CHAPTER SEVEN

Conclusions

Populations are commonly defined by self-identified ethnicity and cultural identity that only describe the history of a population for a few hundred years, whereas genetic variation describes the evolutionary history of a population over a few thousand years. Genetically defined population history therefore predates the cultural heritage of populations and is critical to enable a valid and meaningful interpretation of the evolutionary history of a population. As the connection between morphological characteristics and mtDNA genetic distances is weak and a poor indicator of ethnic ancestry, it is only through the genetic diversity of populations that an accurate estimate can be obtained of evolutionary history and genetic similarity (Nei, 1982). By determination of the extant genetic diversity of present-day African populations, it is possible to gain evidence of human origin 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 human ethnicity, is to advance the studies of medicinal and developmental biology. It allows for the identification of population-specific genetic variation in African populations that has 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 vague (Campbell and Tishkoff, 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 concept of genomics, which underlies the processes of evolution at several loci or genomic regions in the context of broad population behaviour, such as the origin and the demographic history of populations (Luikart *et al.*, 2003), was applied to the Tswana cohort of this investigation to address the aims of this study. These aims were to compare the genetic variability of a Tswana population of South Africa in relation to other African and non-African individuals and to determine a consensus mtDNA sequence for the Tswana-speaking population under investigation, as a mechanism to infer the evolutionary history of this population. The mtDNA genetic diversity of the Tswana population of this

investigation was determined by whole mtDNA sequencing of the 50 individuals of the Tswana cohort. Based on the principle that genetic diversity is determined by the maternal germ line mutations in the mtDNA that are transferred to the offspring, which are under pressure of selection and genetic drift, i.e. they will either become established or lost within the population, the genetic variation within the present-day Tswana population can be regarded as evidence of the evolutionary history of this South African Bantu-speaking population. The genetic diversity of Tswana-speaking individuals of this investigation was used to identify the genetic similarities or differences compared to other African and non-African populations, taking cognisance of the fact that it had affected genetic disease and phenotype. The genomic variation of the population under investigation was considered in the context of its phylogenetic relationships with the observed genomic variation of other African and non-African individuals.

In order to understand the implications of the genetic variation observed in the Tswana population under investigation, the genetic variance of the population was measured by identifying the different patterns of sequence variation within the population, naming these according to two standard classification systems, determining the frequencies of the patterns of variance within the different populations and determining the phylogenetic relationships of those patterns. In addition, statistical and computational tools were used to estimate the effect of natural selection and genetic drift on the gene pool under investigation.

Genomic evolution was thus determined by measuring mitochondrial genome-wide heterogeneity and modelling it against the backdrop of population behaviour over time. This implied the identification of genome-wide effects of selection and mutation, as well as the identification of whole genome effects of genetic drift or bottlenecks and gene flow. The genetic variance of the population under investigation was therefore ascribed to dynamics and parameters of the individual genomes, as well as the behaviour of the population of genomes. Two components, the individual component within the genome due to ancient and recent mutations as well as selection; and the population component due to demography, genetic drift and gene flow, comprised the fundamental processes that needed to be measured and modelled in order to understand the gene pool of the population (Luikart *et al.*, 2003). According to this approach, the genetic variation of the Tswana population of this investigation will be interpreted in the context of evolutionary history by discussing the individual parameters that caused the genetic diversity and their

implications and by discussing the population dynamics that interplayed with the evolutionary mutational events to shape evolutionary history.

7.1 STANDARDS USED FOR DETERMINATION OF GENETIC VARIATION

The rCRS (Andrews *et al.*, 1999) has been widely used since 1981 in mitochondrial studies of human evolution and disease and was used as a standard sequence with which the mtDNA sequences of the Tswana-speaking study group were compared to identify sequence alterations. It has long been accepted that, based on the radiation of maternal lineages into different continental populations, mutations evolved sequentially and accumulated at high frequencies to form region-specific mtDNA sequence alteration profiles (Wallace *et al.*, 1999). These sequence alterations could be identified by comparison to the standard rCRS and organised into groups characterised by sets of sequence motifs, which were assigned to haplogroups (Wallace *et al.*, 1999). The rCRS therefore provided a standard against which sequence variants of the mitochondrial genomes could be identified and classified into haplogroups, as well as providing a uniform numbering structure that was used to identify the positions at which sequence alterations were observed in the mtDNA sequences of the Tswana-speaking cohort.

The assignment of haplogroups was based on the presence of sequence variants in the mtDNA genomes under investigation, as identified by the nucleotide positions of the rCRS at which it differed (Wallace *et al.*, 1999). Furthermore, it was important to know the phylogenetic relationship of mtDNA variants in order to classify sequence variants into appropriate and phylogenetically correct haplogroups (Van Oven and Kayser, 2009). Two classification schemes were used for this purpose in this investigation. The classification system adapted from Wallace (2004) was based on informative SNPs and designated for the use of classification of the major L haplogroups and sub-haplogroups only. It was adapted for purposes of haplogroup assignment of mtDNA coding regions of individuals of African origin by the CGR at the North-West University (Wallace, 2004). The PhyloTree classification system (Van Oven and Kayser, 2009) was used in conjunction with the Wallace classification system in this investigation because it represented a global mtDNA phylogeny based on mitochondrial control and coding region sequence alterations observed in all currently published sequence data and therefore offered a highly resolved global haplogroup hierarchy (Van Oven and Kayser, 2009).

Haplogroups became more resolved over time as sequencing technologies developed and enabled scientists to obtain more sequence variation data to define the regional population haplogroups to a level that was not possible with the initial typing methods, such as high-resolution RFLP analyses (Van Oven and Kayser, 2009). It was therefore not surprising to find that the Wallace classification system did not provide the same level of haplogroup resolution as the PhyloTree classification system (Van Oven and Kayser, 2009). The PhyloTree classification system was constructed from all published full mtDNA sequences that were available at the time of the construction of the global phylogenetic tree and is regularly updated with new published sequence variants, thus ensuring that the haplogroup classification system is current at all times (Van Oven and Kayser, 2009). The incorporation of the 309 full mitochondrial sequences and 315 previously published mtDNA sequences of individuals of African origin by Behar *et al.* (2008) greatly enhanced the resolution of haplogroup L in the PhyloTree classification system (Van Oven and Kayser, 2009) and was therefore of great importance to this investigation.

In contrast to the Wallace classification system (Wallace, 2004), the PhyloTree classification system has been peer-reviewed, is publicly accessible and provides an open platform for constant updates and correspondence with the authors of mitochondrial studies (Van Oven and Kayser, 2009). This has led to the standard use of this classification system in mitochondrial studies after its construction in 2009, which became a critical factor in the development of a standard approach to the discussion of haplogroups and sub-haplogroups between studies (Batini *et al.*, 2011; de Fillipo *et al.*, 2010; Scheinfeldt *et al.*, 2010). On assignment of the mtDNA coding regions of the individuals of the Global African and All African mtDNA sequence datasets of this investigation by using both haplogroup classification systems, the PhyloTree classification system (Van Oven and Kayser, 2009) haplogroup assignments were in agreement with the published haplogroup assignments of the same mtDNA sequences by Pereira *et al.* (2009), whereas the Wallace classification system (2004) haplogroup assignments were not. It was therefore concluded that although the Wallace classification system (2004) is a valid system for the assignment of haplogroups of individuals belonging to haplogroup L lineages, the PhyloTree classification system (Van Oven and Kayser, 2009) is a more resolved and current haplogroup assignment system that has been verified and is openly accessible to all researchers. It was therefore invaluable in this investigation to position the mtDNA sequences of the Tswana-speaking study group relative to other global haplogroups.

7.2 MITOCHONDRIAL VARIATION IN THE TSWANA POPULATION DUE TO INDIVIDUAL FACTORS

The mtDNA sequence variation of a cohort of 50 Tswana-speaking individuals from South Africa was determined by automated sequencing of the full mitochondrial DNA. This provided mtDNA data in which sequence alterations were observed in the form of alterations that were transitions, transversions or sequence variations caused by the deletion or insertion of base pairs in the mtDNA of the study group of Tswana-speaking individuals. The observed alterations that occurred in these Tswana-speaking individuals of this investigation were regarded as segregated homoplasmic alterations that occurred throughout the maternal ancestral lineages of these individuals, as a non-recombining locus and therefore reflected the evolutionary history of the Tswana-speaking individuals.

The observed alterations were expected to reflect neutral changes to the mitochondrial genome, or weak or mildly deleterious effects caused by nonsynonymous changes in amino acids as opposed to severely deleterious mutations that caused lethal disease and would most likely have presented as new heteroplasmic mutations that were not used in this study. The characterisation of the sequence variants observed in the Tswana cohort under investigation provided information about population-specific sequence alterations that played a role in gene function and possible susceptibility to disease within the Tswana-speaking population under investigation.

7.2.1 Sequence variation displayed in the mtDNA genomes

Four hundred and forty sequence alterations were observed in the full mitochondrial genomes of the 50 Tswana-speaking individuals, consisting of 404 transitions, 12 transversions and 24 indels. A transversion:transition count ratio (Lutz-Bonengel *et al.*, 2003) of 1:33.6 was therefore displayed by the full genomes of the mtDNA sequences of the Tswana-speaking cohort, which demonstrated that only about 3% of the sequence variants were transversions, as opposed to a 97% proportion of transitions. The rarity of transversions in the evolution of humans has long been recognised by other studies (Moilanen and Majamaa, 2003; Pereira *et al.*, 2009) and was confirmed by the findings of this investigation. This phenomenon is ascribed to the free energies of mispairing during DNA replication between the template DNA and the incoming base in combination with pressure to maintain the double helix DNA structure, which in general favours transition mispairing above transversion mispairing of nucleotides (Wakeley, 1993).

When considering only the protein-coding regions of the mitochondrial DNA of the Tswana-speaking individuals of this investigation, 293 sequence alterations are observed, which constitute about 3% of the total of 11,395 base pairs that make up the gene regions. The low frequency of sequence alterations observed in the protein-coding regions of these Tswana-speaking individuals is in agreement with a study conducted by Pereira *et al*. (2009), in which it was demonstrated that most of the sequence variants that contributed to the high levels of genetic diversity in the human mitochondrial genome did not occur in the coding regions of the mtDNA. The level of sequence alterations that occurred in the coding regions of the Tswana-speaking individuals is however remarkably lower than the level of sequence alterations of the protein-coding regions of a large global mtDNA dataset of 5,140 sequences observed by Pereira *et al.* (2009). This phenomenon is ascribed to the small size of the Tswana-speaking cohort when compared to the size of the dataset used in the study performed by Pereira *et al.* (2009) and highlights the necessity of large sample sets to screen populations for rare sequence variants. The global dataset used in the study by Pereira *et al.* (2009) further contained a broad spectrum of mtDNA sequences belonging to individuals that originated from many different global populations, as opposed to the singular regional population group under investigation in this study, and therefore it was concluded that the low frequency of sequence alterations in the coding regions was not unexpected for the Tswana dataset. The impact on the sequence variation when investigating a single regional population is further demonstrated by the fact that the observed sequence alterations in the Tswana cohort are not distributed evenly between the 13 genes of the mitochondrial genomes, as was expected (Pereira *et al.,* 2009). The observed number of sequence alterations rangesfrom only six sequence alterations in the *ND4L* gene region to 45 sequence alterations observed in the *ND5* gene region and again highlights that a large sample set would be necessary to identify all rare sequence alterations and this may ultimately lead to more evenly distributed numbers of sequence alterations between the gene regions under an assumption of neutral selective pressure. The uneven distribution of sequence alterations between the genes could, however, also have been caused by the presence of strong adaptive selective forces in some of the genes, which would have favoured the fixation of advantageous nonsynonymous mutations and therefore could have displayed a high frequency of sequence alterations in the genes under selection. The presence of selection in the mtDNA genomes of the Tswana-speaking individuals was investigated and the conclusions with regard to the effect of selection on the genetic diversity observed are discussed in Section 7.3.5.

