the proportion of the mutant mtDNAs to the total mtDNAs, thereby the heteroplasmic level ranges from 0% (wild type homoplasy) to 100% (mutant homoplasy).

Section 1.3: Mitochondrial diseases caused by mtDNA mutations

Clinical features of mitochondrial disease caused by a pathogenic mtDNA mutation are generally presented as adult-onset with a progressive course similar to an age-related disease phenotype. Pathogenic mtDNA mutations causing disease presents both as a multi-systemic and a tissue-specific disease [38]. Typically, disease-related tissues are among the ones requiring high energy levels to maintain their normal function such as brain, heart, skeletal muscle and kidney. Specific symptoms include blindness, deafness, movement disorder, cardiovascular disease and renal dysfunction [24]. The accumulation of mutated mtDNA in specific cell or tissue with age is proposed as an explanation of the adult-onset with progressive course of disease associated with mtDNA mutations [24]. Some pathogenic mtDNA mutations have been observed to cause a reduction of human reproductive fitness; for example the m.3243A>G mutation can cause male infertility by slowing sperm motility [39]. This mutation has been suspected to increase a risk of spontaneous abortion among females carrying this mutation [40-45], although one study indicated that there is non-significant difference of female reproductive fitness between carriers and the general population [46].

Pathogenic mtDNA mutation can be classified into three categories based on the type of mtDNA mutation and its position in the mitochondrial genome: mtDNA rearrangement, mtDNA point mutation in protein coding region, and mtDNA point mutation in RNA coding region [24, 47].

In general, pathogenic mtDNA rearrangement mostly refers to a large-scale deletion of mtDNA, which has a low recurrence risk leading to the general point of view that it is not transmitted through maternal lineage. A single large deletion of mtDNA, approximately 5 kb in
size, has been observed in patients with Kearns-Sayre syndrome (KSS; OMIM#530000) and Pearson marrow pancreas syndrome (OMIM#557000). An affected individual with KSS or Pearson syndrome usually has healthy parents and siblings indicating that this mutation has not been maternally transmitted [48]. However, rare cases of pathogenic mtDNA deletion being transmitted from the mother to her offspring have been reported [49-50]. The actual transmitted molecule that leads to mtDNA deletion in the offspring was thought to be mtDNA duplication rather than the mtDNA deletion itself [51]. Although the observation of a mother and her son both carrying an identical single large deletion associated with KSS leads to a possibility that a single large mtDNA deletion may be maternally inherited as well [49-50].

On the other hand, pathogenic mtDNA point mutations, either in protein-coding region or RNA coding region, is usually maternally inherited, such as m.11778G>A (OMIM*516003.0001), m.3243A>G (OMIM*590050.0001), and m.8344A>G (OMIM*590060.0001). However, there are some pathogenic mtDNA point mutations that were not transmitted to the next generation. The occurrence of a sporadic pathogenic mtDNA mutation may be caused either by de novo mutation occurring only in an affected somatic tissue or by selection against a severe mutated mtDNA in germ line generating an extremely large shift in heteroplasmy level between a mother and her offspring [47, 52-54].

Typically, an affected individual carries a pathogenic mtDNA mutation in a heteroplasmic condition, although affected individuals who carry mutant homoplasm of a particular pathogenic mtDNA mutation have been observed as well. The pathogenic mtDNA mutations observed in mutant homoplasmic individuals usually cause tissue-specific disease phenotypes, for example m.11778G>A and m.3460G>A, the pathogenic mtDNA mutations observed in LHON patients affect retinal ganglion cells causing blindness [55].

Mitochondrial diseases caused by pathogenic mtDNA mutations generally have incomplete penetrance indicating the requirement of other factors to determine an expression of the disease associated with the mutation. Primarily, the expression of the disease phenotype is determined by the mtDNA heteroplasmy level in which the proportion of mutant mtDNA needs to exceed the threshold level. The threshold level varies between different pathogenic mtDNA mutations and different tissues. For example, the threshold level for m.3460G>A mtDNA mutation is approximately 60% [56], while the threshold level for m.3243A>G mtDNA mutation is approximately 95% [57]. Whether the mitochondrial dysfunction is simply caused by inadequate number of wild type mtDNA depends on the type of pathogenic mtDNA mutation [58].

Even though exceeding the threshold level is necessary, in most cases, it may not be sufficient to determine an expression of disease phenotype primarily caused by the pathogenic mtDNA mutation. The modifying factors include co-segregation of secondary mtDNA mutation, nDNA modifiers, or mtDNA haplotype. These modifiers can either determine whether the disease phenotype is expressed or modulate the severity of disease phenotype. For example, coexistence of m.14693A>G with the primary LHON mutation, m.3460G>A, observed in a Han Chinese family may increase a risk of being blind among kindred of this family because the m.14693A>G can alter the tertiary structure tRNA^Glu^ causing tRNA^Glu^ function to change [59]. Identification of X-chromosomal haplotype interacting with specific MTND mtDNA mutation causing visual failure in LHON patients supports the notion that the expression of mitochondrial disease can be modified by a nDNA modifier [60]. The observation of an increased risk of
visual failure associated with m.11778G>A or m.14494T>C among haplogroup J2 and J1, respectively, supports that the expression of disease phenotype may partly be determined by a genetic background of mtDNA [60].

Therefore, although an understanding of human mtDNA heteroplasmy transmission may help predict offspring heteroplasmy level, identifying other factors, both genetic and environmental factors, is perhaps more important. There are some fundamental questions yet to be answered: How does the secondary mtDNA mutation modify the effect of the primary mtDNA mutation? How does nDNA interact with mtDNA to modify an expression of the disease of interest? Can we identify primary-secondary mtDNA interaction and/or nuclear-mitochondrial interaction and how can we do it? Does mitochondrial haplotype influence the prevalence of a particular mtDNA mutation in the population?

Section 1.4: Pathogenic mtDNA mutations in human population

The minimum prevalence of mitochondrial disease both in adult and childhood populations estimated from several published literatures is approximately 1 in 5000 individuals [61]. This estimate includes both nDNA and mtDNA mutations causing mitochondrial disease in human population; the prevalence of mitochondrial disease caused by pathogenic mtDNA mutation, therefore should be less than this estimate. However, the prevalence of the pathogenic mtDNA mutation in human population cannot accurately be implied from this estimate because the prevalence of affected individuals with mitochondrial disease is subject to an ascertainment bias.

Perhaps, the most common pathogenic mtDNA mutation is the m.3243A>G mutation associated with a variety of human diseases, such as type II diabetes and MELAS. Therefore, there have been a number of epidemiological studies of this mutation in several human populations.

The first epidemiological study to estimate the prevalence of the m.3243A>G mutation was done in an adult population of Northern Ostrobothnia in Finland, yielding the prevalence of this mutation at approximately 16.3 in 100,000 [62]. The more recent report on the prevalence of pathogenic mtDNA mutations in human population was estimated from the working-age population of the Northeast of England. The results showed that approximately 9.2 in 100,000 affected individuals with mitochondrial disease and 16.5 in 100,000 unaffected first-degree relatives of the affected individuals carry pathogenic mtDNA mutations. Therefore, the prevalence of pathogenic mtDNA mutations in this population is approximately 25.7 in 100,000 individuals with approximately 7.69 in 100,000 individuals carrying m.3243A>G mutation [63]. The prevalence of m.3243A>G mutation in the Northeast of England seems to be half that of the Finnish population.

Another epidemiological study in the childhood population of the Northern Ostrobothnia in Finland reported the prevalence of the m.3243A>G mutation at approximately 18.4 in 100,000 individuals. This population was ascertained through their disease phenotype identified in the category of mitochondrial disease [64]. The epidemiological study in the population of the Australian cohort, considered a Caucasian-based population, yields the prevalence of this mutation in general population at approximately 236 in 100,000 individuals [65]. This epidemiological study reliably indicates the prevalence of this mutation in general Australian
population because it was estimated from all consenting individuals, both affected and healthy, minimizing the effect of ascertainment bias.

Another common pathogenic mtDNA mutations are the ones associated with LHON diseases. The prevalence of LHON disease in Finland estimated from the entire population is approximately 2 in 100,000 of the population with the prevalence of LHON carriers is 1 in 9000 of the population. These carriers carry one of the most common LHON mutations: m.11778G>A, m.3460G>A, and m.14484T>C [66]. The epidemiological study of LHON disease in the population of North East of England yields the prevalence of visual failure caused by LHON disease at approximately 3.22 in 100,000 of the population. The prevalence of the three most common LHON mutations in this population is approximately 11.82 in 100,000 of the population [67]. Both the prevalence of LHON disease and LHON mutations in these two populations are quite similar.

The most recent study of the prevalence of pathogenic mtDNA mutation in the general population was done prospectively in the population of North Cumbria in England by screening through the neonatal-cord-blood samples of live births. The result of this study presents the prevalence of 10 pathogenic mtDNA mutations including MELAS and LHON mutations: m.1555A>G, m.3243A>G, m.3460G>A, m.7445A>G, m.8344A>G, m.8993T>G, m.11778G>A, m.13513G>A, m.14459G>A, and m.14484T>C, in the general population at approximately 1 in 200 individuals with mutation rate at approximately 1 in 1000 individuals [68]. Overall prevalence of the pathogenic mtDNA mutation in the general population may be more common than previously thought.

The diverse prevalence of pathogenic mtDNA mutations from different studies may be the result of the individuals in each study being ascertained to different clinical features, the data being collected in different timelines with variation of individuals’ age, and different genetic backgrounds of individuals in the study. Thus the prevalence estimated from one population may not reliably predict the prevalence of the same pathogenic mtDNA mutation in other populations.

Section 1.5: Transmission of human mtDNA

1.5.1 Maternal transmission of human mtDNA

Human mtDNA is almost exclusively maternally inherited [69]. The degradation of ubiquitin-tagged mammalian sperm mtDNA at an early pre-implantation embryo supports uniparental inheritance of mtDNA [70]. Until now, only a single case of paternal leakage of human mtDNA has been observed in a muscle sample of an affected male with myopathy [29]. The systematic studies in other patients with sporadic mitochondrial myopathy indicate no transmission of paternal pathogenic mtDNA mutation implicating a very low probability of paternal leakage of human mtDNA [71-72]. Therefore in terms of clinical concern, only mothers who carry a pathogenic mtDNA mutation can pass her mtDNA defect to her offspring.

1.5.2 Transmission of pathogenic mtDNA mutation in human family

A mother carrying a pathogenic mtDNA mutation in homoplasmic condition always has offspring carrying mutant homoplasmy. This transmission pattern is frequently observed in
human family carrying LHON mutations [73-75] because LHON mutations generally have low penetrance requiring modifying factors, such as a secondary mtDNA mutation and nDNA mutation, to determine an expression of the disease phenotype [60, 76-77]. Phenotypic variation of LHON disease is generally observed among family members carrying the same heteroplasmy level [73, 78].

A mother carrying a heteroplasmic pathogenic mtDNA mutation transmits a varied proportion of mutant mtDNA to her offspring generating a large random shift in heteroplasmy level between herself and her offspring as well as among her offspring siblings [79-81]. Both intergenerational and intra-generational variation of heteroplasmy level complicates a prediction of recurrence risk in a family carrying a pathogenic mtDNA mutation. This complication leads to an uncertainty in genetic counseling for a family carrying a pathogenic mtDNA mutation [82-88].

Whether the large shift in heteroplasmy level between a mother and her offspring is purely random, or subject to positive selection, remained to be revealed. The transmission of single generation human mtDNA heteroplasmy level was studied from human clinical pedigrees carrying one of the six common pathogenic mtDNA mutations (m.8344A>G, m.3243A>G, m.8993T>G/C, m.11778G>A, and m.3460G>A). The results showed that the transmission of m.3243A>G, m.8993T>G and m.11778G>A is in favor of inheriting mutant mtDNA, while the transmission of m.8344A>G seemed to be in favor of inheriting wild type genome [89]. The preferential transmission of the m.3243A>G and m.8993T>G mtDNA mutations was also observed in other independent human clinical pedigree data. The result presented statistically significant differences in heteroplasmy level between mother and her offspring carrying these two pathogenic mtDNA mutations [90]. The transmission of m.8993T>C and m.11778G>A seems to be in favor of inheriting the mutant genome but the deviation of the distribution of their intergenerational difference of heteroplasmy level does not reach statistical significance [89]. There is no gender bias of human mtDNA heteroplasmy transmission observed from the study of human clinical pedigree data of m.3243A>G, m.8363G>A, m.8344A>G, and m.8993T>G/C [90].

The fact that these studies were done retrospectively via human clinical pedigree data, the interpretation drawn from these analyses can easily be deceived by an ascertainment bias [89]. Generally for the study of human mtDNA heteroplasmy transmission, the ascertainment bias is simply adjusted by excluding all probands from the analysis, although this simple procedure is not an effective way to eliminate this confounding factor. Therefore the refined process to eliminate the ascertainment bias from human clinical pedigree data used for the study of human mtDNA heteroplasmy transmission is needed to be developed since this type of human data plays a key role in the study of pathogenic mtDNA heteroplasmy transmission.

The variation of mtDNA heteroplasmy level between different tissues and its change over time would be another source of concern. The measurement of m.3243A>G heteroplasmy level from different tissues of an individual indicated a large variation of this mtDNA mutation level between different tissues [91]. As well, the exponential decrease of blood m.3243A>G heteroplasmy level with age has been reported [92-95]. This exponential decrease can be explained by selection against blood stem cells carrying a high proportion of this mutant mtDNA [93]. However, the measurement of m.8993T>G/C heteroplasmy level from several tissues does not show variation of heteroplasmy levels between different tissue types. The longitudinal study
of this mutation based on blood sample measurement also presents non-significant change with age [96]. Thereby different pathogenic mtDNA mutation may have different pattern of heteroplasmy segregations between tissues and its heteroplasmy level may or may not change with age.

Different pathogenic mtDNA mutations present different size of intergenerational random shift in heteroplasmy level. A rapid shift is commonly observed in the families carrying m.8993T>G mutation. The fixation of mutant mtDNA may happen within a single generation [96-97]. However, the rapid shift is less likely to be observed in the family carrying the m.8344A>G mtDNA mutation [79]. The difference of the intergenerational random shift in heteroplasmy level between different pathogenic mtDNA mutations may imply an observation of different size of pedigrees carrying different mutations. The size of m.8993T>G pedigree tends to be smaller than the size of m.8344A>G pedigree. The size difference of random shift between different pathogenic mtDNA mutations also suggests different transmission mechanisms generating offspring heteroplasmy level.

1.5.3 MtDNA heteroplasmy inheritance and mitochondrial genetic bottleneck

In principle, the differences between an organelle and a nuclear genome transmission are genotypic segregation during mitotic cell division, relaxed replication of organelle DNA, stochastic partition of organelle DNA, random intracellular selection, uniparental inheritance, and reduced recombination [98]. These characteristics of the organelle DNA transmission make it distinct from the nuclear DNA transmission, thereby the principle of Mendelian genetics cannot be applied to predict an outcome of the organelle genome transmission.

The study of maternal transmission of a neutral mtDNA polymorphism in Holstein cattle is the first study presenting a large random shift in allele frequency of mtDNA between mother and her offspring leading to a fixation of a particular allele within a few generations [99]. The hypothesis of mitochondrial genetic bottleneck is raised to explain this rapid shift of mitochondrial allele in that the restricted number of mtDNAs repopulates the next generation mtDNA population. The hypothesis of mitochondrial genetic bottleneck has gained supported by the experimental result in a mouse model carrying a neutral NZB/BALB mtDNA. The comparison of the average mtDNA heteroplasmy level of the offspring to the maternal heteroplasmy level showed that these two values are approximately equal, supporting the random segregation of mtDNA heteroplasmy level. The comparison of mtDNA heteroplasmy variance across different stages of female germ line development- primordial germ cell (PGC), primary oocyte, and mature oocyte, until the offspring- indicated that the largest variation of heteroplasmy level is generated during the proliferation of PGCs to generate primary oocytes [100]. The stochastic partition and relaxed replication of mtDNA during the proliferation of PGCs generates variation in heteroplasmy level among primary oocytes of the mother, as shown in Figure 3. However whether the restricted number of mtDNA molecules repopulates the next generation mtDNA population, remains to be explored.
Figure 3: Transmission of mtDNA heteroplasmy level during prenatal development of female germ line. After fertilization, zygote undergoes multiple cell divisions without mtDNA replication to generate blastocyst causing a small copy numbers of mtDNA molecules per blastomere. At implantation, some blastomeres are recruited to be primordial germ cells (PGCs). These PGCs experience serial cell divisions to generate a population of primary oocytes. The mtDNA replication is resumed at an early post-implantation development. The random partition of mtDNA molecules during this prenatal development of female germ lines and stochastic replication of mtDNA molecules during post-implantation development generate variation of heteroplasmy level in the population of primary oocytes.

Due to technological advances, the measurement of mtDNA copy number in a single cell can be done, which will allow us to directly test the mitochondrial genetic bottleneck hypothesis. In 2007, Cao L et al. measured mtDNA copy number per cell in various stages of early oogenesis starting from pre-implantation embryo to mature oocyte. They observed non-significant changes of mtDNA copy number between different stages during this early germ-line development, then they concluded that the mitochondrial genetic bottleneck is a segregation of a small number of mitochondrial segregating units, either nucleoids or actively replicating clusters of mtDNA molecules, without a drastic reduction of mtDNA copy number per cell during this prenatal germ-line development [101]. In 2009, this conclusion was supported by the experimental result in the mouse model observed by the same research group [102]. However, this conclusion is contradicted by our experimental observation in the GFP-Stella mouse model in that there is a drastic reduction of mtDNA copy number per cell during a blastocyst formation generating a small mtDNA copy number in a primordial germ cell. The detail of this study is presented in Chapter 2. Therefore, based on our observation, we concluded that mtDNA
molecules well serves as the mitochondrial segregating unit in which its random partition during pre- and early-post implantation development largely generates offspring heteroplasmy variation [103].

The reduction of mtDNA copy number per cell during the blastocyst formation generating a small mtDNA copy number per cell in a primordial germ cell has gained supported by an independent study done by Wai T et al. in 2009 [104]. They measured mtDNA copy number per cell and mtDNA heteroplasmy level from germ line cells generated during both prenatal and postnatal oocyte developments. Their result showed that only the mtDNA heteroplasmy level variance measured from the germ line cells before postnatal day 11 significantly differs from the value measured from the germ line cells after postnatal day 11. Therefore, they concluded that the variation of offspring heteroplasmy levels has been largely generated during postnatal development of the female germ line cells by random segregation of mitochondrial nucleoids [104].

Even though the mitochondrial genetic bottleneck theory seems to be the principle mechanism determining offspring heteroplasmy level, there are several important issues left to debate, especially about what the mitochondrial segregating unit is and when the mtDNA heteroplasmy variance is generated [105].

1.5.4 Does selection play a role?

The different rate of mtDNA heteroplasmy segregation between different pathogenic mtDNA mutations implies that different mechanisms regulate transmission of pathogenic mtDNA heteroplasmy level. The variation of the sizes of intergenerational random shift in heteroplasmy level may be caused by different degrees of selection.

An experiment in mutator mice showed that nonsynonymous mutation in mitochondrial protein coding region is subject to a purifying selection. This mutator mouse was generated by introducing mutated proofreading PolG gene into the founder female mouse, which then was mated with a normal male mouse to eliminate the mutator gene [106] allowing Stewart et al. (2008) to study the inheritance of mtDNA mutations [107]. The rate of non-synonymous mutation loss is as fast as within two generations, while the synonymous mtDNA mutation is still segregated [107]. They also observed a high mutation rate in tRNA and rRNA coding regions in this mutator mouse indicating that mutation in these regions may be non-lethal or more compatible to life than the mutation in protein coding region; however the pattern of mutation rate of these regions in the mutator mouse is inconsistent with the pattern observed in natural mouse strains and in humans. Within the category of mitochondrial protein coding gene, some genes presented a higher rate of mutations than others such as the MTCYB, MTATP6 and MTATP8 genes. These excess mutated genes can be observed from both the mutator mouse and human. The excess of mutation in MTATP6 gene may explain why we observed a large number of de novo mutations and rapid segregation of mtDNA heteroplasmy level of the pathogenic m.8993T>G mutation [96].

Section 1.6: Application of population genetic theory

The population genetic theory of random drift can mathematically describe random segregation of mtDNA heteroplasmy level. The mathematical formula of random genetic drift theory was
first introduced in the field of mitochondrial genetics by Solignac et al. in 1984 [108]. They applied the Sewall-Wright variance formula, shown in equation 1, to describe the variation of mtDNA heteroplasmy level calculated from various generations descendant of *Drosophila mauritiana* and to predict the evolution of the variation of the descendents’ mtDNA heteroplasmy level. In population genetic theory, this variance formula mathematically explains the variation of neutral allele frequencies in a finite population that its allele frequency is changed by sampling error [109].

\[
V_t = p_0(1 - p_0) \left[ 1 - \left( 1 - \frac{1}{2N_{eff}} \right)^t \right] \approx p_0(1 - p_0) \left[ 1 - e^{-t/2N_{eff}} \right]
\]

The variance of neutral allele frequencies \( (V_t) \) at any time \( t \) depends on a founder population allele frequency \( (p_0) \), an effective population size \( (N_{eff}) \), and the duration time of the development of allele frequency simply counted as a number of generations \( (t) \). Practically, in the field of mtDNA heteroplasmy inheritance, we can directly measure the heteroplasmy variance at a particular generation \( (V_t) \) and the founder female heteroplasmy level \( (p_0) \). The number of generations \( (t) \) can be either the number of organism generations as used in the study of *D. mauritiana* mtDNA heteroplasmy segregation [108] or the number of cell divisions required to generate a population of female mouse germ line cells [100].

The general use of the Sewall-Wright variance formula to understand mtDNA heteroplasmy inheritance is to estimate the size of a mitochondrial bottleneck which is the effective population size in the equation \( (N_{eff}) \). The effective population size can be calculated from the formula, shown in equation 1, by substituting the heteroplasmy variance value \( (V_t) \), the founder female heteroplasmy level \( (p_0) \) and the number of generations \( (t) \) in the equation. By assuming the number of generations to be equal to 15 corresponded to the estimated number of cell divisions required to produce a population of mouse primary oocytes, the bottleneck size of germ line mouse is approximately 185 [100]. In human, the estimated number of cell divisions required to produce a population of human primary oocytes is equal to 23. By using this estimated value for the parameter \( t \) in equation 1, the predicted human germ line bottleneck size is 173, which is close to the bottleneck size in the neutral mouse model [110].

The Sewall-Wright variance formula has a limitation. It cannot describe the distribution pattern of allele frequencies in the population of interest. However, there is an attempt to apply this variance equation to generate the distribution of mtDNA heteroplasmy levels. Poulton et al. in 1998 developed a repeat-selection model based on this variance formula. Given known mitochondrial bottleneck size and assumed number of generations, a discrete distribution of mtDNA heteroplasmy levels can be estimated [111]. Theoretically, the distribution of allele frequencies in the population subject to random drift process can be obtained from the Wright-Fisher model. This model uses binomial sampling process to describe the segregation of allele frequencies in the population given that the parameter \( N_{eff} \) and \( t \) in the Sewall-Wright variance formula are known or can be estimated [112]. However, the issues of what the mitochondrial segregating unit is and how many generations the mtDNA heteroplasmy is segregated cause complications to estimate values for the parameters \( N_{eff} \) and \( t \), respectively. In addition, the discrete distribution of mtDNA heteroplasmy levels estimated from either the Poulton’s model or the theoretical Wright-Fisher model does not correspond to the continuous distribution of
Section 1.7: Application of statistical and computational approaches

The statistical inferences and computational simulation models based on Monte-Carlo method have been applied through all the studies presented in this dissertation. These two approaches have different advantages and were used for different purposes. Most of the times, these two approaches complement each other.

The advantage of applying statistical analysis in the study is the biological interpretation is drawn from the actual measurements either from clinical or experimental data, although the interpretation is limited to availability and quality of the actual measurements. In the field of mitochondrial genetics, the general statistical analysis was applied to a large number of human clinical pedigrees in order to understand human mtDNA heteroplasmy inheritance. For example, a relationship between a maternal heteroplasmy level and a probability of having an affected offspring was statistically estimated from a collection of human clinical pedigree data. The student’s t test and the $\chi^2$ analysis were applied to examine an impact of a possible confounding factor and the relationship between the maternal heteroplasmy level and the probability of having an affected offspring, respectively [113]. Testing whether the average of the difference heteroplasmy level between a mother and her offspring significantly differs from zero and whether its distribution significantly deviates from the normal distribution help to determine the mechanism of mtDNA heteroplasmy transmission [89]. Perhaps, the most important issue of human clinical pedigree data analysis is the confounding effect of the ascertainment bias. Until recently, there still has no effective way to eliminate this confounding factor from the study of human mtDNA heteroplasmy inheritance.

The computational simulation study has an advantage of conquering the limitation of the actual measurements both in terms of quality and quantity. Its flexibility would be useful for testing some hypotheses without having to have the actual measurements. However the application of the simulation model depends on its assumptions. The pedigree simulation of an inheritance of complex trait including mitochondrial genome was designed based on a set of theoretically mathematical functions. This model would be of benefit to the study of nuclear-mitochondrial interaction in determining an expression of a human disease, although this model did not take into account the effect of mtDNA heteroplasmy level [114]. Therefore, the intergenerational segregation of mtDNA heteroplasmy levels still needs to be modeled.

Throughout the dissertation, the statistical analysis is performed using a variety of software such as Microsoft Excel (Microsoft software), Origin (OriginLab Corporation), and R (R foundation for statistical computing). The simulation code is implemented in C/C++ programming using a sequential design. The general purpose of all simulation models presented in this dissertation is to generate synthetic data of mtDNA heteroplasmy levels for hypothesis testing. The synthetic data is a set of random numbers drawn from a particular distribution function. The parameters of the distribution were either arbitrarily determined or estimated from the actual measurements of mtDNA heteroplasmy level.
Section 1.8: Research motivations

The previous sections present the general background that is relevant to human mtDNA heteroplasmy inheritance, the basic questions in the research field and the brief methodological approaches being applied in the study. This section aims to list the specific questions addressed in the dissertation and briefly explain how each study helps to gain understanding of human mtDNA heteroplasmy inheritance.

