CHAPTER ONE

Introduction

Evolutionary studies aim at the interpretation of observed biological patterns for specific purposes such as the reconstruction of human evolutionary history. Investigation of human nuclear and mitochondrial DNA sequences indicates that patterns of DNA sequence variation can be identified and interpreted to uncover past evolutionary events and their impact on biological systems. Evolutionary studies are critical in eliciting the fundamental principles of living systems and understanding the phylogeny of these biologically complex systems within and among populations of living organisms.

The quest to determine the origins of humans and the evolutionary processes that created *Homo sapiens* have continued for many years. The most direct evidence has been obtained from fossils, which have, for example, been crucial in the establishment of the origin of archaic humans in Africa (Mellars, 2006). Anthropology and archaeology are additional fields of research that have contributed greatly to information on the human race and culture, society and human physical traits over thousands of years. The type of fossil anthropological and archaeological evidence that is available depends on the type of discoveries that are made and as a consequence there are some gaps in the evolutionary history when it is based on only this type of evidence. Genetic data offer a mechanism to view human evolution without the constraints of discovery. By measuring genetic variation, it is possible to assess the origins and behaviour of past populations based on the genetic signatures left by major demographic events, such as population migrations, expansions and bottlenecks, and evolutionary forces such as natural selection and genetic drift. The imprints of these evolutionary events are transmitted to future generations to create an extant human genome that contains invaluable information about past evolutionary events (Jorde *et al*., 1998). The field of genomics therefore combines the knowledge of genome-wide sequence variation with the principles of population genetics and evolutionary history and constitutes the interpretation of processes of microevolution such as mutation, natural selection and genetic drift in context with the processes of macroevolution, such as population expansion, bottlenecks or gene flow (Luikart *et al*., 2003).

The process of mutation is a major evolutionary force that governs genetic variability and results in substitutions of nucleotides with incorrect nucleotide incorporations or the deletion or insertion of incorrect nucleotides, mostly owing to errors in the DNA replicative processes in the cell, or in a lower number of cases because of exposure to negative environmental agents (Lynch, 2008). The rate of mutation differs between different regions of the nuclear genome and between different organelles, for example between the nuclear genome and the mitochondrial genome. Germ line mutations are carried over from one generation to the next and are therefore the basis of evolutionary change, as opposed to mutations that have occurred in the somatic cells, which are not carried over to the next generation (Page and Holmes, 1998). The process of conversion of a mutation in a single individual to a fixed state as a polymorphism in a population is dependent on the presence of evolutionary forces such as natural selection and genetic drift that would ensure that a mutation is retained and radiated to future generations or lost from a population in order to shape the genetic variability within populations of individuals.

The great evolutionists, Charles Darwin and Alfred Wallace, were the first scientists to propose a theory to explain the effect of evolutionary processes on living organisms (Darwin and Wallace, 1858; Wallace, 1858). They hypothesised that the fundamental changes that were observed between organisms could be ascribed to positive selection that enabled the organisms to adapt to their environmental challenges and thus survive. The theories of Darwin were further extended by scientists such as Dobzhansky, Huxley, Mayr and Simpson (Dobzhansky, 1937; Huxley, 1863, Mayr, 1944; Simpson, 1944), who developed a theory of Modern Synthesis that described the process of genetic drift as an evolutionary force that governed the fixation of mutations in the genomes of living entities. This was followed by an elaboration of the concepts of genetic drift by Kimura (1971), who proposed that genetic variation was mainly driven by the evolutionary processes of genetic drift and not natural selection in view of the selective neutrality of most mutations in living organisms. Natural selection became a point of contention, as some controversy developed over the role of natural selection and genetic drift in shaping genetic variation (Nielsen, 2005). Although many of the debates over the role of evolutionary forces continue, these different models of evolution provided a mechanism to study the evolutionary genetics of a population. By modelling a set of genetic sequence variation data against the assumptions of a specific evolutionary model, it is possible to interpret the genetic signals of human genome variation (Jobling *et al.*, 2004).

In addition to statistical modelling of evolutionary theory, phylogenetics, which is a field of study that developed in conjunction with the concepts of macroevolution, is fundamentally based on mathematical evolutionary models that incorporate biological, biochemical and evolutionary information to model the genetic relatedness of variations of a set of individuals or entities (Whelan *et al.,* 2001). Phylogenetic tree analysis is a popular method used to analyse genetic variation within or among human populations and represents the genetic diversity between individuals, as well as the fissures and behaviour of a human population over a long period of time (Jobling *et al*., 2004).