Two hundred and thirteen of the protein-coding sequence substitutions are synonymous and 80 are nonsynonymous, which leads to a ratio of 1:2.7 nonsynonymous:synonymous substitutions. The observed number of nonsynonymous substitutions is lower than the observed number of synonymous substitutions, which supports the theory that substitution changes in the mitochondrial genome are dependent on factors such as the consequences of amino acid changes for the viability of the organism, the sequence composition, variable rates of substitutions and selective constraints (Xia *et al*., 1997). Twenty-seven percent of the substitution changes are nonsynonymous, which in agreement with the observed rate of 28.4% nonsynonymous:substitutions in a large dataset of European mtDNA sequences (Moilanen and Majamaa, 2003). Seventeen percent of the sequence variants observed in the protein-coding regions are located in first codon positions, 13% are located in second codon positions and 70% located in third codon positions. It can therefore be concluded that the mtDNA sequences of the Tswana-speaking cohort display a sequence variance profile similar to that observed in other much larger global populations and suggests that the mutational processes behaved as would be expected for the human mitochondrial genome (Moilanen and Majamaa, 2003; Pereira *et al.*, 2009).

The control region of the mitochondrial genome of the Tswana-speaking individuals display 104 sequence alterations, which constitute about 9% of the total number of 1,122 base pairs of the control region, of which 77 are transitions, 13 transversions and 14 indels. The higher proportion of mutations in the control region testifies to the higher rate of mutation within this region of the mitochondrial genome, as well as to the higher prevalence of transitions compared to transversions and indels.

The findings of the sequence variation observed in the rRNA and tRNA genes of the mtDNA sequences of the Tswana-speaking cohort demonstrate that the secondary structure conformation of these genes is the main reason for the uneven distribution of mtDNA sequence diversity in the rRNA and tRNA regions. The rRNA coding regions of the Tswana-speaking individuals of this investigation display 32 sequence alterations, of which 29 are located in the loop regions of the 12S and 16S rRNA secondary structures. This testifies to the reported prevalence of mutations in the loop regions of the rRNA gene structures because of the conservation of the functionality in the stem regions (Pereira *et al.*, 2009). The tRNA-coding regions of the Tswana-speaking study group display 22 sequence alterations in total, which is lower than in the other regions of the mitochondrial genome, as was reported in other studies (Vilmi *et al.*, 2005; Pereira *et al.*, 2009), and

confirms that tRNA genes display higher levels of conservation than synonymous sites in protein-coding regions (Moilanen and Majamaa, 2003). The locations of the sequence variants observed in the tRNA genes of the Tswana-speaking cohort are evenly distributed between the stem regions and loop regions, which is contradictory to the functional constraints present in the stem regions of the rRNA genes of the mitochondrial genomes. The redundancy of the substitutions and the fact that a larger portion of the tRNA gene consists of stem regions are considered as possible reasons for this phenomenon (Pereira *et al*., 2009). The Tswana-speaking study group display an uneven distribution of sequence variants between the different tRNA genes. The high number of sequence variants observed in the tRNA threonine region is in agreement with the observation of tRNA sequence variation reported by other studies (Vilmi *et al*., 2005). It has been concluded that the tRNA genes of the Tswana-speaking cohort display sequence variation in agreement with other mtDNA sequence datasets (Moilanen and Majamaa, 2003; Vilmi *et al*., 2005; Pereira *et al*., 2009).

About 5% of the sequence alterations observed in the full mitochondrial genomes of the Tswana-speaking study group consists of indels, which is in agreement with other studies of mtDNA sequence variation (Moilanen and Majamaa, 2003; Pereira *et al.*, 2009). The 14 indels in the control region were mostly localised in regions of homopolymer C stretches or hypervariable regions, which have been identified as regions that often display insertions or deletions (Bandelt and Parson, 2008; Van Oven and Kayser, 2009). Another stretch of deletions are located in the non-coding region from np 8281 to np 8289 and is a reported marker associated with the Bantu migrations about 4,000 ybp (Soodyall *et al.*, 1996), which was expected to be present in a Bantu-speaking population. It has been concluded that the number and positions of indels observed in the Tswana-speaking cohort were expected and in agreement with published data.

7.2.1.1 Novel sequence variants observed in this investigation

Seventeen novel sequence variants have been observed in the Tswana-speaking study group. Single novel sequence variants are observed in the control region of the mtDNA, the *ND2* gene region, the tRNA cysteine coding region, the *COI* gene region, the *COIII* gene region and the *Cytb* gene region. Two novel mutations are observed in each of the following gene regions of the Tswana-speaking cohort: *ND1*, *ND4* and *ND6*, whereas the *ND5* gene region displayed a total of five novel sequence variants.

It is concluded that most of the novel sequence variants are private mutations based on the fact that 15 of the novel sequence variants only occur once in the Tswana-speaking study group. The size of the Tswana cohort is a limiting factor in terms of representing rare variants and the possibility that some of these novel sequence variants are in fact haplogroup-defining mutations and present in higher frequencies in the larger Tswana population of South Africa than that observed in this investigation must be considered.

Ten of the novel mutations in the protein-coding regions resulted in synonymous amino acid changes and five resulted in nonsynonymous amino acid changes. The hypothesis that private mutations are more likely to result in nonsynonymous amino acid changes than non-private mutations (Moilanen and Majamaa, 2003) was supported by the fact that each of the four novel nonsynonymous substitutions is observed in a single Tswana-speaking individual, suggesting the possibility that those mutations were examples of private mutations. The findings of this investigation further support the hypothesis that non-conservative mildly deleterious changes in the mitochondrial genome will prevail in the form of single private mutations and will be removed from the mitochondrial genome over time by negative selection (Kivisild *et al.*, 2006).

The novel sequence variant observed at np 13473 in the *ND5* gene region in two Tswana-speaking individuals of this investigation is not clustered together in the phylogenetic trees of this investigation and belongs to haplogroups L0k and L0d1 respectively. It has been concluded that this novel sequence variant arose independently in the two Tswana-speaking individuals of this investigation and did not hold a possibility of defining a new haplogroup.

The novel sequence variant at np 12436 in the *ND5* gene region that was observed in five Tswana-speaking individuals belonging to haplogroup L0d1b forms a distinct cluster in all of the phylogenetic trees of this investigation. It is concluded that this novel sequence variant is fixed in the Tswana-speaking population of South Africa and therefore could be indicative of a new haplogroup (Van Oven and Kayser, 2009).

Twenty-seven sequence variants that are novel to haplogroup L lineages were observed in the Tswana-speaking study group. Two novel haplogroup L sequence variants were observed in the control region of the mtDNA of the Tswana-speaking cohort. Two novel haplogroup L sequence variants were observed in the 12S rRNA coding region and one in the 16S rRNA coding region. The tRNA coding regions displayed only one novel haplogroup L sequence variant in the tRNA asparagine region. The *COI*, *COIII*, *ND5* and *ND6* gene regions displayed one novel haplogroup L sequence variant each, the *ND1*, *ND2*, *COII*, *ND3* and *ND4* gene regions displayed two novel haplogroup L sequence variants each and the *Cytb* and *ATP6* gene regions displayed three and four novel haplogroup L sequence variants respectively. Fifteen of the protein-coding region novel haplogroup L sequence variants were synonymous and five were nonsynonymous changes.

All of the novel haplogroup L sequence variants except for the transitions at np 10128 and np 15337 were each observed in a single Tswana-speaking individual, suggesting that these sequence variants were private mutations that had occurred recently. These sequence variants could not be considered as haplogroup-defining, as they were not present in more than one individual of the population under investigation. The novel haplogroup L sequence variant at np 10128 occurred within the *ND3* gene regions of five Tswana-speaking study group, which belonged to haplogroup L0d3 and clustered together in a separate clade within the L0d3 clades of the Global African NJ and MP phylogenetic trees of this investigation, indicating a separate and distinct sequence motif that was shared by the Tswana-speaking individuals of this investigation only. Therefore it was concluded that this sequence variant was a valid candidate for a sub-haplogroup of haplogroup L0d3.

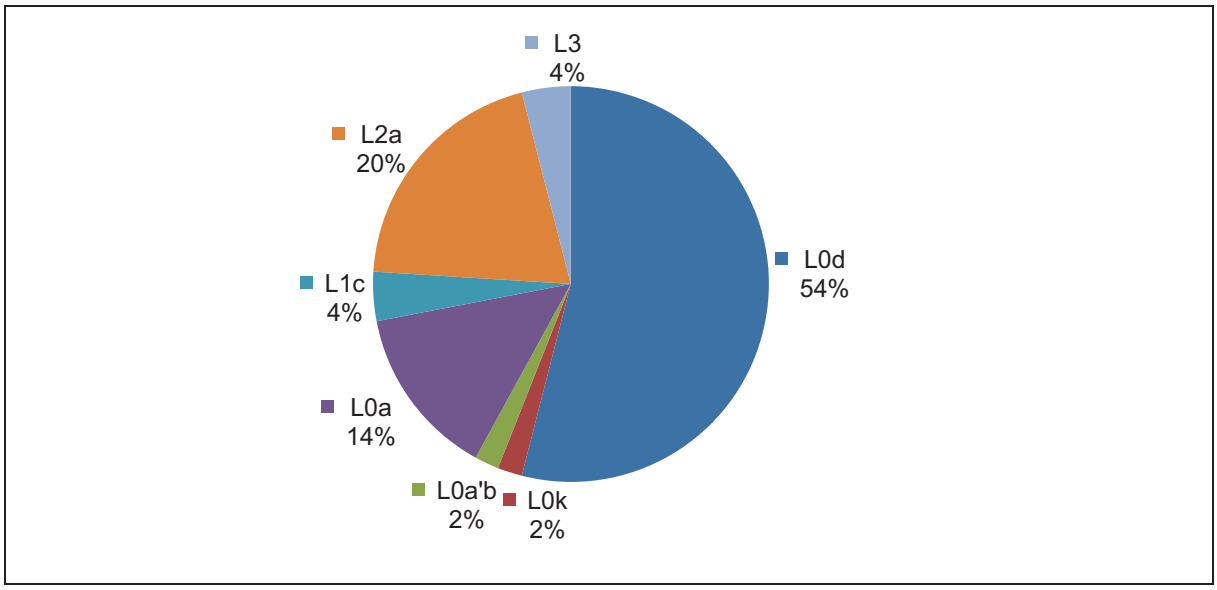
The haplogroup L novel sequence variant at np 15337 was observed in the same individuals that displayed the novel transition at np 12436, as discussed earlier in this section, and they were grouped together in the L0d1b sub-clade in all the NJ and MP phylogenetic trees of this investigation. It was concluded that this sequence variant in

conjunction with the previous novel sequence mutation would be a valid candidate sequence motif for the definition of a new haplogroup within the haplogroup L0d1b lineages.

7.2.2 Sequence variation displayed in haplogroups observed in this investigation

Populations diffuse geographically or migrate over time and the frequency pattern of haplogroups in geographical regions serves as a good signature of evolutionary history. The observed outcome of the phylogenetic trees only represents a single view of the evolutionary history of that set of data, which in actual fact constitutes thousands of possibilities. In addition, many different evolutionary truths could be provided to explain a single phylogenetic outcome. Therefore it was critical to use geographical distribution information of haplogroups and haplogroup frequency in conjunction with the phylogenetic outcomes of this investigation to determine a reasonable and valid evolutionary history for the Tswana cohort. The incorporation of this information into the interpretation of the phylogeny of a population provides a mechanism to determine the most likely evolutionary solution to sequence variation and distribution observed in a population. Therefore it is useful to draw conclusions by comparing different phylogenetic trees that present different phylogenetic relationships for the same data, as in the case of this investigation (Excoffier, 1994). The haplogroup types and haplogroup frequencies of the Tswana-speaking study group are outlined in Figure 7.1.

Figure 7.1 Pie chart distribution of haplogroups observed in the Tswana population of this investigation



Haplogroups assigned by using the PhyloTree classification system (Van Oven and Kayser, 2009). Percentages calculated for the Tswana cohort of 50 individuals only. L0a'b refers to an unresolved haplogroup that could not be classified further than the root for haplogroups L0a and L0b.