Human mtDNA is almost exclusively maternally inherited presenting a large random shift in proportion of mutant mtDNA between a mother and her offspring and a variation of this proportion among siblings. The mitochondrial genetic bottleneck is broadly hypothesized as the main mechanism generating this two dimensional variation [89, 99]. The experiment in a neutral mouse model shed some light on when this mechanism happened during mammalian embryogenesis and how large the bottleneck size should possibly be, although neither the direct measurements of mtDNA copy numbers nor the complete picture of how the variation being generated by the mitochondrial genetic bottleneck had been presented [100]. Thus this left us an important question which is,

How does the mitochondrial genetic bottleneck happen during mammalian embryogenesis generating mtDNA heteroplasmy variance?

By working closely with our experimental collaborators, we can address this question by building an embryogenesis simulation model based on the mtDNA copy number directly measured from the mouse model during its embryogenesis. The effectiveness of the simulation model in estimating mtDNA heteroplasmy variation is examined by the comparison of the simulated germ line heteroplasmy variation to the actual measured mouse mtDNA heteroplasmy variation. Answering this question using the experimental mouse model helps us gain insight into the cellular mechanism of mammalian mtDNA heteroplasmy transmission, which has an implication for the human mtDNA heteroplasmy transmission. The detail study is presented in Chapter 2 [103].

Another approach to understand the human mtDNA heteroplasmy transmission is through indentifying the distribution pattern of mtDNA heteroplasmy levels among siblings of a single maternal ancestor, leading to the question:

What is the distribution pattern of an offspring’s mtDNA heteroplasmy level?

We addressed this question by applying a set of probability distribution functions developed in the field of population genetics to describe the distribution pattern of mtDNA heteroplasmy levels. These probability distribution functions were developed on the basis of random drift theory. The efficacy of these probability functions to describe the distribution of mtDNA heteroplasmy levels is demonstrated by the comparison of the theoretical distribution to the actual mtDNA heteroplasmy distribution measured across several germ line developmental stages and among different organisms. The detail of adapting this set of mathematical functions from the population genetic theory and testing its potential application to mitochondrial genetic is shown in Chapter 3 [115].
The comparison of measured mtDNA heteroplasmy variance between different stages of germ line development plays a key role in demonstrating how offspring’s heteroplasmy variation has been generated. Generally, the mtDNA heteroplasmy variance was calculated from a limited modest sample size of mtDNA heteroplasmy measurements, which may not be sufficient to represent the mtDNA heteroplasmy variance in the population. Lack of reporting of error bars of sampled mtDNA heteroplasmy variance could, at least in an extreme case, distort the biological interpretation drawn from the mtDNA heteroplasmy variance value. This awareness leads us to ask the question:

**How can we estimate the sampling error of the actual mtDNA heteroplasmy variance measurement?**

To answer this question, we adapt the general basis of the standard error of variance calculation and develop it to calculate an error bar of mtDNA heteroplasmy variance. Adding the variance error bar to the experimental mouse mtDNA heteroplasmy variances measured from various stages of mouse germ line development revises the biological interpretation drawn from these measurements. We also add the variance error bar calculation to help us in answering the question:

**Does the mitochondrial bottleneck size in a mouse model differ from that in humans?**

To improve the experimental measurement of mtDNA heteroplasmy variance, we can use the variance error bar calculation to estimate the proper sample size for this higher-order statistic. Therefore, the variance error bar calculation plays an important role in the study of mtDNA heteroplasmy transmission by adding a necessary piece of information to the biological interpretation drawn from the heteroplasmy variance measurement and helping in an experimental design related to the heteroplasmy variance calculation. The detail of how to calculate the heteroplasmy variance error bar, how to apply it to the actual heteroplasmy variance, and how to estimate a proper sample size for the experimental design are presented in Chapter 4 [116].

The m.3243A>G mutation is a high penetrance pathogenic mtDNA mutation associated with a variety of human diseases. Intergenerational transmission analysis of this mutation implicated the preference of inheriting mutant mtDNA [89-90]. However, the reduction of blood heteroplasmy level with age could lead to the conclusion that the offspring’s heteroplasmy level tends to be higher than the mother’s heteroplasmy level without the preference of inheriting mutant mtDNA [89-90], thereby, to clarify this statement, we ask the question:

**Is the intergenerational increase of heteroplasmy level of m.3243A>G caused by the preference of transmitting mutant mtDNA?**

We have answered this question by testing whether the reduction of blood heteroplasmy level with age generates an increase of an average heteroplasmy level between generations and whether the size of random shift statistically deviates from zero. We also test whether this pathogenic mtDNA mutation affects reproduction of female carriers. We ask the questions:

**Does the pathogenic mtDNA mutation affect female reproductive fitness?**
Does the effect of this mutation on female reproduction bias against a particular offspring gender?

and

Does the effect of this mutation on female reproduction depend on maternal heteroplasmy level?

In an attempt to answer these questions, we statistically analyze the number of live births of the female carrying this pathogenic mtDNA mutation, separated by offspring gender and maternal heteroplasmy level. The detail analysis of both the intergenerational mtDNA heteroplasmy transmission and the female reproductive effect of this mutation are presented in Chapter 5.
Chapter 2: A Reduction of Mitochondrial DNA Molecules During Embryogenesis Explains the Rapid Segregation of Genotypes

Lynsey M. Cree\(^1\), David C. Samuels\(^2\), Susana Chuva de Sousa Lopes\(^3\), Harsha Karur Rajasimha\(^2\), Passorn Wonnaapinij\(^2\), Jeffrey R. Mann\(^4,7\), Hans-Henrik M. Dahl\(^5\) & Patrick F. Chinnery\(^1,6\).

\(^1\)Mitochondrial Research Group, Newcastle University, NE2 4HH, UK.
\(^2\)Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, VA, 24061, USA
\(^3\)Wellcome Trust/Cancer Research UK Gurdon Institute, University of Cambridge, Tennis Court Road, Cambridge CB2 1QN, UK.
\(^4\)Division of Biology, City of Hope, 1500 E. Duarte Rd, Duarte, California 91010, USA. Present address: Department of Zoology, The University of Melbourne, Vic. 3010, Australia.
\(^5\)The Murdoch Children’s Research Institute & Dept. of Paediatrics (Univ. of Melbourne), Royal Children’s Hospital, Parkville Melbourne, Victoria. 3052, Australia.
\(^6\)Institute of Human Genetics, Newcastle University, Newcastle NE2 4HH, UK
\(^7\)Present address: Department of Zoology, The University of Melbourne, Victoria 3010, Australia

Reprinted from “Nature Genetics 2008, 40, 249-254” with permission from Nature Publishing Group

To whom correspondence should be conducted: *p.f.chinnery@ncl.ac.uk
Preface

The aim of chapter 2 is to understand how mitochondrial genetic bottleneck generates variation in offspring’s heteroplasmy levels. In particular, we first experimentally observed changes of mtDNA copy number per cell during mouse embryogenesis to examine whether there is a physical bottleneck of mtDNAs, a small number of mtDNA molecules repopulating the next generation mtDNA population. Then, we used the simulation model to investigate how these changes of mtDNA copy number per cell generates offspring’s heteroplasmy variance.

This project has been done in collaborative manner. The overall experimental study was designed by Dr. Patrick F. Chinnery and Dr. Lynsey M. Cree from University of Newcastle upon Tyne. The measurements of mtDNA copy number per cell to observe the changes of mtDNA molecules in a single cell during embryogenesis from the GFP-Stella mouse model was done by Dr. Lynsey M. Cree. These fluorescence-tagged mice were generated in Dr. Susana Chuva de Sousa Lopes’s lab from the University of Cambridge. The actual mouse offspring’s mtDNA heteroplasmy level was measured from the NZB/C57BL.6J mouse model developed by Dr. Hans-Henrik M. Dahl and Dr. Jeffrey R. Mann from the University of Melbourne.

Dr. David C. Samuels designed the mouse embryogenesis simulation model. Dr. Hasha Karur Rajasimha programmed this model in C/C++ language. Dr. Samuels and I carried out this simulation model to study how the variation in female germ line heteroplasmy levels has been generated by the changes of mtDNA copy number per cell during mouse embryogenesis. This model initiated with a single simulated embryo carrying 500,000 mtDNA molecules with 50% heteroplasmy level. We first validated the simulation model by comparing the mtDNA copy number per simulated cell to the actual mtDNA copy number per cell observed from the experimental GFP-Stella mouse model. Then, we calculated mtDNA heteroplasmy variance from the simulated cells at each stage of mouse embryogenesis to study the generation of germ line cells’ heteroplasmy variance. We also applied this simulation model to estimate 95% confident limit of the offspring’s heteroplasmy level of the mother carrying mutant mtDNA in heteroplasmic condition with the maternal heteroplasmy level is in the range of 0% to 100%.
A reduction of mitochondrial DNA molecules during embryogenesis explains the rapid segregation of genotypes

Lynsey M Cree¹, David C Samuels², Susana Chuva de Sousa Lopes³, Harsha Karur Rajasimha², Passorn Wonnapinij², Jeffrey R Mann⁴,⁷, Hans-Henrik M Dahl⁵ & Patrick F Chinnery¹,⁶

Mammalian mitochondrial DNA (mtDNA) is inherited principally down the maternal line, but the mechanisms involved are not fully understood. Females harboring a mixture of mutant and wild-type mtDNA (heteroplasmy) transmit a varying proportion of mutant mtDNA to their offspring. In humans with mtDNA disorders, the proportion of mutated mtDNA inherited from the mother correlates with disease severity¹–⁴. Rapid changes in allele frequency can occur in a single generation⁵–⁶. This could be due to a marked reduction in the number of mtDNA molecules being transmitted from mother to offspring (the mitochondrial genetic bottleneck), to the partitioning of mtDNA into homoplasmic segregating units, or to the selection of a group of mtDNA molecules to re-populate the next generation. Here we show that the partitioning of mtDNA molecules into different cells before and after implantation, followed by the segregation of replicating mtDNA between proliferating primordial germ cells, is responsible for the different levels of heteroplasmy seen in the offspring of heteroplasmic female mice.

Rapid changes in mitochondrial allele frequency were first observed in the offspring of Holstein cows⁵,⁶. Similar changes have subsequently been described in many mammalian species, including humans transmitting pathogenic mtDNA mutations¹,³,⁴,⁷. In heteroplasmic mice transmitting neutral mtDNA polymorphisms, the percentage level of heteroplasmy seen in the offspring is determined at an early stage during oogenesis in the developing mother, before the formation of the primary oocytes⁸. The same process appears to explain the transmission of highly pathogenic mtDNA mutations⁹,¹⁰, present in ~1 in 5,000 of the population.

The size of the mitochondrial genetic bottleneck in mice is predicted to be ~200 segregating units⁸. Given that the total amount of mtDNA remains constant within dividing preimplantation mouse embryos¹¹, we initially determined whether the amount of mtDNA within single blastomeres fell to ~200 immediately before implantation at 5.5 days post coitum (d.p.c.; Table 1). The median number of mtDNA molecules in mature oocytes was 2.28 × 10⁵, and the total amount of mtDNA within the entire preimplantation embryo remained constant (F = 1.82, P = 0.101, Fig. 1a). These values are consistent with previous estimates¹¹,¹² and confirm that mtDNA replication is not required for healthy preimplantation development¹³. With sequential binary cell divisions, the amount of mtDNA in each cell fell to an estimated value of ~4,000 mtDNA molecules immediately before implantation (Fig. 1b, Table 1), some ~20-fold greater than predicted⁵. We therefore concluded that the mtDNA genetic bottleneck is not due to subdivision of a non dividing mtDNA pool in the preimplantation embryo.

Primary oocytes are first detectable at 7.25 d.p.c. and develop from a founder population of 40 primordial germ cell (PGCs) recruited by induction from the epiblast in the posterior-proximal embryonic pole¹⁴. To determine whether the amount of mtDNA in PGCs could explain the mtDNA bottleneck, we studied fluorescently sorted PGCs from mice expressing green fluorescent protein–tagged Stella, which is the most specific marker of PGCs¹⁵. The first discernable PGCs contained substantially lower amounts of mtDNA than the preimplantation blastomeres (Table 2). PGCs isolated from embryos at 8.5–14.5 d.p.c. contained greater amounts of mtDNA, indicating that mtDNA replication had commenced (Fig. 1c), but the average amount of mtDNA within the developing PGCs remained ~100-fold lower than in mature oocytes (median mtDNA copies per PGC at 14.5 d.p.c., 1,529). The median number of mtDNA molecules present within 7.5-d.p.c. PGCs was 203, which closely approximates a previous estimate (185 segregating units⁸) inferred from heteroplasmic mice using a model based upon population genetic theory⁸,¹⁶.

¹Mitochondrial Research Group, Newcastle University, Newcastle NE2 4HH, UK. ²Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061, USA. ³Wellcome Trust/Cancer Research UK Gurdon Institute, University of Cambridge, Tennis Court Road, Cambridge CB2 1QN, UK. ⁴Division of Biology, City of Hope, 1500 E. Duarte Rd., Duarte, California 91010, USA. ⁵The Murdoch Children’s Research Institute & Department of Paediatrics (University of Melbourne), Royal Children’s Hospital, Parkville, Melbourne, Victoria 3052, Australia. ⁶Institute of Human Genetics, Newcastle University, Newcastle NE2 4HH, UK. ⁷Present address: Department of Zoology, The University of Melbourne, Victoria 3010, Australia. Correspondence should be addressed to P.F.C. (p.f.chinnery@ncl.ac.uk).

Received 16 April 2007; accepted 19 October 2007; published online 27 January 2008; doi:10.1038/ng.2007.63
Table 1 Amount of mtDNA in mature oocytes and preimplantation mouse embryos

<table>
<thead>
<tr>
<th>Stage</th>
<th>Mean copy number ($\times 10^3$)</th>
<th>Median copy number ($\times 10^3$)</th>
<th>Range ($\times 10^3$)</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole embryo</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Oocyte</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
</tr>
<tr>
<td>2 cell</td>
<td>24</td>
<td>347.7</td>
<td>312.8</td>
<td>112.9–676.8</td>
</tr>
<tr>
<td>4 cell</td>
<td>13</td>
<td>196</td>
<td>174.3</td>
<td>85.9–483.2</td>
</tr>
<tr>
<td>6 cell</td>
<td>8</td>
<td>308.6</td>
<td>287.8</td>
<td>150.5–547.6</td>
</tr>
<tr>
<td>8 cell</td>
<td>23</td>
<td>244.5</td>
<td>243.0</td>
<td>73.8–488.2</td>
</tr>
<tr>
<td>16–32 cell</td>
<td>12</td>
<td>286.1</td>
<td>267.7</td>
<td>73.8–671.3</td>
</tr>
<tr>
<td>Blastocyst</td>
<td>15</td>
<td>280.8</td>
<td>212.8</td>
<td>56.7–667.1</td>
</tr>
<tr>
<td>Individual blastomeres</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>22</td>
<td>18</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>Mean copy number ($\times 10^3$)</td>
<td>249.4</td>
<td>247</td>
<td>57.7</td>
<td>57.7</td>
</tr>
<tr>
<td>Median copy number ($\times 10^3$)</td>
<td>228.5</td>
<td>245.3</td>
<td>36.3</td>
<td>36.3</td>
</tr>
<tr>
<td>Range ($\times 10^3$)</td>
<td>26.9–645.7</td>
<td>14.6–465.6</td>
<td>15.1–143.6</td>
<td>15.1–143.6</td>
</tr>
<tr>
<td>CV</td>
<td>0.63</td>
<td>0.52</td>
<td>0.9</td>
<td>0.9</td>
</tr>
</tbody>
</table>

CV, coefficient of variance; n, number of cells studied.

Previous estimates of the mitochondrial genetic bottleneck are critically dependent upon the chosen mathematical model. A single-sampling binomial model describes the whole process and has been used to predict the chance of transmitting a particular level of heteroplasmy\(^1^7,18\), but it does not reflect the underlying biology (Fig. 2a). The population genetic model\(^8,16\) assumes that the bottleneck occurs while numerous generations of cells are being formed by binary cell division, with each cell containing the same amount of mtDNA (Fig. 2b). For example, the bottleneck of \(\sim 185\) segregating units was assumed to be present for 15 germ-cell divisions\(^8\), but our observations show that the amount of mtDNA within PGCs increases between 7.5 and 14.5 d.p.c. (Fig. 1c). We therefore adapted these models using the experimental data to model the biological process directly (Fig. 2c).

Before implantation, the model was based on binary cell division of a heteroplasmic mouse embryo with no mtDNA replication, using experimentally determined values for the amount of mtDNA within single blastomeres (Table 1) and the rate of cell division\(^19\). After implantation, the model included mtDNA replication at the minimum rate required to populate the increasing number of PGCs\(^20,21\), and it was also based on experimentally determined values for the median amount of mtDNA within single PGCs (Table 2, Fig. 3a) and the rate of cell division required to produce 25,791 primary oocytes at day 13.5 (ref. 21) (Supplementary Figs. 1 and 2 online).

To determine whether the biological model reliably predicted the transmission of mtDNA heteroplasy, we studied the transmission in 246 offspring from 22 litters born to mothers with different levels of NZB/C57BL.6\(^6\) heteroplasm gener-
ated by cytoplasm transfer (Supplementary Table 1 online). 91% of the offspring had NZB heteroplasmy levels that fell within the predicted 95% confidence interval (CI) for the model (Fig. 3b). The simulation results in Figure 3b are based on a conservative model using the median value for the number of mtDNA copies per cell at 7.5 d.p.c. However, when we simulated a 12-h delay in the onset of mtDNA replication, consistent with the observed variation in implantation time\(^22\), this led to a reduction in mtDNA copies at implantation that approximated the lower end of the measured range in Stella-GFP PGCs (Table 2) and to a greater variance in heteroplasmy values among the offspring.

Together, these observations indicate that the reduction in mtDNA copies we observed during immediate preimplantation and postimplantation development is sufficient to generate the variation in hetero-
plasm levels seen among the offspring of heteroplasmic female mice. Approximately 70% of the heteroplasmy variance was due to the physical partitioning of mtDNA molecules into daughter cells during pre- and early postimplantation develop-
ment, when the amount of mtDNA in each cell fell to low levels (\(\sim 200\)). The remaining 30% developed during the intense proliferation of mtDNA in the exponentially expanding PGC population, where the average amount of mtDNA was \(\sim 1,500\) molecules per PGC (Fig. 3a, c and Supplementary Fig. 2). Variation in heteroplasmy levels will occur without the com-
partmentalization of mtDNA molecules into multicityo segregating units such as nucleoids or mitochondria, or through the selection of particular group of mtDNAs chosen to repopulate the next generation.

Cao et al.\(^23\) reported >1,000 mtDNA molecules in 7.5-d.p.c. PGCs and concluded that the mitochondrial genetic bottleneck is due not to a drastic decline in mtDNA copy number but to a small effective number of effective segregating units, each containing multiple mtDNA molecules. How can we explain the discrepancy between their findings and ours? Differences in our experimental approach explain why we measured a lower number of mtDNA molecules in early PGCs, and our interpretation of these data using a biologically plausible model led us to conclude that the mtDNA molecule is the segregating unit.
Before using Stella-GFP mice, we explored the possibility of using alkaline phosphatase histochemistry to identify PGCs. Like Cao et al., we found that the stain decreased the number of copies measured within individual cells. In our hands, the alkaline phosphatase histochemistry did not affect all cells equally, leading to an increased variability in copy number values. Light microscopy showed that the stain diffused out of PGCs in cryostat sections both at 37 °C and at room temperature, making it difficult to identify which cells were PGCs and which were not. This was particularly a problem when the PGC population was small at 7.5 d.p.c. The correction factor of 1.96 used by Cao et al. when the PGC population was small at 7.5 d.p.c. would not account for the increased variability that we observed. Moreover, as only 95% of alkaline phosphatase–positive PGCs express Stella, it is generally accepted that Stella–GFP is the more accurate marker of the PGC lineage15. It is therefore possible that some of the differences in the parameters governing the genetic bottleneck, such as subtle distortions in the partitioning of mtDNA during cytokinesis.

Table 2 Amount of mtDNA in primordial germ cells at different embryonic stages in mice

<table>
<thead>
<tr>
<th>Stage</th>
<th>n</th>
<th>Mean copy number ($\times 10^3$)</th>
<th>Median copy number ($\times 10^3$)</th>
<th>Range ($\times 10^3$)</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td>7.5 d.p.c.</td>
<td>596</td>
<td>4.51</td>
<td>2.03</td>
<td>0.26–34.02</td>
<td>1.55</td>
</tr>
<tr>
<td>8.5 d.p.c.</td>
<td>165</td>
<td>10.65</td>
<td>7.68</td>
<td>1.28–27.78</td>
<td>0.72</td>
</tr>
<tr>
<td>10.5 d.p.c.</td>
<td>96</td>
<td>18.01</td>
<td>16.30</td>
<td>10.26–21.93</td>
<td>0.18</td>
</tr>
<tr>
<td>14.5 d.p.c., male</td>
<td>1,087</td>
<td>21.52</td>
<td>19.98</td>
<td>7.56–62.41</td>
<td>0.44</td>
</tr>
<tr>
<td>14.5 d.p.c., female</td>
<td>1,528</td>
<td>13.76</td>
<td>12.88</td>
<td>3.42–34.13</td>
<td>0.44</td>
</tr>
</tbody>
</table>

CV, coefficient of variance; n, number of primordial germ cells studied; d.p.c., days post coitum.

Figure 2 Models of the mitochondrial genetic bottleneck. Schematic diagram showing a heteroplasmic fertilized oocyte (top), a model of the mitochondrial genetic bottleneck (middle) and subsequent primary oocytes (bottom). Blue circles, wild-type mtDNA; red circles, mutated mtDNA. Time scale shown on the left in d.p.c. (a) Single-sampling model. A single random sample of mtDNA molecules is assumed to repopulate each primary oocyte. (b) Multiple-sampling model. Using an adaptation of the population genetic model18,16 to our model during the linear phase (after 9 d.p.c.) shows that the relaxed replication of mtDNA contributes considerably to the variation in heteroplasmy levels, particularly during the rapid proliferation of mtDNA molecules in the expanding germ line. Using the model we describe here, the segregation of replicating mtDNA molecules generates sufficient variation in heteroplasmy levels to explain our observations in heteroplasmic mice, indicating that the mtDNA molecule itself is the segregating unit during transmission. Additional variation could arise through the variation in copy number, which was not included in the model. Including these factors would only add weight to our conclusions.

Why is there a mitochondrial genetic bottleneck? mtDNA lacks protective histones and is close to the respiratory chain, a potent source of mutagenic free radicals. Mitochondria are also relatively deficient in DNA repair mechanisms, contributing to the mutation rate of mtDNA. The mitochondrial genetic bottleneck leads to the rapid segregation of new genotypes, which either are lost during transmission or reach very high levels and affect fecundity. This facilitates the rapid removal of deleterious mtDNA mutations from the population, by ‘purifying’ germline mtDNA, and thus prevents the ‘mutational meltdown’ predicted by Muller’s ratchet24. This is not only important for the female germ line, but it is also relevant for the male germ line, where mtDNA mutations can alter sperm motility25 and hence affect fertility. Notably, at 14.5 d.p.c., female PGCs (n = 83, mean mtDNA copies = 1,376, s.d. = 601) contained lower amounts of mtDNA than male PGCs (n = 43, mean mtDNA copies = 2,152, s.d. = 951, two-sample t-test P < 0.001, Fig. 1c), possibly because a stringent bottleneck is more important for the female germ line, with the potential transmission of mutations to subsequent generations. It is also of interest that we observed a range of values for the minimum amount of mtDNA within PGCs (Fig. 1c). This could reflect slight differences in the developmental stage or asymmetric compartmentalization of mtDNA during development. Subtle differences in the parameters governing the genetic bottleneck, such as the timing of the initiation of mtDNA replication, provide an explanation for the apparent difference in bottleneck size seen in...
different human pedigrees transmitting pathogenic mtDNA mutations\(^2\), with some women having offspring with similar levels of heteroplasmy\(^3\) and others having offspring who are markedly different in this regard\(^1\).

**METHODS**

**Isolation of single embryonic cells.** All animal procedures were performed, under license, in accordance with the UK Home Office Animal Act (1986). C57BL/6J.CBA F1 female mice were superovulated by sequential intraperitoneal injection of 5 international units (IU) of pregnant mares’ serum (PMS) (Sigma-Aldrich) and 7.5 IU of human chorionic gonadotrophin (hCG) (Sigma-Aldrich) 48 h apart. Unfertilized eggs were collected 12 h after hCG injection. One-cell zygotes were collected after successful mating with CBA males 24 h after hCG injection. Single-cell embryos were recovered by flushing the oviducts with prewarmed M2 medium (Sigma-Aldrich) as described\(^27\). The freshly recovered embryos were transferred to prewarmed M16 media (Sigma-Aldrich) and maintained at 37 °C in a humidified atmosphere containing 5% CO\(_2\) until they reached the cleavage stage required. Embryos were transferred to sterile PCR tubes and lysed for 16 h in 50 mM Tris-HCl, pH 8.5, 100 mM NaCl, 1 mM EDTA and 0.5% Tween 20 and 100 α-enolase (Sigma-Aldrich) and maintained at 37 °C, followed by heat inactivation at 95 °C for 10 min, a protocol shown previously to maximize the mtDNA yield. Where individual blastomeres were required, the zona pellucida was removed with acid Tyrode’s solution (pH 3.5). The zona-free embryos were then incubated for 10–20 min at 37 °C in calcium- and magnesium-free phosphate-buffered saline. Individual blastomeres were obtained by repeatedly passing the zona-free embryos through a micropipette and lysed as described previously.

