CHAPTER SIX

Results and Discussion

The issues concerning the genetic evolution during the Bantu expansions in Africa have received significant attention over the last few decades (Salas *et al.*, 2002; Atkinson *et al.*, 2009). The investigation of African mtDNA has contributed greatly to the understanding of the demographic settlement of populations globally and also in Africa, while shedding light on the evolutionary development within and between the Bantu-speaking populations of the African continent (Pereira *et al.*, 2001; Salas *et al.*, 2002; Atkinson *et al.*, 2009). The understanding of population dispersal across the African continent is, however, a complex interplay between migration patterns, population expansion and contraction, population bottlenecks and founding effects, as well as gene flow and admixture within and between the African populations, which could only be resolved through information of the genetic diversity of more African Bantu-speaking populations (Atkinson *et al.*, 2009). The aim of this study was to investigate the mtDNA of a Bantu-speaking population of the southern African region with the purpose to contribute novel genetic diversity information to the current body of scientific literature and ultimately provide greater insight into the questions relating to the evolutionary and demographic past of these populations. Studying an ethnic African population living in the northern areas of the North-West and Free State provinces in modern-day South Africa would therefore add significantly to the understanding of the demographic settlements of modern-day African populations in relation to the current view of migrational patterns of the historic Bantu groups. A comparison of the mtDNA sequence variation found in the Tswana population under investigation in this study with current phylogenetic schemes and information about the geographic and ethnic distribution of African haplogroups and sub-groups will provide information about the possible migratory pattern and population admixture of this population (Torroni *et al.*, 2006).

The aims of this study were achieved by isolating, amplifying and sequencing the full mitochondrial genomes of the 50 Tswana-speaking individuals of this investigation with the purpose of determining the mtDNA sequence variability in the context of the published full mtDNA sequences of a dataset of other African individuals, which was constructed for the purposes of this study. The sequence variation of the Tswana-speaking individuals of this investigation were examined and compared with the published genetic variability of other

African populations, while haplogroups were assigned to the full mitochondrial sequences of the African individuals contained in the dataset of this investigation, as well as to the Tswana-speaking individuals of this investigation, with the purpose of establishing the maternal ancestry between and within the populations. The mtDNA sequences of the Tswana-speaking cohort of this investigation and the datasets containing the mtDNA sequences of African individuals were further submitted for phylogenetic analyses to determine the phylogenetic positioning of the Tswana-speaking individuals of this investigation in the context of other African individuals. The genetic variability of the Tswana-speaking population under investigation was further investigated by using specific population genetics methods that consisted of statistical analyses to provide information about the genetic diversity, the size of the populations over time, the effect of selection in shaping the genetic diversity and the population structure and finally an attempt was made to determine the coalescence times of the haplogroups of the Tswana-speaking individuals of this investigation.

Finally, a novel Tswana consensus sequence was constructed based on the sequence variance observed in the full mitochondrial genomes of the 50 Tswana-speaking individuals of this investigation. The purpose of the consensus sequence was to provide a baseline for the sequence variance that is present in a Tswana-speaking population of South Africa and as a representation of the genetic diversity of the maternal ancestral genetic pool of a Bantu-speaking population of South Africa.

6.1 POLYMERASE CHAIN REACTION

The aim of determining the PCR of the Tswana sample set was to generate good quality amplified DNA for the eight regions of the full mitochondrial genome. This entailed generating PCR products that were consistent in yield, had low background and no artefacts. A standard PCR protocol was used for the PCR reactions and is presented in Section 5.4.2. The Promega GoTaq®[1]Flexi DNA Polymerase kit was used in this investigation and contained GoTaq®DNA Polymerase that was used in conjunction with the 5 X Colorless GoTaq®Flexi Buffer and 25 millimolar $MgCl_2$. The forward and reverse primers used are described in Section 5.4.1.

1 GoTaq® is a registered trademark of Promega Corporation, USA.

6.1.1 Primers

The primers used in this investigation were selected from a set of 32 forward and 32 reverse primers published by Maca-Meyer *et al.* in 2001 and presented in Table 6.1. The primers were selected to amplify target DNA lengths of ~2 kb, which are regarded as short fragment lengths for PCR reactions (Cha and Thilly, 1993) and therefore allow for optimal PCR efficiency. The selection of the primers was performed by M. Koekemoer (2010) to ensure consistency between projects.

Table 6.1 Primer pairs used in this investigation

Primer region	Primer name	Primer sequence	Product size (bp)	Overlap (bp)
1	F32:mtL15996	5'-ctc cac cat tag cac cca aag c-3'	2,103	564
	R3:mtH1487	5'-gta tac ttg agg agg gtg acg g-3'		
2	F3:mtL923	5'-gtc aca cga tta acc caa gtc a-3'	2,789	26
	R7:mtH3670	5'-ggc gta gtt tga gtt tga tgc-3'		
3	F8:mtL3644	5'-gcc acc tct agc cta gcc gt-3'	2,227	554
	R11:mtH5832	5'-gac agg ggt tag gcc tct tt-3'		
4	F11:mtL5278	5'-tgg gcc att atc gaa gaa tt-3'	2,679	36
	R15:mtH7918	5'-aga tta gtc cgc cgt agt cg-3'		
5	F16:mtL7882	5'-tcc ctc cct tac cat caa atc a-3'	2,089	42
	R19:mtH9928	5'-aac cac atc tac aaa atg cca gt-3'		
6	F20:mtL9886	5'-tcc gcc aac taa tat ttc act t-3'	2,231	590
	R23:mtH12076	5'-gga gaa tgg ggg ata ggt gt-3'		
7	F23:mtL11486	5'-aaa act agg cgg cta tgg ta-3'	2,740	574
	R27:mtH14186	5'-tgg ttg aac att gtt tgt tgg-3'		
8	F27:mtL13612	5'-aag cgc cta tag cac tcg aa-3'	2,828	405
	R32:mtH16401	5'-tga ttt cac gga gga tgg tg-3'		

From Koekemoer (2010) and Maca-Meyer *et al.,* 2001. Primer region numbers 1 - 8 refer to amplified segment as discussed in Section 5.4.1. F = forward primer and R = reverse primer. L = light strand and H = heavy strand. Primer numbers refer to the nucleotide position according to the CRS where amplification is started and product size to the length between the starting points of the two primers in a pair. T_m = melting temperature of the primer and the mean T_m referring to the mean of the T_m values for both primers. Overlap lengths refer to the overlap between the region indicated and the next region.

6.1.2 PCR optimisation

Optimisation of the PCR process ensures high yields of amplification product and specificity of amplification. This entailed finding the optimal balance between reaction components and cycling parameters to prevent mispriming and generating optimal DNA product yield. Since a commercial PCR kit (Promega GoTaq® Flexi DNA Polymerase kit) was used, it was assumed that the enzyme concentration and buffer composition had

been validated for optimal PCR yield and therefore did not have components that had to be adjusted. Cycling parameters, more specifically the annealing temperature (T_a), were adjusted as first line of optimisation. The kit used allowed the $MgCl_2$ to be added separately and in the absence of successful optimisation this would have been the second line of optimisation to be followed.

Three Caucasian samples were used to optimise PCR conditions of the eight pairs of primers to be used for the amplification of the full mitochondrial genome. A negative control was included in each of the optimisation runs to ensure that there was no contamination during reaction set-up and to rule out false positives caused by contamination. On completion of the PCR cycling, the PCR products were loaded onto an agarose gel and after electrophoresis visualised by using a UV light source (see Section 5.5). Primer pairs were regarded as optimised when single PCR product fragments of a certain molecular weight were present on the gels. Negative controls had to be clean of any DNA product.

The optimisation of the annealing temperatures started with using the calculated melting temperatures (T_m) of the primer-template pairs. The initial T_a was determined by using a melting temperature (T_m) for each individual primer as determined by using the OligoCalc: Oligonucleotide Properties Calculator software version 3.07 (Kibbe, 2007). This calculation was performed by a former student of the CGR, Dr M. Koekemoer (2010), in order to ensure consistency in the methods used. The T_m of each of the primer pairs was averaged and used as a starting point for the optimisation of the primer pairs.

6.1.3 PCR efficiency

The efficiency of the PCR for the different PCR regions varied, as indicated in Table 6.2. The values varied between 49.5 $ng.\mu L^{-1}$ and 73.5 $ng.\mu L^{-1}$ of DNA under optimised conditions and although not the same, these values did not indicate extreme variation in the DNA yield of the respective PCR reactions.

Table 6.2 Average DNA quantity for the different PCR regions under optimised conditions

PCR region	Average DNA quantity
Region 1	59.3 ng.μL^{-1}
Region 2	66.3 ng.μL^{-1}
Region 3	72.2 ng.μL^{-1}
Region 4	57.6 ng.μL^{-1}
Region 5	70.4 ng.μL^{-1}
Region 6	60.5 ng.μL^{-1}
Region 7	73.5 ng.μL^{-1}
Region 8	49.5 ng.μL^{-1}

DNA quantities given here are averaged for three (3) samples used during PCR optimisation. Values used were for optimised conditions.

The efficiency and yield of the PCR are usually better for smaller sized DNA fragments that are less than ~2 kb in length (Cha and Thilly, 1993). The average length of the target fragments in this investigation was 2,461 kb. For this reason it was not expected that the fragment lengths in this investigation would lead to unacceptably low levels of amplification efficiency, as was confirmed by the results in Table 6.2.

Although the target, the primer sequences and the concentrations of other components in the reaction, such as the dNTPs and the primers, can influence the optimality of the PCR buffer used, this was not regarded as a significant factor that influenced the efficiency of the PCR in this investigation, as the buffers used in commercially available PCR kits are validated for the reaction components.

Since divalent cations are essential for PCR, the concentration of the $MgCl_2$ plays a major role in the efficiency of the PCR and must be adjusted according to the concentration of the dNTPs because the negative charge of the phosphate backbone of the dNTPs will affect the availability of the Mg^{2+}. For this reason the PCR kit used in this investigation allowed for $MgCl_2$ to be added separately, making this type of optimisation a possibility.

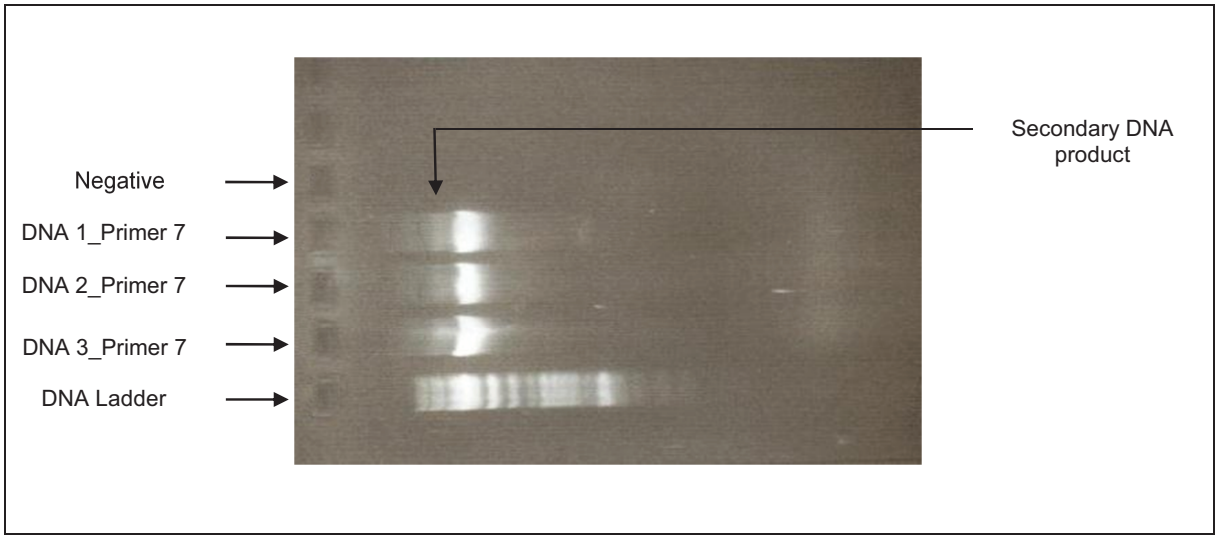
PCR efficiency is ensured by using nonlimiting amounts of primers and dNTPs. Excess amounts of primer will ensure that the template DNA binds with primer instead of with other DNA targets during denaturation. If the concentration of primer is, however, too high in relation to the concentration of the template, secondary DNA products and primer-dimers will form. It is therefore critical to find a balance. A standard quantity of DNA template and primer, as prescribed by the PCR protocol of the CGR of the the North-West University, was used in this investigation. The PCR yield of the different PCR regions was

not regarded as too extreme to necessitate the adjustment of the primer or input DNA template quantities.

6.1.4 Secondary PCR product

Unspecific amplified product can be formed when primers anneal to non-specific sites in the DNA template or due to overamplification. Misprimed products compete with DNA targets for the use of dNTPs and primers and can affect the eventual DNA product yield. Prevention of mispriming due to incorrect annealing of primers can be established by the inclusion of agents such as formamide or glycerol, lowering the pH and the concentrations of dNTPs, primers and $MgCl_2$. Raising the annealing temperature and shortening the annealing time is an effective mechanism to ensure the specific binding of primers to a target area. In this investigation, mispriming was prevented by the use of stringent annealing temperatures, low concentrations of primer and an optimal concentration of $MgCl_2$. The PCR set-up was done with reagents and samples on ice. Not working at ambient temperatures was also a measure taken to prevent secondary PCR products from forming in reaction to mispriming and primer oligomerisation. Low T_a was, however, determined to be the general cause of secondary product in this investigation. Regions 4, 7 and 8 were affected by the presence of secondary product at low T_a. This problem was overcome in all cases by increasing the T_a.

Figure 6.1 Photographic representation of secondary product of Region 7 optimisation at low T_a



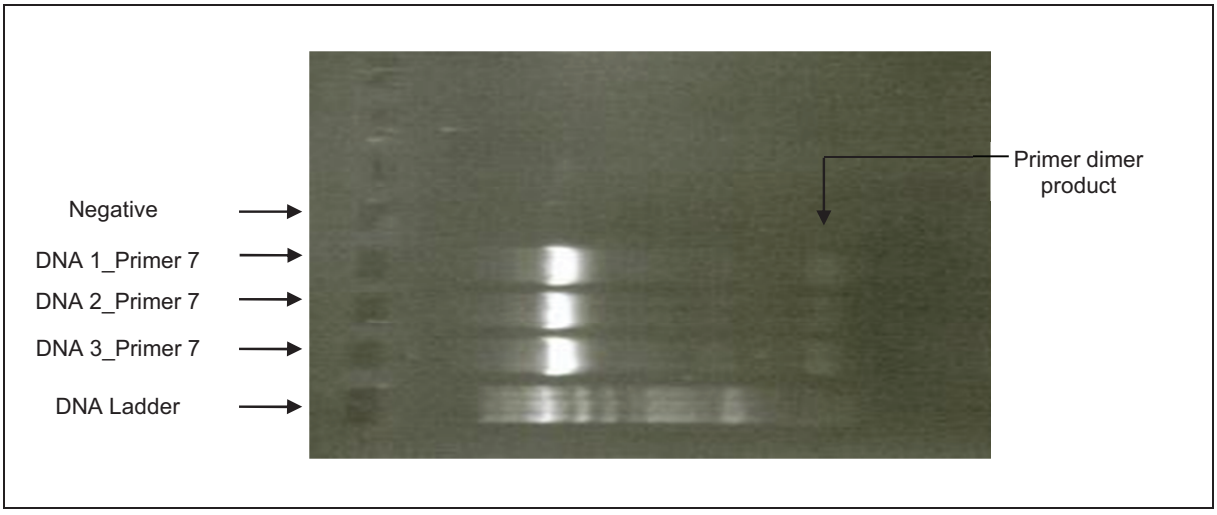
Agarose gel (0.9%) run for 30 min at 100 V and 50 mA. Gel loaded with a FastRuler™High Range DNA Ladder and three (3) samples of DNA sample that were amplified with Primer 7 at T_a 60°C. Negative control was included.

6.1.5 Primer-dimers

When primers interact with each other and not with the template, primer-dimers are formed. This happens because primers are present in the PCR reaction mix in high concentrations and will interact even if there is only one complementary nucleotide present. This will be greatly enhanced after 30 cycles (Brownie *et al.*, 1997). Inter-primer extensions are good substrates for amplification in PCR cycles and create a mixture of primer artefacts.

Primer-dimers were reduced by using well-designed primers with minimal complementarity and stringent PCR cycling conditions. Preparing the PCR set-up at room temperature could have contributed to the formation of primer dimers and for this reason the preparation of the PCR reactions was performed on ice. This problem is much more pronounced in multiplex systems and in this study the singular amplification of eight regions with only one pair of primers each contributed greatly to lessening the opportunity for finding this type of artefact. Region 7 was the only region in which a low annealing temperature led to very faint primer-dimer products, as indicated in Figure 6.2.

Figure 6.2 Photographic representation of primer-dimer product of Region 7 optimisation at low T_a

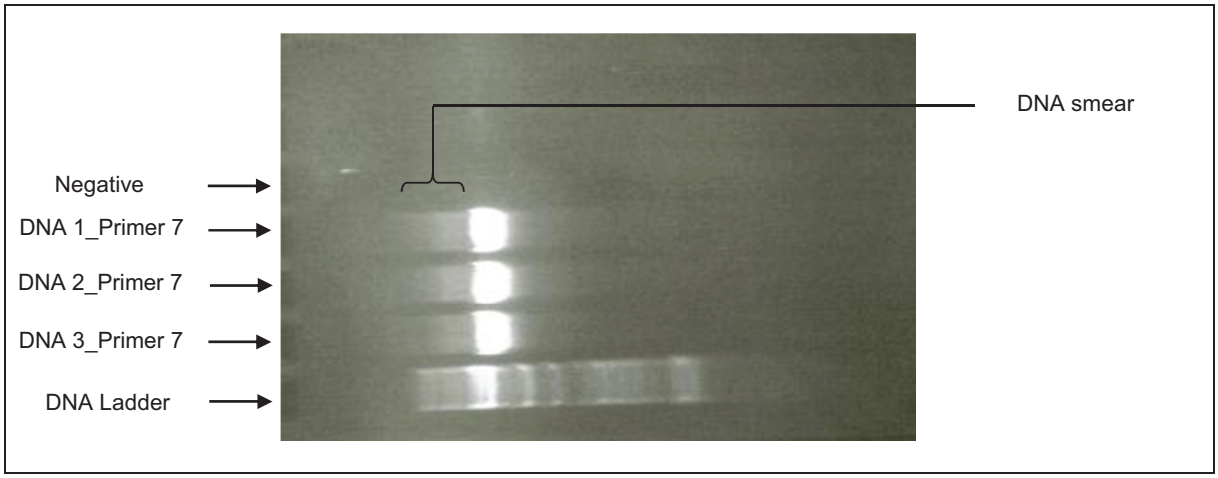


Agarose gel (0.9%) run for 30 min at 100 V and 50 mA. Gel loaded with a FastRuler™High Range DNA Ladder and three (3) samples of DNA sample that were amplified with Primer 7 at T_a 57°C. Negative control was included.

6.1.6 PCR product smearing

Smears and spurious fragments are produced by too many temperature cycles or when the level of starting template is too high. Smears can also be formed when the priming is non-specific and a range of secondary DNA products of different sizes are formed. Smears were detected with optimisation of region 7 and region 8. This could be resolved with the optimisation of the T_a for region 8 but could not be resolved with region 7. Judging by the intensity of the fragments as indicated in Figure 6.3, it was assumed that the level of starting DNA for this region was extremely high and that this was the cause of the continuous presence of the faint smear with the PCR electropherogram.

Figure 6.3 Photographic representation of the smear found with optimisation of region 7 primers



Agarose gel (0.9%) run for 30 min at 100 V and 50 mA. Gel loaded with a FastRuler™High Range DNA Ladder and three (3) samples of DNA sample that were amplified with Primer 7 at T_a 57°C. Negative control was included.

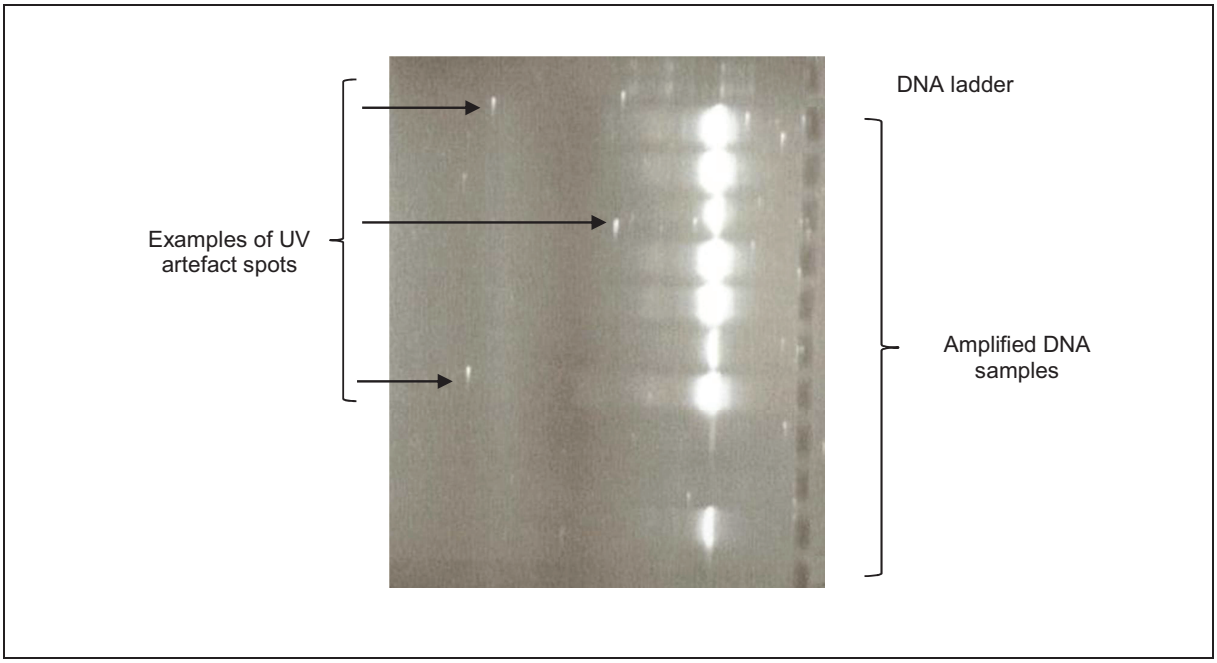
6.2 AGAROSE GEL ELECTROPHORESIS

The PCR product was loaded onto a 0.9% agarose gel and electrophoresed at 100 V and 50 mA for 30 minutes in order to separate the fragments and evaluate the quality of the PCR product. A casting tray large enough to take a sample comb for 25 samples was used routinely to load two batches of ten (10) samples each. A FastRuler™[1] High Range DNA Ladder (Fermentas) of range 100 – 10,000 bp was included in the first lane, in the middle 13th lane and the last lanes of the gel. The negative controls were loaded in lanes 12 and 24 for the two batches respectively.

1 FastRuler™ is a registered trademark of Fermentas International, Inc., Ontario, Canada.

The agarose gel electrophoresis was robust and the artefacts determined did not interfere with the results. Artefacts included some bright UV spots on the gels due to pieces of lint that came from the paper towels used to clean the electrophoresis casting trays. These spots could easily be identified as artefacts, as can be seen in Figure 6.4, and did not interfere negatively with the evaluation of the DNA fragments on the gels.

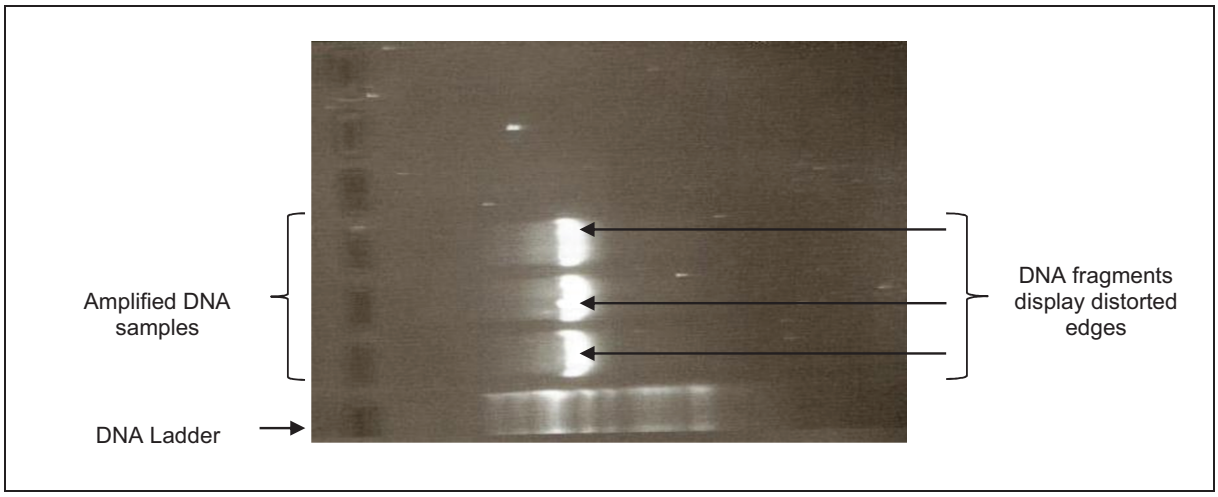
Figure 6.4 Photographic image of the UV artefact spots observed in some of the gels



Agarose gel (0.9%) run for 30 min at 100 V and 50 mA.

Some of the gels showed samples that were distorted, most probably because the samples were not loaded neatly into the gel well or the wells were not washed clean from agarose prior to loading the samples. Again the artefact did not interfere with providing a good estimate of the quality of the PCR product for the purposes of further cycle-sequencing reactions. See Figure 6.5 for a comparison between good and distorted samples from two different gels.

Figure 6.5 Example of distorted DNA fragments on a gel



Agarose gel (0.9%) run for 30 min at 100 V and 50 mA.

6.3 DNA PURITY AND QUANTITY

The PCR product was purified prior to cycle sequencing by using the Zymo Research DNA Clean & Concentrator™-5 kit (see Section 5.6) and the concentration was determined by optical density measures by using the Eppendorf®[1] BioPhotometer 6131 instrument (see Section 5.7). The purity of the sample could be estimated by the absorbance values of the A_{260}/A_{280} ratio. Values of the A_{260}/A_{280} ratio less than 1.7 were regarded as indicative of contamination by protein or organic chemicals.

Table 6.3 Average DNA quantity obtained for the eight different PCR regions

PCR Region	DNA quantity (ng.µL^{-1})	PCR Region	DNA quantity (ng.µL^{-1})
Region 1	44.5	Region 5	47.1
Region 2	54.0	Region 6	46.8
Region 3	40.6	Region 7	34.0
Region 4	30.2	Region 8	31.8

DNA quantities averaged for all 50 Tswana samples per region.

6.4 AUTOMATED DNA SEQUENCING OF THE FULL MITOCHONDRIAL GENOME

The full mitochondrial genome sequencing of the 50 Tswana samples in this study represents enough data from which accurate estimations about evolutionary events such as base composition, codon usage, insertion and deletion processes and selective processes can be made (Pollock *et al.*, 2000). Studying only a single gene in phylogenetic

1 Eppendorf® is a trademark of Eppendorf AG, Hamburg, Germany.

analyss limits the level of accuracy achieved. For this reason it was decided to analyse the whole mtDNA genome of 50 individuals from a Tswana population.

The full mitochondrial genome was sequenced by employing a cycle-sequencing method using the BigDye® Terminator v3.1 Cycle Sequencing Kit. Purified PCR product was used for sequencing reactions as described in Section 6.3. The cycle sequencing-reaction protocol used is described in Section 5.8.2. Input PCR product concentration was optimised for this investigation as presented in Section 5.8.2. The Thermo Hybaid®[1] MBS 0.5S thermocycler was used for the sequence cycling of the DNA product and after completion of the sequence cycling, the sequencing reactions were treated with SDS to remove excess dye terminators, thereby preventing the dye blobs from interfering with the end sequencing results. As mentioned in Section 5.8.5, the electrophoresis of the sequence extension products was not performed on site but by another institution as per contract with the North-West University. On completion of the electrophoresis, the raw data files of the samples were returned to the CGR electronically.

6.4.1 Sequencing strategy

All eight PCR regions, consisting of DNA fragments of between 2,103 bp and 2,828 bp long, were sequenced in four overlapping fragments using four forward primers designed for each PCR region. These primers were based on the primers used in Maca-Meyer *et al.* (2001) as described in Section 5.8.1. The four forward primers and fragment lengths required, without taking the overlap sequence into account for the eight PCR regions, are indicated in Figure 6.6.

1 Thermo Hybaid® is a registered trademark of Hybaid Ltd., Ashford, Middlesex, UK.

Figure 6.6 Sequencing primers for the eight PCR regions

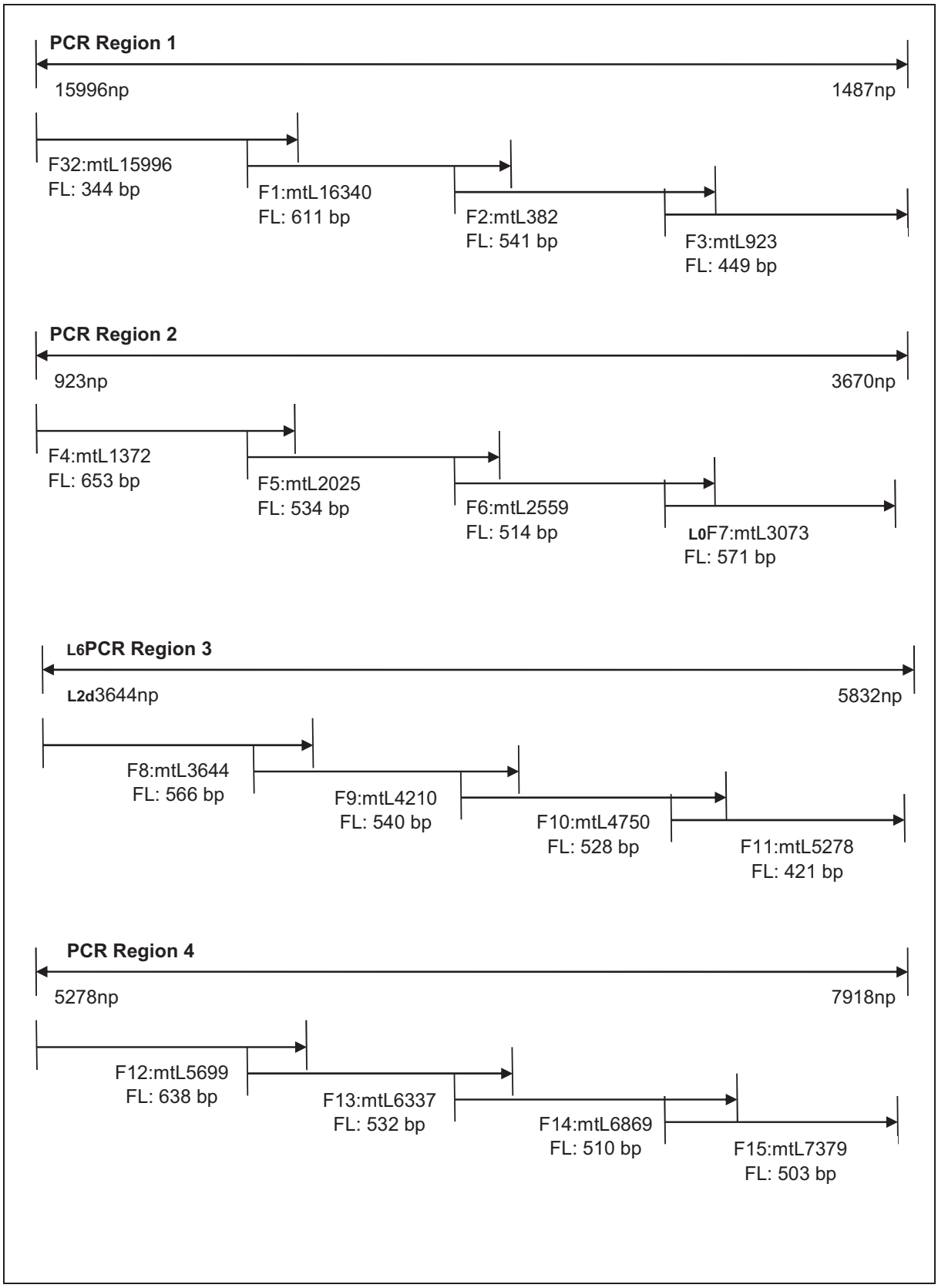
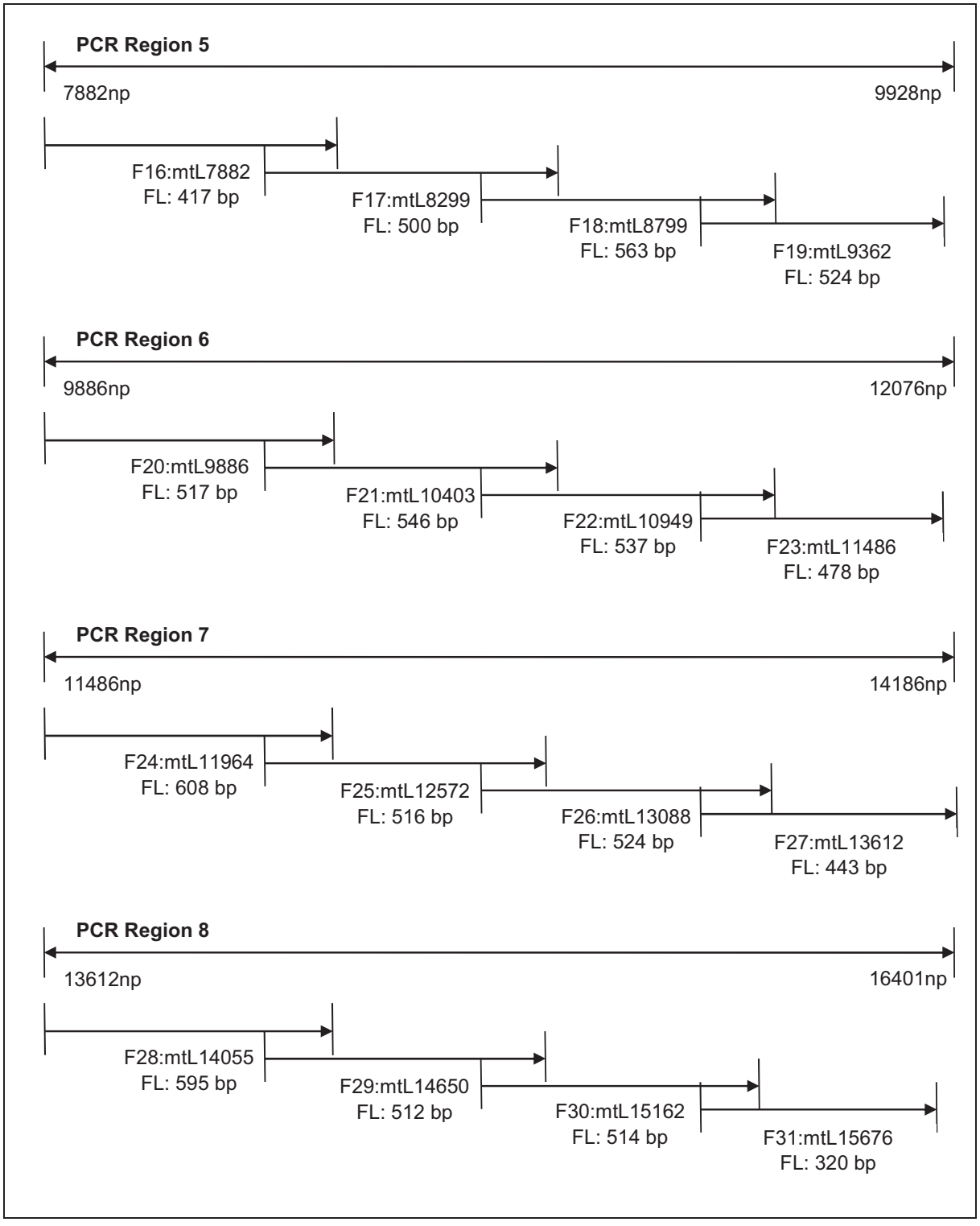


Figure 6.6 Continued…



np = nucleotide position, FL = fragment length, F= forward primer, L= light strand, bp = base pairs. All numbers refer to the nucleotide position according to the rCRS of Andrews *et al.* (2001).

6.5 DATA ANALYSIS RESULTS

A universal standard nomenclature of the mitochondrial genome was used in this investigation for the analysis of the sequence data. It is based on positions and base identities allocated for the first published complete mitochondrial DNA sequence, referred to as the rCRS or Anderson sequence (Tully, *et al.*, 2001). The standard mitochondrial DNA nomenclature refers to nucleotide bases on the light strand (L-Strand) of the rCRS. Sample sequences are reported according to the differences with the revised Cambridge sequence by quoting the base on the light strand of the revised Cambridge sequence, the position of the revised Cambridge sequence, and the changed base of the sample sequence e.g. A3243G.

6.5.1 Sequence alignment

The mitochondrial sequences of this study were aligned by using the CLUSTAL X Multiple Sequence Alignment Program version 2.0.12 (Larkin *et al.*, 2007) according to an algorithm that constructs a distance matrix between pairs of sequences based on the pairwise sequence alignment similarity scores and the penalties for deletions and insertions (Tamura *et al.*, 2007). As discussed in more detail in Section 4.1.1, the algorithm uses a progressive method whereby the evolutionary relationships between homologous sequences are used to align them by constructing a rough NJ tree with midpoint rooting and using it as a guide to group the sequences that are closely related and those that are more distant, giving direction to the alignment (Larkin *et al.*, 2007; Levasseur *et al.*, 2008). There are, however, some problems with the progressive approach. The algorithm greedily adds sequences in order of relatedness based on the clade tree topology and errors that are made in these alignments are not corrected later in the process, and will therefore become progressively more pronounced. Any errors in the preliminary constructed tree will lead to potential alignment errors and the more divergent the sample set, the larger the impact of these errors (Thompson *et al.*, 1994). However, in the case of this study, the sequences came from populations that could be expected to be closely related, in addition to the fact that they consisted of full genome sequences of known length, making the alignment errors obvious and therefore less likely to occur.

6.5.2 DNA contiguous sequences

Contiguous sequences (contigs) were constructed by joining the DNA fragment lengths of adjacent regions. The contigs were assembled by using DNA fragments ranging between 320 bp and 653 bp in length, as presented in Figure 6.6. In order to position the DNA fragments correctly, overlaps of the DNA fragments of at least 45 bp were required, which ensured that the two adjacent pieces of DNA fragments shared the primer length as well as about 25 bp of fragment length. On average, fragment lengths of 691 bp were obtained and were sufficient to ensure accurate overlap between DNA fragments.

Overlap between the PCR regions are shown in Table 6.4. Overlap between regions 2 and 3, between regions 4 and 5, between regions 5 and 6 and between regions 7 and 8 did not allow for the 45 bp overlap used between sequencing primers. It was decided to use a minimum of ten bp overlap within these regions. Since the DNA fragments were aligned with the rCRS, it was anticipated that the position of the fragments would be correct. The overlap was necessary to indicate the follow-on between the respective fragments and for this reason it was decided that an overlap of ten bp was sufficient.

Table 6.4 Overlap between PCR regions

Overlap between regions	Reverse primer of first region	Forward primer of second region	Overlap (bp)
Region 1 and Region 2	R3:mtH1487	F4:mtL1372	115
Region 2 and Region 3	R7:mtH3670	F8:mtL3644	26
Region 3 and Region 4	R11:mtH5832	F12:mtL5699	133
Region 4 and Region 5	R15:mtH7918	F16:mtL7882	36
Region 5 and Region 6	R19:mtH9928	F20:mtL9886	42
Region 6 and Region 7	R23:mtH12076	F24:mtL11964	112
Region 7 and Region 8	R27:mtH14186	F28:mtL14055	31
Region 8 and Region 1	R32:mtH16401	F32:mtL15996	405

Regions refer to PCR regions in which the full mitochondrial genome was amplified and sequenced. R = reverse primer, F = forward primer, bp = base pairs.

6.5.3 Data quality

The quality of the sequencing data was ensured by using appropriate software (BioEdit version 7.0.5.2, Hall, 2001) for visual inspection of electropherograms, base calling and editing of sequence traces. Good quality data were characterised by well-defined peak resolution, uniform peak spacing and high signal-to-noise ratios (Applied Biosystems, 2009).

A poor quality template is one of the main reasons for poor sequencing results and can be caused by salts or organic chemicals that are carried over from the PCR and sequencing reactions or by contamination with cellular components such as proteins. This problem was addressed by purifying the PCR products and subsequently removing these contaminants. The quality of the template was further verified by running the amplified DNA product on an agarose gel to identify any secondary or chromosomal DNA, which would have presented as multiple extra fragments. Samples that presented with secondary products on the agarose gels were re-amplified under more optimised conditions to prevent the formation of secondary products. By ensuring the quality of the template in this manner, a poor quality template in the sequencing reactions was prevented and therefore no sequencing problems were experienced in this regard. To rule out the possibility of a poor quality template owing to the degradation of DNA in storage, the PCR product was sequenced within one week of amplification. Degradation of the DNA template was further prevented by limiting repeated freeze-thaw cycles of the DNA template and not freezing the PCR product and sequencing reactions at any point during the sequencing process.

6.5.4 Sequencing errors and artefacts

Sequencing errors in evolutionary studies have resulted in serious concerns about the possibility of recombination in non-recombinant mitochondrial DNA (Eyre-Walker *et al.*, 1999; Hagelberg *et al.*, 1999) and incorrect estimations of time depths of the development and migration of haplogroups (Stenico *et al.*, 1996). Bandelt and Kivisild (2006) reported on several published mtDNA sequencing datasets that were fraught with sequencing errors, which resulted in false mtDNA variation estimates and subsequently false time estimates and raised the importance of ensuring the quality of sequencing data when performing a study of mtDNA for the purpose of determination of the evolutionary past of a cohort or for other purposes, such as forensic identity determination (Bandelt *et al.*, 2001a; Bandelt and Kivisild, 2006).

One way in which errors could occur is by an alignment or column shift when preparing a data table, which would cause one or several positions to be misscored (Salas *et al.*, 2005). This type of error was avoided in this study because the full genomes of individuals were sequenced through the use of eight primer regions as described in Section 6.4 and constructing contigs that consisted of DNA fragments that overlapped between the primer regions. The separate alignment of the DNA fragments generated by the different primers

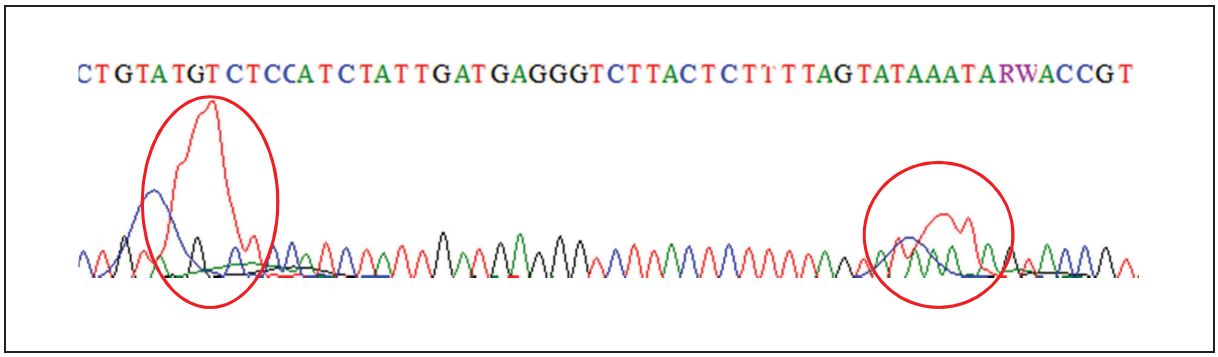
and subsequent verification of the position of the respective fragments during the construction of the contigs made it possible to identify misalignments and correct these.

Other types of errors include an error of reference bias, which occurs when sequence variants are not detected (Salas *et al.*, 2005). This was prevented by editing all electropherograms visually and by following a protocol in which each individual base score was inspected and confirmed. Phantom mutations are caused by errors in the sequencing process, the use of incorrect reading software, the incorrect interpretation of electropherograms and post-mortem DNA damage in old samples (Salas *et al.*, 2005) and were prevented by following the protocol of visual inspection and individual base score inspection and confirmation in order to identify any weak peak morphology. Therefore peak morphology was carefully inspected and only clear, well-defined peaks were accepted for base calling. The sequencing artefacts that were observed within the mtDNA sequences of the Tswana-speaking cohort of this investigation are discussed in more depth in the following sections. Electropherograms were inspected visually for artefacts that might have interfered with base calling, as is discussed in Section 6.5.5. The editing and visual inspection of the electropherograms was performed by using BioEdit version 7.0.5.2 (Hall, 2001).

6.5.4.1 Dye blobs

Dye blobs or excess dye peaks are caused by poor ethanol precipitation, and were generally observed early in the read length of the electropherograms at between 60 and 80 base pairs with another small artefact that was observed between 100 and 115 base pairs, as is presented in Figure 6.7. The dye blobs obscured the sequence data at the positions where they occurred and therefore had to be re-sequenced. The sudden high signal peaks of the dye blobs were caused by excess dye terminators that were present in the sequencing reaction that were not incorporated during the sequencing reaction or because of excess dye-labelled terminators that were not successfully removed from the sequencing reaction during the purification of extension products (Applied Biosystems, 2009).

Figure 6.7 Dye blobs



Example of electropherogram of sample: H02_8_6_1_E_016.ab1. Dye blob nucleotide positions indicated by the red circles.

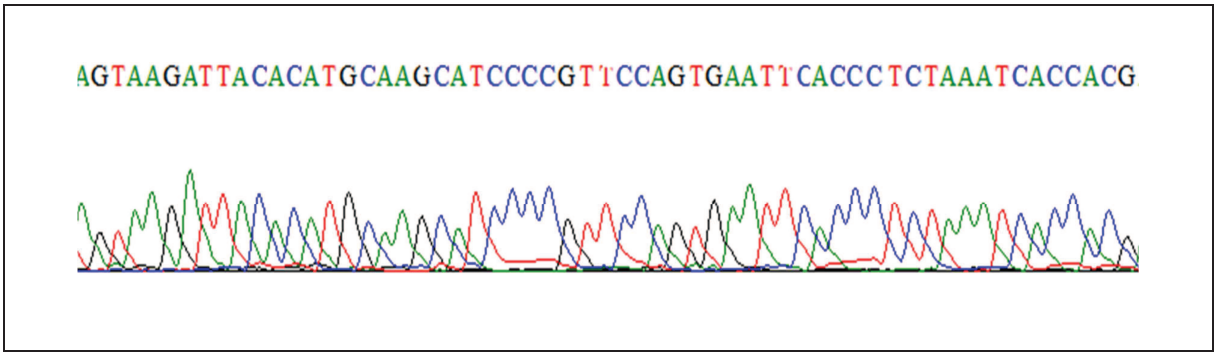
6.5.4.2 Weak signal

Electropherogram peaks that displayed a height of less than 150 relative fluorescent units (rfu) were regarded as displaying a weak signal. Low peak heights could be ascribed to several reasons. The simplest reason would be the lack of sufficient quantities of DNA because of poor amplification during PCR. Other possible reasons include the presence of contaminants in the template, which results in the inhibition of the sequencing reaction, thereby causing a weak signal, human error during the set-up of either of the reactions or poor template quality. The automated sequencers could also contribute to a low signal when insufficient sample is injected into the capillary or when ions in the sample result in poor sample injection or any other problem with the injection or autosampler (Applied Biosystems, 2009). A generally weak signal across all the peaks of an electropherogram that presented as clean peaks with low peak height was interpreted as an indication of low quantities of DNA in the sample, either due to human error or because of contaminants that inhibited the sequencing reactions to automated sequencer error, and was therefore re-sequenced. The possibility of contaminants in the samples was ruled out by the precipitation of the template with ethanol prior to sequencing. Electropherograms that presented with low peak height peaks with considerable background noise were interpreted as containing poor quality DNA and subsequently re-amplified if the problem persisted. In general, however, the peak heights of the mtDNA of the Tswana-speaking individuals of this investigation displayed good peak signals.

6.5.4.3 Trailing peaks

Trailing peaks rendered specific peak morphology unreadable where the base of the peaks were broad and trailing towards the end of the sequence, as indicated in Figure 6.8. This artefact usually appeared in batches where the capillary was used continually for more than a 100 runs, resulting in the build-up of contaminants on the capillary wall over time. These contaminants caused active sites along the capillary wall, which could result in the DNA in the sample adsorbing to the wall and thus creating a trail in the electropherogram. In addition, the active sites could further cause electro-osmotic flow that would have interfered with the efficient flow of the sample DNA through the capillary and result in loss of resolution and trailing of peaks (Butler *et al.*, 2004). The problem was addressed by replacing the capillary.

Figure 6.8 Example of trailing peaks



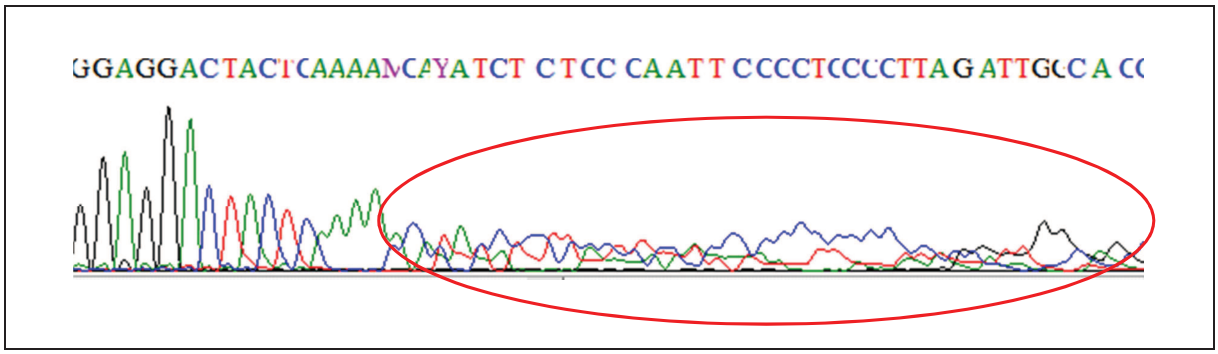
Example of electropherogram of sample: D08_27_1_3_E_008.ab1. Artefact is visible in all the peaks of this electropherogram.

6.5.4.4 Truncated sequence

A truncated sequence is displayed when a good quality sequence is terminated with a sudden truncation of the strong signal peaks and replaced with small, low-level peaks because of the presence of a secondary structure in the sample DNA that causes the inhibition of further sequencing of the DNA fragment. Other reasons include having too much input DNA or primer, which will cause the dNTPs to become exhausted, leading to severe impairment of reaction components to continue the sequencing after a certain point. Furthermore, salt contamination can also cause this type of artefact because it leads to the reannealing of the DNA or the inhibition of the Taq polymerase (Butler *et al.*, 2004). Under the assumption that the quantity of the DNA samples was correct, as was determined before setting up the sequencing reactions; and that the quality of the DNA was good, as was determined by running the amplified DNA products on an agarose gel,

the DNA samples that displayed this artefact were cleaned up by precipitation in ethanol prior to sequencing and re-sequenced to rid them from possible contaminants.

Figure 6.9 Truncated sequence

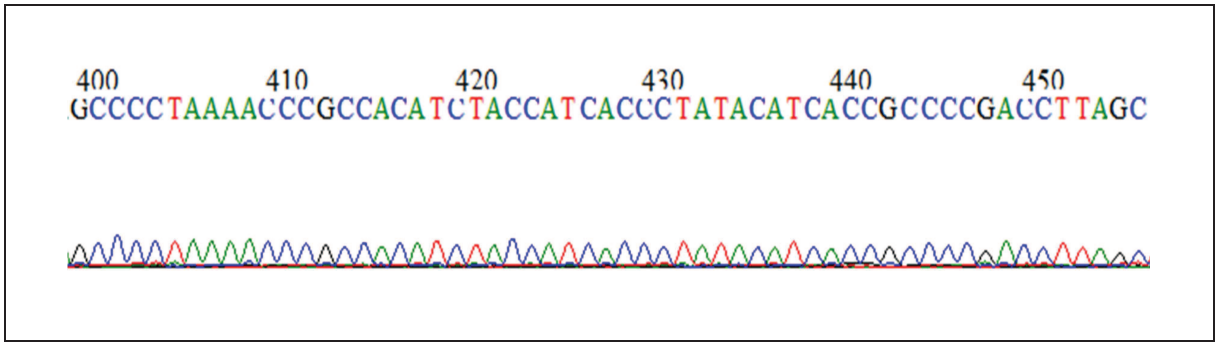


Example of electropherogram of sample: G08_12_50_7_3_E_014.ab1. The start of the truncation of the peaks is indicated by the red circle.

6.5.4.5 Signal loss at the end of the sequence

Figure 6.10 illustrates the low peak height of a sample at around 400 base pairs in length. This artefact was only present in a few batches and when it occurred it always affected the whole batch. A possible reason could be that insufficient amounts of Ready Reaction Mix were used in the sequencing reaction because of human error, which limited the reaction components necessary for optimal sequencing reaction to take place, therefore limiting the product towards the later phase of the sequencing reaction. The problem was solved by re-sequencing the batches.

Figure 6.10 Example of signal loss towards the end of the sequence

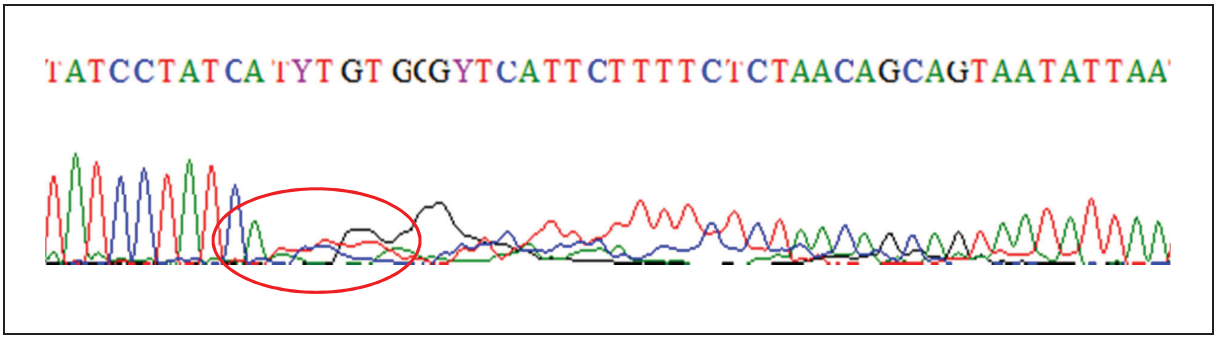


Example of electropherogram of sample: D06_33_2_4_E_008.ab1. The loss of signal occurred from around 400 base pairs, as indicated by the numbering above the sequence. The nature of the loss was gradual and could therefore not be fully presented.

6.5.4.6 Sudden signal loss in the middle of the sequence

The artefact presented in Figure 6.11 indicates a high quality sequence followed by a sudden loss of peak height and peak quality for about 30 bases, after which the signals start to regain quality until they become of good quality sequence again. This artefact was observed within electrophoresis batches and was due to the presence of a contaminant or air bubble that moved through the capillary during electrophoresis. It can also be caused by contamination of the capillary electrophoresis instrument with chemicals during cleaning or by the incomplete replacement of polymer between runs (Applied Biosystem, 2009). This type of artefact was corrected by re-electrophoresis of the batches.

Figure 6.11 Example of loss of signal in the middle of the sequence



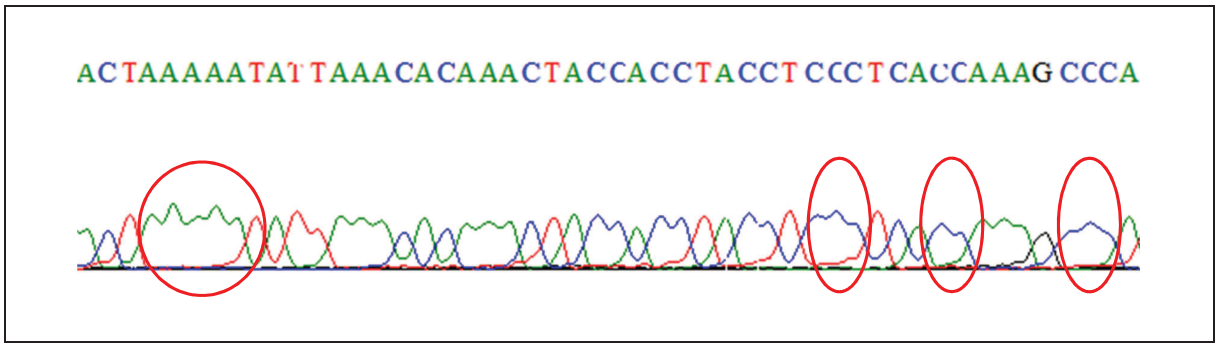
Example of electropherogram of sample: D07_47_4_3_E_007.ab1. The loss of signal is indicated in the area of the red circle.

6.5.4.7 Poor resolution

Only peaks that were well defined, of sufficient signal strength and clear were accepted for base calling. On average, electropherograms contained high-quality peaks up to about 700 bases, from which point the resolution would degenerate until no unambiguous base calling could be performed. In some cases, however, the peak resolution started failing earlier in the read length. This artefact is referred to as poor resolution electropherograms. Poor resolution was identified by poorly defined peaks that tended to become broad and asymmetric and could not be well resolved from each other. Possible reasons for poor resolution are poor capillary performance, old polymer being used, long injection times, incorrect buffer or polymer composition, electrophoresis voltage that is set too high, a sample that is too concentrated, incomplete strand separation due to poor heat denaturation, a sample contaminated by mineral oil, a sample being degraded or use of poor quality water (Applied Biosystems, 2009). In this investigation, poor resolution was identified in batches of samples rather than in individual samples and therefore indicated that the resolution artefacts were mainly caused by electrophoresis problems. Poor

resolution was successfully resolved by re-electrophoresing the samples. It was never necessary to re-PCR a sample to overcome the poor resolution artefacts.

Figure 6.12 Example of poor peak resolution

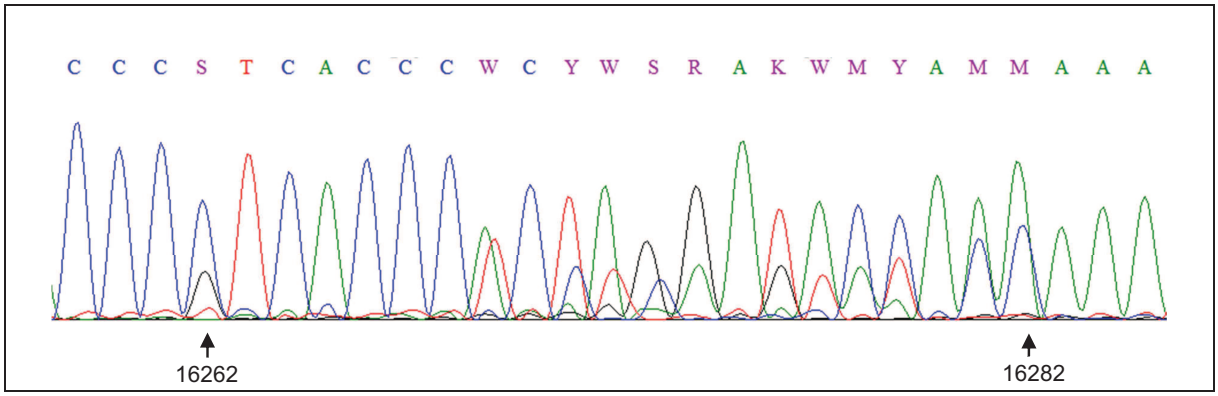


Example of electropherogram of sample: A12_11_5_1_E_002.ab1. Peaks in this electropherogram are generally poorly resolved. Examples of the the worst resolved peaks indicated by red circles.

6.5.4.8 Double sequence between np 16262 and 16282

This artefact affected 50% of the samples in this investigation. This artefact was observed in sequences that consisted of good quality sequence up to nucleotide position 16269, from which point secondary low level peaks were detected, which posed as a second contaminant profile or several successive heteroplasmic peaks. A double peak was present in some of the artefacts at position 16262.

Figure 6.13 Double sequence between nucleotide positions 16262 and 16282

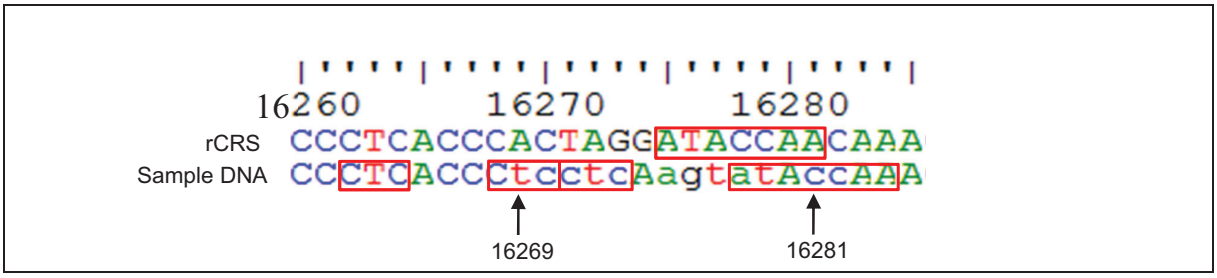


Example of electropherogram of sample: B04_27_1_1_E_004.ab1. Two peaks at nucleotide position 16262 as observed in some of the mtDNA sequences that displayed this artefact and double peaks from nucleotide positions 16269 to 16282.

The overall peak height of the profile was lower in the region of the artefact, as would be expected if the artefact had been caused by heteroplasmy. Detecting heteroplasmy at more than two nucleotide positions in a single mtDNA sequence fragment, however, is highly unlikely and it more often indicates the presence of phantom mutations or contamination (Salas *et al.*, 2005). Because of the limited length at which the extra peaks

of this artefact presented in the samples of this investigation, contamination could be ruled out and the incidence of phantom mutations therefore needed to be considered. Since the artefact was not linked to electrophoresis batches, it was not linked to the electrophoresis process either. Furthermore, it was unlikely that the artefact had been caused by error in sequencing interpretation or reading error by the software because of the high incidence and repeatability of this artefact in terms of peak heights and peak morphology in this dataset. A possible cause is the presence of a secondary structure in the DNA that could have caused the DNA polymerase difficulty in sequencing this sequence segment. It was noted that there was a "ctc" repeat in the affected area that might have caused enzyme slippage. Also interesting was the "ataccaa" sequence that was present in the affected area but did not align with the rCRS. All of these characteristics indicated the possibility of a mutation, indel or secondary structure that might have caused a secondary DNA structure to form at that position. No information about this type of artefact at this specific position in the human mtDNA could be detected in the literature.

Figure 6.14 Alignment of the sequence segment containing the artefact with the rCRS

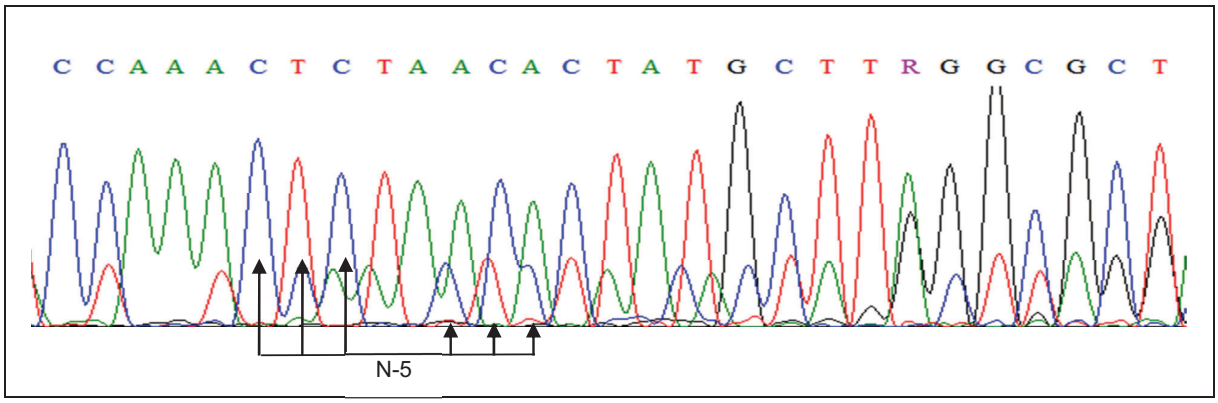


Representation of the alignment of the sample DNA containing the artefact with the rCRS using BioEdit software. The ruler-like scale at the top of the alignment represents the nucleotide position of the DNA aligned according to the rCRS. The small lettered bases are edited bases that were unambiguously called by the software and were manually called during the editing process. The bases in capital letters were called by the BioEdit software.

6.5.4.9 N-5 peaks

This artefact consisted of two identical sequences in one DNA sample profile with the one sequence being five base pairs shorter than the other (N-5), as indicated in Figure 6.15. This could be caused by the presence of primers that differ by five bases in size and therefore cause the formation of sequence fragments that are five bases shorter than the true sequence. If the N-5 primer is present in 40% of the primer stock, it will lead to ambiguities in base calling. The more likely reason for this artefact, however, is the slippage of the Taq polymerase because of the presence of a homopolymer at the beginning of the sequence (Applied Biosystems, 2009). To resolve the problem, primer solutions were remade and the sequencing reaction re-electrophoresed.

Figure 6.15 Example of N-5 artefact



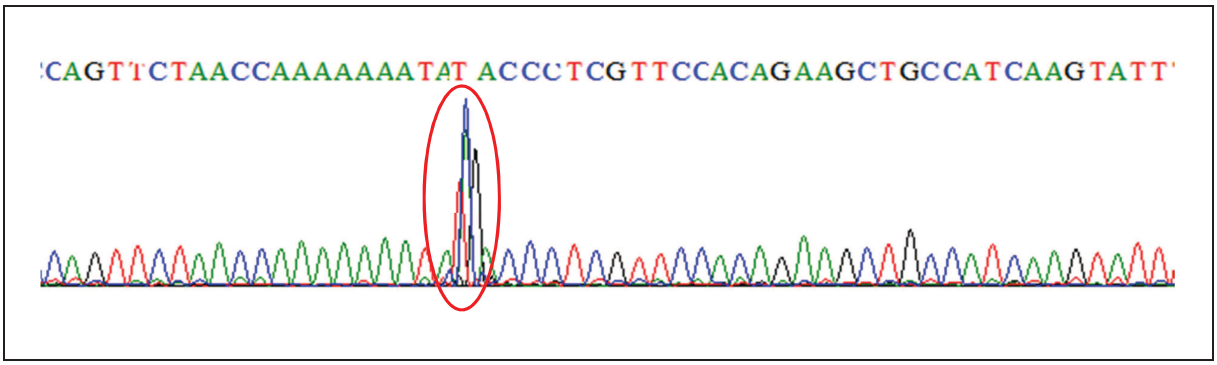
Example of electropherogram of sample: A08_13_7_3_E_064.ab1. The presence of a minor sequence, five nucleotides downstream, that resembles the major sequence, is presented. Three bases in each of the sequences are indicated here by the arrows that differ by five nucleotides in size are but identical in sequence composition.

On investigation, it was, however, determined that the N-5 artefacts had been caused by "bleed through" from other capillaries during electrophoresis. This type of artefact is seen especially on the 96 capillary instruments, as used in this study. The profiles seen in capillaries containing high signal (bright) samples often "bleed through" to capillaries with no signal or low signal. Since the samples that were run per batch were all from the same sequencing reaction, they were primed by the same primer and were therefore similar, which ruled out the cause being two different lengths of primers. The samples differed only at the points of mutation and looked like N-5 profiles due to "bleed through". The signal strength of adjacent samples was confirmed to be much higher than the sample indicated above, hence the cause was identified as "bleed through".

6.5.4.10 Spikes

A spike in the electropherogram was displayed by a short segment of sudden high peaks in contrast to the other peaks of normal signal intensity surrounding it. This is probably caused by the presence of matter in the capillary polymer that scatters the laser light when passing the detection window. This artefact was seldom detected in this investigation and easily addressed by re-electrophoresing the sample. An example is presented in Figure 6.16.

Figure 6.16 Example of spike peaks



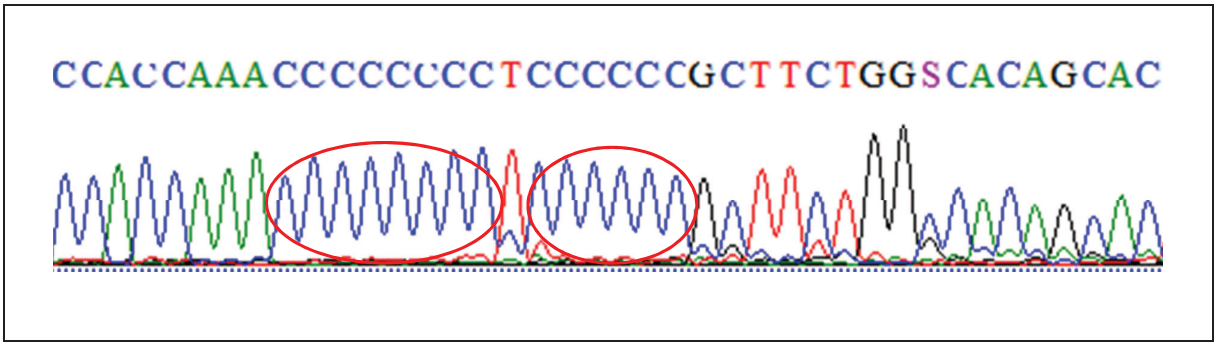
Example of electropherogram of sample: C09_26_3_2_E_005.ab1. The spike artefacts are indicated by the red circle.

6.5.4.11 Homopolymeric tracks

Homopolymeric regions consist of strings of the same bases that cause the polymerase enzyme to "slip", causing noisy data after this region due to multiple sequence peaks. It has been reported that the incidence of ambiguities after homopolymeric tracks is high (Salas *et al.*, 2005) and therefore these regions in the mtDNA sequences of this study were reverse-sequenced to address any possible ambiguities.

Four different homopolymer C regions were identified in the samples of this investigation. The first homopolymer was identified between nucleotide positions 303 and 315 in reference to the rCRS (Andrews *et al.*, 1999) and was present in three of the 50 samples investigated. An example is presented in Figure 6.17.

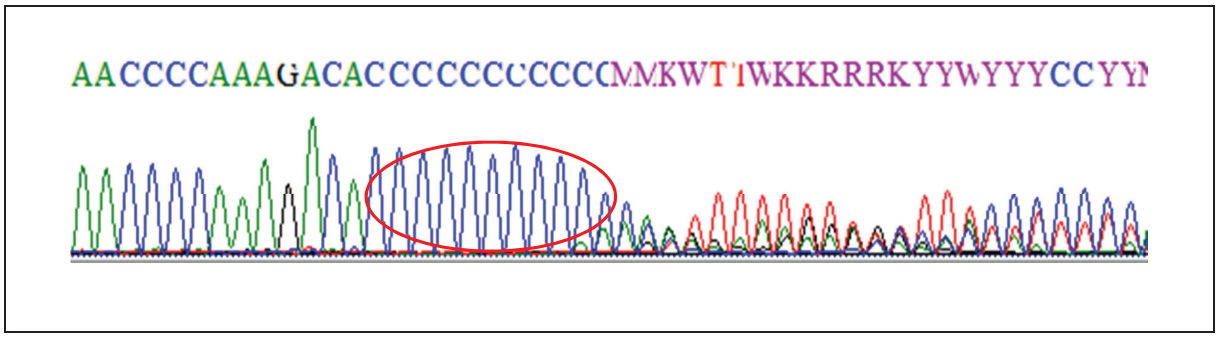
Figure 6.17 Homopolymer regions between nucleotide positions 303 and 315



Example of electropherogram of sample D12_47_1_2_E_008.ab1. The homopolymer regions are indicated by the red circles. The region after the homopolymer displayed noisy data because of the slippage of the Taq polymerase.

The second homopolymer was observed between nucleotide positions 568 and 573 and was present in four of the 50 samples in this investigation. An example is presented in Figure 6.18.

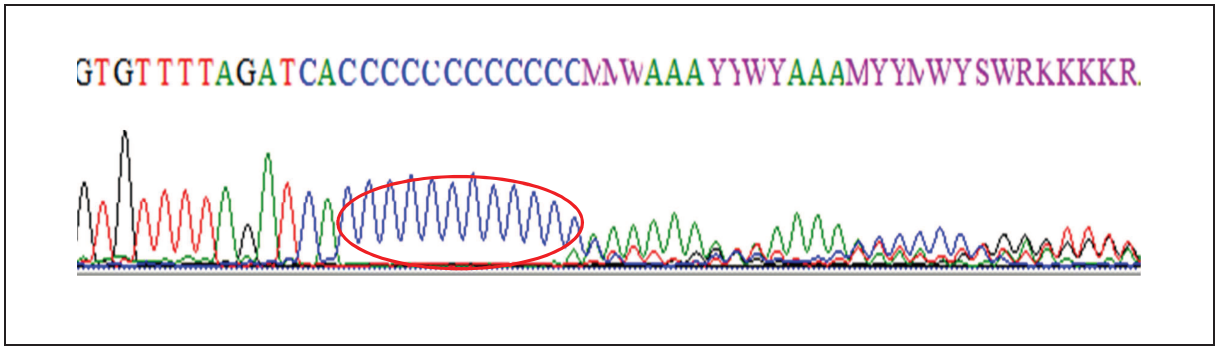
Figure 6.18 Homopolymer region between nucleotide positions 568 and 573



Example of electropherogram of sample H07_2_1_3_E_015.ab1. The homopolymer region is indicated by the red circle. The region after the homopolymer displayed noisy data because of the slippage of the Taq Polymerase.

The third homopolymer region was identified between nucleotide positions 957 and 966 and was present in three of the 50 samples in this investigation. An example is presented in Figure 6.19.

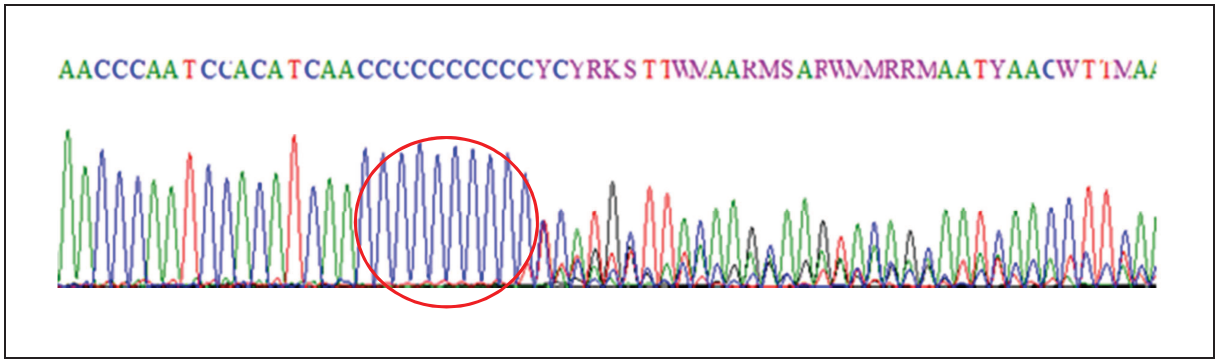
Figure 6.19 Homopolymer region between nucleotide positions 957 and 966



Example of electropherogram of sample E11_9_1_3_E_009.ab1. The homopolymer region is indicated by the red circle. The region after the homopolymer displayed noisy data because of the slippage of the Taq polymerase.

The fourth homopolymeric region was identified between nucleotide positions 16184 and 16193 and was present in seven of the 50 samples in this investigation. An example is presented in Figure 6.20.

Figure 6.20 Homopolymer region between nucleotide positions 16184 and 16193



Example of electropherogram of sample F01_43_1_1_E_001.ab1. The homopolymer region is indicated by the red circle. The region after the homopolymer displayed noisy data because of the slippage of the Taq polymerase.

All of the homopolymeric regions were determined in the control region of the mitochondrial genome. To overcome the issue of ambigious base calling after the homopolymer, reverse primers were used to sequence the complementary strand. The homopolymeric region gave rise to strand slippage in the reverse direction, as in the case of the forward primer, and overlap could only be established in the limited region of the homopolymer. An example of the contig overlap sequence is presented in Figure 6.21. To ensure that the homopolymeric regions were correctly called, 100% correct overlap was required before accepting the sequence data of those regions.

Figure 6.21 Sequence overlap of homopolymer region between forward and reverse primed sequences

Example of BioEdit software application used to construct contig sequences. The numbering at the top indicates the nucleotide positions of the sequence. Top sequence is the standard rCRS to which the sample DNA sequences are aligned. Sample 9_1_3 E consists of the amplified DNA of primer region 1 sequenced with sequencing primer 4 of sample 9 of the Tswana-speaking cohort of this investigation. Sample 9_1_4 R consists of the amplified DNA of primer region 1 sequenced with the reverse primer 4 of sample 9 of the Tswana-speaking cohort of this investigation. Sample 9_1_3 E and Sample 9_1_4 R both end in the homopolymer region starting at nucleotide position 957. An overlap in the homopolymer region between these two sequences is achieved in order to construct the contig sequence.

The reverse primers used for purposes of sequencing the complementary strand of the homopolymeric regions in order to obtain ambiguous sequence in the forward direction after the homopolymeric regions, are listed in Table 6.5. The primer names and sequence of the four primers were obtained from Maca-Meyer *et al.*, 2001. Since the homopolymeric artefacts were only present in PCR region 1, four reverse primers were chosen for this region only.

Table 6.5 Reverse primers used in this investigation

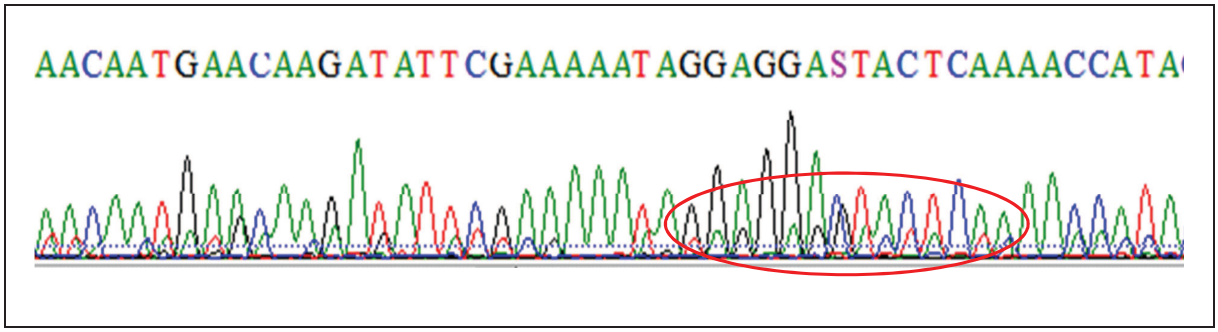
Primer region	Primer name	Primer sequence
1	R1:mtH16401	**5'-** tga ttt cac gga gga tgg tg **-3'**
	R2:mtH408	**5'-** tgt taa aag tgc ata ccg cca **-3'**
	R3:mtH945	**5'-** ggg agg ggg tga tct aaa ac **-3'**
	R4:mtH1487	**5'-** gta tac ttg agg agg gtg acg g **-3'**

R= reverse primer, mt = mitochondrial, H = heavy strand

6.5.4.12 Noisy data

Noisy data can be described as electropherograms with background peaks in addition to the actual peaks of the sample DNA as presented in Figure 6.22. Noise can be present throughout the electropherogram or only up to a certain point. When the noisy peaks are high in relation to the true peaks it is impossible to distinguish between the true sequence and the artefactual sequence. For this reason background noise in raw data makes it impossible to identify heteroplasmy or a new variant. Heteroplasmies are uncommon and if present in high frequencies can indicate sequencing problems (Salas *et al.*, 2005).

Figure 6.22 Example of noisy data



Example of electropherogram of a sample DNA that displays noisy data. The red circle indicates the peaks within peaks that are generally indicative of noisy data.

Noisy data are generally detected when the fluorescent signal of the sample DNA is low or alternatively when the fluorescent signal of the sample DNA is good in combination with the presence of the fluorescent signal from contaminants in the sample. The sequencing kit used in this investigation, however, was developed to prevent unnecessarily noisy data owing to the little spectral overlap caused by narrow emission spectra of the fluorescent dyes in the BigDye® terminators and BigDye® primers.

In terms of the sample, noisy data could be caused by contaminated samples or samples that contain multiple templates in the sequencing reaction. The sequencing reaction contributes to noisy data through the presence of dye-labelled and unlabelled reaction components that interfere with the electrophoretic separation and data analysis. Fluorescent signals from unincorporated dye-labelled terminators can obscure the desired signal of the extension products and interfere with base calling. Noisy data can also be generated by salts that interfere with the sequencing reaction, or problems related to capillary electrokinetic injection or electrophoresis. Expired reagents and failed capillary electrophoresis can also cause noisy data. In terms of the instruments, error and failure of instruments and software on many levels can cause noisy data as well as thermal cycling

and capillary electrophoresis failure. Other reasons include the collection of data with an incorrect run module and incorrect matrix files. In these cases, peaks within peaks are typical because of the inefficiency of the software to distinguish between the fluorescent signals that were detected (Applied Biosystems, 2009).

To prevent noisy data in this investigation, measures were taken to detect contamination of reagents and samples with foreign DNA by using negative controls in the PCR and sequencing reaction batches, inspecting amplification results for the presence of foreign DNA and adhering to strict laboratory practices that prevented contamination with foreign amplified DNA, which was discussed in detail in Chapter 5. Therefore the occurrence of noisy data in this study was treated as problems with the sequencing reactions and/or the electrophoretic process during the sequencing of the samples. Samples of this study that exhibited noisy data were therefore re-sequenced to eliminate the multiple possibilities that could have caused the noisy data.

6.5.4.13 Failed reactions

Failed reactions were characterised by high levels of noise and no well-defined peaks with signal strengths that were below the threshold of analysis i.e. less than 20 rfu. Dye peaks were often present as the only signal that was read in a failed reaction because of the absence of true DNA sequence peaks and electropherograms of failed reactions presented as flat lines.

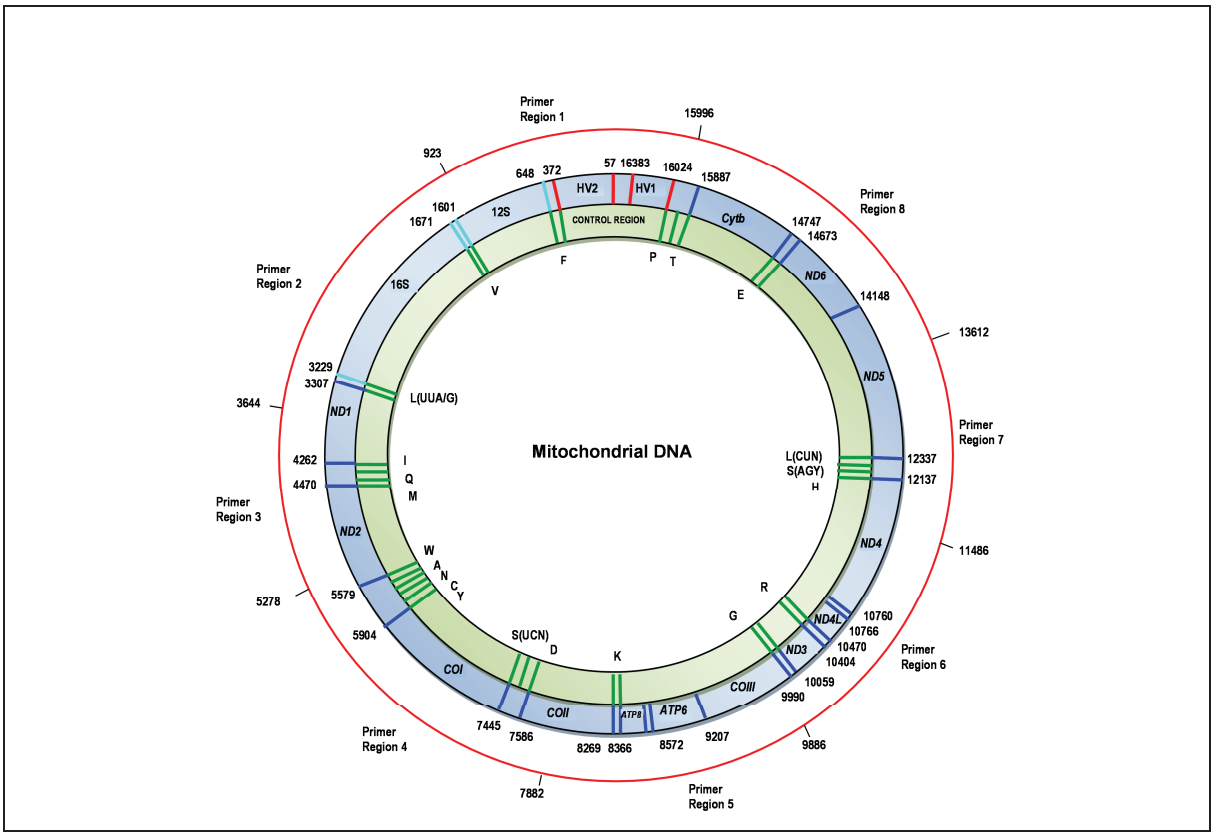
Possible causes for failed reactions are insufficient or contaminated templates, insufficient primer, old reagents, cycler failure, extension products being lost during reaction clean-up or not re-suspended properly, lane tracking failure or electrokinetic failure (Applied Biosystems, 2009). As discussed in Section 6.1, the procedures followed in the laboratory during the amplification of the DNA samples and setting up the sequencing reaction made it more likely that the occurrence of failed reactions were due to errors with the sample clean-up and electrophoresis process. Therefore, samples that failed were electrophoresed again and if the result was still not optimal, the sequencing set-up was repeated to rule out possible human error during the first set-up. In batches where samples failed on a large scale, the whole batch was investigated for an error in the sequencing reaction set-up or electrophoresis.

6.6 SEQUENCING RESULTS OF ALL PCR REGIONS

The full mitochondrial DNA sequences of all 50 Tswana individuals of this investigation were aligned with the rCRS and the sequence variants are reported at the nucleotide positions which occurred relevant to the rCRS. The L-strand base was reported according to the nomenclature guidelines of Tully *et al.* (2001).

The sequencing results are discussed as notated for each of the eight primer regions used in the PCR strategy of this investigation, as was explained in Section 6.1. For purposes of discussion of the sequencing results, the primer regions are defined according to the forward primers of the light strands, as indicated in Figure 6.23.

Figure 6.23 Primer regions and a map of the functional areas of mitochondrial DNA



Red = Control region. Blue = Coding region. **Light blue** = rRNA. **Green** = tRNA. **Circular green band** represents the L strand; **blue band** represents the H strand and the red line on the outer edge the demarcation of the primer regions. Primer regions indicated according to forward primer of the L strand. Numbers refer to base pair positions relative to rCRS (Andrews *et al.*, 199). HV1: Hypervariable segment 1; HV2: Hypervariable segment 2; 12S: 12S ribosomal RNA; 16S: 16S ribosomal RNA; *ND1*: NADH dehydrogenase subunit 1 gene; *COI*: Cytochrome c oxidase subunit I gene; *COII*: Cytochrome c oxidase subunit II gene; *ATP8*: ATP synthase F0 subunit 8 gene; *ATP6*: ATP synthase F0 subunit 6 gene; *COIII*: Cytochrome c oxidase subunit III gene; *ND2*: NADH dehydrogenase subunit 2 gene; *ND3*: NADH dehydrogenase subunit 3 gene; *ND4L*: NADH dehydrogenase subunit 4L gene; *ND4*: NADH dehydrogenase subunit 4 gene; *ND5*: NADH dehydrogenase subunit 5 gene; *ND6*: NADH dehydrogenase subunit 6 gene; *Cytb*: Cytochrome b gene; Control region, including displacement loop; HV1: Hypervariable segment 1; F: tRNA phenylalanine; V: tRNA valine; L(UUA/G): tRNA leucine 1; I: tRNA isoleucine; Q: tRNA glutamine; M: tRNA methionine; W: tRNA tryptophan; A: tRNA alanine; N: tRNA asparagine; C: tRNA cysteine; Y: tRNA tyrosine; S: tRNA serine 1; D: tRNA aspartic acid; K: tRNA lysine; G: tRNA glycine; R: tRNA arginine; H: tRNA histidine; S(UCN): tRNA serine2; L(CUN): tRNA leucine 2; E: tRNA glutamic acid; T: tRNA threonine; P: tRNA proline. Adapted from MITOMAP: A Human Mitochondrial Genome Database. http://www.mitomap.org, 2011. Accessed 16 Feb 2011.

Sequence variation was notated in the context of the functional locations of the coding region, which included the genes for proteins and coding regions for the tRNAs and rRNAs, as well as the non-coding or control region, which included the hypervariable segments 1 and 2. Functional positions for the respective regions are provided in Table 6.6.

Table 6.6 Functional locations of mitochondrial DNA

Locus code	Locus name	Sequence position
CR/D loop	Control region / D-loop	16024 – 576
HV1	Hypervariable segment 1	16024 – 16383
HV2	Hypervariable segment 2	57 – 372
F	tRNA phenylalanine	577 – 647
12S	12S ribosomal RNA	648 – 1601
V	tRNA valine	1602 – 1670
16S	16S ribosomal RNA	1671 – 3229
L(UUA/G)	tRNA leucine 1	3230 – 3304
ND1	NADH dehydrogenase subunit 1	3307 – 4262
I	tRNA isoleucine	4263 – 4331
Q	tRNA glutamine	4329 – 4400
M	tRNA methionine	4402 – 4469
ND2	NADH dehydrogenase subunit 2	4470 – 5579
W	tRNA tryptophan	5512 – 5579
A	tRNA alanine	5587 – 5655
N	tRNA asparagine	5657 – 5729
C	tRNA cysteine	5761 – 5826
Y	tRNA tyrosine	5826 – 5891
COI	Cytochrome c oxidase subunit 1	5904 – 7445
S(UCN)	tRNA serine 1	7446 – 7514
D	tRNA aspartic acid	7518 – 7585
COII	Cytochrome c oxidase subunit 2	7586 – 8269
K	tRNA lysine	8259 – 8364
ATP8	ATP synthase F0 subunit 8	8366 – 8572
ATP6	ATP synthase F0 subunit 6	8527 – 9207
COIII	Cytochrome c oxidase subunit 3	9207 – 9990
G	tRNA glycine	9991 – 10058
ND3	NADH dehydrogenase subunit 3	10059 – 10404
R	tRNA arginine	10405 – 10469
ND4L	NADH dehydrogenase subunit 4L	10470 – 10766
ND4	NADH dehydrogenase subunit 4	10760 – 12137
H	tRNA histidine	12138 – 12206
S(AGY)	tRNA serine 2	12207 – 12265
L(CUN)	tRNA leucine 2	12266 – 12336
ND5	NADH dehydrogenase subunit 5	12337 – 14148
ND6	NADH dehydrogenase subunit 6	14149 – 14673
E	tRNA glutamic acid	14674 – 14742

Table 6.6 Continued…

Locus code	Locus name	Sequence position
Cytb	Cytochrome b	14747 – 15887
T	tRNA threonine	15888 – 15953
P	tRNA proline	15956 – 16023

CR = Control region / D-loop here refers to the non-coding region between positions 16024 – 576. Locus codes and names are the same as used in Figure 6.23 as reported in the MITOMAP database, www.mitomap.org ; 12S: 12S ribosomal RNA; 16S: 16S ribosomal RNA; *ND1*: NADH dehydrogenase subunit 1; *COI*: Cytochrome c oxidase subunit I; COII: Cytochrome c oxidase subunit II; *ATP8*: ATP synthase F0 subunit 8; *ATP6*: ATP synthase F0 subunit 6; *COIII*: Cytochrome c oxidase subunit III; *ND2*: NADH dehydrogenase subunit 2; *ND3*: NADH dehydrogenase subunit 3; *ND4L*: NADH dehydrogenase subunit 4L; *ND4*: NADH dehydrogenase subunit 4; *ND5*: NADH dehydrogenase subunit 5; *ND6*: NADH dehydrogenase subunit 6; *Cytb*: Cytochrome b; Control region, including displacement loop; HV1: Hypervariable segment 1; F: tRNA phenylalanine; V: tRNA valine; L(UUA/G): tRNA leucine 1; I: tRNA isoleucine; Q: tRNA glutamine; M: tRNA methionine; W: tRNA tryptophan; A: tRNA alanine; N: tRNA asparagine; C: tRNA cysteine; Y: tRNA tyrosine; S: tRNA serine 1; D: tRNA aspartic acid; K: tRNA lysine; G: tRNA glycine; R: tRNA arginine; H: tRNA histidine; S(AGY): tRNA serine2; L(CUN): tRNA leucine 2; E: tRNA glutamic acid; T: tRNA threonine; P: tRNA proline. Sequence positions correspond to the rCRS positions (Andrews *et al.*, 1999).