The mtDNA genetic composition of African populations consists of a diverse set of lineages that originated in east, west and west-central Africa (Salas *et al.*, 2002; Salas *et al.*, 2004; Quintana-Murci *et al.*, 2008). This investigation demonstrated that the haplogroups identified in the Tswana population under investigation, reflect a combination of regional haplogroups observed in the Bantu-speaking populations that are associated with the Bantu migration to the southeastern regions of Africa, as well as a major component (56%) belonging to haplogroups L0d and L0k, which are associated with the Khoi-San populations of South Africa, Angola and Tanzania (Watson *et al.*, 1997; Chen *et al.*, 2000; Salas *et al.*, 2002; Kivisild *et al.*, 2004; Tishkoff *et al.*, 2007; Behar *et al.*, 2008).

Seventy-two percent of the Tswana-speaking cohort belong to haplogroup L0, which is in agreement with the reportedly high frequency of the haplogroup L0 in southern Africa and eastern or southeastern Africa (Salas *et al.*, 2002; Gonder *et al.*, 2007). Seven Tswana-speaking individuals (14%) belong to haplogroup L0a, which is associated with the 9 bp deletion that has been identified as an important marker of Bantu migration to southern Africa (Soodyall *et al.*, 1996) and has been reported to be common in southeastern African Bantu speakers (Salas *et al.*, 2002). Two separate maternal lineages of haplogroup L0a have been observed within the Tswana-speaking study group and one Tswana-speaking individual of this investigation belongs to an unresolved haplogroup,

namely L0a'b, and contains alterations that differ markedly from the other L0 lineages of the Tswana-speaking individuals. These alterations, as yet, have not been confirmed to define a haplogroup by the PhyloTree classification system (Van Oven and Kayser, 2009). This mtDNA sequence could represent an undefined haplogroup that originated within the Tswana-speaking Bantu population or a rare variant within a Bantu-speaking population that reached the southern region of Africa. The presence of a rare haplogroup variant in a population is indicative of the fact that the haplogroup has been present in the population over a long period of time, therefore implying, in this instance, that gene flow took place with a Bantu-speaking population in which this haplogroup was well established or that the haplogroup has been present in the Tswana population under investigation for a long time. The gene flow with the Bantu-speaking populations was affirmed by the close phylogenetic relatedness of the respective L0a maternal lineages observed in these Tswana-speaking individuals.

The component of the Tswana-speaking cohort that belongs to haplogroup L0d (54%) represents many different lineages and it could be concluded that this Tswana study population not only displays this ancient maternal lineage at high frequencies, but it also displays a widely representative distribution of the sub-haplogroups, which indicates that the haplogroup itself has most probably been present in the population for a long time (Excoffier, 1994). Based on the principle that the type of haplogroups present in a population is an indicator of the geographical origin of the individuals of the population and that the haplogroup frequency within a population is an indication of the amount of gene flow over history (Excoffier, 1994), it could be concluded that the Tswana population under investigation originated in the southern regions of Africa and that marked gene flow took place with the Khoi-San populations of southern Africa. The timeframe in which the gene flow took place is, however, elusive. Nine different maternal lineages have been observed within the mtDNA sequences of the Tswana-speaking cohort belonging to haplogroup L0d; four of these lineages contain alterations that indicate the possibility of new haplogroups. These rare haplogroup variants, caused by more recent polymorphisms occurring in the lineage, are more closely related to the common haplogroups within the existing population than to other rare variants observed in the Global African and All African phylogenetic trees, as would be expected in a population that has evolved over time and is sufficiently large to sustain and fix these rare variants. New polymorphisms, as observed within the Tswana-speaking study group belonging to haplogroup L0d, are more likely to remain in the population of origin if the population resides in a single geographical location over a long period of time than in cases where there are high levels of genetic flow with

distant populations for some specific reasons. It can be concluded that the Tswana population under investigation harbours a large number of haplogroup L0d sub-groups, indicating a marked gene flow with a large population of Khoi-San individuals, and that it has sustained these haplogroup variants or more probably developed the rare variants over time.

The second largest haplogroup component (20%) within the Tswana cohort consists of individuals belonging to haplogroup L2a. Haplogroup L2a is regarded as the most common haplogroup present in western and southeastern Africa and was introduced to the southern regions by a major Bantu expansion that preceded the Bantu migration to these regions (Salas *et al.*, 2002; Atkinson *et al.*, 2009; Rosa and Brehm, 2011). It could be concluded that the Tswana study population originated from or underwent gene flow with the Bantu-speaking populations that reached the southern regions of Africa, which contained a high frequency of haplogroup L2a. Five different maternal L2a lineages have been observed within the mtDNA sequences of the Tswana-speaking cohort that indicates gene flow with a large population of Bantu-speaking individuals containing many of the haplogroup L2a variants. In contrast to the representation of haplogroup L0d within the Tswana study population, no rare variants or undefined variants of haplogroup L2a have been observed in the Tswana cohort, indicating a more recent evolutionary history or a more limited gene flow with the Bantu-speaking populations. The L2a lineages observed in the Tswana population are also phylogenetically closely related and therefore could be an indication of the presence of these lineages in the population over a long time.

South Africa presents a high level of ethnic diversity and it was therefore not unexpected to observe high levels of nucleotide diversity in the Tswana population under investigation. The level of nucleotide diversity is comparable to the levels of genetic diversity across the African continent owing to the distribution of several haplogroups that originated in the early stage of modern human development and migration. The presence of many ancient haplogroups that are common to the southern African Khoi-San populations in the Tswana population indicates an admixture with a population that had had a large effective population size for a considerable period of time, in which rare haplogroups diverged and were preserved (Gonder *et al.*, 2007).

7.3 MITOCHONDRIAL VARIATION IN THE TSWANA POPULATION DUE TO POPULATION BEHAVIOUR

The present-day patterns of mitochondrial variation are not only determined by the mutational events of the mitochondrial genome but are shaped by different evolutionary processes connected to the size and composition of human populations over thousands of years. These factors include demographic influences, effective population size, genetic drift, gene flow due to the migration of populations and selection of or against certain genetic variants to ensure the survival of a population in a certain set of circumstances. The genetic variation observed in current populations often comprise a complex mix of a number of these factors, which can only be estimated by making inferences through modelling assumptions about evolutionary processes against the genetic diversity of a population or group of populations.

7.3.1 Genetic diversity

Modern humans originated from Africa and have lived on the African continent continuously since 200 kya, resulting in African populations displaying the highest genetic diversity of all populations in the world (Ingman *et al.*, 2000). The African continent furthermore displays a wide range of diverse environments that vary from tropical rainforests to deserts, and it has undergone drastic changes over the course of human evolution (Balloux *et al.*, 2009). In response to the environmental variability, the evolving humans developed a wide array of different lifestyles ranging from hunter-gatherers to pastoralism (de Filippo *et al.*, 2010). The diversity of the people of Africa is further evident in the fact that nearly one third of the world's languages are spoken by the inhabitants of the African continent (Campbell and Tishkoff, 2010). Environmental or cultural changes often result in demographic population behaviours such as migration, admixture and population sub-structuring, which in congruence with molecular evolutionary processes such as selection, play a further role in shaping genetic patterns through adaptation to changing environments, diets and exposure to infectious diseases (Campbell and Tishkoff, 2010; Donnely, 2007). Extant genetic diversity patterns are therefore evidence of past population behaviours and this investigation aimed to determine the evolutionary history of the Tswana population based on these principles and by comparing the genetic diversity observed in the Tswana cohort to other reported African populations.

Genetic diversity was measured in all the datasets of the investigation by the number of segregating sites observed per dataset, the average number of nucleotide differences

(Tajima, 1983) and the nucleotide diversity according to the average number of nucleotide differences per site between sequences (Nei and Jin, 1989). The findings of this investigation confirmed that the African populations display higher levels of genetic diversity than non-Africans (Ingman *et al.*, 2000; Kivisild *et al.*, 2006; Gonder *et al.*, 2007). The levels of nucleotide diversity for all the datasets of this investigation are similar, with the All African dataset displaying the highest level of nucleotide diversity. Based on this finding it is concluded that the genetic diversity displayed by the African datasets of this investigation reflect the ancient origin of humans in Africa where they lived exclusively for an extensive period of time until the rest of the world was populated through a single population that migrated out of Africa (Cann *et al.*, 1984; Vigilant *et al.*, 1991; Ingman *et al.*, 2000).

The level of nucleotide diversity observed in the Tswana-speaking cohort is slightly lower than the nucleotide diversity displayed by the Global African, All African and Eastern African datasets of this investigation (see Appendices B, C and D) and slightly higher than the nucleotide diversity displayed by the Western and Southern African datasets (see Appendix D) of this investigation. The reasons for the observed differences in the levels of nucleotide diversity in the respective datasets could be attributed to either genetic drift or natural selection, which shaped the genetic variance of the respective populations differently. These possibilities were investigated and are discussed in Section 7.3.2 and Section 7.3.5. It could, however be concluded that the Tswana-speaking study group display high levels of nucleotide diversity, which is to be expected for a population consisting of ancient haplogroup L lineages. In addition, it is concluded that the Tswana-speaking cohort is ancient because genetic diversity increases over time as the number of mutations accumulates in the mtDNA genome and therefore it is generally accepted that the higher the level of genetic diversity in a population, the older the population is. In Section 7.4 the TMRCA was investigated to corroborate this outcome.

7.3.2 Genetic drift

The directed process of genetic change in a population through genetic drift is regarded as one of the major evolutionary forces that shape genetic diversity and is an important factor to evaluate when investigating the genetic diversity of a population (Jobling *et al.*, 2004). The sampling of alleles that are carried to the next generation is ultimately dependent on the effective sample size of the population. The determination of effective population size is, however, difficult to achieve because population generations often overlap, populations

are rarely constant in size and mating is seldom random in large populations (Jobling et al., 2004). The effective population size of human populations was determined to be large, consisting of about 15,000 individuals in African populations and 7,500 individuals in non-African populations, based on data from a 10 kb autosomal non-coding region of the human genome (Zhao *et al.*, 2006). When based on data from the X chromosome, lower estimates were made of about 2,300 to 9,000 individuals in African populations and 300 to 3,300 individuals in non-African populations (Cox *et al.*, 2008) and Ingman *et al.* (2000) determined the effective population size of humans based on the mitochondrial coding region to be 8,200 individuals. The complex evolutionary process of genetic drift was investigated in this study by modelling assumptions of population expansion and genetic structure against the genetic diversity observed in the Global African, All African, Western, Eastern and Southern African datasets, as well as the Tswana dataset of this investigation. The findings of these analyses, as discussed in Section 7.3.3 and Section 7.3.4, made it possible to interpret and make conclusions about the factors that influence genetic drift in populations. These factors include population size changes, the possibility of founder effects and population bottlenecks over time and ultimately the impact that these factors might have had on shaping the genetic diversity observed in the Tswana-speaking study group.

7.3.3 Population size and migration

The history of the demographic movement of human populations in Africa is intricate and includes several population expansions and migration events that have shaped the genetic diversity of modern-day African and global populations. The human mitochondrial genome has been exposed to the effects of evolutionary processes over hundreds of thousands of years and therefore modern-day genetic signals of population behaviours are a complex combination of different major evolutionary genetic impacts within the mtDNA of a population, as was observed in this study (Rosa and Brehm, 2011).

The detection of population growth signals obtained through Fu's F_S statistics, R_2 statistics and mismatch distributions and parameters indicate that large and sudden expansions are evident in the Global African, All African, Eastern and Western African datasets. This observation supports the previously reported signs of the Pleistocene expansion of small, isolated human populations in eastern Africa between 100 kya and 30 kya (Sherry *et al.*, 1994; Rogers, 1995). Further population expansion events in Africa included the expansion of a small human population that contained haplogroup L3 around 60 kya to

80 kya into Eurasia (Slatkin and Hudson, 1991; Harpending, 1994; Rogers, 1997; Watson *et al.*, 1997) and a further major human expansion of Bantu-speaking populations in response to favourable climatic changes around 4 kya from western Africa across central Africa and then via two different routes along the eastern and western regions of Africa to the southern regions of Africa (Pereira *et al.*, 2001; Salas *et al.*, 2002).