The human mitochondrial genome displays unique characteristics that make it suitable for use as an evolutionary marker in population genetics studies. Mitochondria are present in high numbers in the cells of many organisms, including humans, and are essential for life because they are fundamentally involved in cellular energy metabolism through the role they play in adenosine-5-triphosphate (ATP) production via the oxidative phosphorylation system (OXPHOS), fatty acid β oxidation and the urea cycle (Chinnery and Schon, 2003). Mitochondria contain a DNA genome which consists of 16,569 nucleotide pairs (np) in a closed double-stranded circular structure (Anderson *et al.*, 1981). The mitochondrial DNA (mtDNA) encodes for 13 polypeptides that are involved in the mitochondrial energy-generating OXPHOS system, two (2) ribosomal RNAs (rRNA) and 22 transfer RNAs (tRNA) that are used to translate the polypeptides that are encoded by the mtDNA (Anderson *et al.*, 1981; Wallace *et al.*, 1995). Most of the genes that code for the mitochondrial proteins were transferred to the nuclear DNA and only the genes that are responsible for the functioning of the inner membrane OXPHOS enzyme complexes were retained in the mitochondria (Wallace *et al.*, 1999).

The mtDNA has several unique features based on the critical role it plays in the survival of an organism, which makes it highly suitable as an evolutionary marker. The mtDNA has high copy numbers in the human cell and is located outside the nucleus in the mitochondrial inner-membrane space, making it easily accessible for isolation and assay (Pakendorf and Stoneking *et al.*, 2005). The mtDNA displays a high mutation rate, especially in the non-coding regions of the genome, which produce high levels of genetic variation. This high rate of evolution in the mitochondrial genome is believed to be a mechanism to ensure the rapid adaptability of an individual to changing energy needs in reaction to changing environmental conditions, such as climate and diet (Wallace, 1994). The high mutation rate of the mitochondrial genome results in human populations that

contain high levels of population-specific polymorphisms, which are well suited to the study of human evolution.

The mitochondria of the sperm are destroyed by the oocyte during reproduction, making the inheritance of the mtDNA strictly maternal (Giles *et al.*, 1980; Sutovsky *et al.*, 2000). The uniparental mechanism of inheritance rules out the decay caused by deleterious mutations introduced through sexual reproduction and recombination and conserves the ability of the mitochondrial genome to provide the most effective energy solution to the environment (Felsenstein, 1974). In addition, the direct transmission of genetic variation from mother to child without the confounding effects of recombination makes it possible to trace ancestral lineages back to a single ancestor, which is ideal for the study of evolution.

MtDNA variation is displayed in the human genome as haplogroups that consist of shared mutations of the human mitochondrial genome, which are acquired through the sequential accumulation of mutations through maternal lineages that have developed over time. mtDNA haplogroups are suitable for the study of evolution because they consist of sequence motifs that are closely linked, which show very little, if any recombination and have ancient origins. MtDNA haplogroups provide a molecular record of the genealogical history of a population as well as a molecular trace of migration patterns of humans over time. The radiating maternal lineages also provide good evidence of the extent of genetic subdivision among populations of regions and continents (Torroni *et al.,* 2006).

MtDNA variation also plays an important part in understanding age-related, metabolic and neurodegenerative diseases, the process of ageing and cancer. Deficiency in energy metabolism is the primary cause of compromised reproduction and growth of living entities and mitochondrial disease generally induces this lack of cellular energy (Wallace, 2007). Several connections have been made between mitochondrial haplogroups and various clinical conditions, such as the high penetrance of Leber's hereditary optic neuropathy (LHON) in individuals that belong to haplogroup J (Brown *et al.*, 1994; Torroni *et al.*, 1997). The study of mtDNA variation of a human population would therefore not only contribute to elucidating its evolutionary past, but also to the identification of possible pathogenic mutations that could have a direct influence on the identification of mtDNA disease associations with possible haplogroup associations, disease aetiology and treatment approaches.