Sequence variance was reported as any nucleotide that differed from the nucleotide at the same position in the rCRS or as an insertion or deletion of a nucleotide that corresponded to the rCRS. These nucleotide substitutions were classified as either transitions or transversions, of which the transversions were expected to be in the minority. It was therefore expected that most substitutions present in the Tswana sequences would be transitions followed by transversions and indels (insertions or deletions). A high number of transversion or indel occurrences were regarded as suspect and were evaluated for error.

A novel mutation is a mitochondrial DNA mutation or polymorphism that has not been reported previously (Bandelt *et al.*, 2006b). To determine the novelty of alterations in this study, the MITOMAP (Brandon *et al.*, 2005; http://www.mitomap.org/), Uppsala mtDB database (Ingman and Gyllensten, 2006; http://www.genpat.uu.se/mtDB/) and PhyloTree (Van Oven and Kayser, 2009, http:/www.phylotree.org/) public databases, GenBank®[1] and the internet were searched. GenBank® was searched with the aid of a study regarding nucleotide variation in the mtDNA (Pereira *et al.*, 2009).

It was expected that the nonsynonymous substitutions could have affected the function of a gene in the coding region and therefore the nucleotide substitutions and indels in the protein-coding genes were indicated as synonymous or nonsynonymous polymorphisms. It was further expected that the nucleotide positions of alterations in the regions of the tRNAs and rRNAs could have affected the secondary structure of the tRNAs and rRNAs, and the function of the tRNAs and rRNAs, and these were therefore indicated in the discussion of the sequence results. It is not the aim of this investigation to identify

1 GenBank® is a registered trademark of the US Department of Health and Human Services, Maryland, USA.

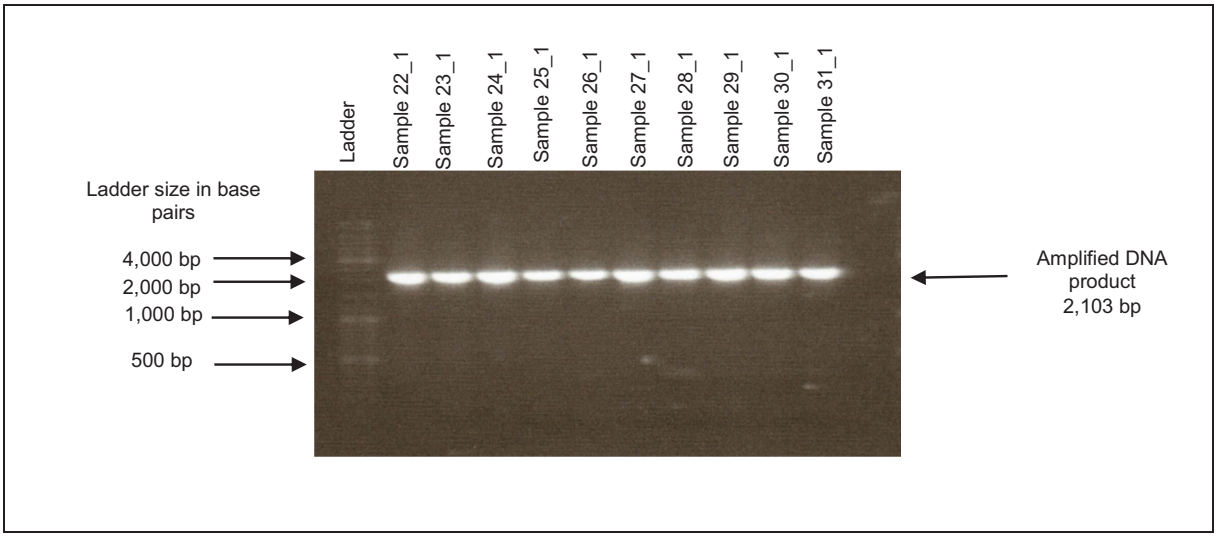
pathogenic mutations present in the mitochondrial DNA of the Tswana individuals, but the possibility of pathogenic mutations was evaluated and discussed on a preliminary basis.

6.6.1 Primer region 1

Primer region 1 consisted of 1,496 base pairs (bp) starting at position 15,996 and ending at position 923. This region included the control region, which was located between positions 16024 to 576 and consisted of the hypervariable region 1 (HV1) at positions 16024 to 16383 and hypervariable region 2 (HV2) at positions 57 to 372. It also contained the coding region for tRNA phenylalanine (F) at positions 577 to 647 and a 275 bp segment of the 12S rRNA coding region.

The region outlined above was amplified by PCR, as discussed in Section 5.4, and the PCR products were electrophoresed on an agarose gel to ascertain the quality of the product. A representative example of the mtDNA products for primer region 1, as visualised by the UVIvue ultraviolet transilluminator, is presented in Figure 6.24.

Figure 6.24 Photographic representation of the amplified mtDNA product of primer region 1



Photograph of the agarose gel on which the mtDNA amplified product was electrophoresed at 100 volts (V) and 50 mAmperes (mA) for 30 minutes as discussed in Section 5.5; ladder = FastRuler™[1] High Range DNA Ladder (Fermentas) of range 100 – 10,000 bp; included in the first lane of the gel; sample names refer to the Tswana-speaking individuals of this investigation.

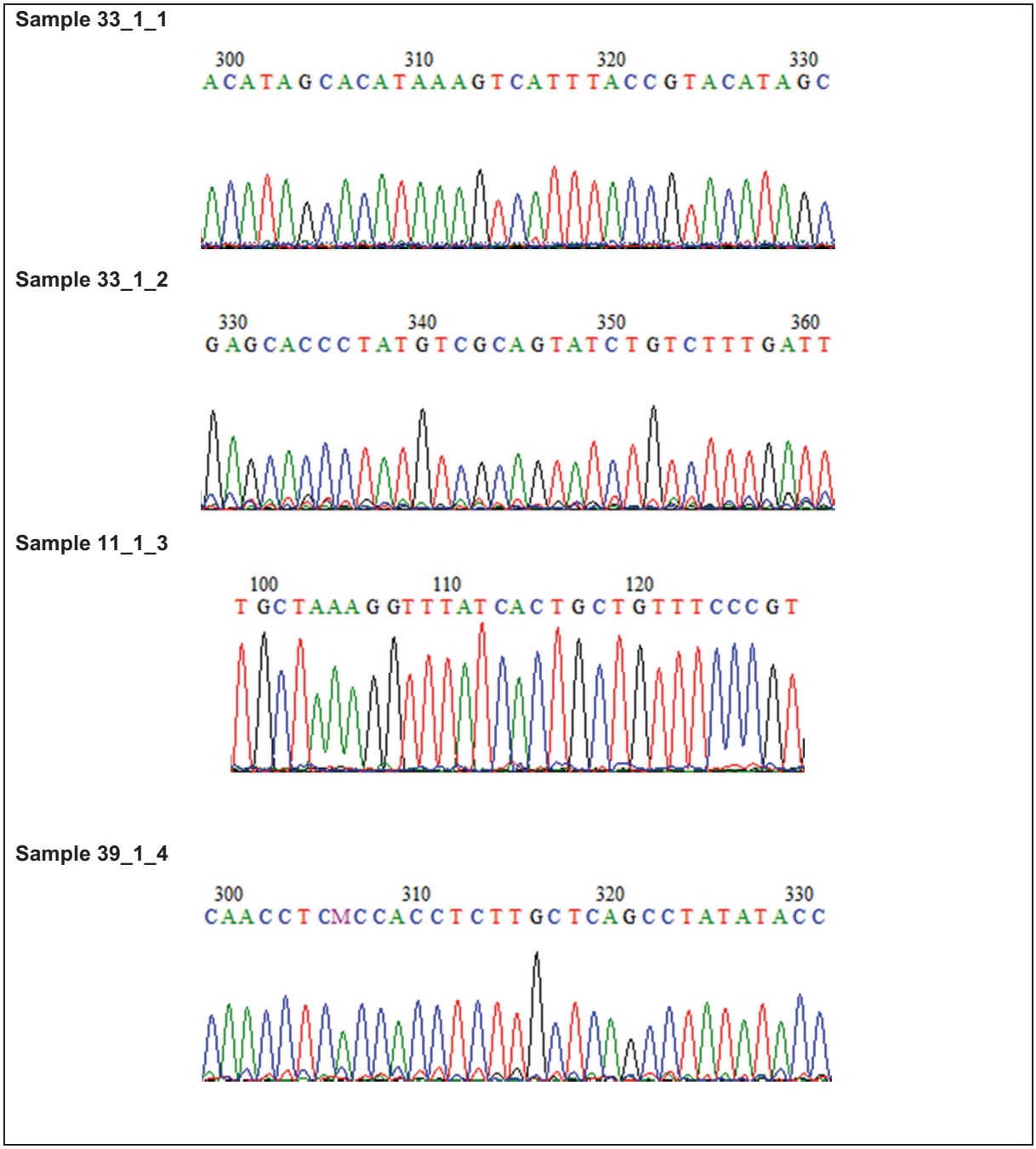
The full mitochondrial genome of the Tswana-speaking individuals of this investigation was sequenced by using the BigDye®[2] Terminator v3.1 Cycle Sequencing Kit and subsequently

1 FastRuler™ is a registered trademark of Fermentas International, Inc., Ontario, Canada.

2 BigDye® Terminator v3.1 Cycle Sequencing Kit is a registered trademark of Applied Biosystems, Foster City, CA, USA.

analysed by manual editing of the electropherograms by using the BioEdit software version 7.0.5.2 (Hall, 2001). A representative example of the electropherograms for primer region 1 is presented in Figure 6.25.

Figure 6.25 Representative electropherograms of the sequence generated for primer region 1 using the forward primers 1-4



Examples of electropherogram data with peaks depicting nucleotides in the sequence region of primer 1; A = adenine; T = thymine; C = cytosine; G = guanine; numbering at the top of the electropherogram represents the numbering of the nucleotides as a sequenced fragment before alignment with the rCRS and therefore does not correspond to the nucleotide positions of primer region 1.

Homopolymeric C regions were identified in some of the samples of this investigation between nucleotide positions 16184 and 16193, nucleotide positions 303 and 315 and nucleotide positions 568 and 573. Examples of mtDNA sequences in this investigation that display homopolymeric regions are discussed and presented in Section 6.5.4.11. Reverse primers have been used in this region to resolve those sequences and representative samples of electropherograms of these sequenced regions are presented in Figure 6.26.

Figure 6.26 Electropherograms of sequences that were sequenced by reverse primers in primer region 1

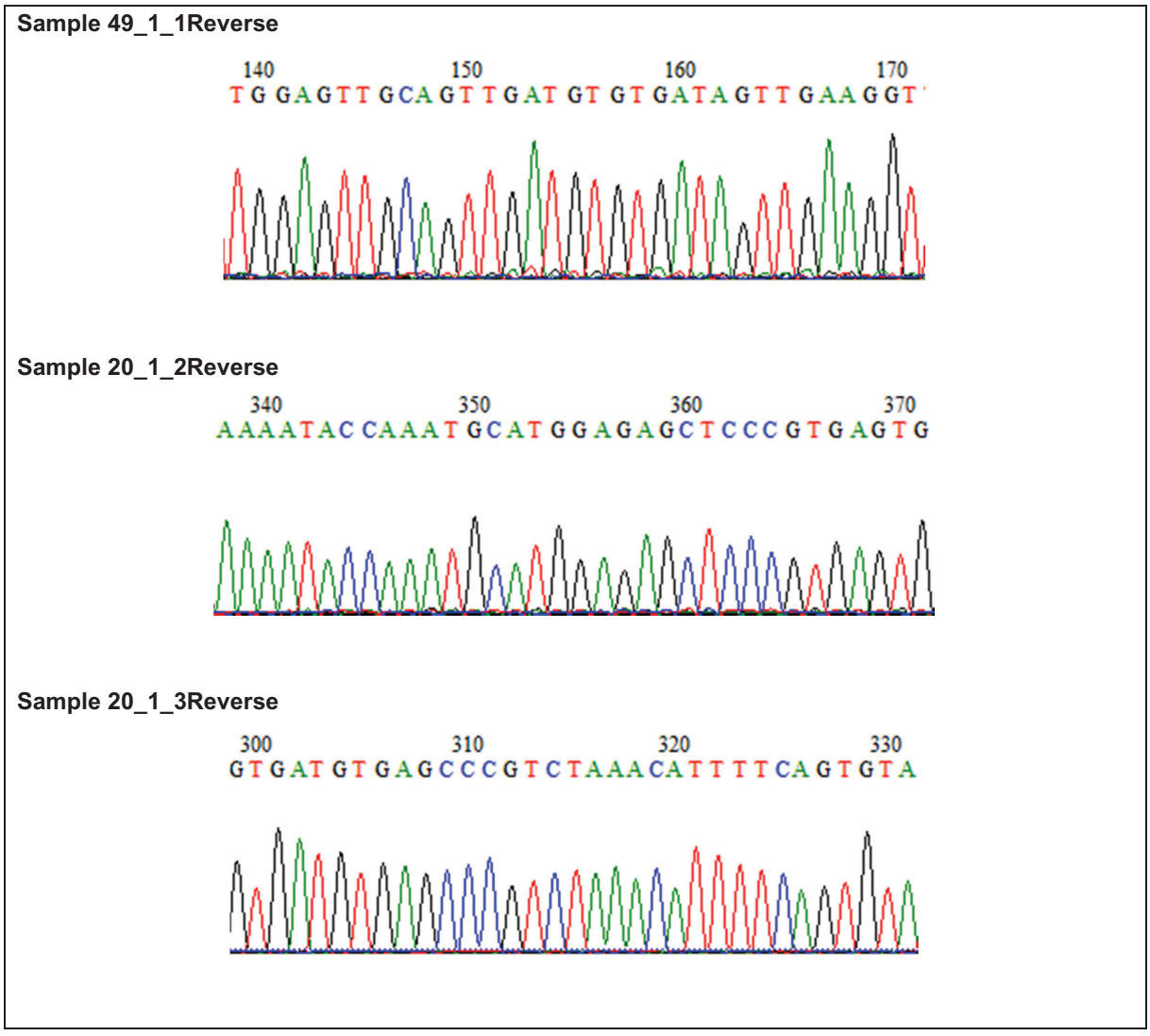
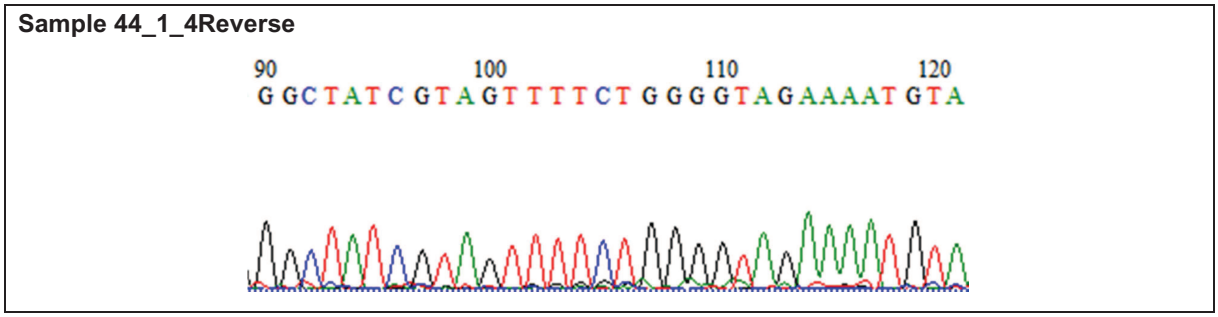


Figure 6.26 Continued…



Examples of electropherogram data with peaks depicting nucleotides in the sequence region of primer 1; A = adenine; T = thymine; C = cytosine; G = guanine; numbering at the top of the electropherogram represents the nucleotide positions.

6.6.1.1 Sequence alterations observed in primer region 1

A region of the 12S rRNA is partly in primer region 2 and for the purposes of this discussion, the sequence variation in the 12S rRNA segment will be dealt with as a single unit, and thus this sequence variation will be indicated in Section 6.6.2, in which primer region 2 is discussed. The sequence alterations determined for primer region 1 are presented in Table 6.7.

Table 6.7 Sequence alterations observed between the complete mitochondrial DNA of the Tswana individuals included in this study and the rCRS in primer region 1

Position	Sequence alteration	Gene/ region	Frequency	Reference
16037	A-G	Control region	1	Behar *et al*., 2008
16086	T-C	Control region	1	Behar *et al*., 2008
16093	T-C	Control region	1	Behar *et al*., 2008
16111	C-T	Control region	1	Behar *et al*., 2008
16124	T-C	Control region	1	Salas *et al*., 2004
16129	G-A	Control region	24	Behar *et al*., 2008
16148	C-T	Control region	8	Behar *et al*., 2008
16166	A-C	Control region	1	Ingmann *et al*., 2000
16168	C-T	Control region	3	Salas *et al*., 2004
16169	C-T	Control region	1	Behar *et al*., 2008
16172	T-C	Control region	11	Batini *et al*., 2011
16174	C-T	Control region	2	Behar *et al*., 2008
16181	delA	Control region	3	Thangaraj *et al*., 2009
16182	delA	Control region	3	Thangaraj *et al*., 2009
16184	C-T	Control region	1	Behar *et al*., 2008
16187	C-T	Control region	36	Behar *et al*., 2008
16188	C-G	Control region	8	Behar *et al*., 2008
16189	T-C	Control region	44	Batini *et al*., 2011
16192	C-T	Control region	1	Behar *et al*., 2008

Table 6.7 Continued…

Position	Sequence alteration	Gene/ region	Frequency	Reference
16193	insC	Control region	1	Bandelt and Parson, 2008
16209	T-C	Control region	1	Behar *et al*., 2008
16212	A-G	Control region	10	Behar *et al*., 2008
16214	C-T	Control region	2	Behar *et al*., 2008
16223	C-T	Control region	46	Behar *et al*., 2008
16230	A-G	Control region	35	Behar *et al*., 2008
16234	C-T	Control region	6	Quintana-Murci *et al.*, 2008
16239	C-T	Control region	7	Behar *et al*., 2008
16242	C-T	Control region	1	Behar *et al*., 2008
16243	T-C	Control region	26	Behar *et al*., 2008
16249	T-C	Control region	2	Quintana-Murci *et al.*, 2008
16265	A-C	Control region	2	Quintana-Murci *et al.*, 2008
16266	C-T	Control region	1	Bandelt *et al.*, 2001(b)
16266	C-G	Control region	2	Behar *et al*., 2008
16266	C-A	Control region	2	Behar *et al*., 2008
16274	G-A	Control region	5	Behar *et al*., 2008
16278	C-T	Control region	22	Behar *et al*., 2008
16286	C-G	Control region	2	Quintana-Murci *et al.*, 2008
16286	C-T	Control region	3	Bandelt *et al.*, 2001(b)
16290	C-T	Control region	8	Behar *et al*., 2008
16291	C-G	Control region	1	Behar *et al*., 2008
16293	A-G	Control region	3	Behar *et al*., 2008
16294	C-T	Control region	20	Behar *et al*., 2008
16300	A-G	Control region	5	Behar *et al*., 2008
16309	A-G	Control region	9	Behar *et al*., 2008
16311	T-C	Control region	39	Behar *et al*., 2008
16319	G-A	Control region	1	Maca-Meyer *et al.,* 2001
16320	C-T	Control region	11	Behar *et al*., 2008
16325	T-C	Control region	1	Herrnstadt *et al.*, 2002
16325	delT	Control region	1	Salas *et al.*, 2004
16360	C-T	Control region	2	Behar *et al*., 2008
16368	T-C	Control region	1	Behar *et al*., 2008
16390	G-A	Control region	19	Behar *et al*., 2008
16399	A-G	Control region	1	Behar *et al*., 2008
16519	T-C	Control region	39	Mishmar *et al.*, 2003
16527	C-T	Control region	2	Quintana-Murci *et al.*, 2008
64	C-T	Control region	4	Behar *et al*., 2008
73	A-G	Control region	42	Olivieri *et al.*, 2006
93	A-G	Control region	8	Behar *et al*., 2008
95	A-C	Control region	3	Behar *et al*., 2008
125	T-C	Control region	1	Vigilant *et al.*, 1989
127	T-C	Control region	1	Vigilant *et al.*, 1989

Table 6.7 Continued…

Position	Sequence alteration	Gene/ region	Frequency	Reference
143	G-A	Control region	4	Behar *et al*., 2008
146	T-C	Control region	39	Behar *et al*., 2008
150	C-T	Control region	7	Behar *et al*., 2008
151	C-T	Control region	2	Behar *et al*., 2008
152	T-C	Control region	37	Behar *et al*., 2008
182	C-T	Control region	3	Behar *et al*., 2008
185	G-A	Control region	5	Behar *et al*., 2008
186	C-A	Control region	2	Quintana-Murci *et al*., 2008
188	A-G	Control region	2	Olivieri *et al*., 2006
189C	A-C	Control region	2	Quintana-Murci *et al*., 2008
189G	A-G	Control region	11	Behar *et al*., 2008
195	T-C	Control region	39	Behar *et al*., 2008
198	C-T	Control region	14	Behar *et al*., 2008
199	T-C	Control region	33	Behar *et al*., 2008
200	A-G	Control region	2	Olivieri *et al*., 2006
204	T-C	Control region	6	Behar *et al*., 2008
207	G-A	Control region	66	Behar *et al*., 2008
211	A-G	Control region	1	Kong *et al*., 2006
236	T-C	Control region	88	Behar *et al*., 2008
247	G-A	Control region	38	Behar *et al*., 2008
263	A-G	Control region	22	Behar *et al*., 2008
267	T-C	Control region	1	Behar *et al*., 2006
297	A-G	Control region	2	Quintana-Murci *et al*., 2008
309.1	insC	Control region	22	Gonder *et al*., 2007
315.1	insC	Control region	50	Behar *et al*., 2008
316	G-A	Control region	8	Behar *et al*., 2008
385	A-G	Control region	2	Behar *et al*., 2008
408	T-A	Control region	1	Behar *et al*., 2008
456	C-T	Control region	33	Behar *et al*., 2008
471	T-C	Control region	2	Behar *et al*., 2008
498	delC	Control region	22	Gonder *et al*., 2007
499	G-A	Control region	1	Just *et al*., 2008
523	delA	Control region	23	Brandstätter *et al*., 2004
524	delC	Control region	23	Behar *et al*., 2008
524.1	insA	Control region	1	Bandelt and Parson, 2008
524.2	insC	Control region	1	Bandelt and Parson., 2008
524.3	insA	Control region	1	Bandelt and Parson, 2008
524.4	insC	Control region	1	Bandelt and Parson, 2008
538	A-G	Control region	1	Torroni *et al*., 2006
567	A-C	Control region	1	Gonder *et al*., 2007
576	**A-T**	**Control region**	**1**	**Current investigation**
597	C-T	F (D-loop)	9	Behar *et al*., 2008

Table 6.7 Continued…

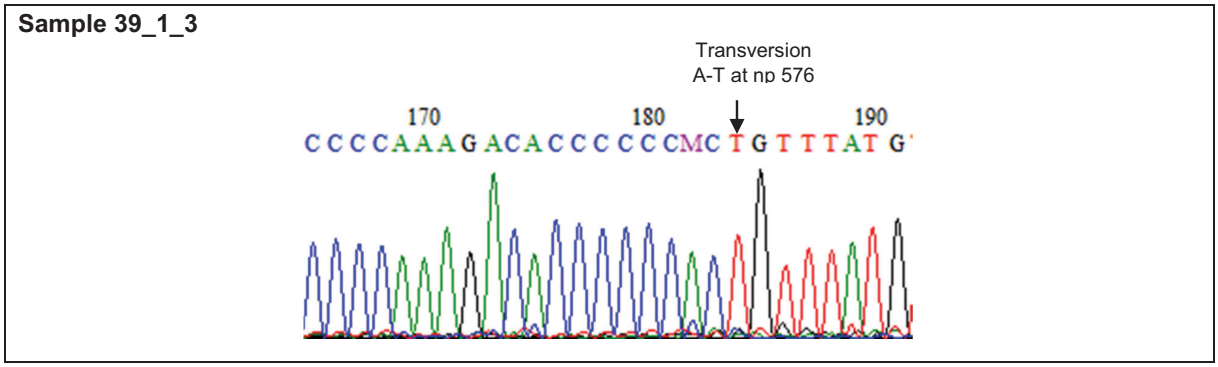
Position	Sequence alteration	Gene/ Region	Frequency	Reference
143	G-A	Control region	4	Behar *et al*., 2008
146	T-C	Control region	39	Behar *et al*., 2008

Transitions indicated in blue and transversions indicated in red. The frequency is indicated as the number of times the sequence variation was observed within the total Tswana cohort of 50 individuals. Deletions are indicated by "del" followed by the sequence position and insertions indicated by "ins" followed by the sequence position. The structural region where the sequence variation occurs within rRNA and tRNA coding regions is indicated in brackets following the region name; F (D-loop) = tRNA phenylalanine in the displacement loop region; novel mutations are indicated in **bold**.

A total of 104 sites of sequence variation in the control region (positions 16024 – 576) were observed in the Tswana dataset of 50 individuals. One variant allele was observed for most of the nucleotide positions (np) except for np16266 where three variant alleles were observed and np16286, np16325, np189 and np524 where two variant alleles were observed at each site. Seventy-seven transitions, 13 transversions and 14 indels were observed, making the transitions the most prevalent type of alteration observed, as is expected (Pereira *et al.*, 2009). This region displayed a transversion:transition count ratio (Lutz-Bonengel *et al.*, 2003) of 1:6, which correlated with a study undertaken by Pereira *et al*. (2009).

One novel alteration was observed in primer region 1, which was present at np576 and consisted of a transversion A-T alteration. It did not fall within the regions of hypervariable length variation or any mutational hotspots. This sequence variation was observed in only one individual of the Tswana dataset (TS_4106), as presented in Figure 6.27.

Figure 6.27 Representative electropherogram of the sequence data generated indicating a transversion at np 576



Sample name 39_1_3 refers to sample TS_4106_39, primer region 1, sequence primer 3; sequence region presented containing a transversion alteration before alignment with the rCRS. Therefore the numbering presented above is not in accordance with the alteration position.

The sequence alterations at np 16181, np 16182, np 211 and np 267, as indicated in Table 6.7, were only observed in non-African individuals. The sequence alterations at np 1618,

np 16182 and np 211 were observed in individuals of Asian origin and the sequence alteration at np 267 was observed in an individual of Jewish origin, as presented in Table 6.8.

Table 6.8 Observed sequence alterations in individuals that did not belong to the L macrohaplogroup

Position	Sequence alteration	Frequency	Population	Reference
16181	delA	3	Asian population	Thangaraj *et al.*, 2009
16182	delA	3	Asian population	Thangaraj *et al.*, 2009
211	A-G	1	East Asian population	Kong *et al.*, 2006
267	T-C	1	Jewish population	Behar *et al.*, 2006

The frequency is indicated as the number of times the sequence variation was observed in the total Tswana dataset of 50 individuals. Deletions are indicated by "del" followed by the type of nucleotide.

The possibility that the alterations were novel in the context of the L macrohaplogroups had to be considered. Since there were two well-described regions of length variations in the control region around the long C-tracts in the regions adjacent to np 16189, and np 16181 had been identified as a nucleotide position of high variability and was therefore not regarded as a valid haplogroup defining site, the mutations observed at np 16181 and np 16182 were not deemed to be novel mutations without further investigation into the occurrence of these two deletions (Bandelt *et al.* , 2009; Van Oven and Kayser, 2009).

The two transitions at np 211 and np 267 were respectively present in two different Tswana-speaking individuals of this investigation. The electropherograms containing the sequence data of the two transitions in two different individuals are presented in Figure 6.28.

Figure 6.28 Representative electropherograms of the sequence data generated indicating transitions at np 211 and np 267

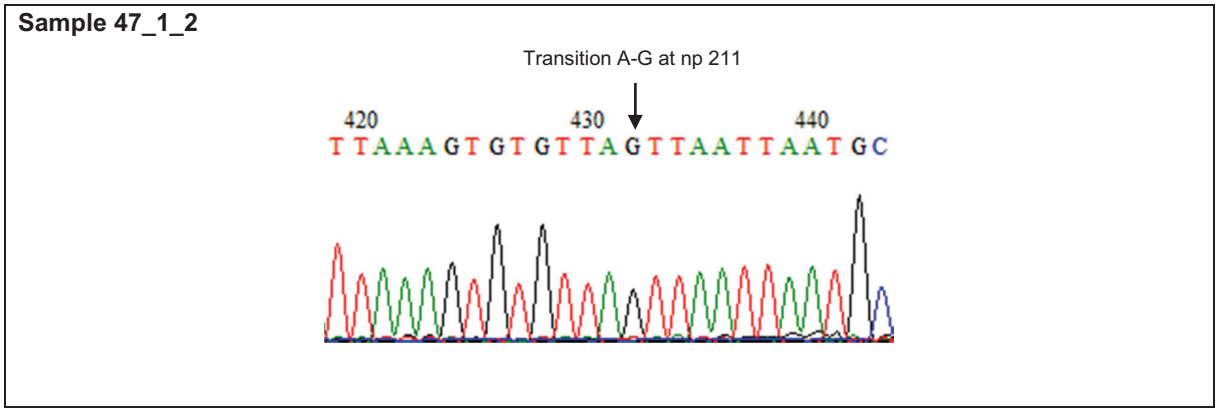
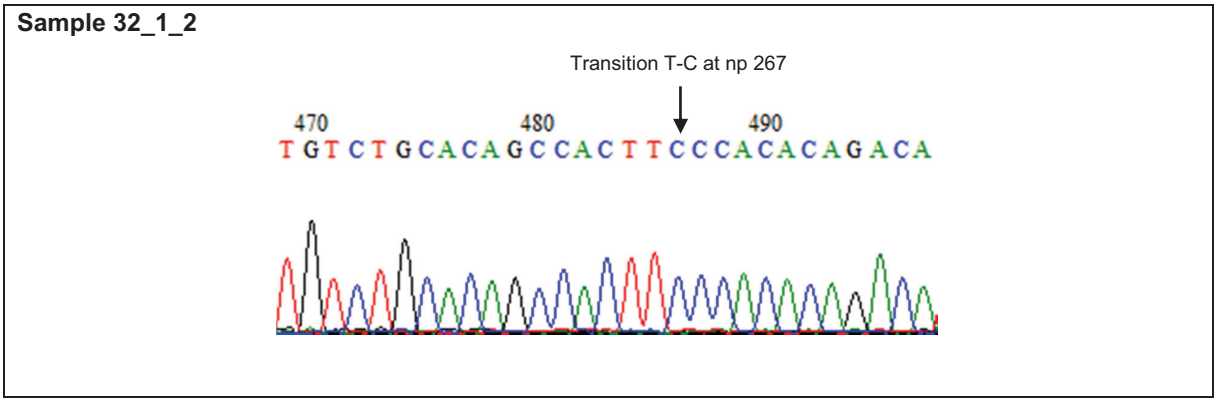


Figure 6.28 Continued…

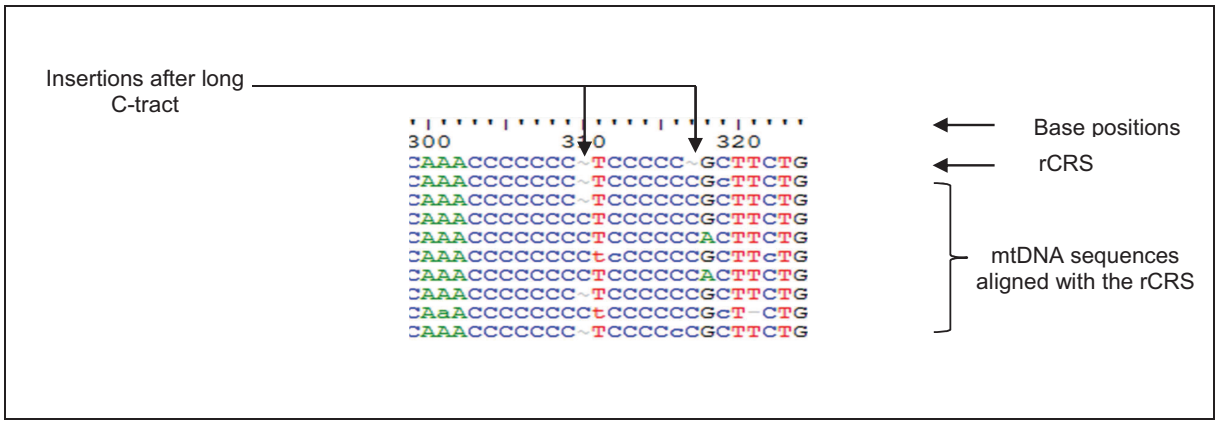


Sample name 47_1_2 refers to sample TS_5083_47, primer region 1, sequence primer 2; sample name 32_1_2 refers to sample TS_4056_32, primer region 1, sequence primer 2; sequence regions presented here containing transitions before alignment with rCRS. Therefore the numbering presented above is not in accordance with the nucleotide position of the alteration.

As was evident from the electropherograms, the peak morphology was acceptable in terms of quality and the electropherograms did not display any signs of background noise or artefacts that could have provided a reason for doubting the accuracy of the sequence results. Both of these sequence alterations were regarded as haplogroup-defining sites in the PhyloTree classification system (Van Oven and Kayser, 2009) for haplogroups H1o and M29 and were therefore associated with a range of substitutions that were specific to haplogroups H and M. These substitutions were not associated with any of the haplogroup L lineages and have not been identified in any studies of populations that were of African origin. The alterations further did not occur in regions that have been reported to contain high incidences of mutational hotspots or in homopolymer regions. Therefore they were regarded as novel mutations for the L haplogroup lineages.

Length variability was observed in the region of np 309 – np 315. A 309.1insC occurred in 22 of the Tswana-speaking individuals of this investigation and a 315.1insC occurred in all of the individuals of the Tswana dataset. These two length variations were demonstrated in an example of Tswana-speaking individuals of this investigation who harboured these alterations in Figure 6.29.

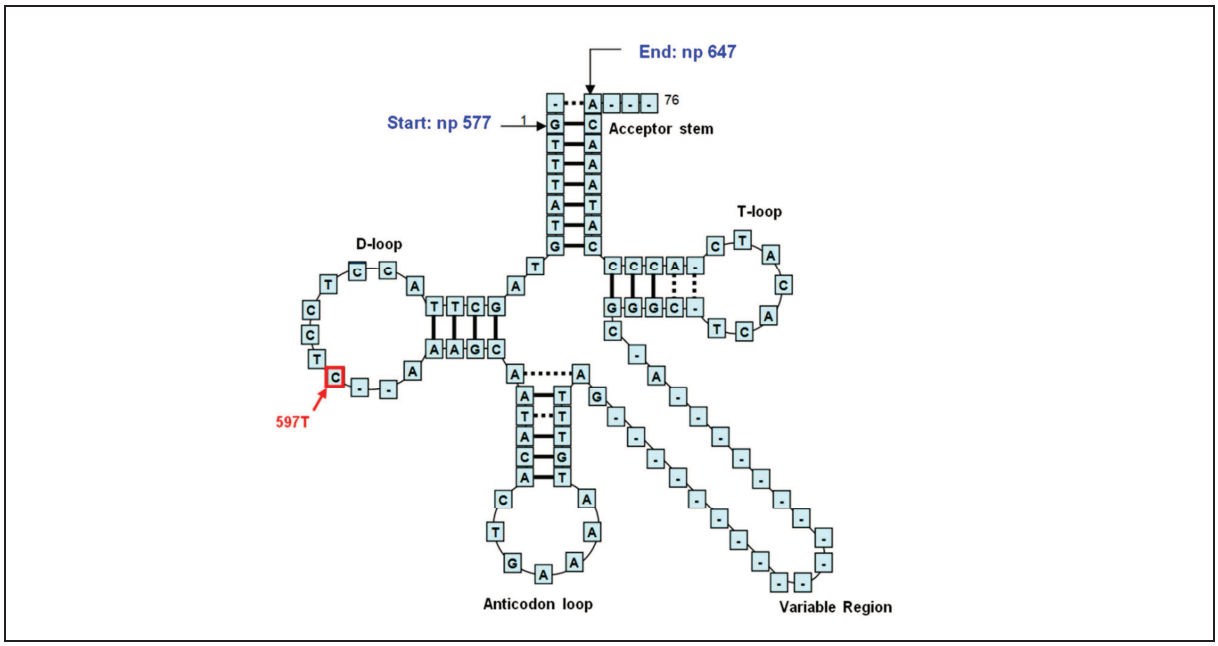
Figure 6.29 Length variation between np 309 and np 315



Electropherogram view of alignment of mtDNA sequences with the rCRS.

Numerous deletions and an insertion were observed within the dinucleotide tract between np 515 and np 524. A 523delA and 524delC were observed in 23 of the individuals in the dataset and can be ascribed to the presence of the AC dinucleotide repeats that preceded it. The deletion at np 523 and np 524 has not been described previously in African populations. As with the length variations, this type of sequence variation should not be used in phylogenetic analysis of a dataset in view of its high variability (Salas *et al.*, 2007). A transition was observed in nine individuals at np 597 of the coding region for tRNA phenylalanine (F) in the D-loop structure.

Figure 6.30 Structure of the tRNA phenylalanine (F) and observed sequence variation of the Tswana-speaking individuals of this investigation



Structure of a typical tRNA phenylalanine (F) molecule; ▬100% Watson-Crick pairs;▪▪ 100% mismatches; numbers alongside the tRNA structure indicate nucleotide positions within the tRNA molecule and not the numbering according to the rCRS; nucleotide starting and ending positions of the tRNA according to the rCRS indicated in **blue** ink; nucleotide position indicated in **red** ink refers to the nucleotide variation observed in the Tswana-speaking individuals of this investigation. From Jühling *et al.*, 2009; Pereira *et al.*, 2009.

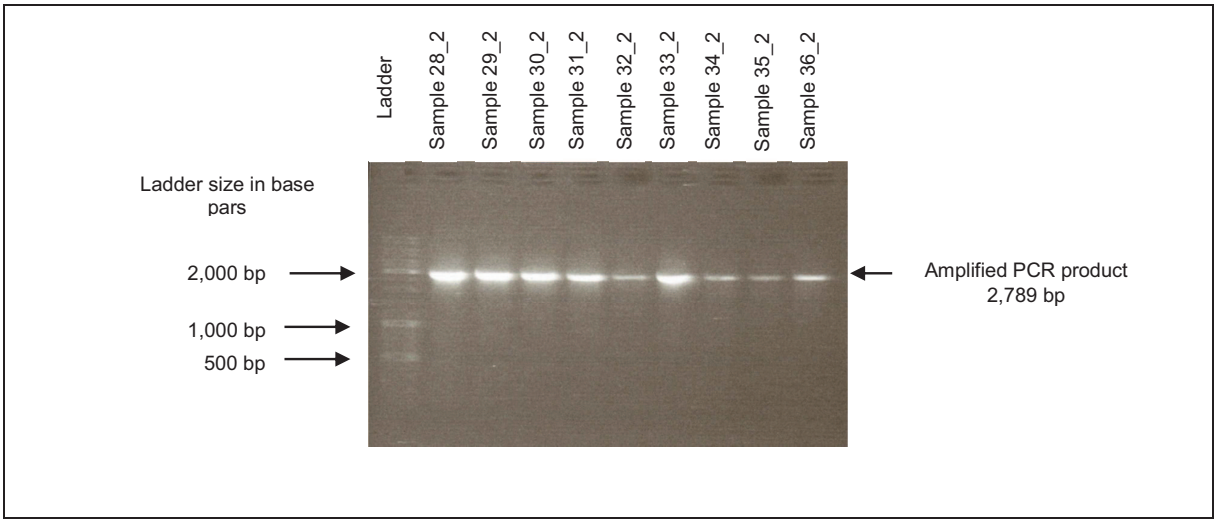
The np 597 transition alteration that was observed in the Tswana-speaking individuals of this investigation was a 100% pyrimidine site in the consensus human tRNA phenylalanine (F) sequence (Jühling *et al.*, 2009) and not reported as a pathogenic mutation. The np 597 has been reported to exhibit recurrent mutations (Kivisild *et al.*, 2006) and it was therefore not unexpected to find sequence variation at this position in the Tswana cohort of this investigation.

6.6.2 Primer region 2

Primer region 2 consists of 2,721 bp starting at nucleotide position 924 and ending at nucleotide position 3644. This region contains the coding region for a segment of the 12S ribosomal RNA gene that stretches from position 924 to 1601, the coding region of the tRNA valine (V) gene at nucleotide positions 1602 to 1670, the coding region of the 16S rRNA gene at nucleotide positions 1671 to 3229 and the tRNA leucine 1 (L(UUA/G)) gene at nucleotide positions 3230 to 3304. It further contains a 337 bp segment of the coding region for the *ND1* gene.

As was the case with all the primer regions, the region outlined above was amplified by PCR, as discussed in Section 5.4, and the PCR products were electrophoresed on an agarose gel to ascertain the quality of the product. A representative example of the mtDNA products for primer region 2, as visualised by the UVIvue ultraviolet transilluminator, is presented in Figure 6.31.

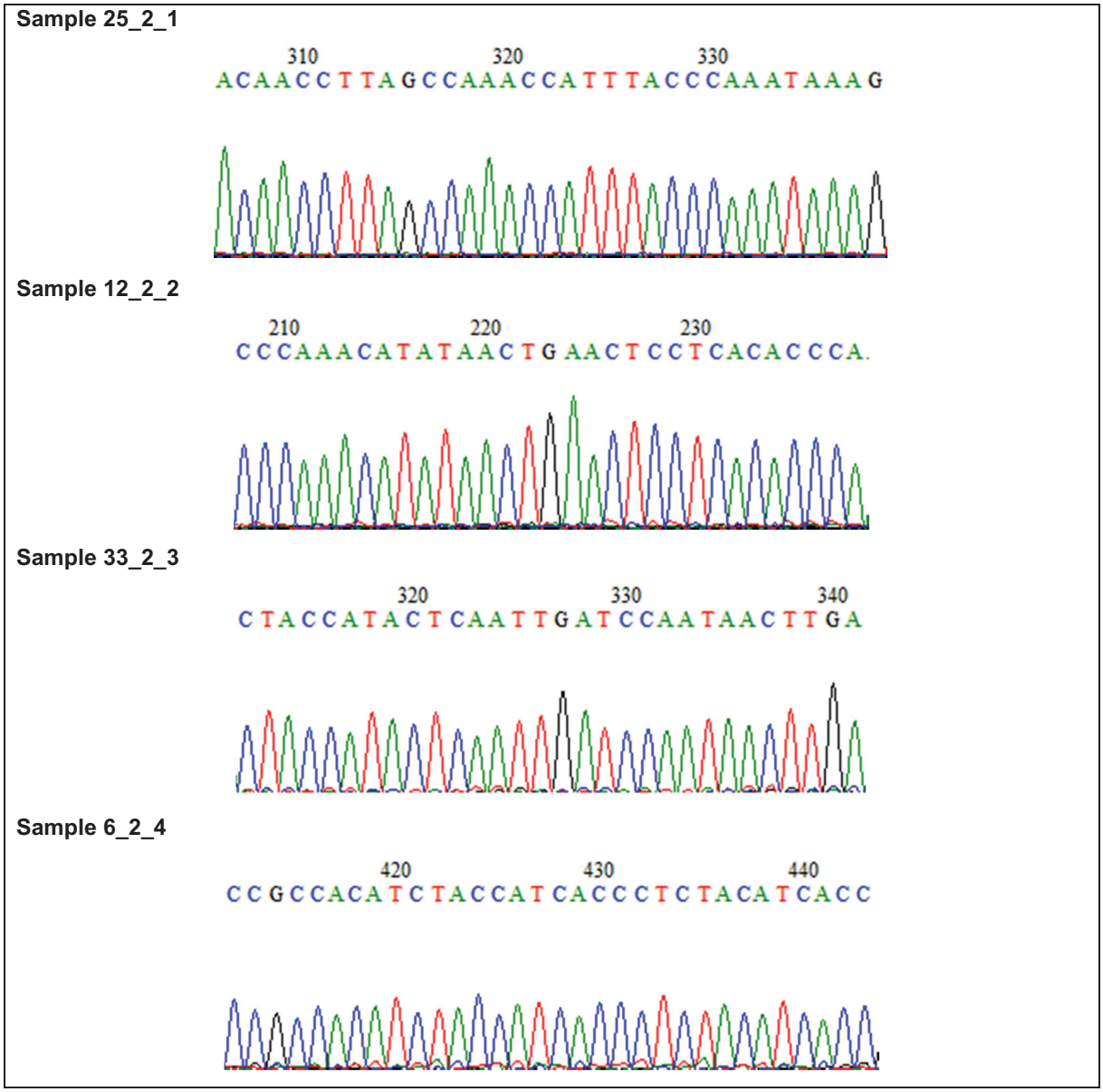
Figure 6.31 Photographic representation of the amplified mtDNA product of primer region 2



Photograph of the agarose gel on which the mtDNA amplified product was electrophoresed at 100 volts (V) and 50 mAmperes (mA) for 30 minutes as discussed in Section 5.5; ladder = FastRuler™ High Range DNA Ladder (Fermentas) of range 100 – 10,000 bp; included in the first lane of the gel; sample names refer to the Tswana-speaking individuals of this investigation.

Representative examples of the electropherograms of the sequence data generated with the BigDye®[1] Terminator v3.1 Cycle Sequencing Kit for the primer region 2 sequences are presented in Figure 6.32. These results were viewed and edited using the BioEdit software version 7.0.5.2 (Hall, 2001).

Figure 6.32 Representative electropherograms of the sequence generated for primer region 2 using the forward primers 1-4



Examples of electropherogram data with peaks depicting nucleotides in the sequence region of primer region 2; A = adenine; T = thymine; C = cytosine; G = guanine; numbering at the top of the electropherogram represents the numbering of the nucleotides as a sequenced fragment before alignment with the rCRS and therefore does not correspond to the nucleotide positions of primer region 2.

1 BigDye® Terminator v3.1 Cycle Sequencing Kit is a registered trademark of Applied Biosystems, Foster City, CA, USA.

6.6.2.1 Sequence alterations observed in primer region 2

The coding region of the 12S rRNA gene stretches from np 648 – np 1601. The primer region 2 segment starts at np 924 and therefore does not contain the mtDNA segment from the start of the 12S rRNA coding region. To enable discussion of the 12S rRNA coding region as a single unit, the 276 base pairs of this gene contained in the primer region 1 segment are included in this section for discussion. The same applies to the *ND1* gene, which is partially covered in primer region 3 at np 4262, causing a 618 base pair region of the gene to fall within the primer region 3 segment. For the purpose of this discussion, the *ND1* gene will be dealt with as a single unit, in Section 6.9.3. The sequence variation present in primer region 2, inclusive of the 12S rRNA coding region, is presented in Table 6.9.

Table 6.9 Sequence alterations observed between the complete mitochondrial DNA of the Tswana individuals included in this study and the rCRS in primer region 2

Position	Sequence alteration	Gene/ region	Frequency	Reference
709	G-A	12S (Non-stem)	1	Kivisild *et al.*, 2006
719	G-A	12S (Non-stem)	12	Kivisild *et al.*, 2006
721	T-C	12S (Non-stem)	5	Behar *et al.*, 2008
750	A-G	12S (Non-stem)	50	Mishmar *et al.*, 2003
769	G-A	12S (Non-stem)	48	Olivieri *et al.*, 2006
825A	T-A	12S (Non-stem)	38	Ingman *et al.*, 2000
850	T-C	12S (Stem)	1	Ingman *et al.*, 2000
921	T-C	12S (Non-stem)	1	Kivisild *et al.*, 2006
961	T-C	12S (Non-stem)	2	Behar *et al.*, 2008
980	T-C	12S (Non-stem)	1	Hernstadt *et al.*, 2002
1008	A-G	12S (Non-stem)	1	Behar *et al.*, 2008
1018	G-A	12S (Non-stem)	48	Mishmar *et al.*, 2003
1048	C-T	12S (Non-stem)	38	Mishmar *et al.*, 2003
1243	T-C	12S (Stem)	10	Ingman *et al.*, 2000
1415	G-A	12S (Non-stem)	1	Mishmar *et al.*, 2003
1420	T-C	12S (Non-stem)	2	Ingman *et al.*, 2000
1438	A-G	12S (Stem)	23	Mishmar *et al.*, 2003
1503	G-A	12S (Non-stem)	1	Kivisild *et al.*, 2006
1676	A-G	16S (Non-stem)	1	Behar *et al.*, 2008
1719	G-A	16S (Non-stem)	1	Kivisild *et al.*, 2006
1900	A-G	16S (Non-stem)	1	Hernstadt *et al.*, 2002
2245	A-G	16S (Non-stem)	8	Ingman *et al.*, 2000
2352	T-C	16S (Non-stem)	1	Mishmar *et al.*, 2003
2416	T-C	16S (Non-stem)	10	Mishmar *et al.*, 2003

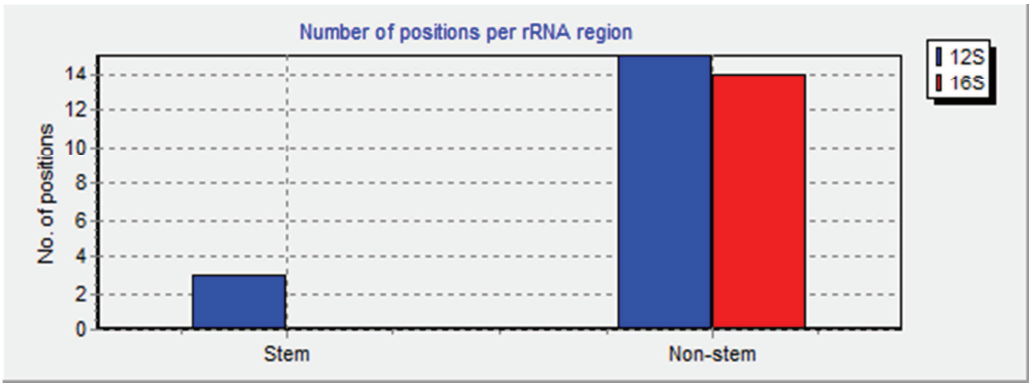
Table 6.9 Continued…

Position	Sequence alteration	Gene/ region	Frequency	Reference
2706	A-G	16S (Non-stem)	38	Mishmar *et al.*, 2003
2755	A-G	16S (Non-stem)	5	Ingman *et al.*, 2000
2758	G-A	16S (Non-stem)	38	Mishmar *et al.*, 2003
2789	C-T	16S (Non-stem)	10	Mishmar *et al.*, 2003
2836	C-A	16S (Non-stem)	1	Ingman *et al.*, 2000
2885	T-C	16S (Non-stem)	38	Kivisild *et al.*, 2006
3010	G-A	16S (Non-stem)	2	Kivisild *et al.*, 2006
3202	T-C	16S (Non-stem)	1	Mishmar et al., 2003

Transitions are indicated in blue and transversions are indicated in red. The frequency is indicated as the number of times the sequence variation is observed within the total Tswana dataset of 50 individuals. The structural region where the sequence variation occurs within rRNA and tRNA coding regions is indicated in brackets following the region name. 12S = 12 Svedberg units; 16S = 16 Svedberg units

The Tswana individuals of this investigation display 32 variable nucleotide sites in primer region 2. The sequence alterations consist of 30 transitions and two transversions presented in blue ink and red ink respectively in Table 6.9 and thus display a transversion:transition count ratio (Lutz-Bonengel *et al.*, 2003) of 1:15. As expected, the presence of transversions is much lower than that of transitions. No indels have been observed. Three of the sequence alterations within the 12S rRNA sequence are present in the stem region while the other 15 alterations are present in the non-stem region. All of the 16S rRNA sequence alterations are present in the non-stem region, as indicated in Figure 6.33. The low number of mutations in the stem regions have been expected and could be explained by the fact that the stem regions of rRNAs are generally reported to be conserved in order to maintain the secondary structure of the molecule and therefore its functionality (Pereira *et al.*, 2009).

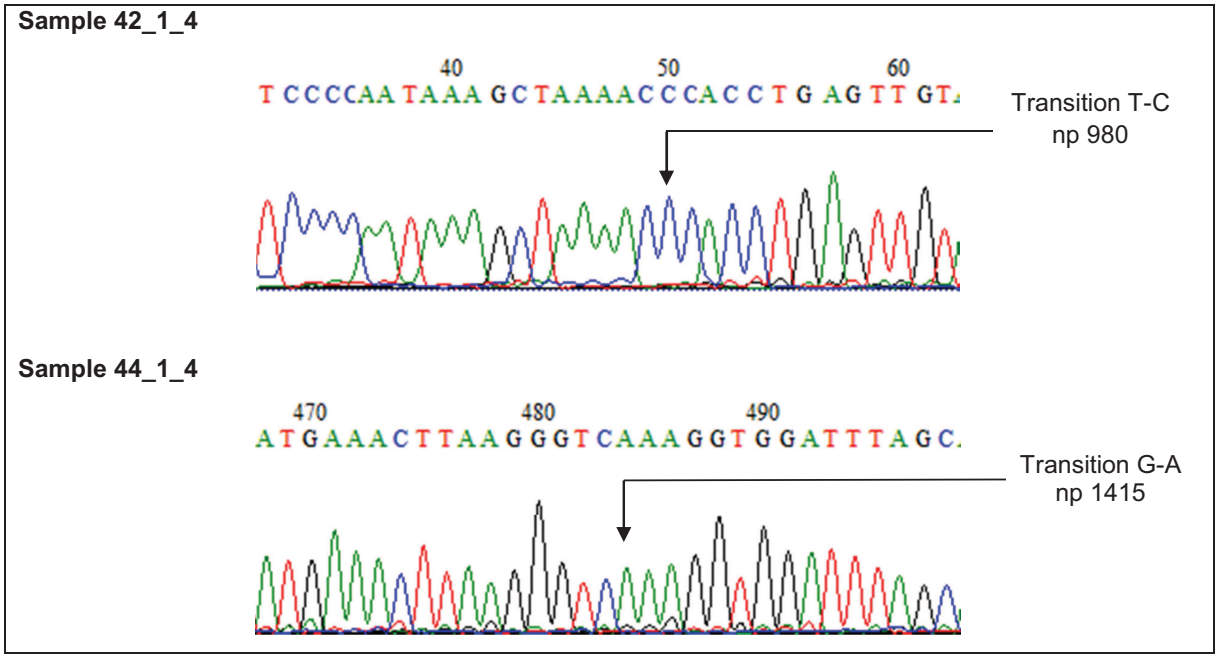
Figure 6.33 Locations of the sequence alterations within the 12S rRNA and 16S rRNA sequences of the Tswana cohort of this investigation



Graph constructed by the mtDNA-GeneSyn tool as described by Pereira *et al.* (2009); number of positions refers to the number of sequence alterations detected in the Tswana-speaking individuals of this investigation.

No novel mutations have been observed in primer region 2 of the Tswana cohort of this investigation. Three sequence alterations at np 980, np 1415 and np 3202 respectively, were, however only reported in individuals who belonged to other haplogroups than the L haplogroup and who did not reside in or originate from Africa. The sequence alterations at np 980 and np 1415 consist of transitions, i.e. T-C and G-A respectively, in the non-stem region of the 12S rRNA and are cited as haplogroup-defining for haplogroups M61, M27a and U7 and haplogroups C5d1 and X2a2 respectively by the PhyloTree classification system (Van Oven and Kayser, 2009). The electropherograms of sample TS_5060, which contains an alteration at np 980, and sample TS_5063, which contains an alteration at np 1415, do not display any signs of background noise, artefacts or bad peak morphologies and it has been assumed that the sequence data do not display sequence errors, nor are these influenced by sequencing artefacts. The sequence alterations at np 980 and np 1415 are therefore regarded as novel sequence variants within the L haplogroup lineages.

Figure 6.34 Representative electropherograms of the sequence data generated indicating transitions at np 980 and np 1415

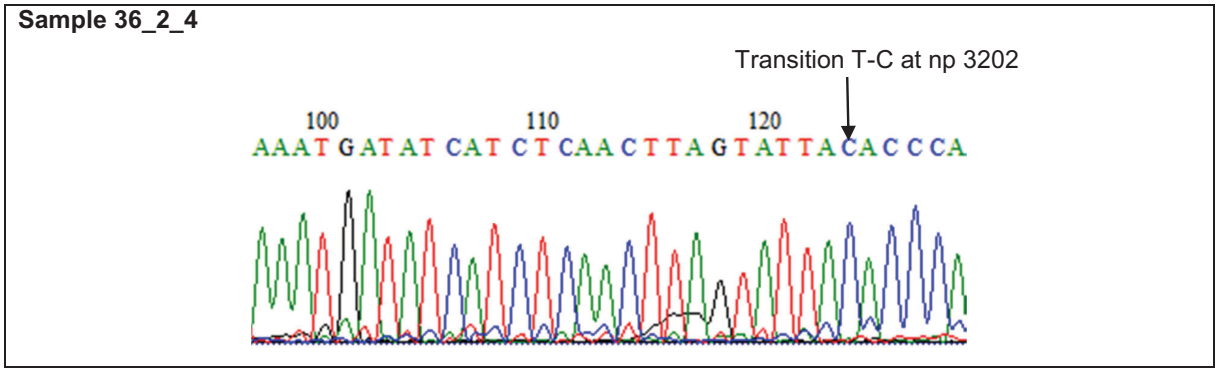


Sample name 42_1_4 refers to sample TS_5060_42, primer region 1, sequence primer 4; sample name 44_1_4 refers to TS_5063_44, primer region 1, sequence primer 4; sequence regions presented here containing the transitions have not been aligned with the rCRS, therefore numbering presented above is not in accordance with alteration position as determined by comparison to the rCRS.

The sequence alteration at np 3202 consists of a transition in the non-stem region of the 16S rRNA coding region of sample TS_4083 and is cited as a haplogroup-defining sequence variant for haplogroup A2k by the PhyloTree classification system (Van Oven and Kayser, 2009). The electropherogram of sample TS_4083 is presented in Figure 6.35

and although it displays low levels of background noise, the peak morphology of the transition at np 3202 is good and it is therefore assumed that the transition has not been caused by sequence error or sequencing artefacts. Based on this observation, the transition at np 3202 was also regarded as a novel sequence alteration within the L haplogroup lineages.

Figure 6.35 Representative electropherograms of the sequence data generated indicating transitions at np 3202



Sample name 36_2_4 refers to sample TS_4083_36, primer region 2, sequence primer 4; sequence region presented containing transition before alignment with rCRS, therefore numbering presented above is not in accordance with alteration position as determined by comparison to the rCRS.

Three highly variable mutational hotspots (Kivisild *et al.*, 2006) are observed in primer region 2 of the Tswana-speaking individuals of this investigation. They are all G-A transitions and occur at np 709, np 1719 and np 3010. The transition at np 709 occurs in the 12S rRNA region and the other two occur in the 16S rRNA region.

Four sequence alterations that have been reported to have pathological associations have been observed in the 12S rRNA coding region and three observed in the 16S rRNA. These sequence alterations and their associations are presented in Table 6.10.

Table 6.10 Reported mtDNA sequence alterations with pathological associations within primer region 2

Locus	Sequence alteration	Number of individuals	Disease	Reference
12S rRNA	T721C	5	Possibly LVNC-associated	Tang *et al.*, 2010
	T850C	1	Possibly LVNC-associated	Tang *et al.*, 2010
	T921C	1	Possibly LVNC-associated	Tang *et al.*, 2010
	T961C	2	Possibly LVNC-associated DEAF	Li *et al.*, 2005 Yao *et al.*, 2006 Mkaouar-Rebai *et al.*, 2008 Lu *et al.*, 2010

Table 6.10 Continued…

16S rRNA	T2352C	1	Possibly LVNC-associated	Tang *et al.*, 2010
	A2755G	5	Possibly LVNC-associated	Tang *et al.*, 2010
	G3010A	2	Cyclic Vomiting Syndrome with migraine	Boles *et al.*, 2009 Zaki *et al.*, 2009

Sequence alterations are displayed as the position at which the mutation occurred with the wild type nucleotide indicated in front of the np and the mutant type nucleotide indicated after the np; the number of individuals = number of Tswana-speaking individuals of this investigation that displayed the mutation; disease associations were reported in one or more publications and thus these alterations have been considered as possibly pathological and are reported as such in MITOMAP; LVNC = left ventricular noncompaction; DEAF = maternally inherited DEAFness or aminoglycoside-induced DEAFness. Adapted from MITOMAP: A Human Mitochondrial Genome Database. http://www.mitomap.org, 2011.

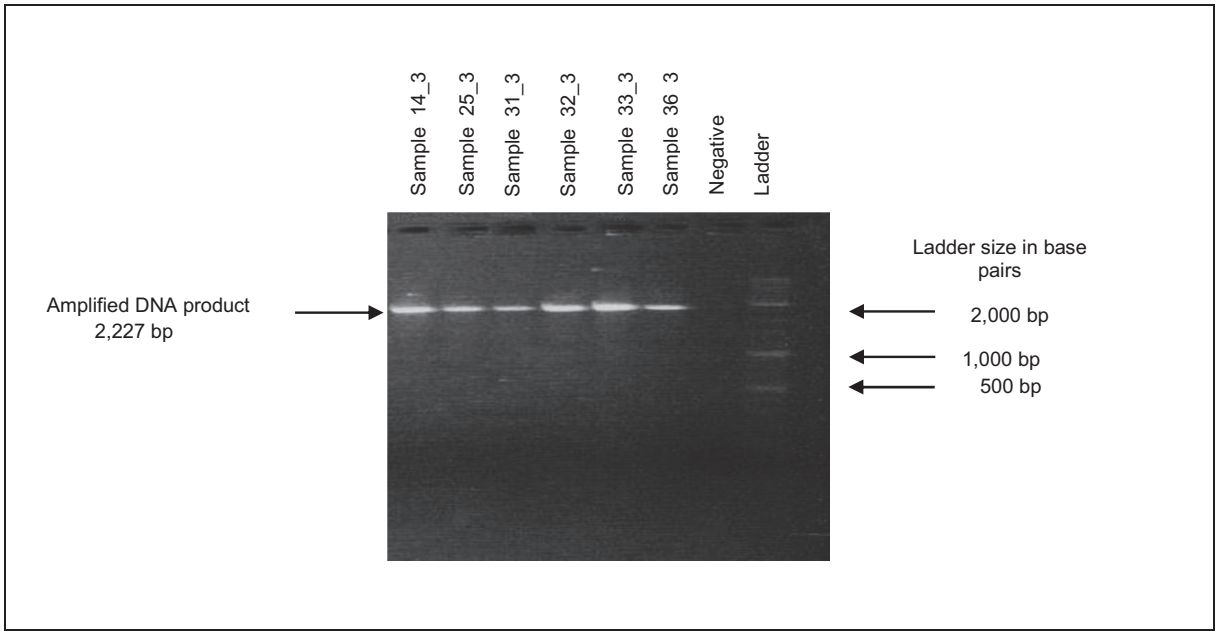
Six of the disease-associated sequence alterations observed in the Tswana-speaking individuals of this investigation are associated with the left ventricular noncompaction (LVNC) syndrome. The np 850 mutation has been described as extremely rare in the mtDNA and could reportedly affect the structure of the 12S rRNA (Tang *et al.*, 2010). The sequence alteration at np 921 is also present in the stem region of the 12S rRNA but has not been reported to change the structure of the 12S rRNA. The other sequence alterations at np 721, np 961 and np 2755 are situated in the loop regions of the 12S rRNA and 16S rRNA structures, which would probably not affect the structures of the rRNA molecules. Although the sequence alteration at np 2352 is rarely observed in LVNC patients, it has been reported in patients that belong to haplogroups L1 and L3. The np 2755 has also been reported in patients belonging to L1c and L0 (Tang *et al.*, 2010). The association of the sequence alteration at np 961 with maternally inherited DEAFness or aminoglycoside-induced DEAFness is uncertain (MITOMAP, 2011).

6.6.3 Primer region 3

Primer region 3 starts at np 3644 and ends at np 5278. It contains a 618 base pair segment of the *ND1* gene, which ranges from np 3307 to np 4262. It further contains the coding regions for tRNA isoleucine (I) gene (np 4263-np 4331), tRNA glutamine (Q) gene (np 4329-np 4400) and tRNA methionine (M) gene (np 4402-np 4469) and a 233 base pair fragment of the *ND2* gene, which stretches from np 4470-np 5579.

The region outlined above was amplified by PCR, as discussed in Section 5.4, and the PCR products were electrophoresed on an agarose gel to ascertain the quality of the product. A representative example of the mtDNA products for primer region 3, as visualised by the UVIvue ultraviolet transilluminator, are presented in Figure 6.36.

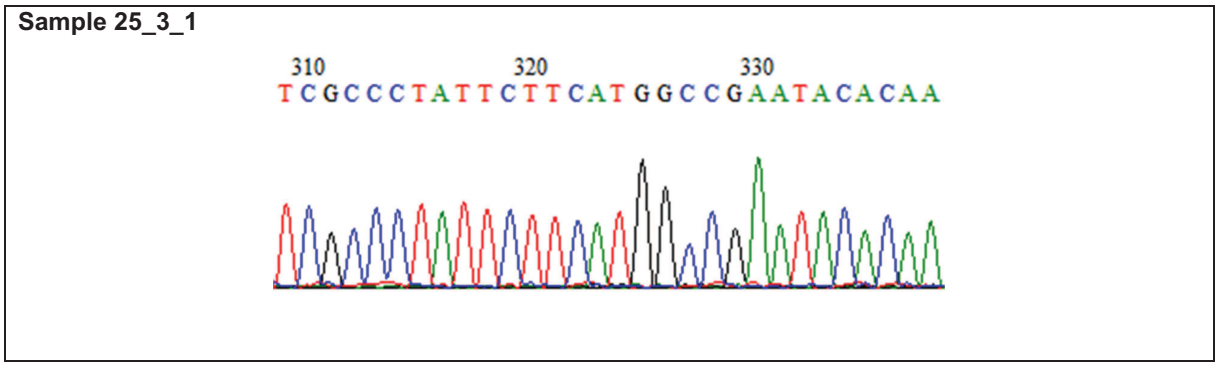
Figure 6.36 Photographic representation of the amplified mtDNA product of primer region 3



Photograph of the agarose gel electropherogram on which the mtDNA amplified product was electrophoresed at 100 volts (V) and 50 mAmperes (mA) for 30 minutes as discussed in Section 5.5; ladder = FastRuler™ High Range DNA Ladder (Fermentas) of range 100 – 10,000 bp; included in the first and last lane of the gel; negative = negative control; sample names refer to the Tswana-speaking individuals of this investigation.

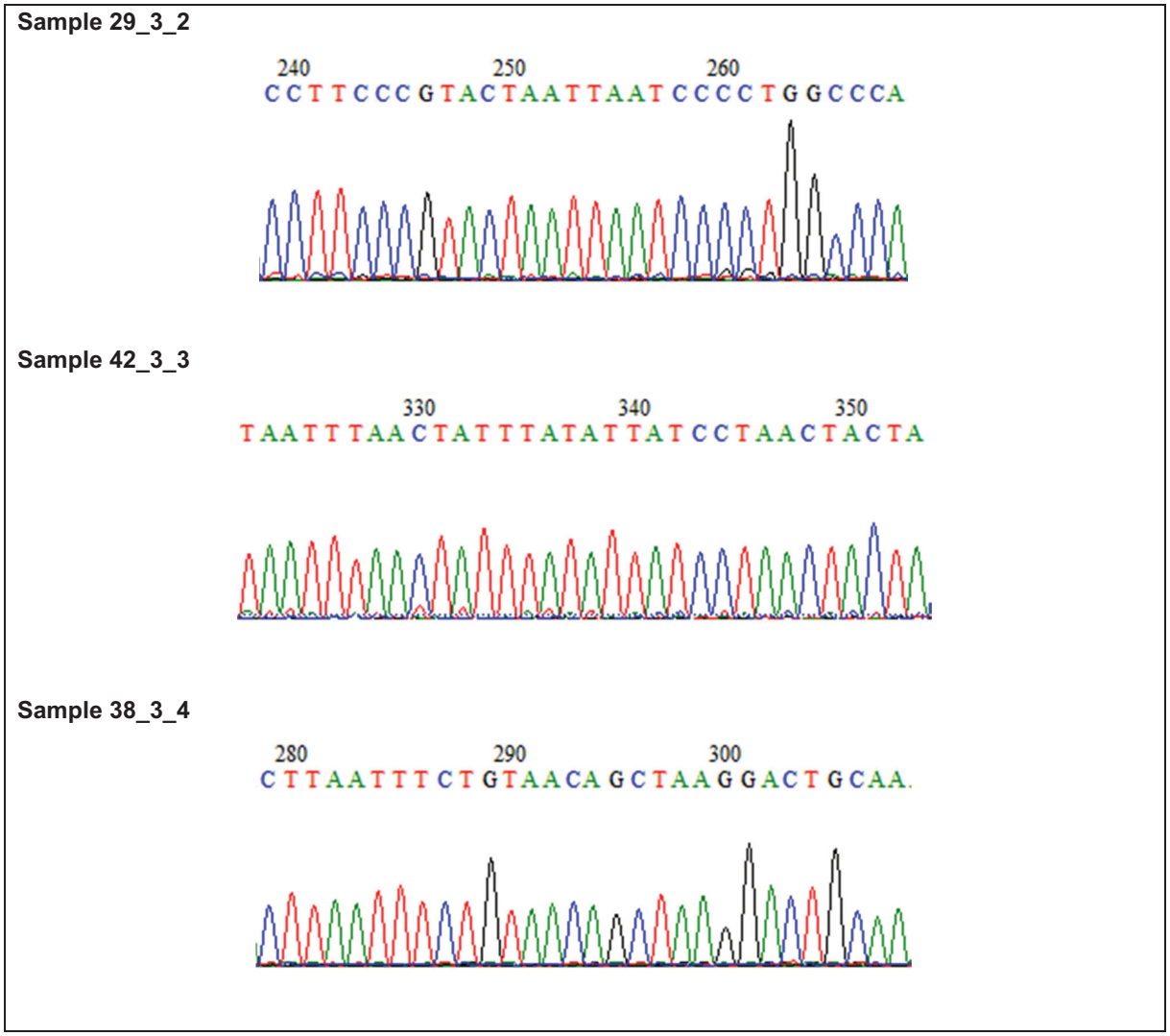
Representative examples of the sequence electropherograms generated by the BigDye®[1] Terminator v3.1 Cycle Sequencing Kit for primer region 3 are represented in Figure 6.37. These results were viewed and edited using the BioEdit software version 7.0.5.2 (Hall, 2001).

Figure 6.37 Representative electropherograms of the sequence generated for primer region 2 using the forward primers 1-4



1 BigDye® Terminator v3.1 Cycle Sequencing Kit is a registered trademark of Applied Biosystems, Foster City, CA, USA.

Figure 6.37 Continued…



Examples of electropherogram data with peaks depicting nucleotides in the sequence region of primer region 3; A = adenine; T = thymine; C = cytosine; G = guanine; numbering at the top of the electropherogram represents the numbering of the nucleotides as a sequenced fragment before alignment with the rCRS and therefore does not correspond to the nucleotide positions of primer region 3.

6.6.3.1 Sequence alterations observed in primer region 3

The sequence variance of the whole *ND1* gene will be discussed in this section, including the 337 base pair segment contained in primer region 2. This is followed by a discussion of the sequence variance observed in the tRNA isoleucine (I), tRNA glutamine (Q) and the tRNA methionine (M) coding regions. The *ND2* gene stretches from np 4470 to np 5579 and therefore contains a segment of 301 base pairs that is only amplified in primer region 4. In order to discuss the *ND2* gene as a single unit, the sequence variation of the *ND2* gene will be provided in Section 6.6.4. The sequence variance for the mtDNA region from np 3307, which includes the *ND1* gene, to the end of the tRNA methionine (M) at np 4469, is presented in Table 6.11.

Table 6.11 Sequence alterations observed between the complete mitochondrial DNA of the Tswana individuals included in this study and the rCRS in primer region 3

Position	Sequence alteration	Gene/ region	Frequency	Reference
3438	G-A	*ND1*	12	Kivisild *et al.*, 2006
3516	C-A	*ND1*	36	Ingman *et al.*, 2000
3579	A-G	*ND1*	1	Gonder *et al.*, 2007
3594	C-T	*ND1*	48	Ingman *et al.*, 2000
3618	T-C	*ND1*	6	Kivisild *et al.*, 2006
3660	**A-G**	***ND1***	**1**	**Current investigation**
3666	G-A	*ND1*	2	Mishmar *et al.*, 2003
3756	A-G	*ND1*	22	Mishmar *et al.*, 2003
3918	G-A	*ND1*	2	Kivisild *et al.*, 2006
3981	A-G	*ND1*	10	Mishmar *et al.*, 2003
4011	C-T	*ND1*	1	Olivieri *et al.*, 2006 non-African
4023	T-C	*ND1*	1	Derenko *et al*., 2007 non-African
4025	C-T	*ND1*	10	Mishmar *et al.*, 2003
4044	A-G	*ND1*	10	Mishmar *et al.*, 2003
4048	G-A	*ND1*	1	Kivisild *et al.*, 2006
4048	**G-C**	***ND1***	**1**	**Current investigation**
4092	G-A	*ND1*	1	Kivisild *et al.*, 2006
4104	A-G	*ND1*	48	Mishmar *et al.*, 2003
4197	C-T	*ND1*	2	Kivisild *et al.*, 2006
4203	A-G	*ND1*	1	Ingman *et al.*, 2000
4225	A-G	*ND1*	9	Mishmar *et al.*, 2003
4232	T-C	*ND1*	27	Mishmar *et al.*, 2003
4312	C-T	I (T-loop)	36	Behar *et al.*, 2008

Transitions are indicated in blue and transversions are indicated in red. *ND1* = NADH dehydrogenase subunit 1 gene; I = tRNA isoleucine ; T-loop = telomere loop.The frequency is indicated as the number of times the sequence variation was observed within the total Tswana dataset of 50 individuals. The structural region where the sequence variation occurs within either rRNA or tRNA coding regions is indicated in brackets following the region name; novel mutations are indicated in **bold**.

The NDI gene region was further investigated by the mtDNA-GeneSyn tool developed by Pereira *et al*. (2009). The nature of the effect of the mutations on the *ND1* gene coding functionality was determined and is presented in Table 6.12.

Table 6.12 Sequence variation within *ND1* gene

Position	Sequence alteration	Gene	Syn/ Non-syn	Codon	New codon	Codon position	Amino acid	New amino acid
3438	G-A	*ND1*	Syn	GGG	GGA	3	G	G
3516	C-A	*ND1*	Syn	CUC	CUA	3	L	L
3579	A-G	*ND1*	Syn	AUA	AUG	3	M	M
3594	C-T	*ND1*	Syn	GUC	GUU	3	V	V
3618	T-C	*ND1*	Syn	UUU	UUC	3	F	F
3660	A-G	*ND1*	Syn	UGA	UGG	3	W	W

Table 6.12 Continued…

Position	Sequence alteration	Gene	Syn/ Non-syn	Codon	New codon	Codon position	Amino acid	New amino acid
3666	G-A	*ND1*	Syn	GGG	GGA	3	G	G
3756	A-G	*ND1*	Syn	CUA	CUG	3	L	L
3918	G-A	*ND1*	Syn	GAG	GAA	3	E	E
3981	A-G	*ND1*	Syn	AUA	AUG	3	M	M
4011	C-T	*ND1*	Syn	AAC	AAU	3	N	N
4023	T-C	*ND1*	Syn	ACU	ACC	3	T	T
4025	C-T	*ND1*	Non-syn	ACA	AUA	2	T	M
4044	A-G	*ND1*	Syn	ACA	ACG	3	T	T
4048	G-A	*ND1*	Non-syn	GAC	AAC	1	D	N
4048C	G-C	*ND1*	Non-syn	GAC	CAC	1	D	H
4092	G-A	*ND1*	Syn	AAG	AAA	3	K	K
4104	A-G	*ND1*	Syn	CUA	CUG	3	L	L
4197	C-T	*ND1*	Syn	ACC	ACU	3	T	T
4203	A-G	*ND1*	Syn	GCA	GCG	3	A	A
4225	A-G	*ND1*	Non-syn	AUA	GUA	1	M	V
4232	T-C	*ND1*	Non-syn	AUU	ACU	2	I	T

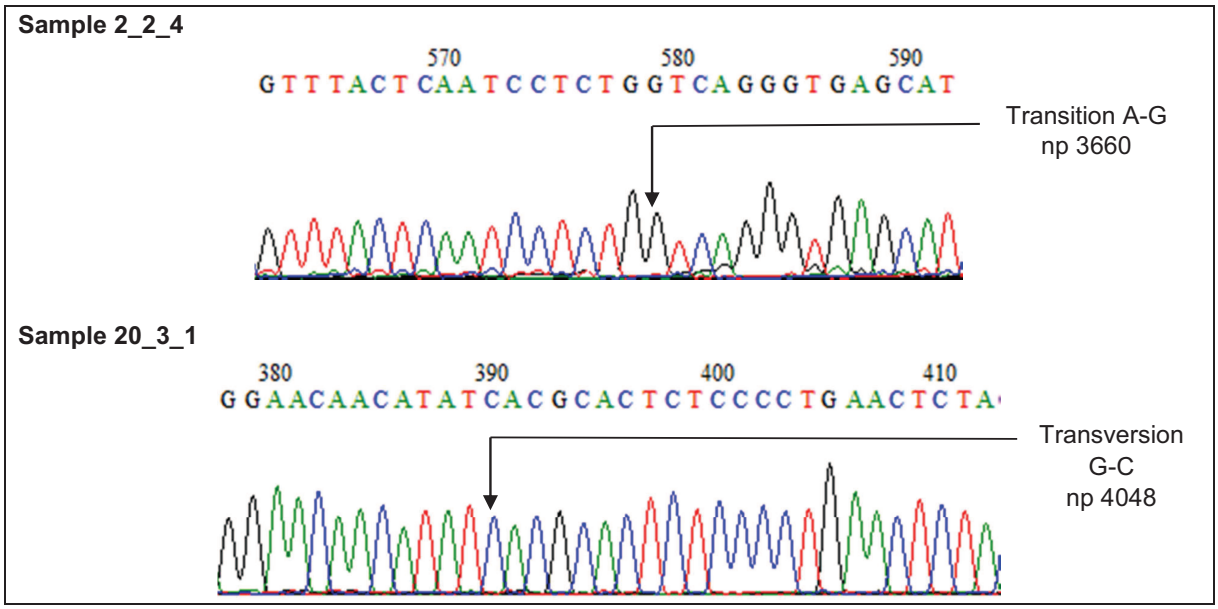
ND1 = NADH dehydrogenase subunit 1 gene; the sequence alteration described in terms of the synonymous or nonsynonymous nature of the change, the codon position that was affected and the new amino acid that was coded ; A = adenine; T = thymine; C = cytosine; G = guanine; syn = synonymous; nonsyn = nonsynonymous; G = Glycine; L = Leucine; M = Methionine; V = Valine; F = Phenylalanine; W = Tryptophan; E = Glutamic acid (Glutamate); N = Asparagine; D = Aspartic acid (Aspartate); K = Lysine; T = Threonine; A = Alanine; I = Isoleucine; H = Histidine; nonsynonymous mutations indicated in grey highlight.

The sequence alterations consist of 21 transitions and two transversions, thus giving a transversion:transition ratio (Lutz-Bonengel *et al.*, 2003) of 1:10.5, which indicates an excess of transitions, as is expected and was also displayed by the previous primer regions. As in primer region 2, no indels have been observed, which is not unexpected, since the occurrence of indels is reported to be rare (Pereira *et al*., 2009). The *ND1* gene displays 22 of these sequence alterations in this study, of which 17 are localised to the third codon positions and are therefore synonymous. Five of the sequence mutations are nonsynonymous, of which two are localised to the second codon position and three are localised to the first codon position. The number of sequence alterations i.e. 22, is much lower than the total number of sequence alterations of 116 that was observed in an investigation of 840 complete mitochondrial genomes by Moilanen and Majamaa (2003). This is attributed to the small population size of the Tswana cohort of this investigation.

Two novel sequence alterations have been observed in the *ND1* gene in primer region 3 at np 3660 and np 4048 of the Tswana-speaking individuals of this investigation. The electropherograms of these two samples of the Tswana cohort of this investigation, which display the novel sequence variants respectively, i.e. TS_2075 and TS_3466, are

presented in Figure 6.38. Both electropherograms display good peak morphologies and no evidence of sequencing artefacts, which rules out the possibility that the peaks are called because of sequencing error or background noise. This, however, does not rule out the possibility of human error or laboratory error and the presence of the sequence alteration should be verified by re-sequencing these mtDNA regions in these two individuals.

Figure 6.38 Representative electropherograms of the sequence data generated indicating a transition at np 3660 and a transversion at np 4048



Sample name 2_2_4 refers to sample TS_2075_2, primer region 2, sequence primer 4; sample name 20_3_1 refers to sample TS_3466_20, primer region 3, sequence primer 1; sequence regions presented here containing transition and transversion before alignment with rCRS, therefore numbering presented above is not in accordance with alteration position as determined by comparison to the rCRS.

The transition at np 3660 is located at the third codon position, which causes the UGA codon to change to a UGG codon and does not affect the amino acid, tryptophan, which it codes for. The transversion at np 4048 does, however, cause a nonsynonymous change at codon position one that codes for a histidine amino acid instead of an aspartic acid and could therefore be a candidate for pathogenicity because of the possibility that the functioning of the gene could have been affected. Both of these novel mutations are only present in a single Tswana-speaking individual from this investigation and would therefore have to be confirmed by repeating the sequencing of this region to verify its presence. Further investigation of the mutations should also identify whether these mutations are private mutations or constitute a rare fixed mutational event that defines a new sub-haplogroup among the Tswana-speaking individuals of this investigation or any other population group. To make any predictions about pathogenicity would be risky before a full control group study of these mutations in conjunction with other disease associated mutations has been undertaken.

The transition within the *ND1* gene at np 4025 has been reported as a candidate for a mutation involved in LHON disease but has only been reported in three cases in one family (Huoponen *et al.*, 1993) and therefore its pathogenicity is highly uncertain. None of the other sequence alterations observed in the Tswana-speaking study group have been reported to have disease associations.

Two sequence alterations have been observed in haplogroups other than the L haplogroups that are associated with individuals from Africa and have therefore been investigated for novelty within the L haplogroup lineages. Both are transitions at np 4011 and np 4023 respectively and each has only been observed in a single Tswana individual of this investigation. The np 4011 transition has been reported as a haplogroup-defining site of haplogroups M1a1a and R12 and the alteration at np 4023 has been reported as a haplogroup-defining site for haplogroup D3 by the PhyloTree classification system (Van Oven and Kayser, 2009). The sequence electropherograms of samples TS_4080 and TS_3002, which contain these sequence alterations, are presented in Figure 6.39. The electropherograms do not display any background noise or sequencing artefacts, which rules out the possibility of a false result due to sequencing errors. Based on these results, the alterations at np 4011 and np 4023 are interpreted as novel for the L haplogroup lineages. As discussed earlier in this section, however, further investigation of these mutations is warranted before any conclusive interpretation of these results can be made.

Figure 6.39 Representative electropherograms of the sequence data generated indicating transitions at np 4011 and np 4023

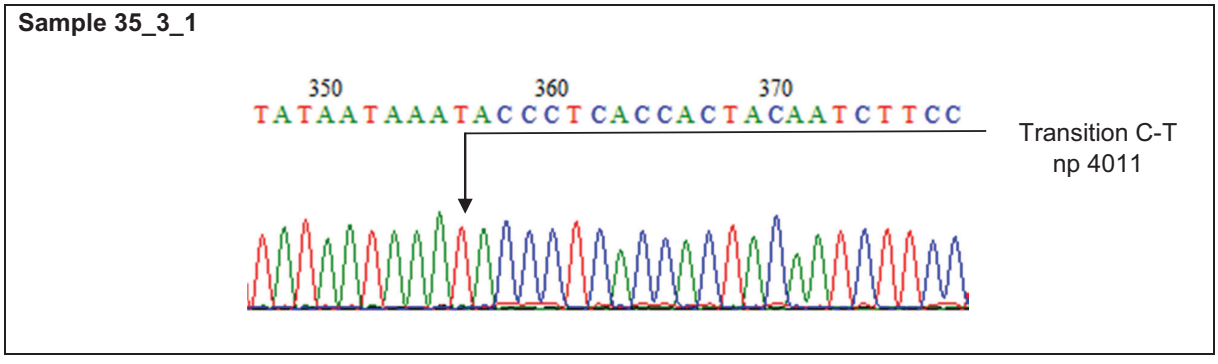
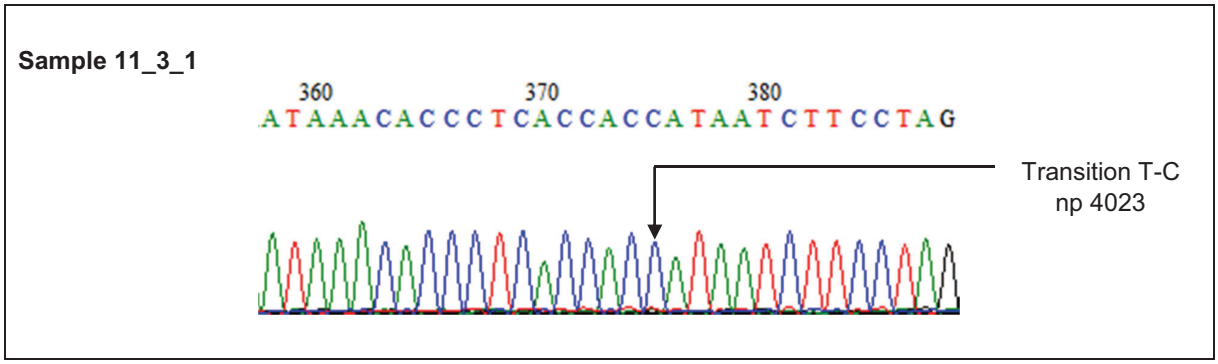


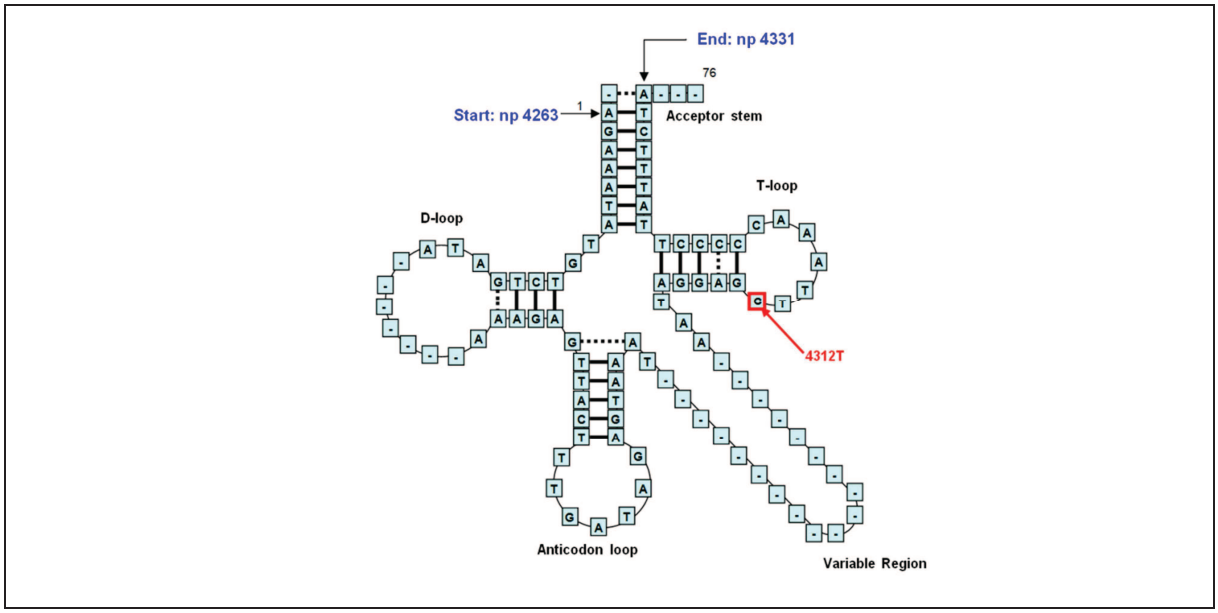
Figure 6.39 Continued…



Sample name 35_3_1 refers to sample TS_4080_35, primer region 3, sequence primer 1; sample name 11_3_1 refers to sample TS_3002_11, primer region 3, sequence primer 1; sequence regions presented here containing transitions before alignment with rCRS, therefore numbering presented above is not in accordance with alteration position as determined by comparison to the rCRS.

The coding region of the tRNA isoleucine (I) gene displays one transition at np 4312, which is localised to the T-loop region of the tRNA in 36 Tswana-speaking individuals of this investigation. The tRNA glutamine (Q) and tRNA methionine (M) genes display no sequence alterations. This observation is in support of the reported low sequence variance that is displayed by mitochondrial tRNA-coding regions when compared to the control region, the rRNA-coding regions and protein-coding regions of the mitochondrial genome (Vilmi *et al.*, 2005). The secondary structure of the tRNA isoleucine (I) molecule is presented in Figure 6.40 to indicate the position of the mutation observed within the Tswana cohort of this investigation.

Figure 6.40 Structure of the tRNA isoleucine (I) and observed sequence variation of the Tswana-speaking individuals of this investigation



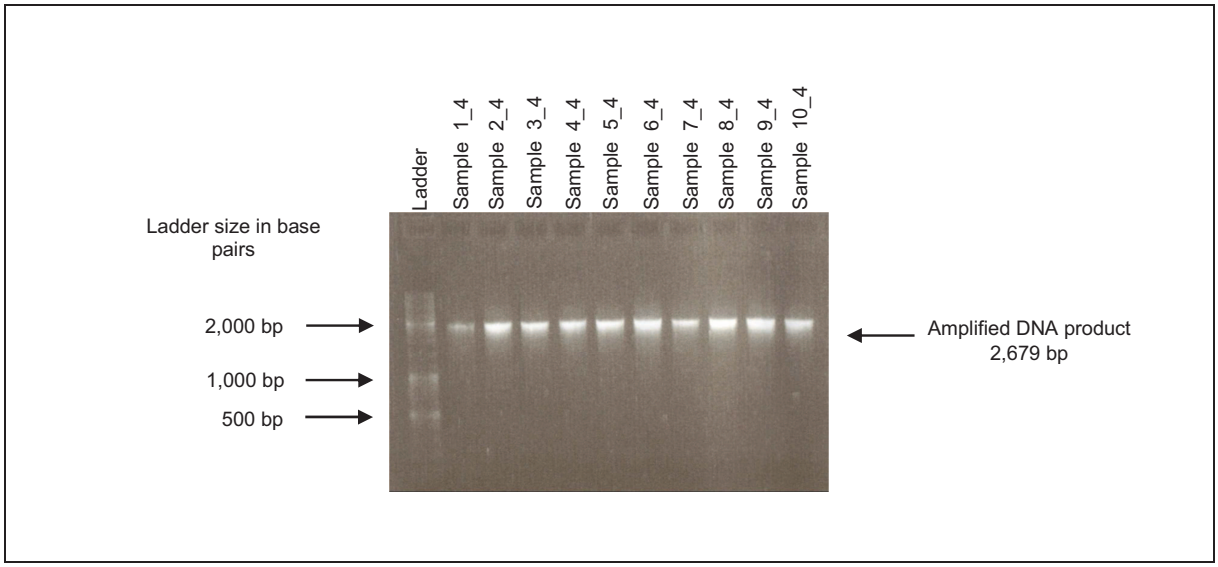
Structure of a typical tRNA isoleucine (L); ▬100% Watson Crick pairs; ▪▪100% mismatches; numbers alongside the tRNA structure indicate nucleotide positions within the tRNA molecule and not according to the rCRS; nucleotide starting and ending positions of the tRNA according to the rCRS indicated in blue ink; nucleotide position indicated in red ink refers to the nucleotide variation observed in the Tswana-speaking individuals of this investigation. From Jühling *et al.*, 2009; Pereira *et al.*, 2009.