The complexity of the population histories of the different populations contained in the large pooled datasets, such as the Global African and All African datasets of this investigation, is evident from the presence of bimodal peaks in the mismatch distributions. It highlights the fact that the genetic signals displayed through pairwise nucleotide differences could be affected by more factors than mutation, genetic drift and population expansion, as assumed under the methods of Rogers and Harpending (1992). These factors refer to adaptation through natural selection as well as demographic and environmental changes that could have caused population bottlenecks, founder effects or fragmentation of populations into subpopulations with limited or extensive migration events having an impact on the genetic diversity (Bertorelle and Slatkin, 1995; Aris-Brosou and Excoffier, 1996). They also reflect the local differentiation of human populations within a regional area through the ethnicity of the individuals contained in the datasets of this investigation. The modern ethnic populations are much younger than the observed genetic variation within these datasets and portray ancient population histories before the establishment of the current ethnic groupings.

The signals of population expansion in the Global African and All African datasets are further ascribed to the climatic changes of the LGM 23-15 kya that had a significant impact on the size and distribution of human populations in Africa. The development of region-specific mtDNA haplogroups was caused by cycles of expansion and retraction of the equatorial forest and corresponding movement of human populations who settled in small pockets of fertile land south of the Sahelian strip. It is believed that modern humans spread from this point at a later stage. The warmer conditions ~9 kya supported further human migrations to new regions with subsequent gene flow between populations, and were accompanied by population growth. By 6 kya, agriculture was well established and this in turn supported further population growth (Clark *et al.*, 2003). Population growth was therefore driven by favourable climatic conditions, as well as the introduction of agriculture and iron-smelting techniques (Bandelt *et al.*, 2001b).

Although the Eastern African dataset displays strong evidence of past population expansion events, it does not fit the model of sudden expansion as well as in the case of the Global African and All African datasets. This has been ascribed to the composition of the Eastern African dataset, which reflects different population histories. The mismatch distribution of this dataset displays a pronounced bimodal distribution that can be explained by the high proportion of mtDNA sequences belonging to haplogroup L3 contained in this dataset, which carries signals of a major expansion event that started human globalisation about 8-12 kya (Mellars, 2006; Atkinson *et al.*, 2009). The Western African dataset displays an even lower signal of population expansion than that observed in the Eastern African dataset and reflects the impact of different population groupings on the genetic signals of the pooled dataset. The Western African dataset contains a large component of mtDNA sequences belonging to individuals of hunter-gatherer populations such as the Pygmies and reflects a stronger component of stationary or isolated population behaviour. Similar observations with regard to populations of western Africa have been made by other studies, where population expansion was determined to have taken place in clusters rather than in the total western African population at once (Graven *et al.*, 1995; Watson *et al.*, 1997).

The southern African Bantu populations were expected to display signals of population growth in response to the expansion of Bantu-speaking populations about 4 kya from an area that at present forms the border between Nigeria and Cameroon in a westerly direction along the coast and riverine sites and into an easterly direction where the agriculturalists settled at the interlacustrine region in the present-day Uganda (Beleza *et al.*, 2005). The migrations that reached the southern African regions through the western route were more gradual than the migrations along the eastern route. A second expansion event took place from the Great Lakes region to the south in two directions and it is believed that a large component of current Bantu-speaking populations in southern Africa migrated to the southern regions via this route. One migration moved along the coast and the other migration moved through the present-day eastern Zimbabwe to southern Africa (Newman and Roberts, 1995). Studies of mtDNA sequence variation supported these Bantu migration theories and more specifically implicated the L0a, L2 and L3b mtDNA lineages in the eastern stream of the Bantu migrations (Pereira *et al.*, 2001; Salas *et al.,* 2002; 2004). The presence of L0a and L2a haplogroups in the Tswana population under investigation, provides evidence that the Tswana population partly originated via an ancestral component of Bantu speakers that migrated through the eastern route to southern Africa.

In contrast to the Eastern and Western African regional datasets, the Southern African and Tswana datasets of this investigation do not display evidence of major population expansion events in past population history. The Tswana study population contains a component of the Bantu-speaking genetic pool and as with the Eastern and Western African datasets, it was expected that some evidence of population expansions within this dataset would be observed. The observed signal of a constant population was, however, not wholly unexpected and was explained by the presence of the large component of Khoi-San lineages present in the current Tswana cohort. Earlier studies reported a similar phenomenon in hunter-gatherer populations of Africa, such as the Pygmies and Khoi-San populations, in which the signals of early population expansions were erased from the population mismatch distributions because of population bottlenecks or fragmentation of populations that occurred during the period after the Neolithic period; therefore in a period after the early Pleistocene expansions (Watson *et al.*, 1997; Excoffier and Schneider, 1999). Studies further demonstrated that more recent bottlenecks could have had the same effect on the genetic signals of previous population expansions (Excoffier and Schneider, 1999).

Since the Southern African and Tswana datasets of this investigation consist of a large component of L0d and L0k haplogroups, it has been concluded that the ancient Khoi-San hunter-gatherer populations, which migrated to the southern African regions early, were strongly represented in these populations. It has also been concluded that the Tswana study population originated from the admixture of hunter-gatherer Khoi-San females in southern Africa with Bantu-speaking males that reached southern Africa via the eastern migration route, as is evident from the large component of L0d, L0a and L2a haplogroups present in the current Tswana cohort. The ragged mismatch distributions and statistical measures of constant population size observed in the Tswana population reflect a slow and constant growth pattern, suggesting that the demography of the Tswana population has been stable over a long period of time and that the population is at a mutation-drift equilibrium (Slatkin *et al.*, 1985). This is attributed to the presence of the large hunter-gatherer Khoi-San population component that underwent early population bottlenecks, which erased signs of early important population expansions. It is further concluded that the signals of past population expansions that were expected to be displayed by the Bantu-speaking ancestral component have been erased by the maintenance of constant population sizes and the presence of a strong genetic component of Khoi-San maternal ancestry. The evidence of constant population size further also indicates that no recent bottlenecks occurred within this population and it is postulated that

the Tswana study population is an example of the admixture between the Khoi-San and Bantu-speaking populations that came in contact after the migration of the Bantu speakers to the southern African regions and maintained a stable and constant-sized population structure over hundreds of years.

7.3.4 Population genetic structure

The coalescence times of the deepest roots of the human phylogenetic tree have provided evidence for the origin of modern humans in Africa and of ancient human population structures before migration to other continents of the world (Cann, 1987). Ancient population substructure is further supported by the anthropological evidence of high variability between the cranial shapes of early modern humans from Africa and the Middle East between 200 kya and 60 kya, which suggests that human populations of different geographical regions developed separately through evolutionary processes of selection, genetic drift, gene flow and population behaviours such as population expansions or contractions due to human migrations and environmental effects (Campbell and Tishkoff, 2010). Genetic differentiation between populations are most pronounced within populations that were isolated over long periods of time and experienced strong genetic drift because of small effective population sizes. In contrast to isolation, human migration has had an opposite effect, in that it resulted in gene flow between populations and therefore migrated populations display smaller genetic distances than populations that were in isolation (Nei, 1982).

Genetic variation within and between populations influences the physical attributes, susceptibility to disease and type of treatment required of individuals both collectively within and among populations (Bamshad *et al.*, 2003). Knowledge about the genetic differentiation and substructure of a population is important to determine the impact of environmental factors or disease. Physical traits, cultural practices or linguistics are often used to classify and define populations or ethnicities for these purposes without corroborating these inferences with genetic data. The classification of black residents of Africa as Bantu speakers of a certain ethnic origin based on their place of birth, language, culture or self-perception of ethnicity is often misleading, as was demonstrated by the discordant combination of ethnic characteristics displayed by the people of southern Angola in the study of Coelho *et al*. (2009). This holds true for the Bantu-speaking Tswana population of South Africa; most of its history has been based on archaeological, cultural and linguistic data and very little on genetic information. It was therefore important to

investigate the Tswana population of this study for genetic differentiation from other African populations.

Molecular perspectives on the early Bantu expansions from western Africa through central and eastern Africa to the southern regions of the continent have elicited different theories. Tishkoff *et al.* (2009) reported genetic evidence for the migration of Bantu-speaking populations through Africa to the southern regions based on polymorphic markers that reflected a genetic relationship between a range of sub-Saharan populations belonging to the Niger-Congo language groups. This interpretation was further supported by the lack of differentiation between the Y chromosome haplogroups of the western and the eastern African populations, indicating a genetic connection between the Bantu-speaking populations of these regions. This theory was also supported by the results of analyses of the HV1 and HV2 regions of mtDNA and Y chromosome regions of Bantu-speaking populations of Zambia in Central Africa where the Bantu-speaking populations possibly came together before their final dispersal to the southern regions (de Filippo *et al.*, 2010). Recent studies performed by Sikora *et al.* (2011), however, reported that the dispersal of the Bantu languages in Africa was not connected to the migration of Bantu-speaking populations but rather expanded through the assimilation of local groups who acquired the Bantu languages.

The presence of the sub-groups of the haplogroups L0a and L2a in the Tswana cohort, supports the theory that the southern migration of Bantu-speakers involved a founder effect of individuals belonging to these haplogroups and resulted in the diversification into sub-groups, as was observed in Bantu-speaking populations of southeastern Africa (Salas *et al.*, 2002). Furthermore, the results of this investigation suggest that the western and Eastern African datasets display low genetic differentiation and therefore are in agreement with the results obtained by Tishkoff *et al.* (2009) and de Filippo *et al.* (2011) that the Bantu-speaking populations are genetically connected through gene flow that took place during the migrations of the Bantu-speaking populations to the southern regions of Africa. This theory is further corroborated by the high level of differentiation displayed between the Western- and Eastern African datasets of this investigation and the Southern African dataset that contains mtDNA haplogroups that have mainly been observed in the Khoi-San-speaking populations of southern Africa. The high level of differentiation displayed between these two regional groups i.e. the western-eastern regional group and the southern regional group, is therefore in agreement with the theory of an early split and isolated development of these population groups (Behar *et al.*, 2008).

The Tswana population under investigation displays low levels of genetic differentiation from the Southern African dataset, suggesting that the gene flow between these populations must have been recent, based on the lack of large genetic distances between these two populations. The Tswana population is however, differentiated from the Southern African dataset, which implies that these populations have not wholly been assimilated into each other and that the large component of genetic haplogroups that are similar to those observed in the Khoi-San speaking individuals, have probably been acquired by the assimilation of Khoi-San females into the patrilocal Tswana tribes. Archaeological models describe three phases of interaction between pastoral Bantu-speaking populations and early hunter-gatherer populations, starting with initial limited contact, followed by settlement of the pastoral populations in the same regions as the hunter-gatherers, with an expansion of farming activities that resulted in an increase in socio-political interaction and eventual assimilation or loss of the hunter-gatherers because of limited access to resources (de Fillippo *et al.*, 2010). It is believed that this happened to the Khoi-San speaking populations that were spread widely throughout Botswana where they came into contact with the Tswana populations that migrated from the current North-West Province of South Africa to Botswana (Osaki, 2001). Archaeological evidence points towards the settlement of Bantu-speaking populations in the current Zimbabwe and Botswana as early as 190 AD, where they could have come into contact with one another and with the Khoi-San speakers that lived in those regions from 17 kya. Archaeology has further provided evidence of the presence of the Khalagari chiefdoms that formed the most western dialect group of Sotho/Tswana speakers and whose contact with the Khoi-San is evident from the cattle raising and hunting lifestyles that they led rather than farming traditions. Evidence has also been obtained for the presence of Sotho/Tswana tribes in the eastern and central regions of Botswana between 600-700 AD and 1200-1300 AD. These tribes migrated westwards into the Kalahari and eastward into the Limpopo regions of South Africa.