**Quantification of mtDNA copy number.** Absolute quantification of mtDNA copy number was performed by real-time PCR using iQ5Syr Green on the Bio-Rad IQcycler to a target template spanning nt12789–nt12876 of the MTND5 gene. Absolute quantification was performed by the standard curve method. A PCR-generated template was created spanning the mitochondrial genome between nt12705 and nt13834. The template was purified by gel extraction (Qiagen), quantified by UV absorbance at 260 nm and serially diluted to generate a standard curve for quantification of mtDNA content in samples. All samples and standards were measured in triplicate. Before studying mouse embryos, we compared three different real-time PCR assays directed at different regions of the mitochondrial genome: ND5 (nt12789–nt12876), ND4 (nt11031–nt1174), and ND1 (nt2751–nt3709). The results obtained from each assay were tightly correlated (r\(^2\) > 0.99) over a wide range of mtDNA concentrations, encompassing the values obtained in subsequent single cell studies (Supplementary Fig. 3 online). We arbitrarily chose the ND5 assay for subsequent experiments.

**Isolation of primordial germ cells.** Traditionally, primordial germ cells (PGCs) have been identified by their characteristic tissue nonspecific alkaline phosphatase (TNP) staining\(^2\). However, TNP is also present in somatic cells that surround the PGCs, and the TNP staining procedure inhibits the mtDNA quantification assay. We initially studied unfixed TNP-stained cryostat sections, capturing PGCs by laser microdissection. However, copy number measurements were highly variable, partly because the reaction product inhibited the real-time PCR reaction and partly because the reaction product diffused out of the PGCs into the adjacent tissue, making the identification of PGCs difficult. All postimplantation studies were therefore carried out in Stella-GFP transgenic mice to allow unambiguous detection of the PGCs. C57BL/6J.CBA F1 females were mated with Stella-GFP BAC-homozygous C57BL/6J.CBA males\(^15\). Noon on the day of vaginal plug detection was designated 0.5 d.p.c. Stella-GFP heterozygous embryos were collected at 7.5 d.p.c., 8.5 d.p.c., 10.5 d.p.c. and 14.5 d.p.c. in Dulbecco’s minimal essential medium (DMEM, Invitrogen) supplemented with 7.5% FCS and 10 mM HEPES. The posterior parts of the 7.5-d.p.c. and 8.5-d.p.c. embryos and the gonadal ridges of 10.5-d.p.c. and 14.5-d.p.c. embryos were dissected using tungsten needles, pooled by age and trypsinized for 15 min at 37 °C. Morphological gender typing was possible after 12.5 d.p.c. The tissue was resuspended under the microscope using an equal volume of FCS and centrifuged for 3 min at 2,000g, and the final pellet was resuspended in DMEM with 7.5% FCS and 10 mM HEPES. Samples were kept on ice until FACs sorting.

Single PGCs and sets of 5, 10, 50 and 100 PGCs were unidirectionally sorted into single wells of a 96-well plate using a BD FACSARia (Becton Dickinson). GFP was detected using a 100-nW sapphire laser and GFP-positive PGCs sorted at 20 p.s.i. using a 100-μm nozzle at a sort rate of 2,000 events per second. Instrument sensitivity was proved stable between sorts by internal QC procedures. Plates were stored at –80 °C until required. PGCs were lysed and mtDNA content quantified as described above.

**Generation of heteroplasmic mice.** Initially mouse strains were backcrossed to maximize reproductive performance and maintain a pure C57BL/6J or NZB mtDNA genotype. Thus, C57BL/6J.C3H.F1 females were crossed with C57BL/6J.C3H.F1 males, and the resulting F2 females backcrossed with C57BL/6J.C3H.F1 males. The F1 mice therefore had a C57BL/6J mtDNA genotype on a mixed C57BL/6J, C3H and NZB nuclear background. These were designated ‘C57mt’ mice. Likewise, NZB.NZW F1 females (obtained from the Jackson Laboratory) were crossed with C57BL/6J.C3H.F1 males, and the F2 females backcrossed with C57BL/6J.C3H.F1 males. The F3 mice had a NZB mtDNA genotype on a mixed C57BL/6J, C3H and NZB nuclear background and were designated ‘NZBmt’ mice.

Fertilized oocytes were obtained from both C57mt and NZBmt mice using standard techniques. Approximately 10–30% of cytoplasm was removed by
micropipette from a C57mt oocyte, and an equivalent amount of cytoplasm was removed from an NZBmt oocyte. The cytoplasm removed from the C57mt oocyte was microinjected under the zona pellucida of the NZBmt oocyte and vice versa. The donor and recipient cytoplasts were fused by electroporation, and oocytes were implanted into a pseudo-pregnant dam. Seven founder mice (four females and three males) were obtained from the transfer of cytoplasm from an NZBmt oocyte to a C57mt oocyte. Four founder mice (three females and one male) were obtained from the transfer of cytoplasm from a C57mt oocyte to an NZBmt oocyte. The C57 mt founder females were mated to C57BL/6j males to breed heteroplasmic mice with a C57mt nuclear background. The NZBmt founder females were mated with NZB males, but pups were not obtained from any of these females.

Quantification of levels of heteroplasmy. The transmission of heteroplasm between a dam and her pups was investigated by measuring the percentage levels of NZB mtDNA in tail biopsies at weaning. There was no difference in the percentage level of NZB mtDNA at this stage between different tissues. Tail biopsies were taken by removing approximately 1 cm of tail with a sterile scalpel blade under anesthesia. Genomic DNA was extracted by incubating the 1.5 ml Salting Out Lysis Buffer (10 mM Tris-Cl, pH 7.5; 0.4 M NaCl; 2 mM EDTA), 200 µl 20% SDS and 250 µl of 5 mg/ml proteinase K (Boehringer Mannheim) for 16 h at 50 °C. Protein was extracted using 25:24:1 phenol/chloroform/isoamyl alcohol and then 24:1 chloroform/isoamyl alcohol. An ethanol precipitation was performed on each sample and the DNA was redissolved in sterile distilled, deionized water. The concentration was quantified by measuring the absorbance at 260 nm. PCRs were performed in 50-µl reaction volumes containing 1× Taq Polymerase Reaction Buffer (Boehringer Mannheim), 0.2 mM dNTPs, 0.5 µg each of primers m3558-F and m3940-R (see Supplementary Table 2 online for primer sequences), 50 ng of genomic DNA and 2.5 units of Taq Polymerase (Boehringer Mannheim). The reactions were overlaid with 100 µl of paraffin oil and amplified for 25 cycles of 95 °C for 1 min, 51 °C for 1 min and 72 °C for 1 min on a Corbett Research FTS-320 Thermal Sequencer. The PCR was paused at the start of the 25th cycle (at 95 °C), 10 µCi of [33P]dCTP was added to each reaction, and the 25th cycle was then completed.

Restriction endonuclease digests were performed in 20-µl volumes containing 15 µl of PCR volume, 1× Buffer L (Boehringer Mannheim) and 1.5 units of Rsal restriction endonuclease (Boehringer Mannheim). The digests were incubated at 37 °C for 16 h, and 5 µl of 10% gel loading buffer was added to stop each reaction. Rsal cuts C57BL/6j mtDNA twice to yield fragments of 218 base pairs (bp), 132 bp and 32 bp, and cuts NZB mtDNA once to yield fragments of 350 bp and 32 bp. Undigested mtDNA can be detected by the presence of a 382-bp fragment.

A 10-µl aliquot of each sample was electrophoresed through an 8% nondenaturing polyacrylamide gel for 2 h at 50 V. The gel was soaked in fixing solution (10% glacial acetic acid and 10% methanol) for 10 min, placed on blotting paper (Schleicher & Schuell) and dried at 80 °C for 2 h. The dried gels were placed against a phosphorimagery screen (Molecular Dynamics) for 2 d, and the screen was scanned using a Storm PhosphorImager (Molecular Dynamics). The intensities for the 350-bp (NZB mtDNA) and the 218-bp and 132-bp (C57BL/6j mtDNA) fragments were quantified using ImageQuant software (Molecular Dynamics) and their heteroplasmy levels quantified.

Modeling. Previous studies8 used an adapted population genetic model10 to model the random segregation of mtDNA genotypes over a number of cell divisions, where the number of copies of mtDNA remained constant over the cell generations. Our experimental observations of the mtDNA copy number per cell showed that the copy number is not constant. To account for the varying mtDNA content of different germline precursors, we simulated the process of mtDNA segregation directly based on the population genetic principles of random segregation. However, by directly simulating the process, we were able to alter the number of mtDNA copies at each stage of PGC development, reflecting our laboratory observations.

The embryogenesis simulation began with a single cell, the fertilized oocyte, that underwent a series of rapid cell divisions, during which mtDNA replication did not occur. Cell divisions were set to occur at constant intervals of 15 h, in agreement with experimental values (Supplementary Fig. 1). The existing mtDNA molecules were simply partitioned randomly between the two daughter cells at each cell division, and the mtDNA content per cell therefore dropped exponentially during this phase, to a minimum of 314 ± 24 in accordance with the experimental values (Table 2). To generate the >200,000 mtDNAs we observed in a mature oocyte (Table 1), mtDNA replication must resume at some point during oocyte development and markedly increase the number of copies per cell over a sustained period of exponential mtDNA proliferation20,21. In our simulation, mtDNA replication was initiated at 7.5 d.p.c. This value was chosen to simulate the median number of mtDNA molecules that we measured at 7.5 d.p.c. PGCs. It should be noted that there was considerable variation in the measured amount of mtDNA within single PGCs at this time point, with a mean value that was greater than the median. The observed variation in copy number at this point was not incorporated into the model.

At this stage, after ten cell divisions, the total number of cells was 1,024. We modeled the development of a specialized PGC lineage at this stage by randomly choosing 40 of these simulated cells, while discarding the others from the simulation. These 40 cells continued to divide with a period of 15 h. We followed this process over a further 10 cell divisions, to a final population at 13.4 d.p.c.

The initial cell in the simulation had a heteroplasmic mtDNA population, set at 50% mutant and 50% wild type. The random processes of distribution of wild-type and mutant mtDNA to daughter cells, mtDNA degradation and mtDNA replication all caused the heteroplasmy level in each simulated cell to drift over time, generating a heteroplasmy variance across all of the simulated cells. During the first phase of embryogenesis, while the mtDNA copy number per cell was still relatively high, very little heteroplasmy variance was generated. As the mtDNA copy number per cell fell to a few hundred copies, heteroplasmy variance began to rise. After mtDNA replication began, and the PGC line was formed, mtDNA copy number per cell rose again, to approximately 1,500 per cell. The heteroplasmy variance slowed its rate of increase, but it continued to increase at a steady rate as long as the PGC cell divisions continue at a rapid pace.

We used the actual measurements of mtDNA content within single cells (Tables 1 and 2) to simulate the formation of PGCs in mothers with different heteroplasmy values ranging from 1% to 100% (Fig. 3c). Each simulation was carried out ten times, resulting in a total of 204,800 cells, allowing us to determine the 95% CI for the heteroplasmy value in individual PGCs for that mother. The upper and lower 95% CI plotted on Figure 3c (solids circles) show the likely range of heteroplasmy values across the whole range (solids lines). At birth there was no difference in the proportion of NZB mtDNA between different tissues, but postnatal segregation of mtDNA genotypes led to significant differences between the tail mtDNA and other tissues in adult mice (S.L. White, W. Hutchinson, D.R. Thorburn, V.A. Collins, K. Fowler & N.M.C. unpublished data and ref. 30). Given that the transmission of NZB mtDNA heteroplasmy is determined by random genetic drift, with an equal likelihood of an increase or decrease in the proportion of mtDNA, we estimated the percentage NZB in the maternal primordial germ cell founders as the mean level of heteroplasmy from the offspring at birth.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

P.E.C. is a Wellcome Trust Senior Fellow in Clinical Science and also receives funding from the United Mitochondrial Diseases Foundation and from the EU FP6 program EUmitocombat and MITOCIRCLE. H.-H.M.D. is a National Health and Medical Research Council (Australia) Principal Research Fellow, and his affiliations are with The Murdoch Children’s Research Institute and the Department of Paediatrics (University of Melbourne), Royal Children’s Hospital, Melbourne, Australia. We thank I. Dimmick for his assistance with the flow cytometry, D. Turnbull and B. Lightowlers for discussions, A. McLaren for her expertise on the cell dynamics of mouse development, and M. Azim Surani for providing the Stella-GFP mice. We also thank D. Thorburn for discussions and advice while studying the heteroplasmic mice, and both W. Hutchinson and S. White for experimental work on the heteroplasmic mice.

AUTHOR CONTRIBUTIONS

This laboratory study was designed by P.E.C. and L.M.C. and carried out by L.M.C. The in silico modeling was designed by D.C.S., programmed by H.K.R. and carried out by D.C.S., H.K.R. and P.W. GFP-Stella mice were produced in the

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

**Supplementary information**

**Supplementary Table 1** Variation in the level of heteroplasmy between the offspring of heteroplasmic female mice.

<table>
<thead>
<tr>
<th>Proportion of NZB genotype in mother</th>
<th>Number of offspring</th>
<th>Variance in heteroplasmy amongst offspring</th>
<th>Coefficient of variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>9</td>
<td>0.0001</td>
<td>0.9274</td>
</tr>
<tr>
<td>0.01</td>
<td>17</td>
<td>0.0000</td>
<td>0.6856</td>
</tr>
<tr>
<td>0.03</td>
<td>9</td>
<td>0.0086</td>
<td>3.0872</td>
</tr>
<tr>
<td>0.05</td>
<td>11</td>
<td>0.0420</td>
<td>4.0999</td>
</tr>
<tr>
<td>0.1</td>
<td>6</td>
<td>0.0135</td>
<td>1.1605</td>
</tr>
<tr>
<td>0.11</td>
<td>11</td>
<td>0.0068</td>
<td>0.7478</td>
</tr>
<tr>
<td>0.16</td>
<td>8</td>
<td>0.0266</td>
<td>1.0475</td>
</tr>
<tr>
<td>0.16</td>
<td>4</td>
<td>0.0281</td>
<td>1.0201</td>
</tr>
<tr>
<td>0.16</td>
<td>16</td>
<td>0.0227</td>
<td>0.9412</td>
</tr>
<tr>
<td>0.18</td>
<td>15</td>
<td>0.0065</td>
<td>0.3219</td>
</tr>
<tr>
<td>0.18</td>
<td>7</td>
<td>0.0034</td>
<td>0.4462</td>
</tr>
<tr>
<td>0.26</td>
<td>20</td>
<td>0.0199</td>
<td>0.5428</td>
</tr>
<tr>
<td>0.27</td>
<td>10</td>
<td>0.0066</td>
<td>0.3013</td>
</tr>
<tr>
<td>0.28</td>
<td>10</td>
<td>0.0244</td>
<td>0.4586</td>
</tr>
<tr>
<td>0.28</td>
<td>4</td>
<td>0.0165</td>
<td>0.5579</td>
</tr>
<tr>
<td>0.3</td>
<td>5</td>
<td>0.0215</td>
<td>0.4882</td>
</tr>
<tr>
<td>0.31</td>
<td>10</td>
<td>0.0233</td>
<td>0.4921</td>
</tr>
<tr>
<td>0.34</td>
<td>8</td>
<td>0.0115</td>
<td>0.3155</td>
</tr>
<tr>
<td>0.35</td>
<td>18</td>
<td>0.0208</td>
<td>0.4122</td>
</tr>
<tr>
<td>0.4</td>
<td>17</td>
<td>0.0093</td>
<td>0.2411</td>
</tr>
<tr>
<td>0.48</td>
<td>17</td>
<td>0.0125</td>
<td>0.2333</td>
</tr>
<tr>
<td>0.58</td>
<td>14</td>
<td>0.0118</td>
<td>0.1869</td>
</tr>
</tbody>
</table>
Supplementary Table 2 Oligonucleotide primer sequences used to quantify NZB/C57BL.6J mtDNA heteroplasmy

<table>
<thead>
<tr>
<th>Forward primer</th>
<th>Reverse primer</th>
</tr>
</thead>
<tbody>
<tr>
<td>5′-ACATTCCTATGGATCGAGC-3′</td>
<td>5′-GATGATGGCAAGGTGATAG-3′</td>
</tr>
</tbody>
</table>
Supplementary Figure 1 Comparison between the number of cells simulated and the actual number of cells observed experimentally. The rate of cell division in pre-implantation embryos was derived from Ref 1, and for post-implantation embryos from Refs 2,3, based upon published values for alkaline phosphatase stained primordial germ cells (PGCs). PGCs are first discernable at 7.25 days post coitus (dpc) during the mid primitive streak stage, just posterior to the primitive streak in the extra-embryonic mesoderm. Note the logarithmic Y-axis indicating exponential growth to a final figure of 25,791 primary oocytes at day 13.5 3.
Supplementary Figure 2 The total amount of mtDNA in the simulated mouse embryos up to 7dpc, followed by the total amount of mtDNA in the entire PGC population. Dpc = days post coitus.
Supplementary Figure 3 Comparison of three real-time PCR assays targeting different regions of the mitochondrial genome: ND5 (nt12789 to nt12876), ND4 (nt1031 to nt11174), and ND1 (nt2751 to nt3709). Total genomic DNA was extracted from the tail of a C57Bl6 mouse and serially diluted. Each dilution was assayed in quadruplicate using the same methods employed in the manuscript to determine the absolute amount of mtDNA in each sample. The graphs shows two independent serial dilutions of a genomic DNA template serially diluted and measured by two different assays: (a) comparison of ND4 with ND5, (b) comparison of ND1 with ND5, (c) comparison of ND4 with ND1. Pearson’s correlation co-efficient showed a tight correlation in the measured copy number values using each assay ($R^2 > 0.99$). The ND5 assay was used in the experiments described in the manuscript.
a)

\[ R^2 = 0.9999 \]

\[ R^2 = 0.9991 \]

b)

\[ R^2 = 0.9997 \]

\[ R^2 = 0.9931 \]

c)

\[ R^2 = 0.9963 \]

\[ R^2 = 0.9990 \]
References for the supplementary material


3 Chapter 3: The Distribution of Mitochondrial DNA Heteroplasmy due to Random Genetic Drift

Passorn Wonnapinij¹, Patrick F. Chinnery², and David C. Samuels¹

¹ Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, VA, 24061, USA
² Mitochondrial Research Group and Institute of Human Genetics, Newcastle University, The Medical School, Newcastle-upon-Tyne NE2 4HH, UK

Reprinted from “The American Journal of Human Genetics 2008, 83, 582-593” with permission from Elsevier

To whom correspondence should be conducted: dsamuels@vbi.vt.edu
The Distribution of Mitochondrial DNA Heteroplasmy Due to Random Genetic Drift

Passorn Wonnapinij,1 Patrick F. Chinnery,2 and David C. Samuels1,*

Cells containing pathogenic mutations in mitochondrial DNA (mtDNA) generally also contain the wild-type mtDNA, a condition called heteroplasmy. The amount of mutant mtDNA in a cell, called the heteroplasmy level, is an important factor in determining the amount of mitochondrial dysfunction and therefore the disease severity. mtDNA is inherited maternally, and there are large random shifts in heteroplasmy level between mother and offspring. Understanding the distribution in heteroplasmy levels across a group of offspring is an important step in understanding the inheritance of diseases caused by mtDNA mutations. Previously, our understanding of the heteroplasmy distribution has been limited to just the mean and variance of the distribution. Here we give equations, adapted from the work of Kimura on random genetic drift, for the full mtDNA heteroplasmy distribution. We describe how to use the Kimura distribution in mitochondrial genetics, and we test the Kimura distribution against human, mouse, and Drosophila data sets.

Introduction

Mitochondrial DNA (mtDNA) encodes several subunits of the electron transfer chain. Defects in human mtDNA cause a wide range of disease conditions, mainly resulting from the impairment of ATP production in the cell. Some examples of the inheritable pathogenic point mutations in mtDNA are the m.3243A > G (MIM #590050.0001) mutation causing mitochondrial encephalomyopathies lactic acidosis and stroke-like episodes (MELAS, MIM #540000),1 the m.8344A > G (MIM #590060.0001) mutation causing mitochondrial encephalomyopathies lactic acidosis and stroke-like episodes (MELAS, MIM #540000),1 the m.8993T > G (MIM #516060.0001) mutation causing myoclonic epilepsy with ragged-red fiber (MERRF, MIM #545000),2 the m.8834G > A (MIM #590050.0001) mutation causing Leber’s hereditary optic neuropathy (LHON, MIM #535000).3,5–7

Any individual cell contains many copies of the mitochondrial genome. The mtDNA copy number per cell ranges from a few hundred to a few hundred thousand copies. Generally, cells containing a pathogenic mtDNA mutation also contain the wild-type genome, a condition called heteroplasmy. Important exceptions to this rule are the mitochondrial diseases such as LHON, which have a low penetrance of the disease phenotype within families carrying the mutation. Individuals may be homoplasmic for these particular pathogenic mutations, often while remaining asymptomatic, and this is generally attributed to the lack of some necessary pathogenesis cofactor, either genetic or environmental.

MtDNA is transmitted through the maternal lineage in humans.8 In pedigrees with an inheritable heteroplasmic mtDNA mutation, the measured heteroplasmy level often shifts by large and apparently random amounts between mother and offspring.9–11 These variations cause complications in estimating the recurrence risks of these genetic diseases and therefore in giving accurate genetic counseling to a female carrying a pathogenic mtDNA mutation.12–14

The inheritance of mtDNA heteroplasm is described by the expected probability distribution of heteroplasmy values in a sibling group. Until now, our ability to predict heteroplasmy distributions has been limited to predicting the mean value and the variance, the two lowest-order statistics. On the basis of neutral genetic drift and standard haploid population genetics, we have been able to predict that the mean heteroplasmy in the offspring should be equal to the mother’s heteroplasmy and that the variance of the offspring heteroplasmy should have the following form15:

\[
V(t) = \frac{p_0(1-p_0)}{N_{eff}} \left[ 1 - e^{-t/N_{eff}} \right] 
\]

The variance of heteroplasmy, \(V\), in a group of individuals with a single common maternal ancestor after \(t\) generations can be calculated from the initial gene frequency, \(p_0\), and the effective population size, \(N_{eff}\). This variance equation is generally referred to in this field as the Sewell-Wright formula. We note again that these equations are based on the assumption of random genetic drift.

Although the mean and variance of the heteroplasmy distribution in a population is useful information, it is very limited information. It does not give us the heteroplasm distribution itself. In particular, this is a problem if the heteroplasmy distribution is not symmetric, which must be the case at high and low heteroplasmy levels, two extremes of enormous practical importance. Ideally, we would want to be able to predict the entire heteroplasmy probability distribution. Fortunately, this problem was solved in 1955 by Motoo Kimura.16 His solution was for gene frequency probabilities in diploid populations, but the application of this theory to mitochondrial heteroplasm is straightforward. The variance in Equation 1 can

---

1Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061, USA; 2Mitochondrial Research Group and Institute of Human Genetics, Newcastle University, The Medical School, Newcastle-upon-Tyne NE2 4HH, UK
*Correspondence: dsamuels@vbi.vt.edu
DOI 10.1016/j.ajhg.2008.10.007. ©2008 by The American Society of Human Genetics. All rights reserved.

582 The American Journal of Human Genetics 83, 582–593, November 7, 2008
be derived from Kimura’s theory, so the full Kimura theory does not displace the previous work that has been done in mitochondrial genetics on the basis of this variance equation. Instead, it greatly extends our capability to calculate the full heteroplasmy distribution.

Kimura derived a set of probability distribution functions to explain the gene frequency distribution of populations under pure random genetic drift. The underlying assumptions of this derivation are nonoverlapping generations, no selection, no migration, no de novo mutation, and a finite and steady population size. Kimura made the assumption of a constant population size to simplify the mathematics. Other work has shown that this assumption is not necessary. If the population size is allowed to vary, either through fluctuations or through events such as population bottlenecks, then the definition of the effective population size in terms of the actual population size becomes complicated. That complication does not concern us here because we will treat the effective population size, Neff, merely as a parameter of the model. The solution of this model consists of three equations: a probability distribution on the wild-type or mutant, and a probability distribution function for losing an allele, a probability distribution function for fixing on that allele, and a probability distribution function that the allele is present at frequency x in the population.

\[ f(0,t) = (1-p_0) + \sum_{i=1}^{n} (2i+1)p_0(1-p_0)(-1)^i F(1-i,i+2,1-p_0)e^{-\frac{(i+1)}{2N_{eff}}t} \]

\[ \phi(x,t) = \sum_{i=1}^{n} i(i+1)(2i+1)p_0(1-p_0)F(1-i,i+2,2,x) F(1-i,i+2,2,p_0)e^{-\frac{(i+1)}{2N_{eff}}t} \]

\[ f(1,t) = p_0 + \sum_{i=1}^{n} (2i+1)p_0(1-p_0)(-1)^i F(1-i,i+2,2,p_0)e^{-\frac{(i+1)}{2N_{eff}}t} \]

The meaning of each variable in these equations is the same as for the Sewall-Wright variance formula (Equation 1). The interpretation in terms of mitochondrial heteroplasmy is straightforward; p0 is the mtDNA heteroplasmy level in the maternal lineage founder and is also the mean heteroplasmy in the offspring distribution, f(0,t) and f(1,t) are the probability distributions on the wild-type or mutant, respectively, in generation t, and x is the offspring heteroplasmy level. The function F(1-i,i+2,2,x) is the hypergeometric function. For simplicity, we will refer to Equations 2-4 as the Kimura distribution. Because this is a probability distribution, the integration of all three terms is equal to unity.

\[ f(0,t) + \int_{0}^{1} \phi(x,t)dx + f(1,t) = 1 \]

Although the mathematical form of the Kimura distribution is certainly complicated, and although care must be taken in the numerical calculation of these equations, the distribution values can be calculated. In this paper, we apply the Kimura distribution to measurements of the mtDNA heteroplasmy distributions in humans, mice, and Drosophila.

**Material and Methods**

**Experimental Data**

The observed heteroplasmy distributions used in this paper have been collected from several sources in the published literature. For experimental data that were available only in graphical form, we used the software Engauge Digitizer to determine approximate numerical values. The experimental data sets analyzed here covered three organisms; human, mouse, and Drosophila. The human study protocol was approved by the participating institutional review boards.