6.6.4 Primer region 4

Primer region 4 starts at np 5279 and ends at np 7882. It contains a 232 base pair segment of the *ND2* gene of which the first part of the gene, consisting of an 809 base pair segment, resides in the primer 3 region. As was discussed in Section 6.6.3, the sequence alterations observed in the *ND2* segment within primer region 3 will be discussed in this section in order to consider the *ND2* gene as a single unit. The *ND2* gene ranges from np 4470 to np 5511. This region further contains the coding regions for tRNA tryptophan (W) gene (np 5512-np 5579), tRNA alanine (A) gene (np 5587-np 5655), tRNA asparagine (N) gene (np 5657-np 5729), tRNA cysteine (C) gene (np 5761-np 5826), tRNA tyrosine (Y) gene (np 5826-np 5891), tRNA serine 1 (S(UCN)) gene (np 7446-np 7514) and tRNA aspartic acid (D) gene (np 7518-np 7585). It also contains the COI protein coding gene that rangesfrom np 5904-np 7445 and a 387 base pair segment of the COII gene that ranges from np 7586-np 8269.

The region outlined above was amplified by PCR, as discussed in Section 5.4, and the PCR products were electrophoresed on an agarose gel to ascertain the quality of the product. A representative example of the mtDNA products for primer region 4, as visualised by the UVIvue ultraviolet transilluminator, is presented in Figure 6.41.

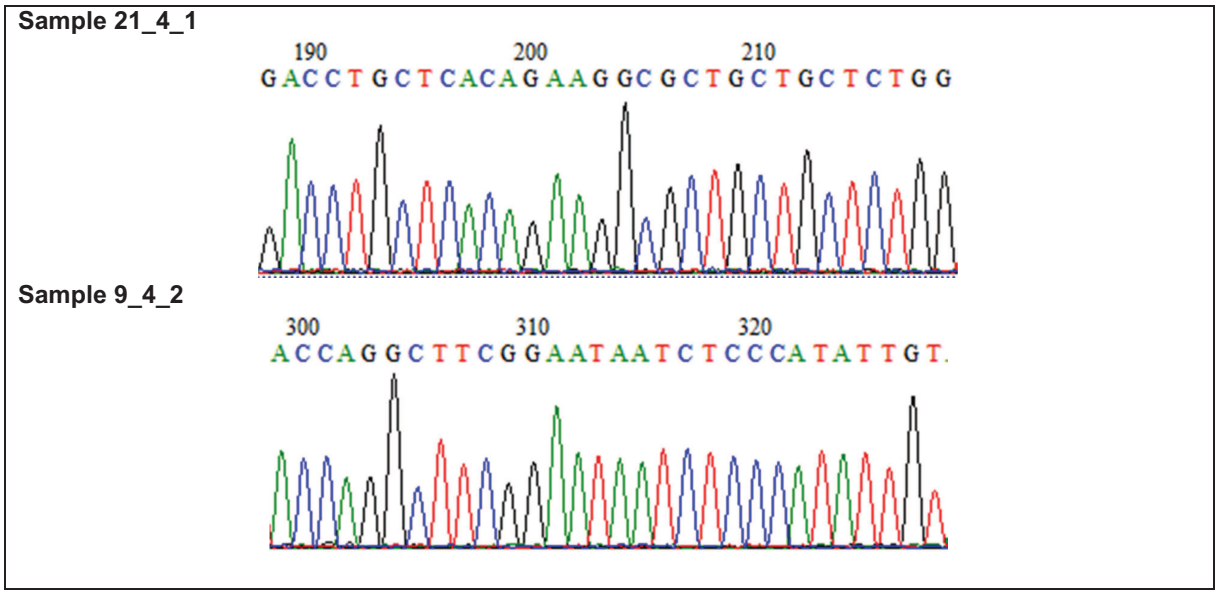
Figure 6.41 Photographic representation of the amplified mtDNA product of primer region 4



Photograph of the agarose gel on which the mtDNA amplified product was electrophoresed at 100 volts (V) and 50 mAmperes (mA) for 30 minutes as discussed in Section 5.5; ladder = FastRuler™ High Range DNA Ladder (Fermentas) of range 100 – 10,000 bp; included in the first lane of the gel; sample names refer to the Tswana-speaking individuals of this investigation.

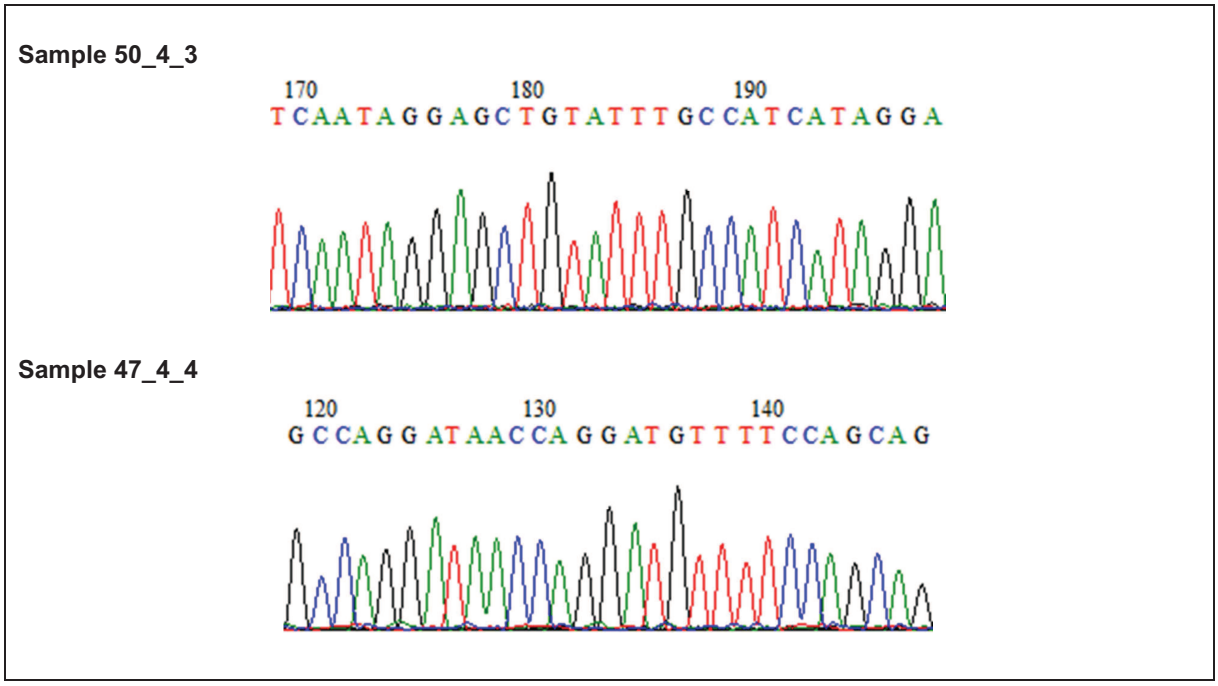
Representative examples of the electropherograms of the sequence data generated by the BigDye®[1] Terminator v3.1 Cycle Sequencing Kit for primer region 4 sequences are represented in Figure 6.42. These results were viewed and edited in BioEdit software version 7.0.5.2 (Hall, 2001).

Figure 6.42 Representative electropherograms of the sequence generated for primer region 4 using the forward primers 1-4



1 BigDye® Terminator v3.1 Cycle Sequencing Kit is a registered trademark of Applied Biosystems, Foster City, CA, USA.

Figure 6.42 Continued…



Examples of electropherogram data with peaks depicting nucleotides in the sequence region of primer 4; A = adenine; T = thymine; C = cytosine; G = guanine; numbering at the top of the electropherogram represents the numbering of the nucleotides as a sequenced fragment before alignment with the rCRS and therefore does not correspond to the nucleotide positions of primer region 4.

6.6.4.1 Sequence alterations observed in primer region 4

For the purpose of discussion, the observed sequence alterations of the *ND2* gene as a single unit, i.e. the entire segment of the *ND2* gene that starts at np 4470 up to the start of primer region 4 at np 5279, are included in this section. The full *COII* gene has not been amplified within the primer region 4 and only a segment that stretches from the start of the *COII* gene at np 7586 to the end of primer 4 region at np 7918 has been included. In order to discuss the *COII* gene as a single unit, the 332 base pair segment of the *COII* gene in primer region 4 is discussed in Section 6.6.5. The sequence alterations observed within primer region 4 in the Tswana-speaking individuals of this investigation, which includes the *ND2* segment of primer region 3 from np 4470 up to the start of *COII* gene at np 7586, are presented in Table 6.13.

Table 6.13 Sequence alterations observed between the complete mitochondrial DNA of the Tswana individuals included in this study and the rCRS in primer region 4

Position	Sequence alteration	Gene/ region	Frequency	Reference
4541	G-A	*ND2*	1	Gonder *et al.*, 2007
4586	T-C	*ND2*	9	Gonder *et al.*, 2007
4769	A-G	*ND2*	50	Gonder *et al.*, 2007

Table 6.13 Continued…

Position	Sequence alteration	Gene/ regionR	Frequency	Reference
4814	C-T	*ND2*	1	Friedlaender *et al.*, 2007
4896	**T-C**	***ND2***	**1**	**Current investigation**
4907	T-C	*ND2*	1	Mishmar *et al.*, 2003
4943	A-G	*ND2*	1	Sun *et al.*, 2006
5090	T-C	*ND2*	3	Behar *et al.*, 2008
5096	T-C	*ND2*	3	Mishmar *et al.*, 2003
5147	G-A	*ND2*	5	Mishmar *et al.*, 2003
5153	A-G	*ND2*	9	Mishmar *et al.*, 2003
5196	T-C	*ND2*	1	Kivisild *et al.*, 2006
5231	G-A	*ND2*	8	Kivisild *et al.*, 2006
5285	A-G	*ND2*	2	Behar *et al.*, 2008
5441	A-G	*ND2*	1	Torroni *et al.*, 2006
5442	T-C	*ND2*	36	Gonder *et al.*, 2007
5460	G-A	*ND2*	12	Gonder *et al.*, 2007
5471	G-A	*ND2*	2	Gonder *et al.*, 2007
5553	T-C	W (Ac-stem)	1	Kivisild *et al.*, 2006
5581	A-G	Non-coding region	2	Kivisild *et al.*, 2006
5582	A-G	Non-coding region	1	Behar *et al.*, 2008
5603	C-T	A (T-loop)	7	Gonder *et al.*, 2007
5656	A-G	Non-coding region	1	Torroni *et al.*, 2006
5711	A-G	N (D-loop)	4	Behar *et al.*, 2008
5773	G-A	C (T-loop)	5	Mishmar *et al.*, 2003
5782	**T-C**	**C (T-stem)**	**1**	**Current investigation**
5811	A-G	C (D-loop)	1	Ingmann *et al.*, 2000
5899.1	insC	Non-coding region	3	Behar *et al.*, 2008
5911	C-T	*COI*	3	Behar *et al.*, 2008
5951	A-G	*COI*	2	Mishmar *et al.*, 2003
6038	**C-T**	***COI***	**1**	**Current investigation**
6071	T-C	*COI*	2	Gonder *et al.*, 2007
6150	G-A	*COI*	2	Kivisild *et al.*, 2006
6152	T-C	*COI*	2	Gonder *et al.*, 2007
6170	C-T	*COI*	5	Behar *et al.*, 2008
6185	T-C	*COI*	36	Gonder *et al.*, 2007
6221	T-C	*COI*	1	Gonder *et al.*, 2007
6253	T-C	*COI*	2	Behar *et al.*, 2008
6257	G-A	*COI*	4	Kivisild *et al.*, 2006
6261	G-A	*COI*	1	Behar *et al.*, 2008
6266	A-G	*COI*	12	Gonder *et al.*, 2007
6267	G-T	*COI*	1	Herrnstadt *et al.*, 2002
6377	C-T	*COI*	6	Gonder *et al.*, 2007
6383	G-A	*COI*	5	Kivisild *et al.*, 2006
6587	C-T	*COI*	1	Behar *et al.*, 2008
6663	A-G	*COI*	2	Gonder *et al.*, 2007

Table 6.13 Continued…

Position	Sequence alteration	Gene/ regionR	Frequency	Reference
6680	T-C	*COI*	1	Behar *et al.*, 2008
6723	G-A	*COI*	2	Behar *et al.*, 2008
6815	T-C	*COI*	28	Gonder *et al.*, 2007
6938	C-T	*COI*	1	Behar *et al.*, 2008
7028	C-T	*COI*	50	Mishmar *et al.*, 2003
7046	A-G	*COI*	1	Just *et al.*, 2008
7055	A-G	*COI*	2	Mishmar *et al.*, 2003
7076	A-G	*COI*	2	Behar *et al.*, 2008
7119	G-A	*COI*	6	Behar *et al.*, 2008
7146	A-G	*COI*	38	Ingmann *et al.*, 2000
7154	A-G	*COI*	10	Gonder *et al.*, 2007
7175	T-C	*COI*	11	Gonder *et al.*, 2007
7256	C-T	*COI*	48	Mishmar *et al.*, 2003
7257	A-G	*COI*	1	Behar *et al.*, 2008
7274	C-T	*COI*	10	Mishmar *et al.*, 2003
7278	T-C	*COI*	1	Gonder *et al.*, 2007
7283	T-C	*COI*	7	Kivisild *et al.*, 2006
7310	T-C	*COI*	1	Behar *et al.*, 2008
7337	G-A	*COI*	2	Behar *et al.*, 2008
7389	T-C	*COI*	4	Kivisild *et al.*, 2006
7424	A-G	*COI*	1	Mishmar *et al.*, 2003
7521	G-A	D Acc-stem	48	Mishmar *et al.*, 2003

Transitions are indicated in blue and transversions are indicated in red. The frequency is indicated as the number of times the sequence variation was observed within the total Tswana dataset of 50 individuals. The structural region where the sequence variation occurs within rRNA and tRNA coding regions is indicated in brackets following the region name; D Acc stem = acceptor stem tRNA. *ND2* = NADH dehydrogenase subunit 2 gene; *COI* = cytochrome c oxidase subunit I gene

The *ND2* gene, which is 1,041 base pairs long, displays 18 sequence alterations all consisting of transitions, with no transversions or indels observed. The tRNA tryptophan, tRNA alanine, tRNA asparagine and tRNA aspartic acid coding regions display a transition each and no transversions or indels. The tRNA cysteine coding region displays three transitions and no transversions or indels. The *COI* gene, which is 1,541 base pairs long, displays 38 transitions, one transversion and no indels. An insertion is observed in the non-coding region between the tRNA asparagine coding region and the *COI* gene. In total, primer region 4 displays a high number of transitions with a transversion:transition count ratio (Lutz-Bonengel *et al.*, 2003) of 1:67. The near absence of transversions and indels could be explained by the fact that the size of the sample is small and limited to a population of one geographical region and will therefore not harbour a high level of sequence variation, as would be expected from a global sample.

The *ND2* and *COI* gene regions were further investigated by the mtDNA-GeneSyn tool developed by Pereira *et al*. (2009). The impact of the sequence alterations on the *ND2* and *COI* gene coding functionality was determined and is presented in Table 6.14.

Table 6.14 Sequence variation within *ND2* and *COI* genes

Position	Sequence alterations	Gene	Syn/ Non-syn	Codon	New codon	Codon position	Amino acid	New amino acid
4541	G-A	*ND2*	Syn	UCG	UCA	3	S	S
4586	T-C	*ND2*	Syn	GCU	GCC	3	A	A
4769	A-G	*ND2*	Syn	AUA	AUG	3	M	M
4814	C-T	*ND2*	Syn	GUC	GUU	3	V	V
4896	T-C	*ND2*	Non-syn	UAC	CAC	1	Y	H
4907	T-C	*ND2*	Syn	UCU	UCC	3	S	S
4943	A-G	*ND2*	Syn	UCA	UCG	3	S	S
5090	T-C	*ND2*	Syn	AUU	AUC	3	I	I
5096	T-C	*ND2*	Syn	AUU	AUC	3	I	I
5147	G-A	*ND2*	Syn	ACG	ACA	3	T	T
5153	A-G	*ND2*	Syn	CUA	CUG	3	L	L
5196	T-C	*ND2*	Syn	UUA	CUA	1	L	L
5231	G-A	*ND2*	Syn	CUG	CUA	3	L	L
5285	A-G	*ND2*	Syn	AAA	AAG	3	K	K
5441	A-G	*ND2*	Syn	CCA	CCG	3	P	P
5442	T-C	*ND2*	Non-syn	UUC	CUC	1	F	L
5460	G-A	*ND2*	Non-syn	GCC	ACC	1	A	T
5471	G-A	*ND2*	Syn	ACG	ACA	3	T	T
5911	C-T	*COI*	Non-syn	GCC	GUC	2	A	V
5951	A-G	*COI*	Syn	GGA	GGG	3	G	G
6038	C-T	*COI*	Syn	GGC	GGU	3	G	G
6071	T-C	*COI*	Syn	GUU	GUC	3	V	V
6150	G-A	*COI*	Non-syn	GUU	AUU	1	V	I
6152	T-C	*COI*	Syn	GUU	GUC	3	V	V
6170	C-T	*COI*	Syn	GCC	GCU	3	A	A
6185	T-C	*COI*	Syn	UUU	UUC	3	F	F
6221	T-C	*COI*	Syn	CCU	CCC	3	P	P
6253	T-C	*COI*	Non-syn	AUA	ACA	2	M	T
6257	G-A	*COI*	Syn	GUG	GUA	3	V	V
6261	G-A	*COI*	Non-syn	GCC	ACC	1	A	T
6266	A-G	*COI*	Syn	GGA	GGG	3	G	G
6267	G-T	*COI*	Non-syn	GCA	UCA	1	A	S
6377	C-T	*COI*	Syn	AUC	AUU	3	I	I
6383	G-A	*COI*	Syn	GGG	GGA	3	G	G
6587	C-T	*COI*	Syn	CCC	CCU	3	P	P
6663	A-G	*COI*	Non-syn	AUC	GUC	1	I	V
6680	T-C	*COI*	Syn	ACU	ACC	3	T	T

Table 6.14 Continued…

Position	Sequence alterations	Gene	Syn/ Non-syn	Codon	New codon	Codon position	Amino acid	New amino acid
6723	G-A	*COI*	Non-syn	GUC	AUC	1	V	I
6815	T-C	*COI*	Syn	UAU	UAC	3	Y	Y
6938	C-T	*COI*	Syn	AUC	AUU	3	I	I
7028	C-T	*COI*	Syn	GCC	GCU	3	A	A
7046	A-G	*COI*	Syn	CUA	CUG	3	L	L
7055	A-G	*COI*	Syn	GGA	GGG	3	G	G
7076	A-G	*COI*	Syn	GGA	GGG	3	G	G
7119	G-A	*COI*	Non-syn	GAC	AAC	1	D	N
7146	A-G	*COI*	Non-syn	ACU	GCU	1	T	A
7154	A-G	*COI*	Syn	AUA	AUG	3	M	M
7175	T-C	*COI*	Syn	ACU	ACC	3	T	T
7256	C-T	*COI*	Syn	AAC	AAU	3	N	N
7257	A-G	*COI*	Non-syn	AUC	GUC	1	I	V
7274	C-T	*COI*	Syn	GGC	GGU	3	G	G
7278	T-C	*COI*	Non-syn	UUC	CUC	1	F	L
7283	T-C	*COI*	Syn	AUU	AUC	3	I	I
7310	T-C	*COI*	Syn	AUU	AUC	3	I	I
7337	G-A	*COI*	Syn	UCG	UCA	3	S	S
7389	T-C	*COI*	Non-syn	UAU	CAU	1	Y	H
7424	A-G	*COI*	Syn	GAA	GAG	3	E	E

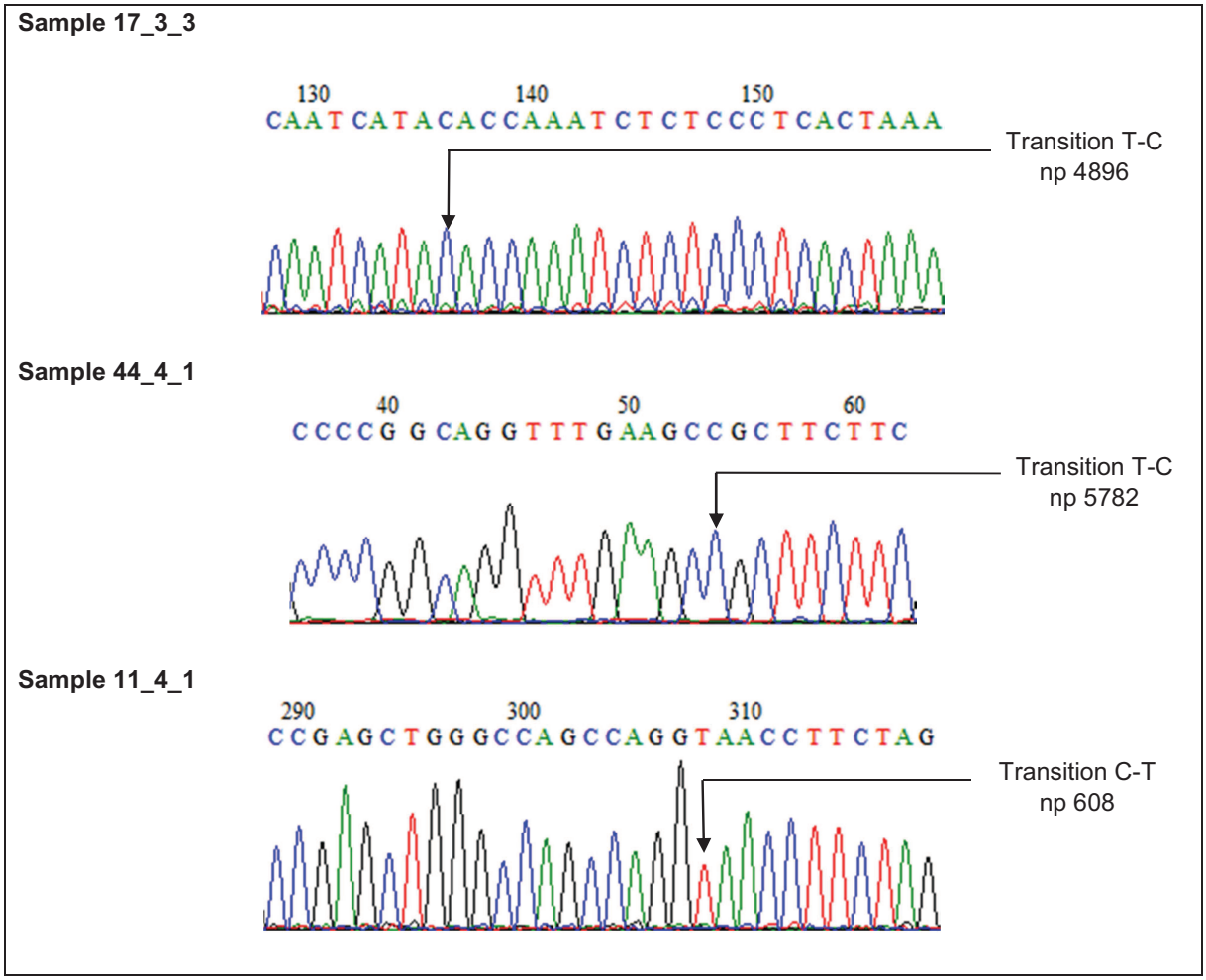
ND2 = NADH dehydrogenase subunit 2 gene; *COI* = cytochrome c oxidase subunit I; the sequence alterations described in terms of the synonymous or nonsynonymous nature of the changes, the codon position that was affected and the new amino acid that was coded for, if so; the nonsynonymous changes were indicated in grey highlight; A = adenine; T = thymine; C = cytosine; G = guanine; synonymous = synonymous; nonsyn = nonsynonymous; G = Glycine; L = Leucine; M = Methionine; V = Valine; F = Phenylalanine; W = Tryptophan; E = Glutamic acid (Glutamate); N = Asparagine; D = Aspartic acid (Aspartate); K = Lysine; T = Threonine; A = Alanine; I = Isoleucine; H = Histidine; S = Serine; Y = Tyrosine; P = Proline.

The *ND2* gene displays 15 synonymous substitutions in the third codon position and three nonsynonymous substitutions, which are all localised to the first codon position. The *COI* gene displays 27 synonymous substitutions in the third codon position and 12 nonsynonymous substitutions. of which ten are localised to the first codon position and two to the second codon position.

Three novel mutations are observed in the Tswana-speaking individuals of this investigation at np 4896 within the *ND2* gene region, at np 5782 in the tRNA cysteine coding region and at np 6038 within the *COI* gene region. Each of these mutations is only observed in a single Tswana-speaking individual. The novel mutation within the *ND2* gene region consists of a transition at the first codon position, which causes a nonsynonymous change that codes for a histidine amino acid instead of a tyrosine amino acid. The novel mutation within the *COI* gene region also consists of a transition, which is located at the

third codon position and thus results in a synonymous alteration. The novel mutation within the tRNA cysteine coding region consists of a transition, which occurs within the T-stem region of the tRNA molecule and could therefore be detrimental to the functioning of the tRNA. The sequence electropherograms of the regions that contain the novel mutations of samples TS_3117, TS_5063 and TS_3002 are presented in Figure 6.43. The electropherograms do not display any sequencing artefacts or background noise and display good peak morphologies, thus it is unlikely that these peaks have been called in response to sequencing errors or sequencing artefacts. Since human error and laboratory error cannot be ruled out, these novel mutations must be verified by re-sequencing the regions of the samples in which they have been observed.

Figure 6.43 Representative electropherograms of the sequence data generated indicating transitions at np 4896, np 5782 and np 6083

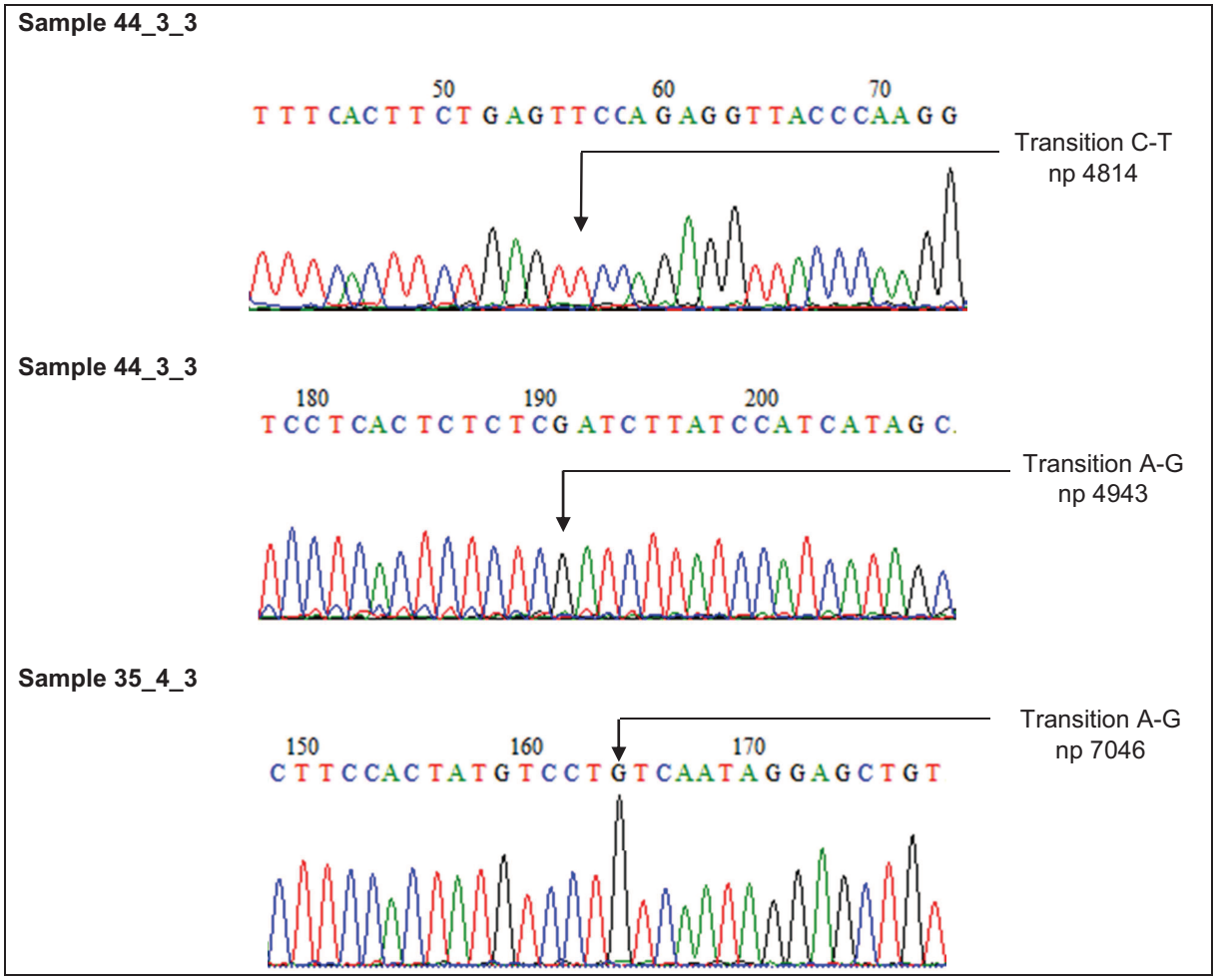


Sample name 17_3_3 refers to sample TS_3117_17, primer region 3, sequence primer 3; sample name 44_4_1 refers to sample TS_5063_44, primer region 4, sequence primer 1; sample name 11_4_1 refers to sample TS_3002_11, primer region 4, sequence primer 1; sequence regions presented containing transitions before alignment with rCRS, therefore numbering presented above is not in accordance with alteration position as determined by comparison to the rCRS.

Two transitions in the *ND2* gene region and one transition in the *COI* gene region have only been reported in an mtDNA sequence that belongs to other haplogroups than those in

macrohaplogroup L. These mutations are therefore possibly novel within the haplogroup L lineages. The transitions at np 4814 and np 4943 are located in a single Tswana-speaking individual of this investigation (TS_5063) within the *ND2* gene at third codon positions and are thus synonymous alterations. The transition at np 7046, which is located in the *COI* gene, is also at a third codon position in one Tswana-speaking individual (TS_4080). The transition at np 4814 is cited as a haplogroup-defining substitution for haplogroup M33b and haplogroup P4a1 by the PhyloTree classification system (Van Oven and Kayser, 2009). The transition at np 4943, which has been observed in the same individual as the previous alteration, is not cited as haplogroup-defining and has only been reported by one other study (Sun *et al.*, 2006). The electropherograms of the samples that contain these mutations, TS_5063 and TS_4080, are presented in Figure 6.44 and display no sequencing artefacts or background noise.

Figure 6.44 Representative electropherograms of the sequence data generated indicating transitions at np 4814, np 4943 and np 7046



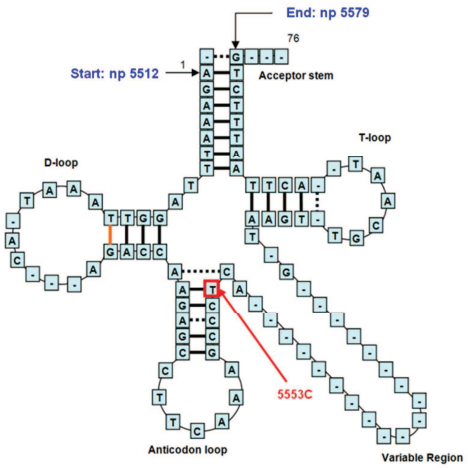
Sample name 44_3_3 refers to sample TS_5063_44, primer region 3, sequence primer 3; sample name 35_4_3 refers to sample TS_4080_35, primer region 4, sequence primer 3; sequence regions presented containing transitions before alignment with rCRS, therefore numbering presented above is not in accordance with alteration position as determined by comparison to the rCRS.

The morphologies of all the peaks are good and the possibility that these peaks have been falsely called has therefore been ruled out. The mutations, however, must be verified by resampling the individual and re-sequencing to rule out the possibility of human or laboratory error, especially since two of these mutations occur in a single individual (TS_5063), who also harbours the novel transition at np 5782 and a novel L haplogroup transition at np 1415, as discussed in Section 6.9.3.1. It is unlikely that a single individual would harbour so many novel mutations and it is critical that the mtDNA genome of this individual be re-sampled and re-sequenced before making final conclusions about the novelty of the mutations observed. To contribute further to this uncertainty, the haplogroup assignment of this individual, TS_5063, cannot be resolved and it has been positioned at the L0a'b'f root in all of the phylogenetic trees of this investigation, suggesting that it does not share maternal ancestors to the level of sub-haplogroup with the other Tswana-speaking individuals of this investigation. The deep level of unresolved assignment within the phylogenetic trees of this individual is a further reason to investigate the validity of the sequencing results, since they suggest that a lineage existed that diverged early in the history of modern humans, which has not yet been identified. However, the fact that the novel L haplogroup mutations in this individual have been determined by two different amplification reactions (primer 2 and primer 3) provides evidence that the observed substitutions are in fact true. If sequencing or human error had been at play, it would be expected that a high number of novel or uniquely combined sets of sequence alterations would have been observed in more than one individual within the datasets of this investigation, which was not the case. It is therefore premature to come to a conclusion about the reasons for this phenomenon and it should be resolved by further investigation.

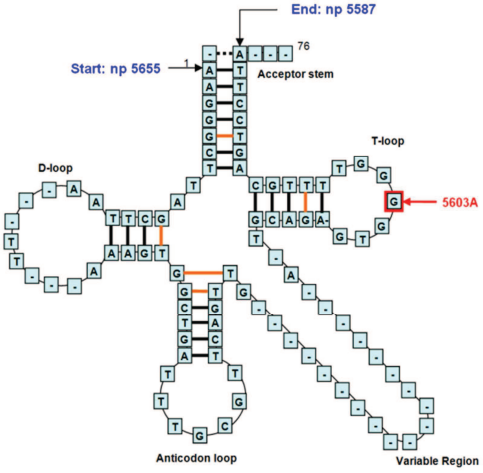
Three of the mutations that occur within the tRNA coding regions of primer region 4, i.e. those at np 5553, np 5782, and np 7521, are localised in the stem regions of the tRNA secondary structure and could possibly affect the functioning of the tRNA, since this molecule is dependent on its secondary structure for proper functioning. The other four alterations at np 5603, np 5711, np 5773 and np 5811 occur in the loop regions, which are the more unconserved areas in the tRNA molecules. The location of the sequence alterations observed in the Tswana-speaking individuals of this investigation within the tRNA coding regions of primer region 4 are presented in Figure 6.45.

Figure 6.45 Structure of the tRNA tryptophan, tRNA alanine, tRNA asparagine, tRNA cysteine and tRNA aspartic acid

tRNA tryptophan



tRNA alanine



tRNA asparagine

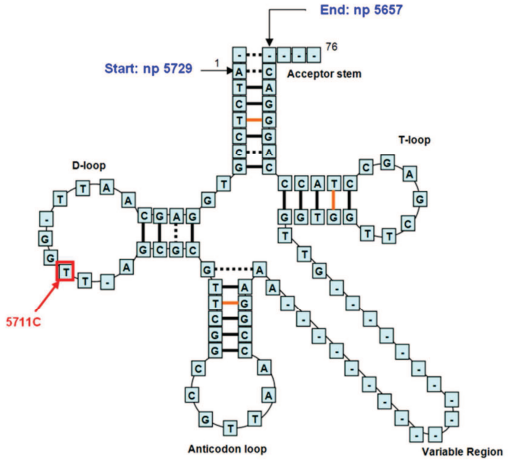
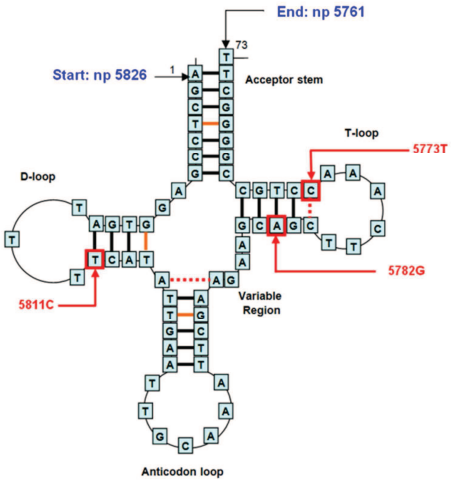
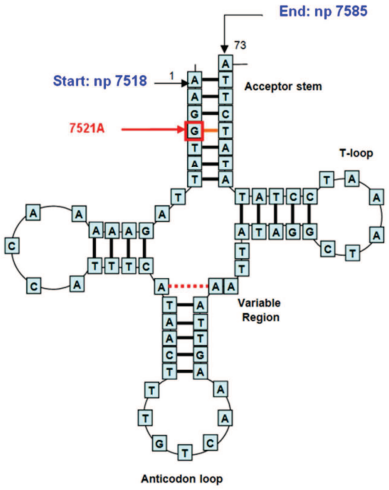


Figure 6.45 Continued…

tRNA cysteine



tRNA aspartic acid



Structure of a typical tRNA tryptophan, tRNA alanine, tRNA asparagine, tRNA cysteine and tRNA aspartic acid; ▬ 100% Watson Crick pairs; ▪▪100% mismatches;▬ 100% GT pairs; numbers alongside the tRNA structure indicate nucleotide positions within the tRNA molecule and are not according to the rCRS; nucleotide starting and ending positions of the tRNA according to the rCRS indicated in blue ink; nucleotide position indicated in red ink refers to the nucleotide variation observed in the Tswana-speaking individuals of this investigation. From Jühling *et al.* (2009); Pereira *et al.*, 2009.

No disease-associated mutations within the *ND2* gene or any of the tRNA coding regions of primer region 4 have been observed in the Tswana cohort of this investigation. A list of the reported disease-associated mutations within the *COI* gene region of the Tswana-speaking individuals of this investigation is presented in Table 6.15.

Table 6.15 Reported mtDNA sequence alterations with disease associations within primer region 4

Locus	Sequence alteration	Number of individuals	Disease	Reference
COI	C5911T	3	Prostate Cancer	Petros *et al.*, 2005
COI	G6150A	2	Prostate Cancer	Petros *et al.*, 2005
COI	T6253C	2	Prostate Cancer	Petros *et al.*, 2005
COI	G6261A	1	Prostate Cancer LHON	Petros *et al.*, 2005 Abu-Amero and Bosley, 2006
COI	G6267T	1	Prostate Cancer	Petros *et al.*, 2005
COI	A6663G	2	Prostate Cancer	Petros *et al.*, 2005

Sequence alterations are displayed as the position at which the mutation occurred with the wild nucleotide type indicated in front of the np and the mutant nucleotide type indicated after the np; the number of individuals = number of Tswana-speaking individuals of this investigation who displayed the mutation; disease associations were reported in one or more publications and have been considered as possibly pathologic and therefore are only a reflection of literature as stated in MITOMAP. Adapted from MITOMAP: A Human Mitochondrial Genome Database. http://www.mitomap.org, 2011.

Although pathogenic mutations are uncommon in the *COI* gene (Petros *et al.*, 2005), six mutations reportedly associated with prostate cancer have been observed in 11 Tswana-speaking individuals of this investigation. Five *COI* mutations, i.e. G6150A, T6253C, G6261A, G6267T and A6663G, have been reported to be associated with prostate cancer (Petros *et al.*, 2005) in individuals belonging to haplogroup L. All five of those alterations have been observed in the Tswana cohort of this investigation. This disease phenotype association may be incidental, since these same polymorphisms are African-specific. The sixth mutation observed in the Tswana cohort is the transition at np 5911 that has been reported to be present in individuals belonging to haplogroup U (Petros *et al*., 2005). This large number of disease-associated mutations in the *COI* gene region is not surprising, since missense polymorphisms in the African haplogroup L lineages are more commonly observed than in individuals from other parts of the world (Mishmar *et al.*, 2003). Furthermore, it has been speculated that the high *COI* sequence variation displayed by individuals belonging to haplogroup L, especially haplogroup L0, may have caused an increased predisposition to prostate cancer in African men (Petros *et al.*, 2005). Although all the prostate-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 one individual at a time. The mtDNA *COI* gene has been reported to play an important role in the etiology of prostate cancer and would therefore be an important pathogenic possibility to investigate in the Tswana-speaking individuals of South Africa, based on the presence of these reported mutations in this population. The etiology of prostate cancer, however, consists of more than *COI* gene mutations and if any conclusions about the prevalence of prostate cancer in the Tswana cohort of this

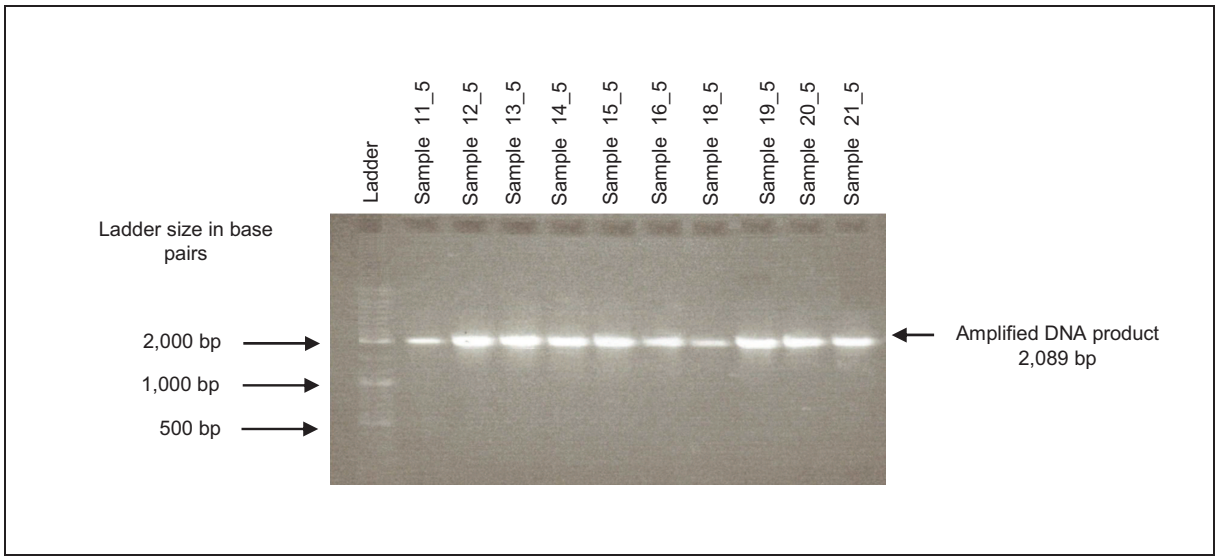
investigation need to be made, other mtDNA mutations would have to be investigated in conjunction with the *COI* gene mutations, as well as using appropriate control groups (Petros *et al.*, 2005).

6.6.5 Primer region 5

Primer region 5 starts at np 7883 and ends at np 9886. The *COII* gene, which starts in primer region 4 at np 7586 and ends in primer region 5 at np 8269, will be discussed as a single unit in this section. Therefore the sequence alterations that occur in the 296 base pair segment of the *COII* gene in primer region 4 will be included in the discussion of the observed sequence alterations of primer region 5. This region further containsd the coding region for the tRNA lysine (K) gene (np 8295-np 8364), the ATP synthase F0 subunit 8 (*ATP8*) gene (np 8366-np 8572), the ATP synthase F0 subunit 6 (*ATP6*) gene (np 8527-np 9207) and the *COIII* gene, which starts at np 9207 and ends at np 9990. Primer region 5 ends at np 9886 and therefore does not include the full region of the *COIII* gene.

The region outlined above has been amplified by PCR, as discussed in Section 5.4, and the PCR have been electrophoresed on an agarose gel to ascertain the quality of the amplified product. A representative example of the mtDNA products for primer region 5, as visualised by the UVIvue ultraviolet transilluminator, are presented in Figure 6.46.

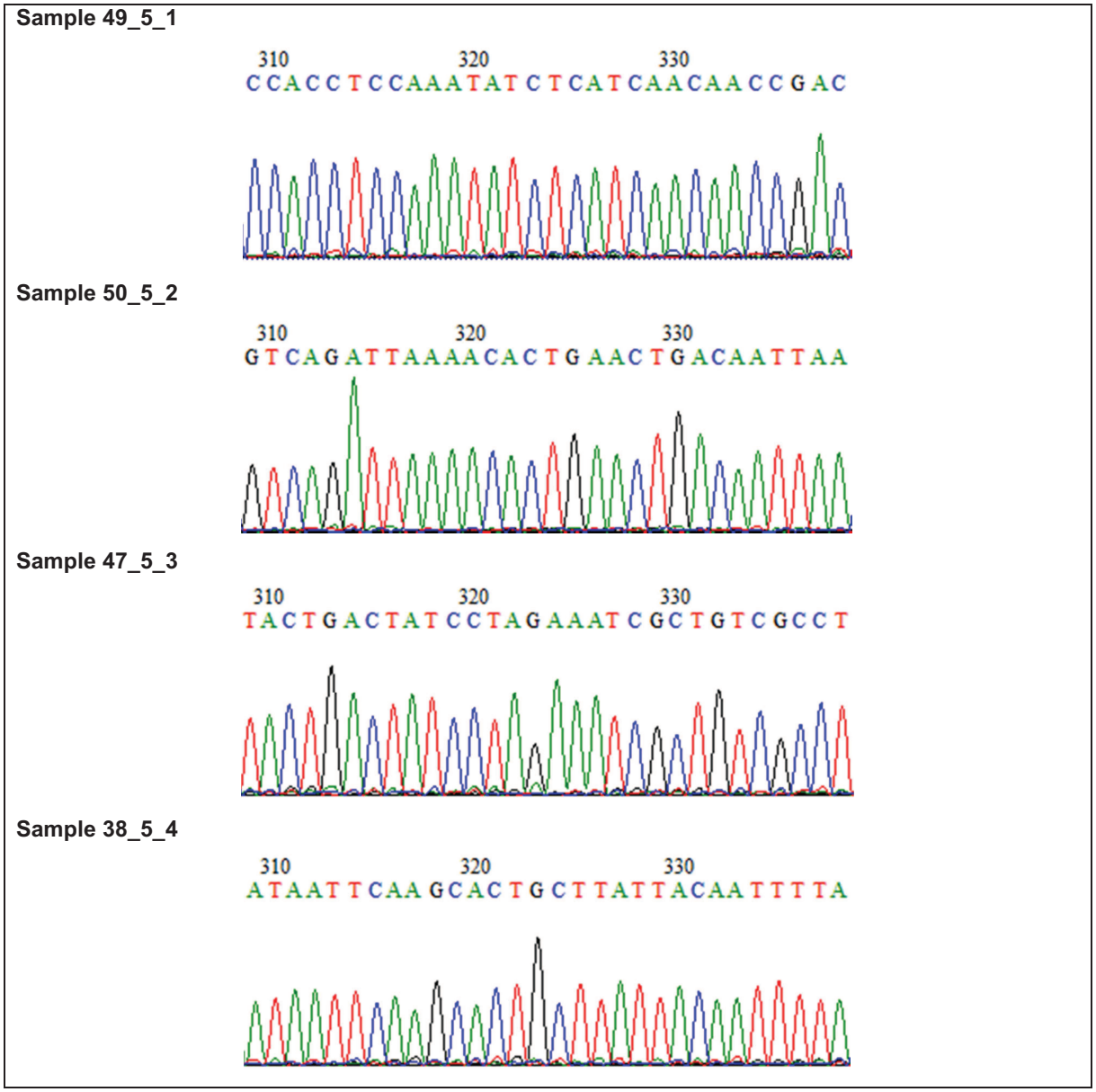
Figure 6.46 Photographic representation of the amplified mtDNA product of primer region 5



Photograph of the agarose gel on which the mtDNA amplified product was electrophoresed at 100 volts (V) and 50 mAmperes (mA) for 30 minutes, as discussed in Section 5.5; ladder = FastRuler™ High Range DNA Ladder (Fermentas) of range 100 – 10,000 bp; included in the first lane of the gel; sample names refer to the Tswana-speaking individuals of this investigation.

Representative examples of the sequence electropherograms generated by the BigDye®[1] Terminator v3.1 Cycle Sequencing Kit for primer region 5 are represented in Figure 6.47. These results were viewed and edited by making use of BioEdit software version 7.0.5.2 (Hall, 2001).

Figure 6.47 Representative electropherograms of the sequence generated for primer region 5 using the forward primers 1-4



Examples of electropherogram data with peaks depicting nucleotides in the sequence region of primer 5; A = adenine; T = thymine; C = cytosine; G = guanine; numbering at the top of the electropherogram represents the numbering of the nucleotide as a sequenced fragment before alignment with the rCRS and therefore does not correspond to the nucleotide positions of primer region 5.

1 BigDye® Terminator v3.1 Cycle Sequencing Kit is a registered trademark of Applied Biosystems, Foster City, CA, USA.

6.6.5.1 Sequence alterations observed in primer region 5

Primer region 5 consists of 2,004 base pairs that include a section of the *COII* gene, the tRNA lysine (K) coding region, the *ATP8* gene, *ATP6* gene and a section of the *COIII* gene. The 296 base pairs of the *COII* gene contained in primer region 4 are discussed in this section in order to consider the *COII* gene region as a single unit. The *COIII* gene region stretches beyond the primer region 5 segment, with 104 base pairs into primer region 6 and is discussed as a single unit in Section 6.6.6. The observed sequence alterations of the *COII* gene region, starting at np 7586, the tRNA Lys (K) coding region, the *ATP8* gene region and the *ATP6* gene region that ends at np 9207, are presented in this section in Table 6.16.

Table 6.16 Sequence alterations observed between the complete mitochondrial DNA of the Tswana individuals included in this study and the rCRS in primer region 5

Position	Sequence alteration	Gene/ region	Frequency	Reference
7648	C-T	*COII*	1	Kivisild *et al.*, 2006
7741	T-C	*COII*	2	Coble *et al.*, 2004
7765	A-G	*COII*	1	Mishmar *et al.*, 2003
7771	A-G	*COII*	10	Mishmar *et al.*, 2003
8014	A-G	*COII*	1	Ingman *et al.*, 2000
8027	G-A	*COII*	2	Gonder *et al.*, 2007
8047	T-C	*COII*	1	Kivisild *et al.*, 2006
8113	C-A	*COII*	27	Mishmar *et al.*, 2003
8152	G-A	*COII*	27	Mishmar *et al.*, 2003
8191	A-G	*COII*	3	Behar *et al.*, 2008
8206	G-A	*COII*	10	Mishmar *et al.*, 2003
8248	A-G	*COII*	1	Mishmar *et al.*, 2003
8251	G-A	*COII*	27	Gonder *et al.*, 2007
8281	delC	Non-coding region	4	Soodyall *et al.*, 1996
8282	delC	Non-coding region	4	Soodyall *et al.*, 1996
8283	delC	Non-coding region	4	Soodyall *et al.*, 1996
8284	delC	Non-coding region	4	Soodyall *et al.*, 1996
8285	delC	Non-coding region	4	Soodyall *et al.*, 1996
8286	delT	Non-coding region	4	Soodyall *et al.*, 1996
8287	delC	Non-coding region	4	Soodyall *et al.*, 1996
8288	delT	Non-coding region	4	Soodyall *et al.*, 1996
8289	delA	Non-coding region	4	Soodyall *et al.*, 1996
8290	G-A	Non-coding region	5	Behar *et al.* 2008
8383	T-C	*ATP8*	1	Kivisild *et al.*, 2006
8392	G-A	*ATP8*	10	Mishmar *et al.*, 2003
8428	C-T	*ATP8*	8	Mishmar *et al.*, 2003

Table 6.16 Continued…

Position	Sequence alteration	Gene/ region	Frequency	Reference
8459	A-G	*ATP8*	5	Behar *et al.*, 2008
8460	A-G	*ATP8*	4	Behar *et al.,* 2008
8468	C-T	*ATP8*	37	Mishmar *et al.*, 2003
8485	G-A	*ATP8*	1	Kivisild *et al.*, 2006
8545	G-A	*ATP8 / 6*	9	Mishmar *et al.*, 2003
8566	A-G	*ATP8 / 6*	8	Mishmar *et al.*, 2003
8577	A-G	*ATP6*	2	Behar *et al.* 2008
8598	T-C	*ATP6*	5	Behar *et al.* 2008
8618	T-C	*ATP6*	1	Mishmar *et al.*, 2003
8655	C-T	*ATP6*	38	Mishmar *et al.*, 2003
8701	A-G	*ATP6*	50	Mishmar *et al.*, 2003
8784	A-G	*ATP6*	2	Ingman *et al.*, 2000
8793	T-C	*ATP6*	1	Kivisild *et al.*, 2006
8860	A-G	*ATP6*	50	Mishmar *et al.*, 2003
8877	T-C	*ATP6*	2	Ingman *et al.*, 2000
8911	T-C	*ATP6*	1	Ingman *et al.*, 2000
8987	T-C	*ATP6*	1	Herrnstadt *et al.*, 2002
8994	G-A	*ATP6*	3	Ingman *et al.*, 2000
9027	C-T	*ATP6*	5	Gonder *et al.*, 2007
9039	G-A	*ATP6*	1	Tanaka *et al.*, 2004
9042	C-T	*ATP6*	36	Mishmar *et al.*, 2003
9058	A-G	*ATP6*	1	Tanaka *et al.*, 2004
9072	A-G	*ATP6*	2	Mishmar *et al.*, 2003
9111	T-C	*ATP6*	5	Behar *et al.* 2008
9136	A-G	*ATP6*	1	Behar *et al.* 2008
9150	A-G	*ATP6*	2	Behar *et al.* 2008
9181	A-G	*ATP6*	1	Derenko *et al.*, 2007

Transitions indicated in blue and transversions indicated in red. The frequency is indicated as the number of times the sequence variation was observed within the total Tswana dataset of 50 individuals; del = deletion indicated by the "del" followed by the nucleotide that was deleted.*COII* = cytochrome c oxidase subunit II gene; *ATP 8* = ATP synthase F0 subunit 8 gene; *ATP 6* = ATP synthase F0 subunit 6 gene.

The *COII* gene region displays a total of 13 sequence alterations that consist of 12 transitions, one transversion and no indels. The tRNA Lysine (K) region displays no sequence alterations. The *ATP8* gene region displays nine sequence alterations that consist only of transitions. The *ATP6* gene region displays 23 sequence alterations that also only consist of transitions. Two of the transitions, at np 8545 and np 8566, are shared between the *ATP8* and *ATP6* genes because of the overlap between the two gene regions. In total, 53 sequence alterations are observed in the 1,621 base pair primer region 5, which consists of 43 transitions, one transversion and nine deletions. The deletions are located in the non-coding region from np 8281 to np 8289 and are a reported

marker associated with the Bantu migrations about 4,000 ybp (Soodyall *et al.*, 1996). The deletions usually occur as a nine base pair deletions of the segment between np 8281 to np 8289 and are observed as reported in four of the Tswana-speaking individuals of this investigation. The non-coding region further displays one transition. Primer region 5 therefore displays a high number of transitions with a transversion:transition count ratio (Lutz-Bonengel *et al.*, 2003) of 1:53.

The *COII*, *ATP8* and *ATP6* gene regions have further been investigated by the mtDNA-GeneSyn tool developed by Pereira *et al*. (2009). The impact of the sequence alterations on the *COII*, *ATP8* and *ATP6* gene coding functionality has been determined and is presented in Table 6.17.

Table 6.17 Sequence variation within *COII*, *ATP8* and *ATP6* genes

Position	Sequence alteration	Gene	Syn/ Non-syn	Codon	New codon	Codon position	Amino acid	New amino acid
7648	C-T	*COII*	Syn	AUC	AUU	3	I	I
7741	T-C	*COII*	Syn	AAU	AAC	3	N	N
7765	A-G	*COII*	Syn	GAA	GAG	3	E	E
7771	A-G	*COII*	Syn	GAA	GAG	3	E	E
8014	A-G	*COII*	Syn	GUA	GUG	3	V	V
8027	G-A	*COII*	Non-syn	GCC	ACC	1	A	T
8047	T-C	*COII*	Syn	AUU	AUC	3	I	I
8113	C-A	*COII*	Syn	CCC	CCA	3	P	P
8152	G-A	*COII*	Syn	CCG	CCA	3	P	P
8191	A-G	*COII*	Syn	GCA	GCG	3	A	A
8206	G-A	*COII*	Syn	AUG	AUA	3	M	M
8248	A-G	*COII*	Syn	AUA	AUG	3	M	M
8251	G-A	*COII*	Syn	GGG	GGA	3	G	G
8383	T-C	*ATP8*	Syn	ACU	ACC	3	T	T
8392	G-A	*ATP8*	Syn	UGG	UGA	3	W	W
8428	C-T	*ATP8*	Syn	UUC	UUU	3	F	F
8459	A-G	*ATP8*	Non-syn	AAC	GAC	1	N	D
8460	A-G	*ATP8*	Non-syn	AAC	AGC	2	N	S
8468	C-T	*ATP8*	Syn	CUA	UUA	1	L	L
8485	G-A	*ATP8*	Syn	AAG	AAA	3	K	K
8545	G-A	*ATP8*	Non-syn	GCU	ACU	1	A	T
8545	G-A	*ATP6*	Syn	UCG	UCA	3	S	S
8566	A-G	*ATP8*	Non-syn	AUC	GUC	1	I	V
8566	A-G	*ATP6*	Syn	CAA	CAG	3	Q	Q
8577	A-G	*ATP6*	Syn	CUA	CUG	3	L	L
8598	T-C	*ATP6*	Syn	AUU	AUC	3	I	I
8618	T-C	*ATP6*	Non-syn	AUC	ACC	2	I	T
8655	C-T	*ATP6*	Syn	AUC	AUU	3	I	I

Table 6.17 Continued…

Position	Sequence alteration	Gene	Syn/ Non-syn	Codon	New codon	Codon position	Amino acid	New amino acid
8701	A-G	*ATP6*	Non-syn	ACC	GCC	1	T	A
8784	A-G	*ATP6*	Syn	GGA	GGG	3	G	G
8793	T-C	*ATP6*	Syn	CCU	CCC	3	P	P
8860	A-G	*ATP6*	Non-syn	ACA	GCA	1	T	A
8877	T-C	*ATP6*	Syn	UUU	UUC	3	F	F
8911	T-C	*ATP6*	Syn	UUA	CUA	1	L	L
8987	T-C	*ATP6*	Non-syn	AUA	ACA	2	M	T
8994	G-A	*ATP6*	Syn	CUG	CUA	3	L	L
9027	C-T	*ATP6*	Syn	GGC	GGU	3	G	G
9039	G-A	*ATP6*	Syn	AUG	AUA	3	M	M
9042	C-T	*ATP6*	Syn	CAC	CAU	3	H	H
9058	A-G	*ATP6*	Non-syn	ACC	GCC	1	T	A
9072	A-G	*ATP6*	Syn	UCA	UCG	3	S	S
9111	T-C	*ATP6*	Syn	AUU	AUC	3	I	I
9136	A-G	*ATP6*	Non-syn	AUC	GUC	1	I	V
9150	A-G	*ATP6*	Syn	UUA	UUG	3	L	L
9181	A-G	*ATP6*	Non-Syn	AGC	GGC	1	S	G

COII = Cytochrome c oxidase subunit 2; *ATP8* = ATP synthase F0 subunit 8; *ATP6* = ATP synthase F0 subunit 6; the sequence alterations described in terms of the synonymous or nonsynonymous nature of the changes, the codon position that was affected and the new amino acid that was coded for. A = adenine; T = thymine; C = cytosine; G = guanine; syn = synonymous; nonsyn = nonsynonymous; G = Glycine; L = Leucine; M = Methionine; V = Valine; F = Phenylalanine; W = Tryptophan; E = Glutamic acid (Glutamate); N = Asparagine; K = Lysine; T = Threonine; A = Alanine; I = Isoleucine; H = Histidine; S = Serine; P = Proline; Q= Glutamine.

Twelve of the sequence alterations in the *COII* gene are synonymous and one transition is nonsynonymous. The transition of G to A at np 8027 in the first codon position changes the coding for an alanine amino acid to a threonine amino acid. The *ATP8* gene displays five synonymous and four nonsynonymous sequence alterations. Three of the nonsynonymous sequence alterations are localised at first codon positions and one is localised at the second codon position. The *ATP6* gene displays 16 synonymous and seven nonsynonymous sequence alterations. Five of the nonsynonymous changes are localised at first codon positions and two at second codon positions. In total, the sequence alterations for the *COII*, *ATP8* and *ATP6* genes in primer region 5 consist of 33 synonymous and 12 nonsynonymous sequence alterations.Nine of the nonsynonymous changes are localised to first codon positions and four to second codon positions. The *ATP8* and *ATP6* genes display higher numbers of nonsynonymous sequence alterations than observed within the *COII* gene, which is expected according to the results of a large study of global mtDNA sequences by Pereira *et al.* (2009).

No novel sequence alterations have been observed in primer region 5. Six sequence alterations have, however, not been reported as being present in individuals harbouring L haplogroups but have been reported in individuals belonging to population groups other than African populations. All of the sequence alterations are transitions, except for the transversion at np 8113. Two transitions occur within the *COII* gene region at np 7741 and np 8014 and four transitions in the *ATP6* gene region at np 8793, np 9039, np 9058 and np 9181. Electropherograms of samples that contain these transitions, namely TS_2093, TS_5083, TS_3505, TS_3002, TS_5063 and TS_2075, are presented in Figure 6.48. No sequencing artefacts or background noise has been observed in the electropherograms of any of these samples. In addition, good peak morphology has been observed for all of these transitions, which provides reason to believe that these peaks have been called correctly. However, since human error and laboratory error cannot be ruled out, these novel mutations must be verified by re-sequencing of the regions of the samples in which they were observed.

Figure 6.48 Representative electropherograms of the sequence data generated indicating transitions at np 7741, np 8014, np 8793, np 9039, np 9058 and np 9181

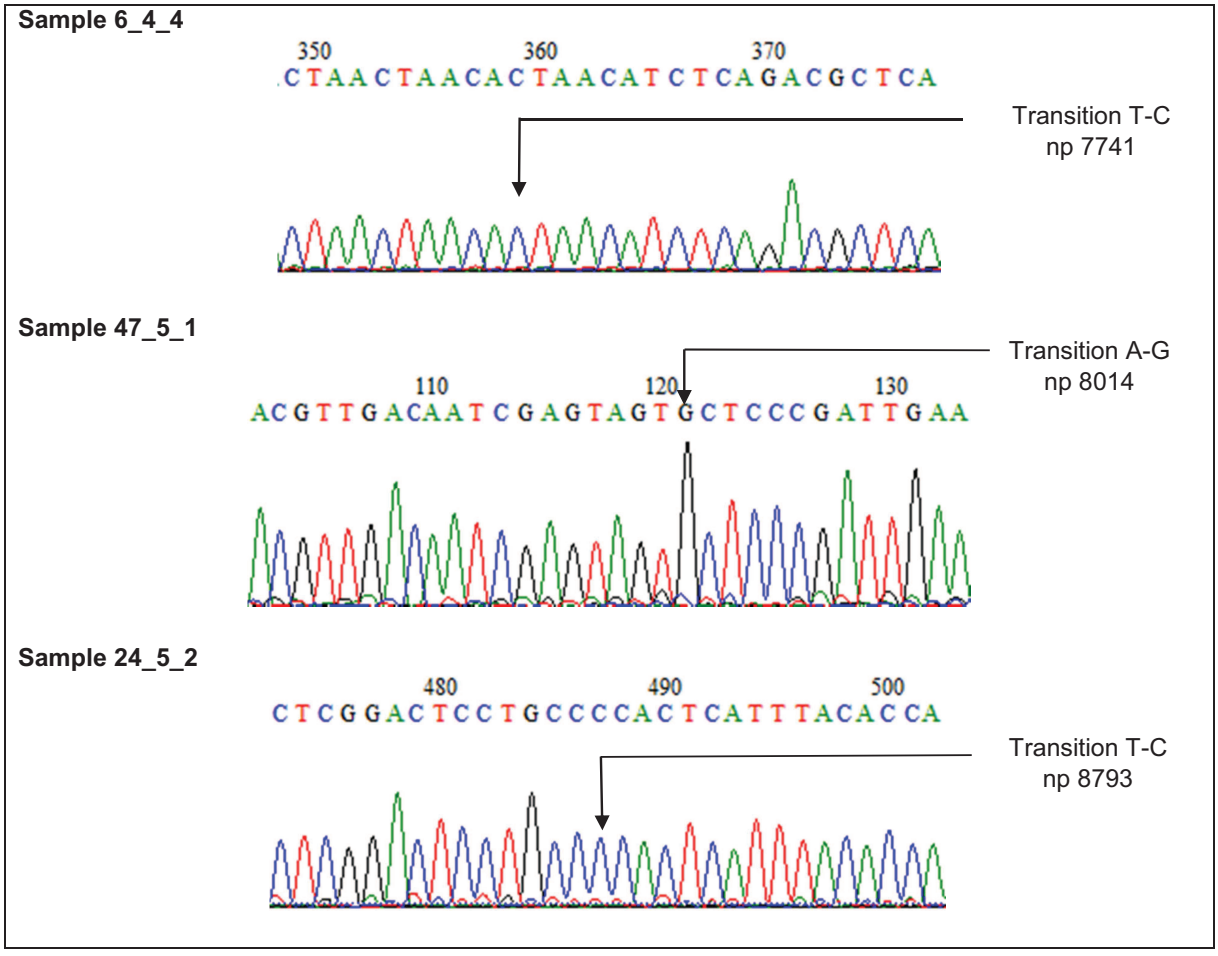
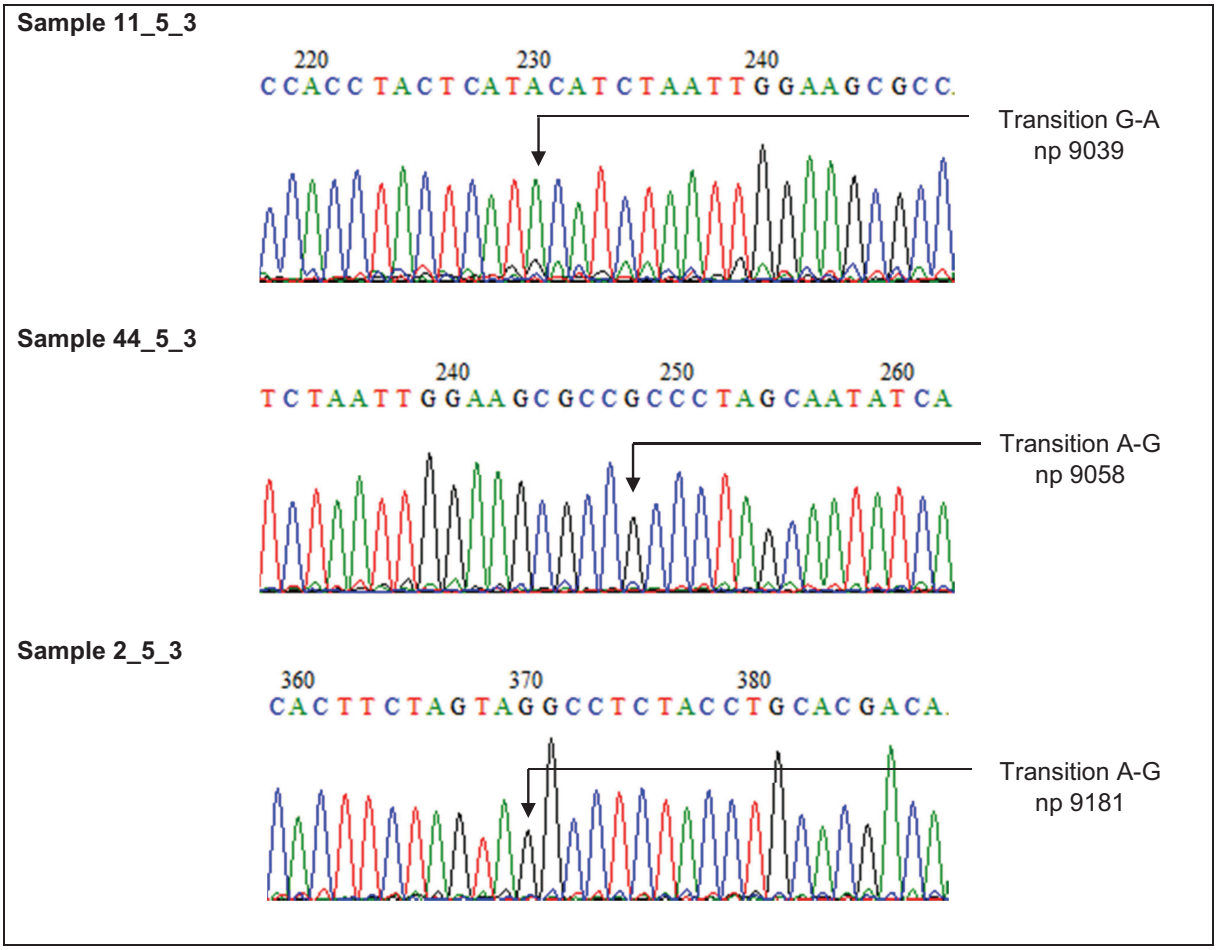


Figure 6.48 Continued…



Sample name 6_4_4 refers to sample TS_2093_6 primer region 4, sequence primer 4; sample name 47_5_1 refers to sample TS_5083_47, primer region 5, sequence primer 1; sample name 24_5_2 refers to sample TS_3505_24, primer region 5, sequence primer 2; sample name 11_5_3 refers to sample TS_3002_11, primer region 5, sequence primer 3; sample name 44_5_3 refers to sample TS_5063_44, primer region 5, sequence primer 3; sample name 2_5_3 refers to sample TS_2075_2, primer region 5, sequence primer 3; sequence regions presented containing transitions before alignment with rCRS, therefore numbering presented above is not in accordance with alteration position as determined by comparison to the rCRS.

The transition at np 8793 is cited as a haplogroup-defining sequence mutation for haplogroups C4a2, M10 and M42a and the transition at np 9181 is cited as a haplogroup-defining mutation for haplogroup D2b. Both of these transitions are cited by the PhyloTree classification system (Van Oven and Kayser, 2009). The transition at np 7741 was reported by Coble *et al.* (2004) in a single individual of European descent that belonged to haplogroup H. The transition at np 8014 was reported in individuals that originated from Evenki and belonged to haplogroup C2a and C4a, individuals from Italy that belonged to haplogroup H1 and Aboriginal Australians that belonged to haplogroup S1 (Ingman *et al*., 2000; Starikovskaya *et al.*, 2005; van Holst Pellekaan *et al.,* 2006; Kivisild *et al.*, 2006). The transition at np 9039 was reported in Japanese individuals that belonged to haplogroup D and in Finnish individuals that belonged to haplogroup H (Tanaka *et al*., 2004; Annunen-Rasila *et al*., 2006). The transition at np 9058 was also reported in Japanese individuals that belonged to haplogroup D, Chinese individuals that belonged to

haplogroup A and in individuals of European descent that belonged to haplogroup H1 (Herrnstadt *et al.*, 2002; Kong *et al.*, 2003; Tanaka *et al.*, 2004). The transition at np 9181 was reported in Jewish individuals that belonged to haplogroup T2f, individuals of European descent that belonged to haplogroup J and individuals from Evenk in Asia that belonged to haplogroup D2b (Coble *et al.*, 2004; Tamm *et al.*, 2007; Behar *et al.*, 2008). The presence of these transitions in haplogroup L lineages in individuals of African origin have not been reported and is therefore novel for the haplogroup L lineages. The presence of one of these novel transitions in individual TS_5063 is of some concern, as discussed in Section 6.6.4.1, where concerns were raised about the high number of novel mutations observed in this individual. As mentioned, it is strongly suggested that this mtDNA genome be re-sequenced to verify the sequencing results.

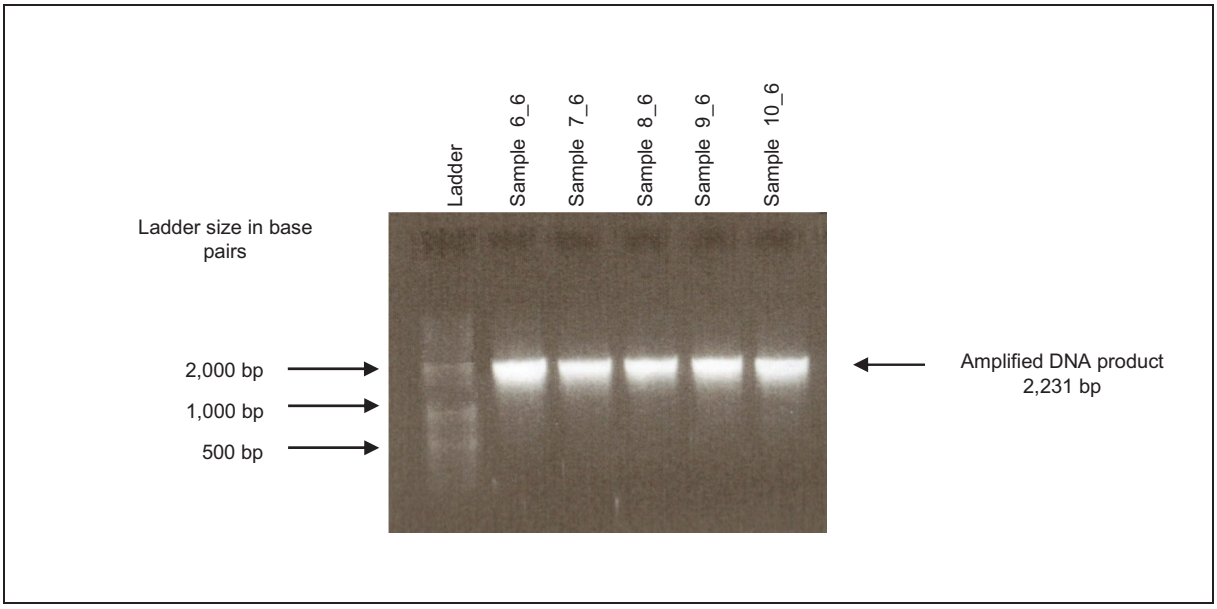
Only one of the transitions observed in primer region 5 is associated with disease. The transition at np 9058, which is also one of the mutations that is novel to haplogroup L lineages, has been reported possibly to be associated with LVNC cardiomyopathy-associated disease (Tang *et al.,* 2010). Other mutations associated with LVNC have been observed in the 12S rRNA and 16S rRNA-coding regions of the Tswana-speaking individuals of this investigation, as discussed in Section 6.6.2.1, and reaffirm the necessity of investigating the association of this disease with the variation present in the Tswana-speaking population of South Africa.

6.6.6 Primer region 6

Primer region 6 starts at np 9886 and ends at np 12076. It contains a 104 base pair segment of the *COIII* gene that stretches from np 9207 within the primer 5 region to np 9990 that is located in primer 6 region. Primer region 6 also contains the NADH *ND3* gene (np 10059-np 10404), *ND4L* gene (np 10470-np 10766) and a 1,316 base pair segment of *ND4* gene that stretches from np 10760 in primer region 6 to np 12137 that is located in primer region 7. Primer region 6 further contains the coding regions for the tRNA glycine (G) gene (np 9991-np 10058) and the tRNA arginine (R) gene (np 10405-np 10469). Primer region 6 containsno non-coding regions.

The region outlined above was amplified by PCR, as discussed in Section 5.4, and the PCR products were electrophoresed on an agarose gel to ascertain the quality of the product. A representative example of the mtDNA products for primer region 6, as visualised by the UVIvue ultraviolet transilluminator, are presented in Figure 6.49.

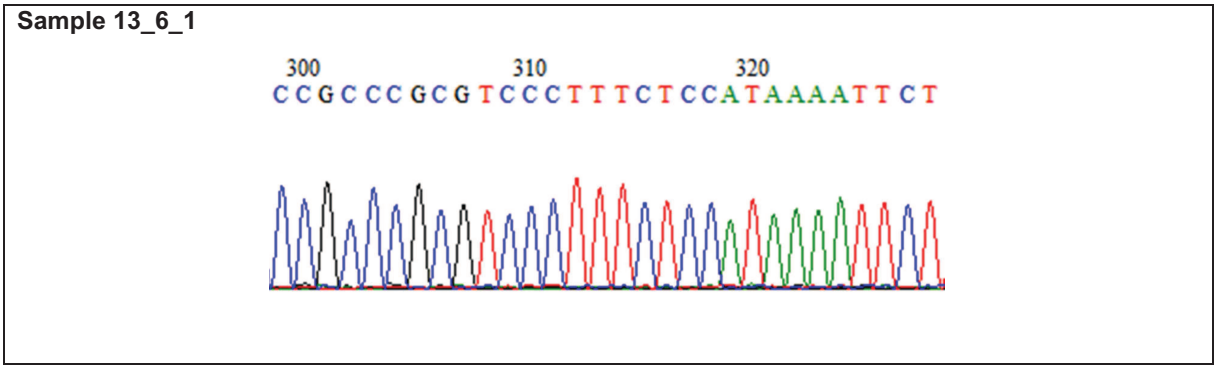
Figure 6.49 Photographic representation of the amplified mtDNA product of primer region 6



Photograph of the agarose gel on which the mtDNA amplified product was electrophoresed at 100 volts (V) and 50 mAmperes (mA) for thirty (30) minutes as discussed in Section 5.5; ladder = FastRuler™ High Range DNA Ladder (Fermentas) of range 100 – 10,000 bp; included in the first lane of the gel; sample names refer to the Tswana-speaking individuals of this investigation.

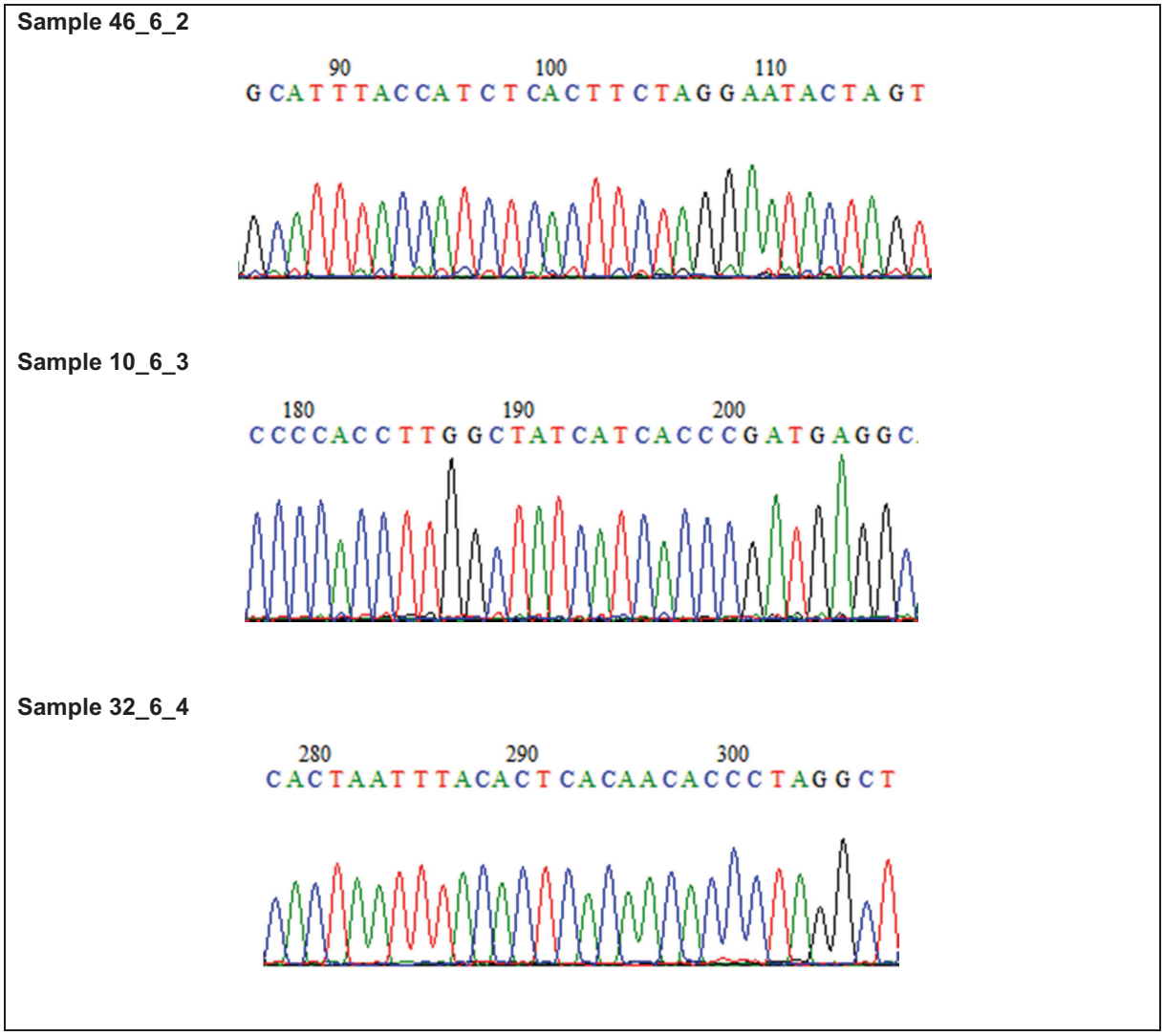
Representative examples of the electropherograms of the sequence data for primer region 6 generated using the BigDye®[1] Terminator v3.1 Cycle Sequencing Kit are indicated in Figure 6.50. These results were viewed and edited using the BioEdit software version 7.0.5.2 (Hall, 2001).

Figure 6.50 Representative electropherograms of the sequence generated for primer region 6 using the forward primers 1-4



1 BigDye® Terminator v3.1 Cycle Sequencing Kit is a registered trademark of Applied Biosystems, Foster City, CA, USA.

Figure 6.50 Continued…



Examples of electropherogram data with peaks depicting nucleotides in the sequence region of primer 6; A = adenine; T = thymine; C = cytosine; G = guanine; numbering at the top of the electropherogram represents the numbering of the nucleotides according to the sequenced fragment before alignment with the rCRS and therefore does not correspond to the nucleotide position of primer region 6.

6.6.6.1 Sequence alterations observed in primer region 6

Primer region 6 stretches from np 9886 to np 12076 and contains a segment of *COIII* gene, the *ND3* gene, the *ND4L* gene and a segment of the *ND4* gene. The *COIII* gene starts at np 9207, which is located within the primer region 5 segment and ends at np 9990, which is located in primer region 6. For the purpose of this discussion, the *COIII* gene is presented in this section as a single unit and the sequence alterations within the 679 base pairs of the *COIII* gene located within primer region 5 are included in this section. The *ND4* gene that starts at np 10760 and ends in primer 7 region at np 12137 is discussed as a single unit in Section 6.6.7 and thus the 1,316 base pair segment of *ND4* that is located in primer region 6 is not included in the discussion of sequence alterations in this section. The sequence alterations between np 9207 and np 10766, which include

the *COIII* gene, *ND3* gene, *ND4L* gene and the coding regions of tRNA glycine (G) and tRNA arginine (R), are presented in Table 6.18.

Table 6.18 Sequence alterations observed between the complete mitochondrial DNA of the Tswana individuals included in this study and the rCRS in primer region 6

Position	Sequence alteration	Gene/ region	Frequency	Reference
9221	A-G	*COIII*	10	Mishmar *et al.*, 2003
9278	C-T	*COIII*	1	Tanaka *et al.*, 2004
9297	**C-T**	***COIII***	**1**	**Current investigation**
9347	A-G	*COIII*	36	Mishmar *et al.*, 2003
9488	C-T	*COIII*	5	Gonder *et al.*, 2007
9530	T-C	*COIII*	1	Behar *et al.*, 2008
9540	T-C	*COIII*	50	Mishmar *et al.*, 2003
9545	A-G	*COIII*	4	Kivisild *et al.*, 2006
9554	G-A	*COIII*	4	Kivisild *et al.*, 2006
9755	G-A	*COIII*	30	Mishmar *et al.*, 2003
9758	T-C	*COIII*	1	Behar *et al.*, 2008
9801	G-A	*COIII*	1	Olivieri *et al.*, 2006
9818	C-T	*COIII*	9	Mishmar *et al.*, 2003
9833	T-C	*COIII*	1	Kivisild *et al.*, 2006
9860	C-T	*COIII*	1	Behar *et al.*, 2008
9932	G-A	*COIII*	1	Mishmar *et al.*, 2003
9950	T-C	*COIII*	4	Behar *et al.*, 2008
10084	T-C	*ND3*	1	Behar *et al.*, 2008
10114	T-C	*ND3*	5	Behar *et al.*, 2008
10115	T-C	*ND3*	10	Mishmar *et al.*, 2003
10128	C-A	*ND3*	5	Just *et al.*, 2008
10143	G-A	*ND3*	3	Gonder *et al.*, 2007
10237	T-C	*ND3*	1	Finnila *et al.*, 2001
10321	T-C	*ND3*	2	Mishmar *et al.*, 2003
10398	A-G	*ND3*	50	Mishmar *et al.*, 2003
10410	T-C	R (Acc-stem)	1	Kivisild *et al.*, 2006
10427	G-A	R (other position)	1	Tanaka *et al.*, 2004
10499	A-G	*ND4L*	1	Behar *et al.*, 2008
10586	G-A	*ND4L*	2	Behar *et al.*, 2008
10589	G-A	*ND4L*	36	Mishmar *et al.*, 2003
10640	T-C	*ND4L*	1	Behar *et al.*, 2008
10664	C-T	*ND4L*	36	Behar *et al.*, 2008
10688	G-A	*ND4L*	38	Mishmar *et al.*, 2003

Transitions are indicated in blue and transversions are indicated in red. The frequency is indicated as the number of times the sequence variation was observed within the total Tswana dataset of 50 individuals; del = deletion indicated by "del" followed by the nucleotide that was deleted; acc-stem = tRNA acceptor stem; other position = in another position than the D loop, the anticodon loop, the T-loop, the variable region and the acceptor stem; novel sequence alterations are indicated in **bold**.

The *COIII* gene, which is 783 base pairs long, displays 17 sequence alterations that consist of transitions only. The *ND3* gene, which is 345 base pairs long, displays eight sequence alterations that consist of seven transitions and one transversion. The tRNA glycine coding region displays no sequence alterations and the tRNA arginine coding region displays two sequence alterations, which consist of transitions only. The *ND4L* gene, which is 296 base pairs long, displays six sequence alterations that consist of transitions only. No indels have been observed in primer region 6. Therefore, the total number of observed sequence alterations in primer region 6 is 33, of which 32 are transitions and one is a transversion. Primer region 6 therefore displays a high number of transitions and a transversion:transition count ratio (Lutz Bonengel *et al.*, 2003) of 1:32.

The *COIII*, *ND3* and *ND4L* gene regions were further investigated by the mtDNA-GeneSyn tool developed by Pereira *et al*. (2009). The impact of the sequence alterations on the *COIII*, *ND3* and *ND4L* gene coding functionality was determined and is presented in Table 6.19.

Table 6.19 Sequence variation within *COIII*, *ND3* and *ND4L* genes

Position	Sequence alteration	Gene	Syn/ Non-syn	Codon	New codon	Codon position	Amino acid	New amino acid
9221	A-G	*COIII*	Syn	UCA	UCG	3	S	S
9278	C-T	*COIII*	Syn	GCC	GCU	3	A	A
9297	C-T	*COIII*	Syn	CUA	UUA	1	L	L
9347	A-G	*COIII*	Syn	CUA	CUG	3	L	L
9488	C-T	*COIII*	Syn	UUC	UUU	3	F	F
9530	T-C	*COIII*	Syn	CCU	CCC	3	P	P
9540	T-C	*COIII*	Syn	UUA	CUA	1	L	L
9545	A-G	*COIII*	Syn	GGA	GGG	3	G	G
9554	G-A	*COIII*	Syn	UGG	UGA	3	W	W
9755	G-A	*COIII*	Syn	GAG	GAA	3	E	E
9758	T-C	*COIII*	Syn	UCU	UCC	3	S	S
9801	G-A	*COIII*	Non-syn	GUA	AUA	1	V	M
9818	C-T	*COIII*	Syn	CAC	CAU	3	H	H
9833	T-C	*COIII*	Syn	AUU	AUC	3	I	I
9860	C-T	*COIII*	Syn	UGC	UGU	3	C	C
9932	G-A	*COIII*	Syn	UGG	UGA	3	W	W
9950	T-C	*COIII*	Syn	GUU	GUC	3	V	V
10084	T-C	*ND3*	Non-syn	AUC	ACC	2	I	T
10114	T-C	*ND3*	Non-syn	AUU	ACU	2	I	T
10115	T-C	*ND3*	Syn	AUU	AUC	3	I	I

Table 6.19 Continued…

Position	Sequence alteration	Gene	Syn/ Non-syn	Codon	New codon	Codon position	Amino acid	New amino acid
10128	C-A	*ND3*	Non-syn	CUA	AUA	1	L	M
10143	G-A	*ND3*	Non-syn	GGC	AGC	1	G	S
10237	T-C	*ND3*	Non-syn	AUU	ACU	2	I	T
10321	T-C	*ND3*	Non-syn	GUU	GCU	2	V	A
10398	A-G	*ND3*	Non-syn	ACC	GCC	1	T	A
10499	A-G	*ND4L*	Syn	CUA	CUG	3	L	L
10586	G-A	*ND4L*	Syn	UCG	UCA	3	S	S
10589	G-A	*ND4L*	Syn	CUG	CUA	3	L	L
10640	T-C	*ND4L*	Syn	AAU	AAC	3	N	N
10664	C-T	*ND4L*	Syn	GUC	GUU	3	V	V
10688	G-A	*ND4L*	Syn	GUG	GUA	3	V	V

COIII = Cytochrome c oxidase subunit 3; *ND3* = NADH dehydrogenase subunit 3; *ND4L* = NADH dehydrogenase subunit 4L; the sequence alterations described in terms of the synonymous or nonsynonymous nature of the changes, the codon position that was affected and the new amino acid that was coded for; A = adenine; T = thymine; C = cytosine; G = guanine; syn = synonymous; nonsyn = nonsynonymous; C = Cysteine; G = Glycine; L = Leucine; M = Methionine; V = Valine; F = Phenylalanine; W = Tryptophan; E = Glutamic acid (Glutamate); N = Asparagine; T = Threonine; A = Alanine; I = Isoleucine; H = Histidine; S = Serine; P = Proline.

The *COIII* gene contains one nonsynonymous transition at np 9801, which is located at the first codon position and changes the amino acid from valine to methionine. The other 16 transitions observed in this region are synonymous changes at the third codon position. All the transitions observed in the *ND4L* gene are synonymous changes at the third codon position.