It is believed that the early ancestors of the Sotho/Tswana populations of southern Africa originated from eastern and central Africa and that their movement into the Hartbeespoort area of the Gauteng Province, followed by a southern migration to the Groot Marico area of the current North-West Province of South Africa, took place during the Late Iron Age in the thirteenth century (Huffman, 1989; Boeyens, 2003). The Tswana-speaking populations were driven from what was referred to as the Western Transvaal to Bechuanaland around 1853 because of inter-tribal wars and colonisation by the Boers from the Cape Colony. After they had lived with the Khoi-San speaking populations in Bechuanaland for about 20

years, some of the Tswana-speaking populations returned to the North-West Province of South Africa, where the Tswana population of this investigation currently resides (Schapera, 1946; Osaki, 2001). The genetic description of the Tswana population under investigation as a genetically differentiated population that displays signs of recent admixture with a genetic pool that is common to the Khoi-San populations of southern Africa, is therefore in agreement with the history of the Tswana populations of southern Africa as described by the archaeological and cultural evidence discussed here.

Some of the Khoe-Kwadi language family of the Khoi-San populations of southern Africa resided in Botswana and were classified by Barnard (1992) into four divisions of Khoe-speaking Bushmen: the Western Khoe Bushmen that were in close contact with the !Kung, the Central Khoe Bushmen that included the G/wi and G//ana of the Kalahari Game Reserve of South Africa, the Northern Khoe Bushmen that resided in the Okavango and the Eastern Khoe Bushmen that resided in the eastern regions of Botswana. The last-named group linguistically constituted the Shua and the Tshwa language groups and have been greatly influenced by the Tswana culture. It has also been reported that large numbers of G//ana relocated to the eastern regions of Botswana where they lived in close contact with the Tswana people (Barnard, 1992). The Khoi-San speaking populations reportedly had diverse relationships with the Tswana-speaking populations and were sometimes traded with as equal partners and sometimes treated as low-class individuals or even enslaved. Most important, however, is the fact that the Khoi-San speakers seldom embraced the traditions and practices of other populations and that although they intermarried with the Bantu-speaking populations, as is evident from the genetic lineages shared with the Tswana study group, they seldom became fully absorbed into those populations, as observed in the genetically differentiated Southern African and Tswana populations of this investigation. This interpretation of the results was done with caution, however, since the Southern African dataset of this investigation consists of San and !Kung individuals from South Africa (Gonder *et al.*, 2007), !Kung from Namibia (Koekemoer, 2010), !Kung from Angola (Koekemoer, 2010) and a few Bantu-speaking individuals, as described in Appendix B. The dataset is not representative of Khoi-San speakers of the Botswana regions, as discussed in this section, and the genetic distance of the Tswana-speaking populations of this investigation from the Khoi-San speaking populations that originated from Botswana might be different and needs to be investigated further.

7.3.5 Selection

Interest in the presence of natural selection in the mtDNA genome has been important in the field of population genetics. Selective neutrality is commonly assumed in studies of the evolutionary history of human populations and has an impact on the interpretation of population genetic signals with regard to demographic events and lineage (Ptak and Przeworski, 2002; Hammer *et al.*, 2004; Kivisild *et al.*, 2004).

The fact that natural selection has shaped the genetic diversity within the human mitochondrial genome has been widely reported (Cann *et al.*, 1984; Nachman *et al.*, 1994; Ingman *et al.*, 2000; Mishmar *et al.*, 2003; Elson *et al.*, 2004; Ruiz-Pesini *et al.*, 2004; Ingman and Gyllensten, 2007). The way in which selection has shaped the mitochondrial genome, however, has been disputed. Two theories are generally considered for the explanation of selection in human mtDNA. Some studies report that the genetic diversity of the mtDNA has been created and maintained by adaptive selection that enabled populations within certain climatic regions to adapt and survive in colder climates (Mishmar *et al.*, 2003; Ruiz-Pesini *et al.*, 2004). Other studies have refuted this theory and suggest that the current mtDNA diversity is a function of purifying selection, which removed deleterious NS substitutions from the mitochondrial genome in combination with genetic drift (Elson *et al.*, 2004; Kivisild *et al.*, 2006).

In this study, deviation from neutrality was displayed by the Global African, All African and Eastern African datasets, as well as within all the major L haplogroups of the All African dataset. The strong signals of Tajima's *D* and Fu and Li's D* and F* test statistics displayed by these datasets indicate that the mtDNA contains high numbers of rare sequence variants and singletons, which would be expected in populations that had been under selective pressure, had undergone population growth or displayed population substructure or division. The Global African and All African datasets display strong signals of past population growth, as discussed in Section 7.3.3, and the possibility that the *D*, D* and F* values were influenced by the population expansion event had to be considered. The opposite was also true in that the signals of demographic expansion could have been caused by pronounced selective sweeps. The results of the investigation of the NS/S private and haplogroup-associated substitutions within the All African dataset of this investigation suggest evidence of purifying selection within the mtDNA coding regions of this population and therefore reject the possibility of a positive selective sweep. Based on this evidence, it has been concluded that the Global African, All African and Eastern

African datasets do display evidence of a major population expansion event and it is further postulated that the mitochondrial genomes of individuals of African origin display evidence of weak purifying selection that is slow in the removal of NS substitutions, resulting in genomes that contain high numbers of rare variants. Previous bottlenecks would have enhanced this effect of rare variants and could therefore not be ruled out.

The negative signals of Tajima's *D* and Fu and Li's D* and F* test statistics of the major L haplogroups suggest that the regional datasets of this investigation should display similar results, since they consist of individuals belonging to the L haplogroups. This is, however, not the case. This discrepancy could be explained by the possible presence of population substructure within the Western, Southern African and Tswana datasets, which has been reported to elevate *D*, D* and F* statistics. This phenomenon is caused by the higher levels of average pairwise nucleotide differences between pairs of sequences in subdivided and structured populations (Simonsen *et al.*, 1995). The presence of population substructure, especially in the deeply rooted haplogroup L0 populations, has been reported in other studies and is therefore a likely possibility to explain the results observed in this investigation (Tishkoff *et al.*, 2009).

In contrast to the significantly negative Tajima's *D* and Fu and Li's D* and F* test statistics displayed by the major L haplogroups of this investigation, haplogroups L0, L2 and L3 do not display significant neutral index values when the NS/S ratios of the private substitutions and the haplogroup-associated substitutions are considered. A significant NI value is, however, displayed by haplogroup L1. All of the L haplogroups display NI values greater than one, which are interpreted as evidence of signs of weak purifying selection, although neutrality could not be rejected. The non-significant NI values of L haplogroups are ascribed to the smaller sizes of the datasets, since, when the datasets are pooled, the All African dataset displays a significant NI value greater than one, thus confirming the presence of weak purifying selection within the mitochondrial genomes of the African individuals of this investigation.

The individual 13 protein-coding genes of the mtDNA display different directions and strengths of selection throughout the genome. It has been demonstrated that selection modified the substitution patterns of nearly all of the individual 13 protein-coding genes of the mtDNA in varied ways. The *ND2*, *ND3* and *CytB* genes demonstrate signs of positive selection although these results are not statistically significant. Neutrality has been rejected for the *ATP8*, *COII*, *ND4* and *ND6* genes and negative selection is thus assumed,

whereas the *ND1*, *COI*, *ATP6*, *COIII* and *ND4L* genes also present with evidence of negative selection, but this is not statistically significant and neutrality has therefore not been rejected. The *ND5* gene displayes absolute neutrality. Some genes are therefore more vulnerable than others to selection. Differentiating between the different types of selection according to substitution patterns is risky because of the complex manner in which the effects of selection are displayed in the genome. Negative selection could, for instance, also be interpreted as the relaxation of selective constraints and positive selection might be interpreted as the relaxation of negative selection. Studies focusing on the different regions of the genes of the human mtDNA molecule provide evidence that the functional domains of the genes are much more conserved than the remaining part of the gene and that the study of the NS to S substitution patterns of the whole gene complement could therefore present misleading signals of selection (Kivisild *et al.*, 2006). It is suggested that such studies should be undertaken into the genes of the human mtDNA of representative samples of populations that reside in different geographical regions to truly determine the effects of selective forces on the substitution patterns displayed by these samples. Overall however it is concluded that most of the individual genes of the mitochondrial genome display signs of weak purifying selection and therefore support the theory of weak purifying selection based on the results of the larger All African dataset.

The Tswana population under investigation consists of haplogroups that display deviation from neutrality and non-significant signs of purifying selection over the whole genome, as well as non-significant signs of selection of different directions and strengths within the individual genes of the mitochondrial genome. Deviation from neutrality is not indicated by the negative Tajima's *D* and Fu and Li's D* and F* test statistics on analysis of the Tswana dataset, which is ascribed to population substructure. The larger African dataset, which includes the Tswana population, displays significant signs of deviation from neutrality and purifying selection. Significant signs of purifying selection have been determined for the *ATP8*, *COII*, *ND4* and *ND6* genes of the mitochondrial genome. Based on the overall significant signs of deviation from neutrality and the presence of purifying selection within the large All African dataset of this investigation, it has been concluded that the Tswana population under investigation is also subject to these selective forces and that the discrepancies between the test results for deviation from neutrality are ascribed to the smaller size of the Tswana population under investigation and possible substructure within this population.

7.4 COALESCENCE-TIME ESTIMATIONS

Demography is a critical aspect to be taken into account when making conclusions about coalescence-time estimates of lineages observed within a population. It has been proven that factors such as population bottlenecks, founder effects and changes in population sizes, which would affect the amount of genetic variability within a population as well as the effect of genetic drift, contribute to the uncertainty of molecular dates obtained by using a model-free lineage-based approach to determine the coalescence-time estimates (Cox, 2008). Although the factors that could contribute to inaccurate date estimates, such as mutation rates, the validity of using a linear molecular-clock, the uniformity of mutation rates, the risks posed by the effects of homoplasy and mutational hotspots in the mtDNA, the choice of mtDNA region for analysis and the effects of selection, have been addressed by the methodology used in this investigation, the time estimates must still be interpreted in the context of the differences between the nature of the samples, the sample sizes used and the methods employed for determination of coalescence-time estimates between the different studies. The standard deviation of the coalescence-time estimates is broad because of these last-named factors and the broad time frames are used to conclude the meaning of the genetic dates in the context of other reported evidence, with the aim of identifying the prehistory of the Tswana population of this investigation.

The coalescence-times of the L0 haplogroup of the All African MP tree of this investigation reaffirm earlier findings that the L0 haplogroup was the first emerging subset of maternal variation displayed by ancient African populations, which diverged from the haplogroup L1'5 sister clade between 100,000 ybp and 150,000 ybp (Salas *et al.*, 2002; Behar *et al.*, 2008; Rosa and Brehm, 2011). Based on human fossil remains discovered in southern Africa at the Border Cave and Klasies River Mouth Caves, that indicated human population behaviour at about the same time, it is believed that an early moderate expansion of human populations took place (Watson *et al.*, 1997). Further human migrations and isolated settlements of early humans have been indicated by paleoclimatology and archaeological evidence occurring over the following few thousand years, most probably due to the improvement in the climate during the Marine Isotope Stage 5 (Forster, 2004; Henshilwood *et al.*, 2002; Endicott *et al.*, 2009).

Fifty-four percent of the Tswana-speaking population of this investigation belong to the L0d lineage of the L0 haplogroup, which diverged from the L0a'b'f'k ancestral group early in accordance with the coalescence-time estimates published by other studies (Salas *et al.*,

2002; Behar *et al*., 2008). The coalescence-time estimates therefore confirm that the Tswana-speaking individuals of this investigation harbour lineages that split from the L0a'b'f''k ancestor early in the prehistory of modern humans in Africa and the population is believed to have developed in isolation for between 50,000 ybp and 100,000 ybp before it was admixed with other lineages (Gonder *et al.*, 2007; Tishkoff *et al.*, 2007; Behar *et al.*, 2008). The coalescence-time estimates for the L0d lineages of this investigation place the maternal ancestors of the Tswana-speaking individuals of this investigation in eastern Africa where artefacts of modern humans have been dated between 100,000 ybp and 40,000 ybp (Gonder *et al.*, 2007; Behar *et al.*, 2008). The Middle Stone Age stretched over the period between 250,000 ybp and 40,000 ybp, in which archaeological evidence indicate a change in culture and subsistence styles that were most probably due to population expansions that took place and subsequent migration of populations into new territories, as would have been the case for the ancestors of the L0d lineages (Henshilwood *et al.*, 2002). The sublineages L0d1a, L0d1b, L0d1c, L0d2a and L0d3 are present in the Tswana-speaking individuals of this investigation, of which the ancestral lineages L0d1, L0d2 and L0d3 coalesced during the time frame of 74,158 ybp and 36,603 ybp. Archaeological evidence, including weapons used for hunting, evidence of utilisation of plant material, marine exploitation and the use of red ochre, stone and shell for art and ornaments, which dates to a period 75,000 to 70,000 ybp in eastern, central and southern Africa, suggests that the ancestral populations of the L0d lineages expanded further into Africa (Henshilwood *et al.,* 2002; Mellars, 2006). It can be concluded that the maternal ancestors of the Tswana-speaking individuals of this investigation diverged into the L0d1, L0d2 and L0d3 lineages during the population expansion of the late Middle Stone Age.