**Setting the Parameter Values for Kimura’s Probability Distribution**

The variance formula as it is normally written is a function of three parameters; p0, t, and Neff. However, the form of the equations...
allows us to combine the t and Neff parameters into a single parameter that we call b, as follows.

\[ V = p_0(1 - p_0)[1 - e^{-t/N_{\text{eff}}}] = p_0(1 - p_0)(1 - b) \] (6)

The new parameter b is then defined as

\[ b = e^{-t/N_{\text{eff}}} \] (7)

Substituting these parameters into the Kimura probability density functions simplifies them to a two-parameter model, with parameters p_0 and b, which both range from zero to one.

\[ f(0) = (1 - p_0) + \sum_{i=1}^{2l+1} \frac{(2l+1)p_0(1-p_0)(-1)^i}{i!F(1-i,i+2,2,1-p_0)b^{i+1/2}} \] (8)

\[ \phi(z) = \sum_{i=1}^{x} i(1+i)2(l+1)p_0(1-p_0)F(1-i,i+2,2,x) \] (9)

\[ f(1) = p_0 + \sum_{i=1}^{x} (2l+1)p_0(1-p_0)(-1)^iF(1-i,i+2,2,p_0)b^{i+1/2} \] (10)

Given a data set of mtDNA heteroplasmy values for a set of individuals arising from a common founder, we can fit a Kimura probability distribution to the heteroplasmy values by determining the values for the two parameters p_0 and b. These two parameters can be determined from the two lowest-order statistics of the data set; the mean and the variance. We take the parameter p_0 to be equal to the mean heteroplasmy value of the data set.

Then we can use Equation 6 to determine the parameter b from p_0 and the variance of the data. The entire data set, including heteroplasmy values fixed at the two extremes of 0 and 1, is used in the calculation of the variance and p_0 and then is used in the calculation of b.

### Calculating the Numerical Value of the Hypergeometric Function

Accurately calculating the numerical value of the hypergeometric function F(a,b,c,z) is a difficult technical problem. Because this is a fundamental mathematical function, this issue has been faced in many different scientific fields. Recently, as a solution to this problem occurring in a spectroscopy application, Hoang-Binh developed an accurate and practical algorithm for the numerical calculation of hypergeometric functions, and we have followed this method. This method uses the following recurrence relation:

\[ F(-1) = F(-1,b,c,z) = 1 - (bz/c) \] (11)

\[ F(0) = F(0,b,c,z) = 1 \] (12)

\[ (a - c)F(a - 1) = a(1 - z)[F(a) - F(a + 1)] + (a + bz - c)F(a) \] (13)

---

**Table 1. The Parameters Estimated from Experimental Data: The Mean Heteroplasmy, p_0, and the b Parameter Calculated from the Variance and the p Value Calculated from the KS Test**

<table>
<thead>
<tr>
<th>Organism</th>
<th>Lineage</th>
<th>Sample</th>
<th>Generations</th>
<th>Mean Heteroplasmy, p_0</th>
<th>Variance</th>
<th>N</th>
<th>b</th>
<th>KS Test p Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Humans</td>
<td>N/A</td>
<td>primary oocytes</td>
<td>-</td>
<td>0.1264</td>
<td>0.01432</td>
<td>82</td>
<td>0.8705</td>
<td>0.827</td>
</tr>
<tr>
<td>Mice</td>
<td>515</td>
<td>tail biopsy</td>
<td>1</td>
<td>0.051</td>
<td>0.00240</td>
<td>41*</td>
<td>0.9505</td>
<td>0.666</td>
</tr>
<tr>
<td></td>
<td>515</td>
<td>mature oocytes</td>
<td>-</td>
<td>0.047</td>
<td>0.00646</td>
<td>31</td>
<td>0.8558</td>
<td>0.049*</td>
</tr>
<tr>
<td></td>
<td>517</td>
<td>tail biopsy</td>
<td>1</td>
<td>0.076</td>
<td>0.00447</td>
<td>43</td>
<td>0.9363</td>
<td>0.75</td>
</tr>
<tr>
<td></td>
<td>517</td>
<td>mature oocytes</td>
<td>-</td>
<td>0.094</td>
<td>0.0119</td>
<td>26</td>
<td>0.8604</td>
<td>0.834</td>
</tr>
<tr>
<td></td>
<td>603A</td>
<td>mature oocytes</td>
<td>-</td>
<td>0.009</td>
<td>0.000185</td>
<td>49</td>
<td>0.9793</td>
<td>0.681</td>
</tr>
<tr>
<td></td>
<td>603A</td>
<td>primary oocytes</td>
<td>-</td>
<td>0.011</td>
<td>0.00023</td>
<td>49*</td>
<td>0.9789</td>
<td>0.037*</td>
</tr>
<tr>
<td></td>
<td>603B</td>
<td>mature oocytes</td>
<td>-</td>
<td>0.031</td>
<td>0.00098</td>
<td>31</td>
<td>0.9674</td>
<td>0.435</td>
</tr>
<tr>
<td></td>
<td>603B</td>
<td>primary oocytes</td>
<td>-</td>
<td>0.025</td>
<td>0.000638</td>
<td>46*</td>
<td>0.9738</td>
<td>0.223</td>
</tr>
<tr>
<td>Drosophila mauritiana</td>
<td>H1</td>
<td>unfertilized eggs</td>
<td>30</td>
<td>0.4150</td>
<td>0.18670</td>
<td>60</td>
<td>0.2310</td>
<td>0.004**</td>
</tr>
<tr>
<td></td>
<td>H1-31M</td>
<td>-</td>
<td>3</td>
<td>0.1784</td>
<td>0.00952</td>
<td>59</td>
<td>0.9350</td>
<td>0.587</td>
</tr>
<tr>
<td></td>
<td>H1-18D</td>
<td>-</td>
<td>3</td>
<td>0.4763</td>
<td>0.01711</td>
<td>31</td>
<td>0.9314</td>
<td>0.993</td>
</tr>
<tr>
<td></td>
<td>H1-12B</td>
<td>-</td>
<td>5</td>
<td>0.8155</td>
<td>0.02791</td>
<td>52</td>
<td>0.8146</td>
<td>0.865</td>
</tr>
<tr>
<td></td>
<td>G20-5</td>
<td>-</td>
<td>3</td>
<td>0.3250</td>
<td>0.01257</td>
<td>55</td>
<td>0.9430</td>
<td>0.922</td>
</tr>
<tr>
<td></td>
<td>G71-12</td>
<td>-</td>
<td>3</td>
<td>0.5834</td>
<td>0.01804</td>
<td>50</td>
<td>0.9258</td>
<td>0.542</td>
</tr>
<tr>
<td>Drosophila simulans</td>
<td>6YF16</td>
<td>-</td>
<td>3</td>
<td>0.1270</td>
<td>0.0096</td>
<td>44</td>
<td>0.9131</td>
<td>0.79</td>
</tr>
</tbody>
</table>

The "generations" column gives the number of organism generations, and thus this value is not given for samples of mature oocytes or primary oocytes. The p_0 parameter is obtained from the average of the heteroplasmy measurements. N is the number of samples from the experiment. The p value is the level of significance for the null hypothesis that the experimental heteroplasmy distribution matches the Kimura distribution. The asterisks indicate significance levels: * 0.01 < p < 0.05 and ** * p < 0.01.

* In these three cases, we found discrepancies between the number of samples listed in the cited paper and the number of samples actually given in the data set.
Numerical Calculation of the Kimura Probability Distributions

The infinite series in Equations 8–10 were truncated when the difference between the $i+1$ and $i$ terms became less than $10^{-4}$. Note that the infinite series in Equations 8–10 have oscillating sign terms, so numerical convergence of these series is slow. We tested the accuracy of the resulting probability distributions by calculating the integral in Equation 5; this integral which should be unity. The difference of the numerical calculation from unity in the results presented here was typically on the order of $10^{-5}$, and the maximum difference was less than 0.004. All calculations were carried out in C programs, which are available from the authors (details are given in Web Resources).

Statistical Test

We applied the Kolmogorov-Smirnov (KS) test to compare the experimental data for mtDNA heteroplasmy distributions to the Kimura probability distributions. Because the parameters for the theoretical Kimura probability distributions were determined from the statistics of the experimental data sets, the $p$ values of this comparison had to be determined from Monte-Carlo simulations. For the Monte-Carlo simulations, 1000 simulated data sets with the same population size as the experimental data set were drawn from the theoretical distribution, and the $p$ values were determined from the fraction of simulated data sets whose maximum deviation from the theoretical probability distribution was larger than the maximum deviation of the experimental data set.

Results

The Kimura distribution represents the distribution of heteroplasmy that develops through random genetic drift in a population of cells or individuals who all are descended by an equal number of generations from a single heteroplasmic progenitor cell or individual. To compare the Kimura distribution to experimental data, we need data sets that satisfy this condition and also contain a large number of individual heteroplasmy measurements so that the probability distribution of the heteroplasmy measurements can be determined. From a search of the literature, we identified four publications containing a total of 16 data sets to analyze. For each data set, we used the mean and variance of the heteroplasmy measurements to set the $p_0$ and $b$ parameters in the Kimura distribution as described in the Material and Methods. Then histograms of the measured heteroplasmy distributions were compared to histograms of the fit Kimura distributions. Finally, a KS test comparing the cumulative probability distributions of the experimental data with the theoretical Kimura probability distributions was carried out. The 16 data sets analyzed consisted of one human data set, eight mouse data sets, and seven Drosophila data sets.

Figure 2. The Measured Heteroplasmy Distribution from Offspring and Mature Oocytes in the Heteroplasmic Mouse Line 515 Is Compared to the Kimura Distribution

(A) The heteroplasmy frequency histogram from the offspring.
(B) KS test comparing the offspring heteroplasmy data to the Kimura distribution fit to the data.
(C) The heteroplasmy frequency histogram of the mature oocytes in the 515 line.
(D) KS test for the line 515 mature oocyte data. There is a significant difference between the two distributions in the mature oocyte data.
Human Data

No human pedigree data set is large enough to make a good test of the Kimura distribution. However, there is one human data set that is large enough. Brown et al.\textsuperscript{19} published a study of the heteroplasmy distribution of 82 single primary oocytes derived from an ovary of a female with the pathogenic 3243A > G mtDNA point mutation. This tissue sample was available because this woman underwent a hysterectomy for reasons unrelated to any mitochondrial disease. This woman was asymptomatic and had a mutation level of 18.11% determined from a quadriceps biopsy and of 7.24% of the mutant type in her leukocytes.\textsuperscript{19} We note here that the mutation level of the 3243A > G mutation decreases with age in blood samples.\textsuperscript{25}

Figure 1 presents the comparison of the measured heteroplasmy distribution in the human primary oocytes and the Kimura distribution fit to the data. The Kimura distribution is a very good fit to the measured heteroplasmy distribution, and this is confirmed by the KS test, which gives a p value of 0.827 for the null hypothesis that the experimental data are consistent with the Kimura distribution (Table 1). The limited amount of human heteroplasmy data currently available indicates that the theoretical Kimura probability distribution is a good tool for calculating the distribution of mtDNA heteroplasmy in a population derived from a single founder.

Mouse Data

Given the limited amount of human data available, it is important to extend this analysis to the existing animal models for the inheritance of mtDNA heteroplasmy. Jenuth et al.\textsuperscript{21} published a seminal paper on a mouse model of mtDNA heteroplasmic inheritance. In this study, they used mice that were heteroplasmic for two mtDNA haplogroups, NZB and BALB. These heteroplasmic mice were produced by an electrofusing cytoplasm technique. The data in this study included heteroplasmy measurements on sets of primary oocytes (as in the human data analyzed above), mature oocytes, and tail samples from offspring; each data set was derived from a single founder female.

Figures 2–5 present the comparisons of the Kimura distributions to the heteroplasmy distributions in eight data sets from the mouse model. In six of the eight data sets, the null hypothesis is not rejected, indicating that the Kimura distribution is a good representation of the distribution of the heteroplasmy values in these data sets (Table 1). The null hypothesis was rejected in two of the data sets: the mature oocytes from line 515 (Figures 2C–2D, \(p = 0.049\)) and the primary oocytes from line 603A (Figures 4C and 4D, \(p = 0.037\)). For the data set consisting of line 515 mature oocytes, the difference between the observed heteroplasmy distribution and the fit Kimura distribution
is largest for the number of cells with zero heteroplasmy for the BALB mtDNA haplotype, and fewer of these cells were observed in the experiment than were predicted by the Kimura distribution. For the data set of primary oocytes from the 603A mouse line (Figures 4C and 4D), the largest difference between the observed heteroplasmy distribution and the Kimura distribution is the lack of observation of any cells with NZB haplotype heteroplasmy in the range 0.1%–0.5%, despite the large number of cells with levels of 0% and 0.5%–1.0% in the neighboring bins. Jenuth et al. remarked on this odd result of the missing heteroplasmy values. For the other six mouse data sets, the Kimura distributions do provide a good representation of the observed mtDNA heteroplasmy distributions (Figures 2–5).

**Drosophila Data**
The *Drosophila* data sets consist of data from two species, *D. mauritiana* and *D. simulans*. In both cases the heteroplasmy measurements were made in a sample of unfertilized eggs. For *D. mauritiana* we had six data sets for which the mtDNA heteroplasmy was defined by the difference in the length of an A+T-rich region of the mitochondrial genome. Figures 6 and 7 present the comparisons of the fit Kimura distributions to the measured *Drosophila* heteroplasmy distributions. For five of the six data sets, the null hypothesis is not rejected (Table 1), and the fit Kimura heteroplasmy distributions show a very good correspondence to the observed heteroplasmy distributions. For one data set (Figures 6A and 6B, p = 0.004), the differences between the Kimura distribution and the measured heteroplasmy distribution are quite large. This is interesting because this data set is unique in another way: The number of generations from the founder in this data set is very large at 30 generations, about ten times larger than the number of generations in the other five data sets.

The *D. simulans* data consist of a single data set where the mtDNA heteroplasmy was generated by cytoplasmic injection forming a mixture of the siIII and sII mtDNA genomes, two naturally occurring mtDNA sequences in this species. The comparison of the data to the Kimura distribution is given in Figure 8. Here the null hypothesis is not rejected, and the Kimura distribution is a good representation of the observed mtDNA heteroplasmy distribution.

**Discussion**

In the field of mitochondrial genetics, the Sewall-Wright variance formula has been generally used as the primary data analysis method for determining the effect of random genetic drift on mtDNA heteroplasmy values. Researchers...
have used this simple function both to examine the ability of random genetic drift to explain mtDNA segregation and to predict the rate of mtDNA segregation from assumptions about the size of the mtDNA segregating unit.\textsuperscript{19,21,22} The advantage of the Sewall-Wright variance formula is its simplicity and ability to estimate most parameters from experimental data (although estimating the effective population size $N_{\text{eff}}$ in Equation 1 has always been a problem).

However, the weakness of this simple approach is that it concentrates on just the two lowest-order statistics, the mean and the variance, and it ignores the rest of the information that is present in the total heteroplasmy distribution. This is of particular importance when the heteroplasmy distribution is not symmetric (not a normal distribution), as it must be at the extremes of low and high heteroplasmy. The shape of the heteroplasmy distribution at high-mutation heteroplasmy values is important for understanding the consequences of pathogenic effects, which generally only appear in individuals with a high level of the mtDNA mutation. The distribution at low heteroplasmy values is important because this range is directly affected by any de novo mutation rate. The heteroplasmy distribution near zero also is important for determining the clearance of a pathogenic mutation from a population. Because these extremes are arguably the most important parts of the heteroplasmy range, any approach that implicitly assumes a normal distribution has severe limitations. Using the Kimura distribution as a model for the heteroplasmy distribution across its entire range from 0\%–100\%, as well as the fixation rate on the extremes, frees us from those limitations and gives us a significant new tool in our analysis of mtDNA heteroplasmy inheritance.

The additional information that we can get from using the Kimura distribution comes at a cost: the increased mathematical complexity of Equations 2–5. These equations are difficult to use, and the numerical computation must be done carefully if accuracy problems are to be avoided.\textsuperscript{23} Two possible alternatives to the Kimura distribution are the normal distribution and the binomial distribution. Examples of the Kimura distribution, the normal distribution, and the binomial distribution with equal values for the mean and the variance in all three distributions are given in Figure 9. As discussed above, normal distributions (Figure 9B) do not correctly describe heteroplasmy distributions over the finite range of 0\%–100\% and do not address the important question of fixation. Although binomial distributions are nonsymmetric, cover only a finite heteroplasmy range, and can deal with fixation, they assume that heteroplasmy values come only in discrete steps (Figure 9C), which is not consistent with the available heteroplasmy distribution data. Despite its mathematical complexity, the Kimura distribution is the best available tool for describing mtDNA heteroplasmy distributions.
An alternative computational approach to determining heteroplasmy distributions is the use of direct simulation models. These include simulations of mtDNA replication in individual cells, simulations of mtDNA dynamics in embryogenesis, and relatively simple multiple sampling models. We note that Poulton presented one heteroplasmy distribution from a multiple-sampling simulation model that at least qualitatively resembles the Kimura distribution. Direct simulation models have the advantage of flexibility in that additional mechanisms such as selection effects and de-novo mutations can easily be added to the simulation, but they have the limitation of only presenting results for specific parameter values. The equations of the Kimura distribution have the advantage of explicit definition (something that is often not clear in a simulation) and the presentation of results for all possible parameter values. These two computational approaches are complimentary. Indeed, as discussed below, the Kimura distributions can be used as a tool in developing population-level simulation models of mitochondrial genetics.

Figure 6. The Measured Heteroplasmy Distribution from Unfertilized Eggs in the Heteroplasmic Drosophila mauritiana Lines H1, G20-5, and G71-12 Is Compared to the Kimura Distribution

(A) The heteroplasmy frequency histogram of the Drosophila line H1 and the Kimura distribution fit to the mean and variance values from these data.

(B) The KS test comparing the data with the Kimura distribution. There is a significant difference between the two distributions for line H1.

(C) The heteroplasmy frequency histogram for the Drosophila line G20-5 is compared to the Kimura distribution.

(D) KS test comparing the data for Drosophila line G20-5 to the Kimura distribution.

(E) The heteroplasmy frequency histogram from the Drosophila line G71-12 is compared to the Kimura distribution.

(F) KS test comparing the data for Drosophila line G71-12 to the Kimura distribution.

Only one human data set was large enough to allow a useful comparison against the Kimura distribution. It would be extremely useful to have further human data sets of this type, covering a wide range of mean heteroplasmy values, in order to more thoroughly test the application of the Kimura distribution to human mtDNA heteroplasmy distributions. Further human data sets would also allow us to explore important questions such as how much the b parameter in this model varies across the population (essentially, this corresponds to how variable the inheritance bottleneck is in the human population). With the limited human data currently available, and the data from the mouse and Drosophila models, the Kimura distributions are consistent with the experimental data in 13 of the 16 data sets analyzed. Because the Kimura distribution only represents the effects of random genetic drift, deviations from that distribution may give us information about the other mechanisms that are occurring, most importantly selection effects and de novo mutation. Of the three data sets in which the null hypothesis was rejected, the data in Figures 6A and 6B are of particular interest. These data are from the 30th generation after the founder female, by far the longest generational separation in any of these data sets. It is reasonable to assume that this large number of generations would accentuate effects such as selection or de novo mutations, which might be negligible over shorter time spans. With the very large variance in this data set (Table 1), the theoretical distribution is relatively flat, and
there are sharp peaks at the fixed points 0% and 100%, which act as absorbing states in the random-drift model (in other words, once a female individual fixes at either extreme, all descendents remain at that fixed state). In contrast, the observed heteroplasmy distribution has a “U” shape, such that the probability distribution rises toward each end of the heteroplasmy extreme. It is difficult to construct a mechanism whereby selection could form such a distribution, unless one were to argue for a selection mechanism that had maximum effect at around 50% heteroplasmy and low effects at either heteroplasmy extreme. A more plausible explanation would be that the two fixed states in this case were not absolutely fixed and that there was some production of heteroplasmic descendents from homoplasmic females in both fixed states. These de novo mutation mechanisms, acting over 30 generations, could form the U-shaped distribution seen in Figure 6A. One could also speculate that the shape of the observed heteroplasmy distribution in Figure 6A suggests that the de novo mutation rate of the formation of the longer genome from the shorter genome (i.e., away from the fixed state at heteroplasmy 0%) is the larger of the two mutation rates.

Finally, let us discuss the roles of the parameters $p_0$ and $b$. We defined the parameter $b$ (Equation 7) to replace a combination of the parameter $t$, the number of generations, and the parameter $N_{eff}$, a statistical parameter related to the number of segregating units of mtDNA (though not necessarily directly equal to it). In this paper we have analyzed only a single generation at a time, and we have not applied this analysis to follow the heteroplasmy distribution over multiple generations. One could certainly use the Kimura distribution to follow the distribution over multiple generations, in which case the formulation of Equations 2–5, which are written in terms of $t$ and $N_{eff}$, should be used. The parameter $p_0$ can be interpreted as either the mean heteroplasmy in the data set or the heteroplasmy in the founder. In the case of pure random drift, the two are the same, but other effects may cause a shift in mean heteroplasmy over the generations. This distinction in the definition of $p_0$ may be important in some cases. One example of this is the $D. simulans$ data set (Figure 8), Even though the Kimura distribution fit to this data is a good model of the heteroplasmy distribution ($p = 0.79$), in that experiment the mean heteroplasmy was observed to shift from an initial value of 38.5% in the founder to a value of 12.7% in the third generation.20 This was reasonably interpreted as indicating a selection effect in this experiment. Despite the apparently strong selection effect, the heteroplasmy distribution in the third generation is still well described by a Kimura distribution with the value

$\text{Figure 7. The Measured Heteroplasmy Distribution from Unfertilized Eggs in the Heteroplasmic Drosophila mauritiana Lines H1-31M, H1-18D, and H1-12B Is Compared to the Kimura Distribution}$

(A) The heteroplasmy frequency histogram from the Drosophila line H1-31M and the Kimura distribution.

(B) The KS test comparing the heteroplasmy data for Drosophila line H1-31M to the Kimura distribution.

(C) The heteroplasmy frequency histogram from the Drosophila line H1-18D is compared to the Kimura distribution.

(D) KS test comparing the Drosophila line H1-18D to the Kimura distribution.

(E) The heteroplasmy frequency histogram from Drosophila line H1-12B is compared to the Kimura distribution.

(F) KS test comparing data for Drosophila line H1-12B and the Kimura distribution.
The lesson here is that even if a Kimura distribution, derived from neutral-drift theory, fits the observed heteroplasmy distribution, this is not enough in itself to allow us to determine that neutral drift alone has shaped that heteroplasmy distribution. Instead, the old standard method of measuring the changes in the mean heteroplasmy over a number of generations must continue to be used. The use of the Kimura distribution adds valuable information to our previous analysis techniques, but it does not invalidate them.

What the Kimura-distribution theory presented here allows us to do that we could not do before is to predict the complete probability distribution, including the probability of fixing on the wild-type and on the mutant mtDNA, for mtDNA heteroplasmy values in a group of offspring. Although this predictive ability is under the assumption of random genetic drift, this is a necessary first step to which important complications such as selection effects and de novo mutations may then be added in further development of this theoretical model. The comparisons of the Kimura distributions to the experimental data sets presented in this paper are one use of these equations, but these comparisons are primarily made here as a validation of the application of this theory to mitochondrial genetics. The Kimura distribution equations give us a theoretical framework for the field of mitochondrial heteroplasmy.

A recent study by Elliot et al. of the prevalence of a set of ten pathogenic mtDNA point mutations has shown that these pathogenic mutations are relatively common in the general population, where it has been measured that 1 in 200 individuals carries one of these ten mtDNA mutations. With this new appreciation of how widespread mtDNA heteroplasmy actually is, the ability to calculate the complete heteroplasmy distribution by using the Kimura distribution as a model of random genetic drift is an important tool for understanding the heteroplasmy distribution in the general population.

One potential application of this new theoretical tool is the calculation of simulated pedigrees. These simulated pedigrees may be used as tools for analyzing clinical pedigrees, for example in a Monte-Carlo test to define a p value for a particular clinical pedigree tested against the null hypothesis of random genetic drift. One could also use the theoretical heteroplasmy distribution to calculate disease occurrence probabilities, based on a heteroplasmy threshold for the disease phenotype, for use in genetic counseling. Further testing of the theory, and in particular more human data such as that in Figure 1, will be needed before that becomes a practical application. Finally, the calculation of simulated pedigrees based on this theoretical heteroplasmy distribution could be extended to model large-scale populations. That model could be tested against recent and future measurements of the occurrence of mtDNA heteroplasmy and will help us understand the
development and spread of pathogenic mtDNA mutations in the human population.

Acknowledgments

P.W. is supported by Thailand’s Commission on Higher Education, Royal Thai Government under the program “Strategic Scholarship for Frontier Research Network.” P.E.C. is a Welcome Trust Senior Clinical Research Fellow.

Received: August 21, 2008
Revised: October 8, 2008
Accepted: October 14, 2008
Published online: October 30, 2008

Web Resources

The URL for data presented herein is as follows:


References


Chapter 4: Previous Estimates of Mitochondrial DNA Mutation Level Variance Did Not Account for Sampling Error: Comparing the MtDNA Genetic Bottleneck in Mice and Humans

Passorn Wonnapinij\textsuperscript{1,2}, Patrick F. Chinnery\textsuperscript{3}, and David C. Samuels\textsuperscript{1}

\textsuperscript{1} Center of Human Genetics Research, Department of Molecular Physiology and Biophysics, Vanderbilt University Medical Center, Nashville, TN, 37232, USA

\textsuperscript{2} Interdisciplinary Program in Genetics, Bioinformatics, and Computational Biology, Virginia Polytechnic Institute and State University, Blacksburg, VA, 24060, USA

\textsuperscript{3} Mitochondrial Research Group and Institute of Human Genetics, Newcastle University, The Medical School, Newcastle-upon-Tyne NE2 4HH, UK

Reprinted from “The American Journal of Human Genetics 2010, 86, 540-550” with permission from Elsevier

To whom correspondence should be conducted:
david.samuels@chgr.mc.vanderbilt.edu
Previous Estimates of Mitochondrial DNA Mutation Level Variance Did Not Account for Sampling Error: Comparing the mtDNA Genetic Bottleneck in Mice and Humans

Passorn Wonnapinij,1,2 Patrick F. Chinnery,3 and David C. Samuels1,*

In cases of inherited pathogenic mitochondrial DNA (mtDNA) mutations, a mother and her offspring generally have large and seemingly random differences in the amount of mutated mtDNA that they carry. Comparisons of measured mtDNA mutation level variance values have become an important issue in determining the mechanisms that cause these large random shifts in mutation level. These variance measurements have been made with samples of quite modest size, which should be a source of concern because higher-order statistics, such as variance, are poorly estimated from small sample sizes. We have developed an analysis of the standard error of variance from a sample of size n, and we have defined error bars for variance measurements based on this standard error. We calculate variance error bars for several published sets of measurements of mtDNA mutation level variance and show how the addition of the error bars alters the interpretation of these experimental results. We compare variance measurements from human clinical data and from mouse models and show that the mutation level variance is clearly higher in the human data than it is in the mouse models at both the primary oocyte and offspring stages of inheritance. We discuss how the standard error of variance can be used in the design of experiments measuring mtDNA mutation level variance. Our results show that variance measurements based on fewer than 20 measurements are generally unreliable and ideally more than 50 measurements are required to reliably compare variances with less than a 2-fold difference.