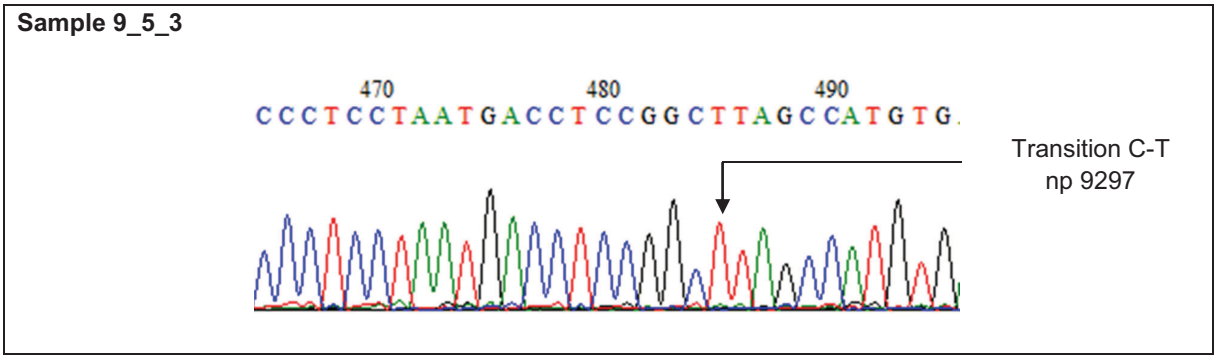
The *ND3* gene contains seven nonsynonymous and one synonymous sequence alteration, of which three are located in the first codon position and four in the second codon position. The large number of nonsynonymous substitutions was not expected, as the frequencies of synonymous substitutions were reported to be at least double or more than nonsynonymous substitutions in the protein-coding regions of the mitochondrial genome (Moilanen and Majamaa, 2003; Pereira *et al*., 2009). Nonsynonymous substitutions that occur in mtDNA sequences that are positioned at a deep level in a phylogenetic tree and are therefore older, have been observed less often than nonsynonymous changes observed in mtDNA sequences positioned at the tips of the phylogenetic tree (Elson *et al*., 2004). This is ascribed to negative selection that removes deleterious substitutions over time. It is to be expected that most of the deleterious nonsynonymous substitutions that arise in individual mtDNA sequences as private mutations, will be removed from mtDNA sequences over time. The large number of nonsynonymous substitutions observed in the *ND3* gene of the Tswana-speaking individuals of this investigation could be ascribed to recent private mutations in the Tswana-speaking individuals that have not yet been

removed by positive selection. This possibility is supported by Kivisild *et al.* (2006) who report that the proportion of nonsynonymous substitutions to synonymous substitutions rises drastically when younger clades are observed, as could be the case in this investigation.

The presence of the large number of nonsynonymous substitutions observed in the *ND3* gene region of this investigation could, however, also be the result of positive selection that has conserved advantageous mutations in a changing environment or diet (Nielsen *et al*., 2007). Although many previous studies have rejected claims of directional positive selection in the protein-coding regions of the mtDNA genome (Mishmar *et al.*, 2003; Elson *et al.*, 2004), several nucleotide positions in the protein-coding regions of the mtDNA genome have been identified as hotspots for mutational activity and are susceptible to site-specific positive selection (Kivisild *et al.*, 2006). Four of the nonsynonymous substitutions within the *ND3* gene of the Tswana-speaking individuals of this investigation, at np 10398, np 10084, np 10143 and np 10321, have been reported as mutational hotspots (Kivisild *et al.*, 2006) and are very likely to have been subjected to positive selection, thus preserving the nonsynonymous substitutions in the lineages.

One novel mutation is observed in primer region 6 at np 9297 within the *COIII* gene region at the third codon position. The electropherogram of the mtDNA sequence that contains this transition, which is observed in one Tswana-speaking individual of this investigation i.e. TS_ 4063, is presented in Figure 6.51. The electropherogram displays good peak morphology and no evidence of sequencing artefacts or background noise. The possibility that the transition has been called incorrectly because of the sequencing artefacts or noisy data has been ruled out although the mtDNA sequence must be re-sequenced to rule out the possibility of human or laboratory error.

Figure 6.51 Representative electropherograms of the sequence generated for the novel sequence alteration at np 9297



Sample name 9_5_3 refers to sample TS_4063_9 primer region 5, sequence primer 3; sequence region presented containing transition before alignment with rCRS, therefore numbering presented above is not in accordance with rCRS alteration position as determined by comparison to the rCRS.

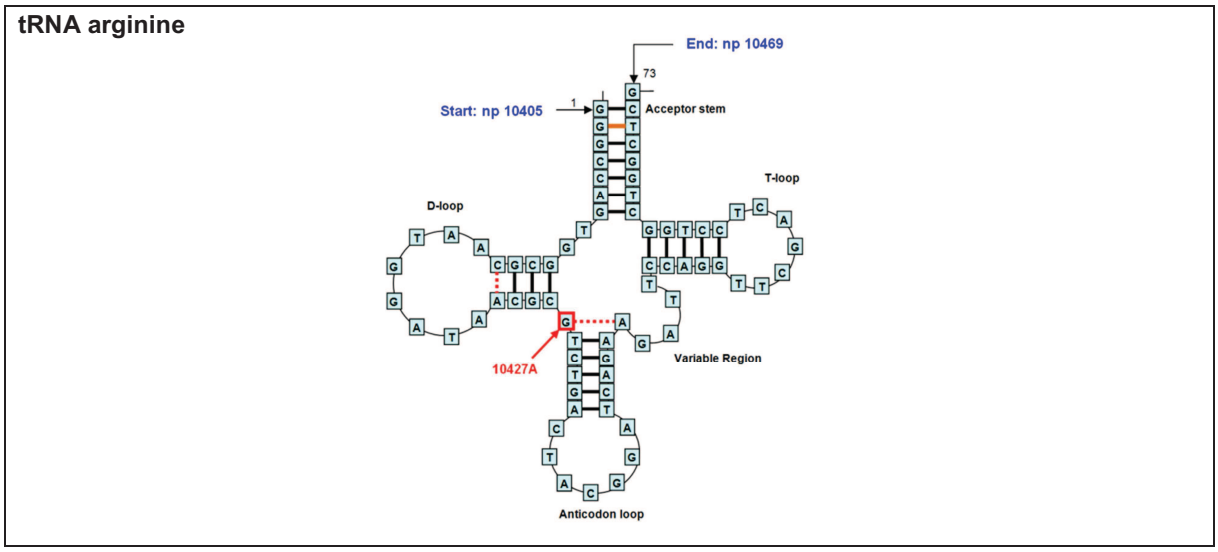
Three transitions, at np 9278 within the *COIII* gene region, np 10427 within the tRNA aspargine (R) coding region and at np 10237 within the *ND3* gene region, and a transversion at np 10128 within the *ND3* gene region have not previously been reported in haplogroup L-harbouring individuals of African origin. Thus, these alterations can be considered novel for haplogroup L lineages. The transition at np 9278 occurs at a third codon position within the *COIII* gene and is therefore synonymous. It has been reported in Japanese individuals belonging to haplogroup N as well as in Jewish individuals belonging to haplogroup W1d (Tanaka *et al.*, 2004; Behar *et al.*, 2008) and has been cited as a haplogroup-defining mutation for haplogroup W1d by the PhyloTree classification system (Van Oven and Kayser, 2009). It has been observed in one Tswana-speaking individual of this investigation i.e. TS_ 3066, and therefore most probably constitutes a private mutation, which is not connected to any of the haplogroups in which it has been reported.

The transversion at np 10128 occurs at a first codon position within the *ND3* gene, which causes a nonsynonymous amino acid change from leucine to methionine. It has been reported in American Hispanic individuals belongingto haplogroup A2 (Just *et al.*, 2008) and is not cited as haplogroup-defining by the PhyloTree classification system (Van Oven and Kayser, 2009). It is present in five Tswana-speaking individuals of this investigation, all belongingto haplogroup L0d3, and could therefore suggest a new sub-haplogroup within the L0d3 lineages. This sequence variant should be considered as a haplogroup-defining site in the haplogroup L0d3 clade.

The transition at np 10237 occurs at a second codon position in the *ND3* gene region of one Tswana-speaking individual of this investigation i.e. TS_ 4075. The transition at this position causes a nonsynonymous amino acid change from isoleucine to threonine. It has

been reported in Finnish individuals belonging to haplogroup J2 and has been cited as a haplogroup-defining mutation for haplogroup J2 by the PhyloTree classification system (Van Oven and Kayser, 2009). The presence of this sequence variant in a single Tswana-speaking individual suggests that it might have been a private mutation and not connected to haplogroup J2. Although novel to the haplogroup L lineages, it is possible that if the effect of the nonsynonymous change is deleterious, the mutation will be removed over time. The transition at np 10427 occurs within the coding region of the tRNA arginine (R) at position 26 between the D-loop and the anticodon loop, as presented in Figure 6.52.

Figure 6.52 Structure of the tRNA arginine



Structure of a typical tRNA arginine; ▬100% Watson Crick pairs; ▪ ▪100% mismatches; ▬100% GC pairs; numbers alongside the tRNA structure indicates nucleotide positions within the tRNA molecule and not according to the rCRS; nucleotide starting and ending positions of the tRNA according to the rCRS indicated in blue ink; nucleotide position indicated in red ink refers to the nucleotide variation observed in the Tswana-speaking individuals of this investigation. From Jühling *et al.*, 2009; Pereira *et al.*, 2009.

This mutation has been reported in Japanese individuals belongingto haplogroup D4 and D5, in Russian individuals belongingto haplogroup D7 and individuals from India belongingto haplogroup G3 (Ingman *et al.,* 2000; Tanaka *et al.*, 2004; Volodko *et al.,* 2008) and has been cited as a haplogroup-defining mutation for haplogroup D4 by the PhyloTree classification system (Van Oven and Kayser, 2009). It occurs in one Tswana-speaking individual of this investigation i.e., TS_ 5063, which is cause for concern, as was discussed in Section 6.6.4.1. This individual displays four haplogroup L novel sequence alterations and one novel sequence alteration in addition to the latter transition in the tRNA arginine coding region, which is highly unlikely for a single individual. As was discussed, it is strongly suggested that this mtDNA genome be re-sequenced to verify the sequencing results.

The electropherograms of individuals harbouring the novel mutations for haplogroup L lineages observed in the Tswana-speaking individuals of this investigation as discussed previously are presented in Figure 6.53. The substitutions at np 9278 are observed in individual TS_ 3066, the alteration at np 10128 is observed in individual TS_2082, the alteration at np 10237 is observed in individual TS_ 4075 and the alteration at np 10427 is observed in individual TS_ 5063. No sequencing artefacts or background noise has been observed in any of the electropherograms, which rules out the possibility that the peaks have been called incorrectly because of sequencing artefacts, noise or poor peak morphology. Re-sequencing the samples would verify the novel mutations and rule out the possibility of human or laboratory error.

Figure 6.53 Representative electropherograms of the sequence data generated indicating transitions at np 9278, np 10237 and np 10427 and a transversion at np 10128

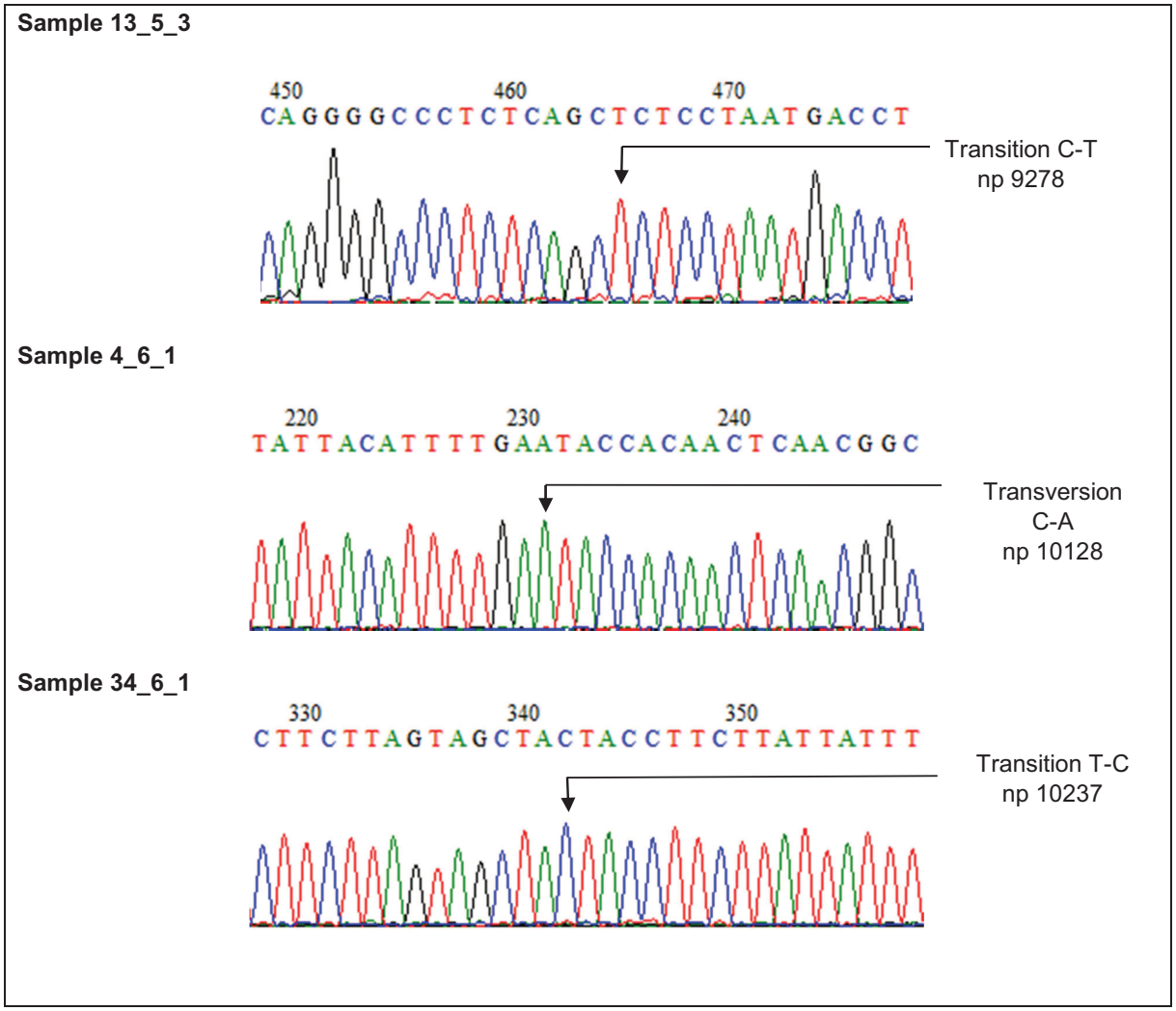
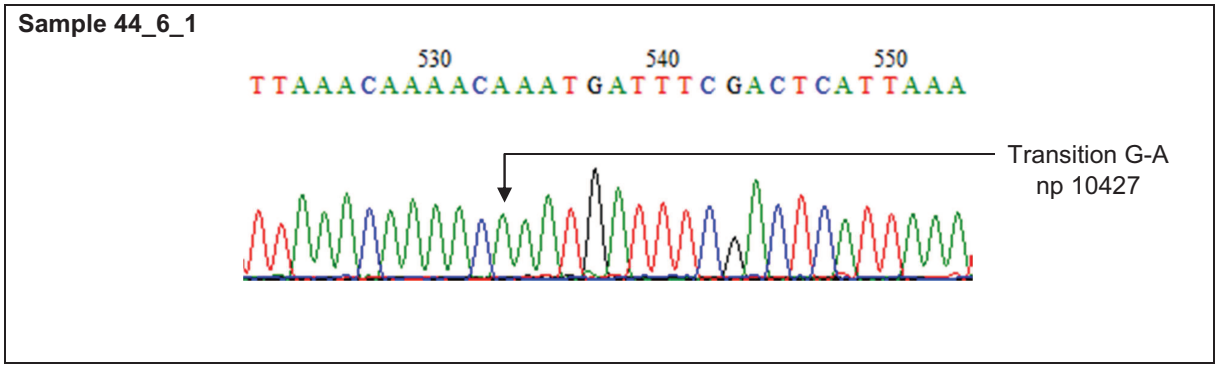


Figure 6.53 Continued…



Sample name 13_5_3 refers to sample TS_3066_13 primer region 5, sequence primer 3; sample name 4_6_1 refers to sample TS_2082_4, primer region 6, sequence primer 1; sample name 34_6_1 refers to sample TS_4075_34, primer region 6, sequence primer 1; sample name 44_6_1 refers to sample TS_5063_44, primer region 6, sequence primer 1; sequence regions presented containing transitions before alignment with rCRS, therefore numbering presented above is not in accordance with rCRS alteration position as determined by comparison to the rCRS.

The sequence variants that were reported to be associated with diseases are presented in Table 6.20. Two mutations, at np 10237 and np 10398, which are located in the *ND3* gene region, have previously been associated with diseases.

Table 6.20 Reported mtDNA sequence alterations within primer region 6 with disease associations

Locus	Sequence alteration	Number of individuals	Disease	Reference
ND3	T10237C	1	LHON	Horvath *et al.*, 2002
ND3	A10398G	50	PD Breast cancer Longevity	Pyle *et al.*, 2005 Covarrubias *et al.*, 2008 Niemi *et al.*, 2005

Sequence alterations are displayed as the position at which the mutation occurred with the wild type nucleotide indicated in front of the np and the mutant type nucleotide indicated after the np; the number of individuals = number of Tswana-speaking individuals of this investigation which displayed the mutation; disease associations were reported in one or more publications and have been considered as possibly pathologic as stated in MITOMAP. LHON = Leber hereditary optic neuropathy; PD = Parkinson's disease. Adapted from MITOMAP: A Human Mitochondrial Genome Database. http://www.mitomap.org, 2011.

Studies of Hungarian patients suffering from LHON presented evidence of a transition at np 10237 that was a strong candidate for involvement in the LHON disease etiology. However, further investigation about its occurrence in LHON and control populations is required to verify its role in the pathogenesis of the disease. It is therefore not likely that this transition in a single Tswana-speaking individual of this investigation has provided strong evidence for a connection between haplogroup L lineages and this disease.

The *ND3* gene is an important factor in the aetiology of Parkinson's disease in Caucasian individuals that belong to haplogroups U, J, K, and T and the transition of G to A at np 10398 is associated with a protective effect against Parkinson's disease (van der Walt *et al.*, 2003). In a further study based on the presence of the latter mutation in cases of

Parkinson's disease, haplogroups U, K, J and T have been associated with less risk of the development of Parkinson's disease, which further includes the T4216C and A4366G mutations (Pyle *et al.*, 2005). None of these mutations have been observed in the Tswana-speaking individuals of this investigation and no studies have reported on similar observations of reduced risk of Parkinson's disease with the presence of the np 10398 mutation in the African haplogroup L lineages. This is not surprising, since np 10398 distinguishes macrohaplogroup N from L an M. Further studies would be required to determine whether there is a possibility of a similar protective effect of this mutation against Parkinson's disease in haplogroup L lineages.

The 10398A variant is associated with susceptibility to breast cancer, especially in African-American women, according to a study by Canter *et al*. (2005) and breast and oesophageal cancer in Indians (Darvishi *et al.*, 2007). The transition of A to G at np 10398 is associated with a higher risk of developing breast cancer in European American women (Bai *et al*., 2007). Subsequent studies have cast some doubt on these reports when the results of increased susceptibility of breast cancer for both alleles of np 10398 could not be duplicated (Setiawan *et al.*, 2008). The discrepancies in the results have been ascribed to the fact that the type of sequence variants at np 10398 that affect women's susceptibility to developing breast cancer depend on the haplogroup to which they belong (Setiawan *et al.*, 2008). The sequence variant at np 10398 is involved in defining some sub-clades in haplogroup L1, L3 and N according to the PhyloTree classification system (Van Oven and Kayser, 2009) and as was demonstrated in the study of Setiawan *et al*. (2008), is an important factor to take into account when investigating the pathogenicity of this sequence variant for breast cancer. Since all the Tswana-speaking individuals of this investigation display this sequence variant, indicating that it has been present in the African haplogroup L lineages since early times, the 10398G variant should be regarded as the reference nucleotide and the 10398A allele should be viewed as the mutant in studies to investigate susceptibility to breast cancer in light of the reported disease associations.

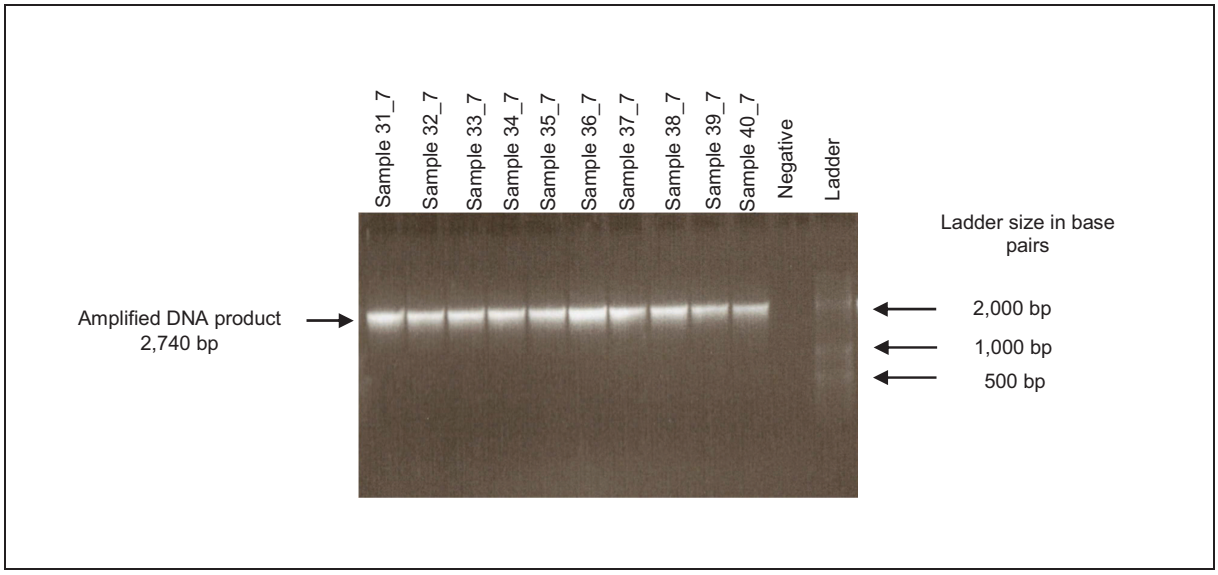
The sequence variant at np 10398 has also been associated with longevity. Longevity is, however, a complex trait that depends on the interaction between several different mtDNA genes and loci and on environmental exposure during life. Three nucleotide positions, i.e. 150T, 489C and 10398G, were identified in Finnish and Japanese populations, which played a role in longevity (Niemi *et al.*, 2005). No studies on longevity in African haplogroup L individuals have been performed that included np 10398.

6.6.7 Primer region 7

Primer region 7 starts at np 11487 and ends at np 13612. It contains a 651 base pair segment of the *ND4* gene (np 10760-np 12137), the coding regions of the tRNA histidine (H) at np 12138-np 12206, tRNA serine2 (S(AGY)) at np 12207-np 12265 and tRNA leucine2 (L(CUN)) at np 12266-np 12336. It also contains a 1,275 base pair segment of the *ND5* gene that starts at np 12337 and ends within primer region 8 at np 14148.

The region outlined above was amplified by PCR, as discussed in Section 5.4, and the PCR products were electrophoresed on an agarose gel to ascertain the quality of the product. A representative example of the mtDNA products for primer region 7, as visualised by the UVIvue ultraviolet transilluminator, is presented in Figure 6.54.

Figure 6.54 Photographic representation of the amplified mtDNA product of primer region 7

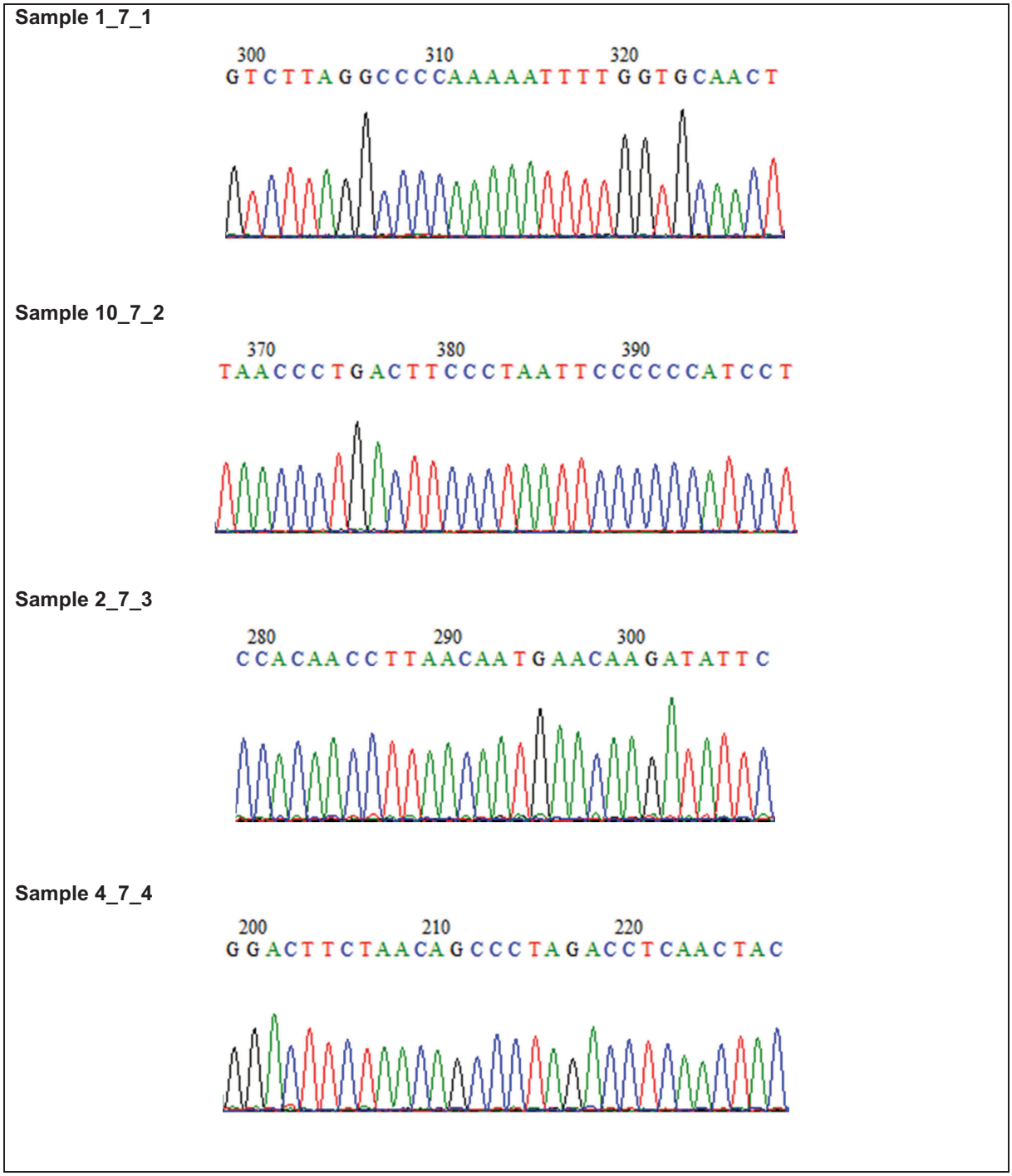


Photograph of the agarose gel on which the mtDNA amplified product was electrophoresed at 100 volts (V) and 50 mAmperes (mA) for 30 minutes as discussed in Section 5.5; ladder = FastRuler™ High Range DNA ladder (Fermentas) of range 100 – 10,000 bp; included in the last lane of the gel; negative = negative control; sample names refer to the Tswana-speaking individuals of this investigation.

Representative examples of the sequence electropherograms generated for primer region 7 using the BigDye®[1] Terminator v3.1 Cycle Sequencing Kit are indicated in Figure 6.55. These results were viewed and edited in BioEdit software version 7.0.5.2 (Hall, 2001).

1 BigDye® Terminator v3.1 Cycle Sequencing Kit is a registered trademark of Applied Biosystems, Foster City, CA, USA.

Figure 6.55 Representative electropherograms of the sequence generated for primer region 7 using the forward primers 1-4



Examples of electropherogram data with peaks depicting nucleotides in the sequence region of primer 7; A = adenine; T = thymine; C = cytosine; G = guanine; numbering at the top of the electropherogram represents the numbering of the nucleotides according to the sequenced fragment before alignment with the rCRS and therefore does not correspond to the nucleotide positions of primer region 7.

6.6.7.1 Sequence alterations observed in primer region 7

Primer region 7 contains a 651 base pair segment of the *ND4* gene that starts in primer region 6 at np 10760 and ends in primer region 7 at np 12137. For the purpose of this discussion, the *ND4* gene region will be treated as a single unit. Therefore sequence

alterations observed from np 10760 are presented in this section, which includes the *ND4* gene, the tRNA histidine (H), tRNA serine 2 (S(AGY)) and tRNA leucine 2 (L(CUN)) coding regions. The *ND5* gene region starts at np 12337 in primer region 7 and ends at np 14148, which is located in primer region 8. The 1,275 base pair segment of the *ND5* gene located in primer region 7 is considered in Section 6.6.8 to enable the discussion of the sequence alterations of the *ND5* gene as a single unit. Therefore, the observed sequence alterations from np 10760 to np 12337 are discussed in this section as presented in Table 6.21.

Table 6.21 Sequence alterations observed between the complete mitochondrial DNA of the Tswana individuals included in this study and the rCRS in primer region 7

Position	Sequence alteration	Gene/ region	Frequency	Reference
10792	A-G	*ND4*	2	Ingman *et al.*, 2000
10793	C-T	*ND4*	2	Ingman *et al.*, 2000
10810	T-C	*ND4*	38	Mishmar *et al.*, 2003
10819	A-G	*ND4*	1	Kivisild *et al.*, 2006
10873	T-C	*ND4*	50	Mishmar *et al.*, 2003
10876	A-G	*ND4*	1	Ingman *et al.*, 2000
10915	T-C	*ND4*	37	Mishmar *et al.*, 2003
10920	C-T	*ND4*	6	Ingman *et al.*, 2000
10939	C-T	*ND4*	1	Ingman *et al.*, 2000
10948	**C-T**	***ND4***	**1**	**Current investigation**
10966	T-C	*ND4*	1	Mishmar *et al.*, 2003
11002	A-G	*ND4*	1	Mishmar *et al.*, 2003
11061	C-T	*ND4*	5	Behar *et al.*, 2008
11101	A-G	*ND4*	1	Behar *et al.*, 2008
11143	C-T	*ND4*	4	Gonder *et al.*, 2007
11167	A-G	*ND4*	2	Quintana-Murci *et al.*, 2008
11172	A-G	*ND4*	4	Kivisild *et al.*, 2006
11176	G-A	*ND4*	12	Mishmar *et al.*, 2003
11260	T-C	*ND4*	1	Behar *et al.*, 2008
11296	C-T	*ND4*	1	Gonder *et al.*, 2007
11299	T-C	*ND4*	1	Behar *et al.*, 2008
11329	A-G	*ND4*	2	Behar *et al.*, 2008
11386	T-C	*ND4*	1	Behar *et al.*, 2008
11437	T-C	*ND4*	2	Gonder *et al.*, 2007
11557	A-G	*ND4*	1	Derenko *et al.*, 2007
11611	G-A	*ND4*	1	Herrnstadt *et al.*, 2002
11641	A-G	*ND4*	8	Gonder *et al.*, 2007
11653	A-G	*ND4*	1	Behar *et al.*, 2008
11654	A-G	*ND4*	2	Behar *et al.*, 2008
11719	G-A	*ND4*	50	Mishmar *et al.*, 2003

Table 6.21 Continued…

Position	Sequence alteration	Gene/ region	Frequency	Reference
11854	T-C	*ND4*	10	Behar *et al.*, 2008
11864	T-C	*ND4*	1	Kivisild *et al.*, 2006
11887	G-A	*ND4*	1	Kivisild *et al.*, 2006
11914	G-A	*ND4*	46	Mishmar *et al.*, 2003
11944	T-C	*ND4*	10	Mishmar *et al.*, 2003
12004	**A-G**	***ND4***	**1**	**Current investigation**
12007	G-A	*ND4*	36	Mishmar *et al.*, 2003
12049	C-T	*ND4*	2	Ingman *et al.*, 2000
12070	G-A	*ND4*	1	Ingman *et al.*, 2000
12121	T-C	*ND4*	27	Mishmar *et al.*, 2003
12127	G-A	*ND4*	3	Behar *et al.*, 2008
12142	A-G	H (Acc-stem)	4	Behar *et al.*, 2008
12172	A-G	H (Anticd-loop)	10	Mishmar *et al.*, 2003
12192	G-A	H (T-loop)	1	Behar *et al.*, 2008
12234	A-G	S(AGY) (Ac-stem)	9	Mishmar *et al.*, 2003
12235	T-C	S(AGY) (Ac-stem)	2	Behar *et al.*, 2008

Transitions are indicated in blue and transversions are indicated in red. The frequency is indicated as the number of times the sequence variation was observed within the total Tswana dataset of 50 individuals; *ND4* = NADH dehydrogenase subunit 4 gene; acc-stem = tRNA acceptor stem; Anticd-loop = tRNA anticodon loop; Ac-stem = tRNA anticodon stem.

Forty-six sequence alterations have been observed in the 1,576 base pairs in primer region 7 of the Tswana-speaking individuals of this investigation, of which all are transitions. No transversions or indels have been observed in this primer region. Forty-one of these transitions have been observed in the *ND4* gene, three transitions in the acceptor-stem, aniticodon-loop and T-loop regions, respectively of the tRNA histidine (H) gene, and two transitions in the tRNA serine 2(S(AGY)) acceptor-stem region.

The *ND4* gene region was further investigated by the mtDNA-GeneSyn tool developed by Pereira *et al*. (2009). The impact of the sequence alterations on the *ND4* gene coding functionality was determined and is presented in Table 6.22.

Table 6.22 Sequence variation within the *ND4* gene

Position	Sequence alteration	Gene	Syn/ Non-syn	Codon	New codon	Codon position	Amino acid	New amino acid
10792	A-G	*ND4*	Syn	UUA	UUG	3	L	L
10793	C-T	*ND4*	Syn	CUA	UUA	1	L	L
10810	T-C	*ND4*	Syn	CUU	CUC	3	L	L
10819	A-G	*ND4*	Syn	AAA	AAG	3	K	K
10873	T-C	*ND4*	Syn	CCU	CCC	3	P	P

Table 6.22 Continue…

Position	Sequence alteration	Gene	Syn/ Non-syn	Codon	New codon	Codon position	Amino acid	New amino acid
10876	A-G	*ND4*	Syn	CUA	CUG	3	L	L
10915	T-C	*ND4*	Syn	UGU	UGC	3	C	C
10920	C-T	*ND4*	Non-syn	CCA	CUA	2	P	L
10939	C-T	*ND4*	Syn	CCC	CCU	3	P	P
10948	C-T	*ND4*	Syn	ACC	ACU	3	T	T
10966	T-C	*ND4*	Syn	ACU	ACC	3	T	T
11002	A-G	*ND4*	Syn	CAA	CAG	3	Q	Q
11061	C-T	*ND4*	Non-syn	UCC	UUC	2	S	F
11101	A-G	*ND4*	Syn	GAA	GAG	3	E	E
11143	C-T	*ND4*	Syn	CCC	CCU	3	P	P
11167	A-G	*ND4*	Syn	UGA	UGG	3	W	W
11172	A-G	*ND4*	Non-syn	AAC	AGC	2	N	S
11176	G-A	*ND4*	Syn	CAG	CAA	3	Q	Q
11260	T-C	*ND4*	Syn	ACU	ACC	3	T	T
11296	C-T	*ND4*	Syn	CUC	CUU	3	L	L
11299	T-C	*ND4*	Syn	ACU	ACC	3	T	T
11329	A-G	*ND4*	Syn	UGA	UGG	3	W	W
11386	T-C	*ND4*	Syn	CUU	CUC	3	L	L
11437	T-C	*ND4*	Syn	GCU	GCC	3	A	A
11557	A-G	*ND4*	Syn	CUA	CUG	3	L	L
11611	G-A	*ND4*	Syn	UCG	UCA	3	S	S
11641	A-G	*ND4*	Syn	AUA	AUG	3	M	M
11653	A-G	*ND4*	Syn	GUA	GUG	3	V	V
11654	A-G	*ND4*	Non-syn	ACA	GCA	1	T	A
11719	G-A	*ND4*	Syn	GGG	GGA	3	G	G
11854	T-C	*ND4*	Syn	GCU	GCC	3	A	A
11864	T-C	*ND4*	Syn	UUA	CUA	1	L	L
11887	G-A	*ND4*	Syn	CUG	CUA	3	L	L
11914	G-A	*ND4*	Syn	ACG	ACA	3	T	T
11944	T-C	*ND4*	Syn	CUU	CUC	3	L	L
12004	A-G	*ND4*	Syn	CAA	CAG	3	Q	Q
12007	G-A	*ND4*	Syn	UGG	UGA	3	W	W
12049	C-T	*ND4*	Syn	UUC	UUU	3	F	F
12070	G-A	*ND4*	Syn	AUG	AUA	3	M	M
12121	T-C	*ND4*	Syn	AUU	AUC	3	I	I
12127	G-A	*ND4*	Syn	GGG	GGA	3	G	G

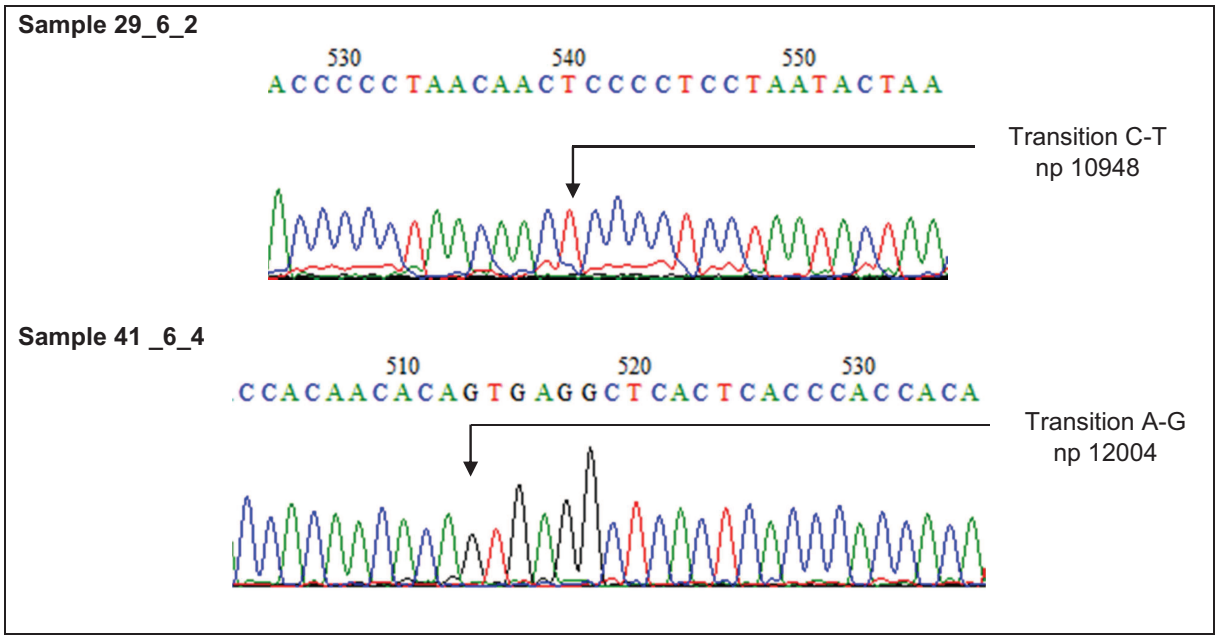
ND4 = NADH dehydrogenase subunit 4; the sequence alterations described in terms of the synonymous or nonsynonymous nature of the changes, the codon position that was affected and the new amino acid that was coded for; A = adenine; T = thymine; C = cytosine; G = guanine; syn = synonymous; nonsyn = nonsynonymous; A = Alanine; C = Cysteine; E = Glutamic acid (Glutamate); F = Phenylalanine; G = Glycine; I = Isoleucine; K = Lysine; L = Leucine; M = Methionine; N = Asparagine; P = Proline; S = Serine; T = Threonine; V = Valine; W = Tryptophan; Q = Glutamine.

The *ND4* gene displays 37 synonymous substitutions, of which 35 are located in the third codon position and two (2) located in the first codon position. Four (4) nonsynonymous

substitutions have been observed, of which one is located in a first codon position and three in a second codon position.

Two novel sequence variants have been observed in primer region 7, which are both located in the *ND4* gene region at np 10948 and np 12004 and are transitions, which cause synonymous amino acid changes.Each of them is present in one Tswana-speaking individual of this investigation, i.e. TS_4037 and TS_5044. The presence of these novel mutations, each in a single individual, suggests that these are recent private mutations and not suitable for consideration as novel haplogroup-defining sites. The electropherograms of the samples that contain these novel mutations are presented in Figure 6.56. Although some traces of pull-up are visible in the electropherogram of sample 29_6_2 (TS_4037), it does not interfere with the peak morphology and the possibility that this sequence variant has been incorrectly called has been ruled out. The electropherogram of sample_41_6_4 (TS_5044) displays no sequencing artefacts or background noise and the validity of this peak has also been confirmed. In order to verify the novelty of both the transitions, re-sequencing of the samples would be necessary.

Figure 6.56 Representative electropherograms of the sequence generated of novel sequence alterations at np 10948 and np 12004

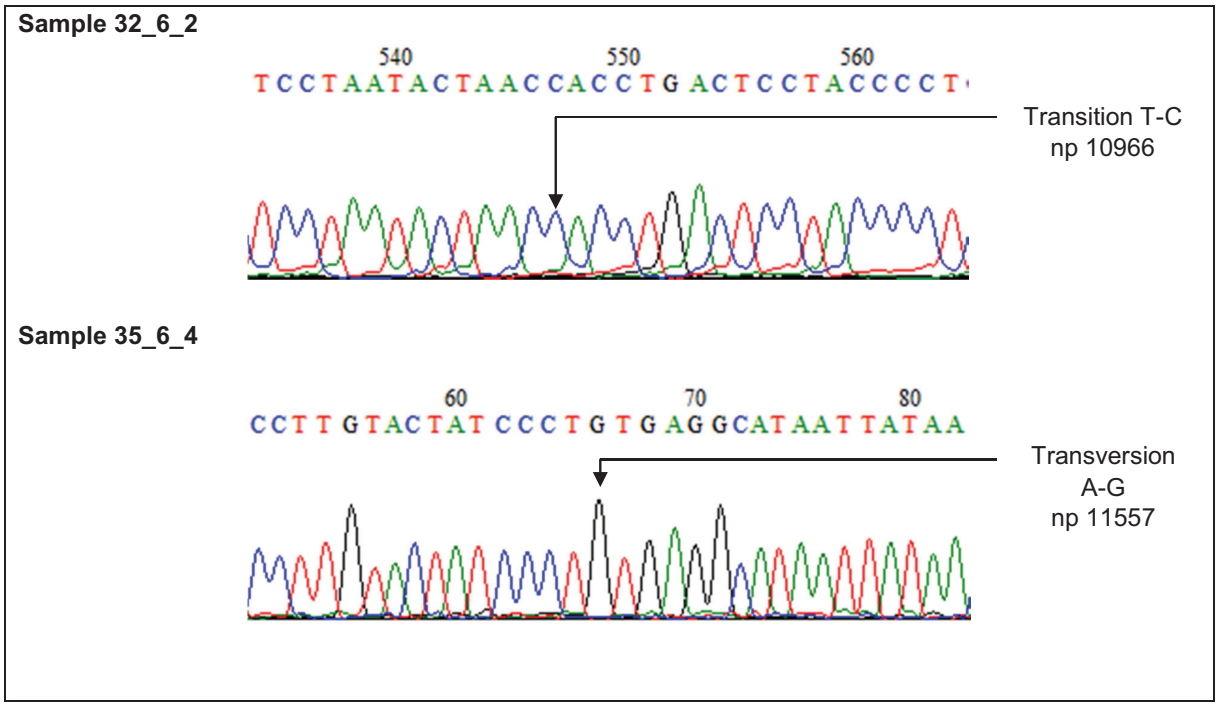


Sample name 29_6_2 refers to sample TS_4037_29 primer region 6, sequence primer 2; sample name 41_6_4 refers to sample TS_5044_41 primer region 6, sequence primer 4; sequence region presented contained transition before alignment with rCRS, therefore numbering presented above is not in accordance with rCRS alteration position as determined by comparison to the rCRS.

Two sequence alterations in the *ND4* gene region of the Tswana-speaking individuals of this investigation have not previously been reported in individuals that belonged to

macrohaplogroup L. Therefore, these sequence alterations were considered to be novel within the haplogroup L lineages. A transition at np 10966 has been reported in Russian individuals belonging to haplogroup M7a2 (Derenko *et al.*, 2007) and Europeans belonging to haplogroup J (Mishmar *et al.*, 2003) and has been cited by the PhyloTree classification system (Van Oven and Kayser, 2009) as a defining mutation for haplogroup J21a3. The second transition, which has been observed at np 11557, has been reported in Russian individuals belonging to haplogroup A4 (Derenko *et al.*, 2007) and has not been cited as a haplogroup-defining mutation by the PhyloTree classification system (Van Oven and Kayser, 2009). Each of these transitions is present in a single Tswana-speaking individual of this investigation i.e. TS_4056 and TS_4080, which suggests that they are both recent private mutations and not suitable for consideration as haplogroup-defining sites. The electropherograms of the samples that display these transitions, presented in Figure 6.57, do not display any sequencing artefacts or background noise and therefore the possibility that these peaks have been called incorrectly has been dispelled.

Figure 6.57 Representative electropherograms of the sequence data generated indicating a transition at np 10966 and a transversion at np 11557

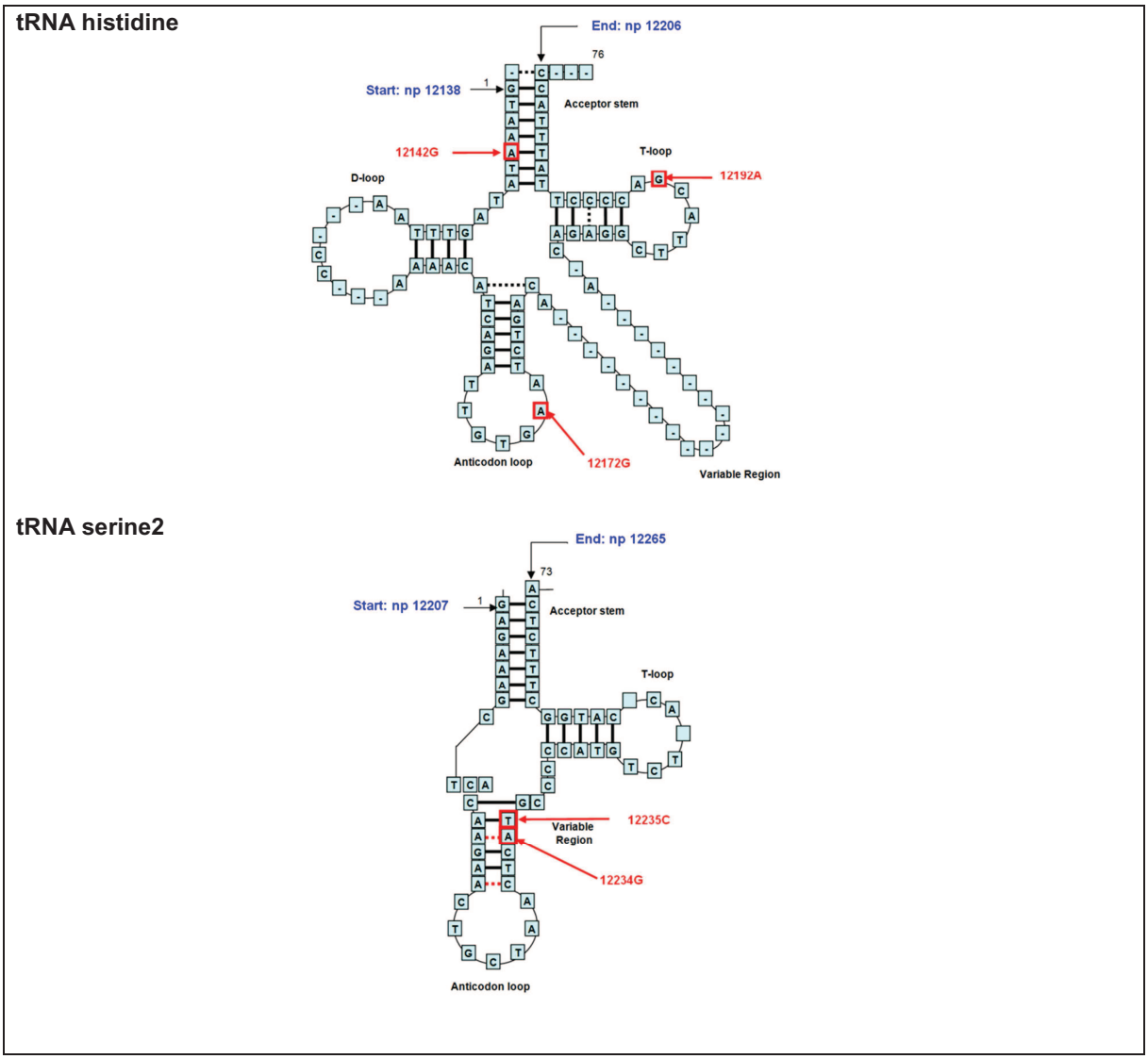


Sample name 32_6_2 refers to sample TS_4056_32 primer region 6, sequence primer 2; sample name 35_6_4 refers to sample TS_4080_35, primer region 6, sequence primer 4; sequence regions presented containing transitions before alignment with rCRS, therefore numbering presented above is not in accordance with rCRS alteration position as determined by comparison to the rCRS.

Five sequence alterations are observed in the coding regions of the tRNA histidine (H) and tRNA serine2 (S(AGY)) of primer region 7. No sequence alterations are observed in the tRNA leucine2 (L(CUN)) coding region. Three of the transitions, i.e. at np 12142, 12234

and 12235, are localised in the stem regions of the tRNA secondary structures, possibly affecting the functioning of the tRNAs. Two of these transitions are localised in the stem region of the tRNA serine2 (S(AGY)) molecule, which has been reported to display higher numbers of mutations in this stem region (Pereira *et al.*, 2009). The other two mutations, at np 12172 and np 12192, occur in the more unconserved loop regions of the tRNAs. The location of the sequence alterations observed in the Tswana-speaking individuals of this investigation within the tRNA coding regions of primer region 7 are presented in Figure 6.58.

Figure 6.58 Structure of the tRNA histidine and tRNA serine2



Structure of a typical tRNA histidine, tRNA serine2, ▬ 100% Watson Crick pairs; ▪ ▪ 100% mismatches; numbers alongside the tRNA structure indicate nucleotide positions within the tRNA molecule and not according to the rCRS; nucleotide starting and ending positions of the tRNA according to the rCRS indicated in blue ink; nucleotide position indicated in red ink refers to the nucleotide variation observed in the Tswana-speaking individuals of this investigation. From Jühling *et al.*, 2009; Pereira *et al*. 2009.

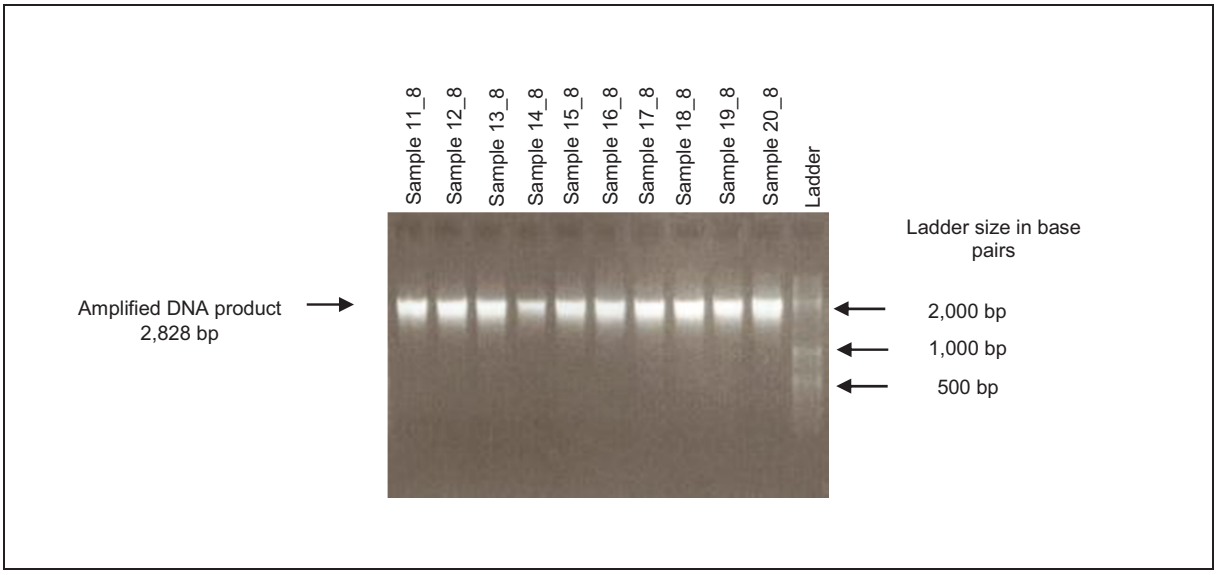
MITOMAP (a Human Mitochondrial Genome Database. http://www.mitomap.org, 2011) was searched for reported mitochondrial DNA sequence alterations that were associated with diseases. One transition, at np 12192, located in the T-loop of the tRNA histidine (H) molecule, was associated with maternally inherited cardiomyopathy (MICM) (Shin *et al.*, 2000). The G-A transition occurred two base pairs from the 3'end of the T-loop of tRNA histidine (H), as presented in Figure 6.58. It is believed that this sequence alteration causes the addition of an A:T base pair and shortens the loop structure to affect the function of the tRNA histidine (H). The study by Shin *et al*. (2000) was performed on Japanese patients only and it was discovered that the transition at np 12192 was connected to MICM throughout the ancestry of these individuals. This makes it highly unlikely that this sequence variant would be associated with the same disease in an African population that has a weak ancestral connection with the Asian populations. Further investigation would be necessary to determine the pathogenicity of this transition in the Tswana population of South Africa.

6.6.8 Primer region 8

Primer region 8 starts at np 13613 and ends at np 15996. It contains the *ND5* gene that starts at np 12337 in primer region 7 and ends at np 14148 in primer region 8. Primer region 8 further contains the *ND6* gene region located between np 14149 and np 14673, the *Cytb* gene region between np 14747 and np 15887 and the coding regions of the tRNA glutamic acid (E) at np 14674-np 14742, tRNA threonine (T) at np 15888-np 15953 and tRNA proline (P) at np 15956-np 16023, which extends into primer region 1. Primer region 8 further contains non-coding regions between np 14742, np 14747, np 15953 and np 15956, as well as an ATT membrane attachment site which stretches from np 15925 to np 499 in primer region 1.

The region outlined above was amplified by PCR, as discussed in Section 5.4, and the PCR products were electrophoresed on an agarose gel to ascertain the quality of the product. A representative example of the mtDNA products for primer region 8, as visualised by the UVIvue ultraviolet transilluminator, are presented in Figure 6.59.

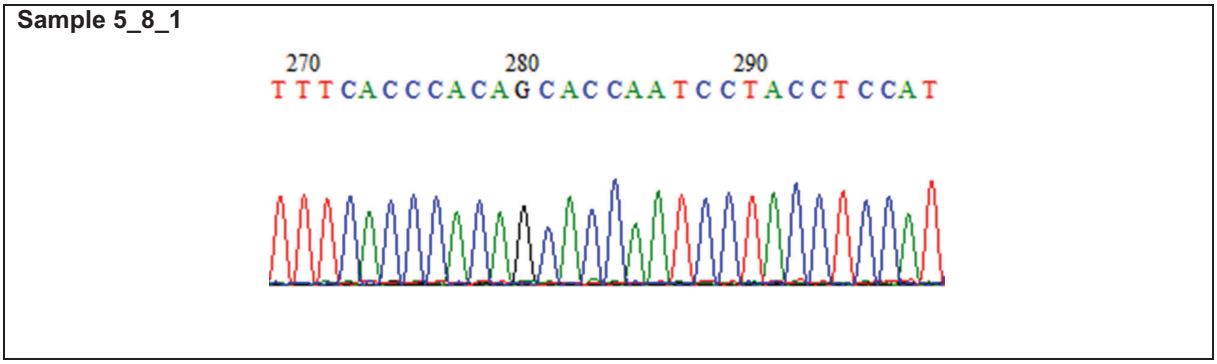
Figure 6.59 Photographic representation of the amplified mtDNA product of primer region 8



Photograph of the agarose gel on which the mtDNA amplified product was electrophoresed at 100 volts (V) and 50 mAmperes (mA) for 30 minutes as discussed in Section 5.5; ladder = FastRuler™ High Range DNA Ladder (Fermentas) of range 100 – 10,000 bp; ladder included in the last lane of the gel; sample names refer to the Tswana-speaking individuals of this investigation.

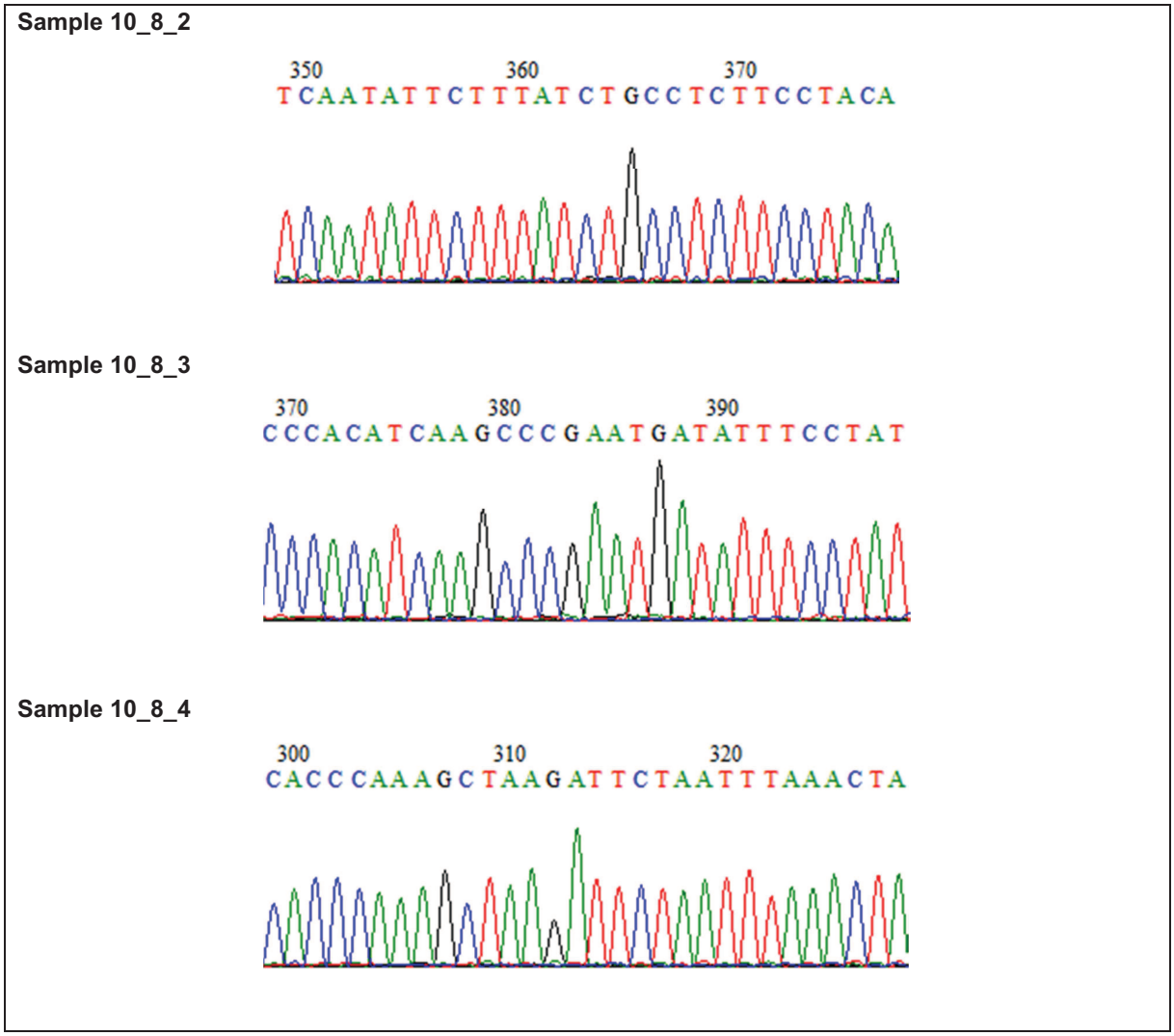
Representative examples of the electropherograms of the sequence data generated for primer region 8 by the BigDye®[1] Terminator v3.1 Cycle Sequencing Kit are represented in Figure 6.60. These results were viewed and edited in BioEdit software version 7.0.5.2 (Hall, 2001).

Figure 6.60 Representative electropherograms of the sequence generated for primer region 8 using the forward primers



1 BigDye® Terminator v3.1 Cycle Sequencing Kit is a registered trademark of Applied Biosystems, Foster City, CA, USA.

Figure 6.60 Continued…



Examples of electropherogram data with peaks depicting nucleotides in the sequence region of primer 8; A = adenine; T = thymine; C = cytosine; G = guanine; numbering at the top of the electropherogram represents the numbering of the nucleotides according to the sequenced fragment before alignment with the rCRS and therefore does not correspond to the nucleotide positions of primer region 8.

6.6.8.1 Sequence alterations observed in primer region 8

Primer region 8 contains a 536 base pair segment of the *ND5* gene, which starts in primer region 7 at np 12337. In order to discuss the observed sequence alterations of the *ND5* gene as a single unit, the 1,275 base pair segment that is located in primer region 7 has been included in the discussion of sequence alterations in this section. The tRNA proline (P) coding region starts at np 15956 and ends at np 16023 in primer region 1. The 27 base pair segment that is located in primer region 1 is included in the discussion of sequence alterations in this section in order to discuss the observed sequence alterations of the tRNA P as a single unit. Therefore, the observed sequence alterations from np 12337 to np 16023 are discussed in this section and are presented in Table 6.23.

Table 6.23 Sequence alterations observed between the complete mitochondrial DNA of the Tswana individuals included in this study and the rCRS in primer region 8

Position	Sequence alteration	Gene/ region	Frequency	Reference
12348	C-T	*ND5*	1	Behar *et al.*, 2008
12414	T-C	*ND5*	1	Behar *et al.*, 2008
12436	**C-T**	***ND5***	**5**	**Current investigation**
12528	G-A	*ND5*	1	Mishmar *et al.*, 2003
12612	A-G	*ND5*	1	Behar *et al.*, 2008
12693	A-G	*ND5*	10	Mishmar *et al.*, 2003
12696	T-C	*ND5*	3	Behar *et al.*, 2008
12705	C-T	*ND5*	49	Mishmar *et al.*, 2003
12720	A-G	*ND5*	35	Mishmar *et al.*, 2003
12798	C-T	*ND5*	4	Mishmar *et al.*, 2003
12810	A-G	*ND5*	11	Mishmar *et al.*, 2003
12952	G-A	*ND5*	1	Behar *et al.*, 2008
12978	A-G	*ND5*	3	Behar *et al.*, 2008
13020	T-C	*ND5*	1	Gonder *et al.*, 2007
13077	**C-A**	***ND5***	**1**	**Current investigation**
13105	A-G	*ND5*	39	Mishmar *et al.*, 2003
13116	C-T	*ND5*	4	Behar *et al.*, 2008
13129	C-T	*ND5*	2	Kivisild *et al.*, 2006
13149	A-G	*ND5*	2	Behar *et al.*, 2008
13276	A-G	*ND5*	35	Ingman *et al.*, 2000
13359	G-A	*ND5*	5	Behar *et al.*, 2008
13431	C-T	*ND5*	1	Olivieri *et al.*, 2006
13473	**A-G**	***ND5***	**2**	**Current investigation**
13488	T-C	*ND5*	1	Just *et al.*, 2008
13506	C-T	*ND5*	38	Mishmar *et al.*, 2003
13590	G-A	*ND5*	11	Mishmar *et al.*, 2003
13604	**G-C**	***ND5***	**1**	**Current investgation**
13650	C-T	*ND5*	48	Mishmar *et al.*, 2003
13708	G-A	*ND5*	2	Kivisild *et al.*, 2006
13759	G-A	*ND5*	12	Kivisild *et al.*, 2006
13767	**C-A**	***ND5***	**1**	**Current investgation**
13789	T-C	*ND5*	2	Mishmar *et al.*, 2003
13803	A-G	*ND5*	10	Mishmar *et al.*, 2003
13818	T-C	*ND5*	1	Ingman *et al.*, 2000
13819	T-C	*ND5*	3	Ingman *et al.*, 2000
13886	T-C	*ND5*	1	Mishmar *et al.*, 2003
13928	G-C	*ND5*	1	Ingman *et al.*, 2000
13934	C-T	*ND5*	1	Behar *et al.*, 2008
14000	T-A	*ND5*	2	Mishmar *et al.*, 2003
14007	A-G	*ND5*	3	Kivisild *et al.*, 2006
14020	T-C	*ND5*	2	Ingman *et al.*, 2000

Table 6.23 Continued…

Position	Sequence alteration	Gene/ Region	Frequency	Reference
14040	G-A	*ND5*	1	Kivisild *et al.*, 2006
14094	T-C	*ND5*	1	Kivisild *et al.*, 2006
14110	T-C	*ND5*	1	Kivisild *et al.*, 2006
14127	A-G	*ND5*	2	Olivieri *et al.*, 2006
14152	A-G	*ND6*	1	Ingman *et al.*, 2000
14163	**C-T**	***ND6***	**1**	**Current investigation**
14178	T-C	*ND6*	2	Mishmar *et al.*, 2003
14182	T-C	*ND6*	1	Ingman *et al.*, 2000
14207	G-A	*ND6*	2	Kivisild *et al.*, 2006
14212	T-C	*ND6*	1	Kivisild *et al.*, 2006
14221	T-C	*ND6*	9	Mishmar *et al.*, 2003
14284	C-T	*ND6*	1	Mishmar *et al.*, 2003
14290	T-C	*ND6*	1	Derenko *et al.,* 2007
14308	T-C	*ND6*	7	Mishmar *et al.*, 2003
14315	C-T	*ND6*	6	Kivisild *et al.*, 2006
14371	T-C	*ND6*	1	Ingman *et al.*, 2000
14374	T-C	*ND6*	1	Ingman *et al.*, 2000
14425	**C-T**	***ND6***	**1**	**Current investigation**
14482	C-T	*ND6*	1	Gonder *et al.*, 2007
14560	G-A	*ND6*	2	Kivisild *et al.*, 2006
14566	A-G	*ND6*	10	Kivisild *et al.*, 2006
14577	T-C	*ND6*	1	Behar *et al.*, 2008
14659	C-T	*ND6*	6	Kivisild *et al.*, 2006
14687	A-G	E (T-loop)	2	Kivisild *et al.*, 2006
14755	A-G	*CYTB*	4	Behar *et al.*, 2008
14766	C-T	*CYTB*	50	Mishmar *et al.*, 2003
14861	G-A	*CYTB*	2	Gonder *et al.*, 2007
14862	C-T	*CYTB*	1	Gonder *et al.*, 2007
14911	C-T	*CYTB*	2	Torroni *et al.*, 2006
14926	A-G	*CYTB*	1	Behar *et al.*, 2008
15016	C-T	*CYTB*	2	Behar *et al.*, 2008
15136	C-T	*CYTB*	8	Mishmar *et al.*, 2003
15140	G-A	*CYTB*	1	Gasparre *et al.*, 2007
15236	A-G	*CYTB*	5	Maca-Meyer *et al.* 2001
15244	A-G	*CYTB*	2	Olivieri *et al.*, 2006
15263	C-T	*CYTB*	1	Behar *et al.*, 2008
15301	G-A	*CYTB*	12	Mishmar *et al.*, 2003
15312	T-C	*CYTB*	5	Behar *et al.*, 2008
15315	C-T	*CYTB*	2	Tamm *et al.*, 2007
15326	A-G	*CYTB*	50	Mishmar *et al.*, 2003
15337	C-T	*CYTB*	5	Finnila *et al.*, 2001
15364	**C-T**	***CYTB***	**1**	**Current investigation**
15431	G-A	*CYTB*	7	Mishmar *et al.*, 2003

Table 6.23 Continued…

Position	Sequence alteration	Gene/ Region	Frequency	Reference
15466	G-A	*CYTB*	28	Mishmar *et al.*, 2003
15629	T-C	*CYTB*	2	Ingman *et al.*, 2000
15670	T-C	*CYTB*	1	Ingman *et al.*, 2000
15692	A-G	*CYTB*	1	Kivisild *et al.*, 2006
15697	T-C	*CYTB*	1	Kivisild *et al.*, 2006
15735	C-T	*CYTB*	3	Kivisild *et al.*, 2006
15758	A-G	*CYTB*	1	Kivisild *et al.*, 2006
15766	A-G	*CYTB*	10	Mishmar *et al.*, 2003
15784	T-C	*CYTB*	12	Mishmar *et al.*, 2003
15808	A-G	*CYTB*	1	Herrnstadt *et al.*, 2002
15812	G-A	*CYTB*	1	Kivisild *et al.*, 2006
15884	G-A	*CYTB*	1	Kivisild *et al.*, 2006
15924	A-G	T (Ac-stem)	1	Mishmar *et al.*, 2003
15930	G-A	T (Variable region)	27	Mishmar *et al.*, 2003
15941	T-C	T (T-loop)	27	Mishmar *et al.*, 2003
15942	T-C	T (T-stem)	1	Ingman *et al.*, 2000
15951	A-G	T (Acc-stem)	2	Kivisild *et al.*, 2006

Transitions are indicated in blue and transversions are indicated in red. The frequency is indicated as the number of times the sequence variation was observed within the total Tswana dataset of 50 individuals; *ND5* = NADH dehydrogenase subunit 5 gene ; *ND6* = NADH dehydrogenase subunit 6 gene; CYTB = cytochrome b; Acc-stem = tRNA acceptor stem; Ac-stem = tRNA anticodon stem.

The 1,811 base pair segment of the *ND5* gene displays 45 sequence alterations that consist of 40 transitions and five transversions. The *ND6* gene region, which consists of 524 base pairs, displays 19 transitions and the *Cytb* gene region that is 1,140 base pairs long, displays 31 transitions. No transversions have been observed in the *ND6* and *Cytb* gene regions. The tRNA glutamic acid (E) coding region displays a single transition in the T-loop and the tRNA threonine (T) coding region displays five transitions, of which three are located in stem regions, one in the T-loop and one in the variable region. In total, primer region 8 displays 101 sequence variants, of which five are transversions and the rest transitions. The transversion:transition count ratio is therefore 1:20 (Lutz Bonengel *et al.*, 2003). No indels have been observed in primer region 8.

The *ND5, ND6* and *CytB* gene regions were further investigated by the mtDNA-GeneSyn tool developed by Pereira *et al*. (2009). The impact of the sequence alterations on the *ND5, ND6* and *CytB* gene coding functionality was determined and is presented in Table 6.24.

Table 6.24 Sequence variation within the *ND5*, *ND6* and *Cytb* genes

Position	Sequence alteration	Gene	Syn/ Non-syn	Codon	New codon	Codon position	Amino acid	New amino acid
12348	C-T	*ND5*	Syn	CAC	CAU	3	H	H
12414	T-C	*ND5*	Syn	CCU	CCC	3	P	P
12436	C-T	*ND5*	Non-syn	CAU	UAU	1	H	Y
12528	G-A	*ND5*	Syn	UCG	UCA	3	S	S
12612	A-G	*ND5*	Syn	GUA	GUG	3	V	V
12693	A-G	*ND5*	Syn	AAA	AAG	3	K	K
12696	T-C	*ND5*	Syn	UAU	UAC	3	Y	Y
12705	C-T	*ND5*	Syn	AUC	AUU	3	I	I
12720	A-G	*ND5*	Syn	AUA	AUG	3	M	M
12798	C-T	*ND5*	Syn	CUC	CUU	3	L	L
12810	A-G	*ND5*	Syn	UGA	UGG	3	W	W
12952	G-A	*ND5*	Non-syn	GCU	ACU	1	A	T
12978	A-G	*ND5*	Syn	CUA	CUG	3	L	L
13020	T-C	*ND5*	Syn	GGU	GGC	3	G	G
13077A	C-A	*ND5*	Syn	CUC	CUA	3	L	L
13105	A-G	*ND5*	Non-syn	AUC	GUC	1	I	V
13116	C-T	*ND5*	Syn	CUC	CUU	3	L	L
13129	C-T	*ND5*	Non-syn	CCC	UCC	1	P	S
13149	A-G	*ND5*	Syn	CCA	CCG	3	P	P
13276	A-G	*ND5*	Non-syn	AUA	GUA	1	M	V
13359	G-A	*ND5*	Syn	AUG	AUA	3	M	M
13431	C-T	*ND5*	Syn	ACC	ACU	3	T	T
13473	A-G	*ND5*	Syn	GCA	GCG	3	A	A
13488	T-C	*ND5*	Syn	CCU	CCC	3	P	P
13506	C-T	*ND5*	Syn	UAC	UAU	3	Y	Y
13590	G-A	*ND5*	Syn	CUG	CUA	3	L	L
13604	G-C	*ND5*	Non-syn	AGC	ACC	2	S	T
13650	C-T	*ND5*	Syn	CCC	CCU	3	P	P
13708	G-A	*ND5*	Non-syn	GCA	ACA	1	A	T
13759	G-A	*ND5*	Non-syn	GCA	ACA	1	A	T
13767	C-A	*ND5*	Syn	CCC	CCA	3	P	P
13789	T-C	*ND5*	Non-syn	UAC	CAC	1	Y	H
13803	A-G	*ND5*	Syn	ACA	ACG	3	T	T
13818	T-C	*ND5*	Syn	ACU	ACC	3	T	T
13819	T-C	*ND5*	Non-syn	UUC	CUC	1	F	L
13886	T-C	*ND5*	Non-syn	CUA	CCA	2	L	P
13928	G-C	*ND5*	Non-syn	AGC	ACC	2	S	T
13934	C-T	*ND5*	Non-syn	ACA	AUA	2	T	M
14000	T-A	*ND5*	Non-syn	CUA	CAA	2	L	Q
14007	A-G	*ND5*	Syn	UGA	UGG	3	W	W
14020	T-C	*ND5*	Syn	UUA	CUA	1	L	L
14040	G-A	*ND5*	Syn	CAG	CAA	3	Q	Q

Table 6.24 Continued…

Position	Sequence alteration	Gene	Syn/ Non-syn	Codon	New codon	Codon position	Amino acid	New amino acid
14094	T-C	*ND5*	Syn	CUU	CUC	3	L	L
14110	T-C	*ND5*	Non-syn	UUC	CUC	1	F	L
14127	A-G	*ND5*	Syn	CUA	CUG	3	L	L
14152	A-G	*ND6*	Syn	AAU	AAC	3	N	N
14163	C-T	*ND6*	Non-syn	GCU	ACU	1	A	T
14178	T-C	*ND6*	Non-syn	AUU	GUU	1	I	V
14182	T-C	*ND6*	Syn	GUA	GUG	3	V	V
14207	G-A	*ND6*	Non-syn	ACU	AUU	2	T	I
14212	T-C	*ND6*	Syn	GUA	GUG	3	V	V
14221	T-C	*ND6*	Syn	UGA	UGG	3	W	W
14284	C-T	*ND6*	Syn	GAG	GAA	3	E	E
14290	T-C	*ND6*	Syn	GAA	GAG	3	E	E
14308	T-C	*ND6*	Syn	GGA	GGG	3	G	G
14315	C-T	*ND6*	Non-syn	AGU	AAU	2	S	N
14371	T-C	*ND6*	Syn	GGA	GGG	3	G	G
14374	T-C	*ND6*	Syn	GUA	GUG	3	V	V
14425	C-T	*ND6*	Syn	GGG	GGA	3	G	G
14482	C-T	*ND6*	Syn	AUG	AUA	3	M	M
14560	G-A	*ND6*	Syn	GGU	GGC	3	V	V
14566	A-G	*ND6*	Syn	GGU	GGC	3	G	G
14577	T-C	*ND6*	Non-syn	AUU	GUU	1	I	V
14659	C-T	*ND6*	Syn	UUG	UUA	3	L	L
14755	A-G	*CYTB*	Syn	CCA	CCG	3	P	P
14766	C-T	*CYTB*	Non-syn	ACU	AUU	2	T	I
14861	G-A	*CYTB*	Non-syn	GCC	ACC	1	A	T
14862	C-T	*CYTB*	Non-syn	GCC	GUC	2	A	V
14911	C-T	*CYTB*	Syn	UAC	UAU	3	Y	Y
14926	A-G	*CYTB*	Syn	UCA	UCG	3	S	S
15016	C-T	*CYTB*	Syn	UUC	UUU	3	F	F
15136	C-T	*CYTB*	Syn	GGC	GGU	3	G	G
15140	G-A	*CYTB*	Non-syn	GUC	AUC	1	V	I
15236	A-G	*CYTB*	Non-syn	AUC	GUC	1	I	V
15244	A-G	*CYTB*	Syn	GGA	GGG	3	G	G
15263	C-T	*CYTB*	Non-syn	CCC	UCC	1	P	S
15301	G-A	*CYTB*	Syn	UUG	UUA	3	L	L
15312	T-C	*CYTB*	Non-Ssyn	AUU	ACU	2	I	T
15315	C-T	*CYTB*	Non-syn	GCA	GUA	2	A	V
15326	A-G	*CYTB*	Non-syn	ACA	GCA	1	T	A
15337	C-T	*CYTB*	Syn	CUC	CUU	3	L	L
15364	C-T	*CYTB*	Syn	AAC	AAU	3	N	N
15431	G-A	*CYTB*	Non-syn	GCC	ACC	1	A	T
15466	G-A	*CYTB*	Syn	AUG	AUA	3	M	M

Table 6.24 Continued…

Position	Sequence alteration	Gene	Syn/ Non-syn	Codon	New codon	Codon position	Amino acid	New amino acid
15629	T-C	*CYTB*	Syn	UUA	CUA	1	L	L
15670	T-C	*CYTB*	Syn	CAU	CAC	3	H	H
15692	A-G	*CYTB*	Non-syn	AUA	GUA	1	M	V
15697	T-C	*CYTB*	Syn	UUU	UUC	3	F	F
15735	C-T	*CYTB*	Non-syn	GCA	GUA	2	A	V
15758	A-G	*CYTB*	Non-syn	AUC	GUC	1	I	V
15766	A-G	*CYTB*	Syn	GGA	GGG	3	G	G
15784	T-C	*CYTB*	Syn	CCU	CCC	3	P	P
15808	A-G	*CYTB*	Syn	GCA	GCG	3	A	A
15812	G-A	*CYTB*	Non-syn	GUA	AUA	1	V	M
15884	G-A	*CYTB*	Non-syn	GCC	ACC	1	A	T

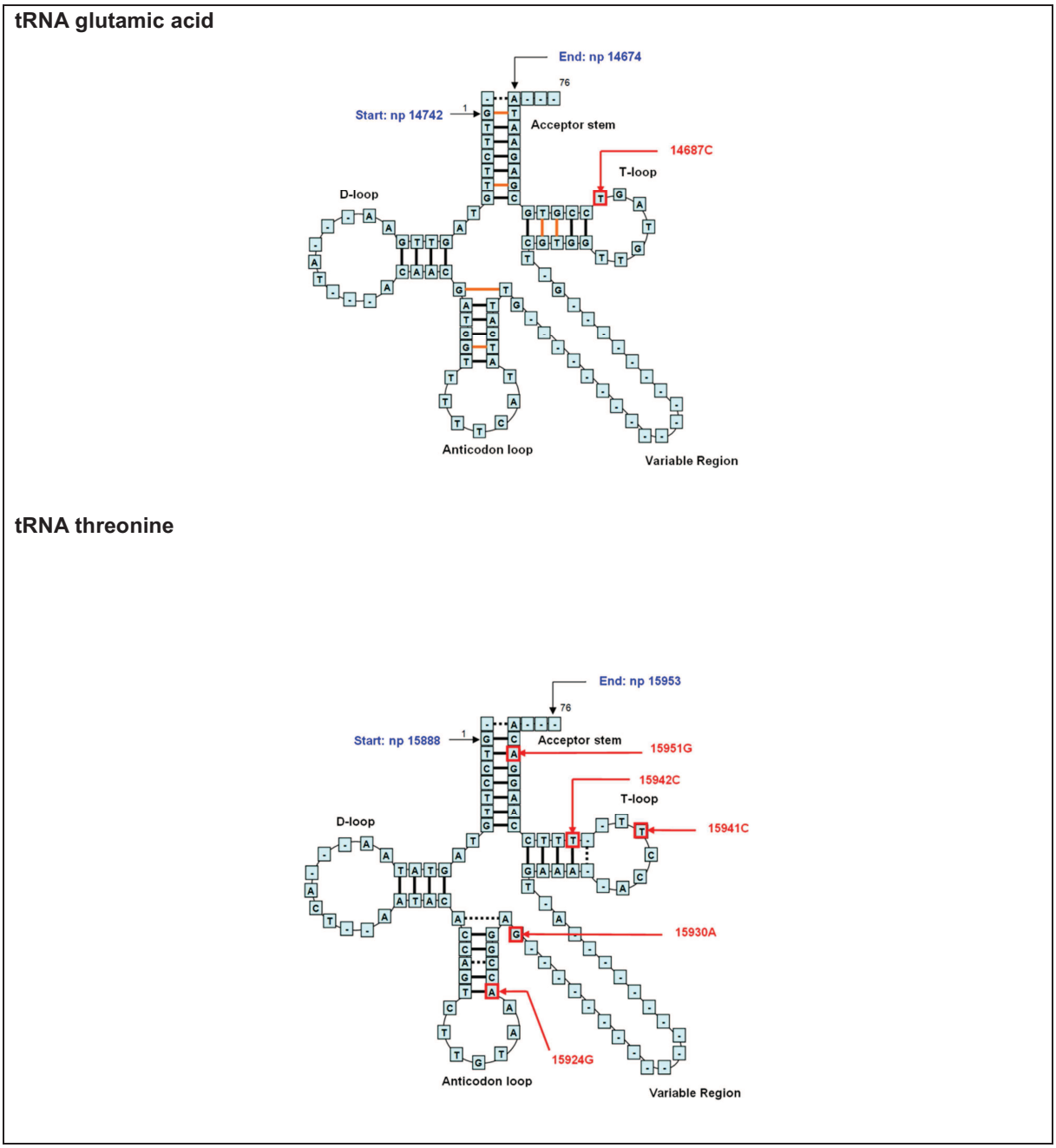
ND5 = NADH dehydrogenase subunit 5; *ND6* = NADH dehydrogenase subunit 6; *CYTB* = Cytochrome b; the sequence alterations described in terms of the synonymous or nonsynonymous nature of the changes, the codon position that was affected and the new amino acid that was coded for; A = adenine; T = thymine; C = cytosine; G = guanine; syn = synonymous; nonsyn = nonsynonymous A= Alanine; E = Glutamic acid (Glutamate); F = Phenylalanine; G = Glycine; H = Histidine; I = Isoleucine; L = Leucine; K = Lysine; M = Methionine; N = Asparagine; P = Proline; Q = Glutamine; S = Serine; T = Threonine; V = Valine; W = Tryptophan;.Y = Tyrosine; nonsynonymous positions indicated in grey highlight.

Thirty of the sequence variants observed in the *ND5* gene region constitute synonymous changes, of which 29 are located in third codon positions and one in a first codon position. Fifteen of the sequence variants in the *ND5* gene region constitute nonsynonymous changes, of which ten are located in first codon positions and five in second codon positions. Five of the 19 transitions in the *ND6* gene and 15 of the 31 transitions in the *Cytb* gene region are nonsynonymous changes, of which three and ten occur at first codon positions and two and five in second codon positions.

One transition is observed in the coding region of the tRNA glutamic acid (E), which is located in the T-loop of the tRNA secondary structure. Another five transitions are observed in the coding region of the tRNA threonine (T), one of each located in the anti-codon stem, variable region, T-loop, T-stem and acceptor-stem of the secondary structure of the tRNA. The large number of sequence variants observed in the tRNA threonine region has been expected, since this region has consistently displayed the highest number of mutations when compared to the other tRNA regions of the mtDNA (Moilanen and Malamaa, 2003; Vilmi *et al.*, 2005). Most of the sequence variants that have been reported for the tRNA threonine are located in the stem regions, as opposed to the loop regions, as is the case in the sequence variants observed in the tRNA threonine in this investigation (Pereira *et al.*, 2009). The positions at which the transitions have been observed in the secondary structures of the tRNA glutamic acid (E) and the tRNA

threonine (T) are presented in Figure 6.61. No sequence variants have been observed in the coding region of the tRNA proline (P).

Figure 6.61 Structure of the tRNA glutamic acid and tRNA threonine



Structure of a typical tRNA glutamic acid and tRNA threonine, ▬ 100% Watson Crick pairs; ▪▪ 100% mismatches; ▬ 100% GT pairs; numbers alongside the tRNA structure indicate nucleotide positions within the tRNA molecule and not according to the rCRS; nucleotide starting and ending positions of the tRNA according to the rCRS indicated in blue ink; nucleotide position indicated in red ink refers to the nucleotide variation observed in the Tswana-speaking individuals of this investigation. From Jühling *et al.*, 2009; Pereira *et al.,* 2009.

Eight novel mutations are observed in the Tswana-speaking individuals of this investigation. Five of the novel mutations are located in the *ND5* gene region at np 12436, np 13077, np 13473, np 13604 and np 13767. Two of these *ND5* gene mutations, at

np 12436 and np 13473, are transitions and the other three, at np 13077, np 13604 and np 13767, are transversions. The novel transversions each occurs in a single Tswana-speaking individual of this investigation, as opposed to the novel transitions at np 12436 and np 13473, which occur in five and two Tswana-speaking individuals respectively. Two of the novel sequence variants within the *ND5* gene region of the Tswana-speaking individuals of this investigation at np 12436 and np 13604 were are nonsynonymous. The electropherograms of the samples in which the novel *ND5* gene mutations have been observed, namely TS_2097, TS_5060, TS_4037, TS_4080 and TS_5063, are presented in Figure 6.62. Those novel transitions that occur in more than one Tswana-speaking individual are represented by a single representative sample. The electropherograms display no sequencing artefacts and no background noise, ruling out the possibility that these peaks have been called incorrectly because of sequencing artefacts or poor peak morphology. The presence of the novel transversion at np 13767 in sample TS_5063 is of some concern, as was discussed in Section 6.6.4 and Section 6.6.5. This Tswana-speaking individual also displays novel transitions in primer regions 4 and 5 and several other novel haplogroup L mutations, which are unlikely all to occur simultaneously in one individual. It is suggested that the sample be further investigated to verify that these novel mutations are true and not due to human or laboratory error.

Figure 6.62 Representative electropherograms of novel sequence alterations observed in the *ND5* gene at np 12436, 13077, 13473, 13604 and 13767

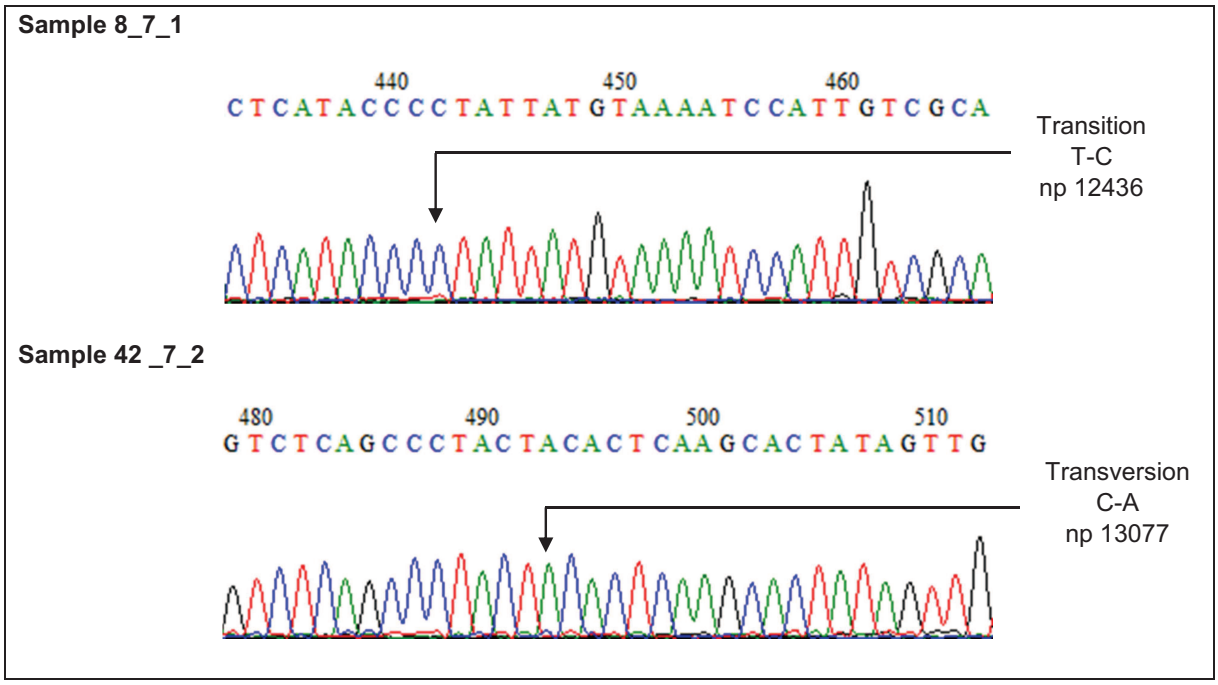
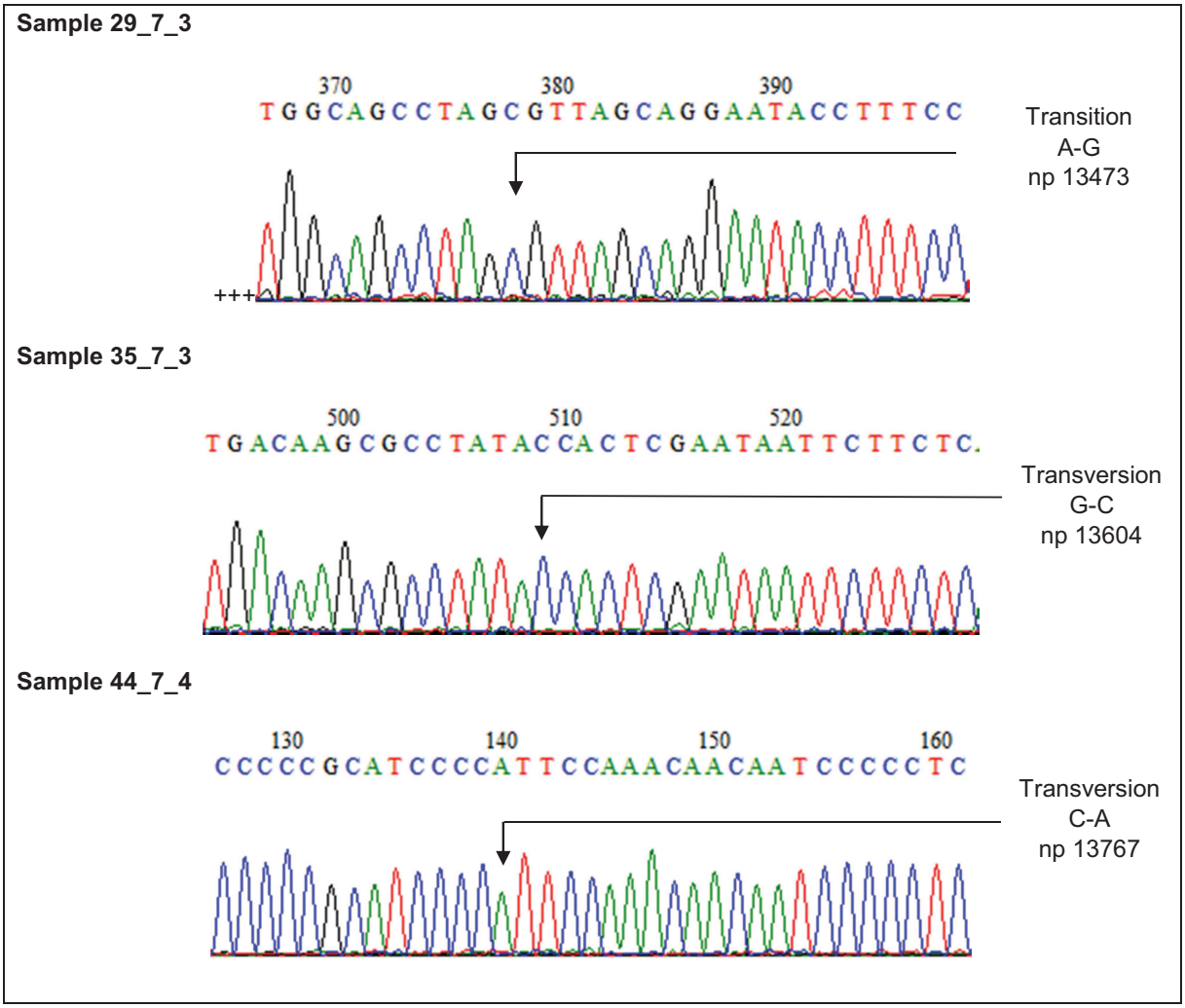


Figure 6.62 Continued…



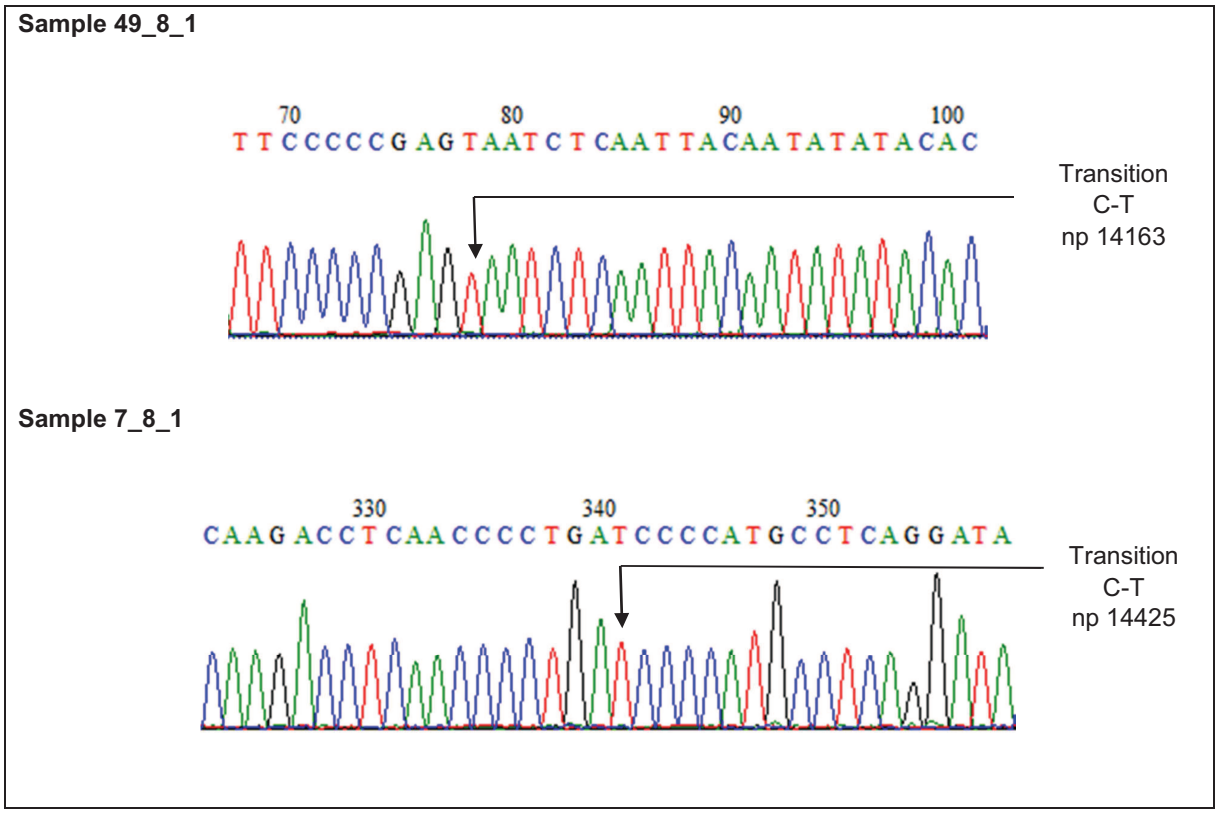
Sample name 8_7_1 refers to sample TS_2097_8 primer region 7, sequence primer 1; sample name 42_7_2 refers to sample TS_5060_42 primer region 7, sequence primer 2; sample name 29_7_3 refers to sample TS_4037_29 primer region 7, sequence primer 3; sample name 35_7_3 refers to sample TS_4080_35 primer region 7, sequence primer 3 sample name 44_7_4 refers to sample TS_5063_44 primer region 7, sequence primer 4; the sequenced region presented contains the transition before alignment with rCRS, therefore numbering presented above is not in accordance with rCRS alteration position as determined by comparison to the rCRS.