1.1 OBJECTIVES OF THIS RESEARCH STUDY

By determination of the extant genetic diversity of present-day African populations, it is possible to gain evidence of human origins and the demographic history of the migrations of modern human populations across Africa. The purpose of recognising DNA variation in human populations, in addition to understanding the evolutionary history of the human species, is to advance the studies of medicinal and developmental biology. This in turn allows for the identification of population-specific genetic variations in African populations that have an effect on gene function as well as on adaptive variation in morphology, behaviour, physiology and susceptibility to disease. The Bantu-speaking populations of South Africa are often phenotypically, linguistically and culturally similar and the definitions of ethnicity are therefore vague (Campbell and Tishkof., 2010), which makes the need for determination of the genetic diversity of these populations critical in order to address the underlying genetic mechanisms of complex diseases within and among these populations.

The broad purpose of this mitochondrial investigation was to determine the mtDNA variation of 50 Tswana-speaking individuals that represent a Bantu-speaking population of South Africa to determine the evolutionary history of this Bantu-speaking cohort and the possibility of pathogenicity of the observed sequence variants, with the associated impact on health and treatment options. The full mitochondrial genomes of the 50 Tswana-speaking individuals of this investigation were studied by using mitochondrial DNA in two distinct approaches. The first was to study the history of mitochondrial DNA lineages by using the mtDNA haplogroups of the Tswana-speaking individuals of this investigation to identify to what extent they were related to the mtDNA sequences of other African and non-African individuals by performing phylogenetic analyses. The other approach was to study the genetic variability of the Tswana-speaking population under investigation by using population genetics methods that generally consisted of statistical methods. These methods used assumptions against which the dataset of observed mtDNA sequence variation of the Tswana-speaking individuals of this investigation, as well as the mtDNA sequence variation of a broad dataset of African and non-African individuals, was tested for validity. Finally, a novel Tswana consensus sequence was constructed based on the sequence variance observed in the full mitochondrial genomes of the 50 Tswana-speaking individuals of this investigation. The purpose of the consensus sequence was to provide a baseline for the sequence variance that is present in a Tswana-speaking population of South Africa and as a representation of the genetic diversity of the maternal ancestral genetic pool of a Bantu-speaking population of South Africa.

1.2 SPECIFIC AIMS OF THE PROJECT

The specific aims of the project were:

a. To determine the full mitochondrial genome sequences of 50 Tswana-speaking individuals of South Africa.
   - To compare the sequence variation observed in the Tswana-speaking individuals of this investigation with other published studies.
   - To identify novel sequence alterations in the Tswana-speaking individuals of this investigation.
   - To identify possible disease-associated sequence variants in the Tswana-speaking individuals of this investigation.

b. To assign mitochondrial haplogroups to the 50 Tswana-speaking individuals of this investigation.
   - To compare the haplogroup composition of the Tswana-speaking individuals of this investigation with the haplogroup composition of a broad set of mtDNA sequences that belong to African and non-African haplogroup L individuals.

c. To investigate the evolutionary relationships of the mtDNA sequences of the Tswana-speaking individuals of this investigation within the Tswana-speaking population and among a broad set of mtDNA sequences of African and non-African individuals to determine the relatedness to other African and non-African haplogroup L populations through the construction of phylogenetic trees.

d. To investigate the population history and past behaviour of the Tswana-speaking population of this investigation by using population genetics methods that are based on the statistical evaluation of the observed sequence variation of the Tswana-speaking population under investigation and the observed sequence variation of a broad set of African and non-African haplogroup L individuals against evolutionary models.
   - To determine the basic statistics of genetic diversity.
   - To determine if the observed sequence variation of the Tswana-speaking population of this investigation and the observed sequence variation of a broad set of mitochondrial sequences of African and non-African haplogroup L individuals, as well as regional populations of African haplogroup L individuals, deviated from the neutral model assumptions, and if so, whether it was caused by the genetic signature of natural selection or population expansion.
   - To investigate the level of genetic differentiation between and among the Tswana-speaking population of this investigation, the broad set of African and

non-African haplogroup L individuals and the regional populations of African haplogroup L individuals.

- To determine the time to the most recent common ancestor (TMRCA) for the haplogroups of the phylogenetic tree that consisted of the mtDNA sequences of the Tswana-speaking individuals of this investigation and a broad set of mtDNA sequences of African and non-African haplogroup L individuals.

e. To construct a novel mitochondrial consensus sequence for the South African Tswana-speaking population of this investigation.