Introduction

Eukaryotic cells typically contain a large number of copies of mitochondrial DNA (mtDNA). Generally, these copies of mtDNA are identical; however, some individuals contain a mixture of two versions of the mtDNA molecule, a condition called heteroplasmy. In the case of inherited mtDNA mutations, this mtDNA heteroplasmy is found in cells throughout the body, but with varying levels of the mutant mtDNA in different tissues.1,2 This variation in mutation level is often also found when comparing multiple cells from the same tissue in the individual.3,4 mtDNA mutation level variations are a major factor underpinning the random mosaic distribution of affected cells that is typically observed in diseases resulting from mtDNA mutation.3,4

Perhaps the most important issue about the mtDNA mutation level variation among cells concerns the variability of the mtDNA mutation levels in the cells of the female germline. Mutation levels of inherited mtDNA mutations are known to vary significantly between the mother and her offspring and among offspring from the same mother.5 This variability is important because the randomness in the inheritance of mtDNA mutations severely limits our ability to provide genetic counseling to affected families.6,7 The processes responsible for this variability in mutation levels among family members and the exact timing of these processes during reproduction are currently a matter of some controversy.8–11 To understand mtDNA mutation inheritance, we must therefore have a reliable means of measuring and comparing the variation generated during the transmission of a heteroplasmic mtDNA mutation, both in the clinical setting and also in several recently developed animal model systems. This understanding will underpin our ability to make predictions about the likelihood of transmitting a particular level of mutation and also provides the analytical tools to study tissue-tissue and cell-cell variability in mtDNA mutation levels, which is fundamental to our understanding of the tissue specificity and clinical progression of mtDNA diseases.

The experimental approach is based upon an estimation of the distribution of mtDNA mutation in a particular sample, which is typically reported as the variance of the mutation level in the sample. As for all statistical estimations, our confidence in the measured variance is critically dependent upon the number of individual measurements—in this case mutation level values—that must be randomly sampled from the population of interest. However, determining the statistical error for a variance measurement is mathematically complex. As a result, the error bars for the measured mtDNA mutation level variance are rarely, if ever, reported.

The mutation level variance is typically estimated from a relatively small sample of cells in the range of 20 cells or even far lower. Major experimental conclusions have been based on comparisons of these measurements of variance, but we currently do not know whether these variance measurements are reliable. In other words, it is not

1Center of Human Genetics Research, Department of Molecular Physiology and Biophysics, Vanderbilt University Medical Center, Nashville, TN 37232, USA; 2Interdisciplinary Program in Genetics, Bioinformatics, and Computational Biology, Virginia Polytechnic Institute and State University, Blacksburg, VA 24060, USA; 3Mitochondrial Research Group and Institute of Human Genetics, Newcastle University, The Medical School, Newcastle-upon-Tyne NE2 4HH, UK
*Correspondence: david.samuels@chg.mc.vanderbilt.edu DOI 10.1016/j.ajhg.2010.02.023. ©2010 by The American Society of Human Genetics. All rights reserved.
known how many individual measurements are required for a reliable estimate of variance with a given statistically defined confidence interval. Here we address this issue from first principles and provide evidence that a far greater number of samples than are generally taken are required to make reliable comparisons of variance between different groups. Central to our approach is a method of reliably calculating the standard error of variance, which will allow these comparisons to be made. With this approach we can confidently conclude that the variation in mutation levels in human pedigrees is greater than that observed in mouse pedigrees transmitting mtDNA heteroplasmy.

**Material and Methods**

**Experimental Data**

Data for mutation level variance measurements, including values for the mutation level variance, the mean mutation level, and the number of measurements (n), in mouse models were gathered from the published literature. The same data for a data set of human primary oocytes was taken from Brown et al. Data for mutation levels in human mother and offspring pairs for various inherited pathogenic mtDNA mutations were gathered from the literature. Probands were excluded from that analysis to minimize ascertainment bias, although it must be kept in mind that this cannot completely remove ascertainment bias.

**Definition of the Standard Error of Variance**

To provide context for the equations for the standard error of variance, we begin with the well-known standard error of the mean. By using a traditional parametric statistical approach, the mean value of a quantity based on n samples from a population has a standard error defined by the well-known equation

\[ SE(p_0) = \sqrt{\frac{s^2}{n}} \]  

(1)

where \( p_0 \) is the mean value (mean percentage level of mutant mtDNA in our case) and \( s^2 \) is the variance of the population that is being sampled. Because the actual variance of the population is not known, nor can it be easily determined, the practical approach is to estimate the population variance \( \sigma^2 \) by measuring the sample variance \( V \) in a randomly selected subgroup of the population. To assign error bars to the measurement of the mean value, one generally follows the practice of setting the error bars to be \( 2 \times SE(p_0) \). This practice arose from the fact that 1.96 \times SE is equal to the 95% confidence intervals for a sample from a normal (Gaussian) distribution. We discuss this practice, and its application to distributions other than the normal distribution, later in this paper.

The corresponding equation for the standard error of the variance based on n samples is less familiar. It is

\[ SE(\sigma^2) = \sqrt{\frac{1}{n} \left( D_4 - \left( \frac{n - 3}{n - 1} \right) \sigma^4 \right)} \]  

(2)

where \( D_4 \) is the fourth central moment of the population.

**Monte-Carlo Simulation Tests of the Statistics**

To test the calculation of the standard error of the variance measurement, we randomly generated a series of independent sets of n values from a chosen probability distribution with a pre-determined mean and variance. This was done for different sample sizes n, ranging from 3 to 100, in order to determine how rapidly estimates of the mean and variance improved as the sample size is increased. By “independent” we mean that a completely new set of values was generated for each value of n. This process was repeated 10,000 times for each sample size allowing the 95% confidence intervals for the mean and variance to be measured as a function of the sample size n. This process was carried out for a normal probability distribution and for a Kimura probability distribution.

**The Levene Test of Paired Simulated Mutation Level Data Sets with Different Variances**

We carried out a Levene test on random simulated data to determine the effect of sample size on the comparison of two data sets with different variances. Sample sizes were varied from 3 to 100 and the samples were independently generated for each sample size from a probability distribution. Pairs of data sets were drawn randomly from a Kimura probability distribution with the same mean mutation level but with different variance values, ranging from 10-fold difference down to equal variances. Although the Levene test is itself complicated, it has the advantage over the standard error of variance that it does not require the calculation of a fourth-order moment or the assumption of a specific underlying probability distribution. The Levene test has several variations, of which the most commonly used is the Brown-Forsyth test, and it is not generally clear which form of the test is the best choice. We evaluated both the standard Levene test and the Brown-Forsyth variation and found that the standard Levene test had a better performance (fewer false negative results) than did the Brown-Forsyth test on our Monte-Carlo simulated data sets.

**Results**

The definition of the standard error of the variance was given as Equation 2 in the Materials and Methods section. This standard error can be used to calculate error bars for a measurement of variance in the same way that the standard error of the mean is used to determine the error bar of a measurement of the mean. The standard error of the variance is a function of the sample size n, the population variance \( \sigma^2 \), and the fourth central moment \( D_4 \) of the population. In the following, we define three different methods of estimating the standard error of variance, based on three different methods of estimating the fourth central moment.

**Model-Free Method of Calculating the Variance Error**

The fourth central moment, \( D_4 \), is not a trivial calculation and this probably accounts for the lack of use of the standard error in a variance measurement. But \( D_4 \) can be calculated in several ways. Most basically, \( D_4 \) can be estimated directly from the n sample measurements. An unbiased estimator for the fourth central moment of the underlying probability distribution is given by

\[ D_4 = \frac{(n - 1)}{n^3} \left( (n^2 - 3n + 3) \mu_4 + 3(2n - 3)\mu_2^2 \right) \]  

(3)
where \( \mu_2 \) and \( \mu_4 \) are defined by

\[
\mu_j = \frac{1}{n} \sum_{i=1}^{n} (p_i - p_0)^j.
\] (4)

With Equations 2–4, the standard error of the variance can be estimated from the data and, as we show in the following sections, the general practice of defining the error bars of the measured variance to be twice the standard error may be followed.

**The Normal Distribution Model**

The value given by Equation 3 for \( D_4 \) is only an estimate of the fourth-order central moment based on a sample of size \( n \). If we are willing to assume that the values of mutation level in the population follow a particular probability distribution, we can use the exact equation for the fourth-order central moment of that distribution. For a normal distribution, the mathematics are particularly simple. The fourth central moment of a normal distribution is simply

\[
D_{4,\text{Normal}} = 3\sigma^4.
\] (5)

Substituting this formula for \( D_4 \) into Equation 2 gives the standard error of variance of a sample of \( n \) data points taken from a normal distribution.

\[
SE(\sigma^2)_{\text{Normal}} = \sigma^2 \sqrt{\frac{2}{n-1}}.
\] (6)

Assuming a normal distribution model has the advantage of greatly simplifying the calculations.

A Monte-Carlo test of the mean and variance values of data sets of size \( n \) drawn from a normal distribution was carried out as described in the Materials and Methods. Figure 1A shows the probability distribution from which the simulated data were chosen, with a mean value of 0.5 and a variance of 0.01. Figure 1B shows the estimates of the mean value as a function of the sample size \( n \), with error bars set to twice the standard error of the mean value (Equation 1). Note how the error bars for the mean values correspond well with the calculated 95% confidence intervals, as expected for a normal distribution. Also note how the variability in the measured mean value corresponds well with the error bars.

These results for the estimates of the mean value and its sample error are well known, and we present them here only to provide context for the corresponding calculation of the sampling error in the estimate of the variance (Figure 1C). As was the case with the mean, setting the error bars of the variance to twice the calculated standard error in the variance is in good agreement with the 95% confidence intervals in the variance measure. The 95% confidence intervals were wide for a variance based on a sample of 20 measurements, especially in comparison to the corresponding confidence interval for the mean value. From Equation 6, when \( n = 20 \) the standard error of the variance is equal to 32% of the variance, meaning that the variance error bars are equal to 64% of the variance values. For a normal distribution, these calculations of the relative size of the variance error bars do not depend on any other parameters, such as mean and variance. In the normal distribution model, a sample of 20 measurements will always have a sampling error of 64% in the estimated variance. As can be seen from Figure 1C, this sampling error increases dramatically as the number of measurements decreases below 20. This raises concerns for studies based on variance values based on 20 individual measurements or less.

**Kimura Distribution Model**

The results given above are general and can be applied to the standard error of the variance of any measured quantity with a normal distribution. Now we specialize to results applicable specifically to mtDNA heteroplasmy. There are two basic features of the normal distribution that make it a poor choice to represent the distribution of
mtDNA mutation level values. First, the normal distribution is defined over the range of minus infinity to plus infinity, whereas mutation level values must be only in the range of zero to one. Second, the normal distribution is always symmetric, whereas mtDNA mutation level distributions can be either symmetric or skewed. For a good example of a skewed distribution of mtDNA mutation level values, see Brown et al. 13 Although the normal distribution can be used as an approximation for the distribution of mtDNA mutation level values, this approximation is good only for distributions with mean values near 0.5 and with very few measurements near either extreme of 0 or 1.

Recently, we defined a probability distribution based on the population genetics theory of Kimura, 41 which can be applied to mtDNA mutation level values. 38 Kimura’s theory of random genetic drift defines the following three equations.

\[
f(0) = (1 - p_0) + \sum_{i=1}^{n} (2i + 1)p_0(1 - p_0)(-1)^i \times F(1 - i, i + 2, 2, 1 - p_0)b^{i(i+1)/2} \tag{7}
\]

\[
\phi(p) = \sum_{i=1}^{n} i(i + 1)(2i + 1)p_0(1 - p_0)F(1 - i, i + 2, 2, p) \times F(1 - i, i + 2, 2, p_0)b^{i(i+1)/2} \tag{8}
\]

\[
f(1) = p_0 + \sum_{i=1}^{n} (2i + 1)p_0(1 - p_0)(-1)^i \times F(1 - i, i + 2, 2, p_0)b^{i(i+1)/2} \tag{9}
\]

The probability of fixing on the wild-type mtDNA is \(f(0)\), the probability of fixing on the mutant is \(f(1)\), and the probability distribution for a mutation level value of \(p\) is \(\phi(p)\). The function \(F(a,b,c,d)\) is the hypergeometric function. We refer to these three equations collectively as the “Kimura distribution.” Despite its complexity, the Kimura distribution is only a two-parameter model, with parameters \(p_0\) and \(b\). Both parameters range from 0 to 1. The parameter \(p_0\) is the mean mutation level and the parameter \(b\) is related to the effective population size and can be referred to as the bottleneck parameter. The effective population size should not be confused with the actual mtDNA copy number 42 and should be interpreted only as a statistical parameter that determines the variance. For further details and comparisons of the Kimura distribution to mtDNA mutation level data, please see Wonnapinij et al. 38 The variance of the Kimura distribution is

\[
\sigma^2 = p_0(1 - p_0)(1 - b). \tag{10}
\]

This variance is equal to the variance equation defined by Sewell-Wright 34,44 and was first used in mitochondrial genetics by Solignac et al. 45

The Kimura distribution has the advantages that it is based solidly on population genetics theory and that it does describe well the existing data on mtDNA mutation level distributions. 38 However, one pays the price for this in its obvious mathematical complexity. For our purposes here, to define the standard error of mtDNA mutation level variance measurements, we need to know only the fourth-order central moment of this distribution. After a significant amount of algebra, this quantity can be calculated as the following.

\[
D_{4,Kimura} = \sigma^4 \left( \left( p_0 - \frac{1}{2} \right)^2 \left( 3(1 - b - b^2) + b^3 + b^4 + b^5 \right) \right)
\]

\[
+ \frac{1}{4} \left( 1 - \left( b + b^2 + b^3 + b^4 + b^5 \right) \right) \tag{11}
\]

Unlike the case of the normal distribution, there is no simplification that occurs when the \(D_{4,Kimura}\) is substituted back into the basic definition of the standard error of the variance, Equation 2. Equations 11 and 2 together define the standard error of the variance in the case that the population that is being sampled follows a Kimura distribution.

As a test of the calculation of the standard error of variance, we carried out a Monte-Carlo test as described in the Materials and Methods. We did this for two cases: a Kimura distribution with \(p_0 = 0.5\) (Figure 2), which is similar to a normal distribution, and a Kimura distribution with \(p_0 = 0.1\) (Figure 3) for which a normal distribution is a poor model. The variance error bars are set to be twice the standard error of the variance, calculated now from Equations 11 and 2. In both examples (Figures 2 and 3), the size of the error bars on both the mean mutation level and the variance corresponded well with the scatter between the independent samples and with the 95% confidence intervals. This validates the use of \(2 \times \text{SE}(V)\) as the variance error bars, even with the Kimura distribution.

As with the normal distribution results, it is concerning how wide the variance measurement error bars and the 95% confidence intervals are in Figures 2 and 3 for relatively common sample sizes, such as \(n = 20\). With the complexity of the Kimura distribution mathematics, the standard error of the variance is not a simple proportion of the variance depending just on the sample size \(n\), as it was in the simpler normal distribution case. Instead, the standard error of the variance depends also on the mean mutation level \(p_0\) and on the bottleneck parameter \(b\). By comparing Figures 2C and 3C, one can see that the variance error (as a proportion of the variance) for the same sample size is larger for extreme values of mean mutation level \(p_0 = 0.1\) in Figure 3) than for moderate mean values \(p_0 = 0.5\) in Figure 2). As a concrete example, consider a sample size of \(n = 20\) with a mean mutation level of \(p_0 = 0.5\) (Figure 2C). In this case, \(2 \times \text{SE}(V)\) is 58% of the variance, in close agreement with the estimate of 64% calculated above for a sample size of 20 assuming a normal distribution. Compare this to the same sample size but with \(p_0 = 0.1\) (Figure 3C). In this case, \(2 \times \text{SE}(V)\) is 86%
of the variance. This will be discussed in more detail later in the paper.

The Standard Error of Variance Shows that There Is a Difference in mtDNA Mutation Level Inheritance between Humans and Mice

The use of the synthetic or simulated data sets in Figures 1–3 allowed us to do idealized tests of the calculation of the standard error of the variance because of sampling effects. However, the true usefulness of this sampling error definition comes from its application to experimentally acquired biological data. Before dealing with the experimental data, though, we must consider an important confounding factor in comparing mtDNA mutation level variances from samples with different mean mutation level values. As the classic Sewell-Wright variance equation (Equation 10) shows, the mutation level variance is a function of the mean mutation level, because it is proportional to $p_0(1-p_0)$. This causes variance to decrease as $p_0$ approaches the extreme values of 0 and 1. In order to correct for this $p_0$ dependence and allow us to compare measured variance values from samples with different mean mutation levels, it is necessary to normalize the variance measurements by dividing them by $p_0(1-p_0)$. The standard error of the variance is then also normalized by dividing it by $p_0(1-p_0)$.

We applied the standard error of variance to the mtDNA mutation level variance data from Jenuth et al., who measured mutation level values from cells sampled from various stages of development of the female germline in a mouse model and the subsequent offspring. Only summary statistics were reported and the full data sets of the mutation level measurement in each cell were not...
given, so the “model-free” method of Equation 3 cannot be used. Instead, we must choose a model for the underlying cell population, and for the reasons given above we chose the Kimura model for this analysis. Usefully, Jenuth et al. reported several repeated independent measurements from each development stage, so we can compare the calculated error bars of the variance to the observed variation in these values across the repeated experiments. The normalized mutation level variance values together with our calculations of the variance error bars are plotted in Figure 4A. The size of the normalized variance error bars corresponds well with the scatter in the measured values within each development stage. Of particular interest are two high normalized variance values reported in the mature oocytes. Based on just the reported variance values (without the error bars), it might be reasonable to conclude that the variance could be fundamentally different in these two samples compared to the other three mature oocyte samples that all had low normalized variances. However, the addition of the variance error bars changes the interpretation of the data. The calculated error bars for the variance in these two samples is very large, and they overlap the error bars for the other three mature oocyte samples. With the variance error bars, the most parsimonious interpretation of the data is that all the normalized variances reported in the mature oocyte data are consistent with each other, with a mean value close to the primary oocyte value. Similarly, though less dramatic, the large scatter in the mutation level variance values reported for the offspring are also shown to be consistent with each other once the sampling error bars are added to the variance values. Finally, the addition of the error bars allows us to interpret the changes in variance between these four stages of development. The increase in variance between the primordial germ cell stage (PGC) and the primary oocyte stage is clear, but no change in the mutation level variance is supported by this data in the comparison of the primary and mature oocytes. Finally, when the error bars are taken into account, one cannot state a firm conclusion about the apparent difference in the mutation level variance between the mature oocyte and offspring. The large error bars in the variance measurements in both of these stages show that the mean variance values in the mouse mature oocytes and the offspring are not significantly different in this experiment.

There is currently only one human data set that is large enough for a reasonable analysis of mtDNA mutation level variance in the female germline cells. Brown et al. reported measurements of mtDNA mutation level in 82 primary oocytes from a woman carrying the m.3243A>G mutation (MIM *590050.0001) who underwent a hysterec-
tomy. We calculated the error bars of the variance for this human oocyte data set and compared it to the mouse primary oocyte data in Figure 4A. The addition of the error bars to the variance measurements supports the conclusion that the mutation level variance in the human oocytes is clearly larger than the variance in the mouse data at the same development stage.

Given that variance is closely linked to the mean mutation level (Equation 10) and that a large number of observations are needed to measure variance, a reliable estimate of the variance can be obtained only from a mother with many offspring or by combining the offspring from mothers with similar mean mutation levels. Published data on mutation levels in mothers and offspring were gathered as described in the Materials and Methods. In order to minimize the differences in variance expected from the Sewell-Wright variance formula (Equation 10), we chose to combine data from mothers with mtDNA mutation levels in the range of 40%–60%, where the differences in the mean mutation level have the least impact on the variance. Data from mothers carrying the A3243G mutation were excluded from this analysis to avoid the potential confounding effects of age on the mutation level measured in blood in this particular mutation. With this approach, we identified 72 human mother-offspring pairs from the published literature. The normalized mutation level variance calculated from these data was significantly higher in the human offspring than it was in the mouse model (Figure 4A), when the variance error bars

---

**Figure 4. Application of the Standard Error of Variance to Data from Human and Mouse Models**

(A) Heteroplasmic mouse model data from Jenuth et al. (circles) at four stages of mtDNA inheritance: primordial germ cells (PGC), primary oocytes, mature oocytes, and offspring. Human data (stars) from Brown et al. for primary oocytes and from numerous sources for offspring data are compared to the mouse data.

(B) mtDNA mutation level variance with error bars measured in 21 mouse lineages. All error bars are twice the standard error calculated from a Kimura distribution. Variance values are normalized by dividing by \(p_0(1 - p_0)\).
are taken into consideration. At both stages of development, primary oocytes and offspring, the human normalized variances are approximately three times larger than the corresponding normalized variance in the mouse model. This is an important point to consider when interpreting the results from any experiment with a mouse model of mtDNA heteroplasmy.

In the recent paper by Cree et al., mutation level variance values in 21 lineages of heteroplasmic mice were reported in the Supplemental Data. Without variance error bars and the proper normalization of the variance, it is difficult to interpret the scatter of the data in Table S1 of Cree et al. In Figure 4B we show our calculated error bars for these normalized data, again based on a Kimura distribution. The error bars show that the normalized variance measurements that are large also have large errors, so that all 21 mouse lineages actually have reasonably consistent normalized variance values.

The development of mtDNA mutation level variance in the female germline of a mouse model was also the subject of a recent paper by Wai et al. In that paper, variance measurements in samples from the female germline were reported over 44 days after birth. Based on these variance measurements, taken from samples with differing mean mutation levels and without correcting for this confounding factor through normalizing the variances, Wai et al. concluded that there was a strong increase in variance in the female germline cells during this postnatal period. They reported statistically significant differences between the variances measured on postnatal day 11 and later compared to the variances measured at postnatal day 8 and earlier. However, within those two periods only the comparison of day 11 to day 29 was statistically significant. In Figure 5 we plot the variance data from Wai et al. with comparison of day 11 to day 29 was statistically significant.

In contrast to the mouse model, we currently have very limited data on the development of mtDNA mutation level variance at different stages of the human female germline. In Figure 4A, we plot normalized variance values for a single human primary oocyte data set and for a group of human offspring. There is a clear difference in the normalized variance values between these two stages of development; however, this difference should be interpreted with caution. The human primary oocyte data were from a single person who carried the A3243G mutation, while that specific mutation was removed from the offspring data set because of the observed decline in the mutation level of the A3243G mutation with age in blood samples. It is possible that the differences in variance in the human data in Figure 4A may be due to the different pathogenic mutations instead of the different stages of development. This question about the human data can be answered only by having more data on the variance of other pathogenic mtDNA mutations at the primary oocyte stage.

**Statistical Tests for the Comparison of Variance Measurements**

The calculation of the standard error of variance is a useful tool for the comparison of measurements of variance values; however, when the full data sets are available it is possible to test for the homogeneity of the variance in different samples via the Levene test as done by Wai et al. We carried out a Levene test of paired simulated data sets drawn from Kimura distributions with the same mean value and different variances, as described in the Materials and Methods. The results are shown in Figure 6 for paired Kimura distributions with a mean mutation level of 0.5. Large variance differences (Figures 6A and 6B) are easily distinguished with significant p values even for small sample sizes. However, moderate variance differences, on the order of 2-fold or less (Figure 6C), can be reliably distinguished only with relatively large samples, and even then there is a high rate of false negative results.
A variance difference of 1.5-fold (Figure 6D) could not be reliably detected even with sample sizes of 100. The test of equal variance samples (Figure 6E) shows approximately 5% false positives, as would be expected. As we showed earlier (Figure 3), sample size effects on variance measurements increase at both large and small mean mutation levels. Figure 7 shows the p value calculations for comparisons of two samples with equal mean mutation level of 0.1. At this low level of mutation, which is not an unusual value in the mouse model data, even 2-fold differences in variance cannot reliably be distinguished with sample sizes of approximately n < 50. Even in the extreme case of a 10-fold variance difference, several false negative results occur (Figure 7A).

Using the Standard Error of Variance for Experiment Design

The analysis we present here can be used to design experiments with sufficient power to reliably detect changes in mtDNA mutation level. If one chooses to assume a normal distribution model, then this process is relatively simple and Equation 6 can be used to determine the necessary sample size n. However, the mathematical complications of the Kimura model mean that its use in experimental design is more difficult than the normal distribution, though we would argue that it is also more accurate. From Equation 11, the standard error of the variance in the Kimura model will depend on the distribution parameter values $p_0$ and $b$, as well as the sample size n. In Figure 8 we plot the standard error of the variance divided by the variance for different values of $p_0$ and $b$, assuming that the value of $b$ is set to 0.9, approximately the value determined from the analysis of the human oocyte data set. Figure 8A shows how the standard error of the variance rises rapidly for small sample sizes (n below about 20). Figure 8B shows that at extreme values of mutation level, below about 0.1 and above about 0.9, the standard error in the variance measurement is much greater. In between these extreme values of $p_0$, the standard error of variance is relatively insensitive to different mutation level values. In this intermediate $p_0$ range, the normal distribution is often a good approximation for the Kimura distribution, and the values in Figure 8B correspond well with those calculated from Equation 6 for the normal distribution.