The novel transition at np 12436 occurs in five Tswana-speaking individuals of this investigation, namely TS_2097, TS_3027, TS_4027, TS_4083 and TS_5085, which all belong to haplogroup L0d1b and are grouped together in a sub-clade in all of the NJ and MP phylogenetic trees of this investigation (see Section 6.8). This transition is furthermore not shared with any of the other mtDNA sequences of the Global African or All African datasets of this investigation. The fact that the individuals that display this transition are grouped together in a single phylogenetic clade suggeststhat they share a maternal ancestor who contributed the novel transition to these five individuals. The validity of the transition has been verified by its presence in more than one of the Tswana-speaking individuals of this investigation and it is therefore a valid candidate for a new

sub-haplogroup of haplogroup L0d1b. See Section 6.7 for a further discussion of haplogroup classification.

Another two novel sequence alterations have been observed in the *ND6* gene region of the Tswana-speaking individuals of this investigation. The sequence alterations occur at np 14163 and np 14425 and are both transitions that occur in two Tswana-speaking individuals, i.e. TS_5086 and TS_2095 respectively. The transition at np 14163 causes a nonsynonymous amino acid change of alanine to threonine at the first codon position. The transition at np 14425 is located at a synonymous third codon position. The electropherograms of the mtDNA sequences in which these novel transitions occur are presented in Figure 6.63 and display no sequencing artefacts or background noise, thus verifying the validity of the peaks.

Figure 6.63 Representative electropherograms of novel sequence alterations observed in the *ND6* gene at np 14163 and np 14425

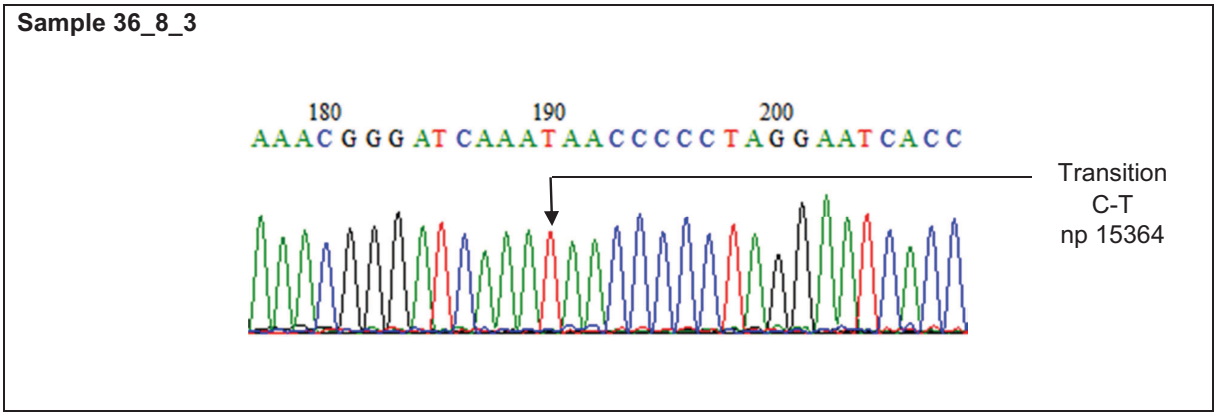


Sample name 49_8_1 refers to sample TS_5086_49 primer region 8, sequence primer 1; sample name 7_8_1 refers to sample TS_2095_7 primer region 8, sequence primer 1; sequence region presented contained transition before alignment with rCRS, therefore numbering presented above is not in accordance with rCRS alteration position as determined by comparison to the rCRS.

One novel synonymous transition is observed at np 15364 in the *Cytb* gene region of one of the Tswana-speaking individuals of this investigation, i.e. TS_4083. The electropherogram of the novel transition at np 15364 does not display any sign of

sequencing artefacts or background noise and has good peak morphology, which rules out the possibility that the peak has been called incorrectly. The electropherogram is presented in Figure 6.64.

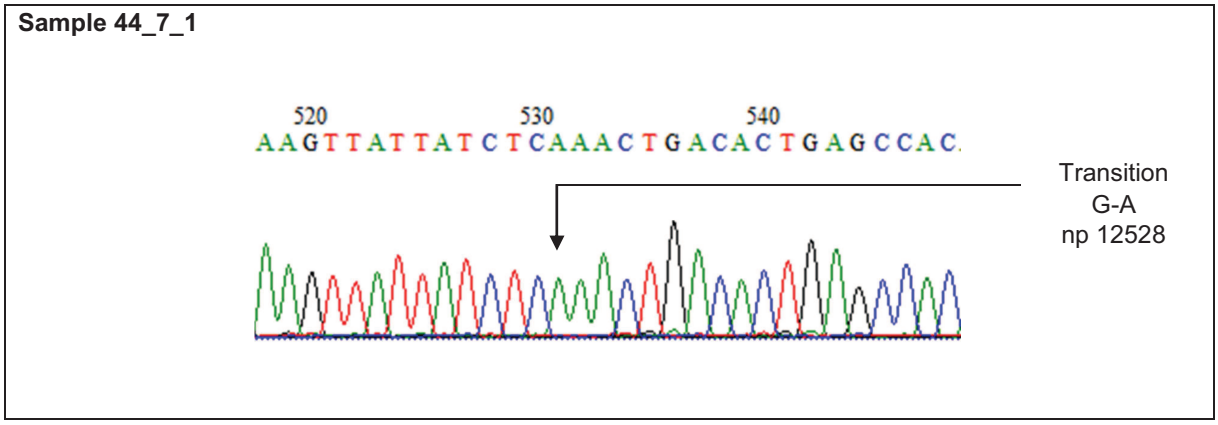
Figure 6.64 Representative electropherograms of a novel sequence alteration observed in the *Cytb* gene at np 15364



Sample name 36_8_3 refers to sample TS_4083_36 primer region 8, sequence primer 3; sequence region presented contained transition before alignment with rCRS, therefore numbering presented above is not in accordance with rCRS alteration position as determined by comparison to the rCRS.

In addition to the novel sequence variants observed in primer region 8, five sequence variants have been observed in the Tswana-speaking individuals of this investigation that have not been reported in individuals belonging to haplogroup L. A synonymous transition at np 12528 has been observed in the *ND5* gene region of one Tswana-speaking individual of this investigation, i.e. TS_5063, which was previously reported in a European individual belonging to haplogroup H (Mishmar *et al.*, 2003). The presence of the haplogroup L novel transition in sample TS_5063 is of some concern, as was discussed in depth in Section 6.6.4 and Section 6.6.5. The presence of a further novel sequence alteration in this individual reiterates the necessity of resampling and re-sequencing the mtDNA of the individual to investigate further and verify that these novel mutations are true and not due to human or laboratory error. The electropherogram of the mtDNA sequence that displays the novel haplogroup L transition is presented in Figure 6.65.

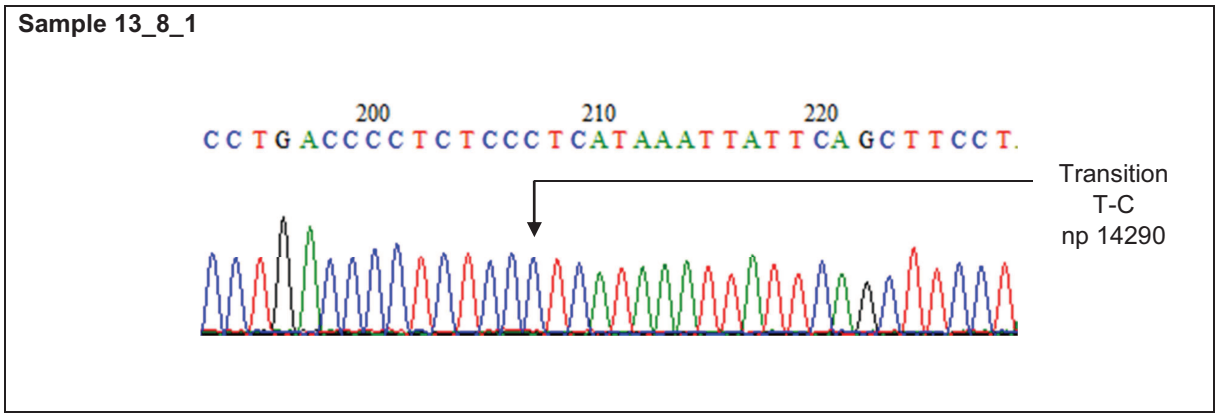
Figure 6.65 Representative electropherogram of a novel sequence alteration observed in the *ND5* gene at np 12528



Sample name 44_7_1 refers to sample TS_5063_44 primer region 7, sequence primer 1; sequence region presented contained transition before alignment with rCRS, therefore numbering presented above is not in accordance with rCRS alteration position.

A synonymous transition at np 14290 occurs in the *ND6* gene region of one Tswana-speaking individual of this investigation, i.e. TS_3066. This transition has been reported in Indian individuals belonging to haplogroup R31a1 (Chaubey *et al.*, 2008), in Russian individuals belonging to haplogroups A1 and A4b (Derenko *et al.*, 2007; Starikovskaya *et al.*, 2005), in Japanese individuals belongingto haplogroup N9b (Tanaka *et al.*, 2004) and in Taiwanese individuals belongingthat belonged to haplogroup B4a1a (Trejaut *et al.*, 2005). In addition, the PhyloTree classification system (Van Oven and Kayser, 2009) cites this sequence variant as a haplogroup-defining mutation for haplogroups A4b and R31a1. The electropherogram of the mtDNA sequence that contains the novel haplogroup L transition is presented in Figure 6.66.

Figure 6.66 Representative electropherogram of sequence alteration observed in the *ND6* gene at np 14290



Sample name 13_8_1 refers to sample TS_3066_13 primer region 8, sequence primer 1; sequence region presented contained transition before alignment with rCRS, therefore numbering presented above is not in accordance with rCRS alteration position.

Three sequence variants that have not previously been reported in haplogroup L individuals are observed in the *Cytb* gene region of the Tswana-speaking individuals of

this investigation. A transition at np 15140 causes a nonsynonymous amino acid change from valine to isoleucine and has been observed in the *Cytb* gene region of one Tswana-speaking individual of this investigation, i.e. TS_2077. It has been reported in Italian individuals belonging to haplogroups HV1 and U4b (Gasparre *et al.*, 2007) and in Columbians belonging to haplogroup C1c (Tamm *et al.*, 2007). A transition at np 15315 occurs in the *Cytb* gene region of two Tswana-speaking individuals of this investigation, i.e. TS_4037 and TS_4117. It occurs at a second codon position and causes a nonsynonymous amino acid change from alanine to valine and was has been reported in individuals from Sardinia belongingto haplogroup H3 (Fraumene *et al*., 2006) as well as in Columbians belonging to haplogroup A2 (Tamm *et al*., 2007). None of these sequence variants is cited by the PhyloTree classification system (Van Oven and Kayser, 2009) as haplogroup-defining. The third transition of the *Cytb* gene region that has not been reported in haplogroup L individuals occurs at np 15337 in five Tswana-speaking individuals of this investigation, namely TS_2097, TS_3027, TS_4027, TS_4083 and TS_5085. These were are the same individuals who display the novel transition at np 12436, as discussed earlier in this section, and are grouped together in the L0d1b sub-clade in all the NJ and MP phylogenetic trees of this investigation. As with the previous novel mutation shared by these individuals, the presence of the transition in more than two individuals could be regarded as verification of its authenticity and as strong evidence for a candidate haplogroup-defining mutation of the sub-haplogroup L0d1b1. See further discussion of haplogroup classification in Section 6.7. The transition has been reported in Finnish individuals belongingto haplogroup U5a (Finnila *et al.*, 2001) and is not cited by the PhyloTree classification system (Van Oven and Kayser, 2009) as a haplogroup-defining mutation. The novel haplogroup L sequence variants observed in the *Cytb* gene region of the Tswana-speaking individuals of this investigation are presented in Figure 6.67.

Figure 6.67 Representative electropherogram of sequence alterations observed in the *Cytb* gene at np 15140, np 15315 and np 15337

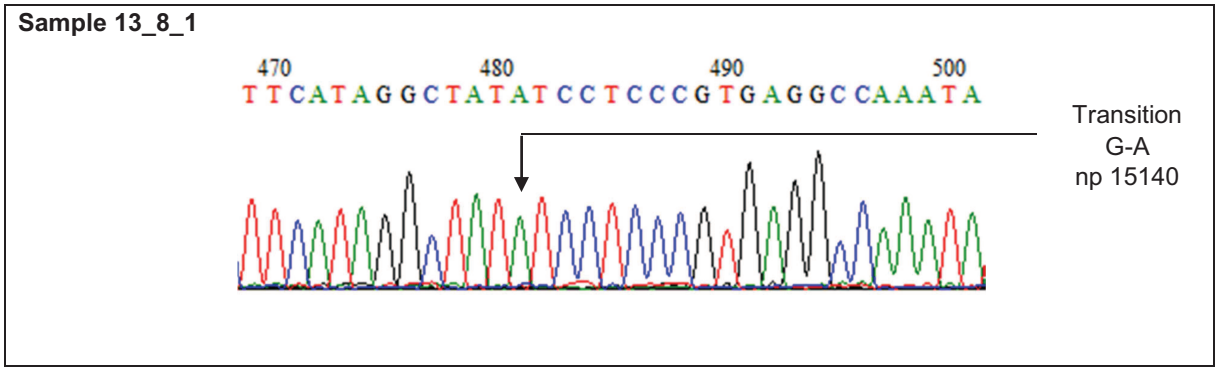
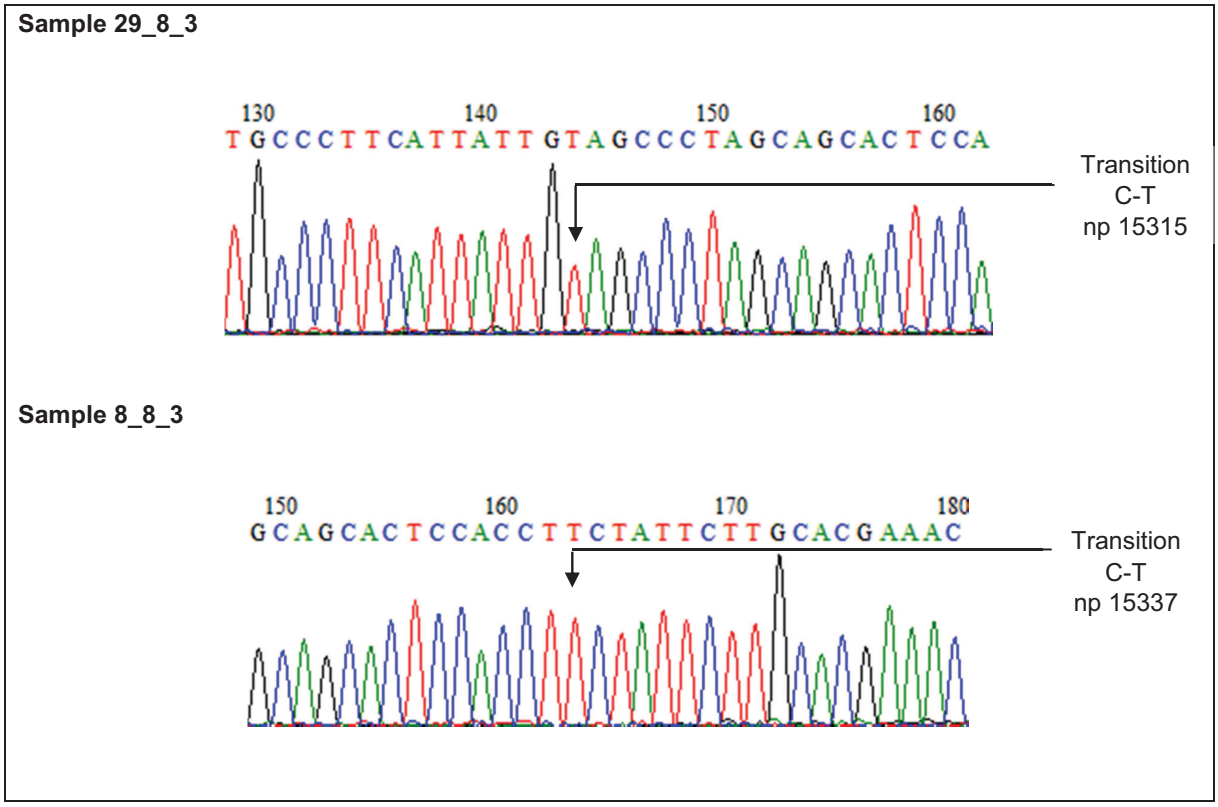


Figure 6.67 Continued…



Sample name 13_8_1 refers to sample TS_3066_13 primer region 8, sequence primer 1; sample name 29_8_3 refers to sample TS_4037_29 primer region 8, sequence primer 3; sample name 8_8_3 refers to sample TS_2097_8 primer region 8, sequence primer 3; sequence region presented contained transition before alignment with rCRS, therefore numbering presented above is not in accordance with rCRS alteration position.

The electropherograms of the novel haplogroup L sequence variants observed in primer region 8 of the Tswana-speaking individuals of this investigation do not display any sequencing artefacts or background noise. The peak morphologies of the sequence variants are good and therefore the possibility is ruled out that the peaks have been called incorrectly because of sequence artefacts or errors. All of these novel haplogroup L sequence variants should, however, be re-sequenced in order to verify their validity.

Four sequence alterations reported to have disease associations have been observed in primer region 8. These sequence alterations are presented in Table 6.25.

Table 6.25 Reported mtDNA sequence alterations with disease associations within primer region 8

Locus	Sequence alteration	Number of individuals	Disease	Reference
ND5	G13708A	2	LHON Increased MS risk PD	Bosley and Abu-Amero, 2010 Yu *et al.*, 2008 Brown *et al*., 1996
CYTB	G15812A	1	LHON	Bosley and Abu-Amero, 2010
tRNA threonine	T15942C	1	Possibly LVNC-associated	Tang *et al*., 2010
	A15951G	2	LHON	Li *et al.*, 2006

Sequence alterations are displayed as the position at which the mutation occurred with the wild type nucleotide indicated in front of the np and the mutant type nucleotide indicated after the np; the number of individuals = number of Tswana-speaking individuals of this investigation that displayed the mutation; disease associations were reported in one or more publications and have been considered as possibly pathologic; LVNC = left ventricular noncompaction; LHON = Leber hereditary optic neuropathy; MS = multiple sclerosis; PD = Parkinson's disease; Adapted from MITOMAP: A Human Mitochondrial Genome Database. http://www.mitomap.org, 2011.

LHON has been described in North American, European, Japanese and Korean populations (Bosley and Abu Amero, 2010) and has primarily been associated with three mutations, at np 3460, 11778 and 14484, that affect the mtDNA complex I genes (Wallace *et al.*, 1988). Although the pathogenicity of these mutations in LHON have been established, the incomplete penetrance of these mutations provides strong evidence that other secondary genetic and environmental factors play a role in the development of the disease. The secondary genetic factors include point mutations and the haplogroups of the affected individuals, such as haplogroup J and H, which have been associated with different levels of penetrance of the disease (Torroni *et al.*, 1997; Howell *et al.*, 2003).

Several mutations with an association with LHON have been observed in different Tswana-speaking individuals of this investigation. These mutations consist of a transition in the *ND1* gene region at np 4025, a transition in the *COI* gene region at np 6261 and a transition in the *ND3* gene region at np 10237. None of these mutations has been reported as a primary LHON disease mutation and has not been interpreted as indicating risk of the prevalence of LHON among the Tswana-speaking individuals of this investigation, as discussed in Section 6.6.4 and Section 6.6.6.

The transitions at np 13708 in the *ND5* gene region and at np 15812 in the *Cytb* gene region of the Tswana-speaking individuals of this investigation have been identified as possible secondary genetic factors in the development of LHON disease (Brown *et al.*, 1994; Bosley and Abu Amero, 2010). The transitions at np 13708 and np 15182 have been reported as secondary because they only moderately affect the conserved amino acids at these encoded sites and are only observed in 10% to 15% of European individuals

presenting with LHON disease and it is therefore unlikely that they cause LHON on their own (Johns, 1991). It is believed that these transitions play a role in the expression of the disease rather than in its cause. Another transition observed in two Tswana-speaking individuals of this investigation at np 15951 in the tRNA threonine coding region has been reported in Chinese LHON patients and is associated with the Asian haplogroup D4. This transition was localised to a highly conserved region of the tRNA threonine secondary structure that affects the pre-tRNA processing, leading to drastically lower levels of tRNA threonine. This is believed to be the reason for the higher prevalence of LHON in Chinese individuals (Li *et al.*, 2005). This transition has therefore, as in the case of the other LHON-associated mutations observed in this investigation, been connected to a specific ethnic group and specific haplogroups, which does not necessarily imply that the same would be true for the Tswana-speaking individuals of this investigation.

The primary LHON mutations at np 3460, 11778 and 14484 have not been observed in the Tswana-speaking individuals of this investigation, which indicates that the LHON-disease-causing mutations are not present in this cohort of Tswana-speakers of South Africa and therefore that the np 13708 and 15182 transitions observed in the three different Tswana-speaking individuals of this investigation probably do not indicate any association with LHON disease. Further investigation of association between LHON disease and haplogroup L individuals would have to include multigenerational family histories, haplogroup associations and secondary genetic mutations specific to African populations (Bosley and Abu Amero, 2010).

The secondary LHON disease mutations, at np 4216 and np 13708, are further associated with susceptibility to multiple sclerosis (MS) in a large Finnish cohort (Yu *et al.*, 2008). The prevalence of the mutation in different ethnic groups suggests that the relevance of the mutation in the disease aetiology is connected to the haplogroups of the individuals and since these studies were performed in Finnish and European cohorts, the pathogenic implications of this mutation in African haplogroup L populations is unclear.

The transition at np 13708 has also been reported to have an association with Parkinson's disease. The pathogenicity of the transition is, however, not considered to be certain since this transition causes a change in a moderately conserved amino acid and is not expected to have a detrimental effect on the individual (Brown *et al*., 1996). The role of this transition in Parkinson's disease has been described as slight and subtle and therefore it is not

regarded in this study as a significant indicator of Parkinson's disease risk in the Tswana-speaking individuals of this investigation.

Many of the structural RNA gene mutations associated with LVNC disease have been observed in the Tswana-speaking individuals of this investigation. These include mutations at np 721, 850, 961, 921 and 2755, which were discussed in Section 6.6.2. The transition at np 15942 could be added to this list of reported LVNC disease-associated mutations, as it is considered possibly to be involved with LVNC disease (Tang *et al.*, 2010). The tRNA threonine is reported as a mutation hotspot for mitochondrial diseases and the mutation at np 15942 has been associated with other diseases caused by mutations in the tRNA threonine, casting doubt on its pathogenicity in terms of LVNC (Tang *et al.*, 2010). Further investigation is necessary to determine if this transition has pathogenic properties in the Tswana-speaking population of South Africa.

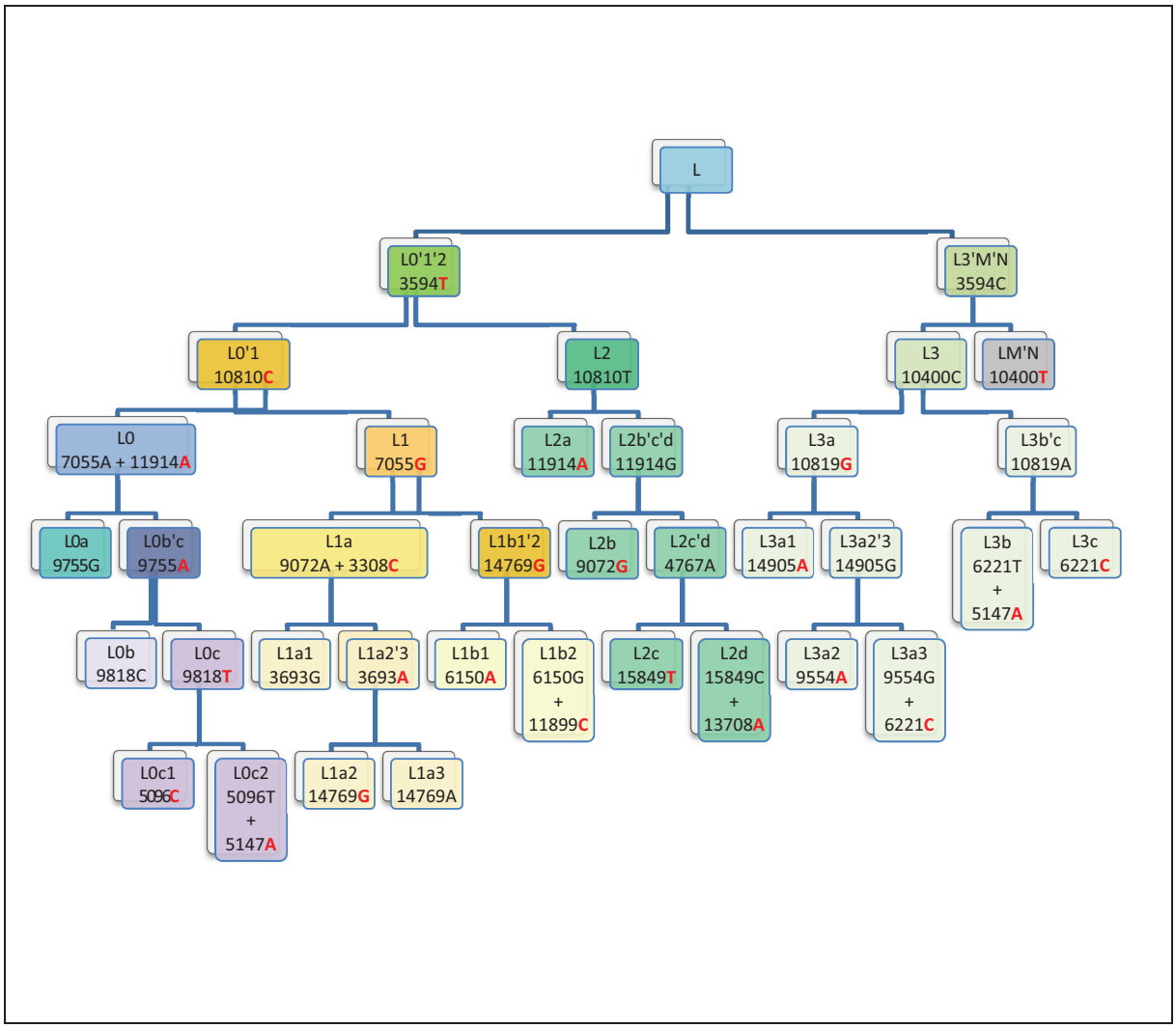
6.7 HAPLOGROUP CLASSIFICATION OF THE TSWANA POPULATION

The assignment of haplogroups is based on the presence of sequence variants in the mtDNA genomes under investigation, as identified by comparison with the rCRS (Wallace, 1994). Two classification schemes, the Wallace classification system (Wallace, 2004) and the PhyloTree classification system (Van Oven and Kayser, 2009), have been used to assign haplogroups to the mtDNA sequences of this investigation. The classification systems are described in Section 5.12. Because of the high mutation rate of the control region of the human mitochondrial genome, which could obscure evolutionary history through a high incidence of reverse mutations (Ingman *et al.*, 2000), only the coding regions of the mtDNA sequences of the Global African, All African and Tswana datasets have been used for haplogroup assignment by both classification systems. The haplogroup assignments of the mtDNA sequences of the Global African, All African and Tswana datasets are presented in Appendix B, in which both the haplogroup assignments of the Wallace classification system (2004) and the PhyloTree classification system (Van Oven and Kayser, 2009) are presented, enabling comparison of the two haplogroup classification schemes per individual.

6.7.1 The haplogroup classification systems used in this investigation

The classification system adapted from Wallace (2004) was based on informative SNPs and designated for use in the classification of the major L haplogroups and sub-haplogroups only. It was adapted for the purpose of haplogroup assignment of mtDNA coding regions of individuals of African origin by the CGR at the North-West University (Wallace, 2004) and was based on informative SNPs identified through high-resolution RFLP analyses. The Wallace classification system (2004) is presented in Figure 6.68.

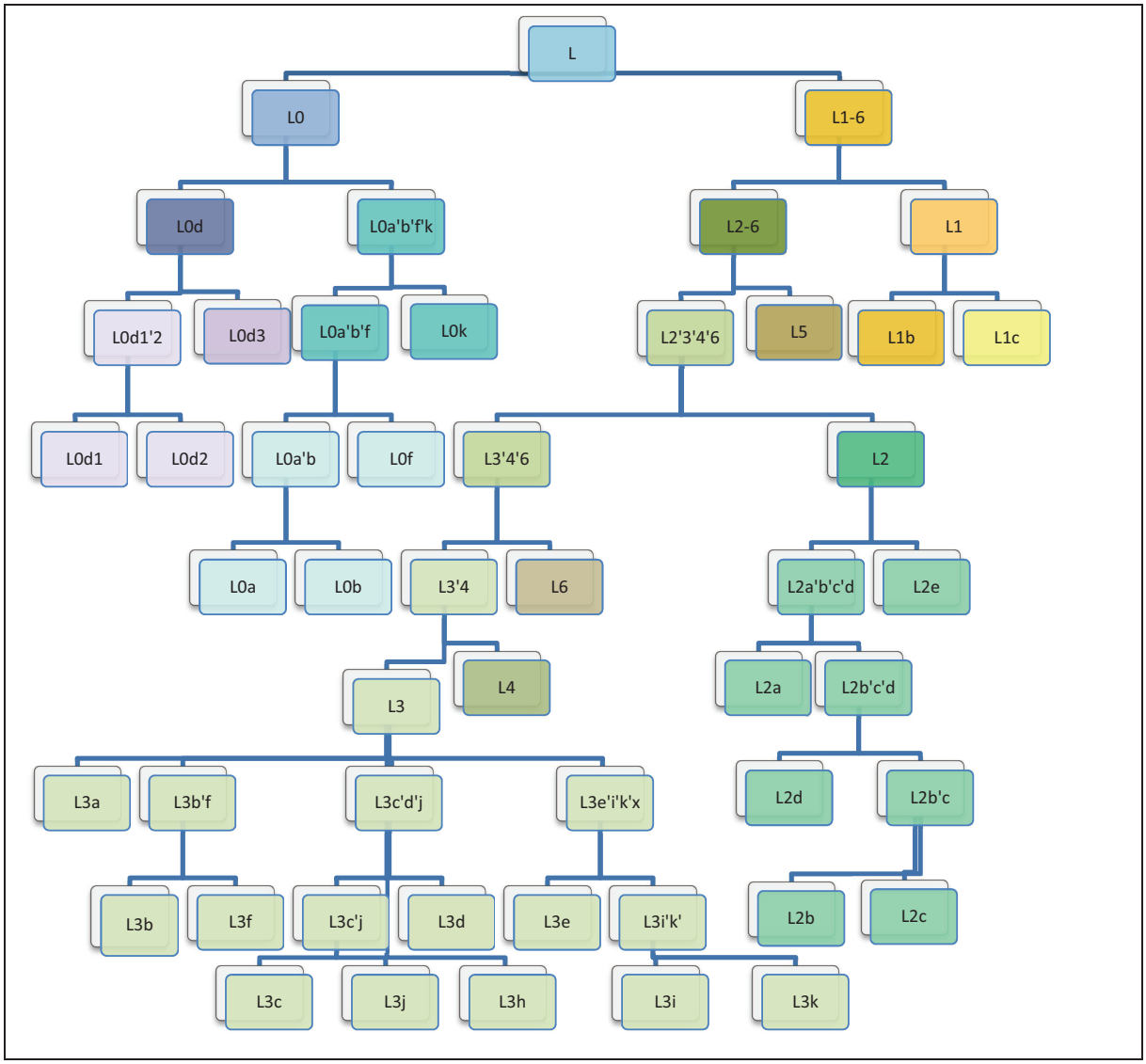
Figure 6.68 Wallace classification system of informative SNPs used to define macrohaplogroup L



Wallace haplogroup classification scheme for macrohaplogroup L. The numbering in the text boxes indicate the informative haplogroup defining nucleotide positions. The letters that follow the nucleotide positions indicate the type of nucleotide that will define the haplogroup at a specific nucleotide position. The **bold, red** letters indicate that the type of nucleotide is different from what was present in the rCRS at that specific nucleotide position and the letters in black indicate that the type of nucleotide is in accordance with the rCRS. Adapted from Wallace, 2004.

The PhyloTree classification system (Van Oven and Kayser, 2009) is a publicly available global human phylogenetic tree that classifies all global haplogroups based on all the currently available complete mtDNA sequence data that have been published and peer-reviewed (Van Oven and Kayser, 2009). This haplogroup system has a high level of haplogroup resolution because it is based on a broad set of global mtDNA sequences and on haplogroup-defining mutations observed in the full mitochondrial genomes of individuals. Because of the accessibility of this classification scheme, mtDNA sequence contributions and sequence variant reports are encouraged, ensuring that the published phylogenetic structure in the PhyloTree classification system is peer-reviewed and therefore scientifically valid and current. This global phylogenetic human tree and haplogroup classification system has greatly contributed to the development of a standard haplogroup nomenclature and classification system that has been used by many global mtDNA studies and has enabled the comparison and discussion of mtDNA sequence variants according to a single haplogroup classification system (Batini *et al.*, 2011; de Fillippo *et al.*, 2010; Scheinfeldt *et al.*, 2010). The extensive collection of mitochondrial DNA sequences used provides for a highly resolved haplogroup classification system that is inclusive of many rare variants. An outline of the basic hierarchy of the macrohaplogroup L of the PhyloTree classification system (Van Oven and Kayser, 2009) is presented in Figure 6.69.

Figure 6.69 Outline of the PhyloTree classification system for macrohaplogroup L



The PhyloTree haplogroup classification scheme for macrohaplogroup L is presented. The colours of the text boxes are correlated with the major L haplogroups and sub-haplogroups. Each haplogroup and sub-haplogroup is defined by several nucleotide positions and can therefore not be presented here. For detail about the haplogroup-defining nucleotides the publicly available PhyloTree can be accessed at http://www.phylotree.org. Adapted from (Van Oven and Kayser, 2009).

6.7.1.1 Overall comparison of the two haplogroup classification systems used in this investigation

The purpose of using two haplogroup classification systems was to ensure a representative assignment of haplogroups based on both the RFLP method and the full genome-sequencing method. The classification systems were designed at different times and therefore reflected some differences in nomenclature, as well as the structure of the hierarchical organisation of the maternal ancestry of macrohaplogroup L. These differences between the two classification systems are highlighted throughout this section as the specific haplogroup results are discussed in more detail.

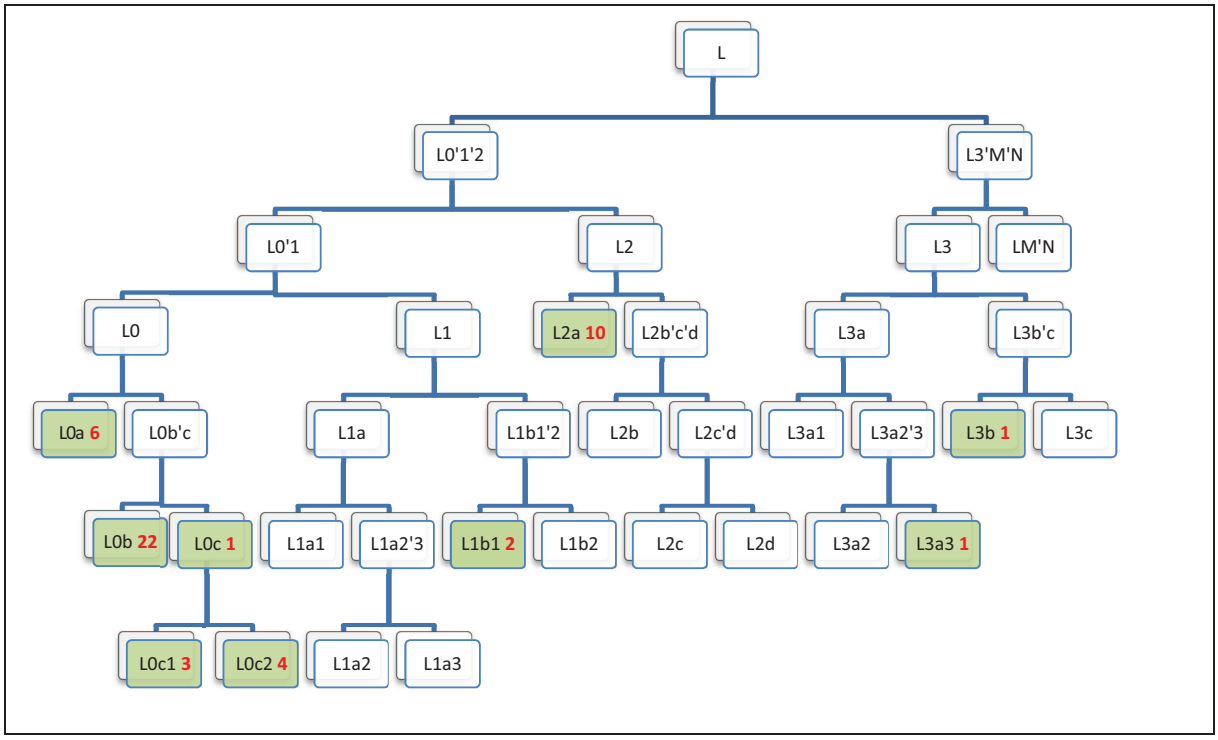
Mitochondrial DNA variation correlating to geographic area and ethnic origin was initially established by the *HpaI* RFLP enzyme restriction site at np3592 owing to a C3594T transition, which was found to be present in African individuals but not in Asian or European individuals (Wallace *et al.*, 1999). The Wallace classification system (2004) is based on this alteration as the unique marker that distinguishes the L3 haplogroup and L0'1'2 macrohaplogroup from each other. The Wallace classification system (2004) then further sub-grouped the L0'1'2 haplogroup by defining the L0'1 haplogroup by the presence of a T10810C transition and subsequently classified the L1 haplogroup by the presence of an A7055G transition. The Wallace classification system (2004) therefore suggested that haplogroups L0, L1 and L2 had a common ancestral sequence alteration not present in haplogroup L3, therefore branching off from the paragroup L0'1'2. This is contradictory to the PhyloTree classification system, where the split in the haplogroup structure is between haplogroup L0 and paragroup L1'-6', indicating different ancestral origins for haplogroups L1 to L6 and the ancient L0 haplogroup, as presented in Figure 6.69.

The discrepancies between the two haplogroup classification systems were, however, not problematic and assisted in providing a detailed and comprehensive haplogroup assignment within this investigation. Differences in haplogroup assignments between the two classification systems were at the level of clades and lineages and in this regard the PhyloTree classification assisted with the finer resolution of haplogroups observed in the Tswana population in this investigation. Both the PhyloTree and Wallace haplogroup classification systems provided good haplostructure to the sequence variation found in the Tswana population. The PhyloTree classification system (Van Oven and Kayser, 2009) was, however, preferred in this investigation based on the high resolution of the haplogroup structure and the incorporation of a broad set of published sequence variants, especially with regard to haplogroup L. It was used extensively in conjunction with the Wallace classification system (2004).

6.7.2 Haplogroups of the Tswana-speaking individuals of this investigation

The haplogroup assignments of the mtDNA sequences of this investigation, as described in Section 5.12, are presented in Appendix A. Nine different haplogroups were observed when using the Wallace classification system (2004) in the Tswana cohort of 50 individuals. The distribution of haplogroups when using the Wallace classification system (2004) is illustrated in Figure 6.70.

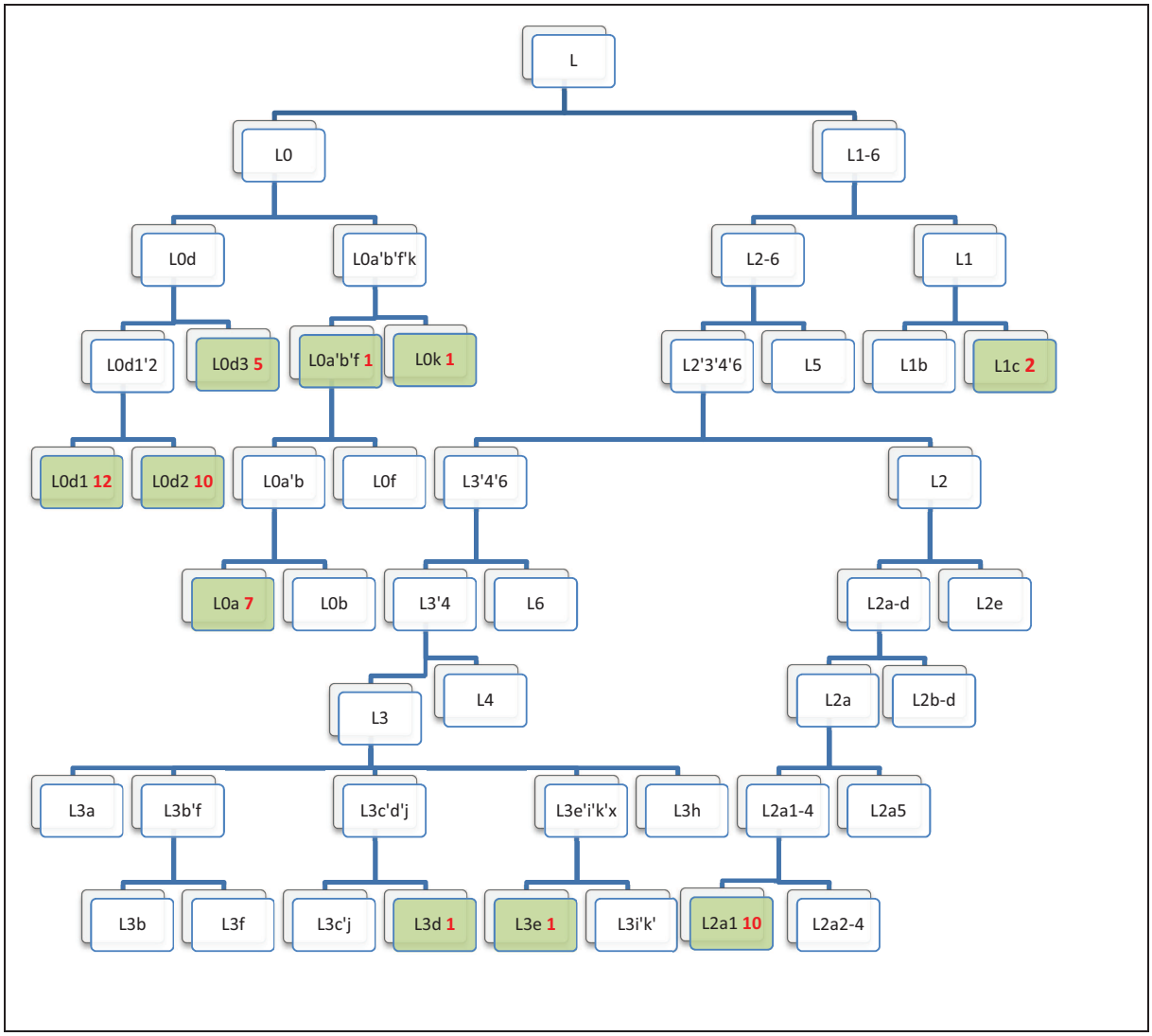
Figure 6.70 Haplogroup distribution of the Tswana-speaking population under investigation according to the Wallace classification system



The haplogroup hierarchical structure is presented based on the Wallace haplogroup classification system (2004) as described in Section 5.12. The haplogroup distribution of the Tswana-speaking individuals of this investigation is indicated by green highlighting and the number of Tswana-speaking individuals of this investigation that were assigned to each of the haplogroups, is indicated by a number in **red**. For the haplogroup-defining mutations, refer to Figure 6.68.

Ten different haplogroups were assigned to the Tswana cohort of 50 individuals when the PhyloTree classification system (Van Oven and Kayser, 2009) was used and the distribution of the haplogroups is illustrated in Figure 6.71.

Figure 6.71 Haplogroup distribution of the Tswana-speaking population under investigation according to the Phylotree classification system



The haplogroup hierarchical structure based on the PhyloTree classification system (Van Oven and Kayser, 2009) as described in Section 5.12. The haplogroup distribution of the Tswana-speaking individuals of this investigation is indicated by green highlighting and the number of Tswana-speaking individuals of this investigation that were assigned to each of the haplogroups, is indicated by a number in **red**. For the haplogroup-defining mutations, refer to Section 5.12.

6.7.3 Haplogroup L0

Although there was migration from Eurasia to north and northeast Africa during the early Palaeolithic period, the M1 and U6 clades carried by these populations never penetrated sub-Saharan Africa, making Africa the richest source of L0 and L1 haplogroups in the world (Behar *et al.*, 2008). Most of the haplogroups present in the Tswana population were assigned to haplogroup L0 with 36 of the 50 (72%) Tswana individuals belonging to one of the sub-groups of haplogroup L0. Haplogroup L0 is defined by the Wallace classification system by the defining SNPs as listed in Table 6.26.

Table 6.26 Tswana-speaking individuals of this investigation assigned to haplogroup L0 by the Wallace classification system

Haplogroup subtype	Defining nucleotide polymorphisms	Tswana dataset sample	Reference
L0a	C3594T T10810C G11914A	TS_2082 TS_3066 TS_3486 TS_4034 TS_4080 TS_5091	Wallace, 2004
L0b	C3594T G9755A T10810C G11914A	TS_2075 TS_2091 TS_2095 TS_2097 TS_3002 TS_3027 TS_3117 TS_3236 TS_3459 TS_3505 TS_4111 TS_4027 TS_4037 TS_4051 TS_4056 TS_4075 TS_4083 TS_4089 TS_4117 TS_5044 TS_5083 TS_5085	
L0c	C3594T G9755A C9818T T10810C G11914A	TS_5063	
L0c1	C3594T T5096C G9755A C9818T T10810C G11914A	TS_4063 TS_3471 TS_4032	
L0c2	C3594T G5147A G9755A C9818T T10810C G11914A	TS_2074 TS_3075 TS_3461 TS_3506	

Polymorphisms are indicated by using nucleotide position numbers according to the rCRS with the original rCRS nucleotide preceding the nucleotide position and the variant allele following the nucleotide position number.

Haplogroups were assigned by the PhyloTree classification system to all the L0 haplogroups classified by the Wallace classification system and are listed in Table 6.27. The haplogroups assigned by the two classification systems differed at the sub-haplogroup level, as presented in Table 6.27.

Table 6.27 Tswana-speaking individuals of this investigation assigned to haplogroup L0 by the PhyloTree classification system

Haplogroup subtype	Defining nucleotide polymorphisms	Tswana dataset sample	Reference
L0d	C1048T A1438G C3516A 4232C T5442C T6185C T6815C C8113A G8152A G8251A C9042T A9347G G10589A G12007A A12720G T12121C G15466A G15930A T15941C	TS_2082 TS_3066 TS_3486 TS_4080 TS_5091 TS_2075 TS_2091 TS_2095 TS_2097 TS_3002 TS_3027 TS_3117 TS_3236 TS_3459 TS_3505 TS_4111 TS_4027 TS_4037 TS_4051 TS_4056 TS_4075 TS_4083 TS_4089 TS_4117 TS_5044 TS_5083 TS_5085	PhyloTree (Van Oven and Kayser, 2009)
L0k	T850C C1048T T1243C C3516A G4541A T4586C T4907C T5442C A5811G T6185C T8911C G8994A C9042T A9347G C9818T G10589A G12007A A12720G A7257G A9136G A10499G A10876G C10920T C11296T T11299C A11653G G13590A T13819C G13928C T14020C T14182C T14371C T14374C	TS_4034	

Table 6.27 Continued…

Haplogroup subtype	Defining nucleotide polymorphisms	Tswana dataset sample	Reference
L0a'b'	C1048T A2245G C3516A T4586C T5442C C5603T T6185C C8428T A8566G C9042T A9347G G9755A C9818T G10589A G12007A A12720G A11641G C15136T	TS_5063	PhyloTree (Van Oven and Kayser, 2009)
L0a	C1048T T14308C A2245G C3516A G5231A T5442C G5460A C5603T T6185C C8428T A8566G C9042T A9347G G9755A C9818T G10589A G12007A A12720G T4586C A11641G C15136T G11176A	TS_4063 TS_3471 TS_4032 TS_2074 TS_3075 TS_3461 TS_3506	

Polymorphisms are indicated by using nucleotide position numbers according to the rCRS with the original rCRS nucleotide preceding the nucleotide position and the variant allele following the nucleotide position number. $ = individual that could not be fully classified according to the PhyloTree, 2009 haplogroup classification scheme.

As discussed in Section 6.6.4.1, to observe high frequencies of novel mutations in a single individual, as displayed by Tswana-speaking individual TS_5063 of this investigation, was not expected. Furthermore, this individual could not be fully assigned to sub-haplogroups by any of the two classification systems because of the absence of the transition G-A at position 5147 in the L0c2 lineage when using the Wallace classification system (2004) and the absence of the transitions G5231A, G5460A, G11176A and T14308C when using the PhyloTree classification system (Van Oven and Kayser, 2009). This finding suggests that the observed sequence motifs in this individual do not correspond to the sequence motifs observed in the haplogroups of either the Wallace classification system (2004) or the PhyloTree classification system (Van Oven and Kayser, 2009) and reiterates the importance of investigating this sample further for laboratory sequencing error or human error, such as sample contamination early in the handling of the sample. Alternatively, if error is dispelled, this individual could have displayed a unique set of sequence alterations, which would require further investigation to determine whether other Tswana-speaking individuals of South Africa harbour the same sequence motif patterns. The sequence alterations of this sample are presented in Table 6.28.

Table 6.28 Sequence alterations observed in the Tswana-speaking individual TS_5063

Sample name	Haplogroup	Sequence variance				
TS_5063	L0a'b (PhyloTree) and L0c (Wallace, 2004)	A750G G1415A T2885C T4586C A5656G C7256T C8468T T8987C G9755A G10589A T10915C G12007A C13431T	G769A A1438G C3516A T5782C A4769G G7521A A8566G C9042T T9758C C10664T G11176A G12527A C13506T 14766T	T825A, A2245G C3594T C4814T T6185C A7765G C8655T A9058G C9818T G10688A A11641G A12720G C13650T C15136T	G1018A A2706G A4104G A4943G C7028T T8047C A8701G A9437G A10398G T10810C G11719A A13105G C13767A A15326G	C1048T G2758A C4312T T5442C A7146G C8428T A8860G T9540C G10427A T10873C G11914A A13276G T14110C G15431A

Sequence variance indicated by the nucleotide as present in rCRS, followed by the nucleotide position and the bp alteration. The light blue highlights indicate the defining SNPs of the Wallace classification system and the grey highlights indicate the defining SNPs of the PhyloTree classification system. The defining SNPs that are shared between the two systems are indicated in green highlight.

Haplogroups L0d and L0k have generally been observed among the African Khoi-San populations and occur at the root of the human mitochondrial DNA tree (Salas *et al.*, 2002) When the PhyloTree classification system was used to assign haplogroups, 27 of the Tswana individuals were assigned to haplogroup L0d. On further sub-grouping, 12 Tswana-speaking individuals of this investigation were found to belong to haplogroup L0d1, ten individuals belonged to haplogroup L0d2 and five individuals belonged to haplogroup L0d3. The Wallace classification system assigned all of the Tswana-speaking individuals of this investigation who had been assigned to haplogroup L0d3 and one Tswana-speaking individual, TS_4034, who had been assigned to haplogroup L0a'b'k by the PhyloTree classification system (Van Oven and Kayser, 2009), to haplogroup L0a based on the shared haplogroup-defining mutations at np 3594, np 10810 and np 11914. It further assigned the Tswana-speaking individuals who belonged to haplogroups L0d1 and L0d2 to haplogroup L0b, based on the shared haplogroup-defining mutation at np 9755, which was also cited by the PhyloTree classification system (Van Oven and Kayser, 2009) as a haplogroup-defining site for haplogroup L0d1'2. The Wallace classification system did not resolve the haplogroup L0b into further sub-groups, as was the case with the PhyloTree classification system (Van Oven and Kayser, 2009).

The Wallace classification system further assigned sample TS_4034 to haplogroup L0a, while the PhyloTree classification system (Van Oven and Kayser, 2009) assigned it to haplogroup L0k. Therefore the Wallace classification could not distinguish between this sample and the rest of the Tswana-speaking individuals belonging to haplogroup L0a, as

opposed to the PhyloTree classification system (Van Oven and Kayser, 2009), which indicated that there was a clear distinction between TS_4034 and the rest of the Tswana-speaking individuals belonging to haplogroup L0d3. This finding indicates that the Wallace classification system (2004) demonstrates lower resolution than the PhyloTree classification system (Van Oven and Kayser, 2009) and therefore is not the preferred haplogroup classification system to be used.

Previous studies reported that the L0d1 and L0d2 lineages of the PhyloTree classification system and equivalently the L0b lineages of the Wallace classification system were primarily present in the Khoi-San population of southern Africa and that the L0d3 lineage of the PhyloTree classification system (Van Oven and Kayser, 2009) and equivalently the L0a lineages of the Wallace classification system (2004) were primarily present in a Tanzanian population (Gonder *et al.*, 2007). The Tswana population of this investigation was therefore associated with the Khoi-San population of southern Africa through the shared haplogroups and it is postulated that the Tswana population obtained these lineages through genetic flow due to population admixture with the Khoi-San populations of southern Africa. The Bantu speakers expanded to the southern regions of Africa about four kya via two main routes (Salas *et al.*, 2002). The one group migrated along the coast and settled in the present-day KwaZulu-Natal about three centuries AD and a second group migrated through modern-day eastern Zimbabwe to reach the northern parts of South Africa around 500 years AD (Pereira *et al.*, 2001). It is postulated that the Bantu populations met up with the Khoi-San populations, followed by a Khoi-San assimilation by the Bantu speakers. The presence of the L0d haplogroup in this population is high and suggests a much higher assimilation of the Khoi-San genetic material into this population, as has been reported for the eastern African Bantu speakers (Salas *et al.*, 2002), which could be a result of high levels of gene flow, most probably because of the close geographical location between the Tswana-speaking population and the Khoi-San population. All but one of the defining polymorphisms for haplogroup L0d according to the PhyloTree classification system (Van Oven and Kayser, 2009) is present in the Tswana individuals. The A1438G alteration has not been observed in any of the Tswana-speaking individuals of this investigation. Recurring mutations are present at this position in other studies (Gonder *et al.*, 2007) and is a likely reason for the observation in this study.

Based on the sequence variation evidence provided in this investigation, as discussed in Section 6.6.8, five Tswana-speaking individuals of this investigation display a novel sequence variant at np 12436 in the *ND5* gene region and a haplogroup L novel sequence

variant at np 15337. On further investigation, it was observed that these five Tswana-speaking individuals were phylogenetically grouped together, as discussed in Section 6.8, and did not share the same sequence motifs as the other individuals that belonged to haplogroup L0d1b according to the PhyloTree classification system. These novel mutations could, according to the criteria for haplogroup-defining mutations in the PhyloTree classification system (Van Oven and Kayser, 2009), be regarded as candidate haplogroup-defining mutations for haplogroup L0d1b2. In order to confirm this, it is suggested that a phylogenetic tree be constructed of the available published mtDNA sequences that belong to haplogroup L0d1b to establish if the mtDNA sequences of the five Tswana-speaking individuals of this investigation, who display the novel transition, group together in the presence of sequence variation over a broad set of mtDNA sequences that belong to this haplogroup, therefore indicating a novel haplogroup-defining mutation.

The same principles discussed in the previous paragraph appy to the novel haplogroup L sequence variant at np 10128, which occurs within the *ND3* gene regions of another five Tswana-speaking individuals of this investigation, who belong to haplogroup L0d3 and phylogenetically cluster together, as discussed in Section 6.8. The sequence variation observed in this investigation and discussed in Section 6.6.6 indicate a separate and distinct sequence motif shared by the Tswana-speaking individuals of this investigation only. Therefore it has been concluded that this sequence variant is a valid candidate for a sub-haplogroup of haplogroup L0d3 in the PhyloTree classification system (Van Oven and Kayser, 2009).

The Tswana-speaking individual, TS_4034, was assigned to haplogroup L0k by the PhyloTree classification system (Van Oven and Kayser, 2009), which shared a maternal ancestor with the L0a, L0b and L0f clades and indicated an independent origin from the L0d lineages. L0k has primarily been observed in the Khoi-San populations of southern Africa (Salas *et al.*, 2002) and it is thought that this haplogroup developed parallel to the L0d lineages early on in human evolutionary history (Behar *et al.*, 2008). All of the PhyloTree defining polymorphisms for haplogroup L0k are present in the Tswana individual, except for the transition at np 12720. This position has also been listed by previous studies as a position of recurrent mutations (Gonder *et al.*, 2007) and is a likely reason for the absence of the alteration at this site. As with the L0d haplogroup, the presence of this haplogroup indicates a probable gene flow between Khoi-San and Tswana populations. As discussed, the Wallace classification system (2004) could not

assign this individual to a haplogroup equivalent to haplogroup L0k as assigned by the PhyloTree classification system.

The L0a haplogroup is believed to be present in many of the African populations and reflects the migratory pattern of the Bantu-speaking populations into eastern and southern Africa over the past few thousand years (Salas *et al.*, 2002). In this investigation, seven of the Tswana individuals are assigned to the L0a haplogroup when using the PhyloTree classification system. All the PhyloTree defining positions for the L0a haplogroup are observed except for the transition at nucleotide position T12270C, which is not observed in three of the Tswana individuals. All of the Tswana-speaking individuals of this investigation that belong to haplogroup L0a according to the PhyloTree classification system (Van Oven and Kayser, 2009), belong to haplogroup L0c when the Wallace classification system is used, based on the shared alteration at np 9818T, which is one of the haplogroup L0a'b'f'k haplogroup-defining mutations cited by the PhyloTree classification system (Van Oven and Kayser, 2009).

The Tswana population is believed to have originated from the Bantu populations that migrated to southern Africa many years ago and the presence of the L0a haplogroups in the Tswana population of this investigation supports this theory. The Tswana haplogroup L0a clades are evenly distributed between the L0a1 and L0a2 lineages of the PhyloTree classification system (Van Oven and Kayser, 2009). There are no individuals that belong to the L0a3 lineage. Further sub-grouping of haplogroup L0a assigns three of the Tswana individuals to the L0a1b lineage and four of the Tswana individuals to the L0a2a2a lineage according to the PhyloTree classification system (Van Oven and Kayser, 2009). The 9 bp deletion in the COII/ tRNA lysine coding region is observed in all the individuals of the L0a2a2a lineage. The L0a haplogroup associated with the 9 bp deletion has been identified as an important marker of Bantu migration to southern Africa (Soodyall *et al.*, 1996) and has been identified to be common in southeastern African Bantu speakers, with its origin in central Africa rather than eastern Africa where L0a has its origin (Salas *et al.*, 2002).

In conjunction with the PhyloTree classification system (Van Oven and Kayser, 2009), the Wallace classification system (2004) assigns the three Tswana-speaking individuals belonging to haplogroup L0a1b to haplogroup L0c1 based on the shared mutation at np 5096. It also assignsthe four Tswana-speaking individuals belongingto haplogroup L0a2 to haplogroup L0c2, based on the shared mutation at np 5147. The Wallace

classification system does not provide further haplogroup-defining mutations for higher resolution of the haplogroups, as in the case of the PhyloTree classification system, and therefore does not take the 9 bp deletion into account.

6.7.4 Haplogroup L1

A small number of Tswana individuals (4%) are assigned to haplogroup L1 when using the Wallace classification system (Wallace, 2004). The Wallace classification system defines the L1 haplogroup and sub-groups by the SNPs listed in Table 6.29. Only two individuals of the Tswana population of 50 individuals are assigned to haplogroup L1b1 when the Wallace classification system (2004) is used.

Table 6.29 Tswana-speaking individuals of this investigation assigned to haplogroup L1 by the Wallace classification system

Haplogroup subtype	Defining nucleotide polymorphisms	Tswana dataset sample	Reference
L1	C3594T T10810C A7055G	---	Wallace, 2004 (Personal communication)
L1b1	C3594T T10810C A7055G A9072G G6150A	TS_2093 TS_3015	

Polymorphisms are indicated by using nucleotide position numbers according to the rCRS with the rCRS bp preceding the bp number and the base substitution following the bp number.

Two individuals (4%) from the Tswana population are assigned to haplogroup L1c2a when using the PhyloTree classification system (Van Oven and Kayser, 2009). These sequences contain all the defining alterations for haplogroups L1, L1c, L1c1'2'4, L1c2'4, L1c2 and L1c2a, except at nucleotide positions A13149G, 5899.1C and 2156.1A.

Table 6.30 Tswana-speaking individuals of this investigation assigned to haplogroup L1 by the PhyloTree classification system

Haplogroup subtype	Defining nucleotide polymorphisms	Tswana dataset sample	Reference
L1	G3666A A7055G G7389A T13789C T14178C G14560A	TS_2093 TS_3015	PhyloTree (Van Oven and Kayser, 2009)
L1c	2395dA A5951G T6071C G8027A A9072G G10586A A12810G C13485T T14000A C14911T		
L1c2'4'6	T10321C		
L1c2'4	5899.1C C12049T A13149G		
L1c2	G6150A T6253C A7076G G7337A A8784G T8877C A10792G C10793T A11654G		
L1c2a	T1420C 2156.1A C15016T T15784C		

Polymorphisms are indicated by using nucleotide position numbers according to the rCRS with the rCRS bp preceding the bp number and the base substitution following the bp number.

The Tswana individuals assigned to haplogroup L1 by the Phylotree classification system display almost all the sequence alterations indicated as haplogroup-defining by the PhyloTree classification system (Van Oven and Kayser, 2009) and are the same two Tswana-speaking individuals also assigned to haplogroup L1 by the Wallace classification system (2004). Current evolutionary models suggest that the early modern humans expanded from East Africa to the rest of the African continent mainly through the L1 haplogroup clade, followed by a later expansion of modern humans towards the eastern and southern regions of Africa that belonged to haplogroup L2 and eventually expanded out of Africa through modern humans that belonged to haplogroup L3 (Behar *et al.*, 2008). L1c haplogroups display evidence of origin in the equatorial regions (Behar *et al.*, 2008), are regarded as markers of Bantu expansions from western Africa to east and southeast Africa (Quintana-Murci *et al.*, 2008) and are rarely present in southern Africa (Gonder *et al.*, 2007). The L1c2 clade originated from the ancestral L1c clade that lived in central Africa (Quintana-Murci *et al.*, 2008) and is highly divergent from the L0d and L0k clades. The presence of the L1c2 clade in the Tswana-speaking individuals of this investigation therefore provides evidence of a maternal ancestor who originated and resided somewhere in central Africa and migrated southwards to reach the southern regions of Africa.

6.7.5 Haplogroup L2

The Wallace classification system (2004) defines haplogroup L2 and sub-group L2a by the SNPs listed in Table 6.31. Twenty percent of the Tswana cohort investigated in this study (10 individuals) is assigned to haplogroup L2a when using the Wallace classification system (2004).

Table 6.31 Tswana-speaking individuals of this investigation assigned to haplogroup L2 by the Wallace classification system

Haplogroup subtype	Defining nucleotide polymorphisms	Tswana dataset sample	Reference
L2a	C3594T G11914A	TS_2077 TS_2103 TS_3107 TS_3466 TS_4013 TS_4106 TS_5060 TS_5062 TS_5066 TS_5086	Wallace, 2004 (Personal communication)

Polymorphisms are indicated by using a nucleotide position according to the rCRS with the original rCRS nucleotide preceding the nucleotide position and the variant allele indicated after the nucleotide position.

All the individuals assigned to haplogroup L2a by the Wallace classification system (2004) are assigned to L2a1 by the PhyloTree classification system (Van Oven and Kayser,

2009). The Tswana-speaking individuals of this investigation contain all the defining alterations for haplogroups L2, L2a-d, L2a, L2a1'2 and L2a1.

Table 6.32 Tswana-speaking individuals of this investigation assigned to haplogroup L2 by the PhyloTree classification system

Haplogroup subtype	Defining nucleotide polymorphisms	Tswana dataset sample	Reference
L2	T2416C G8206A A9221G T10115C G13590A	TS_2077 TS_2103 TS_3107 TS_3466 TS_4013 TS_4106 TS_5060 TS_5062 TS_5066 TS_5086	PhyloTree (Van Oven and Kayser, 2009)
L2a-d	T11944C		
L2a	T7175C		
L2a1'2	C2789T C7274T A7771G G11914A A13803G A14566G		
L2a1	A12693G T15784C		

Polymorphisms are indicated by using nucleotide position numbers according to the rCRS with the rCRS bp preceding the bp number and the base substitution following the bp number.

The L2a haplogroup has been identified as a marker for Bantu expansion and is the most common clade in Africa (Torroni *et al.*, 2001) and one of the most common haplogroups in the central African Bantu populations (Quintana-Murci *et al.*, 2008). The relative frequencies of L2 and L0d as determined by specific mtDNA control region variations and mtDNA-coding region SNPs in previous studies indicated a percentage of between 10% and 36% of haplogroup L2 and up to 90% of haplogroup L0d being present in southern Africa (Salas *et al.*, 2002; Gonder *et al.*, 2007). The haplogroup distribution in this investigation indicates a similar percentage distribution of haplogroup L2, which is represented in 20% of the Tswana population, and haplogroup L0d, which is represented in 54% of the Tswana population. Other studies also reported haplogroup L2a1 to be common in populations of southeastern Africa, with a probable origin in west Africa, followed by a major expansion in southeastern Bantu populations or a population ancestral to the current southeast African populations (Salas *et al.*, 2002).

6.7.6 Haplogroup L3

Two individuals of the Tswana cohort of 50 individuals are assigned to haplogroup L3. TS_3495 is assigned to haplogroup L3a3 and TS_3085 to haplogroup L3b by the Wallace classification system (2004). The sequence site variation that defines haplogroups L3a3 and haplogroup L3b are listed in Table 6.33.

Table 6.33 Tswana-speaking individuals of this investigation assigned to haplogroup L3 by the Wallace classification system

Haplogroup subtype	Defining nucleotide polymorphisms	Tswana dataset sample	Reference
L3a3	3594C 10400C A10819G 14905G 9554G T6221C	TS_3495	Wallace, 2004
L3b	3594C 10400C 10819A 6221T G5147A	TS_3085	

Polymorphisms are indicated by using nucleotide position according to the rCRS with the original rCRS nucleotide preceding the nucleotide position and the variant allele indicated after the nucleotide position. Where there is no nucleotide indicated preceding the bp number it means that no substitution is required in that position and that the nucleotide following the nucleotide positions is therefore the same as the rCRS.

TS_3495 is assigned to haplogroup L3e1 and TS_3085 is assigned to haplogroup L3d1a1a by the PhyloTree classification system (Van Oven and Kayser, 2009). The haplogroup-defining nucleotide positions according to the PhyloTree classification system (Van Oven and Kayser, 2009) observed in the Tswana-speaking individuals of this investigation are presented in Table 6.34.

Table 6.34 Tswana-speaking individuals of this investigation assigned to haplogroup L3 by the PhyloTree classification system

Haplogroup subtype	Defining nucleotide polymorphisms	Tswana dataset sample	Reference
L3c'd	A13105G	TS_3085	PhyloTree (Van Oven and Kayser, 2009)
L3d1-5	T921C		
L3d	G5147A A7424G T8618C T13886C C14284T		
L3d1	T6680C		
L3d1a	G4048A C7648T G11887A		
L3d1a1	G1503A		
L3d1a1a	A4203G G5471A T10640C T10915C		
L3e'l'k'x	A10819G	TS_3495	
L3e	T2352C T14212C		
L3e1	T6221C C6587T A14152G T15670C T15942C		

Polymorphisms are indicated by using nucleotide position numbers according to the rCRS with the rCRS bp preceding the bp number and the base substitution following the bp number.

Haplogroups L3b and L3e are markers of Bantu expansion (Soodyall *et al.*, 1996; Bandelt *et al.*, 2001b). Haplogroup L3b is a common haplogroup reported mainly in west Africa and to a certain extent in southeast Africa; however, it is rarely present in east Africa or central Africa. The Tswana population therefore carries traces of the Bantu expansion that started in western Africa and expanded to central and southern Africa. Haplogroup L3d is

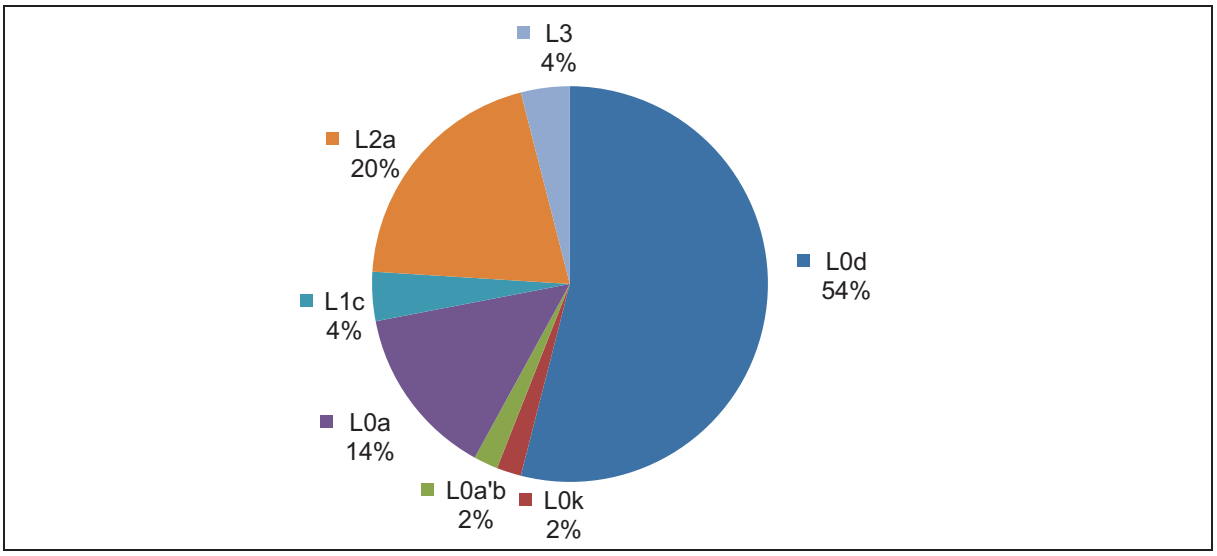
predominantly present in western Africa, with lineage L3d1 being present in southeastern Africa, and L3d is therefore similar in distribution to haplogroup L3b. The lineage L3d3 has been reported in high frequencies in the Khwe and !Kung in the south of Africa and probably spread to the Khoi-San speakers from the Bantu speakers, as was reported (Salas *et al.*, 2002). The presence of the L3d1 haplogroup in the Tswana population indicates genetic flow from the Bantu expansion to southeastern Africa and eventually to southern Africa.

The L3e haplogroup is the most common and ancient clade of the L3 haplogroup and is represented in one third of the L3 haplogroup types in sub-Saharan Africa (Salas *et al*., 2002). L3e1 is commonly reported in southeastern African Bantu speakers with a suggested origin in central Africa (Bandelt *et al.*, 2001). It is postulated that the L3e1 haplogroup originated in central Africa from where it spread via the east of Africa to Ethiopia and downtowards the southeastern regions of Africa. Haplogroup L3e1 is also present at high frequencies in southwest Angola (Coelho *et al.*, 2009), indicating a migration of Bantu speakers that belonged to L3e1 to southwestern Angola from where it spread towards southern Africa. The L3e lineage present in the Tswana population indicates genetic flow from Bantu populations from either southeastern Africa or from southwest Angola. Based on the fact that haplogroup L3e1 has been observed at high frequencies in the Bantu speakers from southeastern Africa (Salas *et al.*, 2002), it is likely that this sub-haplogroup expanded into the rest of South Africa via this route and that the Tswana-speaking individuals of this investigation share a maternal ancestor with the Bantu-speakers that migrated to southern Africa via the eastern route of Africa.

6.7.7 Overview of haplogroups present in the Tswana population

All the haplogroups and sub-groups of the Tswana population of this investigation belong to the African population and represent two distinct components. The first component is the L0d and L0k clades that are representative of the ancestral types harboured by the populations living in southern Africa before the arrival of the Bantu speakers and the second component consists of the rest of the haplogroups and sub-groups that are representative of the haplotypes brought to southern Africa by the Bantu expansion. The distribution of the haplogroups of the Tswana-speaking individuals of this investigation according to the PhyloTree classification system (Van Oven and Kayser, 2009) is presented in Figure 6.72.

Figure 6.72 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 Khoi-San lineages, i.e. L0d and L0k, represent 54% of the haplogroups observed in the Tswana-speaking cohort, whereas the Bantu lineages represent 46% of the haplogroups observed in the Tswana-speaking cohort of this investigation, indicating major population admixture between the Bantu and Khoi-San populations. The sharing of the L0 haplogroup with the Khoi-San population is evidence of gene flow between the two populations. The Bantu-associated marker, i.e. the 9 bp deletion within the CoII/tRNA Lysine gene region (Soodyall *et al.*, 1996), is detected in all of the L0a Tswana haplogroups, indicating the Bantu contribution to the gene pool of the Tswana population. This distribution of haplogroups is not unique to South Africa, as is seen in the data of Soodyall *et al.* (1996), where it was reported that the southern African Xhosa and Zulu populations displayed between 25% and 50% of the Khoi-San lineages, respectively. Further proof of population admixture in South Africa is observed in the presence of about 24% Bantu-specific lineages in the Vasikela !Kung and the phenotypic similarities of the Kwe with the Bantu populations (Chen *et al.*, 2000).

The Bantu lineages in the Tswana population consist of the haplogroups L0 (L0a'b, L0a), L1 (L1b1, L1c2a), L2 (L2a, L2a1) and L3 (L3a3a, L3b, L3e1, L3d1a1) and are representative of Bantu expansion from all regions within Africa (west, central and east) with the emphasis on central and western African origin. This finding is supported by the fact that 75% of the southeastern Bantu lineages can be linked to west/central origins based on linguistic and archeological evidence (Salas *et al.*, 2002). In addition to major

expansion events, the presence of minor haplogroup types in southern Africa can be explained by more recent gene flow or because the founder effects in the migrating Bantu populations were insufficient to eliminate all haplogroups. Other scenarios could also fit this genetic landscape, such as bottlenecks that left no progeny except a specific haplogroup that started to expand and develop into a modern-day haplogroup lineage, as is evident from the haplogroup L1b lineages (Salas *et al.*, 2002; Kivisild *et al.*, 2004). The haplogroup L1c lineages provide evidence of a different evolutionary history that reflect migration and subsequent gene flow between the agricultural Bantu speakers that came into contact with their hunter-gatherer neighbours, the Pygmies, and in current populations carry the evidence of past gene flow between these populations (Quintana-Murci *et al.*, 2008). Haplogroups L2a and L3e are most probably due to major Bantu-speaking population expansions during the sub-Saharan agricultural spread to the southern regions of Africa (Pereira *et al.*, 2001; Salas *et al.*, 2002; Atkinson *et al*., 2009). Many different evolutionary forces played a role in creating the current genetic landscape in South Africa and in the Tswana-speaking population under investigation.

6.8 PHYLOGENETIC ANALYSES

Phylogenetic analysis is widely accepted and a popular methodology used in the study of the evolutionary relationships within and between populations of organisms (Whelan *et al.*, 2001). It is only through the use of accurate models of evolution and statistical tests that information can be obtained about the evolutionary past of a current population of individuals. The coding regions of the mtDNA genomes of 50 Tswana-speaking individuals residing in South Africa were subjected to phylogenetic analyses to establish the maternal lineages and phylogenetic positioning of this population in the context of other African population sequence variation, as contained in the Global African and All African datasets of this investigation.

In line with the lineage approach that was followed for the study of mtDNA sequence variation in this investigation, phylogenetic analyses were used to provide a background to which the haplogroups of the Tswana-speaking individuals under investigation could be positioned relative to other African haplogroups and therefore provide information about the maternal genealogy for the Tswana cohort. Phylogenetic trees were constructed to position the sequence variation observed in the Tswana-speaking individuals of this investigation together with the sequence variation published for a broad geographic group of African individuals as contained in the Global African dataset, as well as a more

confined group of African individuals currently residing in Africa, as contained in the All African dataset discussed in Section 5.11.2. Phylogenetic trees were also constructed of only the mtDNA sequences of the Tswana population of this investigation to determine the positioning of the clades and lineages within the population under investigation.

To prevent the phylogeny from being affected by homoplasy and mutational hotspots, only coding regions of the mtDNA genomes were used for the construction of the phylogenetic trees in the current investigation. In light of previously published reports where haplogroups could not be fully resolved when using only control region site variations (Van Oven and Kayser, 2009), the use of coding region site variation for haplogroup assignment and phylogenetic analyses in this investigation was justified and assisted in providing a highly resolved haplogroup assignment and phylogeny for purposes of studying the evolutionary past of the Tswana cohort under investigation.

Phylogenetic trees were constructed by using three different datasets as discussed in Section 15.11. Dataset 1a consists of 573 mtDNA-coding regions and has been designed to represent individuals of African origin residing both in Africa as well as residing on other continents and is referred to as the Global African dataset. The construction of the dataset is discussed in Section 5.11 and a list of all the mtDNA coding region sequences used in this dataset is described in Appendix B. Dataset 2a consists of the coding regions of the mtDNA of individuals of African origin only residing on the African continent and consists of the coding regions of 390 mtDNA sequences. This dataset is referred to as the All African dataset and is discussed in Section 15.11.2 and described in Appendix C. Dataset 3b is referred to as the Tswana dataset and has been used to represent a Tswana-speaking population of South Africa. This dataset consists of the coding region of mtDNA sequences of the 50 Tswana-speaking individuals investigated in this study and is described in Appendix A.

NJ and MP trees were constructed with the datasets 1a, 2a and 3b respectively, as described in Section 5.13. Two different tree-building approaches with different underlying principles were used for the phylogenetic analyses to validate the phylogenetic relationships identified by this process. The purpose of molecular phylogenetics is to determine a single accurate tree representing the evolutionary history of the cohort of sequences being investigated. Therefore it would be necessary to reject as many trees as possible to end up with the least number of optimal trees, preferably a single, precise tree that is also accurate (Jobling *et al*., 2004). One way of achieving this is to combine tree-

building methods that are based on the principle of stepwise clustering of sequences (NJ) with tree-building methods that search through optimality criteria and select the best fit tree from a set of possibilities (MP). Thus, similar results were expected and regarded as validation of the trees constructed by both methods, given that the datasets used as input were also valid.

Specific parameters were applied in the construction of the trees to ensure that the observed sequence variation reflected the actual evolutionary distances. This was achieved by using the F84 model as discussed in Section 5.13.3.1, which makes provision for the differences in transition and transversion rates and differences in base frequencies. The transition:transversion ratios (Ts/Tv) for the respective datasets were determined by using a subset of 50 samples from each of the datasets 1a and 2a, as discussed in Section 5.11, which in turn were named the Global African dataset 1b and the All Africa dataset 2b respectively. The Tswana dataset 3b consisted of 50 samples and was used for the Ts/Tv ratio calculation. The PHYLIP version 3.6 software was used for the purpose of calculating the Ts/Tv ratio by using the DNA parsimony (Dnapars) program. A Ts/Tv value of 8.566 was obtained for the Global African dataset 1b, while a value of 8.758 was obtained for the All Africa dataset 2b and a value of 7.123 for the Tswana dataset 3b. The Ts/Tv values of the different datasets were used as parameters of the Ts/Tv ratio expected for human populations and thus did not differ significantly (Pereira *et al.*, 2009). The α and gamma values used for the construction of the phylogenetic trees of the datasets of this investigation are displayed in Table 6.35.

Table 6.35 Gamma shaped parameter values for datasets of this investigation

Dataset	Alpha value (α)	Gamma value
Global African dataset (1a)[a]	0.168	2.442
All African dataset (2a)[a]	0.153	2.555
Tswana dataset (3b)[a]	0.03	5.774

a = datasets described in Section 5.11. α value = gamma-shaped parameter. Gamma value as determined by **Error! Reference source not found.**.

The α value shape parameter is small (<1) when a sample of sequences contains mostly invariable nucleotide sites and sequence variation is mainly observed at certain nucleotide sites, such as certain codon positions or within non-coding DNA regions, as is observed in human populations (Gu and Zhang, 1997) and also for the Global African, All African and Tswana datasets of this investigation. The lower α value and subsequently higher gamma value determined for the Tswana dataset when compared to the Global African and All

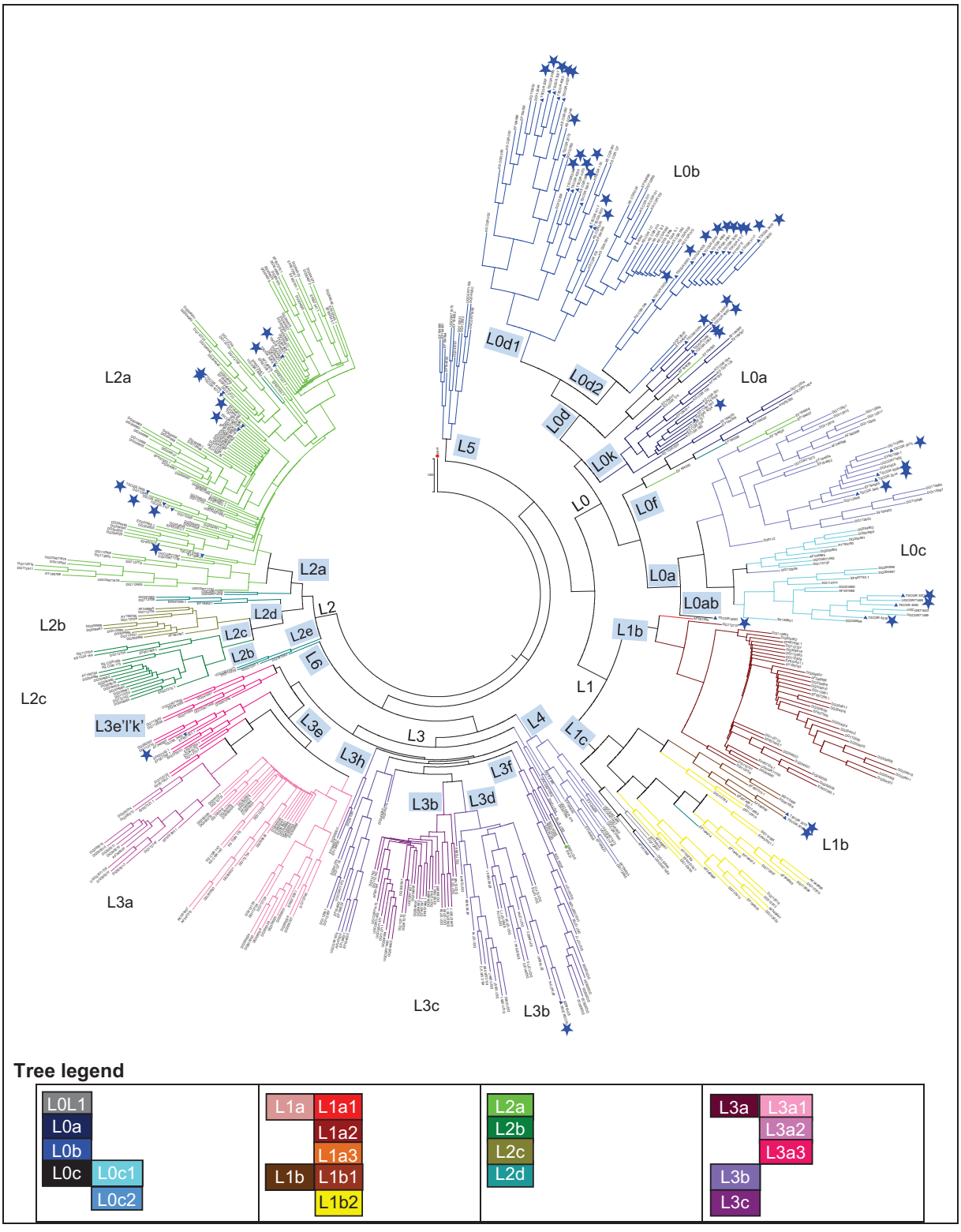
African datasets was expected, based on a reported downward bias observed in smaller sequence datasets (Gu and Zhang, 1997).

Bootstrapping was applied by using 1,000 iterations using random seed numbers for both of the tree-building methods to determine the confidence of branch order determinations. On completion of tree construction, the NJ and MP trees were presented as consensus trees by using the majority rule extended and the majority rule programs, respectively, to provide trees that were representative of the majority of sequence variation that was observed in the datasets.

6.8.1 NJ tree of the Global African dataset

The NJ consensus outtree of the Global African dataset 1a generated by PHYLIP 3.6 was converted to a phylogram format rooted by the outgroup *Pan troglodytes* sequence by using TreeView (Page and Holmes, 1996). PHYLIP consensus program outputs the bootstrap number of each node in the NJ tree as the lengths of the tree branches. Thereafter it was viewed by using MEGA version 5 (Tamura *et al.*, 2011) Tree Explorer in a circular tree format with the bootstrap values indicated on the branches of the tree. The numbering on the branches of the tree in MEGA are therefore not representative of the genetic distances between taxa, but representative of the amount of confidence that can be placed in the branches. The tree is presented in Appendix E in an A3 format with bootstrap values included and in Figure 6.72 without bootstrap values.

Figure 6.72 NJ tree of Global African dataset



This NJ tree represents the coding regions of 573 mtDNA sequences of the Global African dataset 1a as described in Section 5.11.1 according to the tree-construction method described in Section 6.8. The rCRS is indicated with a green diamond (◆) and the *Pan troglodytes* outgroup is represented by a red circle (●). The haplogroups and sub-haplogroups assigned by the Wallace classification system (2004) are presented in colours as described in the legend. Major haplogroup branching is indicated at the branch splits on the tree and sub-haplogroup groupings are indicated on the outside of the tree. The PhyloTree (Van Oven and Kayser, 2009) haplogroup assignments are indicated in blue text boxes. MtDNA sequence identification is presented at the tips of the branches and the Tswana mtDNA sequences of this investigation are indicated by blue triangles (▲) and blue stars (★).

The Global African NJ tree consist of coding regions of 573 mtDNA sequences that are representative of individuals of African origin residing on the African continent as well as individuals of African origin residing outside the African continent. The construction of an NJ tree with mtDNA sequences of a broad population of African individuals presents the phylogenetic relationships of mtDNA sequence variation between different African populations with different ancestral origins. This has assisted in determining the positioning of the Tswana population of this investigation in the context of other African populations within and outside the African continent. The haplogroups according to the Wallace classification system (2004) have been incorporated in Figure 6.72 to indicate the positioning of the haplogroups within the tree. The major haplogroups (L0, L1, L2 and L3) are indicated at the branch splits and the sub-haplogroups are indicated on the outside of the tree. The haplogroups are further presented in colours as described in the figure legend for ease of interpretation. In addition, the haplogroups and sub-groups of all the mtDNA sequences of this dataset have been compared to the published haplogroup assignments as reported by Pereira *et al.* (2009) of each of the mtDNA sequences where possible and confirmed by assigning haplogroups and sub-groups with PhyloTree (Van Oven and Kayser, 2009) as described in Section 6.7. This comprehensive haplogroup assignment is presented in Appendix B.

Major haplogroups L4, L5 and L6 cannot be classified by the Wallace classification system (2004) and have therefore been assigned by using PhyloTree (Van Oven and Kayser, 2009) and published haplogroup assignments (Pereira *et al.*, 2009). The mtDNA sequences assigned to these haplogroups are listed in Table 6.36 and because of the discrepancy in haplogroup assignment between the Wallace classification system (2004), the published haplogroup assignment and PhyloTree (Van Oven and Kayser, 2009), the location of these haplogroups has been identified and indicated in order to interpret the basal topology of the NJ tree.