The sublineages L0d1b and L0d2a of the All African MP tree in this investigation display coalescence-time estimates between 64,224 ybp and 26,204 ybp, which indicates that these two sublineages developed earlier than the others and in the late Middle Stone Age period. The most recent of the sublineages are L0d1a and L0d1c, which date between 39,562 ybp and 10,789 ybp, thus indicating a later development of these lineages, more towards the Neolithic period, which started at about 10,000 ybp. This phase was characterised by the development of agricultural practices and the domestication of animals, which coincided with further population expansion and migration between populations, mainly in the western and central regions of Africa (Scheinfeldt *et al.*, 2010). In the southern regions of Africa however it is believed that the carriers of the L0d lineages were localised hunter-gatherer populations that lived in isolation and demonstrated slow population growth, resulting in constant population size (Atkinson *et al.*, 2009). Nearly 70%

of the current Khoi-San speaking individuals of southern Africa, which here refer to different Khoi and San groups within the click-speaking populations that reside in different regions of southern Africa, harbour either the L0d or L0k lineages that are not displayed by populations in other regions of Africa except in Tanzania in eastern Africa (Salas *et al.*, 2002; Behar *et al.*, 2008). Based on the coalescence-time estimates for these lineages in this investigation it is concluded that the maternal ancestral L0d lineages observed in the Tswana-speaking individuals of this investigation most probably came from the early Khoi-San speaking populations that resided in southern Africa from the late Middle Stone Age.

The L0k lineage is present in one of the Tswana-speaking individuals of this investigation and displays a coalescence-time between 36,993 ybp and 14,387 ybp and divergence from the L0a'b'f ancestor in the All African MP tree between 86,831 ybp and 49,839 ybp. The coalescence-times of these lineages therefore resemble the ancient nature of the L0d lineages and reportedly are only present in Khoi-San speaking populations (Soodyall *et al.*, 2008; Behar *et al.*, 2008). This evidence further supports the conclusion drawn about the admixture between the Tswana-speaking individuals of this investigation and the Khoi-San speaking individuals of southern Africa.

Whether the L0a lineages originated in eastern Africa because of the split between the L0a'b'f, L0d and L0k lineages (Gonder *et al.*, 2007; Tishkoff *et al.*, 2007) or because of a back migration from southern Africa towards eastern Africa where they co-evolved with the L1'5 lineages and dispersed to the west and central regions of Africa (Behar *et al.*, 2008) could not be determined from the results of this investigation. The coalescence-time estimates of the L0a1 and L0a2 lineages present in the Tswana-speaking individuals of this investigation and in the All African MP tree demonstrate that they evolved early between 48,296 ybp and 14,387 ybp. The divergence from the ancestral lineage, L0a'b'f, therefore occurred later than the development of the L0d lineage, which is in agreement with the theory that the L0a lineages developed later and dispersed to western and central Africa with the L1'5 lineage range. The L0a1b lineage observed in the Tswana-speaking individuals of this investigation displayed an earlier coalescence-time (31,855 ybp to 11,303 ybp) than the L0a2a lineage (20,038 ybp to 4,624 ybp), which was also observed in the Tswana-speaking individuals of this investigation. It can therefore be concluded that the Tswana-speaking population of this investigation contain maternal ancestors that originated in the late Middle Stone Age to the Neolithic period in eastern and central Africa respectively, probably with the migration to the western regions of Africa with the L1'5

lineages. Climatic evidence indicates that the tropical forests of central Africa were probably much like the current tropical forest area at about 10,000 ybp and were therefore inhabited by early human populations during that time when the L0a2 lineage developed (Adams and Mortimore, 1997). Evidence of an arid time at about 3,000 ybp in that region provides reason to believe that the climatic conditions motivated the carriers of the L0a2 lineages to migrate towards the central region of Africa where they possibly joined the other Bantu migrations towards the southern region of Africa (Salas *et al.*, 2002; Behar *et al.*, 2008; Soares *et al.*, 2009). The coalescence-time of 20,038 ybp to 4,624 ybp for the L0a2a lineage supports this theory and it was concluded that the L0a1b and L0a2a lineages of the Tswana-speaking individuals of this investigation , which diverged and developed in eastern and central Africa, were both lineages that formed part of the Bantu migration towards the southern regions of Africa about 4,000 ybp.

The fact that only two Tswana-speaking individuals of this investigation belong to the L1 haplogroup suggests that the gene flow with populations that harboured haplogroup L1 was weak or that the lineages disappeared from the Tswana-speaking individuals of this investigation over time owing to genetic drift or population bottlenecks. The presence however, of the L1c2 lineage in this population, provides evidence that this lineage diverged from the L1c ancestor between 128,963 ybp and 82,721 ybp and developed at a coalescence-time of between 69,876 ybp and 36,994 ybp, which is in agreement with the evidence that the L1c lineages originated from a central African ancestral population not later than 70,000 ybp (Salas *et al.*, 2002; Batini *et al.*, 2007; Quintana-Murci *et al.*, 2008).

The second largest haplogroup component of the Tswana-speaking individuals of this investigation belong to haplogroup L2a, which indicates that there was strong gene flow with maternal ancestors belonging to these lineages. The L2 haplogroup lineages comprise about 70% of the sub-Saharan maternal genetic pool (Chen *et al.*, 2000; Torroni *et al.,* 2001). Haplogroup L2a is widespread in Africa and displays comparable coalescence-times of between 55,000 ybp and 45,000 ybp in eastern and western Africa (Salas *et al.*, 2002; Kivisild *et al.*, 2004; Behar *et al.*, 2008). The L2a lineages of the All African MP tree of this investigation demonstrate a coalescence-time estimate of between 57,307 ybp and 35, 353 ybp, which is similar to that reported and it could be concluded that the Tswana-speaking individuals of this investigation contain lineages that coalesced during the late Middle Stone Age. It has been hypothesised that the L2a lineages originated in central Africa from where they spread to the eastern and western regions of Africa along the Sahel corridor during the LGM (Salas *et al.*, 2002). It was

further hypothesised that this period was also a time of pronounced expansion of the L2 lineages into sublineages, which started between 20,000 ybp and 12,000 ybp. The coalescence-times of the L2a1 lineage and the L2a1c and L2a1d sublineages demonstrate a lineage expansion between 40,589 ybp and 12,332 ybp, which indicates that the expansion of these lineages took place earlier than the Bantu dispersal 4,000 ybp to the southern regions of Africa. This is in agreement with the theories that postulate major population expansions of individuals belonging to the L2 haplogroup in central Africa during the LGM. The arid conditions of the enlarged desert area and the conversion of the rain forests to open savannah were supposedly factors that contributed to the migration and expansion of populations in that time (Atkinson *et al.*, 2009) and it is concluded that the L2a1 lineage as well as the L2a1c and L2a1d lineages of the Tswana-speaking individuals of this investigation most probably developed during that phase. This was followed by the Bantu migration to the south at about 4,000 ybp in which many of the L2a lineages were carried to the southern regions of Africa, where the Bantu speakers eventually settled (Atkinson *et al.,* 2009). The coalescence-times of the other L2a1 sublineages of the Tswana-speaking individuals of this investigation, namely L2a1a, L2a1b and L2a1f, are more recent (19,524 ybp to 2,055 ybp) and indicate that the development of these sublineages most probably happened closer to the Bantu migration.

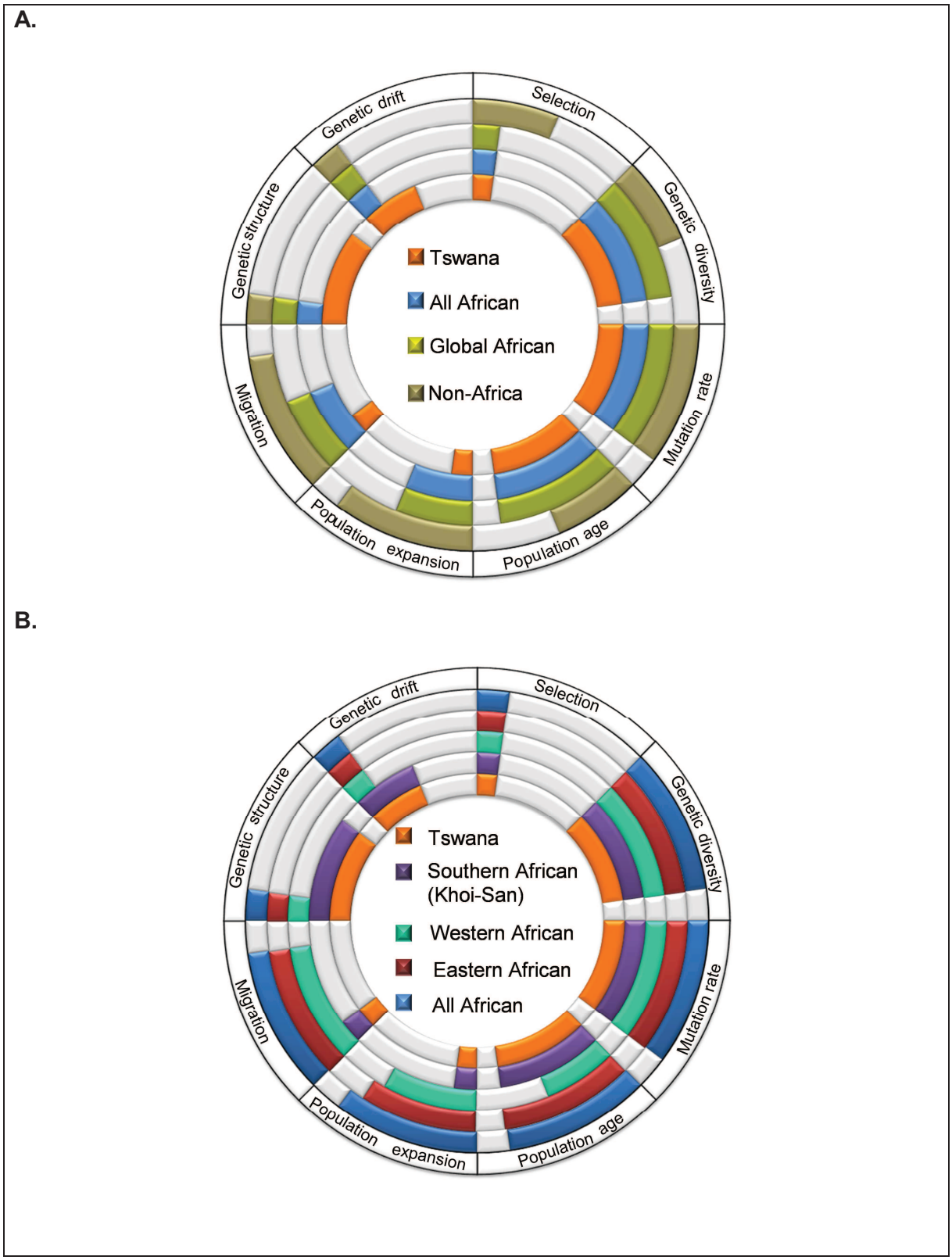
The presence of haplogroup L3 is low in the Tswana-speaking individuals of this investigation, with only two individuals belonging to the L3d1a and L3e1 lineages respectively. This is ascribed to the effects of weak gene flow between the carriers of haplogroup L3 and the Tswana-speaking individuals of this investigation or genetic drift occurring in a Tswana-speaking population with a small effective size possibly because of a bottleneck or founder effect. The fact that the Tswana population under investigation did at some point in their prehistory have contact with Bantu-speaking carriers of these lineages is certain. It is generally believed that haplogroup L3 had its origin in east Africa between 75,000 ybp and 60,000 ybp (Salas *et al.*, 2002; Kivisild *et al.,* 2006; Torroni *et al.*, 2006), and there is evidence of a westward migration towards the western regions of Africa (Watson *et al.*, 1997; Salas *et al.*, 2002). Based on the evidence that the L3d lineages have been commonly observed in the western areas of southern Africa (Cerný *et al.*, 2006; Coelho *et al.,* 2009), it is concluded that the L3d lineages of the All African MP tree of this investigation developed between 63,194 ybp and 32,369 ybp in a western region of southern Africa. The L3e lineages are believed to have arisen in central Africa at about 50,000 ybp to 40,000 ybp (Salas *et al.*, 2002; Torroni *et al.*, 2006; Behar *et al.*, 2008), which is in agreement with the coalescence-time of the L3e lineage of the

All African MP tree of this investigation (66,279 ybp to 34,425 ybp). Based on geographic evidence and the coalescence-time estimate for the L3e1 lineage of this investigation, it has been concluded that the Tswana population of this investigation harbours the L3e1 lineage, which developed in central Africa between 24,148 ybp and 6,680 ybp and became frequent among the Bantu speakers that migrated to the southern areas of Africa via the eastern route.