---

Figure 6. Levene Test p Values for Comparisons of Two Data Sets with Different Variances but Equal Mean Mutation Levels of 0.5

Both data sets were drawn randomly from a Kimura distribution. The distribution for the first data set was set to have $b = 0.9$ while the value of $b$ for the second distribution was lowered according to Equation 10 to give the stated variance difference. p values were calculated with the standard Levene test. The horizontal line indicates a p value of 0.05. Differences in variance are as follows: (A) 10-fold; (B) 5-fold; (C) 2-fold; (D) 50% increase; (E) equal variance.

Figure 7. Levene Test p Values for Comparisons of Two Data Sets with Different Variances but Equal Mean Mutation Levels of 0.1

Other details are the same as in Figure 6.
However, the width of the horizontal region in Figure 8B depends on the bottleneck parameter \( b \). Lower values of \( b \) will correspond to higher variances (as shown by Equation 10), making the normal distribution a poor approximation to the Kimura distribution.

A practical approach would be to calculate the required sample size \( n \) via a general estimate based on the normal distribution approximation of Equation 6. One then needs to keep in mind that this will give a good prediction for the standard error of the variance in data sets with moderate mean \( mtDNA \) mutation levels (near 50%), but that data sets with high or low values of mean mutation level will have even higher relative standard errors of the variance, as illustrated in Figure 8B. The normal distribution method underestimates the standard error of variance for samples with high (>90%) or low (<10%) mean mutation level, so for calculating error bars for the measured variance, either the model-free method (Equations 2–4) or the Kimura model method (Equations 11 and 2) should be used.

**Discussion**

The calculation of \( mtDNA \) mutation level variance values from quite small sample sizes has been an accepted practice, although some have had concerns about this practice. We have addressed these concerns by developing the equations for calculating the standard error of variance measured from a sample of size \( n \). We give three options for doing this calculation. The model-free method uses only the measured data and does not require any assumption of the form of the probability distribution from which the data are sampled. However, the model-free method does require the calculation of the fourth central moment of the data, and high-order moments such as this are difficult to estimate from data. The simplest option is to assume that the population follows a normal distribution, and in that case the standard error of the variance is quite simple to calculate. The difficulty with the normal distribution is that it is a poor description of the mutation level distribution at the high and low extremes and these are often the ranges of great practical interest. To deal with the details of the \( mtDNA \) mutation level distribution, we have developed the Kimura distribution\(^{38}\) based on the theory of neutral genetic drift.\(^{41}\) Although the Kimura distribution is mathematically complicated, it is a reliable description of measured \( mtDNA \) mutation level distributions\(^{38}\) and in this paper we have derived the standard error of variance for mutation level values drawn from a Kimura distribution.

Although the standard error of variance has not been used in this field, the standard error of the mean is, of course, common knowledge. We would argue that in general, assumptions about the reasonable number of samples to take in an experiment have been shaped by our familiarity with the standard error of the mean. However, a number of samples \( n \) that are quite sufficient for the accurate estimation of the mean value can be inadequate for the estimation of higher-order statistics, such as the variance. Comparisons of the confidence intervals for the mean values and for the variance values for both the normal distribution (Figure 1) and the Kimura distribution (Figures 2 and 3) illustrate this difference starkly. Although the confidence intervals for the mean are quite reasonably small for sample sizes of about 20, the corresponding confidence intervals for the variance measurements are disturbingly large at those sample sizes, making it extremely difficult to reliably measure small changes in variance. Because scientific conclusions are being made based on comparisons of these measured variances, it is critical that error bars for these variance measurements be reported and that reliable statistical tests for comparisons of variance measurements, such as the Levene test, should be used.

Based on our Monte-Carlo results (Figures 6 and 7), a good rule of thumb for experimental design is that at moderate mean mutation levels (50%), a 2-fold or greater difference in normalized variance can be reliably detected by >30 measurements, while for low (10%) or high (90%) mean mutation levels, the number of measurements should be increased to 50 or more.
The standard error of variance is a critical tool for assessing the reliability of a variance measurement. With this new capability, we have reinterpreted the experimental data on the development of mtDNA mutation level variance in the female germline. In the mouse model, this reassessment shows that there is no support for the conclusion that the mutation level variance increases greatly during postnatal development (Figures 4A and 5), contrary to the previous interpretation of the data. The addition of the standard error of variance also shows that there is a clear difference between the mouse model and the human data, with humans having a far larger mtDNA mutation level variance than mice in both primary oocytes and offspring.

Acknowledgments

P.W. is supported by Thailand’s Commission on Higher Education, Royal Thai Government under the program “Strategic Scholarship for Frontier Research Network.” P.F.C. is a Wellcome Trust Senior Fellow in Clinical Science who also receives funding from the Medical Research Council (UK), the UK Parkinson’s Disease Society, and the UK NIHR Biomedical Research Centre for Ageing and Age-related disease award to the Newcastle upon Tyne Foundation Hospitals NHS Trust.

Received: December 21, 2009
Revised: February 16, 2010
Accepted: February 23, 2010
Published online: April 1, 2010

References


5 Chapter 5: The Mother Carrying a High Pathogenic A3243G Mutation Level Tends to Have Fewer Daughters than Sons

Passorn Wonnapinij$^{1,2}$, Patrick F. Chinnery$^{3}$, and David C Samuels$^{1}$

1. Center of Human Genetic Research, Department of Molecular Physiology and Biophysics, Vanderbilt University Medical Center, Nashville, TN, 37232, USA

2. Interdisciplinary Program in Genetics, Bioinformatics, and Computational Biology, Virginia Polytechnic Institute and State University, Blacksburg, VA, 24060, USA

3. Mitochondrial Research Group and Institute of Human Genetics, Newcastle University, The Medical School, Newcastle-upon-Tyne, NE2 4HH, UK
The mother carrying a high pathogenic A3243G mutation level tends to have fewer daughters than sons

Passorn Wonnapinij$^{1,2}$, Patrick F. Chinnery$^3$ & David C. Samuels$^1$

$^1$ Center of Human Genetics Research, Department of Molecular Physiology and Biophysics, Vanderbilt University Medical Center, Nashville, TN, 37232, USA

$^2$ Interdisciplinary Program in Genetics, Bioinformatics, and Computational Biology, Virginia Polytechnic Institute and State University, Blacksburg, VA, 24060, USA

$^3$ Mitochondrial Research Group and Institute of Human Genetics, Newcastle University, The Medical School, Newcastle-upon-Tyne NE2 4HH, UK

The pathogenic A3243G mitochondrial DNA (mtDNA) mutation has been widely observed to cause a variety of human diseases, including neurodegenerative disorders affecting multi-systematic tissues such as MELAS$^1$. The prevalence of this mutation is approximately 1 in 1000 of the population, thereby this mutation is relatively common in the general population$^2$. Typically, an individual, both symptomatic and asymptomatic individuals, carries this mutation in a heteroplasmic condition, a mixture of wild type and mutant mtDNA. In this study, we show that the intergenerational increase of mtDNA mutation level is partially the result of no correction for a reduction of blood mtDNA mutation level toward age. We also observed that the transmission of mtDNA mutation level from the mother carrying a high proportion of this mutation is subjected to purifying selection. Importantly, we observed that the mother who carries a high mutation level tends to have more sons than daughters, implying that there is a selection bias against female offspring, naturally preventing transmission of this harmful mtDNA to the next generation.
Until recently, no effective way to treat patients with a mtDNA mutation causing disease has been successfully developed, therefore preventing the transmission of mtDNA mutation has become an important strategy. In general, mothers carrying pathogenic mtDNA mutations transmits a random proportion of mutant mtDNA to their offspring, generating a large random shift in mutation levels between generations\(^3\). In the case of A3243G mutation, the studies of single generation transmission of mtDNA mutation level, based on blood measurements, showed that the transmission of this mutation is in favour of mutant mtDNA\(^{3,4}\). However, an exponential reduction of blood A3243G mutation level with age due to the selection against hematopoietic stem cells could deceptively generate an increase of mutation level across generations\(^{5-8}\). The mathematical formula for correcting the reduction of blood mutation level was provided as shown in Equation 1.

\[ p_{\text{age-corrected}} = p e^{0.02t} \]  

Where \( p \) is the blood A3243G mutation level and \( t \) is the individual’s age when his/her blood was sampled. We have studied the intergenerational transmission of this mutation, both with and without an application of the age-correction to the individual’s blood mutation level.

The comparison of average mutation level between generations shows that the average blood mutation level consistently increases across generations, from the grandparent’s (\( N_{\text{female}} = 13 \) and \( N_{\text{male}} = 5 \)) to the most recent generation (\( N_{\text{female}} = 29 \) and \( N_{\text{male}} = 28 \)), although this pattern disappears after applying the age-correction formula both in females and males, as shown in Figure 1(a) and (b), respectively. Therefore, we can conclude that the application of the age-correction to the individual’s blood measurement help to adjust the confounding effect of the reduction of blood mutation level with age.
that generates the increase of the average mutation level across generations. However the
average grandmother’s age-corrected blood mtDNA mutation level remains to be
significantly different from the average mother’s age-corrected blood mtDNA mutation
level ($N_{mother} = 58$) with p-value of 0.022.

We also statistically analyzed the average value of the mutation level difference
between a mother and her offspring (the average O-M). The mother-offspring pairs in
this analysis, taken across all generations, were divided into two groups based on
maternal age-corrected blood mutation level: a group of mothers carrying a low mutation
level (maternal age-corrected mutation level $\leq 50\%$) and a group of mothers carrying a
high mutation level (maternal age-corrected mutation level $> 50\%$). The average age-
corrected O-M in each mothers’ group has been statistically compared to zero to examine
whether the offspring’s mutation level is significantly different from the mother’s
mutation level. The result, as shown in Figure 1(c), shows that the average O-M of the
mothers carrying a low mutation level, both taken across all offspring and only from
either sons ($N_{mother-son ~pair} = 26$) or daughters ($N_{mother-daughter ~pair} = 35$), presents a non-
significant difference from zero, suggesting that random drift is the mechanism regulating
transmission of this mutation in this mothers’ group. Conversely, the average O-M of the
mothers carrying a high mtDNA mutation level, both taken across all offspring and in
regard to offspring’s gender($N_{mother-daughter ~pair} = N_{mother-son ~pair} = 10$), is a negative value with
significant difference from zero (p-value $< 0.05$), except for the barely missing p-value of
the average O-M taken from the daughters (p-value = 0.057). Hence the transmission of
the A3243G mutation of the mothers carrying a high proportion of this mutation is
subject to a negative selection, which could be expected for the transmission of the
harmful mtDNA mutation. The average O-M of the daughters does not significantly
differ from the average O-M of the sons, both in the group of mothers carrying a low
mutation level and the mothers carrying a high mutation level, indicating no gender bias of the transmission of the A3243G mutation level.

The main analysis of this study is to examine the effect of this mutation on female reproduction which would have an impact on the transmission of this common pathogenic mtDNA mutation. To achieve this particular purpose, we analyze an average number of daughters and sons of the female family members carrying this mutation. As for the analysis of the average O-M, the females were separated into two groups based on their age-corrected blood mutation level. The analysis has been applied only to the females of the parent’s generation who are carriers in order to reduce the effect of maternal lineage extinction, a reduction of the proportion of female family members across generations. In our analyzed pedigrees, the proportion of females decreases from approximately 75% in the grandparent’s (N_{female}=13, N_{male}=5) and the parent’s generation (N_{female}=58, N_{male}=19) to approximately 51% in the most recent generation (N_{female}=29, N_{male}=28).

There is no significant difference of the average number of daughters as well as the average number of sons between the females carrying a low mutation level (N_{female carrier}=43) and the females carrying a high mutation level (N_{female carrier}=26), though the average number of daughters of the females carrying a high proportion of this mutation (0.42±0.27) is approximately half of that of the females carrying a low proportion of this mutation (0.79±0.29), as shown in Figure 2(a).

In the females carrying a low mutation level, the average number of daughters (0.79±0.29) does not significantly differ from the average number of sons (0.81±0.26), supported by p-value of 0.90. On the other hand, the average number of daughters of the females carrying a high mutation level (0.42±0.27) significantly differs from the average number of sons (0.85±0.33) with the p-value of 0.031, as shown in Figure 2(a). This
result indicates that the females carrying a high proportion of this mutation tends to have fewer daughters than sons which may be an alternative form of selection preventing further transmission of this pathogenic mtDNA mutation.

We realized that the number of daughters and sons are not normally distributed which violated an assumption of the two-sample paired t-test that was applied to test the hypothesis of offspring gender bias. Therefore, we reanalyzed the comparison of the number of daughters to the number of sons, both in the group of females carrying a low and a high mutation level, using the two-sample paired sign test, a nonparametric alternative approach of the two-sample paired t-test. The result supports our previous conclusion in the sense that, in the group of females carrying a low mutation level, we are more likely to observe an equal number of the females who have more sons to the females who have more daughters supported by the p-value of 0.70, as shown in Figure 2(b). And, in the group of females carrying a high mutation level, we are less likely to observe an equal number of the females who have more sons to the females who have more daughters supported by the p-value of 0.021, as shown in Figure 2(c).

Even though, the small number of females in the parent’s generation carrying a high proportion of this mutation is a source of concern, the significant p-value of the comparison between the statistical value of the number of daughters and that of the number of sons of the mothers carrying high mutation level indicates a strong effect of this mutation on female reproduction.

One would think that the significant offspring gender bias of the females carrying a high mutation level is possibly subjected to an ascertainment bias in which gender preference of index cases drives the collection of analyzed pedigrees toward the family carrying females who tend to have more offspring with the same gender as the index case’s gender preference. We address this concern by comparing a number of male index
cases to a number of female index cases of the most recent generation. The comparison result shows that the number of female index cases \(N_{\text{female index case}} = 11\) is relatively equal to the number of male index cases \(N_{\text{male index case}} = 12\), therefore, it is less likely that this significant offspring gender bias is misleadingly generated by the index case’s gender preference. In addition, the relative equal average number of daughters to the average number of sons of the females carrying a low mutation level argues for the existence of an offspring’s gender preference in the analyzed pedigrees.

The significant offspring gender bias in the females carrying a high mutation level has an implication for the significant lower average grandmother’s mutation level. The presence of the offspring gender bias reduces a probability of having a daughter in a female carrying a high proportion of this mutation, thus the grandmother needs to carry a low proportion of this mutation to avoid this effect. It is also possible that this effect implies the frequent observation of small pedigrees carrying this mutation.

The offspring gender bias observed in females carrying a high mutation level may suggest a higher frequency of miscarried female foetus than the miscarried male foetus. This would be a clinical concern, although the clinical cause of this offspring gender bias needs to be drawn from more clinical observations, especially the ones that related to gender preference of a miscarried foetus. Therefore, even though we do not completely understand what causes the female carrying high proportion of the A3243G mutation to have more sons than daughters, our result would be of benefit to both genetic counseling and epidemiological study of this common pathogenic mtDNA mutation.

**Method Summary**

Human clinical pedigrees of the family carrying a pathogenic A3243G mtDNA mutation were collected from published literatures as described in an online method
section. The reference list of the resource data is also included in this section. The individual’s mtDNA mutation level was measured from blood sample.

Only the mothers carrying mutant mtDNA and her offspring were included in the analysis. Any individual whose age-corrected blood mtDNA mutation level exceeds 110% was excluded from the analysis. All index cases were excluded from the transmission study, shown in Figure 1, to reduce the effect of ascertainment bias. The study of the effect of this mutation on female reproduction, shown in Figure 2, was restricted only to the mothers of the most recent generation in the pedigree and no female index cases were excluded from this study. The data set used in this paper is presented in Supplementary information.

The two-sample t test for unequal variance (heteroscedastic) was applied to analyze significance of differences of the average mutation level between generations and the comparison of the number of daughters and son between mothers carrying different categories of mutation level. The two-sample paired t test was used for the average O-M analysis and the comparison of the average number of daughters to the average number of sons; the two-sample paired sign test was also applied as an alternative approach. The Fisher’s exact test was used to examine index case’s gender preference. The detail of the statistical analysis is presented in the online method. The significant result is stated when the p-value is less that the critical p-value of 0.05.
References


4 Wong, L. J. C., Wong, H. & Liu, A. Y. Intergenerational transmission of pathogenic heteroplasmic mitochondrial DNA. Genetics in Medicine 4, 78-83 (2002).


Supplementary Information accompanies the paper on www.nature.com/nature.

ACKNOWLEDGEMENTS

P.W. is supported under the program “Strategic Scholarship for Frontier Research Network” by the Thailand’s Commission of Higher Education, Royal Thai Government. P.F.C is a Welcome Trust Senior Fellow in Clinical Research.

AUTHOR CONTRIBUTIONS

This statistical study was designed by D.C.S., P.F.C. and P.W and carried out by P.W.
Correspondence and requests for materials should be addressed to D.C.S. (david.samuels@chgr.mc.vanderbilt.edu).

Figure 1 The intergenerational transmission of A3243G mutation level (a) The average female mutation level per generation was compared across generations from the grandparent’s to the most recent generation. The average mutation level was calculated from both blood and age-corrected blood mutation level. The p-value was calculated from the two-sample t test for unequal variance. (b) The average male mutation level per generation was compared across generations from the grandparent’s to the most recent generation. The average mutation level was calculated from both blood and age-corrected blood mutation level. The p-value was calculated from the two-sample t test for unequal variance. (c) The average age-corrected blood O-M was compared to zero to examine whether the offspring’s, daughter’s, and son’s age-corrected blood mutation level is significantly different from the mother’s age-corrected blood mutation level. The p-value was calculated from the two-sample paired t test.

Figure 2 The effect of A3243G mutation on female reproduction (a) The comparison of the average number of daughters to the average number of sons within the group of females carrying low mutation level (maternal age-corrected mutation level ≤ 50%) and the group of females carrying high mutation level (maternal age-corrected mutation level > 50%). The p-value was calculated from the two-sample paired t test. (b) The distribution of the difference between the number of sons and the number of daughters of the females carrying low mutation level. The p-value was calculated from the two-sample paired sign test. (c) The distribution of the difference between the number of sons and the number
of daughters of the females carrying high mutation level. The p-value was calculated from the two-sample paired sign test.

**Online Method**

**Data collection**

Human pedigrees of the family carrying A3243G mtDNA mutation were collected from published literatures. Each individual’s pedigree position, mtDNA mutation level, age, and relationship with the index case were recorded. Only the number of daughters and sons of the female family members were recorded. Only the information of index cases and their maternal relatives are in the analyzed data. Each individual’s mtDNA mutation level is the blood sample measurement. The age of each individual is the age of the individual when his/her blood was sampled for measuring mtDNA mutation level. Beside the first-degree relatives of the index case, the relationship with the index case of other maternal relatives was recorded as relative. The number of daughter live births and son live births were recorded. The number of daughters and sons of a female family member who bears no child, except the most recent generation females, were recorded as zero. The analyzed data is presented in the Supplementary information with the reference number of the resource literature follows the reference list in this section.

**Data management**

There are two main analyses in this study: the analysis of A3243G mutation level inheritance and the analysis of the effect of this mutation on female reproduction.

Only the mothers carrying a fraction of mutant mtDNA and her offspring were included in the study. The individual whose age-corrected blood mtDNA mutation level exceeds 110% was excluded from the study. All index cases were excluded from the data
set that used for the analysis of the transmission of mtDNA mutation level shown in Figure 1. This process was done to reduce the effect of an ascertainment bias in human clinical data.

The reverse number of generation used in the comparison of the average mtDNA mutation level between generations is the number of generation normalized by the generation number of the most recent generation in the pedigree. Only mother and her offspring whose age and blood mtDNA mutation level were recorded in the resource literatures were included in the analysis of the average different mutation level between a mother and her offspring (average O-M).

No female index case was excluded from the analysis of offspring gender bias of this mutation. This analysis was restricted only to the mothers of the most recent generation to eliminate the effect of maternal lineage extinction. Including other generation mothers would only increase the statistical significance.

**Statistical analysis**

The statistically significant different of an average mutation level between generations, the comparison of an average O-M of the daughters to an average O-M of the sons, and the comparison of the average number of daughters and the average number of sons of the females carrying high mutation level (female’s age-corrected mutation level > 50%) to those of the females carrying low mutation level (female’s age-corrected mutation level ≤ 50%) were examined by the two-sample student’s t test for unequal variance (heteroscedastic).

The two-sample paired t test was applied to test whether the average O-M is significantly different from zero for all offspring, only daughters and only sons. This statistical test was also applied to examine the offspring gender bias of female
reproduction both in the group of females carrying low and high mtDNA mutation level. Both types of the student’s t test were performed using a statistical function “ttest” in Microsoft Excel (Microsoft software).

The two-sample paired sign test was applied to the comparison of the median of the number of daughters to the median number of sons, both in the groups of females carrying low and high mutation level. This statistical test is a nonparametric alternative approach of the two-sample paired t test. This statistical test was performed in OriginPro8 (OriginLab).

The comparison of the number of female index cases to the number of male index cases was statistically tested by the Fisher’s exact test using a function “fisher.test” in R (R foundation for statistical computing). The result is stated as “significant difference” if and only if the p-value is less than the significance level at 0.05.

References


Figure 1

(a) Only Female
- The grandparent's generation
- The parent's generation
- The most recent generation

Average Mutation Level (%)

- Blood Mutation Level (%)
  - p = 1.76 x 10^{-5}
  - p = 9.1 x 10^{-6}

- Age-Corrected Blood Mutation Level (%)
  - p = 0.022
  - p = 0.003

(b) Only Male
- Blood Mutation Level (%)
  - p = 0.33

- Age-Corrected Blood Mutation Level (%)

(c) Age-Corrected Maternal Mutation Level

- Daughter
- Offspring
- Son

Average Age-Corrected O-M (%)

- Low Maternal Mutation Level (≤ 50%)

- High Maternal Mutation Level (> 50%)
  - p = 0.057
  - p = 0.0011
  - p = 0.0086
Figure 2
Supplementary Information

Supplementary Table 1: The human pedigree data used in the analysis of A3243G heteroplasmy transmission and its effect on female reproduction

<table>
<thead>
<tr>
<th>Reference</th>
<th>Gender</th>
<th>Reverse Relationship with the Index Case</th>
<th>Number of Daughters</th>
<th>Number of Sons</th>
<th>Blood Mutation Level</th>
<th>Age</th>
<th>Age-Corrected Blood Mutation Level</th>
<th>Maternal Blood Mutation Level</th>
<th>Age</th>
<th>Maternal Age-Corrected Mutation Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>M</td>
<td>0</td>
<td>S</td>
<td>2</td>
<td>1</td>
<td>70</td>
<td>21</td>
<td>106.54</td>
<td>60</td>
<td>58</td>
</tr>
<tr>
<td>10</td>
<td>F</td>
<td>-2</td>
<td>UI</td>
<td>1</td>
<td>1</td>
<td>16</td>
<td>61</td>
<td>54.20</td>
<td>19</td>
<td>19.40</td>
</tr>
<tr>
<td>10</td>
<td>M</td>
<td>-2</td>
<td>UI</td>
<td>1</td>
<td>1</td>
<td>19</td>
<td>56</td>
<td>58.23</td>
<td>55</td>
<td>58.23</td>
</tr>
<tr>
<td>10</td>
<td>M</td>
<td>-2</td>
<td>UI</td>
<td>1</td>
<td>1</td>
<td>17</td>
<td>51</td>
<td>47.14</td>
<td>55</td>
<td>57.08</td>
</tr>
<tr>
<td>10</td>
<td>M</td>
<td>-2</td>
<td>UI</td>
<td>1</td>
<td>1</td>
<td>13</td>
<td>52</td>
<td>36.78</td>
<td>55</td>
<td>57.08</td>
</tr>
<tr>
<td>10</td>
<td>M</td>
<td>-1</td>
<td>UI</td>
<td>2</td>
<td>1</td>
<td>28</td>
<td>46</td>
<td>70.26</td>
<td>56</td>
<td>58.23</td>
</tr>
<tr>
<td>10</td>
<td>M</td>
<td>-1</td>
<td>UI</td>
<td>1</td>
<td>1</td>
<td>34</td>
<td>41</td>
<td>71.20</td>
<td>56</td>
<td>58.23</td>
</tr>
<tr>
<td>10</td>
<td>F</td>
<td>-1</td>
<td>UI</td>
<td>0</td>
<td>1</td>
<td>45</td>
<td>26</td>
<td>75.69</td>
<td>19</td>
<td>58.23</td>
</tr>
<tr>
<td>10</td>
<td>F</td>
<td>-1</td>
<td>UI</td>
<td>0</td>
<td>0</td>
<td>42</td>
<td>30</td>
<td>76.53</td>
<td>19</td>
<td>58.23</td>
</tr>
<tr>
<td>10</td>
<td>M</td>
<td>0</td>
<td>UI</td>
<td>8</td>
<td>17</td>
<td>11.24</td>
<td>5</td>
<td>13.06</td>
<td>56</td>
<td>13.06</td>
</tr>
<tr>
<td>11</td>
<td>F</td>
<td>-1</td>
<td>UI</td>
<td>0</td>
<td>1</td>
<td>17.6</td>
<td>56</td>
<td>53.94</td>
<td>56</td>
<td>53.94</td>
</tr>
<tr>
<td>11</td>
<td>M</td>
<td>-1</td>
<td>UI</td>
<td>22</td>
<td>44</td>
<td>53.04</td>
<td>19</td>
<td>58.23</td>
<td>56</td>
<td>58.23</td>
</tr>
<tr>
<td>11</td>
<td>F</td>
<td>-1</td>
<td>UI</td>
<td>0</td>
<td>0</td>
<td>29</td>
<td>39</td>
<td>63.26</td>
<td>56</td>
<td>63.26</td>
</tr>
<tr>
<td>11</td>
<td>F</td>
<td>-1</td>
<td>UI</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>57</td>
<td>33.53</td>
<td>56</td>
<td>33.53</td>
</tr>
<tr>
<td>11</td>
<td>M</td>
<td>0</td>
<td>UI</td>
<td>25</td>
<td>26</td>
<td>42.05</td>
<td>17.6</td>
<td>53.94</td>
<td>56</td>
<td>53.94</td>
</tr>
<tr>
<td>12</td>
<td>F</td>
<td>-1</td>
<td>I</td>
<td>0</td>
<td>2</td>
<td>62.7</td>
<td>26</td>
<td>105.46</td>
<td>56</td>
<td>105.46</td>
</tr>
<tr>
<td>12</td>
<td>M</td>
<td>0</td>
<td>O</td>
<td>71.3</td>
<td>4</td>
<td>77.24</td>
<td>62.7</td>
<td>26</td>
<td>105.46</td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>F</td>
<td>-1</td>
<td>M</td>
<td>3</td>
<td>2</td>
<td>28</td>
<td>27</td>
<td>48.05</td>
<td>56</td>
<td>48.05</td>
</tr>
<tr>
<td>13</td>
<td>F</td>
<td>-1</td>
<td>S</td>
<td>3</td>
<td>1</td>
<td>14</td>
<td>55</td>
<td>42.06</td>
<td>56</td>
<td>48.06</td>
</tr>
<tr>
<td>13</td>
<td>F</td>
<td>-1</td>
<td>M</td>
<td>2</td>
<td>2</td>
<td>14</td>
<td>69</td>
<td>55.65</td>
<td>56</td>
<td>55.65</td>
</tr>
<tr>
<td>13</td>
<td>F</td>
<td>-1</td>
<td>S</td>
<td>1</td>
<td>1</td>
<td>24</td>
<td>54</td>
<td>70.67</td>
<td>56</td>
<td>70.67</td>
</tr>
<tr>
<td>13</td>
<td>F</td>
<td>0</td>
<td>I</td>
<td>2</td>
<td>2</td>
<td>81</td>
<td>15</td>
<td>91.12</td>
<td>28</td>
<td>91.12</td>
</tr>
<tr>
<td>13</td>
<td>F</td>
<td>-1</td>
<td>R</td>
<td>1</td>
<td>2</td>
<td>85</td>
<td>3</td>
<td>90.26</td>
<td>28</td>
<td>90.26</td>
</tr>
<tr>
<td>13</td>
<td>F</td>
<td>0</td>
<td>I</td>
<td>14</td>
<td>26</td>
<td>23.55</td>
<td>14</td>
<td>55</td>
<td>14</td>
<td>55</td>
</tr>
<tr>
<td>13</td>
<td>F</td>
<td>0</td>
<td>I</td>
<td>43</td>
<td>46</td>
<td>107.50</td>
<td>14</td>
<td>55.65</td>
<td>14</td>
<td>55.65</td>
</tr>
<tr>
<td>13</td>
<td>F</td>
<td>0</td>
<td>R</td>
<td>42</td>
<td>20</td>
<td>62.66</td>
<td>14</td>
<td>55.65</td>
<td>14</td>
<td>55.65</td>
</tr>
<tr>
<td>13</td>
<td>M</td>
<td>0</td>
<td>I</td>
<td>75</td>
<td>18</td>
<td>107.50</td>
<td>40</td>
<td>87.26</td>
<td>39</td>
<td>87.26</td>
</tr>
<tr>
<td>13</td>
<td>M</td>
<td>0</td>
<td>HS</td>
<td>0</td>
<td>38</td>
<td>0.00</td>
<td>14</td>
<td>55.65</td>
<td>14</td>
<td>55.65</td>
</tr>
</tbody>
</table>