Table 6.36 Haplogroups L4, 5 and 6

GenBank® accession number	GI number	Sequence name/isolate	Ethnicity/geographic origin	Haplogroup/sub-haplogroup classification		
				Published[a]	Wallace, 2004[b]	PhyloTree[c]
DQ112796	70955926	AF07	Lissongo/Africa	L4b2b	L3b$	L4b
DQ112845	70956024	AF18	Khoi-San/Africa	L4b2a2	L3b$	L4b2
DQ341064	84682418	Tor66	Ethiopia	L4a1	L3b$	L4a1
DQ341065	84682432	Tor71	Ethiopia	L4b2a1	L3b$	L4b2
---	---	UG_CGR_71472	Acholi	---	L3b$	L4b2
---	---	UG_CGR_71474	Acholi	---	L3b$	L4a

Table 6.36 Continued…

GenBank® accession number	GI number	Sequence name/isolate	Ethnicity/geographic origin	Haplogroup/sub-haplogroup classification		
				Published[a]	Wallace, 2004[b]	PhyloTree[c]
---	---	UG_CGR_71493	Baganda	---	L3b$	L4a
---	---	UG_CGR_71496	Baganda	---	L3b$	L4b
---	---	UG_CGR_71657	Baganda	---	L3b$	L4b
---	---	UG_CGR_71660	Baganda	---	L3b$	L4b2
---	---	UG_CGR_71708	Lugbara	---	L0L1$	L4b2
DQ112921	70956176	AF99	Pygmy/Zaire	L5a1c	L0$a	L5a1
DQ112922	70956178	AF100	Pygmy/Zaire	L5a1c	L0$a	L5a1
DQ341060	84682362	Tor74	Ethiopia	L5a1a	L0$a	L5a
DQ341061	84682376	Tor68	Ethiopia	L5c1	L0$a	L5
EF184580	133854527	TZSW053_L5	Sandawe/Tanzania	L5	L0$b	L5
EF184581	133854541	TZSW087_L5	Sandawe/Tanzania	L5	L0$b	L5
EF184582	133854555	TZMW009_L5	Mbugwe/Tanzania	L5	L0$a	L5c
EF184583	133854569	TZSW086_L5	Sandawe/Tanzania	L5	L0$b	L5
EF184584	133854583	TZSW055_L5	Sandawe/Tanzania	L5	L0$b	L5
---	---	UG_CGR_71696	Lugbara	---	L0a$	L5a1
---	---	UG_CGR_71709	Lugbara	---	L0L1$	L5
DQ341063	85541074	Tor39	Ethiopia	L6b	L2d$	L6b

GI = GenInfo Identifier sequence identification number used for identification in GenBank®. Where sequences from other unpublished studies were used, the GenBank® identifiers isolate names and published haplogroup assignments are not relevant and indicated by ---. Isolates with the prefix UG_CGR are from a study performed by Isabirye (2010). a = Published haplogroups were obtained from Pereira *et al.* (2009). b = Wallace classification system (2004). c = PhyloTree classification system (Van Oven and Kayser, 2009). $ = MtDNA sequence haplogroup assignment that could not be fully resolved.

The NJ tree represents major haplogroup L2 as the most basal haplogroup of the tree. From there the tree branches into two major groups. The one group consists of two branches, of which one branch consists of only one mtDNA sequence (DQ341063) from an individual with ethnic origin in Ethiopia. DQ341063 has been assigned to haplogroup L2d and cannotfully resolved when using the Wallace classification system (2004). The positioning of an mtDNA sequence belonging to haplogroup L2 at this position in the tree, separate from the other L2 haplogroups, is problematic. It indicates that this sequence shares a common ancestor with the L3 haplogroup individuals and not with the L2 haplogroup individuals. This aspect is resolved when using the PhyloTree haplogroup assignment, which is in accordance with the published assignment, namely haplotype L6b. Haplogroup L6 is only represented by one mtDNA sequence in this dataset, most probably because it is very rare or because the sampling strategies used for the sequences in this dataset bias haplogroups other than L6. Haplogroup L6 demonstrates low haplogroup diversity and is present at high frequencies in Yemenis, at low frequencies in Ethiopians and has not been reported elsewhere in Africa, which provides reason to speculate that it originated from a small founding population, which might not have been sampled for phylogenetic studies yet (Kivisild *et al.*, 2004).

The other branch of this group has bifurcated into two sub-branches. The one sub-branch contains the mtDNA sequences belonging to haplogroup L3 and the other branch contains a group of mtDNA sequences belonging to haplogroup L3b that cannot be fully resolved when using the Wallace classification system (2004). When using the PhyloTree (Van Oven and Kayser, 2009) and published haplogroup assignments, it is determined that these sequences belong to haplogroup L4, as depicted in Table 6.36. It has been concluded that this clade represents haplogroup L4 and shares a common ancestor with haplogroup L3. This is in accordance with the PhyloTree classification system (Van Oven and Kayser, 2009), which indicates np 3594, np 7256 and np 13650 as the common haplogroup root-defining sites shared between haplogroup L3 and L4. Besides the sequences listed in Table 6.36, the rest of the group consists of mtDNA sequences EF184629 and EF184627, both assigned to haplotype L3g and both individuals originating from Tanzania, and EF184633, EF184639 and EF184640, which have all been assigned to haplogroup L3 and of which two individuals originate from Tanzania and one Khoi-San individual from South Africa. These sequences could not be resolved into sub-groups of L3, indicating the lack of common ancestors with the other L3 sub-groups and therefore are most probably positioned correctly in the tree. The L3g haplotype has been described as the sister clade of L4a because of shared ancestral character states at np 769 and np 1018 (Kivisild *et al.*, 2004), as can be seen in the positioning within the tree.

The second major grouping present in the NJ tree consists of two branches. The one branch consists of two separate groups representative of the haplogroups L1 and L0 respectively, indicating that a common ancestor is shared. Studies have reported that the mtDNA phylogeny of major haplogroup L consists of two major branches, L0 and L1’2’3’4’5’6, located on opposite sides of the root and not on the same branch from the root as displayed in the Global African NJ tree (Ingman *et al.*, 2000; Mishmar *et al.*, 2003; Kivisild *et al.*, 2006; Gonder *et al.*, 2007; Behar *et al.*, 2008). The phylogenetic hierarchy in the Global African NJ tree, in which the L2 haplogroup is positioned as the most ancient lineage and haplogroups L0 and L1 are positioned as sharing ancestry, differs when compared to previously published views in which haplogroup L0 is widely regarded as the most ancient lineage that does not share ancestry with haplogroup L1 (Kivisild *et al.*, 2006; Torroni *et al.*, 2006; Gonder *et al.*, 2007; Behar *et al.*, 2008). Although the possibility exists that a new phylogenetic paradigm has been discovered in this investigation, it is highly unlikely, in light of the number of published reports that state otherwise and suggest it not to be the case. A more likely reason for this observation could be ascribed to the method of NJ tree construction in which sequences are grouped together based on genetic

distances only and not on the actual discrete sequence data. In this context, the Global African NJ tree has grouped the haplogroups and distinguished correctly between major haplogroups. The NJ tree construction method relies on clustering sequences in a stepwise manner without being able to determine how well the tree fits the data and therefore the tree topology will be dependent on the order in which the sequences have been processed (Page and Holmes, 1998). For this reason, NJ trees can demonstrate incorrect tree topology. Another possible explanation may be the greedy search method used by the NJ tree-building method, which could have missed the global optimal tree because it works on a principle of local exploration and therefore explores only a small part of the set of possible phylogenies for the given set of mtDNA sequences. In addition, only a single tree topology has been reported for the data and as is typical of greedy search methods, no other alternative tree topologies have been reported from which the most optimal tree topology could be selected, as is the case with MP tree-building methods (Pearson *et al.*, 1999).

Other reasons for incorrect tree topology could be related to the analysis of the dataset used in the construction of the NJ tree. The introduction of large numbers of substitutions to the sequence set owing to the misalignment of the mtDNA sequences could have caused the appearance of incorrect clades in the NJ tree. This is highly unlikely, since the dataset of this investigation was aligned by CLUSTAL (Thompson *et al.*, 1994) that would have identified such a problem. Phylogenetic error could also have been caused by outspoken base bias (Sanderson and Shaffer, 2002). This too has been ruled out as a possible reason for the observed NJ tree topology of this investigation, because no such bias was observed in the mtDNA of the coding regions used in this investigation.

A critical parameter for the construction of the correct topology of the NJ tree-building method is the requirement of a reliable measure of evolutionary distances between sequences. The use of different protein, gene or non-coding sequences may result in different phylogenetic tree reconstructions (Brocchieri, 2001). Reverse mutations, different mutational rates for sequence regions or sequence positions, functional constraints on sequence sites and mutational saturation affect the determination of evolutionary distances between sequences (Brocchieri, 2001). The NJ trees of this investigation were constructed by using only the coding regions of the mtDNA sequence data. The sequence variation present in the control regions of the samples in the dataset of this investigation was not taken into account in the determination of genetic distances between mtDNA sequence pairs for purposes of NJ tree construction. The control region of the mtDNA is

reported to display a mutation rate that is ten times higher than the mutation rate in the coding region and carries a major part of the sequence variation within the full genome of the mtDNA sequence. The control region also displays a high rate of reverse mutations and to prevent the evolutionary history and therefore the true genetic distances from being obscured, it was decided to exclude this region from the phylogenetic analysis in this investigation. Each region of the mtDNA genome or genes provides a specific set of sequence variation that demonstrates different substitution and mutation rates and provides different levels of phylogenetic information (Cummings and Meyer, 2005). This exclusion unfortunately also means the exclusion of a major component of sequence variation from the distance determinations in the tree constructions, which could possibly have been the reason why the tree topology differed from other studies. The studies depicting the L0 and L1-6 haplogroups on separate branches and the haplogroup L0 as the most ancient grouping (Kivisild *et al.*, 2006; Torroni *et al.*, 2006; Gonder *et al.*, 2007; Behar *et al.*, 2008) all use full mtDNA sequence data for the construction of the phylogenetic trees and/or networks. In addition to this, their datasets consist of a broad set of haplogroups representative of more than the L haplogroup and in the case of the Behar *et al.* (2008) study that focused on the L haplogroup only, care was taken to include rare variants of all the known L haplogroups for the purpose of creating a phylogenetic tree that would be based on evolutionary distances representative of the full L haplogroup. Although the dataset of this investigation also included a broad range of residents from African and non-African countries, it was not compiled specifically to include rare variants of the L haplogroup and could have been more limited in terms of the genetic distances portrayed. It was concluded that these factors, in addition to the NJ tree construction methodology, may have contributed to the tree topology of the NJ trees of this investigation.

Bootstrap values, determined by 1,000 iterations, were used to test the reliability of the phylogenetic internal branches of the trees constructed in this investigation. Bootstrap values do not allow for the probability that a clade is a true clade, but only the probability that a particular phylogeny would be obtained upon resampling, thereby indicating how well a tree-building method supports a clade. For this reason, bootstrap values must be interpreted by taking into account variables that can influence the probability that a group represents a true clade. These include the number of characters used, the number of taxa, the evolutionary rate of change, mutational rate among lineages, independence of the characters and the type of tree-building method used (Hillis and Bull, 1993). Bootstrap values have generally proven to be an underestimate of the accuracy of the phylogenetic tree branches under a wide range of these variables (Hillis and Bull, 1993) and it can be

assumed that this will be true for the NJ tree of this investigation. Bootstrap values of 50% are acceptable as sufficient confidence in the validity of a branch or clade (Hillis and Bull, 1993). The bootstrap values obtained for the major haplogroup branches in the Global African NJ tree are listed in Table 6.37.

Table 6.37 Bootstrap values for the Global African NJ tree of this investigation

Haplogroups	Global African NJ tree
L0'1'5	64%
L0	91%
L1	98%
L2	66%
L3'4'6	44%
L3	32%
L4	48% and 30%
L5	67%
L6	100%

L0'1'5 and L3'4'6 refer to root-defining branches that branch into the respective haplogroups. The bootstrap values are listed as percentages by calculating them from the values listed as branch lengths in the MEGA 5 Tree Explorer.

All of the major haplogroups display bootstrap values that indicate confidence in the positioning of the branches except for haplogroups L3 and L4 and the root-defining branch L3'4'6, which display bootstrap values of less than 50% and therefore those branches cannot be accepted with confidence. The low bootstrap values by implication reflect negatively on the method of tree construction, which could not support the phylogeny after resampling. Sub-groups of the major haplogroups demonstrate somewhat higher bootstrap values in correspondence to their positioning and are in accordance with published data.

The other branch of this major group consists of mtDNA sequences that have been assigned to haplogroup L5 by the PhyloTree classification system (Van Oven and Kayser, 2009). This clade only consists of 12 mtDNA sequences, of which 11 have been assigned to haplogroup L5 and are listed in Table 6.36 and one mtDNA sequence, UG_CGR_71675, which has been assigned to haplogroup L3 and belongs to an individual of Lugbara ethnicity. This individual cannot be assigned to a fully resolved haplotype, indicating that the sample contains haplogroup-defining sequence variations that are not shared with the L3 sub-groups. It could therefore be possible that this individual contains a unique haplogroup-defining sequence variation that has not been incorporated into the current haplogroup classification systems for haplogroup L5. All of the individuals belonging to haplogroup L5 were of east African origin and corresponded to the reported geographic location of the L5 haplogroup in Africa (Batini *et al.*, 2011). The L5 haplogroup

was previously characterised as the L1e haplotype (Salas *et al.*, 2002) and has been reported to occupy an intermediate position between haplogroup L1 and L2'3' (Shen *et al.*, 2004). In the Global African NJ tree this clade is more basal than L0, which may be indicative of a very ancient nature of the haplogroup L5, which, however, must be regarded with caution in view of the low number of L5 haplotype samples used in the construction of this tree. The low number of mtDNA sequences in the Global African dataset of this investigation belonging to this haplogroup could be ascribed to the fact that it is a rare haplogroup or that it has not been sampled in the cohorts that were used for the construction of the Global dataset undertaken in this investigation.

The L0 haplogroup is split into two main groups on the Global African NJ tree. The first group consists of mtDNA sequences assigned to haplogroups L0a and L0b by the Wallace classification system (2004) and haplogroups L0k and L0d when using the PhyloTree classification system (Van Oven and Kayser, 2009). The second group consists of mtDNA sequences assigned to haplogroup L0c by the Wallace classification system (2004) and haplogroups L0a and L0f by the PhyloTree classification system (Van Oven and Kayser, 2009). The latter haplogroup assignment by PhyloTree corresponds to what has been reported by other studies (Salas *et al.,* 2002; Mishmar *et al.*, 2003, Salas *et al.*, 2004; Kivisild *et al.*, 2004), in which it is stated that the L0 haplogroup consists of the four sub-haplogroups L0a, L0d, L0f and L0k. On further investigation, when compared to the haplogroups published for these mtDNA sequences, it has been determined that it corresponds to the haplogroup assignments by PhyloTree (Van Oven and Kayser, 2009) and will, for the purposes of this discussion, be used because the PhyloTree classification system (Van Oven and Kayser, 2009) has been validated by the confirmation of haplogroup assignment in literature.

The L0d haplogroup has been reported to lie at the root of the phylogenetic tree of human mtDNA (Chen *et al.*, 1995a; Ingman *et al.*, 2000; Mishmar *et al.*, 2003; Kivisild *et al.*, 2004; Gonder *et al.*, 2007) and lies at the root of the L0 branch of the Global NJ tree of this investigation, suggesting an ancient and independent development from the other haplogroup of the L0 clade. The branch containing the mtDNA sequences assigned to haplogroup L0d has been split into three sub-groups, namely L0d1, L0d2 and L0d3. The group belonging to haplogroup L0d1 contains 12 of the mtDNA sequences of the Tswana population from this investigation. The group further contains 29 mtDNA sequences belonging to Khoi-San individuals of South African origin, three mtDNA sequences belonging to Khoi-San individuals of African origin, one mtDNA sequence belonging to an

individual of Sandawe Tanzanian origin and an mtDNA sequence from a Tswana-speaking individual and one from an individual of Zulu-speaking origin, both of southern African descent. The L0d2 clade consists of ten mtDNA sequences of the Tswana-speaking individuals of this investigation, three mtDNA sequences from Khoi-San individuals of South African descent and an mtDNA sequence from a Tswana-speaking individual and one from a Zulu-speaking individual of southern African origin. The L0d3 clade consists of one sub-clade containing mtDNA sequences of five Tswana-speaking individuals of this investigation and another sub-clade consists of four mtDNA sequences of individuals of Tanzanian origin.

The mtDNA sequence EF184589 is assigned to haplogroup L2a by the Wallace classification system (2004) but it does not fit in this group in the current phylogenetic tree and this discrepancy has been indicated by a different colour (green) than it would have been if it had been part of the other mtDNA sequences in the L2a clade (blue). This sequence was reported as belonging to the L0d haplogroup and this assignment was confirmed by the definition of the sequence as being haplogroup L0d3 by the PhyloTree classification system (Van Oven and Kayser, 2009). All of the other mtDNA sequences of that clade were also assigned to haplogroup L0d3 by the PhyloTree classification system and its grouping into one clade, reflecting a shared ancestral state, was therefore expected. All of the individuals of this clade originated from Tanzania, with one individual belonging to the Burunge ethnic group and the other three belonging to the Sandawe ethnic group. This grouping highlights the subdivision of L0d into separate clades as mtDNA sequences belonging to individuals of Tanzanian origin (Sandawe and Burunge ethnic origin) and Khoi-San individuals of southern African origin are present (Gonder *et al.*, 2007).

This L0d branch of the Global NJ tree consists of 27 mtDNA sequences from the Tswana-speaking individuals of this investigation. Therefore 54% of the Tswana population undergoing investigation has been classified into this group. This is therefore a substantial number of the total Tswana population. These mtDNA sequences group with 33 mtDNA sequences from Khoi-San individuals of southern African origin, two mtDNA sequences of Tswana-speaking individuals, two mtDNA sequences from Zulu-speaking individuals of other studies and five mtDNA sequences from individuals from Tanzania, of which four group into a single clade. It was expected that the Khoi-San individuals would group together, since the L0d haplogroups have only been reported among the Khoi-San of southern African origin (Gonder *et al.*, 2007). The presence of the Tanzanian individuals

in this grouping was also expected, since it has been reported that they share the ancient L0d haplogroup with the Khoi-San individuals, albeit in different clades, as observed in this study (Gonder *et al.*, 2007). From the phylogenetic grouping in this tree it is clear that the Tswana-speaking individuals of this investigation share a great deal of ancestral sequence variation with the Khoi-San individuals of southern African origin. The grouping of four Bantu individuals of southern African origin, two Tswana-speaking individuals (DQ112855, DQ112850) and two Zulu-speaking individuals (DQ112854, DQ112830) in this haplogroup indicates that the ancestry of the Bantu-speaking individuals of southern Africa with the Khoi-San lineages that has been reported in previous studies, is confirmed in this study by a significant number of Tswana individuals that group together in this clade.

The other major group of this L0 haplogroup branch consists of mtDNA sequences assigned to haplogroup L0k. It does not form a clade with L0f and L0a as was previously reported (Gonder *et al.*, 2007) and demonstrates common ancestry with the L0d lineage. This haplogroup has also been connected to the Khoi-San lineages with L0d and L0k, together contributing to 60% of the sequence variation reported in the modern Khoi-San populations of southern Africa (Tishkoff *et al.*, 2007). Higher frequency and increased internal sequence variation within the Khoi-San individuals of a study performed by Behar *et al.* (2008), compared to non-Khoi-San individuals belonging to haplogroup L0k, suggest that the Khoi-San population was is the source of the haplogroup and that non-Khoi-San individuals most probably acquired the haplogroup by gene flow through population admixture. This is a probable case for the L0k haplogroup of the Tswana individual of this investigation.

The second major branch of the L0 haplogroup of the Global African NJ tree consists of mtDNA sequences assigned to the haplogroup L0c by the Wallace classification system (2004) and assigned to haplogroups L0a1, L0a2 and L0f by the PhyloTree classification system (Van Oven and Kayser, 2009). This major branch consists of the L0f and L0a clades, which demonstrate common ancestry.

The L0f clade consists of eight mtDNA sequences of which five are of Tanzanian origin, one is from Ethiopia and two are from Uganda. All of these sequences therefore demonstrate eastern African origin. Two mtDNA sequences (EF184595 and EF184596) in this clade are assigned to haplogroup L2a and L2d respectively by the Wallace classification system and are therefore not positioned correctly in this grouping, as indicated by the different branch colour (green) in comparison to the branch colour of the

other mtDNA sequences (purple) of that clade. On further investigation it was determined that these two sequences, which belonged to individuals of Wafiome and Akie ethnicity from Tanzania, had been assigned to L0f in the literature and this was confirmed by assigning the sequences to L0f2 and L0f respectively when using the PhyloTree classification system (Van Oven and Kayser, 2009).

Alternatively, the L0a clade could be divided into two sub-clades, namely L0a1 and L0a2, and one mtDNA sequence that forms a deep basal branch to the root of the L0a clade. This mtDNA sequence belongesto the Tswana-speaking individual, TS_5063, and could not be fully resolved by either of the haplogroup classification systems used in this investigation. The sequence is assigned to haplogroup L0c by the Wallace classification system (2004) and to L0a'b' with the PhyloTree classification system (Van Oven and Kayser, 2009). This individual lacks the haplogroup-defining transition at np G5147A to be assigned to haplogroup L0c2 with the Wallace classification system (2004) and the transitions at np G5231A, np G5460A and np T14308C to define it as haplogroup L0a with PhyloTree (Van Oven and Kayser, 2009). Haplogroup-defining sites that are only preliminary and need further investigation to be confirmed as valid, have been indicated in the PhyloTree classification system These preliminary haplogroup defining sites were not used in this investigation because of their unconfirmed status, and TS_5063 could therefore not be determined to belong to haplogroup L0b. It is not likely that the mutations at these four positions could all be obscured by reverse mutations and the possibility of a new haplogroup should be considered, especially since it is noted in PhyloTree (Van Oven and Kayser, 2009) that the L0b haplogroup-defining mutations need further resolution. The positioning of the sequence in the Global African NJ tree as a separate branch to the L0a haplogroups highlights this possibility.

The L0a1 and L0a2 sub-clades contain three and four mtDNA sequences of Tswana-speaking individuals of this investigation, respectively. The L0a1 sub-clade is represented by a combination of individuals from west, central and east Africa, as well as mtDNA sequences from individuals of African-American origin. This indicates the shared ancestry that African-Americans have with the Africans from which they originated during the slave trade, which were mostly of eastern African origin (Pereira *et al.*, 2001). There is therefore a small component of the Tswana-speaking individuals in this investigation that demonstrates shared ancestry with the Bantu populations of many different regions of Africa. The L0a2 sub-clade is representative of a haplogroup that is common in Africa (Salas *et al.*, 2002) and is associated with the 9 bp deletion in the COII/tRNA^{Lys} intergenic

region. All of the Tswana-speaking individuals in this investigation that are grouped in this sub-clade display the deletion and their ancestry is represented by the Bantu migration from western Africa to eastern and southern Africa. In total, 14% of the Tswana-speaking individuals of this investigation demonstrate shared ancestry with the haplogroup L0a. A discrepancy between the haplogroup and its position in the tree has been noticed in this sub-clade. The mtDNA sequences EF184608 and EF184607, both of Tanzanian origin, are assigned to haplogroup L2a by the Wallace classification system because of the absence of the transition at np T10810C to classify it to the L0, L1 major haplogroup. For this reason it is indicated in another colour (green) than the rest of the clade (blue). The incorrect assignment to haplogroup L2 by the Wallace classification system (2004) could have been due to a reverse mutation at np 10810. When investigated further, it was determined that both mtDNA sequences had been assigned to haplogroup L0a2 by the PhyloTree classification system (Van Oven and Kayser, 2009) and were assigned to haplogroup L0a in the literature.

The L1 haplogroup branch consists of two groups. The first group represents the mtDNA sequences of the Global African dataset that belongs to haplogroup L1b when using the PhyloTree classification system (Van Oven and Kayser, 2009) and L1a when using the Wallace classification system (2004). None of the mtDNA sequences of the Tswana-speaking individuals of this investigation is grouped in this clade.

The second group consists of mtDNA sequences representative of four different haplogroups when using the PhyloTree classification system (Van Oven and Kayser, 2009), namely L1c/ L1c3, L1c1, L1c2 and L1c4, and two haplogroups when using the Wallace classification system (2004), namely L1b and a group that can only be resolved to the level of L1L0. The deepest branch of this group consists of a clade of mtDNA sequences that cannot be fully resolved using either classification systems. Sequences EF184612, DQ112697, DQ112696 and AF381992 demonstrate a discrepancy with their position in the Global African NJ tree when the Wallace (2004) assignment is considered because these sequences are all assigned to L0L1 and are indicated in grey, which differs from the other sequences, which have been assigned to haplogroup L1b. When using the PhyloTree classification system (Van Oven and Kayser, 2009), all of these sequences are assigned to haplogroup L1c and on investigation of the published haplogroups, it was determined that all had been assigned to L1c3. A sequence from Uganda has also also included in this group (UG_CGR_71676) and assigned to L1c3 with the PhyloTree classification (Van Oven and Kayser, 2009). The position of this group coincides with the

positioning of the L1c clade according to the PhyloTree system (Van Oven and Kayser, 2009) by splitting it into L1c3 and L1c1'2'4; the latter then split into L1c1 and L1c2'4. Sequence EF184614 is grouped with mtDNA sequences assigned to haplogroup L1c1 (PhyloTree, 2009) although it is assigned to haplogroup L2d$ by the Wallace classification system (2004). This discrepancy has been clarified by using the published classification and the PhyloTree (Van Oven and Kayser, 2009) assignment of L1c. Two of the mtDNA sequences of the Tswana-speaking individuals of this investigation have been grouped with the clade representing haplogroup L1c2. Both TS_3015 and TS_2093 are assigned to haplogroup L1c2. L1c is often present in the Western Pygmy groups that live in the central African region of the Cameroon, the Republic of the Congo, Gabon and Central African Republic and consists of the Binga, Baka, Biaka and Aka ethnic groups.

Demographic information demonstrates an early split between the maternal ancestors of the Pygmy and Bantu populations, followed by population growth in the Bantu populations before the Bantu migration started. The presence of this haplogroup in the Tswana population under investigation could be a result of gene flow between the Tswana-speaking population of southern Africa and the modern-day Pygmies or of more ancient gene flow between the Bantu and Pygmy populations prior to the migration to southern Africa. The latter is more likely and it is postulated that the L1c haplogroup was contributed to the Tswana population by the haplogroups and sub-groups carried by the historic Bantu migration to southeastern Africa. This haplogroup is present in only 4% of the Tswana population investigated in this study and therefore does not suggest a major contribution.

The L2 haplogroup was divided into two major groups of which one group consisted of only three mtDNA sequences assigned to haplogroup L2e by the PhyloTree classification system (Van Oven and Kayser, 2009) and to haplogroup L2d by the Wallace classification system (2004). These individuals originated from an unidentified region within Africa, the Dominican Republic and Uganda respectively. This clade only shared common ancestry with the rest of the L2 haplogroups at the root of the haplogroup. The second major group represented the rest of the L2 haplogroup and branched into two groups. One group consisted of the mtDNA sequences belonging to haplogroup L2a as classified by both classification systems used in this investigation, and the other group bifurcated into the L2b'c and L2d haplogroups. The L2b'c branch bifurcated into the L2b and L2c haplogroups as classified by both classification systems. Therefore, according to the Global African NJ tree, the L2b and L2c haplogroups share a common ancestor and developed

independently from the L2a and L2e haplogroups. The L2a clade, L2b and L2c clades and L2d clade share a common ancestor that is not shared with the L2e haplogroup. The PhyloTree classification system (Van Oven and Kayser, 2009) and Salas *et al.* (2002) describe the hierarchy in the L2 haplogroup in the same way (Salas *et al.*, 2002).

The Tswana-speaking individuals of this investigation are only grouped in the L2a1 clade of the Global African NJ tree and consist of ten Tswana-speaking individuals (20%) from this investigation. The L2a clade could further be subdivided into three sub-clades consisting of mtDNA sequences assigned to haplogroups L2a1, L2a2 and L2a4 by the PhyloTree classification system (Van Oven and Kayser, 2009). The Eastern Pygmies often display the L2a2 and L2a4 haplogroups. Haplogroup L2a has been identified as a genetic marker of the Bantu expansions and is common to the Bantu populations of Africa (Quintana-Murci *et al.*, 2008; Coelho *et al.*, 2009). The L2a1 lineage often includes African American and western central African sequences and L2a1a and L2a1b clusters have been observed to be well represented in the southeastern Africans (Batini *et al.*, 2011). The presence of 20% of the Tswana-speaking individuals in this clade is therefore not unexpected. It serves as evidence that although the Tswana population does carry strong evidence of gene flow from the Khoi-San populations in southern Africa, it also harbours the Bantu haplogroups associated with the Bantu, confirming the theory that the modern Tswana-speaking population under investigation consists of a combination of Bantu and Khoi-San genetic components.

A Tswana-speaking individual, TS_5060, groups with one other individual from Tanzania, EF184620, in a clade that branches off at the root of the L2a1 branch. On further investigation, this individual has been assigned to haplogroup L2a1d, defined by transitions at np T5196C, np T9530C, np T11386C, np A12612G and np C13934T and not shared with the other L2a1 mtDNA sequences. TS_3466, TS_3107 and TS_5062 group together with two other mtDNA sequences, a Pedi individual (DQ112849) and a Khoi-San individual (DQ112900), both originating from southern Africa. These sequences display a transition at np G10143A, which is not shared with other mtDNA sequences of the L2a1 clade and are assigned to haplogroup L2a1b by the PhyloTree classification system (Van Oven and Kayser, 2009). The presence of this haplogroup-defining mutation in the Pedi and Khoi-San individuals of the Global African dataset, reaffirms the possibility of gene flow between the early Bantu-speakers and the Khoi-San of southern Africa. The fact that such a significant group of Tswana individuals of this investigation belong to Khoi-San related haplogroups (L0d) is an indication that gene flow between the Tswana and

Khoi-San populations took place, which is also supported by the reported Khoi-San assimilation rate in southeast Bantu speakers of about 5% (Salas *et al.*, 2002). TS_2077, TS_5086 group together with mtDNA sequences that contain the L2a1f haplogroup-defining transition at np A5580G according to the PhyloTree classification system (Van Oven and Kayser, 2009) and group with eight other mtDNA sequences of African American origin and two of Burkina Faso origin. TS_4013 and TS_4106 are assigned to haplogroup L2a1c by the PhyloTree classification system (Van Oven and Kayser, 2009) and are grouped with other mtDNA sequences belonging to the same haplogroup, of which six are of African American origin, four are from Burkina Faso, one is from the Dominican Republic, one of Tanzanian origin and one from Uganda. The mtDNA sequence from Uganda (UGCGR71680) is assigned to haplogroup L2d by the Wallace classification system (2004) and differs from the rest of the sequences in that clade. This discrepancy is clarified when the PhyloTree classification system (Van Oven and Kayser, 2009) is used and haplogroup L2a1c is assigned. These mtDNA sequences all share the haplogroup-defining transitions at np A6663G and np G3010A. TS_2103 and TS_5066 are grouped in a sub-clade with 25 other mtDNA sequences, all belonging to haplogroup L2a1a when using the PhyloTree classification system (Van Oven and Kayser, 2009). Eighteen of the mtDNA sequences in this sub-clade are of African American origin and the other six mtDNA sequences of unknown African origin. The L2a1 clade contains a large component of mtDNA sequences of African American origin with some sequences from the Dominican Republic and Burkina Faso. This observation is supported by the literature, which reports a significant component of non-Africans belonging to haplogroup L2a1, as well as the fact that this haplogroup is regarded as a footprint of the Bantu migration to the southern and eastern parts of Africa (Salas *et al.*, 2002; Gonder *et al.*, 2007; Batini *et al.*, 2011).

Only two of the Tswana-speaking individuals are positioned within the L3 haplogroup of the Global African NJ tree. Haplogroup L3 has its origin in eastern Africa and evidence suggests that the out of Africa migration took place from a source in eastern Africa that gave rise to this haplogroup (Salas *et al.*, 2002). The L3 haplogroup of the Global African NJ tree consists of two major groups. The first group is defined by the PhyloTree classification system (Van Oven and Kayser, 2009) as haplogroup L3e'l'k because of a transition at np A10819G. The mtDNA sequence of TS_3495 groups within this cluster with other mtDNA sequences that are assigned to haplogroup L3a by the Wallace classification system (2004) and to haplogroup L3e1 by the PhyloTree classification

system (Van Oven and Kayser, 2009). The other mtDNA sequences in this clade are from six African individuals of unknown origin and from two non-Africans.

The other major group consists of the L3a, L3b, L3c, L3d, L3f and L3h clades as defined by the PhyloTree classification system (Van Oven and Kayser, 2009), which does not share the np A10819G transition with the first group. The mtDNA sequence of TS_3085 is assigned to haplogroup L3d1a1a by the PhyloTree classification system (Van Oven and Kayser, 2009) and to haplogroup L3b by the Wallace classification system. (2004), and grouped with mtDNA sequences assigned to haplogroup L3d1a1a, which are of Yoruba origin from Burkina Faso and two mtDNA sequences of African origin of unknown ethnicities. Also included in the same sub-clade are mtDNA sequences belonging to haplogroup L3d1d, L3d1c, L3d1b and L3d1b1 from the Dominican Republic, Burkina Faso and other unknown African regions.

Pereira *et al.* (2001) report the L3e haplogroup as a marker of the Bantu dispersal and it is widely regarded as the most common and ancient sub-haplogroup of haplogroup L3 (Salas *et al.*, 2002). Evidence indicates that this haplogroup had its origin in central Africa in the Sudan region and is common in southeastern African Bantu populations (Bandelt *et al.*, 2001b; Kivisild *et al.*, 2004; Coelho *et al.*, 2009). The L3e1 clade especially has been observed in southeastern African Bantu individuals and also in Khwe individuals of southern Africa. The presence of the L3e1 haplogroup in the Khwe has been attributed to gene flow from the Bantu individuals who arrived in southern Africa through the Bantu migrations (Coelho *et al.*, 2009). The introduction of this haplogroup to the Tswana population could therefore have occurred either through direct gene flow with Bantu individuals or indirectly through gene flow from Khoi-San individuals that harboured these haplogroups.

Haplogroup L3d is mainly located in western Africa and has also been reported in non-African individuals. The L3d1 haplogroup, however, has been reported in southeastern African individuals at especially high frequencies in the Fulbe lineage (Salas *et al.*, 2002). Evidence suggests that this lineage originated from western Africa and was introduced to the southeastern parts of Africa through the Bantu dispersals.

The two Tswana-speaking individuals from this investigation that have been assigned to haplogroup L3e1 and L3d1a1a are grouped in separate clades, indicating origin from two different maternal lineages within the L3 haplogroup. Historically it seems that the L3d1

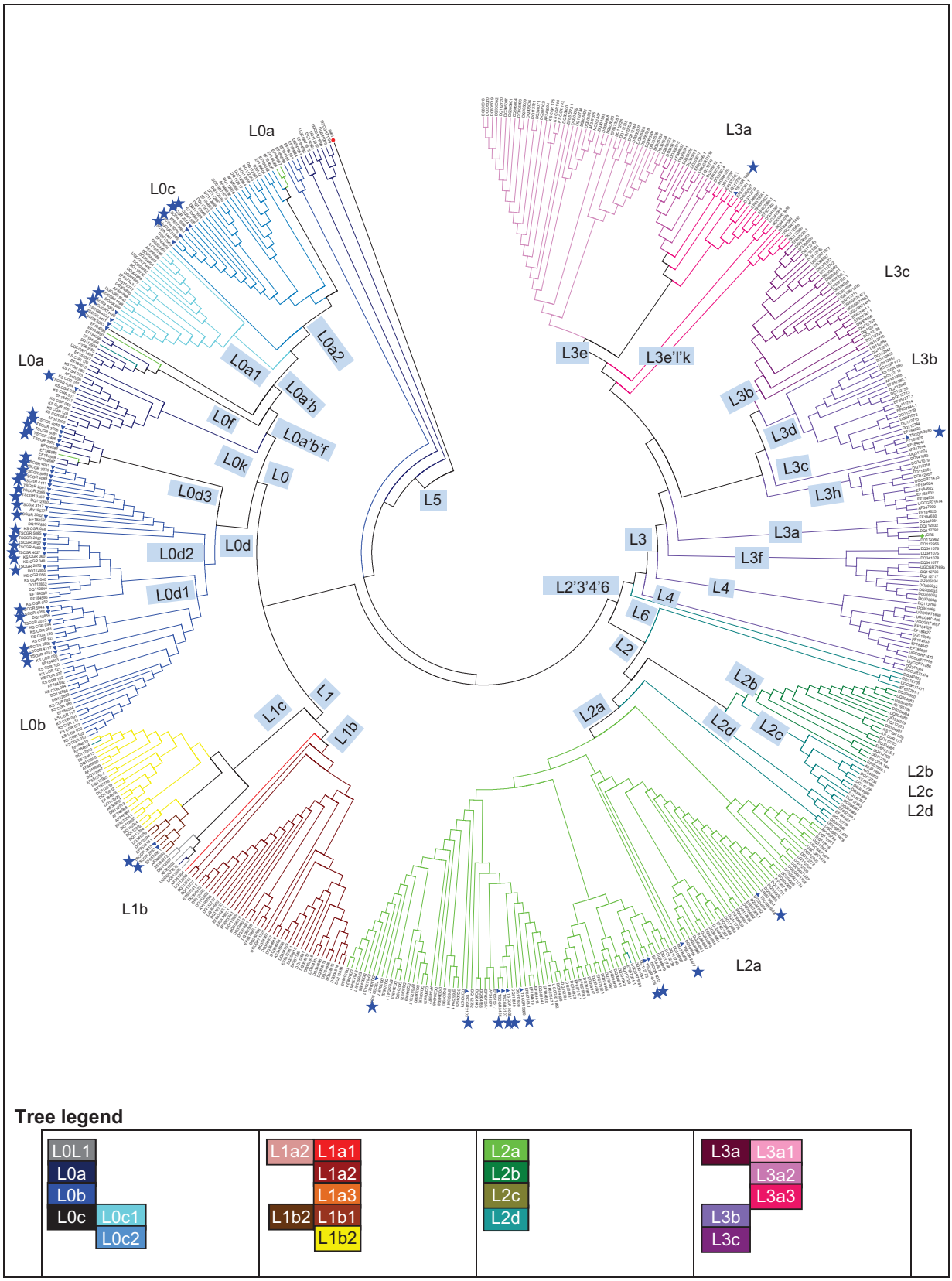
and L3e haplogroups originated in different regions of Africa and reached Africa via different Bantu expansions, as reflected in the positioning of the two haplogroups in the Global African NJ tree.

The Bantu migrations are also a feasible explanation for the presence of the L3d1 haplogroups in the Tswana individuals of this investigation and the history of these haplogroups underlies the fact that they were grouped in two different clades. Because both haplogroups are common to southeastern Africa, it is not surprising that they are observed in the Tswana population.

6.8.2 MP tree of the Global African dataset

The MP tree of the Global African dataset was constructed by using MEGA version 5 (Tamura *et al.*, 2007), which produced an original tree constructed from the Global African dataset and a bootstrap consensus tree constructed from the resampled datasets that represented the confidence of each of the branches of the tree. The original tree was further viewed in MEGA Tree Explorer in a circular format with the bootstrap values indicated on the branches of the tree, which was rooted with the outgroup sequence *Pan troglodytes*. The tree is presented in Appendix E in an A3 format with bootstrap values included and in Figure 6.73 without bootstrap values.

Figure 6.73 MP tree of the Global African dataset



The Global African MP tree represents the coding regions of 574 mtDNA sequences of the Global African dataset as described in Section 5.11.1 according to the MP tree construction method described in Section 6.8. The rCRS is indicated by a green diamond (♦) and the *Pan troglodytes* outgroup is represented by a red circle (●). The haplogroups and sub-haplogroups assigned by the Wallace classification system (2004) are presented in colours as described in the legend. Major haplogroup branching is indicated at the branch splits on the tree and sub-haplogroup groupings are indicated on the outside of the tree. The PhyloTree (Van Oven *et al.*, 2009) haplogroup assignments are indicated in blue text boxes. MtDNA sequence identification is presented at the tips of the branches and the Tswana mtDNA sequences of this investigation are indicated by a blue triangle (▲) and blue stars (★).

This phylogenetic tree provides a good summary of the evolutionary information in the Global African dataset and brings sequence variation into an evolutionary context. Phylogenetic analysis is however always based on a degree of uncertainty owing to the incompleteness of the data because of the limits of sampling all sequence variants and it is therefore important to be able to compare alternative trees to one another to decide on the most optimal topology (Morrison, 1996). For this reason it was decided to construct phylogenetic trees of the same datasets of this investigation, by-tree building methods that differ fundamentally in the tree-construction approach. By comparing the topologies of the NJ and MP trees, the phylogenetic inferences made in this investigation could be validated.

To ensure the accuracy and validity of haplogroup assignment in this investigation, the mtDNA sequences have been assigned to haplogroups by using two different classification systems and comparing the assignments performed in this study to published haplogroup assignments of the sequences (Pereira *et al.*, 2009). The Wallace classification system (2004) assignments are indicated on the MP tree by using colour coding, as described in the tree legend, and labels in black to identify the haplogroups. The haplogroups assigned by the PhyloTree classification system (Van Oven and Kayser, 2009) are indicated by blue highlights and are also displayed on the MP tree branches.

The designation of the major haplogroups L4, L5 and L6 of the MP tree was determined by using the PhyloTree classification system (Van Oven and Kayser, 2009) since the Wallace classification system (2004) does not include haplogroups L4, L5 and L6 in its hierarchy. The mtDNA sequences assigned to these haplogroups are listed in Table 6.36 in Section 6.8.1.

The three most basal branches of the Global African MP tree consist of mtDNA sequences that belong to haplogroup L5 and present as long branches, suggesting an ancient origin. Bootstrap values for these branches are high, indicating confidence in the topology of the clade. In contrast to the Global African NJ tree that groups the L5 haplogroup sequences in one clade, the MP tree groups them in three different clades that are all more basal than the L0 haplogroup, suggesting a more ancient origin. The most ancient clade consists of five mtDNA sequences of which two belong to individuals of Lugbara ethnicity from Uganda, two are of Pygmy origin and one sequence is from an Ethiopian individual. All are assigned to haplogroup L5a, which is a sub-group that has been detected at high

frequencies in Eastern Pygmy populations and is regarded as having an ancient origin in Africa (Batini *et al.*, 2011). The second clade consists of three mtDNA sequences of Ugandan, Ethiopian and Tanzanian origin respectively and all are assigned to haplogroup L5c. The third clade consists of four mtDNA sequences that do not display the L5 defining transitions at np A7972G and A12950G or any of the haplogroup-defining mutations for L5a and L5c. All of the individuals of this clade are of Tanzanian origin (Tishkoff *et al.*, 2007), a region that is reported to harbour the most ancient haplogroups. It is therefore justifiable that it is located at the root of the MP tree. Other Tanzanian mtDNA sequences have been reported to display the absence of these L5 defining polymorphisms, suggesting that it could have been caused by reverse mutations in a maternal ancestor (Gonder *et al.*, 2007). The basal position of the L5 clade could have been influenced by the small number of sequences belonging to this haplogroup in this dataset and the ancient position of this clade in the Global African MP tree should therefore be considered with caution.

The next level of the phylogeny of the Global African MP tree consists of a bifurcation into a branch representing the root for haplogroups L0 and L1 and a branch representing the root for haplogroups L2, L3, L4 and L6. This hierarchy differs from that specified by the Wallace classification system (2004), which proposes a root for the L0, L1 and L2 haplogroups, which is defined by a transition at np C3594T, whereas this transition is listed as one of the root-defining polymorphisms for the paragroup L3'4 in the PhyloTree classification system (Van Oven and Kayser, 2009). PhyloTree proposes a split between the L0 and L1'2'3'4'6 haplogroups, implying the independent development of the L0 haplogroup and a common ancestor between the L1 haplogroup and the L2-6 haplogroups. There is overwhelming evidence in the literature supporting the PhyloTree hierarchy, which supports the hypothesis that the L0 and L1-6 are sister clades that originated in eastern Africa from where there was a major early expansion of modern humans carrying haplogroup L1 to most of the African continent (except the southern African regions) and later followed by a major migration of L2 and L3 clades via different migration routes to the south-eastern regions of Africa and the rest of the world (Ingman *et al.*, 2000; Salas *et al.*, 2002; Kivisild *et al.*, 2006; Torroni *et al.*, 2006; Behar *et al.*, 2008). The tree topology of the Global African MP tree differs from both the classification systems used in this investigation by suggesting a common ancestor between the L0 and L1 haplogroups, which is not shared with haplogroups L2-6. A possible explanation for the tree topology could be that the L0 and L1 clades in this MP tree may seem more similar because of a high number of reverse mutations in the mtDNA sequences, which obscure

the true evolutionary history and cause the MP algorithm to group clades together that are in actual fact more distant than what the sequence variation reflects (Nei, 1996). As only coding-region sequence data have been used in this investigation to rule out the possibility of problems due to reverse mutations, this possibility is not likely. Another possibility to consider is the fact that because the MP tree-building method does not yield a single MP tree, but constructs many equally parsimonious trees based on principles of minimum evolution, the most parsimonious tree might not constructed by the search method applied (Nei and Kumar, 2000). In this investigation, the MP tree was selected by using a heuristic search method in MEGA version 5 (Tamura *et al.*, 2007), termed the CNI method. Using a heuristic search method could be a reason for not obtaining the most parsimonious tree because of the fact that it is not an exhaustive search method and therefore will only search a small proportion of the large number of equally parsimonious trees generated from a large dataset (Nei and Kumar, 2000). There is therefore no guarantee that the most parsimonious tree will be selected. Using the branch-and-bound exhaustive search method would guarantee a more accurate tree topology, but would be computationally too intensive in view of the large dataset used in this investigation. In addition to this, the composition of the variants included in the Global African dataset is more limited, compared to the composition of the datasets used in the studies, which indicates that the L0 and L1-6 haplogroups are on separate branches and that haplogroup L0 is the most ancient grouping. This study includes a broad range of haplogroups that include rare variants and represent as many different haplogroups as possible (Kivisild *et al.*, 2006; Torroni *et al.*, 2006; Gonder *et al.*, 2007; Behar *et al.*, 2008). The composition of the dataset of this investigation is therefore more limited in the sequence variation it portrays than that observed in the other studies mentioned, which could be a possible reason why the tree topology of the Global African MP tree differs from that described in the literature.

The L2’3’4’6 branch in the Global African MP tree leads to a branch that represents haplogroup L2 and another branch that contains the root-defining polymorphisms of haplogroups L3, L4 and L6. According to the positioning of haplogroup L2 in the Global African MP tree, this grouping developed independently from the other haplogroups and suggests an earlier origin. It only shares the more basal ancestry with haplogroups L3, L4 and L6, from where this clade branches off into an independent clade, supporting the theory that Africa was populated by a major expansion of haplogroups L2 and L3 independently (Ingman *et al.*, 2000; Salas *et al.*, 2002; Kivisild *et al.*, 2006; Torroni *et al.*, 2006; Behar *et al.*, 2008). The topology of the Global African MP tree agrees with the haplogroup hierarchy of the PhyloTree classification system (Van Oven and Kayser, 2009)

at this level but differs from that proposed based on the Wallace classification system (2004). The Wallace system proposed a common ancestor for L0, L1 and L2 from where the L2 haplogroup is grouped separately from the other two owing to the absence of the transition at np T10810C. It therefore also indicates an independent development from haplogroup L0 and L1, as reflected in the Global African MP tree, but uses a different paragroup-defining polymorphism from that proposed by the PhyloTree classification system (Van Oven and Kayser, 2009).

The L6 haplogroup branches off from the L3'4'6 root-defining branch in the Global African MP tree, suggesting an earlier independent development from haplogroups L3 and L4. Only one mtDNA sequence (DQ341063) belongs to haplogroup L6 in the Global African dataset, with a 63% bootstrap value, which provides confidence in the tree branching at that position. It agrees with the haplogroup hierarchy of the PhyloTree classification system (Van Oven and Kayser, 2009), which specifies haplogroup root-defining coding region polymorphisms for L3'4 at np 3594, np 7256 and np 13650 that are not shared by individuals belonging to haplogroup L6.

Haplogroup L4 is represented as two separate clades in the Global African MP tree, suggesting the independent development of two sub-clades within the haplogroup. The PhyloTree classification system (Van Oven and Kayser, 2009) presents haplogroup L4 as a single clade defined by a coding region polymorphism at np 5460. Only 15 mtDNA sequences have been assigned to haplogroup L4 within the Global African dataset of this investigation, of which 12 have been grouped in one sub-clade and three (UG_CGR_71474, UG_CGR_71486 and DQ341064) in the other sub-clade. The last-named three mtDNA sequences are the only sequences that display the haplogroup-defining polymorphism at np 5460, as well as all of the haplogroup L4a defining polymorphisms according to the PhyloTree classification system (Van Oven and Kayser, 2009). The other clade of 12 mtDNA sequences does not display the haplogroup-defining polymorphism at np 5460, which is why it has been grouped in a separate clade in the Global African MP tree. The L4a clade was determined using a limited number of mtDNA sequences, of which two are from a study that has not been published yet and could therefore possibly carry a unique sequence variation specific to that region, which has not been taken into account in the construction of the PhyloTree classification system (Van Oven and Kayser, 2009). The other 12 mtDNA sequences of the second clade are listed in Table 6.38 and are grouped together because of the shared haplogroup L4b defining transition at np G3918A and the shared L3 haplogroup polymorphisms at np 769 and np

1018. This group therefore consists of L3 and L4 haplogroup mtDNA sequences when using the PhyloTree classification system (Van Oven and Kayser, 2009).

Table 6.38 Sequences belonging to a sub-clade of haplogroup L4 of the Global African MP tree

GenBank® accession number	GI number	Sequence name/isolate	Ethnicity/geographic origin	Haplogroup/sub-haplogroup classification		
				Published[a]	Wallace, 2004[b]	PhyloTree[c]
---	---	UG_CGR_71474	Acholi	---	L3b$	L4a
---	---	UG_CGR_71486	Baganda	---	L3b$	L4a
DQ341064	84682418	Tor66	Ethiopia	L4a1	L3b$	L4a1
---	---	UG_CGR_71708	Lugbara	---	L0L1$	L4b2
---	---	UG_CGR_71472	Acholi	---	L3b$	L4b2
EF184639	133855349	TZIQ053_L3	Iraqw/Tanzania	L3	L3b$	L3$
EF184640	133855363	TZHZ081_L3	Hadza/Tanzania	L3	L3b$	---
EF184633	133855265	sanC1_L3	San/South Africa	L3	L3b$	L3$
DQ112845	70956024	AF18	Khoi-San/Africa	L4b2a2	L3b$	L4b2
EF184627	133855181	Tzsw011_L3g	Sandawe/Tanzania	L3g	L3b$	L3$
EF184629	133855209	tztr023_L3g	Turu/Tanzania	L3g	L3b$	L3$
---	---	UG_CGR_71657	Baganda	---	L3b$	L4b
---	---	UG_CGR_71660	Baganda	---	L3b$	L4b2
DQ341065	84682432	Tor71	Ethiopia	L4b2a1	L3b$	L4b2
DQ112796	70955926	AF07	Lissongo/Africa	L4b2b	L3b$	L4b

GI = GenInfo Identifier sequence identification number used for identification in GenBank®. Where sequences from other unpublished studies were used, the GenBank® identifiers isolate names and published haplogroup assignments were not relevant and are indicated by ---. Isolates with the prefix UG_CGR were from a study performed by Isabirye (2010). a = Published haplogroups were obtained from Pereira *et al.* (2008). b = Wallace classification system (2004). c = PhyloTree classification system (Van Oven and Kayser, 2009). $ = MtDNA sequence haplogroup assignment that could not be fully resolved. mtDNA sequence EF184640 belonging to an individual of Tanzanian origin could not be assigned to a haplogroup within the PhyloTree classification system (Van Oven and Kayser, 2009) owing to the absence of the haplogroup L3 defining transition at np G769A.

This grouping within the Global African MP tree differs from the PhyloTree classification system (Van Oven and Kayser, 2009) and reflects the unique composition of mtDNA sequences of the Global African dataset of this investigation. The L0 haplogroup forms three major groupings in the Global African MP tree; haplogroup L0a, L0b and L0c, when using the Wallace classification system (2004). These clades have been divided into two groupings, of which one contains the L0a and L0b clades and the other group the L0a and L0c clades. When the Wallace classification system is used, it is suggested that the L0a grouping shares common ancestry with both the L0c and L0b clades. This phylogenetic structure is, however, not supported by the literature, which reports L0a as a single clade that does not share ancestry with L0b and L0c (Torroni *et al.*, 2006; Gonder *et al.*, 2007; Behar *et al.*, 2008) and therefore casts some doubt on the haplogroup assignments by the Wallace classification system (2004).

A further problem when using the Wallace classification system (2004) is that there are five mtDNA sequences that demonstrate discrepancies between their assigned

haplogroup and their position in the Global African MP tree phylogeny. This can be noticed by the incongruous green colour of the branches in contrast to the blue branches of the rest of the L0a branches. Two sequences, EF184607 and EF184608, are assigned to L2a by the Wallace system (2004) but grouped within the L0c clade. On investigation, it was determined that all the sequences had been assigned to haplogroup L0a in the literature (Pereira *et al.*, 2009), as well as when using the PhyloTree classification system (Van Oven and Kayser, 2009). When the PhyloTree classification system (Van Oven and Kayser, 2009) is used, both mtDNA sequences are positioned within the L0a clade and therefore this is a likely possibility. The sequences EF184595 and EF184596 are assigned to haplogroups L2a and L2d respectively by the Wallace classification system, but are grouped within the L0a clade of the tree. These sequences are assigned to L0f in the literature as well as by the PhyloTree classification system (Van Oven and Kayser, 2009). As with the previous sequence, the positioning of these sequences within the phylogeny only makes sense when the PhyloTree classification system (Van Oven and Kayser, 2009) is used and it is grouped with other L0f sequences.

The Wallace classification system (2004) defines haplogroup L0a by the transition at np G11914A in conjunction with a transition at np T10810C and haplogroup L2a by the transition at np G11914A only. The discrepancy in haplogroup assignments of the above-mentioned mtDNA sequences could have been caused by a reverse mutation event at the np T10810C, obscuring the assignment of the sequences to the L0 haplogroup. The tree positioning indicates that the polymorphisms present in these sequences are shared with the mtDNA sequences that belong to haplogroup L0a, indicating an incorrect haplogroup assignment by the Wallace system.

The Wallace classification system (2004) further assigns haplogroup L2a to sequence EF184589 but it is positioned within the L0a clade of the Global African MP tree. The haplogroup for this sequence is published as L0d and assigned to L0d3 when using the PhyloTree classification system (Van Oven and Kayser, 2009). This sequence groups within the L0d clade when the PhyloTree classification system (Van Oven and Kayser, 2009) is used. It is concluded that the latter classification system is more accurate because it can provide an explanation for the positioning of this sequence within the MP tree.

Haplogroup L0 branches into two major groupings in the Global African MP tree to form two sister clades, L0a'b'f'k and L0d, when the PhyloTree classification system (Van Oven

and Kayser, 2009) is used. The basal position of the L0d branch at the root of haplogroup L0 indicates an ancient origin and independent development from the other haplogroups of this clade. The L0a'b'f'k clade bifurcates into a haplogroup L0k branch and a root branch for the L0a'b'f lineages. The split between the L0a'b'f'k and the L0d lineages has been reported as an important event in the structuring of the modern maternal lineages, after which the L0d and L0k clades expanded to southern Africa, giving rise to the ancestors of the contemporary Khoi-San populations (Behar *et al.*, 2008). It has also been hypothesised that the populations that carried the L0d, L0k and L0a'b'f lineages were isolated from each other for a long period of time and that the L0a'b'f lineage migrated back toward the eastern regions of Africa from where it gave rise to the haplogroups L0a, L0b and L0f, which were carried by the modern humans that populated the rest of Africa (Behar *et al.*, 2008). The Global African MP tree of this investigation supports this phylogenetic hierarchy and could be explained by this hypothesis. In further agreement with this theory, two further groupings have been formed from the L0a'b'f branch in the Global African MP tree. These groupings consist of one branch that represents the root for the L0a'b clade and one branch that contains the mtDNA sequences assigned to haplogroup L0f. The L0a'b branch has given rise to the L0a and L0b clades respectively.

Four of the mtDNA sequences of the Tswana-speaking individuals of this investigation are grouped in the L0c clade according to the Wallace classification system and the L0a2 clade according to the PhyloTree classification system (Van Oven and Kayser, 2009). The other mtDNA sequences in this clade originated from the Dominican Republic (five sequences), Tanzania (two sequences), Central African Republic and also included a Pygmy individual, a Sudanese individual and one other Sotho-speaking individual from southern Africa. When the published haplogroups and PhyloTree classification system (Van Oven and Kayser, 2009) are used, these Tswana individuals group in the L0a2 clade, which is the sister clade of the L0a1 haplogroup. This clade is composed of individuals from a wide range of origins, which include a non-African region and central, eastern and southern African regions. This grouping in the Global African MP tree is supported by the literature, which describes haplogroup L0a2 as distributed widely within the Bantu populations in Africa (Kivisild *et al.*, 2004) and as a marker of the major Bantu expansions by its association with the 9 bp deletion in the COII/tRNA^{Lys} region (Salas *et al.*, 2002).

Another three mtDNA sequences from Tswana-speaking individuals are grouped together in a separate clade of haplogroup L0c when the Wallace classification system (2004) is used and in the L0a1 clade when the PhyloTree classification system (Van Oven and

Kayser, 2009) is used. The position of these sequences suggests a separate development from the L0a2 clade but shared ancestry within the L0a'b'f lineage when using the PhyloTree classification system (Van Oven and Kayser, 2009). It also suggests a separate development within the L0c haplogroup when using the Wallace classification system (2004). As with the L0a2 clade, the Tswana-speaking individuals of this clade group with mtDNA sequences from individuals originating from a wide spectrum of African regions and there is also a significant component of individuals of non-African origin. The Tswana-speaking individuals of both clades are located at the shortest terminal branches of the clades and are grouped together, which suggests that the Tswana-speaking population investigated in this study not only carries markers of the early Bantu migrations, but also carries shared polymorphisms, indicating matrilineal commonality between these individuals, which has not been dated in this investigation and therefore it is not possible to comment on the time when this commonality arose. The contribution of ancestry relating to different Bantu populations over a wide range of African regions to the genetic composition of the contemporary Tswana-speaking individuals under investigation in this study is, however, not significant, as only a small proportion of the Tswana-speaking individuals of this investigation (14%) belong to this grouping within the L0 haplogroup of the Global African MP.

One mtDNA sequence of a Tswana-speaking individual (TS_5063) in this investigation is singularly positioned on a long branch that connects to the L0a'b root. The sequence is assigned to haplogroup L0c$ by the Wallace classification system (2004) and to L0a'b with the PhyloTree classification system (Van Oven and Kayser, 2009) and is described in Section 6.8.1. It shares a common ancestry with the L0a clades but does not carry the unique haplogroup-defining polymorphisms that define the different sub-groups of the clade. It is the only sequence that belongs to haplogroup L0a'b within the Global African dataset of this investigation. The positioning of this sequence as a branch of the L0a'b root in the Global African MP tree is in agreement with its positioning in the Global African NJ tree, suggesting that it could belong to haplogroup L0b or be a candidate for an undefined haplogroup, as was explained in Section 6.8.1. The high bootstrap value (91%) of the TS_5063 branch adds to the confidence of the phylogenetic position of this mtDNA sequence within the Global African dataset.

The sequence TS_4034 of this investigation is positioned in a group of mtDNA sequences that are assigned to haplogroup L0a when the Wallace classification system (2004) is used and to haplogroup L0k when the PhyloTree classification system (Van Oven and

Kayser, 2009) is used. This group is a sister clade of the L0a'b'f clade of the PhyloTree classification system and its more basal positioning over haplogroups L0a, L0b and L0f in the Global African MP tree, is in agreement with the literature in which this haplogroup has been reported as one of the lineages that developed independently from the L0a'b'f lineage (Behar *et al.*, 2008). This clade consists of 13 other mtDNA sequences of Khoi-San origin and therefore also supports the theory that it gave rise to the matrilineal ancestry of the Khoi-San gene pool (Behar *et al.*, 2008). The presence of the Tswana-speaking individual within this grouping is significant, since the Khoi-San is currently still a localised population group with distinct maternal ancestry that is not often present in populations of other regions of Africa (Coelho *et al.*, 2009; Behar *et al.*, 2008). The phylogenetic association between the Khoi-San maternal lineages and the Tswana-speaking individuals of this investigation through the presence of high numbers of haplogroup L0d Tswana individuals, is strengthened by the presence of a haplogroup L0k within the Tswana cohort of this investigation and reiterates the likely possibility that these two populations came into contact at some point after the Bantu migrations to the southern regions of Africa and that gene flow took place between these two groups.

Twenty-seven mtDNA sequences of Tswana-speaking individuals of this investigation are grouped within one of the major groupings of the L0 haplogroup of the Global African MP tree. The Wallace classification system (2004) assigns the sequences of this group to two different clades, the L0a and L0b clade. The PhyloTree classification system (Van Oven and Kayser, 2009) assigns all of the sequences in that grouping to haplogroup L0d. The position of this branch is in agreement with the basal nature of the L0d haplogroup reported in other studies (Gonder et al., 2007; Coelho *et al.*, 2009; Behar *et al.*, 2008; Henn *et al.*, 2011) and the larger group itself consists mainly of mtDNA sequences that belong to Khoi-San individuals. The L0d haplogroup is reported to be a major matrilineal contributor to the mtDNA genetic composition of Khoi-San populations and this grouping is therefore not unexpected. Thirty-four (34) of the total of 59 Khoi-San individuals of the Global African dataset are grouped in the L0d clade. Therefore 48 Khoi-San individuals of the Global African dataset are grouped in the L0k and L0d clades in both the Global African MP and NJ trees, indicating that 81% of the Khoi-San individuals carry either the L0d or L0k lineages. This grouping is again in agreement with the L0d and L0k contributions to Khoi-San populations reported in other studies (Chen *et al.*, 2000; Tishkoff *et al.*, 2007; Behar *et al.*, 2008). A significant proportion of 56% of the Tswana-speaking individuals of this investigation carry the Khoi-San mtDNA genetic markers, which is an unequivocal sign of admixture having occurred between the Tswana population and the

local Khoi-San. This theory is supported by evidence that the majority of Bantu-speaking populations of South Africa today have experienced contact with the Ju≠Hoan, Khoe and Tuu speaking Khoi-San hunter-gatherer and herder populations (Mitchell, 2010). It is further supported by previous studies that reported that many of the Bantu-speaking populations of South Africa contain mtDNA lineages derived from Khoi-San sources, most probably through the assimilation of Khoi-San females into patrilocal Bantu-speaking populations (Coelho *et al.*, 2009).

The branching structure of the L1 haplogroup within the Global African MP tree is the same as for the Global African NJ tree and provides confidence in the phylogenetic positioning of this haplogroup within the Global African dataset of this investigation. As with the Global African NJ tree, the L1 haplogroup is split into two major groupings. The one major group represents the haplogroup L1b when the PhyloTree classification system (Van Oven and Kayser, 2009) is used and haplogroup L1a when the Wallace classification system (2004) is used. No Tswana-speaking individuals of this investigation are grouped in this clade. As with the Global African NJ tree, the other major grouping consists of haplogroup L1c divided into four different clades, namely L1c/ L1c3, L1c1, L1c2 and L1c4 when using the PhyloTree classification system (Van Oven and Kayser, 2009) and two +haplogroups, namely L1b and a group that could only be resolved to the level of L1L0, when using the Wallace classification system (2004). The topology of the Global African MP tree further supports the PhyloTree classification system (Van Oven and Kayser, 2009) for this grouping because it displays four defined sub-clades, as is reported for haplogroup L1c and is positioned in a more basal position than the L1b clade, confirming the reported earlier development of the L1c haplogroup (Salas *et al.*, 2002).

The deepest branch of this clade contains four mtDNA sequences (EF184612, DQ112697, DQ112696 and AF381992) that cannot be fully resolved by either classification system. This clade is presented in the same way as in the Global African NJ tree and is discussed in Section 6.8.1. The phylogenetic inferences made in the Global African MP tree agree with the inferences made in the Global African NJ tree that the PhyloTree classification system (Van Oven and Kayser, 2009) should be used and that this sub-clade represents the L1c3 haplogroup.

The mtDNA sequences of two Tswana-speaking individuals of this investigation are grouped in a sub-clade of the L1c haplogroup of the Global African MP tree. MtDNA sequences of two Pygmy individuals and an mtDNA sequence from an individual of

Ethiopian origin are grouped in the same sub-clade, which was expected, since the L1c haplogroup is often associated with Biaka Pygmies and Ethiopians (Salas *et al.*, 2002). The Tswana-speaking individuals of this investigation are positioned as neighbours within the sub-clade, indicating close matrilineal ancestry. Only 4% of the Tswana cohort of this investigation is represented in this clade of the Global African MP tree and this therefore supports the inferences made in the Global African NJ tree with regard to the small matrilineal genetic contribution made by Bantu-speaking populations of central and eastern African origin to the Tswana cohort under investigation. This is discussed in more detail in Section 6.8.1.

The topology of haplogroup L2 of the Global African MP tree is in agreement with the Global African NJ tree and is discussed in Section 6.8.1. Haplogroup L2e and a root branch for haplogroups L2a'b'c'd form the major branching pattern when using the PhyloTree classification system (Van Oven and Kayser, 2009). As in the Global African NJ tree, the L2e clade is the deepest haplogroup and consists of only three mtDNA sequences. One branch leading from the root represents the L2a clade as classified by both classification systems used in this investigation. The other branch bifurcates into a branch representing the L2d clade according to the PhyloTree classification system (Van Oven and Kayser, 2009) and a branch representing the root for the L2b and L2c clades, as classified by both classification systems. It could be inferred, as in the Global African NJ tree, that the L2b and L2c haplogroups share a common ancestor and developed independently from the L2a and L2e haplogroups and that the L2a, L2b, L2c and L2d clades share common ancestry that is not shared with the L2e haplogroup.

The L2a clade consistes of three sub-clades, namely L2a1, L2a2 and L2a4, of which L2a4 is the deepest and only consists of two mtDNA sequences of individuals that originated in Uganda and belonged to the Acholi and Lugbara tribes respectively. This clade is also represented in the Global African NJ tree.

Ten Tswana-speaking individuals of this investigation are positioned within clade L2a1 of the Global African MP tree, which confirms the inference made in the Global African NJ tree that the Tswana population under investigation in this study containes genetic ancestry typical of the Bantu populations of Africa, as discussed in Section 6.8.1 (Quintana-Murci *et al.*, 2008; Coelho *et al.*, 2009). The mtDNA sequences belonging to the Tswana individuals in this clade are not all grouped closely together in the Global African MP tree.

Although the mtDNA sequences of the Tswana individuals TS_2077 and TS_5086 both group with mtDNA sequences that contain the haplogroup L2a1f defining transition at np A5580G according to the PhyloTree classification system (Van Oven and Kayser, 2009), they do not group within the same clade. On further investigation it has been determined that they only differ at two positions within the coding regions. TS_2077 displays a transition at np G15140 and TS_5086 displays a transition at np C14163T, positioning them at opposite ends of the clades. Therefore it can be concluded that these sequences could share the same maternal ancestor and that the respective transitions in each are examples of recent private mutations. The rest of the mtDNA sequences that are assigned to haplogroup L2a1f and group in the same clades as the Tswana sequences belong to individuals residing in non-African countries.

The sequences TS_4013 and TS_4106 are assigned to haplogroup L2a1c by the PhyloTree classification system (Van Oven and Kayser, 2009) and are grouped with other mtDNA sequences belonging to the same haplogroup, of which 13 individuals do not reside in Africa, four are from Burkina Faso, one is from Tanzania and one from Uganda. The mtDNA sequence that belongs to the individual from Uganda displays a discrepancy in terms of its position in the tree and the haplogroup assigned to it. This mtDNA sequence is also present in the Global African NJ tree in the same position and the discrepancy is clarified in Section 6.8.1 and UGCGR71680 has been assigned to haplogroup L2a1c, therefore belonging to this clade.

Other than in the Global African NJ tree, the mtDNA sequence of TS_5060 groups closely with TS_3466, TS_3107 and TS_5062. As was described in Section 6.8.1, TS_5060 belongs to haplogroup L2a1d as assigned by the PhyloTree classification system (Van Oven and Kayser, 2009) and the mtDNA sequences TS_3466, TS_3107 and TS_5062 belong to haplogroup L2a1b, also assigned by the PhyloTree classification system (Van Oven and Kayser, 2009), which is defined by a transition at np G10143A that is not present in TS_5060. As in the case of the Global African NJ tree, TS_3466, TS_3107 and TS_5062 group next to a Pedi individual (DQ112849) and a Khoi-San individual (DQ112900), both originating from southern Africa. The positioning of the Tswana-speaking individuals of this investigation in this clade indicates close maternal ancestry with the mtDNA sequences of the other southern African individuals. It has been noted that ancestry is not only shared between Bantu-speaking individuals of southern Africa but also with a Khoi-San individual, confirming the gene flow between Khoi-San

individuals and Bantu-speaking individuals from southern Africa, as reported in literature (Salas *et al.*, 2002).

The mtDNA sequences of the Tswana-speaking individuals TS_2103 and TS_5066 are both positioned in the sub-clade that represents the mtDNA sequences assigned to haplogroup L2a1 by the PhyloTree classification system (Van Oven and Kayser, 2009). This is in agreement with the Global African NJ tree and indicates shared ancestry with 18 non-African individuals originating from the Dominican Republic and from the United States of America. A small percentage of the Tswana-speaking individuals of this investigation therefore share their ancestry with non-African populations (Salas *et al.*, 2002; Gonder *et al.*, 2007; Batini *et al.*, 2011).

The L3 haplogroup branch consists of clades for haplogroups L3a, L3b, L3c, L3d, L3e, L3f, L3h and L3l'k according to the PhyloTree classification system (Van Oven and Kayser, 2009) and clades only for haplogroups L3a, L3b and L3c when the Wallace classification system (2004) is used. The Wallace classification system (2004) does not define haplogroups L3d, L3e, L3f, L3h, L3l'k and therefore the PhyloTree classification system is used for the discussion of the clades of this haplogroup.

The L3f clade is the most basal branch in the L3 haplogroup of the Global African MP tree and shares ancestry with a branch that consists of the root for haplogroups L3a, L3b, L3c, L3d, L3e, L3h and L3l'k, indicating independent development from these haplogroups within the L3 haplogroup. This major group consists of the L3a clade in the most basal position as sister clade to a root-defining branch for the other L3 haplogroups. This branch has split into another two groups. One group bifurcates into the L3h clade and a root branch for L3b'c'd and the other group bifurcates into the L3e clade and a root branch for L3e'l'k.

Two of the Tswana-speaking individuals of this investigation (TS_3085 and TS_3495) are grouped within the L3 haplogroup of the Global African MP tree and are assigned to haplogroups L3d1a1a and L3e1 by the PhyloTree classification system (Van Oven and Kayser, 2009) and to haplogroups L3b and L3a by the Wallace classification system, (2004) respectively. The Global African MP tree indicates the L3e clade phylogenetically as developing in conjunction with the other L3 haplogroups, whereas the Global African NJ tree indicates the L3e clade in a basal position, developing independently from the other L3 haplogroups. Although this branch displays a bootstrap value of 73% in the MP tree,

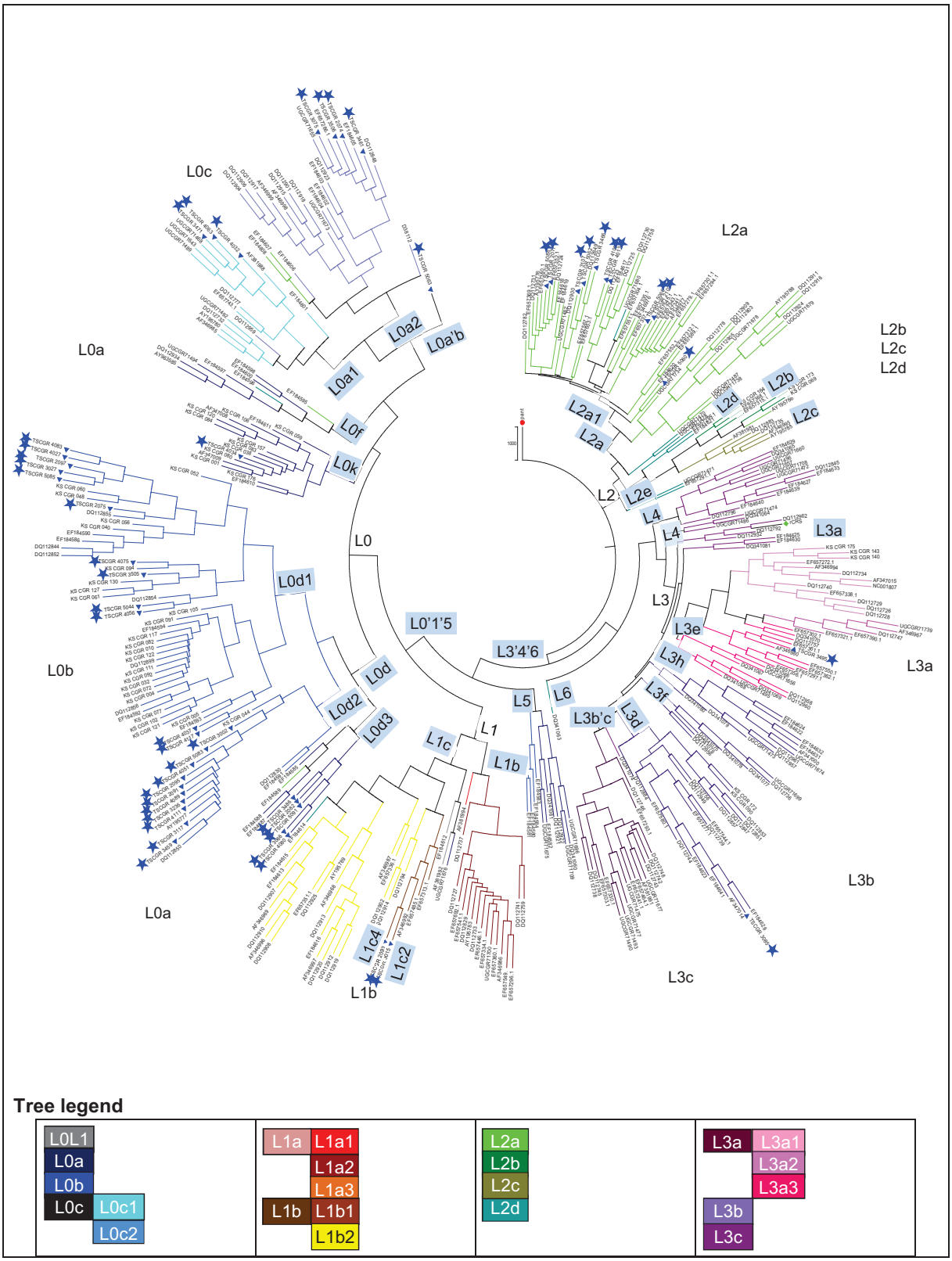
which indicates satisfactory confidence in its positioning within the tree, it does not imply that the positioning is necessarily accurate, as discussed earlier in this section. Upon further investigation of this discrepancy between the Global African MP tree and the Global African NJ tree, it was determined that the PhyloTree classification system (Van Oven and Kayser, 2009) depicts the L3e clade in a more basal position within the haplogroup hierarchy in agreement with the Global African NJ tree. The L3e clade is widely regarded as the most common and ancient clade of the L3 haplogroup (Salas *et al.*, 2002). Therefore, the position of this clade in the Global African MP tree should be regarded with some scepticism.

The transition at np A10819G is the haplogroup-defining site for both haplogroup L3a of the Wallace classification system (2004) and haplogroup L3e'l'k of the PhyloTree classification system (Van Oven and Kayser, 2009) and is therefore present in the mtDNA sequence of TS_3495 and absent in the mtDNA sequence of TS_3085. The sequence of TS_3495 also contains the haplogroup-defining polymorphisms at np 2352 and np 14212, which results in it being assigned to haplogroup L3e by the PhyloTree classification system (Van Oven and Kayser, 2009). These polymorphisms are, however, absent in the mtDNA sequence of TS_3085. The sequence of TS_3085 further contains PhyloTree-defined haplogroup polymorphisms at np 7424, np 8618, np 13886, np 13105 and np 14284 and a Wallace-defined polymorphism at np 5147, which is not shared with the mtDNA sequence of TS_3495. It is concluded that the two Tswana-speaking individuals of this investigation that are positioned within the L3 haplogroup of the Global African MP tree, demonstrate different maternal lineages, which is therefore in agreement with the phylogenetic inferences made in the Global African NJ tree.

6.8.3 NJ tree of the All African dataset

The NJ consensus outtree of the All African dataset was generated by PHYLIP 3.6 (Felsenstein, 1989) and viewed in TreeView (Page, 1996) in a phylogram format, rooted by the outgroup *Pan troglodytes* sequence. The bootstrap number of each node is displayed as the tree length by the PHYLIP 3.6 (Felsenstein, 1989) software and branch lengths are therefore interpreted as bootstrap values and not as indicators of sequence distances between taxa. Thereafter the NJ tree was viewed by using MEGA version 5 (Tamura *et al.*, 2011) Tree Explorer in a circular tree format with the bootstrap values indicated on the branches of the tree. The tree is presented in Appendix E in an A3 format with bootstrap values displayed and in Figure 6.74 without bootstrap values.

Figure 6.74 NJ tree of All African dataset



The All Africa NJ tree represents the coding regions of 385 mtDNA sequences of the All African dataset as described in Section 5.11.2 according to the tree construction method described in Section 6.8. The rCRS is indicated by a green diamond (♦) and the *Pan troglodytes* outgroup is represented by a red circle (●). The haplogroups and sub-haplogroups assigned by the Wallace classification system (2004) are presented in colours as described in the legend. Major haplogroup branching is indicated at the branch splits on the tree and sub-haplogroup groupings are indicated on the outside of the tree. The PhyloTree (Van Oven and Kayser, 2009) haplogroup assignments are indicated in blue text boxes. MtDNA sequence identification is presented at the tips of the branches and the Tswana mtDNA sequences of this investigation are indicated by blue triangles (▲) and blue stars (★)

The All Africa dataset (2a) consists of the mtDNA sequences of individuals of African origin obtained from previous studies and deposited in GenBank®[1], previous PhD phylogenetic studies on African Khoi-San individuals and individuals from Uganda (Koekemoer, 2010; Isabirye, 2010) and the mtDNA sequences of the Tswana-speaking cohort used in this investigation. The All African dataset consists of a total of 390 mtDNA sequences, which have been used to construct an NJ tree to determine the phylogenetic position of the Tswana-speaking cohort of this investigation in the context of only African individuals. The All Africa dataset is representative of all the major regions of Africa as described in Table 5.7 of Section 5.11.6 and serves as representation of the L-haplogroup distribution across Africa, thus providing a backdrop against which the Tswana-speaking population investigated in this study can be positioned in terms of its evolutionary history.

Haplogroups have been assigned to the mtDNA sequences of the All Africa dataset by using the Wallace classification system (2004) and the PhyloTree classification system (Van Oven and Kayser, 2009) as described in Section 6.8. The haplogroups are presented on the All Africa NJ tree by colour-coding the branches of the tree as indicated by the tree legend according to the Wallace classification system and using blue text boxes to indicate the haplogroup assignments of the PhyloTree classification system (Van Oven and Kayser, 2009). As described in Section 6.8, a comprehensive haplogroup assignment is provided in Appendix B of the mtDNA sequences contained in the All Africa NJ tree that consists of the haplogroup assignments by the Wallace classification system (2004), the PhyloTree classification system (Van Oven and Kayser, 2009) as well as the published haplogroups (Pereira *et al.,* 2009) assigned to these sequences.

As noted in the discussions of the previous trees, the Wallace classification system (2004) does not contain haplogroup-defining polymorphic sites for the major haplogroups L4, L5 and L6. For the purpose of discussion of these haplogroups, mtDNA sequences belonging to these haplogroups have been identified by using the PhyloTree classification system (Van Oven and Kayser, 2009).

The phylogenetic structure of the All Africa NJ tree is identical to the phylogenetic structure observed in the Global African NJ tree. Haplogroup L2 is positioned in the deepest position of the tree and suggests an ancient origin. From there two major groups are formed. The

1 GenBank is a registered trademark of the United States Department of Health and Human Services, Bethesda, MD, USA.

first of these groups consists of two sister branches, of which one only contains a single mtDNA sequence belonging to an individual from Ethiopia (DQ341063). This is in accordance with what was observed in the Global African NJ tree and it has been determined that the mtDNA sequence belongs to haplogroup L6b in Section 6.8.1 and is the only mtDNA sequence of this dataset to represent haplogroup L6. This haplogroup has been reported to be rare in Africa (Kivisild *et al.*, 2004).

The second group consists of two branches. One branch represents mtDNA sequences that cannot be resolved further than the L3 haplogroup assignments when using the Wallace classification system (2004), while the other branch splits into a branch containing unresolved haplogroups of L3 and a branch containing mtDNA sequences that are assigned to haplogroup L3. The Global African NJ tree contains the exact clades containing the same mtDNA sequences, which provides high confidence in the grouping of these sequences. As discussed in Section 6.8.1, the mtDNA sequences of these two clades are assigned to haplogroup L4 when using the PhyloTree classification system (Van Oven and Kayser, 2009). The All Africa NJ tree therefore also indicates that haplogroups L3 and L4 share common ancestry, which is in accordance with the PhyloTree haplogroup hierarchy (Van Oven and Kayser, 2009). As in the Global African NJ tree, these two clades contain two mtDNA sequences from Tanzanian individuals (EF184627 and EF184629) that are assigned to haplogroup L3g by the PhyloTree classification system (Van Oven and Kayser, 2009). Haplogroup L3g has been described as a sister clade to haplogroup L4a (Kivisild *et al.*, 2004) and its position in the All Africa NJ tree therefore highlights the common ancestry shared by these two haplogroups.

The clade also contains one mtDNA sequence from a Khoi-San individual from South Africa and two mtDNA sequences belonging to individuals from Tanzania, which are all assigned to haplogroup L3 by the PhyloTree classification system (Van Oven and Kayser, 2009) and to haplogroup L3b by the Wallace classification system (2004). On further investigation, many of the PhyloTree classification system (Van Oven and Kayser, 2009) haplogroup L4b2 defining polymorphisms are observed in these mtDNA sequences. All of these mtDNA sequences, however, lack the L4 haplogroup defining polymorphism at np 5460, which may be a reason why the L4 maternal ancestry has been obscured. This could have been caused by a reverse mutation event at that position within these mtDNA sequences. It has also been noticed that all of these mtDNA sequences containhe haplogroup L3 defining polymorphisms at np 769 and np 1018, indicating shared ancestry with the L3 haplogroup and causing the unresolved haplogroup assignment problem. Only three mtDNA sequences in this dataset display this rare combination of sequence variants

and it may possibly be a rare haplogroup that has not been reported yet or an indication of sequencing errors that occurred in these three mtDNA sequences of the same study. It is, however, likely that these reported mtDNA sequences will have been verified for sequencing error and it is therefore more likely that they represent novel sequence variants.

Although studies have reported that the mtDNA phylogeny of macrohaplogroup L consists of two major branches, L0 and L1'2'3'4'5'6, located on opposite sides of the root (Behar *et al.*, 2008), haplogroups L0'1'3'4'5'6 are grouped together from one root in the All African NJ tree and reasons for the unexpected topology are discussed in depth in Section 6.8.1.

The bootstrap values of the major haplogroup branches of the Global African NJ tree and the All African NJ tree are listed in Table 6.39. Both trees display similar bootstrap values, which is to be expected, since the NJ trees are constructed from similar data and by the same tree-building method. It is concluded that all the branches can be accepted with confidence, except for the branches representative of the L3'4'6 and L3 haplogroups in both trees and the branches representing haplogroup L4 in the Global African NJ tree. This may be due to the method of tree construction that could not support the phylogeny after resampling. Bootstrap values do not provide a probability that a clade is a true clade, but only the probability that a particular phylogeny will be obtained upon re-sampling, thus indicating how well a tree-building method supports a clade. For this reason, bootstrap values must be interpreted by taking into account variables that can influence the probability that a group represents a true clade, such as the number of characters used, the number of taxa, the evolutionary rate of change, mutational rate among lineages, independence of the characters and the type of tree-building method used (Hillis and Bull, 1993), as discussed in Section 6.8.1.

Table 6.39 Bootstrap values for Global African NJ tree and All African NJ tree

Haplogroups	Global African NJ tree	All African NJ tree
L0'1'5	64%	65%
L0	91%	91%
L1	98%	98%
L2	66%	66%
L3'4'6	44%	47%

Table 6.39 Continue…

Haplogroups	Global African NJ tree	All African NJ tree
L3	32%	37%
L4	48% and 30%	86% and 65%
L5	67%	68%
L6	100%	100%

L0'1'5 and L3'4'6 refer to root-defining branches that branch into the respective haplogroups. The bootstrap values are listed as percentages by calculating these from the values listed as branch lengths in the MEGA 5 Tree Explorer.

The other branch of this major group represents haplogroup L5 as defined by the PhyloTree classification system (Van Oven and Kayser, 2009) and consistsof the same 11 mtDNA sequences of which the L5 clade in the Global African NJ tree consists (and which are listed in Table 6.36), as well as the mtDNA sequence of an individual from Uganda (UG_CGR_71675), which has been assigned to haplogroup L3 but contains sequence variations that group with the L5 clades in both trees, confirming the possibility that this individual could contain haplogroup-defining polymorphisms not incorporated into the current haplogroup classification systems. The basal position of this clade in the All African NJ tree also indicates an ancient nature for this group of mtDNA sequences, as in the case of the Global African NJ tree. However, since the clade only consists of 12 mtDNA sequences, it should be interpreted with care and the possibility that the clade could be positioned differently if more L5 haplogroup individuals should be included in the dataset, should be considered.

The L0 haplogroup consists of two major groups. The first group consists of mtDNA sequences assigned to haplogroup L0b by the Wallace classification system (2004) and to haplogroup L0d with the PhyloTree classification system (Van Oven and Kayser, 2009). Since the published haplogroups assigned to the mtDNA sequences of this clade agree with the PhyloTree classification system (Van Oven and Kayser, 2009), the latter classification system has been used to interpret this clade of the All African NJ tree.

The position of the L0d clade is in agreement with what has been reported in the literature and indicates that the mtDNA sequences belonging to haplogroup L0d are of an ancestral nature (Chen *et al.*, 1995a; Ingman *et al.*, 2000; Mishmar *et al.*, 2003; Kivisild *et al.*, 2004; Gonder *et al.*, 2007). This haplogroup was split into sub-groups L0d1, L0d2 and L0d3. Twelve mtDNA sequences are grouped in the L0d1 clade, of which six mtDNA sequences from the Tswana-speaking individuals of this investigation group within a sub-clade that represents haplogroup L0d1b. Five of these mtDNA sequences are grouped as neighbours, indicating close genetic ancestry. These individuals display novel sequence

mutations at np 12436 and np 15337, which suggests that they belong to a candidate new haplogroup - L0d1b2, according to the PhyloTree classification system (Van Oven and Kayser, 2009). The sequence variation is discussed in depth in Section 6.6. The Tswana-speaking individuals of this sub-clade are grouped with eight mtDNA sequences belonging to Khoi-San individuals from southern Africa, one mtDNA sequence belonging to a Sandawe individual from Tanzania and one mtDNA sequence belonging to a southern African Tswana-speaking individual. Four mtDNA sequences from the Tswana-speaking individuals of this investigation are grouped together in the sub-clade, representing haplogroup L0d1a with four mtDNA sequences from Khoi-San individuals and one mtDNA sequence belonging to a Zulu-speaking individual of African origin. Two of the Tswana-speaking individuals group as neighbours, again indicating close genetic ancestry. Another two mtDNA sequences from Tswana-speaking individuals included in this investigation group with 20 other mtDNA sequences from Khoi-San individuals of southern African origin in a sub-clade that represents the haplogroup L0d1c. The composition of this clade is the same as in the Global African NJ tree and therefore provides confidence for this grouping. The L0d2 clade consists of ten mtDNA sequences from Tswana-speaking individuals of this investigation, three mtDNA sequences from Khoi-San individuals of southern African origin, one mtDNA from a Tswana individual and one mtDNA sequence from a Zulu-speaking individual, both of southern African origin, as well as one mtDNA sequence from an individual of unknown African origin. The same Tswana-speaking individuals that group in this clade in the Global African NJ tree are grouped together in this clade in the All African NJ tree. Most of the Tswana-speaking individuals of this clade are grouped as neighbours, again indicating a strong genetic ancestry. The L0d3 clade consists of a branch consisting of four mtDNA sequences from individuals of Tanzanian origin and another branch consisting of five Tswana-speaking individuals of this investigation. The Tswana-speaking individuals of this investigation that are grouped in this clade all contain polymorphisms at np 6170, np 7119, np 8290, np 10114 and np 10128, which are absent from the Tanzanian individuals present in this clade. No sub-groups have been reported for haplogroup L0d3 by the PhyloTree classification system (Van Oven and Kayser, 2009) and it is therefore concluded that the Tswana individuals of this investigation represent a sub-group of haplogroup L0d3 that has not been previously reported and that the Tswana individuals do not share the same ancestry in terms of the sub-grouping as the Tanzanian individuals. It is also clear that the Tswana individuals of this clade share close ancestry, indicating a small number of maternal ancestors that introduced this haplogroup into the Tswana population.

The same was concluded for the L0d clade as determined for the Global African NJ tree (and discussed in Section 6.8.1), that the large number of Tswana-speaking individuals grouped within this clade, which has been reported as an ancient haplogroup that belongs mostly to Khoi-San individuals (Gonder *et al.*, 2007), indicates gene flow between the Tswana population of this investigation and the Khoi-San. The presence of the Bantu-speaking individuals in this clade also confirms the reported gene flow between other southern African Bantu populations and the Khoi-San (Kivisild *et al.*, 2006).

The other major group consists of a branch representing the haplogroup L0k and a branch that represents the root to haplogroups L0a, L0b, and L0f. This hierarchy is different from what has been observed in the Global African NJ tree, but is identical to the haplogroup hierarchy of the PhyloTree classification system (Van Oven and Kayser, 2009) and in agreement with the haplogroup structure reported by Gonder *et al.* (2007).

Other than that which has been observed in the Global African NJ tree, the positioning of the L0k clade within the tree indicates that it developed independently from the L0d haplogroup and shares ancestry with haplogroups L0f, L0a and L0b. One Tswana-speaking individual groups within this clade with ten other mtDNA sequences from Khoi-San individuals of southern African origin and another two mtDNA sequences from Khoi-San individuals of unknown African origin, thus confirming the connection between the Tswana-speaking population of this investigation and the Khoi-San population of southern Africa, as discussed in Section 6.8.1.

The other clade in this major grouping consists of a sub-clade that represents mtDNA sequences belongng to haplogroup L0a when the Wallace classification system (2004) is used, and to haplogroup L0f and serving as a root to haplogroups L0a and L0b when the PhyloTree classification system (Van Oven and Kayser, 2009) is used. This latter hierarchy is in agreement with literature and the PhyloTree classification system (Van Oven and Kayser, 2009), which indicates that haplogroups L0a and L0b did originate from a shared maternal ancestor. It seems that two mtDNA sequences (EF184595 and EF184596) had not been positioned correctly in the L0f clade and after investigation in Section 6.8.1, it was determined that the Wallace classification system assigned the incorrect haplogroup to these mtDNA sequences and that they were correctly positioned when using the PhyloTree classification system (Van Oven and Kayser, 2009).

As with the Global African NJ tree, the L0a'b root branches into a single mtDNA sequence that belongs to haplogroup L0c according to the Wallace classification system (2004) and to an unresolved haplogroup L0a'b when the PhyloTree classification system (Van Oven and Kayser, 2009) is used. The possibility of this mtDNA sequence belonging to haplogroup L0b is discussed in Section 6.8.1 and this is confirmed by its position within the All African NJ tree. The L0a'b root also has a high bootstrap value of 99% and is therefore strongly supported within the tree. The Wallace classification system (2004) only assigns haplogroup L0c to the mtDNA sequences of all the sub-clades within this clade. By using the PhyloTree classification system (Van Oven and Kayser, 2009), a more resolved haplogroup hierarchy can be obtained, in which the groupings of the mtDNA sequences of the Tswana individuals of this investigation can be interpreted more fully.