7.5 MODEL OF GENETIC VARIATION

A visual model is presented in Figure 7.2 that accounts for the empirical observations made in this investigation and provides an overall conceptual framework in which the impact of various evolutionary factors on the patterns of the observed genetic variation of the Tswana population under investigation are juxtaposed against the genetic variation observed in the Global African, All African and regional African datasets of this investigation. It demonstrates the novel contribution of the Tswana-speaking population of this investigation to an understanding of the genetic diversity of the Bantu-speaking populations of South Africa in the context of their evolutionary past.

Figure 7.2 Model of genetic variation



A model of genetic variance, denoting the interplay between the evolutionary factors that shape genetic diversity within human populations, is presented. The evolutionary factors that determine genetic variance are indicated by the segments of the model as indicated on the outside of the circles. Model A draws a comparison between the Tswana-speaking individuals of this investigation and broader African populations i.e. Global African and All African populations of this investigation and non-African populations as reported in literature. Model B draws a comparison between the Tswana-speaking individuals of this investigation and the regional African populations as investigated in this study and reported in literature. Each of the populations is denoted in a specific colour as depicted in the figure legends within each of the models. The size of the bar for each population column indicates the size of the variable for that population as displayed by the proportion of the total length of the column.

The model represents the observed genetic diversity within the populations and the primary evolutionary forces that influenced and shaped the genetic diversity over time to what is observed in extant human populations. The model demonstrates how the observed genetic diversity is connected to the mutation rate and the age of populations. Mutation is the primary generator of genetic variance and takes place at a high rate in the human mitochondrial genome (Jobling *et al.*, 2004). Based on the assumption that the mitochondrial genomes in all populations were subjected to the same mutation rate, genetic diversity increased over time and older populations therefore display higher levels of genetic diversity. The Tswana study cohort demonstrates levels of genetic diversity that are in agreement with those observed for the Global African and All African datasets, as opposed to the lower levels of genetic diversity reported for non-Africans, which originated at a later stage of human evolutionary history when the migration of the first human populations out of Africa is believed to have occurred (Salas *et al*., 2002). The Tswana-speaking study group further contains lineages that are ancient and are believed to have separated from the initial L0 haplogroup lineages early in the history of modern humans. The regional African datasets further support the theory that the current African gene pool resulted from an early expansion of human populations out of their eastern African homeland to most of the African continent with a wave of L2 and L3 haplogroups that developed later, as presented in the western African population in this model (Behar *et al.*, 2008).

The generation of genetic diversity through mutation over time is not a simple process that occurs in isolation. Genetic variations are observed at different frequencies between generations, both within populations and between populations, owing to major evolutionary forces that eliminate and shape genetic diversity because of the finite number of individuals of a population that participate in the formation of the next generation and the probability of survival of individuals of a population based on the forces of natural selection. The model of genetic variance demonstrates how genetic drift contributed to the genetic variance observed in the Tswana cohort in comparison to the effects of genetic drift in the African datasets included in this investigation, as well as in non-African populations.

The variance of allele frequencies between generations in populations is determined by a stochastic process of gamete sampling, which is mostly determined by the effective size of the population. The assumptions of the Wright-Fisher model for effective population size is rejected under real conditions because of factors such as the overlap between

generations, the fact that populations are rarely constant in size and mating is seldom random, leading to genetic substructure within normal extant human populations. Investigation of population size changes and genetic structuring within populations provides information that is critical to understanding the impact of genetic drift on a population. The complex interplay of these factors has been demonstrated in this investigation and is presented in the model of genetic variation. The observed genetic evidence for population expansion observed in the African datasets of this investigation and reported for non-African populations, increases the genetic variance within these populations and in conjunction with high levels of migration and low levels of genetic structure, the effects of genetic drift in these population groupings were expected to be low. The Tswana population of this investigation, however, displayed the opposite. Genetic evidence of constant population size reported low levels of migration and evidence of high levels of genetic structure suggested that the Tswana-speaking individuals of this investigation experienced higher levels of genetic drift and underwent evolutionary changes that were more similar to what was observed for the Khoi-San speaking populations of southern Africa than those observed for the regional and global African datasets.

Genetic diversity was further manipulated by natural selection, as presented in the model of genetic variance. Studies of the mitochondrial coding region have suggested that natural selection plays a fundamental role in shaping the genetic diversity of human populations and that the maintenance of sequence variance is partly due to the impact of natural selection on the human mitochondrial genome, which according to the neutralist theory, primarily removes severely deleterious mutations from the population (Kimura, 1969; Mishmar *et al.*, 2003; Elson *et al.*, 2004; Kivisild *et al.*, 2006). The frequency spectrum of the synonymous and nonsynonymous mutations observed in the datasets of this investigation clearly demonstrates that a certain proportion of the mutations have been affected by purifying selection. The findings further demonstrate that some of the protein-coding genes of the mitochondrial genome are under positive selective pressure, even if only slightly. The pervasiveness of evidence of positive selection in human mitochondrial genomes due to the confounding impact of demographic factors such as population expansion has made it difficult to identify positive selection (Nielsen, 2005). More studies are presenting evidence of the presence of positive selection, especially in the mitochondrial DNA of non-Africans, which has been ascribed to the adaptation of these individuals to colder climates as they migrated to the northern continents (Mishmar *et al.*, 2003). Current versions of the neutral theory now allow for the consideration of the

presence of negative selection and some positive selection and have been described as such in the model of genetic variation. Weak purifying selection has been reported for African populations and is also present in the Tswana-speaking cohort. The non-Africans are presented as being under additional pressure of adaptive selection.

7.6 IMPLICATIONS OF THE MITOCHONDRIAL DNA CONSENSUS SEQUENCE OF THE TSWANA POPULATION

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 benchmark for the sequence variance that is present in a Tswana-speaking population of South African origin and it is a representation of the genetic diversity of the maternal ancestral genetic pool of a Bantu-speaking population of South Africa.

The Tswana consensus sequence and the sequence variants that were observed to differ from the rCRS, are presented in Appendix G and Appendix H. The Tswana consensus sequence contributes greatly to providing a glimpse of the maternal ancestry of a Bantu-speaking population of southern Africa as represented by the accumulation of sequence variants over a long evolutionary history of migration and demographic changes. When compared with the rCRS, it provides evidence of the fundamental genetic differences that exist between populations in different geographical regions of the world and emphasises the importance of making these distinctions in order to understand the evolutionary histories of different global populations and provide a reference for applied sciences such as disease association studies and forensic human identification. For these reasons, consensus sequences should be made publicly available on a platform such as MITOMAP in conjunction with the current reference sequences to provide researchers with access to the important features of sequence variations across different global populations.

The construction of the consensus sequence was based on the full mtDNA sequences of a cohort of 50 Tswana-speaking individuals residing in the Ikageng and Sonderwater urban areas located near the town of Potchefstroom and in the rural areas of Ganyesa and Tklagameng in the North-West province in South Africa and was therefore a representation of the sequence variants expected to be observed in these populations. Investigation of a larger cohort would contribute more sequence variants that may have

been present at low frequencies in this study but upon investigation of a larger number of Tswana-speaking individuals, may be displayed at higher frequencies. Furthermore, this consensus sequence does not necessarily represent the major sequence variants that will be displayed by other Bantu-speaking populations of South Africa. The availability of full genome mtDNA sequence data for other Bantu-speaking populations of South Africa is limited, which makes it impossible to know the maternal genetic divergence between these populations and therefore difficult to predict the degree of sequence variation similarities that could be expected between these populations. Based on Y chromosome and autosomal sequence data from different Bantu-speaking groups of South Africa, the genetic divergence is low (Lane *et al.*, 2002). Whether these findings would be in agreement with the outcomes of a study on the maternal genetic contributions of the respective South African Bantu-speaking populations is, however, an open question.

The Tswana consensus sequence contains the major sequence variants that are present in the Tswana-speaking individuals of this investigation and by implication also represents the haplogroup-defining mutations that distinguish it from the mtDNA sequences of individuals from other geographical origins. The Tswana consensus sequence will therefore contribute to the identification of rare and novel sequence variants that are not associated with haplogroups that are common to this population. As discussed further in Section 7.7.1, the identification of disease-associated mutations involves differentiation between haplogroup-associated mutations, rare or private mutations (Zaragoza *et al.,* 2011) and the consensus sequence would therefore contribute to making this possible for the Tswana-speaking cohort of this investigation.

7.7 IMPLICATIONS OF MITOCHONDRIAL DIVERSITY OF THE TSWANA POPULATION FOR FUTURE STUDIES

Human genetic history and genetic variance not only hold information about the past evolutionary history of humankind, but have developed a diverse interface with many of the different fields of applied science. The increase in size and diversity of genetic datasets has proven to be invaluable in contributing information to medical genetics and disease and the developing field of forensic human identification. The investigation of the genetic variation and diversity of the complete mitochondrial genomes of the Tswana population under investigation provided a rich source of novel genetic data that was not only valuable in terms of providing information about the past evolutionary history of the

Tswana-speaking individuals of this investigation, but could also contribute to the fields of applied science.

7.7.1 Mitochondrial disease

The observation that disease-predisposing polymorphisms are geographically restricted makes the study of genetic diversity within different ethnic populations of critical importance to the identification and management of diseases. To make appropriate health care available, it is necessary to know the distribution of inter-population genetic variation that underlies susceptibility to disease. A recent study by Zaragoza *et al.* (2011) highlighted the importance of identifying rare and novel mutations for investigation of pathogenicity by association of those mutations with haplogroups and therefore identifying the mutations within the framework of a population.

The findings of this investigation identify eight sequence variants in the Tswana-speaking individuals of this investigation that have been associated in the literature with LVNC, one sequence variant with cyclic vomiting syndrome with migraine, six sequence variants with LHON disease, six sequence variants with prostate cancer, one sequence variant with Parkinson's disease, breast cancer and longevity and one sequence variant with maternally inherited cardiomyopathy. Although none of these sequence variants are novel, most of them are rare, each only occurring in a single Tswana-speaking individual of this investigation and therefore most likely not haplogroup-associated. The exception is displayed in two sequence variants that have been associated with LVNC disease, which have been observed in five Tswana-speaking individuals of this investigation that all belong to the same haplogroup L0a clade in the phylogenetic trees of this investigation. This finding therefore strongly suggests that these sequence variants are haplogroup-associated and not rare in the Tswana-speaking population of South Africa. The LHON-associated mutations observed in this investigation, are only associated with the disease in specific non-African ethnic groups and specific non-African haplogroups, which highlights the fact that disease association presents with ethnic-specific profiles and this therefore emphasises that the presence of these mutations in the Tswana-speaking cohort should be investigated further by the inclusion of multigenerational family histories, haplogroup associations and secondary genetic mutations specific to African populations (Bosley and Abu-Amero, 2010) to determine the relevance of the mutations to disease aetiology in this population. Although all six of the reported prostate cancer associated haplogroup L mutations have been observed in the Tswana-speaking individuals of this

investigation, the individuals display different combinations of the mutations and not more than three of these mutations have been observed in a single individual at a time. Therefore these alterations require more investigation in terms of the aetiology of the disease.