Notice
a: Relationship with the index case
  I: Index case
  M: Mother of the index case
  O: Offspring of the index case
  S: Sibling of the index case
  HS: Half sib of the index case
  R: Maternal relative of the index case
  UI: Unidentified relationship with the index case
b: Number of daughters or sons
  UIN: unidentified number of offspring
<table>
<thead>
<tr>
<th>Reference</th>
<th>Gender</th>
<th>Reverse Generation No.</th>
<th>Relationship with the Index Case</th>
<th>Number of Daughters</th>
<th>Number of Sons</th>
<th>Blood Mutation Level</th>
<th>Age</th>
<th>Age-Corrected Blood Mutation Level</th>
<th>Maternal Blood Mutation Level</th>
<th>Maternal Age-Corrected Mutation Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>13</td>
<td>M</td>
<td>0</td>
<td>R</td>
<td>22</td>
<td>38</td>
<td>47.04</td>
<td>28</td>
<td>59</td>
<td>91.12</td>
<td>28</td>
</tr>
<tr>
<td>13</td>
<td>M</td>
<td>0</td>
<td>R</td>
<td>17</td>
<td>31</td>
<td>31.60</td>
<td>28</td>
<td>59</td>
<td>91.12</td>
<td>28</td>
</tr>
<tr>
<td>13</td>
<td>F</td>
<td>0</td>
<td>R</td>
<td>28</td>
<td>32</td>
<td>33.10</td>
<td>28</td>
<td>59</td>
<td>91.12</td>
<td>28</td>
</tr>
<tr>
<td>14</td>
<td>F</td>
<td>2</td>
<td>M</td>
<td>8</td>
<td>2</td>
<td>6</td>
<td>72</td>
<td>25.32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>F</td>
<td>1</td>
<td>S</td>
<td>0</td>
<td>2</td>
<td>7</td>
<td>72</td>
<td>25.32</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>F</td>
<td>1</td>
<td>S</td>
<td>0</td>
<td>1</td>
<td>14</td>
<td>36</td>
<td>28.76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>M</td>
<td>1</td>
<td>S</td>
<td>0</td>
<td>39</td>
<td>0</td>
<td>6</td>
<td>72</td>
<td>25.32</td>
<td>6</td>
</tr>
<tr>
<td>14</td>
<td>M</td>
<td>1</td>
<td>S</td>
<td>0</td>
<td>43</td>
<td>0</td>
<td>6</td>
<td>72</td>
<td>25.32</td>
<td>6</td>
</tr>
<tr>
<td>14</td>
<td>F</td>
<td>1</td>
<td>S</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>47</td>
<td>30.72</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>F</td>
<td>1</td>
<td>S</td>
<td>1</td>
<td>1</td>
<td>48</td>
<td>0</td>
<td>0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>F</td>
<td>1</td>
<td>M</td>
<td>1</td>
<td>1</td>
<td>20</td>
<td>43</td>
<td>47.26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>M</td>
<td>0</td>
<td>I</td>
<td></td>
<td></td>
<td>30</td>
<td>18</td>
<td>43.00</td>
<td></td>
<td>20</td>
</tr>
<tr>
<td>16</td>
<td>F</td>
<td>2</td>
<td>S</td>
<td>0</td>
<td>2</td>
<td>6</td>
<td>72</td>
<td>7.64</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>F</td>
<td>1</td>
<td>O</td>
<td>0</td>
<td>0</td>
<td>44</td>
<td>19</td>
<td>64.34</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>F</td>
<td>0</td>
<td>R</td>
<td></td>
<td></td>
<td>48</td>
<td>23</td>
<td>76.04</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>M</td>
<td>1</td>
<td>S</td>
<td></td>
<td></td>
<td>28</td>
<td>41</td>
<td>63.57</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>F</td>
<td>2</td>
<td>R</td>
<td>2</td>
<td>1</td>
<td>17</td>
<td>49</td>
<td>45.30</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>M</td>
<td>3</td>
<td>S</td>
<td></td>
<td></td>
<td>37</td>
<td>53</td>
<td>106.80</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>M</td>
<td>2</td>
<td>R</td>
<td></td>
<td></td>
<td>17</td>
<td>58</td>
<td>54.23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>F</td>
<td>1</td>
<td>S</td>
<td>1</td>
<td>1</td>
<td>19</td>
<td>48</td>
<td>49.62</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>F</td>
<td>1</td>
<td>S</td>
<td>0</td>
<td>1</td>
<td>23</td>
<td>32</td>
<td>43.62</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>F</td>
<td>2</td>
<td>M</td>
<td>1</td>
<td>1</td>
<td>25</td>
<td>40</td>
<td>55.64</td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>M</td>
<td>0</td>
<td>R</td>
<td></td>
<td></td>
<td>59</td>
<td>15</td>
<td>79.64</td>
<td></td>
<td>19</td>
</tr>
<tr>
<td>17</td>
<td>F</td>
<td>1</td>
<td>S</td>
<td></td>
<td></td>
<td>10</td>
<td>42</td>
<td>23.16</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>F</td>
<td>3</td>
<td>S</td>
<td>0</td>
<td>2</td>
<td>18</td>
<td>25</td>
<td>29.68</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>F</td>
<td>2</td>
<td>S</td>
<td>2</td>
<td>0</td>
<td>28</td>
<td>28</td>
<td>49.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>F</td>
<td>1</td>
<td>R</td>
<td>0</td>
<td>0</td>
<td>36</td>
<td>21</td>
<td>54.79</td>
<td></td>
<td>17</td>
</tr>
<tr>
<td>17</td>
<td>F</td>
<td>1</td>
<td>S</td>
<td>0</td>
<td>1</td>
<td>36</td>
<td>21</td>
<td>54.79</td>
<td></td>
<td>25</td>
</tr>
<tr>
<td>17</td>
<td>F</td>
<td>1</td>
<td>S</td>
<td></td>
<td></td>
<td>46</td>
<td>8</td>
<td>53.98</td>
<td></td>
<td>28</td>
</tr>
<tr>
<td>17</td>
<td>F</td>
<td>1</td>
<td>S</td>
<td></td>
<td></td>
<td>43</td>
<td>21</td>
<td>65.44</td>
<td></td>
<td>14</td>
</tr>
<tr>
<td>17</td>
<td>M</td>
<td>1</td>
<td>I</td>
<td></td>
<td></td>
<td>59</td>
<td>18</td>
<td>84.57</td>
<td></td>
<td>25</td>
</tr>
<tr>
<td>17</td>
<td>F</td>
<td>0</td>
<td>R</td>
<td></td>
<td></td>
<td>50</td>
<td>6</td>
<td>56.37</td>
<td></td>
<td>28</td>
</tr>
<tr>
<td>17</td>
<td>F</td>
<td>1</td>
<td>S</td>
<td>3</td>
<td>0</td>
<td>11</td>
<td>54</td>
<td>32.39</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>F</td>
<td>1</td>
<td>S</td>
<td>1</td>
<td>1</td>
<td>15</td>
<td>49</td>
<td>39.97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>F</td>
<td>1</td>
<td>R</td>
<td>0</td>
<td>2</td>
<td>12</td>
<td>52</td>
<td>33.95</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>M</td>
<td>1</td>
<td>R</td>
<td></td>
<td></td>
<td>4</td>
<td>58</td>
<td>12.76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>F</td>
<td>0</td>
<td>R</td>
<td></td>
<td></td>
<td>15</td>
<td>33</td>
<td>29.02</td>
<td></td>
<td>11</td>
</tr>
<tr>
<td>17</td>
<td>F</td>
<td>0</td>
<td>R</td>
<td></td>
<td></td>
<td>11</td>
<td>30</td>
<td>20.04</td>
<td></td>
<td>11</td>
</tr>
</tbody>
</table>

**Notice**

a: Relationship with the index case
- I: Index case
- M: Mother of the index case
- O: Offspring of the index case
- S: Sibling of the index case
- HS: Half sib of the index case
- R: Maternal relative of the index case
- UI: Unidentified relationship with the index case

b: Number of daughters or sons

UIN: unidentified number of offspring
<table>
<thead>
<tr>
<th>Reference</th>
<th>Gender</th>
<th>Reverse Generation No.</th>
<th>Relationship with the Index Case</th>
<th>Number of Daughters</th>
<th>Number of Sons</th>
<th>Blood Mutation Level</th>
<th>Age</th>
<th>Age-Corrected Blood Mutation Level</th>
<th>Maternal Blood Mutation Level</th>
<th>Maternal Age-Corrected Mutation Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>17</td>
<td>F</td>
<td>0</td>
<td>R</td>
<td>8</td>
<td>28</td>
<td>14.01</td>
<td>11</td>
<td>54</td>
<td>32.39</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>F</td>
<td>0</td>
<td>R</td>
<td>24</td>
<td>27</td>
<td>41.18</td>
<td>15</td>
<td>49</td>
<td>39.97</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>M</td>
<td>0</td>
<td>R</td>
<td>31</td>
<td>24</td>
<td>50.10</td>
<td>15</td>
<td>49</td>
<td>39.97</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>M</td>
<td>0</td>
<td>R</td>
<td>13</td>
<td>26</td>
<td>21.87</td>
<td>12</td>
<td>52</td>
<td>33.95</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>M</td>
<td>0</td>
<td>R</td>
<td>13</td>
<td>26</td>
<td>21.87</td>
<td>12</td>
<td>52</td>
<td>33.95</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>F</td>
<td>-1</td>
<td>I</td>
<td>1</td>
<td>0</td>
<td>19.52</td>
<td>12</td>
<td>52</td>
<td>33.95</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>F</td>
<td>-1</td>
<td>I</td>
<td>0</td>
<td>1</td>
<td>28.48</td>
<td>12</td>
<td>52</td>
<td>33.95</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>F</td>
<td>0</td>
<td>I</td>
<td>0</td>
<td>1</td>
<td>29.17</td>
<td>12</td>
<td>52</td>
<td>33.95</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>F</td>
<td>0</td>
<td>I</td>
<td>0</td>
<td>0</td>
<td>29.17</td>
<td>12</td>
<td>52</td>
<td>33.95</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>F</td>
<td>-2</td>
<td>M</td>
<td>2</td>
<td>3</td>
<td>16.31</td>
<td>6</td>
<td>50</td>
<td>16.31</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>F</td>
<td>-1</td>
<td>S</td>
<td>2</td>
<td>0</td>
<td>5.36</td>
<td>6</td>
<td>50</td>
<td>16.31</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>M</td>
<td>-1</td>
<td>S</td>
<td>0</td>
<td>0</td>
<td>5.36</td>
<td>6</td>
<td>50</td>
<td>16.31</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>F</td>
<td>0</td>
<td>R</td>
<td>18</td>
<td>7</td>
<td>20.70</td>
<td>3</td>
<td>29</td>
<td>5.36</td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>F</td>
<td>0</td>
<td>R</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
<td>3</td>
<td>29</td>
<td>5.36</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>F</td>
<td>-1</td>
<td>M</td>
<td>1</td>
<td>3</td>
<td>34.75</td>
<td>0</td>
<td>78</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>F</td>
<td>-1</td>
<td>R</td>
<td>0</td>
<td>0</td>
<td>0.00</td>
<td>0</td>
<td>78</td>
<td>0.00</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>F</td>
<td>0</td>
<td>S</td>
<td>0</td>
<td>20</td>
<td>0.00</td>
<td>15</td>
<td>42</td>
<td>34.75</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>F</td>
<td>0</td>
<td>S</td>
<td>0</td>
<td>15</td>
<td>0.00</td>
<td>15</td>
<td>42</td>
<td>34.75</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>M</td>
<td>0</td>
<td>S</td>
<td>0</td>
<td>12</td>
<td>0.00</td>
<td>15</td>
<td>42</td>
<td>34.75</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>F</td>
<td>-1</td>
<td>I</td>
<td>1</td>
<td>2</td>
<td>17.44</td>
<td>44</td>
<td>40.99</td>
<td>40.99</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>F</td>
<td>-2</td>
<td>R</td>
<td>1</td>
<td>2</td>
<td>22.70</td>
<td>44</td>
<td>40.99</td>
<td>40.99</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>M</td>
<td>0</td>
<td>O</td>
<td>2</td>
<td>0</td>
<td>22.70</td>
<td>44</td>
<td>40.99</td>
<td>40.99</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>F</td>
<td>0</td>
<td>O</td>
<td>2</td>
<td>0</td>
<td>22.70</td>
<td>44</td>
<td>40.99</td>
<td>40.99</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>F</td>
<td>-1</td>
<td>M</td>
<td>1</td>
<td>1</td>
<td>72.89</td>
<td>31</td>
<td>44</td>
<td>74.74</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>M</td>
<td>0</td>
<td>O</td>
<td>0</td>
<td>15</td>
<td>22.02</td>
<td>31</td>
<td>44</td>
<td>74.74</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>M</td>
<td>0</td>
<td>I</td>
<td>1</td>
<td>1</td>
<td>23.45</td>
<td>31</td>
<td>44</td>
<td>74.74</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>F</td>
<td>0</td>
<td>S</td>
<td>0</td>
<td>20</td>
<td>15.90</td>
<td>31</td>
<td>44</td>
<td>74.74</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>F</td>
<td>-2</td>
<td>M</td>
<td>4</td>
<td>0</td>
<td>22.02</td>
<td>31</td>
<td>44</td>
<td>74.74</td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>F</td>
<td>-2</td>
<td>S</td>
<td>0</td>
<td>6</td>
<td>22.02</td>
<td>31</td>
<td>44</td>
<td>74.74</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>F</td>
<td>-2</td>
<td>S</td>
<td>2</td>
<td>2</td>
<td>36.78</td>
<td>5</td>
<td>61</td>
<td>36.78</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>F</td>
<td>-2</td>
<td>I</td>
<td>1</td>
<td>2</td>
<td>36.78</td>
<td>5</td>
<td>61</td>
<td>36.78</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>M</td>
<td>-1</td>
<td>R</td>
<td>1</td>
<td>3</td>
<td>8.06</td>
<td>61</td>
<td>16.94</td>
<td>16.94</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>F</td>
<td>-1</td>
<td>I</td>
<td>1</td>
<td>1</td>
<td>34.34</td>
<td>61</td>
<td>16.94</td>
<td>16.94</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>F</td>
<td>-1</td>
<td>I</td>
<td>0</td>
<td>1</td>
<td>41.87</td>
<td>4</td>
<td>69</td>
<td>15.90</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>F</td>
<td>-1</td>
<td>R</td>
<td>1</td>
<td>3</td>
<td>24.17</td>
<td>5</td>
<td>61</td>
<td>16.94</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>F</td>
<td>-1</td>
<td>R</td>
<td>0</td>
<td>9</td>
<td>15.76</td>
<td>5</td>
<td>61</td>
<td>16.94</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>F</td>
<td>-1</td>
<td>R</td>
<td>0</td>
<td>17</td>
<td>28.03</td>
<td>5</td>
<td>61</td>
<td>16.94</td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>M</td>
<td>-1</td>
<td>O</td>
<td>2</td>
<td>0</td>
<td>36.44</td>
<td>12</td>
<td>56</td>
<td>36.78</td>
<td></td>
</tr>
</tbody>
</table>

**Notice**

- a: Relationship with the index case
  - I: Index case
  - M: Mother of the index case
  - O: Offspring of the index case
  - S: Sibling of the index case
  - HS: Half sib of the index case
  - R: Maternal relative of the index case
  - UI: Unidentified relationship with the index case

- b: Number of daughters or sons
  - UIN: unidentified number of offspring
<table>
<thead>
<tr>
<th>Reference</th>
<th>Gender</th>
<th>Reverse</th>
<th>Relationship with the Index</th>
<th>Number of Daughters</th>
<th>Number of Sons</th>
<th>Blood Mutation Level</th>
<th>Age</th>
<th>Age-Corrected Blood Mutation Level</th>
<th>Maternal Blood Mutation Level</th>
<th>Maternal Age</th>
<th>Maternal Age-Corrected Mutation Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>22</td>
<td>F</td>
<td>-1</td>
<td>O</td>
<td>0</td>
<td>1</td>
<td>17</td>
<td>26</td>
<td>28.59</td>
<td>12</td>
<td>56</td>
<td>36.78</td>
</tr>
<tr>
<td>23</td>
<td>F</td>
<td>-1</td>
<td>M</td>
<td>0</td>
<td>2</td>
<td>11</td>
<td>38</td>
<td>23.52</td>
<td>0</td>
<td>72</td>
<td>0.00</td>
</tr>
<tr>
<td>23</td>
<td>M</td>
<td>0</td>
<td>I</td>
<td>56</td>
<td>14</td>
<td>74.10</td>
<td>11</td>
<td>38</td>
<td>0</td>
<td>72</td>
<td>23.52</td>
</tr>
<tr>
<td>24</td>
<td>M</td>
<td>0</td>
<td>S</td>
<td>65</td>
<td>8</td>
<td>76.28</td>
<td>11</td>
<td>38</td>
<td>0</td>
<td>72</td>
<td>23.52</td>
</tr>
<tr>
<td>24</td>
<td>F</td>
<td>-1</td>
<td>S</td>
<td>0</td>
<td>0</td>
<td>25</td>
<td>46</td>
<td>62.73</td>
<td>0</td>
<td>72</td>
<td>23.52</td>
</tr>
<tr>
<td>24</td>
<td>F</td>
<td>-1</td>
<td>S</td>
<td>0</td>
<td>0</td>
<td>24</td>
<td>41</td>
<td>54.49</td>
<td>0</td>
<td>72</td>
<td>23.52</td>
</tr>
<tr>
<td>24</td>
<td>F</td>
<td>-1</td>
<td>S</td>
<td>0</td>
<td>1</td>
<td>27</td>
<td>40</td>
<td>60.09</td>
<td>0</td>
<td>72</td>
<td>23.52</td>
</tr>
<tr>
<td>24</td>
<td>F</td>
<td>-1</td>
<td>S</td>
<td>0</td>
<td>1</td>
<td>27</td>
<td>39</td>
<td>58.90</td>
<td>0</td>
<td>72</td>
<td>23.52</td>
</tr>
<tr>
<td>24</td>
<td>M</td>
<td>-1</td>
<td>I</td>
<td>28</td>
<td>35</td>
<td>56.39</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>F</td>
<td>0</td>
<td>R</td>
<td>21</td>
<td>16</td>
<td>28.92</td>
<td>27</td>
<td>40</td>
<td>60.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>F</td>
<td>0</td>
<td>R</td>
<td>33</td>
<td>14</td>
<td>43.66</td>
<td>27</td>
<td>40</td>
<td>60.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>F</td>
<td>0</td>
<td>R</td>
<td>40</td>
<td>13</td>
<td>51.88</td>
<td>27</td>
<td>40</td>
<td>60.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>M</td>
<td>0</td>
<td>R</td>
<td>47</td>
<td>10</td>
<td>57.41</td>
<td>27</td>
<td>40</td>
<td>60.09</td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>M</td>
<td>0</td>
<td>R</td>
<td>42</td>
<td>8</td>
<td>49.29</td>
<td>27</td>
<td>39</td>
<td>58.90</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>F</td>
<td>-1</td>
<td>M</td>
<td>3</td>
<td>2</td>
<td>11</td>
<td>48</td>
<td>28.73</td>
<td>11</td>
<td>48</td>
<td>28.73</td>
</tr>
<tr>
<td>25</td>
<td>F</td>
<td>0</td>
<td>I</td>
<td>32</td>
<td>24</td>
<td>51.71</td>
<td>11</td>
<td>48</td>
<td>28.73</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>F</td>
<td>0</td>
<td>S</td>
<td>19</td>
<td>20</td>
<td>28.34</td>
<td>11</td>
<td>48</td>
<td>28.73</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>M</td>
<td>0</td>
<td>S</td>
<td>21</td>
<td>19</td>
<td>30.71</td>
<td>11</td>
<td>48</td>
<td>28.73</td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>M</td>
<td>0</td>
<td>S</td>
<td>27</td>
<td>17</td>
<td>37.93</td>
<td>11</td>
<td>48</td>
<td>28.73</td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>F</td>
<td>-1</td>
<td>R</td>
<td>0</td>
<td>1</td>
<td>37.8</td>
<td>42</td>
<td>87.56</td>
<td>0</td>
<td>72</td>
<td>23.52</td>
</tr>
<tr>
<td>26</td>
<td>M</td>
<td>-1</td>
<td>I</td>
<td>36.7</td>
<td>36</td>
<td>75.40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>M</td>
<td>-1</td>
<td>S</td>
<td>58.1</td>
<td>28</td>
<td>101.71</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>M</td>
<td>0</td>
<td>R</td>
<td>35.4</td>
<td>0</td>
<td>53.88</td>
<td>37.8</td>
<td>42</td>
<td>87.56</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>F</td>
<td>-2</td>
<td>R</td>
<td>3</td>
<td>2</td>
<td>7.19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>F</td>
<td>-1</td>
<td>S</td>
<td>2</td>
<td>0</td>
<td>14</td>
<td>54</td>
<td>41.23</td>
<td>2</td>
<td>64</td>
<td>7.19</td>
</tr>
<tr>
<td>27</td>
<td>M</td>
<td>-1</td>
<td>R</td>
<td>30</td>
<td>41</td>
<td>68.11</td>
<td>2</td>
<td>64</td>
<td>7.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>F</td>
<td>-1</td>
<td>M</td>
<td>40</td>
<td>39</td>
<td>87.26</td>
<td>2</td>
<td>64</td>
<td>7.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>F</td>
<td>-1</td>
<td>R</td>
<td>38</td>
<td>37</td>
<td>58.69</td>
<td>2</td>
<td>64</td>
<td>7.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>M</td>
<td>-1</td>
<td>I</td>
<td>24</td>
<td>43</td>
<td>56.72</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>F</td>
<td>-1</td>
<td>R</td>
<td>52</td>
<td>25</td>
<td>85.73</td>
<td>2</td>
<td>64</td>
<td>7.19</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>F</td>
<td>0</td>
<td>R</td>
<td>13</td>
<td>25</td>
<td>23.08</td>
<td>14</td>
<td>54</td>
<td>41.23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>F</td>
<td>0</td>
<td>R</td>
<td>42</td>
<td>18</td>
<td>60.20</td>
<td>14</td>
<td>54</td>
<td>41.23</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>M</td>
<td>0</td>
<td>I</td>
<td>75</td>
<td>19</td>
<td>109.67</td>
<td>40</td>
<td>39</td>
<td>87.26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>F</td>
<td>0</td>
<td>S</td>
<td>59</td>
<td>15</td>
<td>79.64</td>
<td>40</td>
<td>39</td>
<td>87.26</td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>F</td>
<td>0</td>
<td>I</td>
<td>25</td>
<td>16</td>
<td>34.44</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>F</td>
<td>0</td>
<td>S</td>
<td>9</td>
<td>39</td>
<td>19.63</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>M</td>
<td>0</td>
<td>I</td>
<td>5</td>
<td>35</td>
<td>10.07</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>F</td>
<td>0</td>
<td>I</td>
<td>13</td>
<td>37</td>
<td>27.25</td>
<td>78</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>M</td>
<td>0</td>
<td>I</td>
<td>1</td>
<td>23</td>
<td>1.58</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notice:  
- a: Relationship with the index case  
  - I: Index case  
  - M: Mother of the index case  
  - O: Offspring of the index case  
  - S: Sibling of the index case  
  - HS: Half sib of the index case  
  - R: Maternal relative of the index case  
  - UI: Unidentified relationship with the index case  