Three mtDNA sequences that belong to Tswana individuals of this investigation are grouped within the L0a1 sub-clade of the L0a clade according to the PhyloTree classification system (Van Oven and Kayser, 2009) and are grouped with mtDNA sequences belonging to individuals from a broad region of Africa as well as to individuals of African-American origin. This confirms the ancestry that is shared between the Tswana-speaking population of this investigation and the Bantu lineages. Only a small number of the Tswana-speaking population belong to this haplogroup, suggesting that the genetic component contributed to this population by the Bantu-speaking lineages has been minor. In addition, these sequences are grouped closely together, indicating close ancestry between the Tswana-speaking individuals within the sub-clade. They also group closely with three mtDNA sequences from individuals of Acholi and Baganda ethnicity from Uganda, thus also indicating close genetic ancestry with the Ugandan individuals.

The L0a2 sub-clade, according to the PhyloTree classification system (Van Oven and Kayser, 2009), consists of two major groups. One group contains four mtDNA sequences belonging to Tswana-speaking individuals of this investigation and the other group consists of three mtDNA sequences that have not been positioned correctly according to the Wallace classification system (2004) and are indicated as such by different colouring of the branches. Since the same grouping was observed in the Global African NJ tree, this discrepancy was investigated in Section 6.8.1 and it was determined that the haplogroup assignment by the Wallace classification system (2004) could not be accepted and that the PhyloTree classification system (Van Oven and Kayser, 2009), which assigns these mtDNA sequences to haplogroup L0a2, positions these mtDNA sequences correctly. The positioning of the mtDNA sequences of the Tswana-speaking individuals of this

investigation in this sub-clade indicates that this population carries markers of the Bantu migration. The mtDNA sequences of the Tswana-speaking individuals were are grouped close together, suggesting close genetic ancestry.

The grouping that represents the L1 haplogroup in the All African NJ tree is composed of fewer mtDNA sequences than those observed in the Global African NJ tree, indicating that a substantial number of the mtDNA sequences in the Global African dataset that originated from non-African countries belong to haplogroup L1 and have therefore been removed from the All African dataset. The hierarchy of the L1 haplogroup in the All African NJ tree is identical to the Global African NJ tree. Haplogroup L1 can be divided into two major groups. One group represents haplogroup L1b and the other group haplogroup L1c when the PhyloTree classification system (Van Oven and Kayser, 2009) is used. In contrast, when the Wallace classification system is used the daughter clades of L1 are L1a and L1b. None of the Tswana-speaking individuals are grouped in the first clade that represents haplogroup L1b (PhyloTree classification) and L1a (Wallace classification) respectively. Only two Tswana-speaking individuals from this investigation are grouped in the L1 haplogroup and both of them group within haplogroup L1c as defined by the PhyloTree classification system or the L1b1 haplogroup when the Wallace classification is used. The Wallace classification system (2004) defines haplogroup L1b1 by a transition at np G6150A and haplogroup L1b2 by a transition at np T11899C. It was therefore not expected that haplogroup L1b2 would be present in the two different clades, especially when it is highlighted that it shares ancestry with haplogroup L1b1 in one of the clades, as is the case in the All African NJ tree. When the PhyloTree classification system (Van Oven and Kayser, 2009) is used, all the mtDNA sequences in the same two branches in the All African NJ tree are assigned to haplogroup L1c. The branch containing the two Tswana-speaking individuals of this investigation is assigned to haplogroup L1c2 and both are defined by the same transition at np 6150, as specified by the Wallace classification system, in addition to ten other coding region polymorphisms. The other branch in this clade consists of mtDNA sequences that are assigned to haplogroup L1c4 that is defined by eight different coding region polymorphisms, as reported in the PhyloTree classification system (Van Oven and Kayser, 2009), of which the transition at np 11899, as used in the Wallace classification system, is not used. Using the PhyloTree classification system (Van Oven and Kayser, 2009) therefore provides a more resolved haplogroup assignment that can better explain the positioning of the mtDNA sequences in the tree.

The haplogroup assignments by the Wallace classification system (2004) and the PhyloTree classification system (Van Oven and Kayser, 2009) agree in most instances in this grouping of the tree, except for mtDNA sequences EF184621, EF657299, UG_CGR_71478, UG_CGR_71470 and UG_CGR_71680, which are assigned to haplogroup L2d by the Wallace classification system (2004), but grouped in different sub-clades of the L2 clade. This is problematic, since the L2d haplogroup is defined by the Wallace classification system (2004) by a transition at np G13708A, which is present in all the mtDNA sequences and it is therefore expected that these sequences will be grouped together. The PhyloTree classification system (Van Oven and Kayser, 2009) defines the L2 sub-clades by many different coding region polymorphisms and when this system is used, it has been determined that the mtDNA sequences that demonstrate discrepancy with its positioning in the All African NJ tree, group correctly when the PhyloTree classification system (Van Oven and Kayser, 2009) is used. The Wallace classification system (2004) is therefore lacking in haplogroup-defining polymorphisms to resolve the haplogroup L2 mtDNA sequences fully. For this reason the PhyloTree classification system (Van Oven and Kayser, 2009) has been used for the interpretation of the positioning of the Tswana-speaking individuals of this investigation within this clade.

The L2a clade consists of three sub-clades, L2a1, L2a2 and L2a3, when the PhyloTree classification system (Van Oven and Kayser, 2009) is used. Ten Tswana-speaking individuals (20%) from this investigation are grouped within the sub-clade L2a1. As discussed in Section 6.8.1, because the L2a haplogroup has been identified as a genetic marker of the Bantu expansions and is common to the Bantu populations of Africa (Quintana-Murci *et al.*, 2008; Coelho *et al.*, 2009), this grouping of 20% of the Tswana-speaking individuals of this investigation is to be expected and confirms that genetic flow has been evident between the early Bantu populations that reached southern Africa through the major Bantu migrations from western central Africa and the modern-day Tswana population of this investigation.

As observed in the Global African NJ tree, one Tswana-speaking individual of this investigation (TS_5060) is grouped with one other individual from Tanzania, EF184620, in a clade that branches off at the root of the L2a1 branch because of unique polymorphisms at np 5196, np 9530, np 11386, np 12612 and np 13934, which are not shared with the other mtDNA sequences within this clade. Three mtDNA sequences from Tswana-speaking individuals from this investigation (TS_3466, TS_3107 and TS_5062) are grouped closely together in a sub-clade with a Pedi individual (DQ112849) and a

Khoi-San individual (DQ112900), both originating from southern Africa. The same cluster is observed in the Global African NJ tree and confirms the shared ancestry of the southern African Pedi-speaking, Khoi-San and Tswana-speaking individuals, which could be due to genetic flow between the Khoi-San populations and the Bantu-speaking populations of southern Africa. Two mtDNA sequences from Tswana-speaking individuals from this investigation (TS_2077 and TS_5086) are grouped together in a sub-clade belonging to haplogroup L2a1f, according to the PhyloTree classification system (Van Oven and Kayser, 2009) and share common ancestry with six other individuals of African origin. Another two Tswana-speaking individuals of this investigation are grouped in a sub-clade that contains mtDNA sequences that belong to haplogroup L2a1c together with four mtDNA sequences from individuals from Burkina Faso and one individual from Tanzania. Two Tswana-speaking individuals of this investigation are grouped in a sub-clade that contains mtDNA sequences that belong to haplogroup L2a1a with eight mtDNA sequences from individuals of African origin.

The hierarchy of haplogroup L3 in the All African NJ tree differs from the hierarchy in the Global African NJ tree in terms of where haplogroup L3a according to the PhyloTree classification system (Van Oven and Kayser, 2009) is grouped. Haplogroup L3 bifurcates into two groups, of which one represents haplogroups U and H and the other the L3 haplogroup according to the PhyloTree classification system (Van Oven and Kayser, 2009). The L3 haplogroup branch splits into one branch that represents the root for haplogroups L3a'e'l'k and another branch that represents the root for haplogroups L3b'c'd'f'h by the PhyloTree classification system (Van Oven and Kayser, 2009). The Wallace classification system (2004) only distinguishes between haplogroups L3a, L3b and L3c. Haplogroup L3 is divided into two groups containing haplogroups L3a and a root for haplogroup L3b'c. The difference in tree topology can therefore only be observed when the better resolved PhyloTree classification system (Van Oven and Kayser, 2009) is used and this system will therefore be used to describe the topology of this haplogroup in the All African NJ tree.

Haplogroup L3a is therefore grouped with the L3e'l'k branch, suggesting a common ancestor between the two clades. In the Global African NJ tree, the L3a haplogroup is grouped with the L3b'c'd'f'h branch. The reason for the difference in the topology between the trees could be the fact that 65 mtDNA sequences have been removed from the Global African dataset to compile the All African dataset and this has resulted in 40% fewer mtDNA sequences that are taken into account in the determination of the genetic

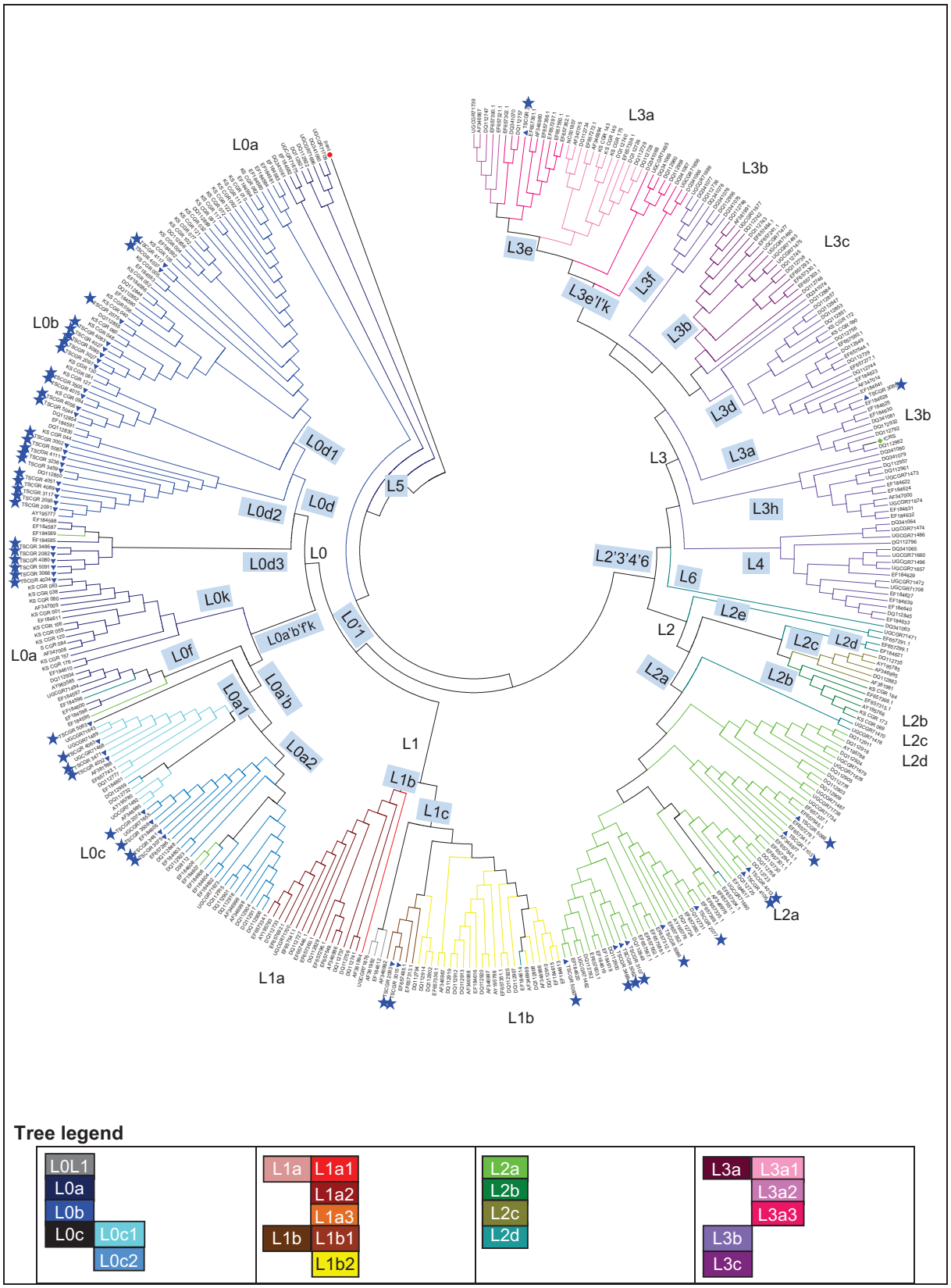
distances and therefore the branch topology. The basal positioning of the haplogroups U and H is also only observed when the PhyloTree classification system (Van Oven and Kayser, 2009) is used and differs from what is observed in the Global African NJ tree, possibly for the same reason.

TS_3495 groups within the clade that contains mtDNA sequences belonginged to haplogroup L3e1 when the PhyloTree classification system (Van Oven and Kayser, 2009) is used and to haplogroup L3a when the Wallace classification is used. As discussed in Section 6.8.1, the presence of a Tswana individual in this clade indicates that the Tswana population under investigation harbours a haplogroup that is regarded as a marker of the Bantu migration and also a haplogroup that has been reported in the Khwe population of southern Africa (Coelho *et al.*, 2009). The mtDNA sequence belonging to TS_3085 groups within a different clade that contains mtDNA sequences belonging to haplogroup L3d1a1a according to PhyloTree and to haplogroup L3b when the Wallace classification system (2004) is used. This confirms the presence of a lineage that was introduced to the southeastern regions of Africa through the Bantu dispersals (Salas *et al.*, 2002).

6.8.4 MP tree of the All African dataset

The All African MP tree was constructed by using the All African dataset (2a) as described in Section 5.11.2. It consists of mtDNA sequences belonging to individuals that reside only within Africa. The purpose of this analysis is to visualise the phylogenetic positioning of the mtDNA sequences of the Tswana individuals of this investigation in the context of the mtDNA sequences of the African phylogeographic groupings. The MP tree of the All African dataset was constructed by using MEGA version 5 (Tamura *et al.*, 2007) and viewed in MEGA Tree Explorer in a circular format rooted at the outgroup sequence *Pan troglodytes*. The bootstrap values are indicated on the tree branches and the tree itself is presented in Appendix E in A3 format with bootstrap values included and in Figure 6.75 without bootstrap values.

Figure 6.75 MP tree of the All African dataset



The All African MP tree represents the coding regions of 387 mtDNA sequences of the All African dataset as described in Section 5.11.2 according to the MP tree construction method described in Section 6.8. The rCRS is indicated by a green diamond (♦) and the *Pan troglodytes* outgroup is represented by a red circle (●). The haplogroups and sub-haplogroups assigned by the Wallace classification system (2004) are presented in colours as described in the legend. Major haplogroup branching is indicated at the branch splits on the tree and sub-haplogroup groupings are indicated on the outside of the tree. The PhyloTree (Van Oven and Kayser, 2009) haplogroup assignments are indicated in blue text boxes. MtDNA sequence identification is presented at the tips of the branches and the Tswana mtDNA sequences of this investigation are indicated by a blue triangle (▲) and blue star (★).

The individuals in the All African MP tree have been haplogrouped by using the two classification systems used for the interpretation of the other phylogenetic trees of this study, namely the Wallace classification system (2004) and the PhyloTree classification system (Van Oven and Kayser, 2009). To differentiate between the two different systems, the tree branches are colour-coded according to the assigned haplogroups by the Wallace classification system (2004), as explained in the tree legend, and the PhyloTree classification system (Van Oven and Kayser, 2009) haplogroup assignments are indicated in blue text boxes on the tree itself. As in the case of the interpretation of the major haplogroups of the other phylogenetic trees in Sections 6.8.1, 6.8.2 and 6.8.3, the PhyloTree classification system (Van Oven and Kayser, 2009) is used to identify and describe the positioning of the major haplogroups L4, L5 and L6 because the Wallace classification system (2004) does not define these haplogroups.

The phylogenetic hierarchy of the major haplogroups in the All African MP tree is identical to the Global African MP tree hierarchy except for the positioning of haplogroup L4. Haplogroup L5 is positioned as the most basal haplogroup in the All African MP tree and consists of three (3) branches that are in exact agreement with those observed in the Global African MP tree. The most basal and therefore most ancient of the three clades consists of the same mtDNA sequences as in the Global African MP tree and as described in Section 6.8.2, all belong to haplogroup L5a, which has been reported to be common to the Eastern Pygmy populations and which has an ancient origin (Batini *et al.*, 2011). The second clade consists of three mtDNA sequences and belongs to haplogroup L5c, while the third clade contains four mtDNA sequences of Tanzanian origin that belong to haplogroup L5, but do not display the L5 haplogroup-defining transitions at np A7972G and A12950G or any of the haplogroup-defining mutations for L5a or L5c. This phylogenetic grouping of the All African MP tree therefore confirms the phylogenetic association between these mtDNA sequences and supports the possibility that it represents an ancient haplogrouppreviously reported in the Tanzanian population (Gonder *et al.*, 2007), as described in Section 6.8.2. The bootstrap values are 40% and 30% respectively for the branches of the L5 haplogroup clades, which does not indicate high confidence in the positioning of these branches. The small number of mtDNA sequences that define these clades could possibly be the reason for the low bootstrap value.

The next major grouping within the All African MP tree consists of one branch that represents a root for the L0 and L1 haplogroups and another branch that represents a root branch for haplogroups L2, L3, L4 and L6. Although this agrees with the hierarchy

observed in the Global African MP tree, it is unexpected because of the evidence in the literature supporting the hypothesis that L0 and L1-6 are sister clades that originated in eastern Africa, from where they spread across the continent (Ingman *et al.*, 2000; Salas *et al.*, 2002; Kivisild *et al.*, 2006; Torroni *et al.*, 2006; Behar *et al.*, 2008). Reasons for obtaining alternative tree topologies when using the MP tree-building method are discussed in Section 6.8.2 and are also applicable to the observed tree topology in the All African MP tree.

The L2'3'4'6 branch bifurcates into a branch that contains mtDNA sequences belonging to haplogroup L2 and a root branch for haplogroups L3, L4 and L6. The presence of the L2 clade in a more basal position than the branches for haplogroup L3, L4 and L6 is in agreement with the Global African MP tree and in agreement with the published theory that Africa was populated by a major expansion of haplogroups L2 and L3 independently (Ingman *et al.*, 2000; Salas *et al.*, 2002; Kivisild *et al.*, 2006; Torroni *et al.*, 2006; Behar *et al.*, 2008). The L3'4'6 root branch bifurcates into a branch that contains a single mtDNA sequence that is assigned to haplogroup L6 by the PhyloTree classification system (Van Oven and Kayser, 2009) and belongs to an individual originating from Ethiopia. As with the Global African MP tree, this hierarchy indicates that haplogroup L6 developed earlier and independently from haplogroups L3 and L4. The other branch represents a root branch for L3 and L4 haplogroups. These haplogroups bifurcate into a branch that contains mtDNA sequences belonging to the L4 haplogroup and another that contains mtDNA sequences belongingto haplogroup L3.

The L4 haplogroup is split into two clades in the Global African MP tree, which is not the case in the All African MP tree, in which the same groupings are split into two sub-clades within the L4 clade. One sub-clade consists of the same three mtDNA sequences that are observed in the one clade of the Global African MP tree L4 haplogroup grouping. It therefore consists of mtDNA sequences that harbour L4 haplogroup-defining polymorphisms at np 5460, as well as all the other coding region polymorphisms describing haplogroup L4a, as defined by the PhyloTree classification system (Van Oven and Kayser, 2009). The other sub-clade within the L4 clade of the All African MP tree consists of 12 mtDNA sequences that agree with the grouping observed in the Global African MP tree and therefore consists of a group of mtDNA sequences that does not display the haplogroup L4 defining polymorphism at np 5460 but does display coding region polymorphisms defined for haplogroup L4b and haplogroup L3 at np 769 and np 1018 according to the PhyloTree classification system (Van Oven and Kayser, 2009).

The hierarchy of the L4 haplogroup of the All African MP tree is in agreement with the hierarchy reported in the PhyloTree classification system (Van Oven and Kayser, 2009). The reason for the difference in topology between the two MP trees could be ascribed to the MP tree-building method that uses a heuristic search method, which does not guarantee the selection of the most parsimonious tree, as discussed in depth in Section 6.8.2.

The bootstrap values of the major haplogroup branches in the Global African MP tree and the All African MP tree are listed in Table 6.40. The bootstrap values of all the major haplogroup branches except for haplogroup L5 in both trees display values higher than 50%, indicating a high level of confidence in the tree topology. The lower bootstrap value of the branches of haplogroup L5 could be due to the small number of mtDNA sequences assigned to this clade and is therefore to be expected.

Table 6.40 Bootstrap values for Global African MP tree and All African MP tree

Haplogroups	Global African NJ tree	All African NJ tree
L0'1	83%	82%
L0	87%	86%
L1	97%	98%
L2	84%	85%
L3'4'6	63%	57%
L3	57%	51%
L4	65% and 17%	65%
L5	40% and 29%	40% and 30%
L6	63%	62%

L0'1 and L3'4'6 refer to root-defining branches that branch into the respective haplogroups. The bootstrap values are listed as percentages displayed in the MEGA 5 Tree Explorer.

The use of the Wallace classification system (2004) to describe the phylogenetic positioning of the mtDNA sequences within the L0 clade in the All African MP tree is problematic for the same reasons as observed in the Global African MP tree. When the Wallace classification system (2004) is used, it displays two groupings within the L0 clade of which one contains L0a and L0b clades and the other group contains the L0a and L0c clades, indicating a shared ancestry between the L0a clade and the L0b and L0c clades simultaneously. This is in contrast to the reported independent development of the L0a clade from the L0b and L0c clades (Torroni *et al*., 2006; Gonder *et al.*, 2007; Behar *et al.*, 2008). When the PhyloTree classification system (Van Oven and Kayser, 2009) is used, the L0a clade groups separately from the L0b clade, as has been reported.

The use of the Wallace classification system (2004) furthermore assigns haplogroups to five mtDNA sequences that do not agree with their positioning within the tree. These mtDNA sequences can be identified by the incongruous green colours of the branches in contrast to the blue colour of the other mtDNA sequences within the L0 clade. A similar discrepancy has been determined in the Global African MP tree. On further investigation, as described in Section 6.8.2, it was determined that mtDNA sequences EF184607 and EF184608 belong to haplogroups L0a according to the PhyloTree classification system (Van Oven and Kayser, 2009) and as reported by Pereira (Pereira *et al.*, 2009) and not to haplogroup L2a, as assigned by the Wallace classification system (2004) and that these sequences are positioned within the clade representative of the haplogroup L0a. It was also determined that EF184595 and EF184596, which are respectively assigned to haplogroups L2a and L2d by the Wallace classification system (2004), belong to haplogroup L0f according to the PhyloTree classification system (Van Oven and Kayser, 2009) and published records, which makes its positioning with L0a in the All African MP tree correct. MtDNA sequence EF184589 is also assigned to haplogroup L2a by the Wallace classification system (2004), which on further investigation, using the PhyloTree classification system (Van Oven and Kayser, 2009), belongs to haplogroup L0d and is positioned correctly in the All African MP tree. This phenomenon of the incorrect assignment of the L2 haplogroup to mtDNA sequences that belong to haplogroup L0 may be explained by a reverse mutation event in these sequences, as described in Section 6.8.2.

The phylogenetic hierarchy of the L0 clade of the All African MP tree is identical to what was observed in the Global African MP tree and discussed in Section 6.8.2. The L0 clade bifurcates into a branch that represents haplogroup L0d and a sister branch that represents a root to the haplogroups L0a, L0b, L0f and L0k according to the PhyloTree classification system (Van Oven and Kayser, 2009). The L0d branch is therefore the most basal node within the L0 clade, which confirms the ancient nature of this haplogroup, as reported in literature (Behar *et al.,* 2008). It also confirms its independent development from the other sub-haplogroups of the major haplogroup L0, providing further evidence for the hypothesised expansion of the ancestors of the Khoi-San to southern Africa (Behar *et al.,* 2008).

The L0a'b'f'k branch bifurcates into a branch that contains mtDNA sequences belonging to haplogroup L0k and another root branch to haplogroups L0a, L0b and L0f when the PhyloTree classification system (Van Oven and Kayser, 2009) is used. According to its

positioning, the L0k clade developed independently from the other L0 haplogroups, which underlies the hypothesis that the L0d and L0k haplogroups developed in response to population migration and isolation events of the early ancestors of the Khoi-San populations of southern Africa (Behar *et al.,* 2008). The L0a'b'f root bifurcates into a branch that represents the L0f clade and the root to the L0a and L0b haplogroups according to the PhyloTree classification system (Van Oven and Kayser, 2009). The phylogenetic hierarchy of the L0 haplogroup of the All African MP tree is therefore in agreement with the popular view that the major haplogroup L0 gave rise to the independent development of the haplogroups L0d, L0k and L0a'b'f and the later independent development of the L0f haplogroup followed by the development of the L0a and L0b haplogroups from the L0a'b'f root (Behar *et al.*, 2008).

A major component (81%) of the Khoi-San individuals of the All African dataset is grouped within the L0d and L0k clades of the All African MP tree, as is also observed within the Global African MP and NJ trees, as well as the All African NJ tree. This is as expected, since it has been reported that haplogroups L0d and L0k contributed greatly to the contemporary Khoi-San genetic makeup (Chen *et al.*, 2000; Tishkoff *et al.*, 2007; Behar *et al.*, 2008). With the largest component of the Tswana individuals of this investigation present in these clades as well, it could be derived that the Khoi-San was a major ancestral genetic contributor to the Tswana population of this investigation.

Twenty-seven mtDNA sequences of the Tswana-speaking individuals in this investigation are grouped within the L0d clade as defined by the PhyloTree classification system (Van Oven and Kayser, 2009) and to the L0a and L0b clades by the Wallace classification system (2004). The Wallace classification system (2004) therefore indicates that the Tswana-speaking individuals of this investigation belongingto these clades originated from two different maternal ancestors of which the group belonging to the L0b clade contained a haplogroup-defining transition at np G9755A and the group belonging to the haplogroup L0a was defined by the absence of that transition. In contrast to this, the PhyloTree classification system (Van Oven and Kayser, 2009) assigns these mtDNA sequences to only one haplogroup, namely L0d. The L0d clade according to the PhyloTree classification system (Van Oven and Kayser, 2009) consists of one branch that represents the L0d3 clade and another branch that bifurcates into the L0d1 and L0d2 clades. The L0d1'2 root is defined by two transitions at np A3756G and np G9755A and the L0d3 haplogroup by a number of coding region polymorphisms, which do not include the transition at np G9755A. Therefore, the discrepancy between the two classification systems could be explained by

the fact that the mtDNA sequences that contain the transition at np G9755A are assigned to haplogroup L0b by the Wallace classification system (2004) and haplogroups L0d1 or L0d2 by the PhyloTree classification system (Van Oven and Kayser, 2009); and the mtDNA sequences in which the transition at np G9755A is absent, are assigned to haplogroup L0a by the Wallace classification system (2004) and to haplogroup L0d3 by the PhyloTree classification system (Van Oven and Kayser, 2009).

Twelve mtDNA sequences of the Tswana-speaking individuals of this investigation are positioned within the clade representing haplogroup L0d1 and consisted of three sub-clades; L0d1a, L0d1b and L0d1c. The individuals TS_4117 and TS_4037 were are positioned within the sub-clade L0d1c with four mtDNA sequences from Khoi-San individuals of South African origin, five mtDNA sequences of !Kung individuals from Namibia, nine mtDNA sequences of !Kung individuals from Angola and one mtDNA sequence of a Khoi-San individual of unknown African origin. The individuals TS_2075, TS_4083, TS_4027, TS_5085 , TS_3027 and TS_2097 are positioned within the L0d1b sub-clade with five !Kung individuals from Angola, one San individual from Namibia, one San individual from South Africa, one Sandawe individual from Tanzania and one Tswana-speaking individual of unknown African origin. The mtDNA sequences of TS_3505, TS_4075, TS_4056 and TS_5044 are grouped within the L0d1a sub-clade with three !Kung individuals from Angola, one !Kung individual from Namibia and one Zulu-speaking individual of unknown African origin. This clade therefore consists of 30 mtDNA sequences from Khoi-San individuals of different regions, a Tanzanian individual and two Bantu-speaking (Tswana and Zulu) individuals. The large Khoi-San component of this clade makes it clear that there is a strong Khoi-San maternal ancestry in the Tswana-speaking individuals of this investigation; genetic flow between the Khoi-San and the Bantu-speaking populations has also been observed in other studies.

The L0d2 clade consists of ten Tswana-speaking individuals of this investigation that group closely together with one San individual from South Africa, one !Kung individual from Angola, one Tswana-speaking and one Zulu-speaking individual of African origin and one individual of unknown African origin. This clade therefore contains a stronger component of Bantu-speaking individuals than the previous clade, in addition to the Khoi-San component that is also present and therefore reflects the genetic connection between the Khoi-San and the Bantu-speaking populations.

The L0d3 clade consists of five Tswana-speaking individuals of this investigation grouped with three mtDNA sequences of Sandawe individuals and one Burunge individual from Tanzania. The Burunge individual from Tanzania (EF184589) is assigned to haplogroup L2a by the Wallace classification system (2004), as is also observed in the Global African MP tree. The probable reason for the incorrect haplogroup assignment by the Wallace classification system (2004) is discussed in Section 6.8.2. This clade therefore consists of only Tswana-speaking individuals of this investigation, as well as Tanzanian individuals. This is in accordance with the study done by the Gonder *et al.* (2007), in which haplogroup L0d3 was identified within the Hadza and Sandawe ethnic groups in Tanzania, and haplogroups L0d1 and L0d2 were observed within the Khoi-San of southern Africa, which implied an ancient genetic connection between these two populations. The Khoi-San possibly developed in eastern Africa and contributed an early maternal lineage to the eastern African ethnic Hadza and Sandawe populations, which lost some of the genetic ancestry owing to genetic drift over time. It is uncertain, however, whether the haplogroup connection between the Tanzanian and Khoi-San populations was caused by a shared common ancestor or by genetic flow (Gonder *et al.*, 2007). The phylogenetic connection between the Tswana-speaking individuals of this investigation and the Tanzanian individuals suggests that genetic flow could have taken place between these two populations directly or indirectly via other populations, such as the Khoi-San, which also harbours this haplogroup.

The L0a'b'f'k root defining branch bifurcates into one branch that harbours the mtDNA sequences that are assigned to haplogroup L0k by the PhyloTree classification system (Van Oven and Kayser, 2009) and a root branch for the haplogroups L0a, L0b and L0f. The L0k clade contains only one mtDNA sequence belonging to a Tswana-speaking individual of this investigation (TS_4034) and six !Kung individuals from Angola, two !Kung individuals from Namibia, one !Kung individual from South Africa, two San individuals from South Africa and two San individuals of African origin. The positioning of the L0k clade in the All African MP tree indicates an earlier and independent development from the L0a, L0b and L0f haplogroups, as was also reported in the literature (Behar *et al.*, 2008). It also represents a clade of Khoi-San speaking individuals and therefore contains a gene pool of matrilineal Khoi-San ancestry, as reported in literature (Behar *et al.*, 2008) and the presence of the Tswana-speaking individual of this investigation in this clade confirms the genetic connection with Khoi-San speaking populations.

The positioning of the mtDNA sequence of one Tswana-speaking individual of this investigation (TS_5063), singularly on a long branch connected to the L0a'b root, is also observed in the Global African NJ and MP tree and All African NJ tree. This sequence is assigned to haplogroup L0a'b by the PhyloTree classification system (Van Oven and Kayser, 2009) because of the haplogroup-defining polymorphisms for the haplogroup L0b if the PhyloTree classification system is preliminary and can therefore not be used to assign the haplogroup in this study. However, as discussed in Sections 6.8.1 and 6.8.2, the possibility that this sequence represents haplogroup L0b or an undefined haplogroup cannot be ruled out.

The other branch of the L0a'b root splits into two groups, the L0a1 and L0a2 clades, when using the PhyloTree classification system (Van Oven and Kayser, 2009), and haplogroup L0c when using the Wallace classification system (2004). The L0a1 clade contains three mtDNA sequences of Tswana-speaking individuals of this investigation and four mtDNA sequences of individuals from Uganda, one mtDNA sequence from an individual from Morocco, two mtDNA sequences from individuals from Sudan, one mtDNA sequence from an individual from Burkina Faso, one mtDNA sequence from an individual from Tanzania and three mtDNA sequences from individuals of unknown African origin. The L0a1 clade therefore consists of individuals from a broad range of African regions, as was the case in the other phylogenetic trees of this investigation, and confirms the previously suggested genetic flow between the Tswana-speaking individuals of this investigation and Bantu-speaking individuals. The broad range of African individuals that harbour haplogroup L0a1 could be indicative of an expansion of the Bantu-speaking populations into eastern Africa within the last few thousand years (Behar *et al.*, 2008). The mtDNA sequences of the Tswana-speaking individuals of this investigation within this clade are not grouped as closely together as was the case in the Global African MP tree, indicating that these sequences do not share as many unique polymorphisms as observed in the Bantu-speaking individuals and as suggested by the Global African MP tree. This suggests that the Tswana population under investigation developed from an extensive admixture with the Bantu-speaking populations that reached southern Africa rather than a single Bantu matrilineal lineage that expanded within the Tswana population in response to population growth.

The L0a2 clade as assigned by the PhyloTree classification system (Van Oven and Kayser, 2009) consists of four mtDNA sequences belonging to Tswana-speaking individuals of this investigation and three mtDNA sequences of individuals from Uganda,

16 mtDNA sequences of Pygmy individuals and two mtDNA sequences of individuals of African origin. Pygmy individuals make up 64% of this clade, which is in agreement with the reported high incidence of haplogroup L0a2 within the Eastern Pygmy populations (Salas *et al.*, 2002; Gonder *et al.*, 2007; Behar *et al.*, 2008; Quintana-Murci *et al.*, 2008; Batini *et al.*, 2011). The mtDNA sequences EF184608 and EF184607 are assigned to haplogroup L2a by the Wallace classification system (2004) and therefore demonstrate a discrepancy with their positioning within the L0a2 clade of the PhyloTree classification system (Van Oven and Kayser, 2009). The discrepancy is resolved in Section 6.8.1 and reasons for the incorrect assignment of haplogroup L2 to sequences that belong to haplogroup L0 have been discussed earlier in this section. Haplogroup L0a2 is associated with the 9 bp deletion in the COII/tRNA^{Lys} intergenic region, which has been reported as a marker of the Bantu migration from western Africa to eastern and southern Africa (Salas *et al.*, 2002) and the high incidence of the Pygmy within this clade therefore does not necessarily suggest that the Tswana-speaking individuals of this investigation harbour haplogroup L0a2 because of direct genetic flow with the Pygmies. As the Tswana-speaking individuals of this investigation also display other markers of the Bantu migration, it has been concluded that it is more likely that haplogroup L0a2 observed in the Tswana population under investigation was contributed by Bantu-speaking individuals via the migration to southern Africa.

Haplogroup L1 forms the second major group that branched off from the L0'1'5 root and is therefore positioned as a sister branch to the L0 haplogroup. The tree topology of haplogroup L1 is identical to the tree topology observed in the Global African MP tree. The L1 haplogroup consists of two major clades, i.e. the L1b and L1c clades, when the PhyloTree classification system (Van Oven and Kayser, 2009) is used and the L1a and L1b clades when the Wallace classification system (2004) is used. The L1c clade consists of two groups, of which one comprises mtDNA sequences that indicate a discrepancy between their haplogroup assignment according to the Wallace classification system (2004) and their position in the All African MP tree. This clade consists of three mtDNA sequences, UG_CGR_71676, AF381992, EF184612, which do not present with the Wallace classification system (2004) haplogroup L1 defining transition at np A7055G and is therefore assigned to haplogroup L0. The reason for the incorrect haplogroup assignment in these three mtDNA sequences could be a possible reverse mutation event at that nucleotide position. This possibility is affirmed with the use of the PhyloTree classification system (Van Oven and Kayser, 2009), which assigns these sequences to haplogroup L1c3. The PhyloTree classification system (Van Oven and Kayser, 2009)

refers to np 7055 as a site where reverse mutation towards a base identical by state to the rCRS has been observed in individuals belonging to haplogroup L1c3, as observed in this study. The phylogenetic positioning of these three mtDNA sequences in a clade and the subsequent re-evaluation of the haplogroup assignment by using the PhyloTree classification system (Van Oven and Kayser, 2009) clarifies the discrepancy observed between the haplogroup assignment of these mtDNA sequences and their positioning within the All African MP tree.

The other group of the haplogroup L1c clade consists of two sub-clades. Two Tswana-speaking individuals from this investigation (TS_2093 and TS_3015) are grouped within one of the L1c sub-clades and are assigned to haplogroup L1b1 by the Wallace classification system (2004) and to haplogroup L1c2 by the PhyloTree classification system (Van Oven and Kayser, 2009) and share ancestry with a clade that contains mtDNA sequences belonging to haplogroup L1b2 according to the Wallace classification system (2004) and to haplogroup L1c4 according to the PhyloTree classification system (Van Oven and Kayser, 2009). The second major clade of this grouping consists of mtDNA sequences that are assigned to haplogroup L1b2 by the Wallace classification system (2004) and to haplogroup L1c1 when using the PhyloTree classification system (Van Oven and Kayser, 2009). The PhyloTree classification system (Van Oven and Kayser, 2009) therefore provides a more resolved haplogroup assignment to the mtDNA sequences than the Wallace classification system (2004) and this also underlies the phylogenetic positioning of the clades of haplogroup L1 more accurately than the Wallace classification system (2004). The mtDNA sequences of TS_2093 and TS_3015 are positioned as neighbours in the clade, which confirms the close matrilineal ancestry between these two mtDNA sequences, which is also observed in the Global African MP tree. Only 4% of the Tswana-speaking individuals of this investigation belongs to haplogroup L1c2, which reportedly originated from Bantu ancestors (Batini *et al.*, 2007).

The All African dataset contains 51% fewer mtDNA sequences than the Global African dataset, which reflects the large number of mtDNA sequences of individuals residing in non-African countries who have not been included in the construction of the All African MP tree. The tree topology of haplogroup L2 is in agreement with the Global African MP tree. The L2 clade consists of two branches, of which one branch only consists of two mtDNA sequences, which are assigned to haplogroup L2e by the PhyloTree classification system (Van Oven and Kayser, 2009) and to L2d by the Wallace classification system (2004). The basal positioning of the L2e clade is also observed in the Global African MP tree,

confirming that this clade developed earlier and independently from the other L2 clades. The other branch contains the sub-clades of the L2 clade, which consists of the haplogroups L2a, L2b, L2c and L2d as classified by both classification systems used in this investigation. The L2a'b'c'd root branch bifurcates into two major groups. One group consists of the sub-clades of haplogroup L2a and the other group of a branch that represents the L2d haplogroup and another branch that bifurcates into the L2b and L2c clades.

Ten Tswana-speaking individuals of this investigation are grouped in the L2a clade of the All African MP tree. The mtDNA sequences of TS_5066 and TS_2103 are positioned within a sub-clade that contains mtDNA sequences belonging to haplogroup L2a1a when the PhyloTree classification system (Van Oven and Kayser, 2009) is used and they group with seven mtDNA sequences belonging to individuals of African origin and one mtDNA sequence belonging to an individual from Nigeria. The mtDNA sequences of TS_4013 and TS_4106 are positioned within another sub-clade that contains mtDNA sequences belonging to haplogroup L2a1c when the PhyloTree classification system (Van Oven and Kayser, 2009) is used. It further consists of four mtDNA sequences belonging to individuals from Burkina Faso, one mtDNA sequence belonging to an individual from Tanzania, one mtDNA sequence belonging to an individual from Uganda and three mtDNA sequences from individuals of African origin. The mtDNA sequences of TS_2077 and TS_5086 are assigned to haplogroup L2a1f by the PhyloTree classification system (Van Oven and Kayser, 2009) and are positioned within a sub-clade consisting of two mtDNA sequences belongingto individuals from Burkina Faso and four mtDNA sequences belonging to individuals of African origin. Individual samples TS_3107, TS_5062, TS_3466 and TS_5060 are positioned in a sub-clade that represents haplogroup L2a1b when using the PhyloTree classification system (Van Oven and Kayser, 2009) and consists of one mtDNA sequence belongingto a Pedi-speaking individual of African origin, one mtDNA sequence belongingto a Khoi-San individual, three mtDNA sequences belongingto individuals from Tanzania, one mtDNA sequence belonging to an individual from Ethiopia, one mtDNA sequence belonging to an individual from Uganda and five mtDNA sequences belonging to individuals of African origin.

Haplogroup L2a has been described as a common haplogroup of Africa, which originated between east and west Africa and dispersed along the Sahel corridor in eastern and western directions. It is regarded as a marker of the Bantu expansion (Salas *et al.*, 2002). The L2a haplogroup has been reported to be well represented in the southeastern Bantu

speakers (Salas *et al.*, 2002) and it was therefore not surprising to observe 20% of the Tswana-speaking individuals of this investigation within this clade. The presence of the L2a haplogroup within the Tswana-speaking individuals of this investigation is evidence of genetic flow with the Bantu-speaking populations that reached southern Africa via the major Bantu expansion.

The major haplogroup L3 consists of clades for haplogroups L3a, L3b, L3c, L3d, L3e, L3f, L3l'k and L3h when the PhyloTree classification system (Van Oven and Kayser, 2009) is used. The Wallace classification system (2004) only resolves haplogroup L3 to the level of haplogroups L3a, L3b and L3c. The L3 clade hierarchy of the All African MP tree consists of one major group belongingto haplogroup L3a and another group mainly belongingto haplogroup L3b and a sub-clade that belongs to haplogroup L3c when the Wallace classification system (2004) is used. The same hierarchy is more detailed when the PhyloTree classification system (Van Oven and Kayser, 2009) is used. It agrees with the published haplogroup hierarchies (Kivisild *et al.*, 2006; Torroni *et al.*, 2006; Gonder *et al.*, 2007; Behar *et al.*, 2008) and is therefore used for the interpretation of the L3 clade in the All African MP tree, as in the case of the Global African MP tree.

The L3h haplogroup clade is the deepest branch of the haplogroup L3 hierarchy of the All African MP tree as opposed to haplogroup L3f that is observed in the most basal position of the Global African MP tree. This branch has a bootstrap value of 51% in the All African MP tree, which is sufficient to accept the branch with confidence. The PhyloTree classification system (Van Oven and Kayser, 2009) indicates haplogroup L3h in a basal position and provides some confidence to the hierarchy observed in the All African MP tree, which is in opposition to the Global African MP tree that indicates the L3f clade as the basal haplogroup. In contrast to the MP trees of this investigation, the Global African NJ tree and All African NJ tree do not indicate haplogroups L3f or L3h in a basal position but rather the branch that represents the H, pre-V and U haplogroups. Except for the differing tree-building methods, as discussed in depth in Section 6.8.2, the differences in the tree topology between the Global African MP tree and the All African MP tree could possibly also be ascribed to the fact that 65 mtDNA sequences were removed from the Global African dataset to compile the All African dataset and resulted in 40% fewer mtDNA sequences that were taken into account in the determination of the evolutionary positioning of the mtDNA sequences.

The sister branch to the L3h clade represents a root for the other clades of haplogroup L3 and bifurcates into a branch that represents haplogroup L3a and haplogroups H, pre-HV and U and a branch that represents a root to the L3b'c'd'e'f'l'k clade. The All African MP tree therefore positions the L3a haplogroup mtDNA sequences phylogenetically separate from the other L3 clades, as has also been observed in the Global African MP tree. The L3b'c'd'e'f'l'k root bifurcates into two major groupings within the L3 clade. One group consists of a branch representing the haplogroup L3f mtDNA sequences and a sister branch representing the root to the haplogroups L3b, L3c and L3d. This root bifurcates into a branch that represents the haplogroup L3d and another branch that splits into the L3b and L3c clades. The All African MP tree therefore indicates shared ancestry between the last-named two haplogroups with the L3d clade, which does not agree with the observed shared ancestry between L3b and L3d in the Global African MP tree. Reasons for a discrepancy of this kind have been discussed earlier in this section.

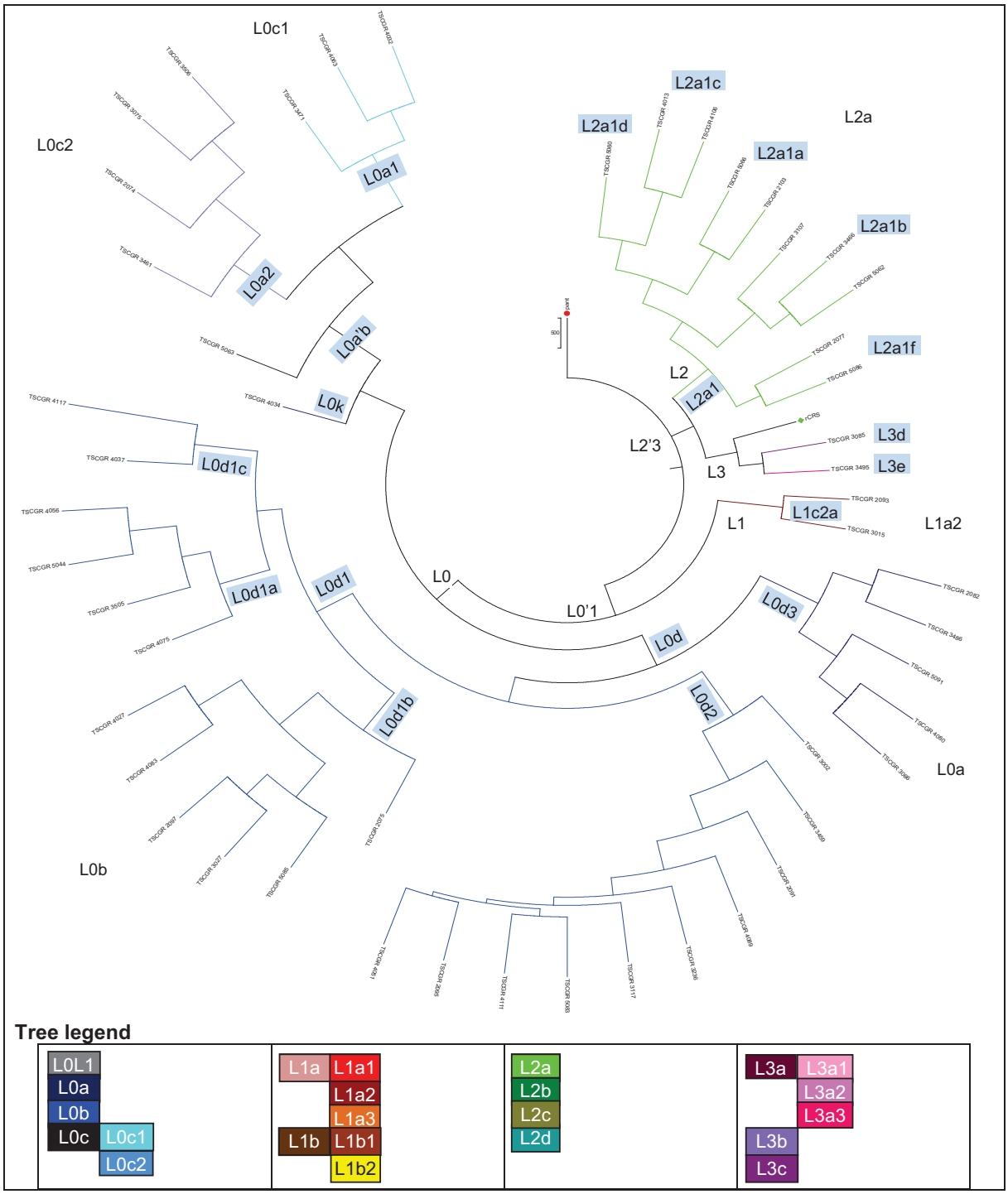
Two Tswana-speaking individuals of this investigation are grouped in the L3 clade. The individual TS_3085 is positioned in the L3d1 clade with three mtDNA sequences belonging to individuals from Tanzania, one mtDNA sequence belonging to an individual from Nigeria, two mtDNA sequences belonging to individuals from Burkina Faso, one mtDNA sequence belonging to an individual from Sudan and three mtDNA sequences belonging to individuals of African origin. Although haplogroup L3d is common in west Africa, haplogroup L3d1 is well represented in southeastern Africa, probably because of a later development in the Bantu populations that reached the southern regions through the major Bantu expansions from west Africa (Salas *et al.*, 2002; Salas *et al.*, 2004). The presence of the Tswana-speaking individual of this investigation within this clade therefore reaffirms the genetic connection between the Tswana population under investigation with the Bantu-speaking populations of the early migrations to the southern regions of Africa.

Individual TS_3495 is assigned to haplogroup L3e1 and positioned in the L3e clade with one mtDNA sequence belonging to an individual from Cameroon and five mtDNA sequences from individuals of African origin. Haplogroup L3e1 is common in southeastern African Bantu-speakers although it appears to have a west central African origin and it is postulated that Cameroon is the source population (Salas *et al.*, 2002). The positioning of the mtDNA sequence from an individual from Cameroon was therefore expected in this clade. The positioning of the Tswana-speaking individual of this investigation in this clade also affirms the genetic flow between the Bantu-speaking populations that reached the southern regions through the early major Bantu expansions.

6.8.5 NJ tree of the Tswana dataset

The NJ outtree of the Tswana dataset was generated by using PHYLIP 3.6 and was converted to a phylogram format in TreeView (Page, 1996). It was rooted by using an outgroup, the *Pan troglodytes* sequence. Thereafter it was viewed by using MEGA version 5 (Tamura *et al.*, 2007) in a circular format. The PHYLIP outtree format converted the bootstrap values of the Tswana NJ tree into branch lengths, which are presented in Appendix F with the bootstrap values included and in Figure 6.76 without the bootstrap values.

Figure 6.76 NJ tree of the Tswana dataset



The Tswana NJ tree represents the coding regions of 50 mtDNA sequences of the Tswana dataset as described in Section 5.11.3 according to the NJ tree construction method described in Section 6.8. The rCRS is indicated by a green diamond (♦) and the *Pan troglodytes* outgroup is represented by a red circle (●). The haplogroups and sub-haplogroups assigned by the Wallace classification system (2004) are presented in colours as described in the legend. Major haplogroup branching is indicated at the branch splits on the tree and sub-haplogroup groupings are indicated on the outside of the tree. The PhyloTree (Van Oven and Kayser, 2009) haplogroup assignments are indicated in blue text boxes.

The Tswana NJ tree was constructed by using the Tswana dataset 3b, which consists of the mtDNA sequences of 50 Tswana-speaking individuals of this investigation. The Global African NJ tree and All African NJ tree present the mtDNA sequences of the Tswana population of this investigation in the context of a broad range of African and non-African L

haplogroups in order to determine the ancestral relationships with other populations. The purpose of this NJ tree, however, is to determine the phylogenetic relationships between the individuals within the Tswana population under investigation as a mechanism to visualise the phylogenetic distances between the individuals within the population.

The haplogroups of the Tswana-speaking individuals of this investigation that have been assigned by the Wallace classification system (2004) are indicated on the Tswana NJ tree by colour-coding the branches as described in the tree legend of Figure 6.76, as well as by indicating the major haplogroups at the branch splits and the sub-haplogroups on the outside of the tree in black. The haplogroups assigned to the Tswana-speaking individuals of this investigation by the PhyloTree classification system (Van Oven and Kayser, 2009) are indicated on the tree in blue text boxes. A concise list of the haplogroups of all the Tswana-speaking individuals of this investigation is presented in Appendix A.

The Tswana dataset was not intended to be representative of the haplogroup L but only representative of the population under investigation and it was therefore anticipated that the tree topology could differ from the other phylogenetic hierarchies observed in this study and in the literature. The Tswana dataset consists of only 50 mtDNA sequences from only one population residing in one region, as opposed to 573 mtDNA sequences of the Global African dataset representing individuals residing in African and non-African countries and the 390 mtDNA sequences representing individuals of broad African origin. The hierarchy of the major haplogroups of the Tswana NJ tree is positioned in two groups. One group consists of a branch that bifurcates into two sister clades representing the major haplogroups L0 and L1 respectively. The other group consists of a branch that also bifurcates into two sister clades representing the major haplogroups L2 and L3.

Studies have reported that the mtDNA phylogeny of major haplogroup L consists of two major branches, L0 and L1’2’3’4’5’6, which are located on opposite sides of the root and that haplogroup L0 is widely regarded as the most ancient lineage, which does not share ancestry with haplogroup L1 (Kivisild *et al.*, 2006; Torroni *et al.*, 2006; Gonder *et al.*, 2007; Behar *et al.*, 2008). The NJ and MP trees of this investigation, however, position the L0 and L1 haplogroups as sister clades, as also observed in the Tswana NJ tree. Reasons for this phylogenetic hierarchy are discussed in Section 6.8.1 and Section 6.8.2 and are also valid reasons for the phylogenetic hierarchy observed in the Tswana NJ tree. In summary, a probable alternative tree topology that has been published, could be ascribed to the NJ tree-building method that clusters sequences stepwise and therefore builds the tree

topology based on the order in which the sequences have been processed and also relies on a greedy search method that uses local exploration principles that explore only a small set of possible phylogenies to identify the final tree topology. The tree topology of phylogenetic trees could also have been affected by the number and type of characters used to construct the trees (Brocchieri, 2001) and the composition of the datasets used to determine the phylogenetic relationships (Cummings and Meyer, 2005). As discussed, the use of only the coding regions in this investigation as opposed to the use of the full mtDNA genome as in other studies, could have had an impact on the determination of genetic distances between sequence pairs because the different sequence regions within the mtDNA genome contain nucleotides that have evolved at different rates and therefore carry different levels of phylogenetic information, leading to different phylogenetic relationship determinations. The relatively small size of the Tswana dataset and the limited representation of the full range of L haplogroup variants especially could also have caused bias in the distance determinations and therefore tree topology.

The shared ancestry indicated by the positioning of haplogroup L2 and haplogroup L3 as sister clades in the Tswana NJ tree, is different from the basal positioning of haplogroup L2 as observed in both the Global African NJ tree and the All African NJ tree. The phylogenetic hierarchy observed in the Tswana NJ tree is, however, shared with the Global African MP tree and All African MP, tree as well as with the phylogenetic hierarchy of the PhyloTree classification system (Van Oven and Kayser, 2009), which indicates confidence in this grouping. The small size and restricted representation of the Tswana NJ dataset in comparison to the Global African and All African datasets could have been a likely reason for this observation. Opinions regarding the impact of the sample size used in phylogenetic molecular studies on the outcome of the proposed evolutionary history have been controversial. Saitou and Nei (1987) reported that a small sample was adequate on condition that a sufficient number of characters were used, as opposed to Excoffier (1994) that reported the opposite.

The bootstrap values for the major haplogroups L0, L1, L2 and L3 of the Tswana NJ tree are 46%, 99%, 73% and 49% respectively. The branches of the L0 and L3 clades display low bootstrap values that do not provide sufficient confidence in their positioning to be accepted as precise. Inaccuracy of the positioning of these branches, however, cannot be implied because bootstrap values only provide confidence in the preciseness of the positioning of a branch by the tree-building method on repeated analyses and not the accuracy of phylogenetic relationships, as discussed in Section 6.8.1.

Haplogroup L0 consists of two major groups in the Tswana NJ tree. One group represents haplogroup L0d when the PhyloTree classification system (Van Oven and Kayser, 2009) is used and haplogroups L0a and L0b when the Wallace classification system (2004) is used. The other group represents haplogroups L0k, L0a'b and L0a when the PhyloTree classification system (Van Oven and Kayser, 2009) is used and haplogroup L0c when the Wallace classification system (2004) is used. This phylogenetic hierarchy is also observed in the other phylogenetic trees of this investigation except in the Global African NJ tree. Reasons for the difference in the tree topology have been discussed extensively in Section 6.8.3. For reasons explained in Section 6.8.1 and 6.8.3, the PhyloTree classification system (Van Oven and Kayser, 2009) has been used for the discussion of the phylogenetic tree topology in the Tswana NJ tree.

The branch containing mtDNA sequences of the Tswana-speaking individuals of this investigation that are assigned to haplogroup L0a by the PhyloTree classification system (Van Oven and Kayser, 2009) and to haplogroup L0c by the Wallace classification system (2004), consists of two sub-branches that contain mtDNA sequences belonging to haplogroup L0a1 and L0a2 and L0c1 and L0c2 respectively. The L0a1 clade according to the PhyloTree system and L0c1 clade according to the Wallace system, contains three mtDNA sequences and is defined by the haplogroup-defining transition at np T5096C by both classification systems. The L0a2 clade according to the PhyloTree system and L0c2 clade according to the Wallace system consists of four mtDNA sequences and is defined by a haplogroup-defining transition at np G5147A that is shared between both classification systems. In addition, the PhyloTree classification system (Van Oven and Kayser, 2009) defines five distinct coding region polymorphisms that are present in the Tswana-speaking individuals of this investigation that are grouped in this clade. The individuals of the two clades therefore exhibit different haplogroup-defining polymorphisms that have been reported in the literature for haplogroups L01a and L02a and that suggest origins from different geographic regions and a separate development (Salas *et al.*, 2002; Kivisild *et al.*, 2004). This suggests that the Tswana population of this investigation exhibits signs of gene flow between different Bantu-speaking groups. The presence of the L0a2 haplogroup, which is regarded as a marker of the Bantu migration in 14% of the Tswana-speaking individuals of this investigation, is also indicative of the connection between the population under investigation and Bantu-speaking individuals that reached southern Africa through the major early Bantu expansion.

The L0a clade according to the PhyloTree classification system (Van Oven and Kayser, 2009) forms a sister clade with one mtDNA sequence, TS_5063, which cannot be resolved by either of the haplogroup classification systems used in this investigation. TS_5063 lacks the haplogroup-defining transition at np G5147A to be assigned to haplogroup L0c2 by the Wallace classification system (2004), which could be ascribed to a reverse mutation event. The possibility of this mtDNA sequence belonging to haplogroup L0b when the PhyloTree classification system (Van Oven and Kayser, 2009) is used has been discussed in Section 6.8.1. It was concluded that this individual displays polymorphisms that are not present in the mtDNA sequences of the individuals of the L0a clade and therefore that it suggests genetic flow with a Bantu-speaking population that originated and developed separately from the populations harbouring haplogroup L0a.

The L0a'b root branch shares common ancestry with a branch containing one mtDNA sequence of a Tswana-speaking individual of this investigation, TS_4034, and belongs to haplogroup L0k. The basal position of this mtDNA sequence within the L0 clade indicates an ancient origin and development separate from the other major group in the L0 clade. Haplogroup L0k has been associated with the Khoi-San populations of southern Africa (Behar *et al.*, 2008) and the presence of this haplogroup in a Tswana-speaking individual indicates genetic flow between the Tswana population under investigation and the Khoi-San.

The other major group within the L0 clade consists of mtDNA sequences belonging to haplogroup L0d when the PhyloTree classification system (Van Oven and Kayser, 2009) is used and to haplogroup L0a and L0b when the Wallace classification system (2004) is used. As discussed in the previous Section 6.8.4, the mtDNA sequences that contain the transition at np G9755A are assigned to haplogroup L0a by the Wallace classification system (2004) and to haplogroup L0d3 by the PhyloTree classification system (Van Oven and Kayser, 2009) and the mtDNA sequences that do not contain the transition np G9755A are assigned to haplogroup L0b by the Wallace classification system (2004) and to haplogroup L0d1 and L0d2 by the PhyloTree classification system (Van Oven and Kayser, 2009). As in the previous sections, the PhyloTree classification system (Van Oven and Kayser, 2009) demonstrates a higher level of resolution in the haplogroup assignment of the mtDNA sequences and is therefore used to discuss the L0d clade of the Tswana NJ tree.

Haplogroup L0d is mainly associated with the Khoi-San populations of Africa as seen in several studies (Salas *et al.*, 2002; Kivisild *et al.*, 2004, Gonder *et al.*, 2007; Behar *et al.*, 2008; Coelho *et al.*, 2009) and the presence of 54% of the Tswana population of this investigation in the L0d clade indicates that this Khoi-San lineage is well established in this population. Three distinct clades of the L0d haplogroup are observed in the Tswana-speaking individuals of this investigation, namely haplogroups L0d1, L0d2 and L0d3. The L0d branch bifurcates into two branches, a branch that represents haplogroup L0d3 and a root branch for haplogroups L0d1 and L0d2.

The L0d1 branch splits into a clade that represents the mtDNA sequences assigned to haplogroup L0d1b and a root branch that bifurcates into the L0d1a and L0d1c clades. The L0d1a clade consists of four mtDNA sequences and the L0d1c clade of two mtDNA sequences and they share a common maternal ancestor that was not shared with the L0d1b clade, which consists of six mtDNA sequences. Each of these haplogroups is defined by the PhyloTree classification system (Van Oven and Kayser, 2009) via the use of distinct haplogroup-defining polymorphisms. The two Tswana individuals that are positioned in the L0d1c clade share the same haplogroup-defining polymorphisms, which indicates a single maternal ancestor. Two of the four Tswana individuals that are positioned within the L0d1a clade display the same haplogroup-defining polymorphisms, which differ from the haplogroup-defining polymorphisms of the other two mtDNA sequences belonging to the same clade, indicating two maternal ancestors from which these individuals originated. The positioning of one of the six mtDNA sequences (TS_2075) of the L0d1b clade indicates that it developed from a different maternal ancestor than the other mtDNA sequences within the clade. The other five mtDNA sequences are positioned in two separate groups, TS_4027 and TS_4083 in one grouping and TS_2097, TS_3027 and TS_5085 in another grouping. On further investigation into this phylogenetic positioning, it was determined that the mtDNA sequences of the L0d1b clade display all the PhyloTree classification system (Van Oven and Kayser, 2009) haplogroup L0d1b defining coding region polymorphisms. In addition, TS_2075 displays the haplogroup L0d1b1 defining coding region polymorphisms, which is the most likely reason for its positioning separately from the other Tswana individuals. The other five mtDNA sequences within this clade further display five coding region polymorphisms that have not been identified as haplogroup-defining by any of the classification systems used in this investigation. The transitions at np G6383A, np T9111C, np C10920T, np C12436T and np G11176A are present in TS_4027, TS_4083, TS_2097, TS_3027 and TS_5085, in addition to the haplogroup-defining polymorphisms specified by the PhyloTree

classification system (Van Oven and Kayser, 2009), hence the close phylogenetic relationship between these individuals. The identification of these additional polymorphisms within the Tswana-speaking individuals of this investigation indicates the existence of a possible undefined haplogroup that is present within the Khoi-San or Tswana-speaking populations of southern Africa. An additional transition at np G11176A is identified in TS_2097, TS_3027 and TS_5085, which is a probable reason for the close phylogenetic relationship displayed by these mtDNA sequences within the L0d1b clade. This transition has not been specified within the PhyloTree classification system (Van Oven and Kayser, 2009) or in published data as a haplogroup-defining polymorphism and also suggests the possibility of a new haplogroup.

The L0d2 clade consists of a branch that represents only one (1) mtDNA sequence (TS_3002) and another branch that represents nine mtDNA sequences, indicating that the majority of the sequences within this clade originate from a single maternal ancestor and that the single mtDNA sequence in the sister clade originates from a different maternal ancestor or has undergone private mutational events that place it on a phylogenetically separate branch. On further investigation of the sequence variation of these mtDNA sequences, it has been determined that the nine Tswana individuals represented on the one branch all belong to haplogroup L0d2a1 and contain all the PhyloTree classification system (Van Oven and Kayser, 2009) L0d2a1 haplogroup defining coding region polymorphisms, which are not present in TS_3002. MtDNA sequence TS_3002 further displays transitions at np T4023C, np G5231A, np C6038T, np G9139A, np G9801A, np G12925A, np T13488C, np T14094C and np G15884A, which are not displayed in the other nine mtDNA sequences, which therefore confirms that TS_3002 does not originate from the same maternal ancestor, but harbours a unique set of polymorphisms that is most likely not only caused by private mutational events but could possibly represent a haplogroup that has not been defined yet. This should be confirmed by the presence of similar polymorphisms in other mtDNA sequences.

The L0d3 clade consists of five (5) mtDNA sequences that all display the PhyloTree classification system (Van Oven and Kayser, 2009) haplogroup-defining polymorphisms for haplogroup L0d3. In addition to these polymorphisms, transitions are also identified at np C6170T, np G7119A, np G8290A and np T10114C and a transversion at np C10128A within all of these mtDNA sequences. These polymorphisms are also observed in an mtDNA sequence (EU092842) belonging to a Khoi individual from South Africa in a study performed by Behar *et al.* (2008) and are therefore not novel (Pereira *et al.*, 2009). The

presence of these polymorphisms in all the Tswana individuals of this investigation belonging to haplogroup L0d3 indicates that this lineage is well established in the Tswana population under investigation and could have been acquired through gene flow with the Khoi-San population of southern Africa. It also indicates the presence of a new sub-haplogroup that has not been identified by the current classification systems.

The phylogeny of haplogroup L0 of the Tswana NJ tree highlights the fact that the Tswana population under investigation consists of similar lineages to the Khoi-San-speaking populations. The Tswana NJ tree is in agreement with the PhyloTree classification system (Van Oven and Kayser, 2009), which identifies haplogroups L0d1, L0d2 and L0d3 with further sub-groups L0d1a, L0d1b and L0d1c and L0d2a, L0d2b and L0d2c. No further sub-groups have been specified for haplogroup L0d3. The Tswana population under investigation therefore represents all of the L0d haplogroups and many of the sub-groups, which suggests that this population originated from a diverse Khoi-San population that consisted of a range of haplogroup variants which would be characteristic of a well-established population and are not due to gene flow with single individuals only.

Haplogroup L1 is the sister clade to haplogroup L0 in the Tswana NJ tree and consists of two mtDNA sequences (TS_3015 and TS_2093) that are assigned to haplogroup L1c2a according to the PhyloTree classification system (Van Oven and Kayser, 2009) and to haplogroup L1a2 by the Wallace classification system (2004). Both of these mtDNA sequences display all of the PhyloTree classification system (Van Oven and Kayser, 2009) haplogroup L1c2a defining coding region polymorphisms. As described in Section 6.8.1, haplogroup L1c and sub-groups are characteristic of the western Pygmy populations and it has been suggested that these haplogroups originated from maternal ancestors to both Pygmy and Bantu-speaking populations and were carried to southern Africa by the early major Bantu expansion (Quintana-Murci *et al.*, 2008). Although haplogroup L1 consists of many different sub-groups, only one lineage is represented in the Tswana population under investigation, which suggests that only one maternal ancestor was involved in the introduction of this lineage to the Tswana population.

Ten Tswana-speaking individuals of this investigation belong to haplogroup L2a according to both classification systems used in this investigation. Haplogroup L2a has been connected to a pronounced population expansion that occurred before the major Bantu expansion, most likely due to favourable environmental changes resulting from the LGMthat caused arid conditions suitable for the migration of human populations (Atkinson

et al., 2009). It has been postulated that the high frequency of haplogroup L2a in Africa, especially in western and southeastern Africa, can be ascribed to this expansion, which led to haplogroup L2 being the most widespread and common haplogroup in Africa (Salas *et al.*, 2002). The presence of high frequencies of specifically haplogroups L2a1a and L2a2 in the southeast populations of Africa is associated with the sub-Saharan agricultural migration of the Bantu-speaking populations (Atkinson *et al.*, 2009; Rosa and Brehm, 2011)

The Wallace classification system (2004) defines haplogroup L2a by a transition at np G11914A, whereas the PhyloTree classification system (Van Oven and Kayser, 2009) defines the root to the haplogroups L2a1 and L2a2 by the same transition at np G11914A, in addition to five other coding region polymorphisms. The ten Tswana-speaking individuals of this investigation belonging to haplogroup L2 are further assigned to haplogroups L2a1a, L2a1b, L2a1c, L2a1d and L2a1f by the PhyloTree classification system (Van Oven and Kayser, 2009), which will be used for the purpose of discussion of the phylogenetic relationships within haplogroup L2 of the Tswana NJ tree because of the better resolution of the PhyloTree system.

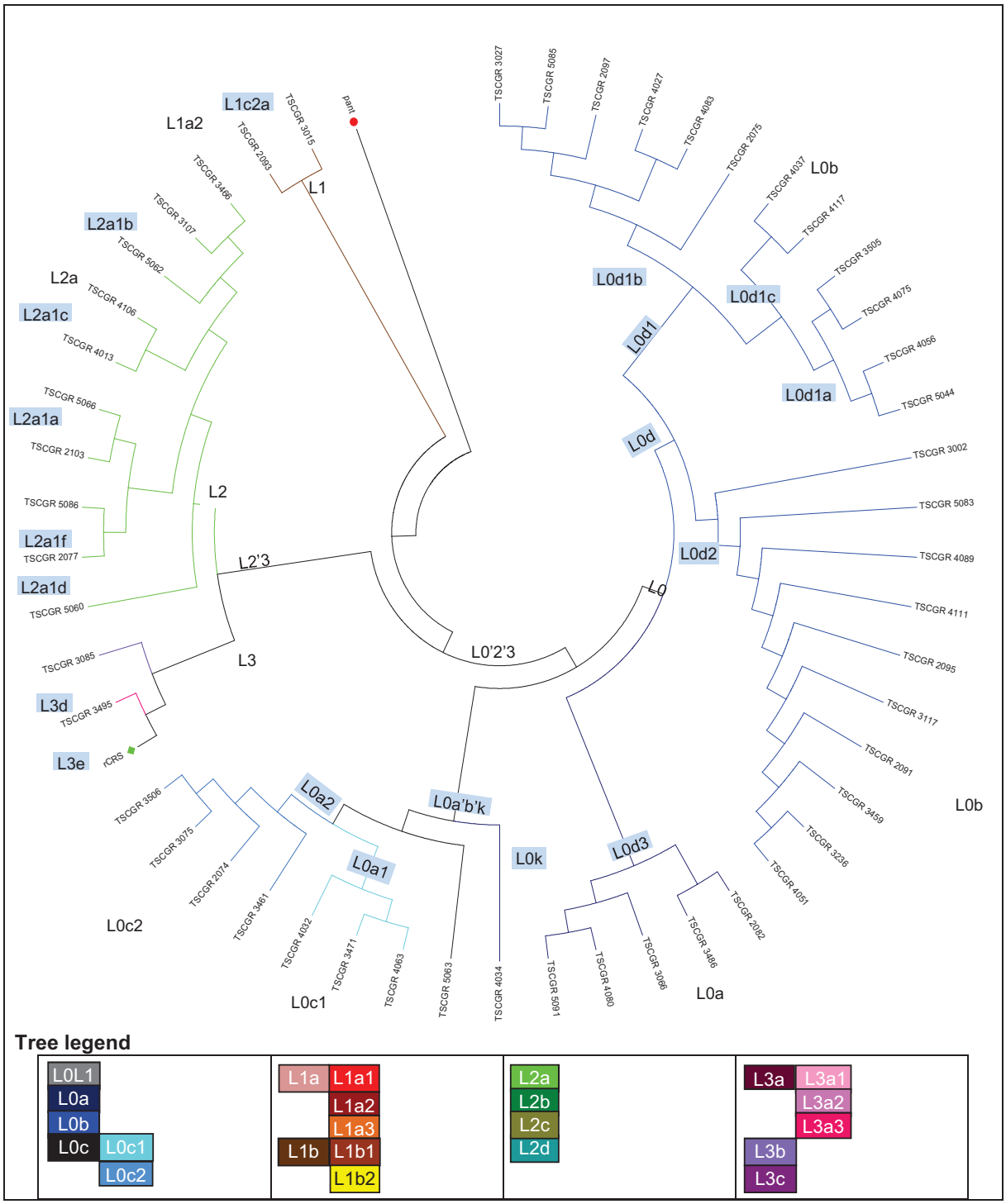
Haplogroup L2 branches into a clade that consists of two mtDNA sequences belonging to haplogroup L2a1f and a sister clade consisting of the root to the haplogroups L2a1a, L2a1b, L2a1c and L2a1d. The Tswana-speaking individuals of this investigation belongingto haplogroup L2a1f therefore carry a lineage that originated earlier than the other L2a lineages. The root branch of L2a1a'b'c'd bifurcates into root branches for haplogroups L2a1a, L2a1c and L2a1d and a clade that consists of three mtDNA sequences belongingto haplogroup L2a1b. The L2a1a'c'd root bifurcates into a clade that consists of two mtDNA sequences belonging to haplogroup L2a1a and a root branch that splits into a clade consisting of two mtDNA sequences belongingto haplogroup L2a1c and a clade consisting of one mtDNA sequence that belongs to haplogroup L2a1d. The clades were are defined by distinct coding region polymorphisms as specified by the PhyloTree classification system (Van Oven and Kayser, 2009), which are displayed in the mtDNA sequences of the respective clades, indicating the presence of several L2 lineages within the Tswana population under investigation. This is interpreted as an indication of a strong genetic component in the Tswana population of the Bantu-speaking population that was involved in a major expansion event in the LGM period and participated in the major westward and eastwards Bantu dispersals along the Sahel corridor.

Two mtDNA sequences, TS_3085 and TS_3495, are grouped within the major haplogroup L3 clade, which forms a sister clade with haplogroup L2 in the Tswana NJ tree, indicating a shared ancestry. TS_3085 is assigned to haplogroup L3b by the Wallace classification system (2004) by a transition at np G5147A and to haplogroup L3d1a1a by the PhyloTree classification system (Van Oven and Kayser, 2009) by the same transition at np G5147A and 14 other coding region polymorphisms. The L3d lineage has been reported as an important component of the South African maternal genetic pool originating in western Africa and reaching the southern African regions through a western route via Angola and through an eastern route via Tanzania (Coelho *et al.*, 2009; Rosa and Brehm, 2011). TS_3495 is assigned to haplogroup L3a3 by the Wallace classification system (2004) by a transition at np T6221C and to haplogroup L3e1 by the PhyloTree classification system (Van Oven and Kayser, 2009) by the same transition at np T6221C and four other coding region polymorphisms. The presence of a high frequency of haplogroup L3e1 in the southern Mozambican Bantu populations indicates that this lineage reached southern Africa through the Bantu migration via an eastern route (Salas *et al.*, 2002). The two Tswana-speaking individuals of this investigation that belong to haplogroup L3, however, harbour different L3 lineages. There are two possibilities through which the Tswana population under investigation could possibly have gained these haplogroups. It could have been through genetic flow with a Bantu-speaking population that reached southern Africa through an eastern route and carried both haplogroups, or it could have been due to genetic flow with separate Bantu-speaking populations that reached southern Africa via an eastern route (carrying haplogroup L3e), and via a western route (carrying the L3d haplogroup).

6.8.6 MP tree of the Tswana dataset

The MP tree of the Tswana dataset of this investigation was generated by using MEGA version 5 (Tamura *et al.*, 2007) and rooted by using the *Pan troglodytes* sequence as an outgroup. The original MP tree was viewed in MEGA version 5 (Tamura *et al.*, 2007) Tree Explorer in a circular phylogram format with bootstrap values indicated on the branches. The Tswana MP tree is presented in Appendix F with the bootstrap values included and in Figure 6.77 without the bootstrap values.

Figure 6.77 MP tree of the Tswana dataset



The Tswana MP tree represents the coding regions of 50 mtDNA sequences of the Tswana dataset as described in Section 5.11.3 according to the MP tree construction method described in Section 6.8. The rCRS is indicated by a green diamond (♦) and the *Pan troglodytes* outgroup is represented by a red circle (●). The haplogroups and sub-haplogroups assigned by the Wallace classification system (2004) are presented in colours as described in the legend. Major haplogroup branching is indicated at the branch splits on the tree and sub-haplogroup groupings are indicated on the outside of the tree. The PhyloTree (Van Oven and Kayser, 2009) haplogroup assignments are indicated in blue text boxes.

The Tswana MP tree was constructed by using the Tswana dataset 3b consisting of 50 mtDNA sequences from Tswana-speaking individuals included in this investigation. The construction of the Tswana MP tree had the same purpose as was discussed for the

Tswana NJ tree in Section 6.8.5 and in addition provided a phylogenetic tree constructed from the same Tswana dataset as the Tswana NJ tree. However, by using a different tree-building method it is possible to verify the phylogenetic positioning of the mtDNA sequences of the Tswana population under investigation.

The Wallace classification system (2004) and the PhyloTree classification system (Van Oven and Kayser, 2009) are used to assign haplogroups to the mtDNA sequences of the Tswana-speaking individuals under investigation. The haplogroups assigned by the Wallace classification system (2004) are indicated on the Tswana MP tree by colour-coding the branches as described in the tree legend of Figure 6.77. The major haplogroups assigned by the Wallace classification system (2004) and the PhyloTree classification system (Van Oven and Kayser, 2009) are indicated at the branch splits and the sub-haplogroups assigned by the Wallace classification system (2004) on the outside of the tree in black. To differentiate between the haplogroups assigned by the Wallace classification system (2004) and the PhyloTree classification system (Van Oven and Kayser, 2009) the haplogroups assigned by the PhyloTree classification system (Van Oven and Kayser, 2009) are indicated on the Tswana MP tree in blue text boxes. A concise list of the haplogroups of all the Tswana-speaking individuals of this investigation is presented in Appendix A.

Haplogroup L1 is positioned as the most basal haplogroup in the phylogenetic hierarchy of the Tswana MP tree and is the only phylogenetic tree of this investigation that depicts this phylogenetic structure. Haplogroup L1 is positioned as a sister clade to haplogroup L0 in the phylogenetic trees of the Global African and All African datasets as well as in the Tswana NJ tree of this investigation, providing strong support for a suggested shared ancestry between these two haplogroups within the datasets of this investigation. The sister clade to the haplogroup L1 clade of the Tswana MP tree consists of a root branch to haplogroups L0, L2 and L3. This branch bifurcates into a branch that represents the mtDNA sequences belonging to haplogroup L0 and a root L2’3 branch. The phylogenetic hierarchy of the Tswana MP tree therefore represent the L2 and L3 haplogroups as sharing ancestry and developing separately from the L0 haplogroup. The literature, however, provides evidence that supports a hypothesis that haplogroups L0 and L1-6 are sister clades, which originated in eastern Africa from where they expanded via modern human migrations to most of the African continent and developed into haplogroup L1 with a later development into haplogroups L2 and L3, which reached southeastern Africa and the rest of the world through major human migration via different migration routes (Ingman

et al., 2000; Salas *et al.*, 2002; Mishmar *et al.*, 2003; Kivisild *et al.*, 2006; Torroni *et al.*, 2006; Behar *et al.*, 2008). It is scientifically accepted that haplogroup L0 is the most ancient and therefore most phylogenetically basal haplogroup of modern humans (Ingman *et al.*, 2000; Salas *et al.*, 2002; Mishmar *et al.*, 2003; Kivisild *et al.*, 2006; Torroni *et al.*, 2006; Behar *et al.*, 2008), which is not what was observed in the phylogenetic hierarchy of the Tswana MP tree. Reasons for phylogenetic MP tree topologies that differ from what was expected have been discussed extensively in Section 6.8.4 and would also apply in this case. The MP tree-building method provides no guarantee for the selection of the most parsimonious tree from the large number of trees constructed, as the heuristic search method employed does not search exhaustively (Nei and Kumar, 2000). In addition, the Tswana dataset does not contain a representative sample of all the variants of the L haplogroup and it is assumed that the topology could differ from the published literature because it does not provide the same phylogenetic informative sites contained in the datasets used in the published studies. As stated in Section 6.8.5, however, the impact of sample size on the topology of phylogenetic trees has been controversial and care should be taken in making conclusions about small datasets (Saitou and Nei, 1987; Excoffier, 1994).

The bootstrap values for the haplogroups L0, L1, L2 and L3 are 87%, 100%, 99% and 81% respectively. These bootstrap values are higher than those observed in the Tswana NJ tree and provide strong confidence in the positioning of the branches.

The phylogenetic hierarchy of the haplogroup L0 in the Tswana MP tree agrees with the observed hierarchy for haplogroup L0 in the Global African MP tree, the All African MP and NJ trees and the Tswana NJ tree, as well as with the haplogroup hierarchy used in the PhyloTree classification system (Van Oven and Kayser, 2009). It consists of a group that contains the mtDNA sequences belonging to haplogroup L0d when the PhyloTree classification system (Van Oven and Kayser, 2009) is used and to haplogroups L0a and L0b when the Wallace classification system (2004) was is used. The sister clade to this group consists of the L0a'b'k root when the PhyloTree classification system (Van Oven and Kayser, 2009) is used and to a branch that represented haplogroup L0c when the Wallace classification system (2004) is used.

The L0a'b'k root branch as defined by the PhyloTree classification system (Van Oven and Kayser, 2009) bifurcates into a branch that consists of one mtDNA sequence (TS_4034) and a branch that contains a root to the L0a'b clade. The Wallace classification system

(2004) does not provide the resolution to distinguish between different haplogroups in these clades. The L0a'b root branch splits into a branch that contains only one mtDNA sequence (TS_5063) and a branch that represents the L0a1 and L0a2 clades when the PhyloTree classification system (Van Oven and Kayser, 2009) is used and into the L0c1 and L0c2 clades when the Wallace classification system (2004) is used. The L0a1 clade according to the PhyloTree classification system (Van Oven and Kayser, 2009) and the L0c1 clade according to the Wallace classification system (2004) contains the same three mtDNA sequences observed in the corresponding clade in the Tswana NJ tree and therefore it is also concluded that these Tswana-speaking individuals share the haplogroup-defining transition at np T5096C. The sister clade that contains mtDNA sequences belonging to the L0a2 clade when the PhyloTree classification system (Van Oven and Kayser, 2009) is used and to the L0c2 clade when the Wallace classification system (2004) is used contains the same four mtDNA sequences observed in the corresponding clade in the Tswana NJ tree and it is therefore concluded that these mtDNA sequences all contain the transition at np G5147A defined by both classification systems and the additional polymorphisms defined by the PhyloTree classification system (Van Oven and Kayser, 2009). This phylogenetic positioning therefore confirms that the mtDNA sequences belonging to the respective clades originated from different maternal ancestors that reportedly originated from different origins and developed separately (Salas *et al.*, 2002; Kivisild *et al.*, 2004) and that the Tswana population under investigation displays signs of gene flow between different Bantu-speaking groups. The presence of the L0a2 haplogroup in the Tswana population further indicates that it carries markers of the major Bantu expansion. The phylogenetic position of TS_5063 as a sister clade to the L0a clade in the Tswana NJ tree is confirmed in the Tswana MP tree. The same positioning is observed in all the other phylogenetic trees of this investigation and, as discussed in Section 6.8.5, contains unique polymorphisms that are not observed in the mtDNA sequences of the sister clade. It therefore confirms the inference made in Section 6.8.5, i.e. that it indicates genetic flow with a Bantu-speaking population that originated from a different maternal ancestor than what was observed for the sister clade.

The sequence of TS_4034 is positioned as a sister clade to the L0a'b clade and is assigned to haplogroup L0k by the PhyloTree classification system (Van Oven and Kayser, 2009). The Wallace classification system (2004) cannot resolve this mtDNA sequence any further. The basal position of this mtDNA sequence is also observed in the Tswana NJ tree as well as in the other phylogenetic trees of this investigation and therefore provides evidence of the presence of a maternal ancestor in the Tswana

population under investigation that harboured an ancient haplogroup. Haplogroup L0k has been associated with the Khoi-San populations of southern Africa (Behar *et al.*, 2008) and the suggested gene flow between the Tswana-speaking individuals of this investigation and the Khoi-San has therefore been confirmed.

The second major group within the L0 clade of the Tswana MP tree consists of mtDNA sequences belonging to the haplogroups L0d1, L0d2 and L0d3 when the PhyloTree classification system (Van Oven and Kayser, 2009) is used and to haplogroups L0a and L0b when the Wallace classification system (2004) is used. As discussed in Section 6.8.5, the Wallace classification system (2004) can only resolve the mtDNA sequences into two haplogroups, in contrast to the PhyloTree classification system (Van Oven and Kayser, 2009), which can define three different haplogroups. The PhyloTree classification system (Van Oven and Kayser, 2009) is therefore used for the discussion of the mtDNA sequences of this clade.

The phylogenetic hierarchy of the L0d clade of the Tswana MP tree is in agreement with the Tswana NJ tree. Both trees group the Tswana-speaking individuals of this investigation belonging to haplogroup L0d into two groups, one group that contains the mtDNA sequences belonging to haplogroup L0d3 and the other group that contains two branches that represent the haplogroups L0d1 and L0d2 respectively.

The L0d1 branch bifurcates into a branch that represents the mtDNA sequences assigned to haplogroup L0d1b and a branch that represents the root of haplogroups L0d1a and L0d1c. The L0d1a clade consists of four mtDNA sequences that are identical to those observed in the Tswana NJ tree and also display the haplogroup-defining polymorphisms specified by the PhyloTree classification system (Van Oven and Kayser, 2009) for haplogroup L0d1a. Two of the mtDNA sequences of this clade (TS_4056 and TS_5044), which are grouped together in a sub-clade, display the haplogroup L0d1a1 defining transitions at np T7389C and np T12696C, which are not present in the mtDNA sequences TS_4075 and TS_3505 of the same clade. The same phenomenon is observed in the Tswana NJ tree and therefore the presence of two different maternal ancestors is confirmed. The mtDNA sequences that belong to haplogroup L0d1c are positioned in a clade that shares ancestry with the L0d1a clade and contains two mtDNA sequences that agree with the corresponding clade in the Tswana NJ tree and therefore also contains the same set of haplogroup-defining polymorphisms indicating a single maternal ancestor. The L0d1b clade of the Tswana MP tree agrees with the phylogenetic structure of the

corresponding clade in the Tswana NJ tree and consistes of the same six mtDNA sequences. The mtDNA sequence TS_2075 is positioned separately from the other mtDNA sequences of this clade because of the presence of the haplogroup L0d1b1 defining polymorphisms that are not present in the other mtDNA sequences of this clade. This positioning is also observed in the Tswana NJ tree and therefore confirms that this individual contains evidence of genetic flow with a maternal ancestor that developed separately from the other maternal ancestors of the individuals in this clade. The other mtDNA sequences of this clade are positioned in two groups, TS_4027 and TS_4083 in one grouping and TS_2097, TS_3027 and TS_5085 in another grouping, as is also observed in the Tswana NJ tree and this confirms the close phylogenetic positioning of these sequences based on the transitions at np G6383A, np T9111C, np C10920T, np C12436T and np G11176A that have not been defined by the PhyloTree classification system (Van Oven and Kayser, 2009) and therefore indicates the possibility of an undefined haplogroup. The close phylogenetic relationship of the TS_2097, TS_3027 and TS_5085 individuals within the L0d1b clade has been confirmed in the Tswana MP tree and, as discussed in Section 6.8.5, is probably due to the unique transition at np G11176A observed within these three mtDNA sequences. The presence of the additional polymorphisms within the Tswana-speaking individuals of this investigation suggests that the Khoi-San or Bantu-speaking populations of southern Africa contained an unidentified haplogroup.

The sister clade to the L0d1a'b'c root branch consists of mtDNA sequences that belong to haplogroup L0d2. The phylogenetic hierarchy of the L0d2 clade is identical to the phylogenetic positioning of the correlating clade within the Tswana NJ tree, which provides strong support for the phylogenetic relationships between the sequences of this clade. The L0d2 clade consists of a branch that represents only one mtDNA sequence (TS_3002) and another branch that represents nine mtDNA sequences belonging to haplogroup L0d2a1. The mtDNA sequence of TS_3002 is positioned separately in the clade owing to the absence of the haplogroup L0d2a1 defining polymorphisms and because it displays nine unique polymorphisms that are absent in the other mtDNA sequences of this clade and are not specified in the PhyloTree classification system (Van Oven and Kayser, 2009). The similar phylogenetic positioning of TS_3002 in both Tswana trees provides support for the unique sequence variation that is observed and therefore the possibility that this individual could represent a maternal ancestor that carried an unidentified haplogroup.

The phylogenetic hierarchy of the L0d3 clade is in agreement with the phylogenetic hierarchy of the corresponding clade in the Tswana NJ tree and therefore consists of the same five mtDNA sequences observed in the Tswana NJ tree. Therefore the close phylogenetic relationship between these mtDNA sequences that contain the transitions at np C6170T, np G7119A, np G8290A and np T10114C and the transversion at np C10128A in addition to the L0d3 haplogroup-defining polymorphisms, is confirmed. As was discussed in Section 6.8.5, this connects the Tswana-speaking individuals of this investigation belonging to haplogroup L0d3 to the Khoi-San population based on the presence of identical polymorphisms in a Khoi individual from South Africa (Behar *et al.,* 2008).

The Tswana MP tree highlights the conclusions made from the phylogeny observed in the Tswana NJ tree with regard to the fact that 54% of the Tswana-speaking individuals of this investigation are represented in the L0d clade and that this is strong evidence for possible genetic flow between the Tswana-speaking population under investigation and the Khoi-San populations of South Africa. It also highlights the fact that these Tswana individuals harbour L0d haplogroups of different maternal ancestries, which suggests strong genetic flow between the Tswana and the Khoi-San populations. Based on the good Tswana representation of the different sub-haplogroups of haplogroup L0d, as described in the PhyloTree classification system (Van Oven and Kayser, 2009), it is concluded that the Khoi-San connected haplogroups are well established in the Tswana population.

Haplogroup L1 consists of two mtDNA sequences that are both assigned to haplogroup L1c2a by the PhyloTree classification system (Van Oven and Kayser, 2009) and to haplogroup L1a2 by the Wallace classification system (2004). Both sequences display the same PhyloTree classification system (Van Oven and Kayser, 2009) haplogroup-defining polymorphisms, which indicates a shared maternal ancestor. Haplogroup L1c has reportedly been carried by the Bantu migrations to the southern regions of Africa from the central African regions where it originated (Quintana-Murci *et al.*, 2008). The presence of haplogroup L1c in the Tswana-speaking individuals of this investigation provides evidence of genetic flow between the Tswana population under investigation and the Bantu speakers of central African origin. Haplogroup L1 is only present in two of the 50 individuals of the Tswana cohort included in this investigation, for which it has been determined that only one maternal ancestor can be identified. This implies that the genetic

flow between the Tswana population under investigation and the Bantu speakers that harboured this haplogroup was not strong and probably only involved single individuals.

The phylogenetic hierarchy of haplogroup L2 of the Tswana MP tree differs from the Tswana NJ tree. Ten Tswana-speaking individuals of this investigation are assigned to haplogroup L2a by both classification systems used in this investigation. Haplogroup L2a has been reported to have originated and expanded mainly before the major Bantu migration that carried this haplogroup in high frequencies to the other regions of Africa, making it the most widespread and common haplogroup in Africa (Salas *et al.*, 2002; Atkinson *et al.*, 2009). The presence of 20% of the Tswana-speaking individuals of this investigation in the L2a clade is therefore not unexpected and confirms the genetic flow between the Bantu-speaking populations and the Tswana population under investigation.

By using the PhyloTree classification system (Van Oven and Kayser, 2009), the mtDNA sequences within this clade can be resolved to the level of haplogroups L2a1a, L2a1b, L2a1c, L2a1d and L2a1f, which is not possible when using the Wallace classification system (2004). As in the case of the Tswana NJ tree, the PhyloTree classification system (Van Oven and Kayser, 2009) is therefore used for the discussion of the phylogenetic positioning of the Tswana-speaking individuals within the L2a1 clade.

The L2 branch bifurcates into a branch that represents one mtDNA sequence (TS_5060) belonging to haplogroup L2a1d and a root branch to haplogroups L2a1a, L2a1b, L2a1c and L2a1f. According to the Tswana MP tree, TS_5060 represents a maternal ancestor that developed earlier and separately from the other L2a1 haplogroups. The L2a1a'b'c'f root bifurcates into a branch that represents a root to the haplogroups L2a1a and L2a1f and a root to the haplogroups L2a1b and L2a1c. The mtDNA sequences all contain the haplogroup-defining polymorphisms specified by the PhyloTree classification system (Van Oven and Kayser, 2009) for the respective haplogroups, which indicates that the Tswana-speaking individuals of this investigation that belong to haplogroup L2a represent five maternal L2a lineages, which indicates genetic flow with a population in which haplogroup L2a1 was well established.