On establishment of the rarity of mutations in a population by the population frequency of the mutation, based on reliable databases such as the updated human sequence database constructed by Pereira *et al*. (2009), rare mutations can be regarded as a candidate for further investigation based on the gene affected, the codon position of the sequence variant, the altered amino acid if the sequence variant is located within a protein-coding region, the conservation index of the altered amino acid and the presence of heteroplasmy. The findings of this investigation therefore contribute to identifying the presence of rare disease-associated mutations in the Tswana-speaking study population. The relatively small sample size of the Tswana-speaking cohort may, however, cast doubt on the true frequencies of the respective disease-associated mutations. It is suggested that the population frequency and haplogroup association of the observed sequence variants of this study be investigated in a larger population of Tswana-speakers from South Africa. Most importantly, the sequence variation data presented in this study for the Tswana-speaking population of South Africa will emphasise the need for large arrays of mtDNA sequence data from around the world for the purpose of the identification of rare and novel sequence variants that could be investigated for pathogenicity in the context of specific populations. It is only through the establishment of large sets of full sequence variations associated with the different haplogroup lineages of the world that it would be possible to identify the rare and novel haplogroup-associated sequence variants in specific populations for further investigation of disease association (Jorde *et al.*, 2001).

7.7.2 Contribution of genetic variation data of the South African Bantu speakers to the global human phylogeny

The publicly accessible global human phylogenetic tree, PhyloTree (Van Oven and Kayser, 2009), which consists of more than 5,000 mtDNA coding region sequences, is widely used in human evolutionary studies for haplogroup assignment and is regularly updated with newly published coding region mtDNA sequence variants. Only about 30 mtDNA coding region sequences of South African Bantu speakers have, however, been used in the construction of this haplogroup classification system. Based on the approach used in the construction of the PhyloTree classification system (Van Oven and Kayser,

2009), which ensured the broad inclusion of all published mtDNA coding region sequences, it could be determined that there were most probably no large datasets of mtDNA coding region sequences of Bantu-speaking populations of South Africa available for use in the construction of the PhyloTree classification system (Van Oven and Kayser, 2009). Secondly, it is evident that the haplogroups of the South African Bantu speakers are not fully represented in the PhyloTree classification system and it is therefore not unexpected that two sequence variants have been observed in the mtDNA sequences of the Tswana-speaking individuals of this investigation that are candidates to be used for new haplogroup-defining mutations within the PhyloTree classification system (Van Oven and Kayser, 2009). Both of these sequence variants have been observed in Tswana-speaking individuals who were phylogenetically positioned within the haplogroup L0d clade and are likely to define sub-haplogroups L0d31 and L0d1b2. A critical need exists for the investigation of larger samples of Tswana-speaking populations of South Africa to investigate the possibility of additional new haplogroups that have developed within these populations over time.

7.7.3 Genetic diversity among Bantu-speaking populations of South Africa

A deficiency of mtDNA coding region sequence data of Bantu-speaking populations of South Africa exists. This highlights the need for a broad dataset of complete mtDNA sequences of different Bantu-speaking populations of South Africa in order to determine the within and between population genetic variance of the Bantu-speaking ethnic groups of South Africa. In addition, the increased movement of rural Bantu-speaking individuals to urban areas in South Africa has led to a marked change in the sociological and economic circumstances of these individuals (Lane *et al.*, 2002) and it would therefore also be of importance to understand what the genetic implications of these environmental changes would be on the genetic variation and ultimately the functional variants that are associated with diseases.

The self-identified description of individuals' ethnicity and heritage is more often based on socio-cultural, linguistic or geographical association than necessarily on genetic ancestral heritage. This problem is especially prevalent in South Africa, which is a country that is home to ethnically diverse population groups and in which nine Bantu languages are officially recognised. Based on the languages, the South African ethnic groups are classified in the Sotho/Tswana group that consists of speakers of South and North Sotho and Tswana, the Nguni group that includes the Xhosa, Zulu and Tsonga/Shangaan

speakers, and the Venda group (Lane *et al*., 2002). These ethnic groups have developed culturally and linguistically into distinct groups that often also reside in specific regions of the country. The study by Lane *et al.* (2002) demonstrates that these Bantu populations do not display high levels of genetic differentiation according to autosomal and Y chromosome data, which made them conclude that the Bantu speakers share ancestry and have not been isolated into the different ethnic groupings long enough to demonstrate genetic differentiation. Large-scale sequencing of the full mitochondrial genomes of the different South African Bantu-speaking populations will contribute greatly to understanding of the maternal contribution to the evolutionary history of these groups in terms of the degree of genetic differentiation among them, as well as provide a large array of sequence data from which sequence variation can be studied in more depth to confirm the findings of Lane *et al.* (2002).

Sufficiently large samples of full genome mtDNA sequence data would further provide valuable information from which the cultural, lifestyle and genetic assimilation between the agriculturist South African Bantu speakers and the traditional hunter-gatherer Khoi-San speakers of southern Africa could be concluded. Moreover, as already discussed in Section 7.7.2, based on the high levels of substructure within the African regional populations, it would be important to investigate the Bantu-speaking populations of South Africa for region-specific functional variants that are associated with disease, drug responses and risk factors. Sequence variance frequency information of populations is furthermore of critical importance in the forensic identification of humans. The extent of genetic diversification between populations is used to accommodate the biases that are caused by population stratification, as is currently the case in the urban areas of South Africa, where a specific socio-economic population may consist of individuals that are from different ethnic origins but are classified as one ethnic grouping within the urban setting. In South Africa, the population groupings that are used to determine the statistical weight of DNA evidence are based on databases that classify the population into different racial groups such as “blacks” and “whites”, which may be insufficient for true differentiation. Better genetic differentiation between different ethnic groupings will be determined when different populations are broadly sampled and compared, thus contributing not only to the determination of sequence variants but also to the subdivision between groupings for forensic purposes.

7.7.4 Genetic diversity within the Tswana-speaking population of South Africa

The Tswana speakers of South Africa are classified as belonging to the larger Sotho/Tswana ethnic group that consists of speakers of South and North Sotho and Tswana, and resides in different regions of the country. The findings of this investigation constitute the genetic variance within a few ethnolinguistic Tswana-speaking settlements of the North West Province of South Africa and do not represent the larger geographic Tswana-speaking populations of South Africa. Therefore it is suggested that an investigation be performed on a more representative cohort of Tswana-speaking individuals of South Africa.

7.7.5 Use of additional markers for the study of genetic diversity within and between South African Bantu-speaking populations

The use of only mtDNA sequence data for the determination of the evolutionary history of different populations provides a one-sided perspective. The evolutionary history of African populations across broad geographical regions has been studied extensively by Y chromosome sequence data (Wood *et al.*, 2005) and by the determination of the autosomal sequence variation of more than 1,000 autosomal markers in over 2,000 African samples (Tishkoff *et al.*, 2009). The differences between the patterns of inheritance between the paternally and the maternally inherited loci provide information about the manner in which humans dispersed and the type of marriages that were practised by the early Bantu-speaking populations. Evidence from early Bantu-speaking populations, including the early Tswana-speaking populations of South Africa, suggests that intermarriages tended to be between hunter-gatherer women and Bantu-speaking farmer men in patrilocal societies (Wood *et al.*, 2005). The findings of this study support this theory through the observation of the high degree of dilution of the Bantu-speaking mtDNA with the mtDNA haplogroups of the hunter-gatherer Khoi-San speaking populations of southern Africa. To verify this theory, it would be important to investigate the Y chromosome patterns of the Tswana-speaking populations of South Africa. Studies of Y chromosome patterns within populations that have been studied for mtDNA sequence variation would further provide valuable information on early expansion events in terms of the movements of the male individuals of the tribes. The study of Y chromosome sequence will contribute to elucidate the history of the early migrations of the Bantu-speaking ethnic groups in South Africa to effect the establishment of the current ethnic groupings. In addition, autosomal SNPs provide high levels of variation due to the large amount of polymorphisms that are available for investigation and especially provide

good resolution of the effects of population changes and proportions of shared ancestry. The use of Y chromosome and autosomal genetic data provides the only way in which comprehensive conclusions can be drawn about demographic events of past populations and would be important in the determination of the evolutionary history of the Bantu-speaking populations of South Africa.

7.7.6 Future population genetic inferences from genomic sequence data

A whole range of next generation sequencing technologies has changed the scope of DNA sequencing projects. Large-scale genomic sequencing has been made possible by the introduction of sequencing technologies such as Illumina sequencing (Bentley *et al.,* 2008), pyrosequencing (Margulies *et al.*, 2005), SOLiD sequencing (McKernan *et al.*, 2009) and cPAL sequencing (Drmanac, 2009). The use of RRSS in combination with the next generation sequencing technologies have provided the means for large-scale generation of genomic sequencing data that will provide answers to many questions of evolutionary histories and past demographic events. It would provide higher resolution of haplogroup information and identification of more rare population-specific sequence variants. This would allow more accurate models of evolution and interpretations of the past demographic history of modern humans.

7.7.7 Investigation of the extent of the effects of selection on genetic variation

Understanding how selection targets human genomes would increase the understanding of evolution, gene function and the basis for genetic disease and is an ongoing priority within the field of molecular phylogenetics. The study of selection within African populations that are genetically and phenotypically highly diverse makes it important to include more ethnically diverse African populations in the quest to understand the evolutionary processes of selection. The interpretation of population genetic neutrality tests, as performed in this investigation, is often inconclusive because of the presence of unknown demographic factors that confound the outcome of the tests for selection. Many authors have suggested that selection plays a fundamental role in the evolution of the mitochondrial genome (Excoffier, 1990; Merriwether *et al.*, 1991; Mishmar *et al.*, 2003; Kivisild *et al.*, 2006) and evidence of selection in the mitochondrial genomes of humans is increasing (Bamshad *et al.*, 2003), which leads scientists to question whether the primary force of genetic diversity can be ascribed to genetic drift and whether genetic variation and effective population size are not primarily governed by models of selective sweep (Nielsen,

2005). The accurate identification of selection in different functional regions within the human genome can furthermore contribute to the successful prediction of putative disease factors (Nielsen *et al.*, 2007). This is therefore of importance for the interpretation of evolutionary history based on molecular data, as well as for the understanding of the genetic signals produced by this investigation. The genetic signals of selection, such as a deficit of sequence variants in regions under negative selection or an excess of rare sequence variants in regions under positive selection, have been incorporated into methods that scan large regions of genomic data for affected loci, such as using the distribution of human SNPs to scan for selective sweeps (Nielsen and Beaumont, 2009). Large population genetic and genomic datasets, as produced by next generation sequencing, will enable the determination of the proportion of mutations affected by positive and negative selection respectively. Using methods that are robust with regard to the underlying demographic factors would enable the genetic signals for selection to be connected to the signals of demographic factors (Nielsen, 2005). Other approaches for to investigating the effects of selection within genomic data rather than only mitochondrial sequence data include the investigation of linkage patterns that are associated with distinct patterns caused by selective sweeps.

This study contributes to understanding the regional phylogenetic topology and ancestral relationships observed in southern Africa, via the mtDNA investigation of a Tswana-speaking cohort. The Tswana-speaking cohort provides insight into the interconnectedness of this ethnic group with some of the other reported sub-groups of Bantu speakers that reside in South Africa. Full understanding of the complex inter-relatedness of the diverse subpopulations of southern Africa will only be achieved if the complete genomes, nuclear and mitochondrial, of the entire region’s diverse populations are known.