- b: Number of daughters or sons  
  - UIN: unidentified number of offspring
<table>
<thead>
<tr>
<th>Reference</th>
<th>Gender</th>
<th>Reverse Generation No.</th>
<th>Relationship with the Index Case</th>
<th>Number of Daughters</th>
<th>Number of Sons</th>
<th>Blood Mutation Level</th>
<th>Age</th>
<th>Age-Corrected Blood Mutation Level</th>
<th>Maternal Blood Mutation Level</th>
<th>Maternal Age</th>
<th>Maternal Age-Corrected Mutation Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>29</td>
<td>F</td>
<td>0</td>
<td>I</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>32</td>
<td>7.59</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>F</td>
<td>0</td>
<td>S</td>
<td>1</td>
<td>0</td>
<td>14</td>
<td>38</td>
<td>29.94</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>F</td>
<td>-1</td>
<td>M</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>69</td>
<td>19.87</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>M</td>
<td>-1</td>
<td>R</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>76</td>
<td>13.72</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>F</td>
<td>-2</td>
<td>R</td>
<td>1</td>
<td>1</td>
<td>38</td>
<td>74</td>
<td>30.75</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>M</td>
<td>0</td>
<td>S</td>
<td>0</td>
<td>0</td>
<td>23</td>
<td>49</td>
<td>74.60</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>F</td>
<td>-1</td>
<td>S</td>
<td>0</td>
<td>2</td>
<td>18</td>
<td>46</td>
<td>45.17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>F</td>
<td>0</td>
<td>I</td>
<td>1</td>
<td>1</td>
<td>13</td>
<td>37</td>
<td>37.09</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>F</td>
<td>-1</td>
<td>M</td>
<td>1</td>
<td>1</td>
<td>18</td>
<td>55</td>
<td>54.07</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>F</td>
<td>-4</td>
<td>R</td>
<td>1</td>
<td>0</td>
<td>17</td>
<td>39</td>
<td>37.09</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>M</td>
<td>-3</td>
<td>R</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>54</td>
<td>11.78</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>F</td>
<td>-1</td>
<td>R</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>55</td>
<td>12.02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>M</td>
<td>1</td>
<td>R</td>
<td>1</td>
<td>1</td>
<td>15</td>
<td>47</td>
<td>38.40</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>M</td>
<td>-1</td>
<td>R</td>
<td>1</td>
<td>0</td>
<td>16</td>
<td>45</td>
<td>39.35</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>F</td>
<td>-3</td>
<td>R</td>
<td>1</td>
<td>0</td>
<td>10</td>
<td>43</td>
<td>23.63</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>M</td>
<td>0</td>
<td>R</td>
<td>2</td>
<td>26</td>
<td>42</td>
<td>46</td>
<td>45.17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>M</td>
<td>0</td>
<td>I</td>
<td>1</td>
<td>0</td>
<td>45</td>
<td>50</td>
<td>54.07</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>M</td>
<td>0</td>
<td>R</td>
<td>1</td>
<td>1</td>
<td>30</td>
<td>46</td>
<td>45.17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>F</td>
<td>0</td>
<td>S</td>
<td>3</td>
<td>28</td>
<td>45</td>
<td>55</td>
<td>54.07</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>F</td>
<td>0</td>
<td>R</td>
<td>1</td>
<td>3</td>
<td>5</td>
<td>8</td>
<td>45.17</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>M</td>
<td>0</td>
<td>R</td>
<td>4</td>
<td>33</td>
<td>7</td>
<td>4</td>
<td>11.78</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>F</td>
<td>0</td>
<td>R</td>
<td>1</td>
<td>1</td>
<td>13</td>
<td>39</td>
<td>19.01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>F</td>
<td>-1</td>
<td>I</td>
<td>0</td>
<td>1</td>
<td>16</td>
<td>52</td>
<td>45.27</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>F</td>
<td>-2</td>
<td>I</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>55</td>
<td>18.03</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>M</td>
<td>0</td>
<td>O</td>
<td>3</td>
<td>24</td>
<td>59</td>
<td>52</td>
<td>45.27</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>M</td>
<td>0</td>
<td>O</td>
<td>4</td>
<td>25</td>
<td>72</td>
<td>55</td>
<td>18.02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>F</td>
<td>-1</td>
<td>M</td>
<td>1</td>
<td>2</td>
<td>15</td>
<td>60</td>
<td>49.80</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>F</td>
<td>-1</td>
<td>R</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>52</td>
<td>42.44</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>M</td>
<td>0</td>
<td>I</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>45</td>
<td>36.89</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>M</td>
<td>0</td>
<td>I</td>
<td>0</td>
<td>0</td>
<td>15</td>
<td>38</td>
<td>0.00</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>F</td>
<td>0</td>
<td>S</td>
<td>2</td>
<td>28</td>
<td>54</td>
<td>60</td>
<td>49.80</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>M</td>
<td>0</td>
<td>M</td>
<td>2</td>
<td>43</td>
<td>88</td>
<td>36</td>
<td>88.34</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>M</td>
<td>0</td>
<td>S</td>
<td>4</td>
<td>49</td>
<td>59</td>
<td>43</td>
<td>88.34</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>F</td>
<td>0</td>
<td>I</td>
<td>2</td>
<td>68</td>
<td>76</td>
<td>43</td>
<td>88.34</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>M</td>
<td>0</td>
<td>I</td>
<td>1</td>
<td>54</td>
<td>65</td>
<td>18</td>
<td>37.18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>F</td>
<td>-1</td>
<td>M</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>45</td>
<td>12.30</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Notice:

a: Relationship with the index case
   I: Index case
   M: Mother of the index case
   O: Offspring of the index case
   S: Sibling of the index case
   HS: Half sib of the index case
   R: Maternal relative of the index case
   UI: Unidentified relationship with the index case

b: Number of daughters or sons
   UIN: unidentified number of offspring
6 Chapter 6: Conclusion and Future Directions

This chapter aims to review original contributions of this dissertation and highlight future research directions. Each section presents a brief conclusion of each research project. The last section is devoted to discussing the possible future research directions.

6.1 Section 6.1: Mitochondrial genetic bottleneck

In Chapter 2, we presented an experimental study using a neutral mouse model complemented by a computational simulation of mouse embryogenesis to investigate how the mitochondrial genetic bottleneck generates a large random shift in heteroplasmy level between a mother and her offspring. The experimental mouse model, designed and carried out by our collaborators, showed a drastic reduction of mtDNA copy numbers per cell during blastocyst formation, generating a small number of mtDNA molecules per cell among progenitor germ cells. This small mtDNA copy number in a germ line precursor is a physical bottleneck of mtDNA population. The embryogenesis simulation, developed in our lab, showed that the offspring heteroplasmy variation is largely generated by a random partition of mtDNA molecules during pre- and early post-implantation development. Our result indicated that an individual mtDNA molecule can well be a mitochondrial segregating unit.

6.2 Section 6.2: Distribution pattern of mtDNA heteroplasmy level

In Chapter 3, we modified a set of probability functions developed by Motoo Kimura in 1955 [117] to describe the distribution of mtDNA heteroplasmy level of a subsequent generation of a single maternal ancestor. The theoretical Kimura distribution was tested against 16 actual mtDNA heteroplasmy distributions measured from several organisms, including human, mouse and Drosophila. The consistency between the theoretical Kimura distribution and the actual mtDNA heteroplasmy distribution presents the potential application of population genetic theory to mitochondrial genetics.

6.3 Section 6.3: Sampling error of heteroplasmy variance

Generally, the heteroplasmy variance has been calculated from a modest number of mtDNA heteroplasmy measurements, which may be sufficient to get a reliable estimate population parameter for this higher-order statistics. In Chapter 4, we addressed this concern by developing several models, varied by the underlying distribution, to calculate a standard error of variance. We also proposed how to normalize heteroplasmy variance. This normalization is necessary for the comparison of the heteroplasmy variance values across different data sets with different heteroplasmy mean values.

The application of both the standard error of variance calculation and the normalization to the variance of the actual heteroplasmy measurements can alter the biological interpretation and reveal the difference in mitochondrial bottleneck between humans and mice. The result clearly shows that, with limited human data, humans have a tighter bottleneck than mice. The bottleneck size difference is present both in primary oocytes and offspring. In addition, the standard error of variance calculation can be applied to estimate a proper sample size for an experimental design related to heteroplasmy variance calculation.
6.4 Section 6.4: The mothers carrying a high proportion of A3243G mutation tends to have fewer daughters than sons

In Chapter 5, we aim to study the effect of the A3243G mutation on female reproduction. We started the analysis by studying the transmission pattern of this mutant heteroplasmy level. The application of an age-correction to blood heteroplasmy measurements revealed that the transmission of this mutant heteroplasmy level of the mothers carrying a low proportion of this mutation follows a neutral random drift architecture, while the transmission of this mutation from the mother carrying a high proportion of this mutation is subject to a purifying selection mechanism.

The effect of this mutation on female reproduction has been shown by the comparison of the average number of daughters to the average number of sons from the mother carrying low and from the mother carrying high mutation level. The average number of daughters is statistically significant lower than the average number of sons of the mother carrying a high proportion of this common pathogenic mtDNA mutation. This offspring gender bias has an implication for why we tend to observe a small pedigree carrying this mutation.

6.5 Section 6.5: Research overview and future directions

The main purpose of this dissertation is to gain an understanding of how human mtDNA heteroplasmy is inherited. This has an implication for an estimation of recurrence risk of having an affected individual with mitochondrial disease in the family carrying a pathogenic mtDNA mutation. In Chapter 2, we have a clearer picture of how mitochondrial genetic bottleneck generates a random shift in heteroplasmy level between a mother and her offspring and variation of this measurement among offspring siblings. Then in Chapter 3, the application of probability functions that were developed to describe the segregation of allele frequencies subject to random genetic drift helps us to describe the intra-generational distribution of mtDNA heteroplasmy level. In Chapter 4, the applications of the normalization method and the standard error of variance calculation to the actual mtDNA heteroplasmy level variance reveal the difference in mitochondrial genetic bottleneck between humans and mice. Finally in Chapter 5, we observed the tendency of bearing sons in the mother carrying a high proportion of m.3243A>G mutation via the statistical analyses of the published human clinical pedigrees.

The purpose of this section is to discuss future research directions, which aims both to answer some fundamental questions in the field of mitochondrial genetics and to develop tools for clinical practice. Some fundamental questions that need to be answered are for example:

**How large is the variation of b parameter in humans?**

The b parameter value can be deduced from the heteroplasmy mean and variance, as shown in Chapter 3. This parameter is necessary for estimating the probability value of an individual carrying mutant mtDNA at a particular level, having an implication for an estimation of a recurrence risk in the family carrying a pathogenic mtDNA mutation. Knowing the range of this parameter value would gain confidence in the probability estimation.

The variation of this parameter value could be presented both at the level of the type of pathogenic mtDNA mutation and at an individual level, thus its numerical value should be
calculated from a set of heteroplasmy measurements of the descendents or germ line cells of a single maternal ancestor. We can calculate this parameter value from published human clinical pedigrees, although each individual mother in the pedigree typically has a very small number of offspring, leading to a requirement to estimate this parameter value from a group of mothers. This is complicated by the fact that the dependency of heteroplasmy variance on the heteroplasmy mean affects the estimation of this parameter. To solve this problem, either the normalization or the transformation method needs to be developed in order to calculate the b parameter across a range of heteroplasmy levels.

**Does selection play a major role in the transmission of pathogenic mtDNA mutation?**

There are observations from two different experimental mouse models indicating that negative selection plays a role in regulating mtDNA heteroplasmy transmission [107, 118]. In fact, to some extent, the transmission of pathogenic mtDNA mutation may be subjected to selection in which its effect would be different from one type of pathogenic mtDNA mutation to another mutation, both in quality and quantity. In particular, this question has been driven by the observation of a significant increase of offspring heteroplasmy level from the mother carrying m.8993T>G (OMIM+516060.0001) mtDNA mutation with no evidence of a reduction of heteroplasmy level toward age.

We can study the effect of selection on the transmission of the pathogenic mtDNA mutation by testing whether the offspring heteroplasmy level significantly differs from the mother’s heteroplasmy level, which the mother and offspring heteroplasmy level can be obtained from a set of published human clinical pedigrees. We can use the same procedure as presented in Chapter 4 to exercise this question. The direction of the heteroplasmy level difference between a mother and offspring can indicate whether the transmission mechanism is a random drift or selection. We can further examine whether the significant difference is the result of the change of heteroplasmy level toward age by comparing the intergenerational heteroplasmy variation to the intra-generational heteroplasmy variation.

**How can we properly adjust for the effect of ascertainment bias in human clinical pedigree data?**

The human clinical pedigree data is typically subjected to an ascertainment bias. The ascertainment bias could distort the segregation pattern of the allele of interest [119]. For the study of nDNA transmission, there is a refined method to adjust for this confounding factor, knowing whether the mutation is dominant or recessive [119]. However, there is no such method to adjust for this confounding factor in the case of mtDNA heteroplasmy transmission. The general accepted procedure is excluding all index cases from the analysis [89-90], although this simple process may not be effective enough to eliminate the ascertainment bias.

We can apply Kimura distribution and defined pedigree structure to build the multigenerational transmission model of mtDNA heteroplasmy level or the pedigree model. In this model, the theoretical Kimura distribution is applied to determine an individual’s heteroplasmy level with b parameter estimated either from human primary oocyte heteroplasmy measurements [110] or from human clinical pedigree data of the mother carrying the intermediate heteroplasmy level as used in Chapter 4. We can simply assume the distribution of the number of offspring per mother to follow the Poisson distribution [120] with the offspring
gender ratio of one to one or to apply the structure extracted from human clinical pedigree data, as presented in Chapter 5.

For this particular question, we can design the pedigree simulation model to make a statistical comparison between the total pedigree families carrying mutant mtDNA and their fraction being brought to clinical attention. The statistical comparison result should help us to develop the method to correct the ascertainment bias for the study of the segregation pattern of mtDNA heteroplasm level.

How does the particular pathogenic mtDNA mutation affect human population?

We can adapt the pedigree model to build the population model of mtDNA heteroplasm transmission by simultaneously generating multiple pedigrees. This model would help us gain insight into how the pathogenic mtDNA mutation segregates in a human population. At a particular time point, we can observe the proportion of carriers and affected individuals in the system. The validity of the population model can be judged by comparing the simulated population prevalence to the actual population prevalence presented in published literatures.

In clinical practice, we need to improve an accuracy of genetic counseling for the family carrying the pathogenic mtDNA mutation. At this point, knowing the b parameter value and the maternal heteroplasm level, we can apply Kimura distribution to estimate the probability of her offspring carrying mutant mtDNA higher than the threshold level. This probability value would, to some extent, determine the probability of her offspring being affected. In medical practice, the b parameter can be estimated from a set of heteroplasm measurements of a maternal tissue sample, which the sampling effect can be addressed by applying standard error of variance calculation, developed in Chapter 4. If the maternal tissue sample is the female germ line cell, this strategy can be used to predict a recurrence risk prior to conception. If the maternal tissue sample is collected after conception, either through pre-implantation genetic diagnosis (PGD) [121] or prenatal diagnosis (PND) process [122-125], application of this idea complemented by the knowledge of the mtDNA heteroplasm segregation between different cells and different tissues would help to estimate the likelihood of the offspring being affected. If there are other known factors determining the severity of the disease phenotype, those modifying factors need to be taken into account to the likelihood estimation of the offspring being affected. Thereby the future research project would be to test the validity of this whole idea of preventive transmission strategy. In particular, we build the computational simulation model by taking all known factors determining an expression of disease phenotype into account- then statistically compare the simulated data against the clinical observation. If the simulation model can well describe the clinical observation, this computational model would be a useful tool for mitochondrial genetic counseling. The simulation can be extended to predict the recurrence risk among other relatives, such as the cousin, by incorporating all factors determining the severity of the disease phenotype into the pedigree model.
References


116. Wonnapinij, P., P.F. Chinnery, and D.C. Samuels, Previous estimates of mitochondrial DNA mutation level variance did not account for sampling error:


Appendix

This section contains information about copyright permission. In particular, the copyright permission letter from Nature Genetics journal and online statement for Elsevier’s journal authors are included in this section.
This is a License Agreement between Passorn Wonnapinij ("You") and Nature Publishing Group ("Nature Publishing Group") provided by Copyright Clearance Center ("CCC"). The license consists of your order details, the terms and conditions provided by Nature Publishing Group, and the payment terms and conditions.

All payments must be made in full to CCC. For payment instructions, please see information listed at the bottom of this form.

| License Number   | 2324260745970 |
| License date     | Dec 08, 2009  |
| Licensed content publisher | Nature Publishing Group |
| Licensed content publication | Nature Genetics |
| Licensed content title | A reduction of mitochondrial DNA molecules during embryogenesis explains the rapid segregation of genotypes |
| Licensed content author | Lynsey M Cree, David C Samuels, Susana Chuva de Sousa Lopes, Harsha Karur Rajasimha, Passorn Wonnapinij, Jeffrey R Mann, Hans-Henrik M Dahl, Patrick F Chinnery |
| Volume number    |               |
| Issue number     |               |
| Pages            |               |
| Year of publication | 2008 |
| Portion used     | Full paper    |
| Requestor type   | Student       |
| Type of Use      | Thesis / Dissertation |
| Billing Type     | Invoice       |
| Company          | Passorn Wonnapinij |
| Billing Address  | 808 Glenavon Ct. |
|                  | Nashville, TN 37220 |
|                  | United States  |
| Customer reference info |         |
| Total            | 0.00 USD      |

Terms and Conditions for Permissions

Nature Publishing Group hereby grants you a non-exclusive license to reproduce this material for this purpose, and for no other use, subject to the conditions below:

1. NPG warrants that it has, to the best of its knowledge, the rights to license reuse of this
material. However, you should ensure that the material you are requesting is original to Nature Publishing Group and does not carry the copyright of another entity (as credited in the published version). If the credit line on any part of the material you have requested indicates that it was reprinted or adapted by NPG with permission from another source, then you should also seek permission from that source to reuse the material.

2. Permission granted free of charge for material in print is also usually granted for any electronic version of that work, provided that the material is incidental to the work as a whole and that the electronic version is essentially equivalent to, or substitutes for, the print version. Where print permission has been granted for a fee, separate permission must be obtained for any additional, electronic re-use (unless, as in the case of a full paper, this has already been accounted for during your initial request in the calculation of a print run). NB: In all cases, web-based use of full-text articles must be authorized separately through the 'Use on a Web Site' option when requesting permission.

3. Permission granted for a first edition does not apply to second and subsequent editions and for editions in other languages (except for signatories to the STM Permissions Guidelines, or where the first edition permission was granted for free).

4. Nature Publishing Group's permission must be acknowledged next to the figure, table or abstract in print. In electronic form, this acknowledgement must be visible at the same time as the figure/table/abstract, and must be hyperlinked to the journal's homepage.

5. The credit line should read:

Reprinted by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication)

For AOP papers, the credit line should read:

Reprinted by permission from Macmillan Publishers Ltd: [JOURNAL NAME], advance online publication, day month year (doi: 10.1038/sj.[JOURNAL ACRONYM].XXXXX)

6. Adaptations of single figures do not require NPG approval. However, the adaptation should be credited as follows:

Adapted by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication)

7. Translations of 401 words up to a whole article require NPG approval. Please visit http://www.macmillanmedicalcommunications.com for more information. Translations of up to a 400 words do not require NPG approval. The translation should be credited as follows:

Translated by permission from Macmillan Publishers Ltd: [JOURNAL NAME] (reference citation), copyright (year of publication).

We are certain that all parties will benefit from this agreement and wish you the best in the
Gratis licenses (referencing $0 in the Total field) are free. Please retain this printable license for your reference. No payment is required.

If you would like to pay for this license now, please remit this license along with your payment made payable to "COPYRIGHT CLEARANCE CENTER" otherwise you will be invoiced within 30 days of the license date. Payment should be in the form of a check or money order referencing your account number and this license number 2324260745970.

If you would prefer to pay for this license by credit card, please go to http://www.copyright.com/creditcard to download our credit card payment authorization form.

Make Payment To:
Copyright Clearance Center
Dept 001
P.O. Box 843006
Boston, MA 02284-3006

If you find copyrighted material related to this license will not be used and wish to cancel, please contact us referencing this license number 2324260745970 and noting the reason for cancellation.

Questions? customercare@copyright.com or +1-877-622-5543 (toll free in the US) or +1-978-646-2777.
What rights do I retain as a journal author*?

As a journal author, you retain rights for large number of author uses, including use by your employing institute or company. These rights are retained and permitted without the need to obtain specific permission from Elsevier. These include:

- the right to make copies (print or electric) of the journal article for their own personal use, including for their own classroom teaching use;
- the right to make copies and distribute copies (including via e-mail) of the journal article to research colleagues, for personal use by such colleagues (but not for Commercial Purposes**, as listed below);
- the right to post a pre-print version of the journal article on Internet web sites including electronic pre-print servers, and to retain indefinitely such version on such servers or sites (see also our information on electronic preprints for a more detailed discussion on these points);
- the right to post a revised personal version of the text of the final journal article (to reflect changes made in the peer review process) on the author’s personal or institutional web site or server, incorporating the complete citation and with a link to the Digital Object Identifier (DOI) of the article;
- the right to present the journal article at a meeting or conference and to distribute copies of such paper or article to the delegates attending the meeting;
- for the author’s employer, if the journal article is a 'work for hire', made within the scope of the author’s employment, the right to use all or part of the information in (any version of) the journal article for other intra-company use (e.g.

---

*limitations apply

**Commercial Purposes: for purposes of sale, rent, lease or other commercial exploitation or distribution (such as for payment or profit)

---

(C) 2007 Elsevier B.V. All rights reserved.
training), including by posting the article on secure, internal corporate intranets;

- patent and trademark rights and rights to any process or procedure described in the journal article;
- the right to include the journal article, in full or in part, in a thesis or dissertation;
- the right to use the journal article or any part thereof in a printed compilation of works of the author, such as collected writings or lecture notes (subsequent to publication of the article in the journal); and
- the right to prepare other derivative works, to extend the journal article into book-length form, or to otherwise re-use portions or excerpts in other works, with full acknowledgement of its original publication in the journal.

*Please Note: The rights listed above apply to journal authors only. For information regarding book author rights, please contact the Global Rights Department.*

Elsevier Global Rights Department
phone (+44) 1865 843 830
fax (+44) 1865 853 333
email: permissions@elsevier.com

Other uses by authors should be authorized by Elsevier through the Global Rights Department, and journal authors are encouraged to let Elsevier know of any particular needs or requirements.

**Commercial Purposes includes the use or posting of articles for commercial gain including the posting by companies or their employee-authored works for use by customers of such companies (e.g. pharmaceutical companies and physician-prescribers); commercial exploitation such as directly associating advertising with such postings; the charging of fees for document delivery or access; or the systematic distribution to others via e-mail lists or list servers (to parties other than known colleagues), whether for a fee or for free.**

When Elsevier changes its journal usage policies, are those changes also retroactive?

When Elsevier changes its policies to enable greater academic use of journal materials (such as the changes several years ago in our web-posting policies) or to clarify the rights retained by journal authors, Elsevier is prepared to extend those rights retroactively with respect to articles published in journal issues produced prior to the policy change.

Elsevier is pleased to confirm that, unless explicitly noted to the contrary, all policies apply retrospectively to previously published journal content. If, after reviewing the material noted above, you have any questions about such rights, please contact Global Rights.

How do I obtain a Journal Publishing Agreement?

You will receive a form automatically by post or email once your article is received by Elsevier's Editorial-Production Department. View a generic example of the agreement here. Some journals will use another variation of this form.

Can I post my article on the Internet?

You can post your version of your journal article on your personal web page or the web site of your institution, provided that you include a link to the journal's home page or the article's DOI and include a complete citation for the article. This means that you can update your version (e.g. the Word or Tex form) to reflect changes.
made during the peer review process.

Complete information about proper usage and citation of journal articles can be found in this pamphlet, published by the Global Rights Department.

**Why does Elsevier request transfer of copyright?**

Elsevier wants to ensure that it has the exclusive distribution rights for all media. Copyright transfer eliminates any ambiguity or uncertainty about Elsevier's ability to distribute, sub-license and protect the article from unauthorized copying or alteration.

**Why does Elsevier believe it needs exclusive rights?**

The research community needs certainty with respect to the validity of scientific papers, which is normally obtained through the editing and peer review processes. The scientific record must be clear and unambiguous. Elsevier believes that, by obtaining exclusive distribution rights, it will always be clear to researchers that when they access an Elsevier site to review a paper, they are reading a final version of the paper which has been edited, peer-reviewed and accepted for publication in an appropriate journal. See also our information on electronic preprints for more detailed discussion on these points.

**Can you provide me with a PDF file of my article?**

Many Elsevier journals are now offering authors e-offprints – free electronic versions of published articles. E-offprints are watermarked PDF versions, and are usually delivered within 24 hours, much quicker than print copies. These PDFs may not be posted to public websites.

For more information, please see your Journal's Guide to Authors or contact sciencereprints@elsevier.com

**What is Elsevier's policy on plagiarism and conflicts of interest?**

While it may not be possible to draft a "code" that applies adequately to all instances and circumstances, Elsevier and the journal editors believe it useful to outline our expectations of authors and procedures that the journal will employ in the event of questions concerning author conduct. Procedures and guidelines with respect to such queries and investigations are outlined in the Elsevier Position on Journal Publishing Ethics and Responsibilities and Conflicts of Interest are incorporated herein by reference, and should be reviewed by authors. The policies of individual journals may differ, please review the relevant journal's instructions to authors.

© Copyright 2010 Elsevier | http://www.elsevier.com