The shared ancestry between haplogroups L2 and L3 that is observed in the Tswana NJ tree has been confirmed in the Tswana MP tree. This observation is in agreement with the literature that reports that haplogroup L2 and haplogroup L3 are closely related in the context of the other African haplogroups and that evidence indicates that the haplogroups

L2 and L3 originated at similar times and underwent a major population expansion, from where they were dispersed to the rest of Africa through the Bantu migrations (Watson *et al.*, 1997; Salas *et al.*, 2002; Behar *et al.*, 2008). The L3 clade of the Tswana MP tree consists of two mtDNA sequences (TS_3085 and TS_3495) that belong to the haplogroups L3b and L3a3 when the Wallace classification system (2004) is used and to haplogroups L3d1a1a and L3e1 when the PhyloTree classification system (Van Oven and Kayser, 2009) is used. The shared haplogroup-defining polymorphisms between the two classification systems used in this investigation for the assignment of the haplogroups to the mtDNA sequences of the L3 clade have been discussed in Section 6.8.5. It is concluded that the two mtDNA sequences belonging to the L3 clade originated from two different maternal ancestors who reached southern Africa through the Bantu migration via a western and an eastern route. The presence of haplogroup L3 in only two Tswana-speaking individuals of this investigation indicates a weak maternal connection with the Bantu-speaking populations that harboured haplogroup L3 and sub-groups.

6.9 STATISTICAL ANALYSES

It is critical that the evolutionary processes that influence the mtDNA genome be integrated in the interpretation of the phylogenetic trees. Using statistical tests for genetic diversity and testing for patterns of selection in the mtDNA genome are essential for determining the effect of evolutionary processes on the mtDNA sequence under investigation and therefore in interpreting the phylogenetic tree structure (Ballard and Rand, 2005).

Statistical analyses were performed on the Global African dataset (1a), the All African dataset (2a) and the Tswana datasets (3a and 3b) as described in Section 5.11. A list of all the mtDNA coding region sequences used in these datasets is provided in Appendix B. The Global African dataset (1a) consists of the coding regions of 573 mtDNA genome sequences of African origin and was designed to represent individuals of African origin that not only resided in Africa but also on other continents, in order to represent a broad population of African individuals that were exposed to many different evolutionary influences. The All African dataset (2a) consists of the coding regions of 386 mtDNA genomes of individuals of African origin residing on the African continent only. This dataset is used to represent a broad population of African individuals that reside in Africa and have therefore been exposed to the evolutionary forces of population behaviour specific to African groupings only. In addition to these datasets, regional subsets of the All African

dataset are also used for statistical analysis. The Western African dataset (4), Eastern African dataset (5) and Southern African dataset (6) provide descriptive statistics of the regional groupings of the African populations that have been investigated phylogenetically and therefore assists in positioning the Tswana population genetically against an African backdrop. The composition of the regional datasets 4, 5 and 6 is presented in Table 6.41, while a comprehensive list of the datasets is provided in Appendix D.

Table 6.41 Composition of regional African mtDNA genome datasets of this investigation

Regional dataset	Number of mtDNA sequences	Countries represented
Western African dataset (4)	60	Burkina Faso, DRC, Nigeria, Cameroon and Ghana
Eastern African dataset (5)	110	Sudan, Tanzania, Ethiopia and Uganda
Southern African dataset (6)	66	South Africa, Namibia and Angola

Only coding regions of the mtDNA genomes are used in all the datasets. The different countries are not represented equally in the datasets. Different ethnicities of the respective countries are included. DRC = Democratic Republic of the Congo.

Datasets 3a and 3b represent a Tswana-speaking population of South Africa that was under investigation in this study. The investigation of a cohort of the Tswana-speaking individuals from South Africa is novel and the evolutionary history and genetic diversity of this population has been unreported. Statistical determinations to determine the genetic diversity and infer population behaviour have been interpreted in the context of the African populations of this investigation to estimate evolutionary history and relationships with other African populations.

Because of the high mutational rates present in the control region of the mtDNA genome, regions of the mtDNA that are prone to mutations suffer from recurrent mutational incidents, which hide true evolutionary events (Tamura and Nei, 1993; Pakendorf and Stoneking *et al.*, 2005). To prevent this phenomenon from affecting the genetic diversity determinations, only coding regions of the mtDNA genome were used in the statistical determinations of genetic diversity in this investigation. The statistical analysis was performed by the use of software programs as described in Section 5.14.

6.9.1 Nucleotide composition

Sequence variation within populations and between populations of the same and of different species relies on silent mutations that are fixed by a combination of selective forces and genetic drift (Knight *et al.*, 2001a). The most basic descriptive measure of sequence variation is determined by the nucleotide composition of a genome and is usually similar within species but not between species (Nikolaou and Almirantis, 2006). The nucleotide composition of the Tswana population under investigation was determined as a basic descriptive measure of the sequence variation that was observed within this novel population and to establish whether the nucleotide composition was similar to that of other human populations. The nucleotide composition of the complete mtDNA genomes of the Tswana dataset (3a) under investigation was determined by using MEGA version 5 (Tamura *et al.*, 2011) and the G+C content of the same dataset was determined by using the DnaSP version 5 (Librado and Rozas, 2009) software programs. The results are presented in Table 6.42.

Table 6.42 Nucleotide composition of the Tswana population of this investigation

Dataset	Total Nucleotide T%	Total Nucleotide C%	Total Nucleotide A%	Total Nucleotide G%	G+C content%
Tswana dataset (3a)	24.7	31.3	30.9	13.1	44.4

Tswana dataset (3a) = complete mtDNA genome sequences of 50 Tswana individuals. The nucleotide values and G+C content are calculated as percentages of the average number of total nucleotides and are averaged over the total number of samples.

The nucleotide composition of the Tswana population under investigation displays a higher percentage of A and C nucleotides in the full mtDNA genomes than G and T nucleotides, which is in agreement with the literature (Nikolaou and Almirantis, 2006). This observation can be explained by the fact that the nucleotide composition of the mitochondrial genome has been reported to be highly variable and does not adhere to Chargaff's Parity Rule type 2 (PR2), which states that within each DNA strand A=T and C=G, under the condition that bias between two DNA strands due to mutation and selection are absent (Rudner *et al.*, 1968). Reasons for this phenomenon are ascribed to the systematic nucleotide composition asymmetry between the coding-rich strand (H strand) that is purine rich and the coding-poor strand (L strand) that is pyrimidine rich. It is also ascribed to the mtDNA duplication process in which both strands are continuously replicated through starting points that are far apart in the mitochondrial genome, leading to the majority of the genome being duplicated in a unimodal direction (Nikolaou and Almirantis, 2006). In addition, evidence indicates that the mtDNA genome contains a strand-specific nucleotide

substitution mechanism that causes substitution rate differences between the four nucleotides in the leading and lagging strands and is caused by a high misincorporation rate, low proofreading functionality in the DNA polymerase complex or the absence of protective histones that protect the mtDNA against oxidative damage (Tanaka and Ozawa, 1994). The fact that one DNA strand is allowed to be single-stranded for a long period during replication is significant in terms of mutational bias, because it exposes the single strand DNA to the effects of oxidation in the mitochondria. Some publications suggest that the considerably higher rate at which C deaminates in single-stranded DNA and leads to a C to T substitution is evidence of single-stranded DNA that provides the opportunity for mutational-driven bias in nucleotide composition (Frank and Lobry, 1999).

Selective pressures are also regarded as possible reasons for PR2 deviations and codon composition patterns observed in the DNA genomes of humans (Sueoka, 2002). According to the neutral theory of evolution, the directional mutation pressure is the major force that causes sequence variation, in conjunction with purifying selection that plays a much smaller role in this process. The four types of nucleotides are unevenly distributed across the genome and therefore exposed to different forces of functional constraint in the context of the neutral evolution theory, therefore leading to deviations from nucleotide equidistribution (Knight *et al.*, 2001b).

Furthermore, selective constraints on the amino acid content of proteins based on protein structure, function or localisation occur through selective constraints on codon positions 1 and 2 and could also be a contributing factor to induce violation of the PR2 (Frank and Lobry, 1999). PR2 biases that dominate transitions or transversions could change the G+C content of the human DNA genome. For this reason, MEGA version 5 (Tamura *et al.*, 2011) was used to determine the nucleotide composition across the mtDNA genome for the respective codon positions of the Tswana population under investigation, as presented in Table 6.43. The G+C content for codon positions 2 and 3 was also determined by using DnaSP version 5 software (Librado and Rozas, 2009) and was in agreement with the values obtained when using the MEGA version 5 (Tamura *et al.*, 2011) software program. The values for the G+C content at the respective codon positions of the Tswana population under investigation were compared with the same measurements performed for

31,745 human mtDNA coding regions obtained from GenBank®[1] that contained 8,998,998 codons (Nakamura *et al.,* 2000) and the results are presented in Table 6.43.

Table 6.43 Nucleotide composition at codon positions for the Tswana dataset of this investigation

Dataset	Codon position # 1				Codon position # 2				Codon position # 3			
	T%	C%	A%	G%	T%	C%	A%	G%	T%	C%	A%	G%
Tswana dataset (3a)	21.0	26.8	30.9	20.9	41.0	27.0	20.0	11.6	18.0	38.4	36.2	7.8
Total G+C	47.7				38.6				46.3			
NCBI-Genbank®	48.11				39.34				47.36			

Tswana dataset (3a) = complete mtDNA genomes of 50 Tswana individuals. Codon positions indicated for all the coding regions of the mtDNA genome. Nucleotide composition indicated as percentages - calculated as the average total number of nucleotides at that codon position over the average total number of that codon position in the coding regions of the mtDNA genome. The NCBI-GenBank® represents the G+C content as a percentage obtained from the "Codon Usage Tabulated from GenBank®"as described by Nakamura *et al.,* 2000.

The Tswana population under investigation displays a bias towards G at codon position 1 when compared to the G content at the other two codon positions, which is in agreement with the literature and is ascribed to the frequent usage of glycine, alanine and valine in amino-acid composition. A bias towards T has been observed at codon position 2, which is described in the literature as a bias towards A or T because of a preference for acidic amino acids (Frank and Lobry, 1999). An intra-genomic decrease in G+C content has been reported in humans and specifically reported as always less than 50% in the mtDNA across different species (Knight *et al.*, 2001b). This phenomenon has also been observed in the Tswana population under investigation with an observed value for the total G+C content for the whole mtDNA genome at 44.4% and although not extremely low, this displays signs of bias. The G+C content is lowest at codon position 2 and similar for codon positions 1 and 3 in the Tswana population and therefore exhibits similarity with the G+C content observed for the large number of GenBank® human mtDNA genomes that were investigated, as presented in Table 6.43. Asymmetric deamination of the C residues occurring in the single-strand state during replication in the mitochondria has been regarded as a cause for the G+C heterogeneity, as well as an effective mechanism within the mitochondria to eliminate active oxygen through the superoxide dismutase-catalase system, leading to a lower occurrence of 8-oxo-dGTP and therefore lowering the G+C content effectively (Sueoka*,* 2002).

The DNA base composition of human genomes has been determined to play a role in the codon usage pattern and is correlated to forces that work on the nucleotide composition of

1 GenBank® is a registered trademark of the US Department of Health and Human Services, Bethesda, MD, USA.

the genome rather than on the selection for specific codons (Wright, 1986). More than half of the codon usage patterns in the mitochondria can be predicted by the nucleotide composition of the mitochondrial genome and therefore the nucleotide composition of the mtDNA is regarded as the single most important factor that determines codon usage (Knight *et al.*, 2001b). Synonymous codon usage is species-specific, usually differs between the different genes within a genome and is ascribed to the G+C variation observed at the synonymous codon positions within the coding regions of the genome. The G+C content of the Tswana population under investigation exhibited similarity with the G+C content observed for a large number of GenBank® human mtDNA genomes that were investigated and it can therefore be concluded that the codon usage pattern of the Tswana population under investigation can be expected to be similar to the codon usage patterns observed within other human mtDNA populations. Codon usage is preferred, as it relates to high levels of gene expression and refers to the efficiency of the translational system (Frank and Lobry, 1999). Mutational bias could cause a bias towards synonymous codon usage and/or G+C content (Sueoka, 2002).

6.9.2 Nucleotide diversity

DnaSP version 5 (Librado and Rozas, 2009) was used to determine different measures of both the individual and total sequence variance within the coding regions of datasets 1a, 2a, 3b, 4, 5 and 6 of this investigation, as was discussed in Section 6.9. DnaSP version 5 estimated the average number of *k* (Tajima, 1983) and its stochastic, sampling and total variances (Tajima, 1993) as well as the π, which was determined by the mean number of observed nucleotide substitutions that differed per site when compared among sequences. Sampling variance and standard error for π were also determined (Nei and Jin, 1989). The determination of the *S* through pairwise DNA sequence comparisons under the assumption that the sites were selectively neutral was also one of the informative measures of nucleotide diversity used in this investigation. The results are listed in Table 6.44.

Table 6.44 MtDNA coding region sequence diversity statistics of African populations of this investigation

Populations	Number of sequences (*n*)	Number of segregating sites (*S*)	Average number of nucleotide differences (*k*)	Total variance of *k*	Nucleotide diversity(π) X 10^{-3}	SD(π) X 10^{-3}
Global African	572	1,569	52.826	523.291	3.43	±0.05
All African	385	1,393	56.747	603.543	3.68	±0.05
Western African	60	406	48.873	460.257	3.17	±0.15
Eastern African	110	827	52.671	526.643	3.42	±0.11
Southern African	66	323	49.203	465.197	3.19	±0.16
Tswana	50	337	49.900	482.233	3.23	±0.15

Global African = dataset 1a; All African = dataset 2a; Tswana = dataset 3b; Western African = dataset 4; Eastern African = dataset 5; Southern African = dataset 6; SD = standard deviation.

The DNA sequence variation within the respective populations as defined by the datasets was measured by the *S* and by *k*. The measurement of *S* does not reflect the frequencies of mutations and would therefore not reflect the presence of deleterious mutations that are present in low frequencies as a result of negative selection. Measurement of *S* as an indicator of nucleotide diversity is therefore recommended for populations that are under neutral selective pressure and where mutation and genetic drift are at equilibrium (Tajima, 1993). A critical factor to take into account when using the number of segregating sites estimate is its dependence on the sample size (Tajima, 1983), as reflected in the results obtained for the measurement of *S* in this investigation. The larger the population sizes were, the larger the number of segregating sites that were observed in each of the respective populations under investigation. The largest population, the Global African population, consisted of 573 mtDNA sequences in which 1569 segregating sites were identified as opposed to the Tswana population that consisted of 50 mtDNA sequences and only displayed 337 segregating sites. The value of the number of segregating sites in terms of an indicator of nucleotide diversity therefore lies in the comparison of the estimates between populations of similar sizes and also for use in further population diversity estimates, such as the mutation parameter. For this reason Tajima (1983) recommended that the number of segregating sites be used to estimate the nucleotide diversity within a population rather than between populations (Tajima, 1983).

The South African dataset consists mainly of Khoi-San speaking individuals, which have been reported to display slightly lower levels of nucleotide diversity owing to the higher proportions of similar or related lineages in these populations. This is ascribed to the food-gathering and hunting lifestyle that was practised by these populations, which caused them to maintain constant population sizes over long periods of time, as well as limiting

gene flow from other human populations (Watson *et al.*, 1996). This is reflected in the slightly lower number of segregating sites that is observed in the Southern African dataset of this investigation when compared with the similarly sized Western African and Tswana datasets.

The difference between the estimates of the number of segregating sites and the estimates of the average number of nucleotide differences is the effect of selection. Negative selection will strongly affect the number of segregating sites because of the removal of the deleterious mutations from the population, whereas the average number of nucleotide differences will accommodate the deleterious mutations and therefore be a more accurate estimate of nucleotide diversity (Tajima, 1989). For populations that do not adhere to the requirement of neutral selection, measurement of *k* is recommended (Tajima, 1993) and in addition, the measurement of the average number of nucleotide differences is not dependent on the sample size. Using both of these measures in this investigation therefore provides a reliable measurement independent of evolutionary assumptions of the genetic diversity observed in the respective populations.

Table 6.45 MtDNA coding region sequence diversity statistics of Global and African populations

Populations	Number of sequences (*n*)	Number of segregating sites (*S*)	Average number of nucleotide differences (*k*)	Nucleotide diversity(π) X 10^{-3}	Reference
Global	320	1,545	41.4	2.69	Gonder *et al.*, 2007
	53	516	43.9	2.80	Ingman *et al.*, 2000
	277	---	---	3.06	Kivisild *et al.*, 2006
Non-African	226	1,068	27.9	1.81	Gonder *et al.*, 2007
	32	255	25.7	1.70	Ingman *et al.*, 2000
	148	---	---	1.85	Kivisild *et al.*, 2006
African	94	758	60.3	3.92	Gonder *et al.*, 2007
	21	290	57.0	3.70	Ingman *et al.*, 2000
	129	---	---	3.79	Kivisild *et al.*, 2006
Western African	70	---	---	3.22	Kivisild *et al.*, 2006
Eastern African	15	---	---	3.16	Kivisild *et al.*, 2006
South Africa	16	---	---	3.59	Kivisild *et al.*, 2006

--- = no results. Global = datasets that represent African and non-African mtDNA sequences, African = datasets that are representative of African mtDNA sequences only, Non-African = datasets that represent mtDNA sequences of individuals that do not belong to haplogroup L, Western, Eastern Africa = datasets that represent the mtDNA sequences from populations residing in western and eastern regions of Africa, South Africa = dataset that represents mtDNA sequences from individuals residing in South Africa. Statistics have been calculated for coding region of mtDNA sequences. Adapted from Ingman *et al.*, 2000; Kivisild *et al.*, 2006; Gonder *et al.*, 2007.

The average number of nucleotide differences observed within the datasets of this study reflects the same pattern of genetic diversity observed in other studies (Ingman *et al.*, 2000; Kivisild *et al.*, 2006; Gonder *et al.*, 2007), as presented in Table 6.45.

The All African dataset of this investigation displays the highest measurement for *k,* followed by the Global African and Eastern African datasets that display similar measurement values, and the Southern African dataset and Tswana datasets that also display similar measurement values. The Western African dataset displays the lowest measurement for *k*.

The nucleotide diversity and average number of nucleotide differences determined for all of the populations represented by the datasets of this study are in agreement with the nucleotide diversity measures of the coding regions of mtDNA genomes of other African populations (Ingman *et al.*, 2000; Kivisild *et al.*, 2006; Gonder *et al.*, 2007), as is presented in Table 6.45. The All African population of this study displays the highest level of genetic diversity, as in the case of the other studies presented in Table 6.45. This observation supports the theory that anatomically modern humans originated in Africa and the models that suggest that modern humans lived exclusively in Africa for an extensive period prior to the migration out of Africa to establish human populations in Eurasia (Cann, 1987; Vigilant *et al.*, 1991; Ingman *et al.*, 2000). Based on the ancient nature of the African populations and the high haplogroup diversity observed within African populations, it was expected that the All African dataset would display the highest level of nucleotide diversity.

This was followed by a slightly lower level of nucleotide diversity as well as a lower average number of nucleotide differences displayed by the Global African dataset. This was supported by the lower levels of nucleotide diversity and *k* observed in other global populations (Ingman *et al.*, 2000; Kivisild *et al.*, 2006; Gonder *et al.*, 2007). The global populations used in other studies were representative of African and non-African individuals and displayed even lower levels of genetic diversity. This could be ascribed to the fact that human populations that resided in Africa had a longer evolutionary history and larger effective population sizes than the population that migrated out of Africa and established the lineages of the Eurasian world (Ingman *et al.*, 2000; Kivisild *et al.*, 2006; Gonder *et al.*, 2007).

The level of nucleotide diversity and average number of nucleotide differences observed in the eastern African population of this investigation are similar to the nucleotide diversity

displayed by the Global African population. This observation is consistent with similar studies that have indicated high levels of genetic diversity in populations from eastern Africa (Gonder *et al.*, 2007). It has been postulated that eastern Africa has been an ancient source of human populations that migrated out of Africa and the high levels of nucleotide diversity therefore reflect the long evolutionary history of the populations that resided in this region of Africa (Watson *et al.*, 1997; Kivisild *et al.*, 2004; Gonder *et al.*, 2007).

The southern African and western African populations of this investigation display the lowest level of nucleotide diversity and average number of nucleotide differences. In contrast to the level of nucleotide diversity of non-African populations as presented in Table 6.45, it is high and therefore indicates a long evolutionary history, as would be expected from African populations. The slightly lower levels of nucleotide diversity and average number of nucleotide differences of the southern African population in comparison to the All African population of this investigation could be explained by the fact that this dataset consists mainly of mtDNA sequences that belong to Khoi-San speaking individuals (89%), which display a high proportion of haplogroups L0d and L0k. Although the internal variation of these haplogroups is evidence of the ancient nature of these lineages and therefore an expected high nucleotide diversity, the nucleotide diversity within the ancient hunter-gatherer populations of Africa often display signs of limited genetic flow due to the isolated and nomadic lifestyles that were practised (Batini *et al.*, 2011). The low measurement value for *k* and the nucleotide diversity in the Western African dataset was, however, unexpected and could be explained by a sample bias caused by the large proportion of mtDNA sequences belonging to individuals from Burkina Faso (50%) or of Pygmy origin (23%). Burkina Faso populations have been described as being similar to the ancient Pygmy and Khoi-San populations in terms of lifestyle and population behaviour (Cerný *et al.*, 2006). As a consequence, the populations from Burkina Faso may display lower levels of genetic diversity based on limited gene flow with other populations. Similarly, the Pygmy populations also demonstrate low levels of genetic diversity (Quintana-Murci *et al.*, 2008), which could be a likely reason for the observed lower average number of nucleotide differences and diversity in this dataset.

The Tswana population of this investigation displays a level of nucleotide diversity and average number of nucleotide differences that are slightly higher than those in the South African population but lower than those in the Eastern African dataset. In the context of other non-African populations, the level of nucleotide diversity is high and typical of an African population. By implication this indicates a long evolutionary history for this

population. The slightly higher level of nucleotide diversity of the Tswana population in comparison to the South African population indicates that the population was probably not as isolated as the typical hunter-gatherer Khoi-San speaking populations and that it contained some ancestral influences from other African lineages. This is evident from the haplogroup composition of the Tswana population, which indicates gene flow with the early Bantu-speaking populations that migrated to southern Africa. It also indicates that the population had a sufficiently large effective population size to maintain these sequence variants over a long period of time. The nucleotide diversity of the Tswana population is, however, not as high as that of the eastern African population, indicating a later origin of this population than the populations that resided in eastern Africa.

6.9.3 Population size

The determination of population expansion signals is based on the principle that the likelihood that two individuals in a population are related through maternal ancestry is greater in a small population than in a large population. Lineages would therefore be greater in larger populations than in smaller populations. When a small population has expanded rapidly, the number of lineages will increase but coalesce back to the point of common ancestry before the expansion took place. The coalescence theory is primarily based on the neutral theory of evolution and provides the structure for most of the statistical tests that are currently used to determine signals of past population growth in modern-day populations (Hudson, 1991; Fu and Wen-Hsiung, 1999). These statistical tests include Fu's F_S statistic, the Ramos-Onsins and Rozas R_2 statistic and the mismatch distribution according to the method of Harpending (1994), in conjunction with the raggedness statistics, and allow for the quantification of the smoothness of the mismatch distribution, as discussed in Section 5.14.3. All these methods were used in this investigation for the determination of the signals of population growth.

Mutation rate heterogeneity between different regions of the mtDNA, as well as high levels of homoplasy, could obscure the signal of population growth within a population (Excoffier, 1994) and therefore only the coding regions of the mtDNA sequences of the respective datasets were used for the determination of population size and growth estimates in this investigation. The datasets of this investigation were designed to represent different regional African populations in addition to the broader sampled Global African and All African datasets that were used to represent the broader African population as a whole. The purpose of determining the statistical signals of population growth for these different

African populations was to position the Tswana population of this investigation against the broader African populations. The Tswana population under investigation is a modern population comprised of the offspring of earlier, more ancient populations that contributed to the current population composition of Tswanas and the population signatures carried by the modern-day Tswana population would thus be a reflection of the past population histories of the respective contributors to this population. Therefore, interpreting the population signature of the Tswana population under investigation in the context of the other African regional and continental groups is important. The composition of the datasets used is described in Section 5.11 and a full list of ethnicity is presented in Appendix B.

Different statistical approaches were followed to determine the population growth signals in the respective datasets of this investigation, as described in Section 5.14.3. Fu's F_S test was used, followed by the mismatch distribution and raggedness statistic (*rg*) that was calculated for each of the datasets and further complemented by using the R_2 statistic (Ramos-Onsins and Rozas, 2002). Confidence intervals were determined for each of the statistical estimates used by computer simulations that were based on the coalescence algorithm under the assumptions of no recombination and the neutral infinite-site model. DnaSP version 5 (Librado and Rozas, 2009) software was used for this purpose and the results are presented in Table 6.46.

Table 6.46 Statistical measures of population growth for the datasets of this investigation

Populations	Global African	All African	Western African	Eastern African	Southern African	Tswana
Dataset	1(a)	2(a)	4	5	6	3(b)
Population size	572	385	60	110	66	50
F_S $(P)^a$	-602.013 **(0.00000)**	-353.445 **(0.00000)**	-11.994 **(0.00800)**	-57.1866 **(0.00000)**	-12.351 **(0.00920)**	-6.486 (0.05260)
R_2 $(P)^b$	0.0161 **(0.00000)**	0.0193 **(0.00020)**	0.0536 **(0.01500)**	0.0277 **(0.00000)**	0.0718 (0.15140)	0.0671 (0.05840)

Global African = individuals of African origin from African and non-African countries; All African = individuals of African origin residing in Africa in all regions; Western African = individuals of western African origin; Eastern African = individuals of eastern African origin; Southern African = individuals of southern African origin; Tswana = Tswana-speaking population under investigation; Population size = the number of individuals contained in the dataset = number of mtDNA coding region sequences; *P* = statistical significance determined by 5,000 coalescence simulations under confidence interval assumption; a = all *P* values < 0.02 for Fu's F_S statistically significant; b = all *P* values < 0.05 for *rg* and R_2 statistically significant; *P* values in **bold** indicate statistical significance; F_S = Fu's F_S statistic; R_2 = Ramos-Onsins and Rozas statistic (Ramos-Onsins and Rozas, 2002).

Fu's F_S test is based on the probability that the number of segregating sites will be equal or larger in a sample taken from an expanding population than the observed number of segregating sites taken from a stationary population that adheres to the parameter $\theta = 2Nu$ where N is the effective population size and u represents the mutation rate and where θ is

equated to the average number of pairwise differences (π). Negative values for the F_S statistic were interpreted as an indication of population expansion (Excoffier and Schneider, 1999) and *P* values of less than 0.02 were regarded as a statistically significant indication of an expansion event in the population under investigation (Fu, 1997). According to these criteria, both the Global African and All African populations display large negative F_S values of statistical significance, which indicate a population expansion event. This is in agreement with the literature, in which it has been reported that the population of modern humans expanded in size in Africa about 60,000 ybp before migrating to other regions of the world and expanding further (Rogers and Harpending, 1992; Sherry *et al.*, 1994; Jorde *et al.*, 1998). Because of this early expansion it was postulated that most of the human populations would exhibit a bell-shaped mismatch distribution and signals of population expansion (Bandelt and Forster, 1997), as in the case of the broad Global African and All African datasets that represent a large number of individuals of African origin. Broad sampling from many different populations of different global regions, however, can result in an upward bias in the number of observed rare alleles or singletons in a dataset (Pilkington *et al.*, 2008). Although the Global African and All African datasets consist of individuals that are all of African origin and belong to haplogroup L, the datasets are large and sampled from a broad range of populations and this could be the reason for the large negative F_S values observed in these two datasets. Natural selection shows a similar type of genetic signature as displayed by demographic processes associated with population expansion and population structure. Therefore the datasets were investigated for selection and thus interpreted in Section 6.9.4.

The R_2 statistic is based on the difference between the average number of nucleotide differences and the number of rare alleles or singletons in a population and reportedly has greater power to detect population expansion signals within a sample of non-recombining sequence data that is either large i.e. >50 or small i.e. <10 (Ramos-Onsins and Rozas, 2002). Significantly low values of the R_2 statistic support the F_S values of the Global African and All African datasets and therefore confirm the hypothesis of population expansion within these two datasets.

Although the sample design of the datasets was developed to avoid problematic issues such as mutation heterogeneity and homoplasy, it has been demonstrated that the interpretation of population behaviour signals from datasets that have been pooled or sampled across different geographical regions is difficult and could seem to be compatible with values expected under a population expansion hypothesis even if it is not true

(Bertorelle and Slatkin, 1995). The subdivision of the All African dataset into regional datasets provides a mechanism to overcome this problem by providing statistical signals of population growth for each of the respective regionally sampled datasets and consequently clarity on the observed population signals of the larger datasets.

The Eastern African dataset displays the highest negative F_S value of the regional datasets, with a *P* value that is statistically significant, which presents strong evidence of population expansion. This signal has been verified by the statistically significant positive R_2 estimate and supports a model of population expansion for the populations of this region within Africa. Although the statistical estimates clearly indicate evidence of population expansion, the signals are not as high as those estimated from the larger datasets. This could be ascribed to the pooled nature of the larger datasets or to the size or nature of the composition of the Eastern African dataset. Furthermore, the results are in agreement with measures of genetic diversity as discussed in Section 6.9.2 where the Eastern African dataset displayed higher genetic diversity than was observed for the other African populations, indicating the presence of a long genetic history and large expanded populations in this region of Africa. It has been reported that haplogroup L3 is commonly present in eastern Africa (Salas *et al.*, 2002) and that eastern Africa was the starting point of the human globalisation event. Studies of the L3 haplogroup have indicated marked growth at 8-12 kya before the global expansion of modern humans took place and it is postulated that the populations in eastern Africa underwent a major population expansion event prior to the migration out of Africa (Atkinson *et al.*, 2009). This event could have been caused by favourable climate changes about 70 kya in eastern Africa (Scholz *et al.*, 2007) in combination with cultural practices that favoured the survival of one population over other populations in the same region. Archaeological evidence of a cultural revolution in eastern Africa at that time (Mellars, 2006) supports this theory of a major population expansion within the eastern African populations that belonged to haplogroup L3. Individuals belonging to haplogroup L3, according to the PhyloTree classification system (Van Oven and Kayser, 2009), comprise 37.2% of the Eastern African dataset of this investigation, making it the predominant haplogroup present in this dataset, as would be expected in light of the data. The strong signals for population expansion within this regional dataset were therefore expected and supported the theories presented in the literature.

The Southern African and Western African datasets displayed very similar statistically significant negative F_S values that were much lower than those displayed by the Eastern

African dataset. The population expansion signal within the Western African dataset was supported by a statistically significant positive R_2 statistic. This, however, was not the case with the Southern African dataset, where the R_2 statistic was not significant.

As described in Section 6.9.2, the Western African dataset consists of a large proportion of mtDNA sequences that belong to individuals from the Foulbe ethnic group in Burkina Faso (50%) as well as individuals of Pygmy descent (23%). The Foulbe ethnic group forms part of the Fulani populations that reside *inter alia* in Burkina Faso where many of these groups have led nomadic lifestyles over many thousands of years. Studies have reported that the nucleotide diversity within these groups is lower than expected within the western African Bantu-speaking populations and, more specifically, it has been reported that these populations also display low F_S values that are often not statistically significant (Cerný *et al.*, 2006). It has been postulated that these populations have undergone reduced demographic expansions that were comparable to the hunter-gatherer Khoi-San-speaking groups and Pygmy groups. The other major component of the Western African dataset is the Pygmy group that reportedly has a constant gene pool due to their cultural practices in which women migrated to neighbouring Bantu-speaking populations to be married, but the opposite seldom occurred (Cerný *et al.*, 2006). Therefore the gene pool on the maternal side of the population has not displayed any expansion within these ethnic groups.

In addition, the dataset represents a large proportion of L0 and L1 haplogroups (45%) as well as a large proportion of L2 and L3 haplogroups (55%). As has been discussed, haplogroup L3 displays strong signals of population expansion. This is also the case for populations that harboured haplogroup L2, which were involved in later population expansions about 12-20 kya, which is associated with a period that predated the Bantu migrations to the southeastern regions of Africa (Atkinson *et al.*, 2009). This population expansion has been associated with the climate changes during the LGM, which enlarged the Sahara and increased the open savannah and woodlands in central Africa, creating an ideal environment for the settlement and subsequent expansion of populations (Salas *et al.*, 2002). In contrast, haplogroups L0 and L1 displayed statistical signals of slow constant growth. This could be attributed to the ancient hunter-gatherer populations, such as the Khoi-San and Pygmy populations that harboured these haplogroups nearly exclusively and displayed signs of deep population structure over a long evolutionary history (Atkinson *et al.,* 2009). The near equal contributions of these two groups of haplogroups to the Western African dataset, as well as the composition of the dataset,

could explain the F_S and R_2 values that, although clearly indicating signs of population expansion, are not as high as what was observed in the Eastern African dataset.

In contrast to the Western African dataset, the R_2 statistic for the Southern Africa dataset does not support a model of population expansion as was observed in the significantly low negative F_S value of -12.351 as indicated in Table 6.46. The Southern African dataset consists of a large component of Khoi-San speakers belonging to haplogroup L0 (77%) from which the signals for population growth could be determined and compared to the Tswana populations that display a similar haplogroup composition. As discussed, haplogroup L0 was reported to display signs of slow and constant growth with evidence of deep population structure early in the human mtDNA tree (Atkinson *et al.*, 2009). This population structure was ascribed to the ancient hunter-gatherer Khoi-San populations that nearly exclusively harboured the haplogroups L0d and L0k and lived in isolation for long periods of time (Behar *et al.*, 2008). Recent studies have furthermore suggested that there was a split between the haplogroup L0 and haplogroups L1'-6 populations at 210-140 kya and that this event was the cause of pronounced population structure and separate development of populations that populated the African continent (Behar *et al.*, 2008). The Khoi-San populations possibly diverged from the other human populations in eastern Africa between 140-90 kya, followed by the independent development of the L0d and L0k haplogroups. Later Bantu migrations to the southern regions of Africa led to the assimilation of the Khoi-San speaking population lineages into the Bantu-speaking populations. The transition from hunter-gatherer lifestyles to food-producing lifestyles by the Bantu speakers in the central African region of the Cross River Valley, in conjunction with the iron-smelting era, created conditions for population growth and migration to the south via two eastward and westward routes (Newman and Roberts, 1995). Genetic evidence suggests strong gene flow between the different Bantu lineages during the migrations (Tishkoff *et al.*, 2007). The southern African sample of this investigation is biased towards the Khoi-San populations and it is therefore expected that the population growth signals would indicate the stationary population behaviour of the Khoi-San populations rather than lean towards a population expansion model. The presence of Bantu speakers in the dataset, albeit a small component, in addition to the assimilation of a component of Bantu lineages into the Khoi-San populations, would explain the contrasting signals of expansion between the F_S statistic and the R_2 statistic. The weak population expansion signal represented by the low negative F_S value therefore represents the slow-growing Khoi-San populations in conjunction with the signals of the Bantu

migration and subsequent assimilation between populations in the southern region of Africa.

The Tswana population of this investigation displays a non-significant low negative value for the F_S statistic and a non-significant positive value for the R_2 statistic. This population has haplogroup and nucleotide diversity, which indicates gene flow with the early Bantu-speaking populations that migrated to southern Africa, as was discussed in Section 6.9.2. This phenomenon is not unusual, as has been observed in studies of other Bantu-speaking populations of southern Africa that display low negative but significant F_S values indicating past population expansions (Salas *et al.*, 2002; Castri *et al.*, 2009). It also displays haplogroup and nucleotide diversity involving the maternal lines of descent that indicate strong gene flow or assimilation with the Khoi-San populations, as is was evident from the large component of haplogroup L0d that is predominantly present in the southern African Khoi-San populations. As discussed, the Khoi-San populations display signals of slow or constant population growth in contrast to the expanding Bantu-speaking populations that migrated to southern Africa. The non-significant F_S and R_2 values displayed by the Tswana population under investigation therefore reflect the combination of population growth signals from the Bantu-speaking genetic component as well as the population growth signals from the Khoi-San genetic component within the Tswana population, which indicates a generally slow-growing population. This observation is in line with the population growth signal obtained from the South African dataset and is to be expected in view of the shared maternal genetic signatures within the Tswana population under investigation.

Studies have indicated that most of the coalescence events within a population can be detected before an expansion took place and individuals of a modern population would therefore be able to trace ancestry back to a common ancestor that existed before the expansion (Slatkin and Hudson, 1991; Rogers and Harpending, 1992). This phenomenon is indicated by a peak in a nucleotide pairwise difference or mismatched distribution and can be used to determine the time at which the population expansion took place (Jorde *et al.*, 1998). The mismatch distribution determinations were therefore used in this investigation to measure the signal left in the distribution of the pairwise differences between nucleotides after a population expansion event and the raggedness statistic was used to measure the smoothness of the mismatch distribution where low values for *rg* were interpreted as signs of population growth (Harpending, 1994). The mismatch distribution investigated the relationship between the observed numbers of pairwise

nucleotide differences within the respective datasets of this investigation with the expected number of pairwise differences under the assumption of a sudden expansion event with the aim of inferring past population behaviour of the sampled population. A population that underwent exponential population growth was expected to display a unimodal Poisson distribution and a population that was stationary was expected to display ragged peaks in the mismatch distribution (Slatkin and Hudson, 1991).

To provide confidence and assist in the interpretation of the findings of the mismatch distributions, the sum of squared deviations was determined by performing a parametric bootstrap of 1,000 simulated samples using the coalescence algorithm (Hudson, 1991) and estimating the *P* value as an approximation of the probability that the simulated SSD value would be larger or equal to the SSD value observed in the population under investigation. The empirical distribution of the SSD goodness-of-fit statistic between the simulated and expected mismatch distribution was compared with the SSD distribution between the observed and expected mismatch distribution to determine the *P* value for the SSD. Thus a small *P*(SSD) value was interpreted as indicative of a poor fit of data to the model estimated by the demographic parameters, which in the case of using Arlequin version 3.5, was a model of sudden expansion. In addition, confidence intervals (CI) at 95% of the mismatch distribution and demographic parameters were determined by a parametric bootstrap method. Only the CI values for the mismatch distributions were used in this investigation, since the demographic parameters were reportedly overly conservative (Schneider and Excoffier, 1999). The 95% CI range values of the theta parameters of the initial population size (θ_0) and the population size after expansion (θ_1) as determined in Arlequin version 3.5 under an assumption of sudden population expansion were used to support the population growth signals obtained from the mismatch distributions of this investigation. An overlap of the 95% CI range values of the theta parameters was regarded as evidence of a stationary population (Pilkington *et al.*, 2008). The mismatch distribution parameters for the respective datasets of this investigation are presented in Table 6.47.

Table 6.47 Mismatch distribution parameters estimated under a sudden expansion model

Datasets	Global African	All African	Western African	Eastern African	Southern African	Tswana
θ_0 qt 95% Lower bound	0.000	0.000	0.000	0.000	0.000	0.000
θ_0 qt 95% Upper bound	9.311	7.871	10.030	85.312	9.946	83.273
θ_1 qt 95% Lower bound	71.777	91.212	62.029	118.767	55.222	76.482
θ_1 *q*t 95% Upper bound	582.487	885.748	644.607	99999.000	466.510	525.935
SSD(*P*)	0.00130 (0.92000)	0.00130 (0.81600)	0.00221606 (0.92600)	0.00373 (0.39200)	0.00656 (0.40900)	0.00828 (0.12600)
Rg index (*P*)[a]	0.00021 (1.00000)	0.00028 (1.00000)	0.00190 (0.88800)	0.00035 (0.99900)	0.00382 (0.21000)	0.00246 (0.79400)
Rg index (*P*)[b]	0.0001 **(0.01120)**	0.0002 (0.59860)	0.0019 (0.15760)	0.0006 (**0.04580**)	0.0027 (0.32440)	0.0028 (0.23280)

Global African = individuals of African origin from African and non-African countries; All African = individuals of African origin residing in Africa in all regions; Western African = individuals of western African origin; Eastern African = individuals of eastern African origin; Southern African = individuals of southern African origin; Tswana = Tswana-speaking population under investigation; θ_0 qt 95% lower bound and upper bound = 95% CI range values for the initial population parameter estimate; θ_1 qt 95% lower bound and upper bound = 95% CI range values for expanded population parameter estimate; SSD = squares of sums of deviations for mismatch distribution; SSD (*P*) = statistical significance based on 1,000 parametric bootstrap simulations performed by Arelquin v 3.5 under sudden expansion population model; *rg* = raggedness statistic (Harpending, 1994) *Rg*(*P*)[a]= statistical significance determined by 1,000 parametric simulations performed by Arlequin v 3.5 under sudden expansion model; *Rg*(*P*)[b]= statistical significance determined by 5,000 coalescence simulations under confidence interval assumption performed by DnaSP v 5 under constant size population model, all *P* values < 0.05 statistically significant; *P* values in **bold** indicate statistical significance.

Even though the mismatch distributions were interpreted with the latter goodness-of-fit statistics and CIs, Schneider and Excoffier (1999) reported that caution was needed when the variance of the expected distribution of possible lineages was high and the population under investigation was stationary and therefore far removed from the expected distribution. The mismatch distribution method of Rogers and Harpending (1992), however, displayed low levels of error when population samples were larger than 50 samples and especially if population expansion took place and therefore the large datasets in this investigation were not regarded as being problematic in this regard (Bertorelle and Slatkin, 1995). The larger the number of genes or sequence regions selected for the investigation of the pairwise nucleotide differences, the less chance there would be of observing these large differences between the variations of expected and observed pairwise differences.

The mismatch distribution method of Rogers and Harpending (1992) further assumes that only mutation, genetic drift or population expansion would be responsible for a unimodal expansion (Bertore and Satkin, 1995). The possibility of selection or high levels of homoplasy causing a unimodal distribution has therefore not been taken into account when using this method. By using the coding regions of the mtDNA sequences only, the

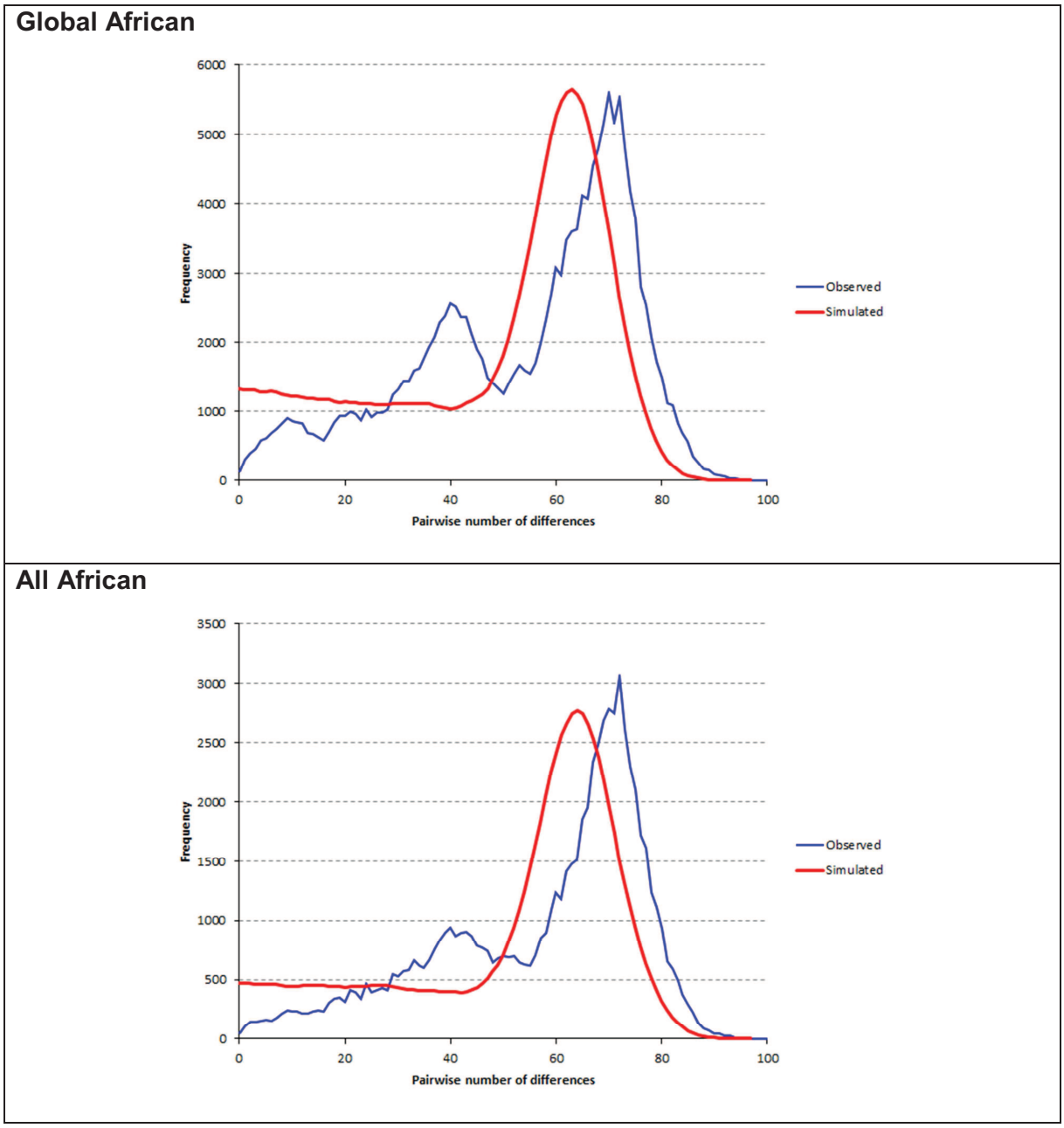
possibility of homoplasy is disregarded and by testing for selection, as discussed in Section 6.9.4, the impact of selection on the genetic signals of population expansion is considered.

The mismatch distributions of the Global African and All African datasets, as presented in Figure 6.78, are similar in shape and in both instances display bimodal peaks. Non-unimodal distributions in combination with insignificant Fu statistics are regarded as evidence of a stationary population under the assumption of neutrality (Harpending, 1993). In this investigation the mismatch distributions for both datasets are expected to be unimodal under the assumption of a sudden expansion model based on the statistically significant large negative F_S values for both datasets, as presented in Table 6.47. The high *P*(SSD) values for both datasets provide further evidence that the data are a good fit to the model of sudden expansion and therefore also suggests a model of population growth for these populations. In addition, the *rg* indices for both datasets are small and statistically significant when determined under a model of constant population size and large when determined under a model of sudden population expansion and thus confirm the presence of a smooth distribution.

The observed departures from the expected unimodal mismatch distribution in the Global African and All African datasets can be explained by adaptation through natural selection if the population size is constant or by demographic changes if neutrality is maintained, or both under different mutational models (Aris-Brosou and Excoffier, 1996). Furthermore, the genetic diversity patterns caused by periods of contraction in the growth of populations can also display unexpected mismatch distributions (Excoffier and Schneider, 1999). Unimodal peaks are expected to be defined by long terminal branches that are representative of an expanding population in which the mutations between the lineages are specific to the respective lineages that developed within the population and are not shared, as would be expected in a population of constant size (Donnelly, 2007). A bimodal peak distribution therefore indicates two such distributions, which could be ascribed to two separate population expansion events that would have led to the development of two different major haplogroup types. Another possibility could be the presence of two or more haplogroups at high frequencies in the datasets due to the effect of selective pressure in the population under investigation or because of the haplogroup compositions of the datasets. Since the datasets have not been sampled to be equally representative of all the haplogroups, the bimodal distributions could be related to the unequal composition of the respective haplogroups present in the datasets, with each of the major peaks representing the

mutational steps involved in the most frequent haplogroup developments. The presence of the bimodal distributions in both datasets, which differ in terms of haplogroup composition, suggests that this possibility is, however, unlikely and that the reasons for the bimodal distribution is more likely to be the complex and long history of the African populations in which contractions and expansions are thought to have occurred in conjunction with selection regimes that mimicked population size changes.

Figure 6.78 Mismatch distributions for the Global African and All African populations under the assumption of a sudden population expansion model



Mismatch distributions under an assumption of a sudden population expansion are presented. The mismatch distributions presented here investigate the relationship between the observed numbers of pairwise nucleotide differences within the Global African and All African datasets of this investigation with the expected number of pairwise differences under the assumption of a sudden expansion. The observed pairwise differences are indicated by a blue line and the expected pairwise differences by a red line. The mismatch distributions have been determined by using Arlequin version 3.5 (Excoffier and Lischer, 2010) as discussed in Section 5.14.

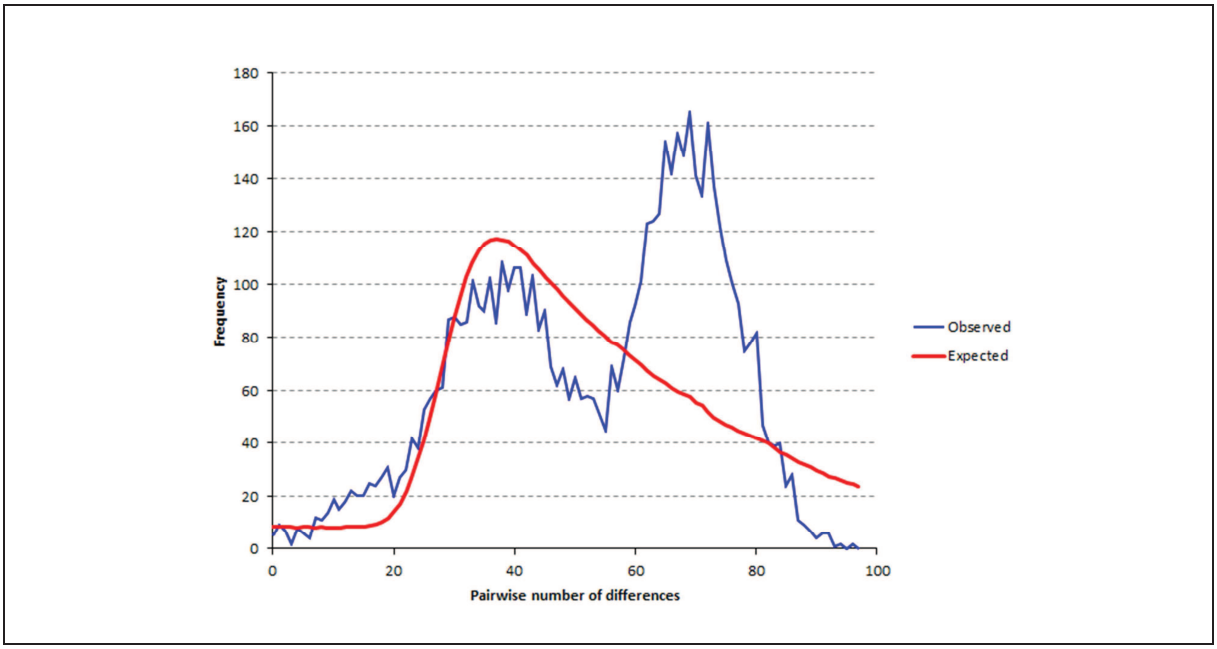
The model of population growth suggested by the mismatch distributions and population parameters for the Global African and All African datasets of this investigation is in agreement with the reported mismatch distributions of most human populations that display evidence of a major human population expansion that took place between 100 kya and 30 kya (Slatkin and Hudson, 1991; Harpending, 1994; Rogers, 1995). It is further believed that a small population that contained haplogroup L3 acquired some behavioural innovation around 60 kya – 80 kya and subsequently expanded into Eurasia as a result of this (Watson *et al.*, 1997). Further climatic changes around 4 kya caused the next major human expansion when the agriculturalist Bantu-speaking populations moved 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).

Two theories exist to explain the unimodal mismatch distribution observed in the mtDNA of human populations. The first is based on the interpretation of the mismatch distribution as evidence of one or more late Pleistocene demographic human population expansion events that were preceded by a population bottleneck, therefore from a small population of early humans (Sherry *et al.*, 1994; Rogers, 1995; Rogers, 1997). The second theory is based on the effect of selective sweeps that favoured the development of specific advantageous mutations within the early human populations. Based on evidence that the pairwise differences between chimpanzees and humans were both supportive of a population expansion model, the theory of a simultaneous selective sweep in both species is not regarded as likely and the model of a population bottleneck caused by some environmental factor that could have affected both species is more probable (Jorde *et al.*, 1995). Further arguments against the selective theory are the fact that autosomal and Y chromosome demographic analyses are supportive of a population expansion theory and that it is highly unlikely that selection would affect both autosomal and Y chromosomes in the same way (Rogers, 1995; Atkinson *et al.*, 2009).

The mismatch distribution of the Eastern African datasets is presented in Figure 6.79 and displays a more pronounced bimodal distribution than what was observed in the Global African and All African datasets. The expected mismatch distribution indicates a unimodal peak positioned more to the left than in the Global African and All African mismatch distributions, suggesting stronger genetic evidence of an earlier human population expansion event or selective sweep within the Eastern African dataset. Although the mismatch parameters of the Eastern African dataset do indicate a population expansion

model, they do not fit the model of sudden expansion as well as in the case of the other two datasets, as is evident from the higher goodness-of-fit SSD statistic and lower *P*(SSD). The signal of population expansion is further supported by the low *rg* indices obtained under both the sudden expansion and constant population assumptions. The *P*(*rg*) is significant when determined under a model of constant population size, supporting the model of population expansion for this dataset. The mismatch distribution parameters are therefore in alignment with the F_S statistic that is not as negative as the larger datasets but indicates strong support for past population expansion events. The same reasons for the presence of a bimodal distribution observed in the Global African and All African mismatch distributions are applicable to this dataset. As discussed earlier in this section, the Eastern African population contains a high percentage of L3 haplogroup individuals for which evidence exists of a major expansion event about 8-12 kya (Mellars, 2006; Atkinson *et al.*, 2009). It is therefore known that this population underwent more than one major population expansion event and this is a likely reason for the more pronounced bimodal distribution observed in this dataset. On removal of the haplogroup L3 individuals from the Eastern African dataset, not presented here, the mismatch distribution peak to the right became less pronounced and therefore the bimodal distribution became less pronounced and thus supported the latter theory. It is further proposed that the significant second peak in the bimodal mismatch distribution also contributes to the lower *P*(SSD) value observed for this dataset because of the large variation between the expected and observed mismatch distributions and the smaller size of the dataset (Schneider and Excoffier, 1999).

Figure 6.79 Mismatch distribution for the Eastern African population under the assumption of a sudden population expansion model

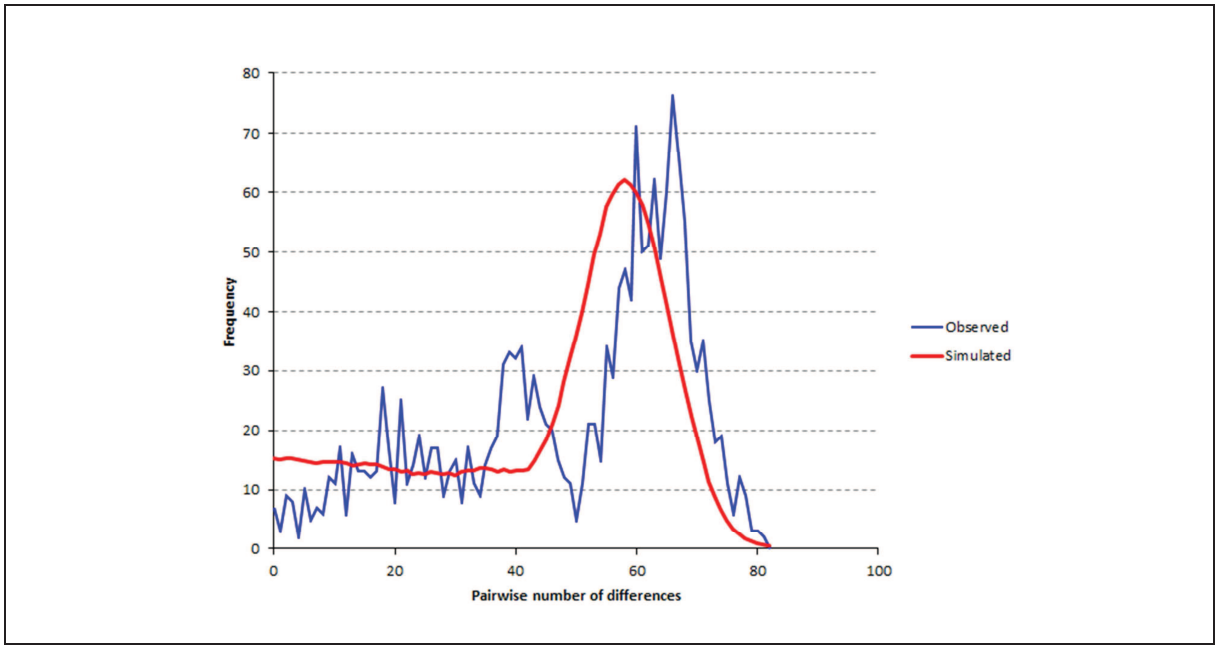


A mismatch distribution under an assumption of a sudden population expansion is presented. The mismatch distribution presented here investigates the relationship between the observed numbers of pairwise nucleotide differences within the Eastern African dataset of this investigation with the expected number of pairwise differences under the assumption of a sudden expansion. The observed pairwise differences are indicated by a blue line and the expected pairwise differences by a red line. The mismatch distributions have been determined by using Arlequin version 3.5 (Excoffier and Lischer, 2010) as discussed in Section 5.14.

The Western African dataset displays a mismatch distribution as presented in Figure 6.79 that is in agreement with the F_S statistics and indicates less support for a model of past population expansion than the Global African, All African and Eastern African datasets. As discussed earlier in this section, the absence of strong evidence for a population expansion model was not expected. This is evident from the *rg* indices determined under sudden expansion and constant size assumptions that are higher than those observed in the other datasets. This therefore suggests that the distribution was less smooth than would be expected of a model of population expansion. The *P*(*rg*) statistic determined under a model of constant population size is not statistically significant, as in the case of the Eastern African dataset, and therefore supports a more ragged mismatch distribution. The *P*(SSD) statistic is, however, similar in value to the observed *P*(SSD) values of the Global African and All African datasets and therefore indicates a model of population expansion for this population. The low population expansion signal for the Western African dataset, indicated by the F_S and R_2 statistical measures, also indicates a weak model of population expansion and is ascribed to the composition of the dataset, as discussed in this section. The same reasons explain the mismatch distribution and parameters obtained for the Western African dataset, which alludes to the large component of Burkina Faso ethnic groups contained in the dataset in addition to the Pygmy population component as the most probable contributors of a population signal of constant size to the mismatch

distribution. Similar observations with regard to populations of western Africa have been made, where population expansion was determined to have taken place in clusters rather than the total Western African population (Graven *et al.*, 1995; Watson *et al.*, 1997). The results of this investigation are therefore supported by these previous observations.

Figure 6.80 Mismatch distribution for the Western African population under the assumption of a sudden population expansion model

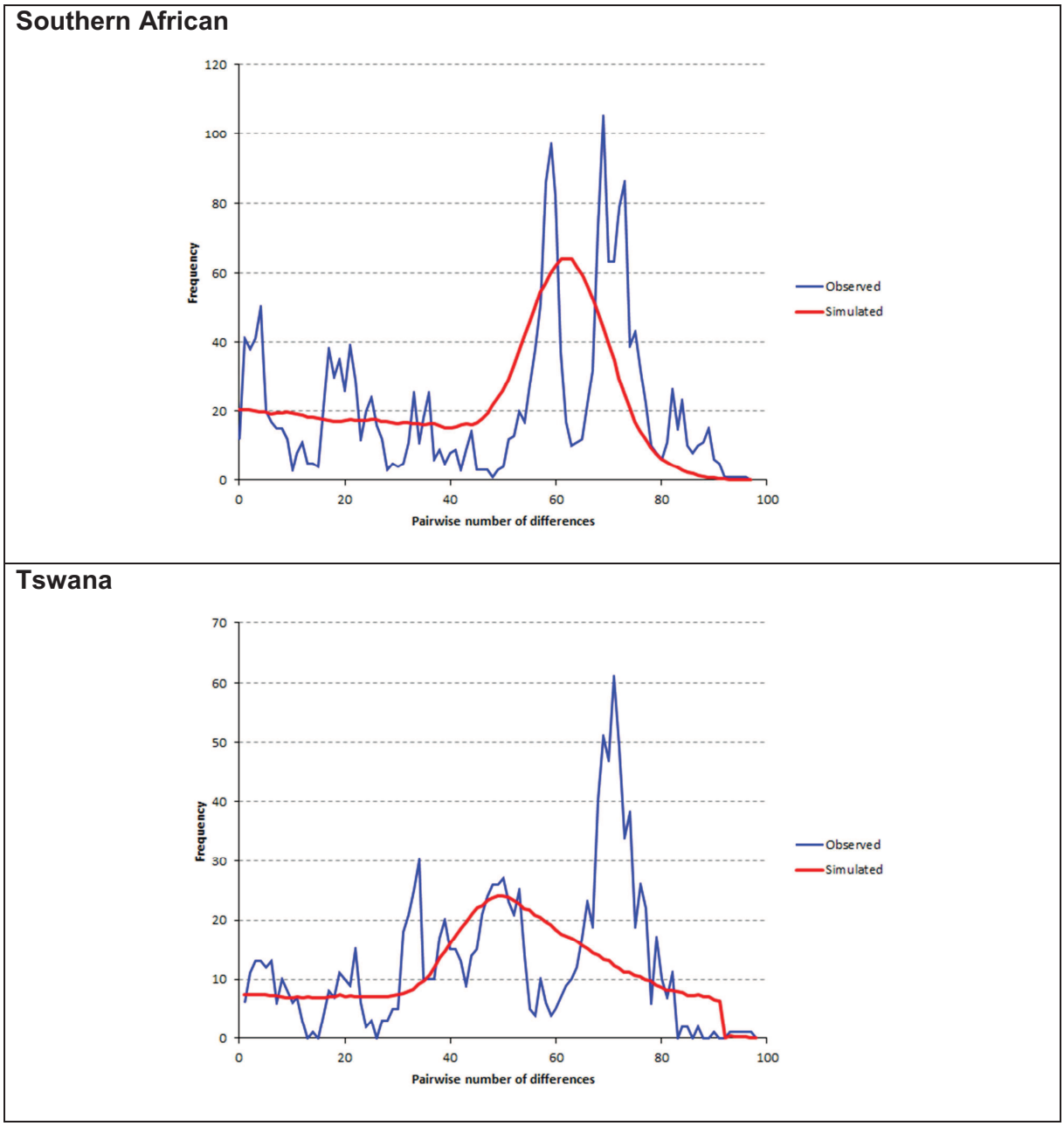


A mismatch distribution under an assumption of a sudden population expansion is presented. The mismatch distribution presented here investigates the relationship between the observed numbers of pairwise nucleotide differences within the Western African datasets of this investigation with the expected number of pairwise differences under the assumption of a sudden expansion. The observed pairwise differences are indicated by a blue line and the expected pairwise differences by a red line. The mismatch distributions have been determined by using Arlequin version 3.5 (Excoffier and Lischer, 2010) as discussed in Section 5.14.

The Southern African and Tswana populations of this investigation both display multimodal mismatch distributions and parameters, presented in Figure 6.81, which indicates a model of constant population size for these two populations. The goodness-of-fit SSD statistic is the highest of all the SSD values obtained for the datasets of this investigation and the *P*(SSD) values are low, especially for the Tswana population. The low *P*(SSD) value of the Tswana population furthermore coincides with an overlap between the 95% CI upper bound and lower bound measurement range of the θ_0 and the θ_1 parameters, providing strong support for a model of constant population size for the Tswana population under investigation. The *rg* indices and *P*(*rg*) measurements are similar for both populations and higher than any of the other *rg* indices for the other datasets investigated, indicating that the mismatch distributions of both datasets are more ragged than the other mismatch distributions. These results are in agreement with the F_S statistics and provide strong

evidence of a model of constant population size for the Southern African and Tswana populations of this investigation.

Figure 6.81 Mismatch distributions for the Southern African and Tswana populations under the assumption of a sudden population expansion model



Mismatch distributions under an assumption of a sudden population expansion are presented. The mismatch distributions presented here investigate the relationship between the observed numbers of pairwise nucleotide differences within the Southern African and Tswana datasets of this investigation with the expected number of pairwise differences under the assumption of a sudden expansion. The observed pairwise differences are indicated by a blue line and the expected pairwise differences by a red line. The mismatch distributions have been determined by using Arlequin version 3.5 (Excoffier and Lischer, 2010) as discussed in Section 5.14.

The composition of the modern-day Tswana population is based on a complex population history over thousands of years that was expected to reflect a combination of the signals of population growth, selection, gene flow and migration of those past populations. Analysing

the population for signals of population expansion or constant size by using statistical measures only, without a phylogenetic dissection of the sequence variation of the population, would have run the risk of conflating the expansion events (Watson *et al.*, 1997). In the case of the Tswana population, this would be true, since the haplogroup composition reflects a major Khoi-San genetic component as well as another important genetic component originating from the Bantu-speaking populations that migrated to southern African regions. Even though the phylogenetic relationships between the individuals of this population supportthe model for constant size, they also assist in highlighting the different genetic components of the population that carry different evolutionary histories. This is evident in the observed L0d clade that displays short terminal branches and extensive mutational development within that lineage that is shared by individuals, as opposed to the expected long terminal branches and high number of mutations within lineages not shared with other individuals of an expanding population. Based on studies by Excoffier and Schneider (1999), hunter-gatherer populations such as the Khoi-San populations of southern Africa commonly present with ragged mismatch distributions that display an increased level of low pairwise difference classes (0 and 1) and insignificant or positive F_S values, indicating models of constant population size without any signatures of past population expansions. This is ascribed to a reduction in effective population sizes within the hunter-gatherer populations caused by post-Neolithic bottlenecks or the infiltration of the agriculturist Bantu-speaking populations into the demographic regions of the hunter-gatherer populations, which caused fragmentation of the hunter-gatherers and therefore smaller effective population sizes (Excoffier and Schneider, 1999). The Tswana population under investigation displays these characteristics even though it contains a component of Bantu-speaking maternal ancestry.

6.9.4 Selection

According to the neutral theory of selection, most evolutionary change at the molecular level is caused by genetic drift of selectively or nearly neutral mutations and not by selective pressure, supporting a neutral mutation-random drift hypothesis (Kimura, 1968; King and Jukes, 1969). Although this approach has been widely accepted, a number of studies have investigated the pattern of selection within the mtDNA coding regions and produced different and sometimes conflicting conclusions with regard to the effect of selective forces in the mtDNA genome (Excoffier, 1990; Weinreich and Rand, 2000; Mishmar *et al.*, 2003; Elson *et al.*, 2004; Ruiz-Pesini *et al.*, 2004). The pattern of polymorphisms in an mtDNA dataset is evidence of evolutionary processes that have had

an impact on the sequence diversity of a population of mtDNA sequences. These evolutionary processes include genetic drift, population growth, gene flow and selection. The mtDNA sequence datasets of this investigation were investigated for signs of population expansion in the previous section. To verify the conclusions about the signals of population growth inferred from the distributions of pairwise nucleotide differences within populations in Section 6.9.3, it was necessary to determine whether the patterns of sequence variation observed could possibly be ascribed to the effect of selection on the mtDNA genomes.

The role of natural selection in mtDNA is important in the conservation of energy production within the cell, to ensure the formation of the essential electron transport chain protein complexes and, in the absence of recombination, fix advantageous polymorphisms that are important to the viability of future mitochondrial generations (Ballard and Rand, 2005). Selection has, for example, been cited as the reason for the higher frequency of conserved amino acids encoded by the mtDNA in human populations that resided in colder climates compared to populations in warmer regions of the world (Mishmar *et al.*, 2003).

The effects of selection on the mtDNA genome of individuals within this study were determined by adopting a whole-genome approach to detect overall violations of the neutral model of evolution. Then a gene-by-gene approach was followed to detect violations of selection within specific regions of the mtDNA genome. Site frequency spectrum methods, which were based on the distribution of nucleotide substitutions within the mtDNA genome, were used to investigate deviation from neutrality in the coding regions of the mtDNA genomes of the Global African, All African, Western, Eastern and Southern African datasets and of the Tswana dataset of this investigation. In addition, the major L haplogroups were also investigated to compensate for the possibility that the regional datasets are not sufficiently representative of all the haplogroup substitutions that have developed within the African populations. The purpose of the broad analysis of the mtDNA of African and non-African individuals in terms of regional location and haplogroup type was to determine the effect of selection on the mitochondrial genome of African individuals in different contexts.

The polymorphism patterns of the mtDNA sequences of the datasets of this investigation were tested for violations to the model of neutrality (Kimura, 1980) by determination of Tajima's *D* statistic (Tajima, 1989) and Fu and Li's D* and F* statistics (Fu and Li, 1993) by

using DnaSP version 5 (Librado and Rozas, 2009). Because of the different effects of the evolutionary forces on sequence variation, it was decided to test not only for the departure of neutrality by testing the sequence frequency spectrum of polymorphisms, but also to test for the excess of rare alleles at low frequencies that have been fixed in a population through selection. Fu and Li's D* and F* statistical tests have been reported to be powerful detectors of background selection and Tajima's *D* statistic has been reported to be a powerful indication of positive selective sweeps and by using these tests in combination, the mechanism responsible for the polymorphism could be investigated. The results of these analyses are presented in Table 6.48.

Table 6.48 Tajima's *D* and Fu and Li's D* and F* test statistics

Population	Number of sequences (*n*)	Tajima's *D*	Fu and Li's D*	Fu and Li's F*
Global African	572	-2.328**	-8.018**	-5.427**
All African	385	-2.550***	-11.738**	-7.711**
Western African	60	-1.568	-2.223	-2.355
Eastern African	110	-2.248**	-4.565**	-4.217**
Southern African	66	-0.976	-1.340	-1.432
Tswana	50	-1.231	-1.824	-1.913
Haplogroup L0	137	-1.911*	-4.807**	-4.126**
Haplogroup L1	49	-1.531	-3.298*	-3.146*
Haplogroup L2	74	-2.284**	-3.860**	-3.852**
Haplogroup L3	97	-2.467**	-5.592**	-5.065**

Statistical significance: *$P < 0.05$; **$P < 0.02$; ***$P < 0.001$; no asterisk = not statistically significant; statistically significant values are highlighted in grey; Global African = dataset 1a; All African = dataset 2a; Tswana = dataset 3a; Western African = dataset 4; Eastern African = dataset 5; Southern African = dataset 6.

The Global African, All African and Eastern African datasets of this investigation display significant levels of negative Tajima *D* and Fu and Li's D* and F* values, which provides evidence of selection or population expansion and/or population subdivision (Simonsen *et al.*, 1995). These results were not unexpected, since the genetic signals of population growth determined in Section 6.9.3 provided evidence of a major population expansion within these respective populations that could have contributed to these results. The negative *D* values further provided evidence of an excess of rare substitutions that have not been removed from the internal branches of the human mtDNA phylogenetic tree, either owing to adaptive selection or because they may be mildly deleterious mutations, which have not yet been removed from the mitochondrial genome by purifying selection. Similar values for global and African populations have been reported elsewhere (Torroni *et al.*, 2001; Kivisild *et al.*, 2006).

It is difficult to isolate selection events from the other population signals and the significant deviations from neutrality indicated by the negative values of Tajima's *D* and Fu and Li's D* and F* test statistics could only be interpreted as evidence that an excess of rare mutations is present in these populations because of adaptive selection, weak purifying selection, population growth or population substructure (Ptak and Przeworski, 2002; Hammer *et al.*, 2004; Kivisild *et al.*, 2004). These findings could be due to the composition of the sample in addition to signals of population growth. This coincides with the high genetic diversity and strong signals of population growth displayed by the All African dataset in Section 6.9.2 and Section 6.9.3.

The Western African dataset does not display significant Tajima *D* or Fu and Li's D* and F* test statistics, which was also expected. As discussed in Section 6.9.3, this dataset consists of a large component of mtDNA sequences belonging to individuals from Burkina Faso, which has demonstrated lower genetic diversity and lower signals of population growth in comparison with the Global African, All African and Eastern African datasets. Kivisild *et al*. (2006) reported the same phenomenon in a Western African population and ascribed it to the substructure of the population, of which one of the subpopulations consisted of individuals that originated from Burkina Faso. The Southern African and Tswana datasets of this investigation also do not display significant negative values for Tajima's *D* and Fu and Li's D* and F* test statistics. There is therefore no evidence that neutrality could be rejected within these populations, which coincides with previously reported signs of constant population growth for these populations, as discussed in Section 6.9.3.

All of the major L haplogroups of the All African dataset of this investigation display statistically significant negative values for Tajima's *D* and Fu and Li's D* and F* tests statistics, providing evidence of selection, population growth and/or population substructure. Haplogroup L1 is the only exception, with a negative Tajima *D* value that is not significant but does display evidence of the rejection of neutrality by the statistically significant negative values for Fu and Li's D* and F* test statistics. The deviation from neutrality detected in the L0 haplogroup is in contrast to the opposite signal being determined in the Southern African and Tswana datasets, which consist of a large component of individuals belonging to the L0 haplogroup. As discussed in Section 6.9.3, low signals for population growth were detected within the Southern African and Tswana datasets that were ascribed to the slow and constant growth of these populations, with specific reference to individuals belonging to haplogroup L0. The deviation from neutrality

detected within the L0 haplogroup therefore most probably indicates the presence of selection or population substructure and not population growth. The same could be inferred for the Western African dataset, which consists of a large component of haplogroup L1 and L3 mtDNA sequences. The non-significant negative Tajima's *D* and Fu and Li's D* and F* values displayed by the Western African dataset were more probably representative of low population growth signals rather than the absence of selection, as opposed to the significant negative Fu and Li's D* and F* values displayed by haplogroups L1 and L3, which reflected selection or population substructure within these datasets. These results were interpreted as evidence that the population signals with regard to population growth in the previous section were indeed reflective of the population size in the case of the Western African, Southern African and Tswana datasets of this investigation.

The mtDNA datasets were further investigated by determining the number of synonymous (S) and nonsynonymous (NS) substitutions for each of the different haplogroups, i.e. L0, L1, L2 and L3, of the All African dataset of this investigation for the reasons discussed in Section 5.14.4. The data for each of the L haplogroups were pooled to determine the presence of selection in a dataset that was representative of the African continent and included the Tswana population under investigation. Selective constraints or negative selection was determined by this approach, based on the principle that the number of NS substitutions in the haplogroup-associated class was expected to be less than the number of private NS substitutions because the older haplogroup-associated substitutions would have been exposed to either purifying or directional selection over a longer period of time than the younger private substitutions.

In this investigation, the number of mutational classes has furthermore been expanded from two classes, S and NS substitutions, to four classes, haplogroup-associated synonymous and haplogroup-associated nonsynonymous substitutions and private synonymous and private nonsynonymous substitutions, to enable greater insight into the selective forces present, as discussed in Section 5.14.4. Private substitutions usually occur in only one individual and are therefore only connected to one other haplotype in the phylogenetic tree, whereas the haplogroup-associated substitutions are shared by more than one haplotype and are connected to two or more haplotypes within a haplogroup. These statistical measures depend on an unbiased and accurate count of the different classes of substitutions identified in the coding regions of the mtDNA sequences of the haplogroups of the All African dataset. The NJ and MP phylogenetic trees as described in

Section 6.8.3 and Section 6.8.4 were used to determine the classes of substitutions. A conservative counting approach was followed by disregarding homoplastic events by counting identical substitutions that occurred multiple times in a haplogroup only once (Elson *et al.*, 2004). The results of the NS/S ratios for both haplogroup-associated and private substitutions of the 13 protein-coding genes of the mtDNA of each of the L haplogroups of the All African dataset are presented in Table 6.49.

Table 6.49 NS/S_H and NS/S_P ratios for the 13 protein coding genes of the mtDNA of African individuals that belonged to haplogroups L0, L1, L2 and L3

Gene	Haplogroup L0 (*n*=137)		Haplogroup L1 (*n*=49)		Haplogroup L2 (*n*=74)		Haplogroup L3 (*n*=97)		Total[b] (n=357)	
	NS/S_H	NS/S_P	NS/S_H	NS/S_P	NS/S_H	NS/S_P	NS/S_H	NS/S_P	NS/S_H	NS/S_P
ND1	6/18 (0.33)	5/8 (0.63)	2/8 (0.25)	5/3 (1.60)	3/9 (0.33)	7/8 (0.88)	3/13 (0.23)	5/18 (0.28)	14/48 (0.29)	22/37 (0.59)
ND2	5/15 (0.33)	4/14 (0.29)	3/6 (0.50)	4/11 (0.36)	4/4 (1.00)	2/11 (0.18)	3/12 (0.25)	7/17 (0.41)	15/38 (0.39)	17/53 (0.32)
COI	6/18 (0.33)	9/22 (0.41)	5/17 (0.29)	0/3 (0.00)	3/11 (0.27)	6/12 (0.50)	1/10 (0.10)	7/21 (0.33)	15/56 (0.27)	22/58 (0.38)
COII	1/8 (0.13)	4/11 (0.36)	1/9 (0.11)	4/3 (1.30)	1/3 (0.33)	2/7 (0.23)	0/8 (0.00)	0/7 (0.00)	3/28 (0.11)	10/28 (0.36)
ATP8	3/6 (0.50)	3/2 (1.50)	0/1 (0.00)	2/3 (0.67)	1/1 (1.00)	2/1 (2.00)	0/2 (0.00)	6/1 (6.00)	5/37 (0.16)	13/33 (0.39)
ATP6	7/11 (0.64)	6/9 (0.67)	2/5 (0.40)	4/4 (1.00)	3/6 (0.50)	5/7 (0.71)	5/6 (0.83)	8/11 (0.73)	17/28 (0.61)	23/31 (0.74)
COIII	2/13 (0.15)	5/12 (0.42)	1/8 (0.13)	4/6 (0.67)	1/7 (0.14)	1/6 (0.17)	1/9 (0.11)	3/9 (0.33)	5/37 (0.16)	13/33 (0.39)
ND3	5/0 (0.00)	1/2 (0.50)	2/0 (0.00)	2/1 (2.00)	1/0 (0.00)	0/1 (0.00)	2/4 (0.50)	5/4 (1.25)	10/4 (2.50)	8/8 (1.00)
ND4L	1/4 (0.25)	1/3 (0.33)	0/2 (0.00)	0/2 (0.00)	0/2 (0.00)	0/2 (0.00)	0/3 (0.00)	1/1 (1.00)	1/11 (0.09)	2/8 (0.25)
ND4	4/26 (0.15)	10/20 (0.50)	2/15 (0.13)	3/13 (0.23)	1/9 (0.11)	5/4 (1.25)	1/11 (0.09)	4/24 (0.17)	8/61 (0.13)	22/61 (0.36)
ND5	9/25 (0.36)	14/26 (0.54)	8/16 (0.50)	9/16 (0.56)	6/11 (0.55)	10/18 (0.56)	7/14 (0.50)	7/28 (0.25)	30/66 (0.45)	40/88 (0.45)
ND6	3/10 (0.30)	1/8 (0.13)	0/1 (0.00)	4/3 (1.30)	0/4 (0.00)	6/6 (1.00)	0/6 (0.00)	8/15 (0.53)	3/21 (0.14)	19/32 (0.59)
CytB	12/12 (1.00)	14/12 (1.17)	5/5 (1.00)	6/15 (0.40)	8/10 (0.80)	10/10 (1.00)	11/12 (0.92)	8/21 (0.38)	36/39 (0.92)	38/58 (0.66)
Total[a]	64/166 (0.39)	77/149 (0.52)	31/91 (0.34)	49/83 (0.59)	32/77 (0.42)	56/93 (0.60)	34/110 (0.31)	69/177 (0.40)	161/447 (0.36)	249/502 (0.50)

ND1: NADH dehydrogenase subunit 1; *COI*: Cytochrome c oxidase subunit I; *COII*: Cytochrome c oxidase subunit II; *ATP8*: ATP synthase F0 subunit 8; *ATP6*: ATP synthase F0 subunit 6; *COIII*: Cytochrome c oxidase subunit III; *ND2*: NADH dehydrogenase subunit 2; *ND3*: NADH dehydrogenase subunit 3; *ND4L*: NADH dehydrogenase subunit 4L; *ND4*: NADH dehydrogenase subunit 4; *ND5*: NADH dehydrogenase subunit 5; *ND6*: NADH dehydrogenase subunit 6; *Cytb*: Cytochrome b; NS/S_H = ratio of haplogroup associated nonsynonymous substitutions to synonymous substitutions; NS/S_P = ratio of private nonsynonymous substitutions to synonymous substitutions; values indicated in parentheses; a = total NS/S for all genes of individuals of respective haplogroups L0, L1, L2 and L3; b = total NS/S for individual genes over all the haplogroups and representative of All African dataset.

One hundred and sixty-one NS_H substitutions, 447 S_H substitutions, 249 NS_P substitutions and 502 S_P substitutions were observed in total for all the genes across all the haplogroups. The higher number of observed private substitutions was ascribed to their

evolutionary younger age in comparison to the haplogroup-associated substitutions, which would have been exposed to selective forces for a much longer time and therefore removed if the substitutions had been deleterious (Elson *et al.*, 2004). The smaller number of NS_H substitutions in contrast to the observed number of NS_P substitutions for the whole All African dataset over all the genes provides evidence that the mtDNA could have been subjected to negative selection or to relaxed selective constraints (Elson *et al*., 2004).

The possibility of positive selection as indicated by higher numbers of NS_H than NS_P is displayed by the *ND2, ND3* and *CytB* genes when the ratios are determined over all the major haplogroups. The ratios for these genes within each of the haplogroups do not always coincide with the ratio determined when the data are pooled. This was expected, since the sample sizes of the individual haplogroup datasets were smaller than the sample size of the pooled L haplogroup. Other studies also reported higher NS_H than NS_P substitutions for *ND2* and *ND3* genes but not for the *CytB* gene and lower numbers of NS_H substitutions than NS_P substitutions for the other genes, as observed in this investigation. This observation confirmes the possibility of purifying selection occurring in the mtDNA genomes of African populations (Elson *et al.*, 2004; Ingman and Gyllenstein, 2007). Ingman and Gyllenstein (2007) further reported a high k*a*/k*s* ratio for *ATP6,* which was considered to be evidence of the possibility of the *ATP6* gene being under positive selection. When the substitutions were, however, evaluated by taking their age into account, as happened in this investigation, positive selection was not evident, although there were a high number of NS_H substitutions present within this gene. When interpreted in the context of the number of NS_P substitutions, it was observed that the NS/S_H ratio was slightly smaller than the NS/S_P ratio, indicating a small deviation from neutrality towards purifying selection, which was also reported by Ingman and Gyllenstein (2007) after using additional methods to analyse the *ATP6* gene.

The ratio of NS/S_H to NS/S_P was smaller for all the haplogroups over all the genes, indicating that the NS substitutions had been removed from the mtDNA genome and therefore purifying selection was present in the mtDNA. This was in accordance with other studies that concluded that purifying selection acts on all mtDNA genes (Elson *et al.*, 2004; Kivisild *et al.*, 2006; Ingman and Gyllenstein, 2007).

The effects of selection were further described by determining the neutrality index (NI) of each of the genes within each of the major haplogroups of the All African dataset of this investigation. The NI values > 1 indicate purifying selection because of the excess of NS_P

substitutions and NI values < 1 indicate directional or positive selection based on the excess of NS_H substitutions. The Fisher's exact test was used to determine the *P* values for each of the NS and S estimates (Templeton, 1996). The results are presented in Table 6.50.

Table 6.50 NI and P values for the 13 protein coding genes of the mtDNA of African individuals that belonged to haplogroups L0, L1, L2 and L3

Gene	Haplogroup L0 (*n*=137)		Haplogroup L1 (*n*=49)		Haplogroup L2 (*n*=74)		Haplogroup L3 (*n*=97)		Total[b] (n=357)	
	NI	***P***	**NI**	***P***	**NI**	***P***	**NI**	***P***	**NI**	***P***
ND1	1.90	0.313	6.40	0.088	2.65	0.226	1.20	0.575	2.03	0.201
ND2	0.88	0.573	0.72	0.539	0.18	0.115	1.65	0.403	0.82	0.382
COI	1.24	0.491	0.00	0.496	1.85	0.368	3.30	0.262	1.41	0.236
COII	2.77	0.356	11.82	0.060	0.87	0.706	0.00	---	3.27	0.072
ATP8	3.00	0.343	0.00	0.667	2.00	0.700	0.00	0.083	4.65	**0.039**
ATP6	1.05	0.614	2.50	0.378	1.42	0.528	0.73	0.579	1.21	0.390
COIII	2.80	0.254	5.15	0.184	1.21	0.733	3.00	0.368	2.44	**0.049**
ND3	0.00	0.107	0.00	0.600	0.00	0.500	2.50	0.378	0.40	0.206
ND4L	1.32	0.722	0.00	---	0.00	---	1.00	0.400	2.78	0.429
ND4	3.33	0.063	1.77	0.470	11.36	**0.049**	1.89	0.523	2.77	**0.017**
ND5	1.50	0.296	1.12	0.542	1.02	0.617	0.50	0.212	1.00	0.556
ND6	0.43	0.450	0.00	0.500	0.00	0.115	0.00	0.114	4.21	**0.024**
CytB	1.17	0.505	0.33	0.221	1.25	0.493	0.41	0.112	0.72	0.172
Total[a]	1.33	0.090	1.72	**0.030**	1.43	0.107	1.29	0.200	1.39	**0.004**

ND1: NADH dehydrogenase subunit 1; *COI*: Cytochrome c oxidase subunit I; *COII*: Cytochrome c oxidase subunit II; *ATP8*: ATP synthase F0 subunit 8; *ATP6*: ATP synthase F0 subunit 6; *COIII*: Cytochrome c oxidase subunit III; *ND2*: NADH dehydrogenase subunit 2; *ND3*: NADH dehydrogenase subunit 3; *ND4L*: NADH dehydrogenase subunit 4L; *ND4*: NADH dehydrogenase subunit 4; *ND5*: NADH dehydrogenase subunit 5; *ND6*: NADH dehydrogenase subunit 6; *Cytb*: Cytochrome b; NI = neutrality index; a = total NI and *P* for all genes of individuals of respective haplogroups L0, L1, L2 and L3; b = total NI and *P* for individual genes over all the haplogroups and representative of All African dataset; *P* values in **bold** are significant, P < 0.05; --- indicates that P values could not be determined.

When all 13 mitochondrial genes were evaluated together within each of the major haplogroups, the results displayed evidence of purifying selection within all the haplogroups. These results were, however, not statistically significant except for haplogroup L1, which displayed a statistically significant NI value of 1.72. It could therefore be concluded that a model of neutrality was rejected and that purifying selection was significantly present in this haplogroup. Neutrality could not be rejected for the other haplogroups, as was also reported in a study by Elson *et al.* (2004), in which statistically significant evidence for the rejection of neutrality for any of the haplogroup L groupings that were analysed was not present. The presence of significant evidence for purifying selection in the L1 haplogroup dataset is surprising and has not been reported elsewhere in the literature. Even though the NI value of the L1 haplogroup is the highest of all the haplogroups, it is similar to that observed in the other haplogroups. The significance of

purifying selection in this haplogroup is ascribed to the smaller size of the dataset in comparison with the other L haplogroup datasets and the consequent presence of four NI values of zero, probably because of the smaller sample of substitutions. Therefore the statistical significance displayed by the differences between the NS and S substitutions in the L1 haplogroup cannot be regarded as absolute evidence that purifying selection is stronger in this haplogroup in comparison to the other major haplogroups of this investigation. When the results were pooled for all the genes over all the haplogroups, a statistically significant NI result of 1.39 was obtained that confirmed the presence of purifying selection in the mitochondrial genome as a whole, as was also reported by other studies (Elson *et al.*, 2004; Kivisild *et al.*, 2006; Ingman and Gyllenstein, 2007). Although the respective major haplogroups L0, L2 and L3 do not display statistically significant signs of purifying selection, the NI values of these haplogroups are greater than one, which therefore does indicate evidence of purifying selection. It is postulated that the nature and size of the samples used was the reason for the discrepancy between the statistical significance of the NI results of the respective haplogroups and the total statistically significant NI result of the pooled haplogroup dataset.

It has further been observed that the NI values vary considerably across genes within the respective haplogroups, which could be interpreted as evidence that the selective forces are not uniform across all the protein-coding genes of the mtDNA and that because of constraints imposed on the patterns of substitutions by the lack of recombination within the mtDNA, it leads to whole genes being affected by the net effect of either directional or purifying selection across the mtDNA (Rand and Kann, 1996). This implies that selective pressure within a gene of the mtDNA may affect the whole gene through linkage, which would be evident in a variation of NS to S ratios over whole genes.

When the data of the individual genes were pooled over all the haplogroups, nine of the genes displayed NI values greater than one; one gene, *ND5*, displayed an NI value equal to one and only three genes, i.e. *ND2*, *ND3* and *CytB*, displayed NI values lower than one. *ND5* therefore displayed absolute neutrality as opposed to the *ND2*, *ND3* and *CytB* genes that displayed low NI values, which suggested that directional selection or relaxed purifying selective constraints were present in those genes. The *P* values of these genes, however, were not significant and neutrality could therefore not be rejected. This observation was in accordance with the reported signs of directional selection in some of the genes of all the major haplogroups within the African datasets of other studies (Ingman and Gyllenstein, 2007; Mishmar *et al.*, 2003; Elson *et al.*, 2004; Ruiz-Pesini *et al.*, 2004).

The majority of the mitochondrial genes displayed higher NI values, which indicated purifying selection. The *ATP8*, *COII*, *ND4* and *ND6* genes displayed statistical significance and neutrality was rejected for those genes. The *ND1*, *COI*, *ATP6*, *COIII* and *ND4L* genes also displayed NI values greater than one, but were not statistically significant and neutrality was therefore not rejected for those genes.

The study of the effects of selection on the human genome is complex. Many of the signals observed in the substitution patterns can be interpreted as evidence of more than one type of selective pressure and the traces of negative selection will be similar to those of relaxed positive selection (Ingman and Gyllenstein, 2007). Furthermore, the functional domain of the gene is more conserved than the remaining parts of the gene owing to the functional dependence on the subunits of the OXPHOS system encoded by the mitochondrial genome (Kivisild *et al.*, 2006). The conserved functional domains are under different levels of purifying selection that are not applied to the rest of the genes and therefore the signal of the whole gene may be misleading in terms of the direction and strength of the selective pressure within the gene (Kivisild *et al.*, 2006). The gene-specific signals observed in this investigation could therefore be explained by these phenomena.

In summary, statistically significant deviation from neutrality is displayed by all the major L haplogroups of the All African dataset as well as by the Global Africa, All African and Eastern African datasets. The Western, Southern African and Tswana datasets do not display statistically significant deviation from neutrality, which could be explained by the absence of selective forces within these populations or more likely, by possible population substructures. The significance of the deviation from neutrality in the other datasets could be ascribed not only to selection, but also to past population expansion events. The further analyses of the NS to S ratios of the private and haplogroup-associated substitutions within the All African dataset and the respective haplogroups of the All African dataset displayed statistically significant evidence of purifying selection over the pooled All African dataset. However, except for haplogroup L1, none of the other L haplogroups displayed signs of statistically significant deviations from neutrality in the direction of purifying selection and it was concluded that the deviation from neutrality displayed by the negative Tajima's *D* and Fu and Li's D* and F* tests statistics was influenced by past population expansion events. Investigation of the individual genes of the mtDNA displayed signals of different selective strengths and directions over the different genes. Only statistically significant signals of purifying selection could be observed in four of the individual genes. Based on the evidence displayed for purifying selection only presented by the larger

pooled All African dataset and haplogroup L1 and in four individual genes, it was postulated that the mitochondrial genome was under weak purifying selection. Based on this conclusion, the signals of population growth observed for the Global African, All African, Eastern and Western African datasets in Section 6.9.3 were true and not influenced by the effects of positive selection. The Tswana population of this investigation would therefore have been under weak purifying selection and constant slow population growth.

6.9.5 Population genetic structure

It is generally accepted that AMH originated about 200 kya in sub-Saharan Africa and dispersed across the continent of Africa and eventually to the rest of the world (Salas *et al.*, 2002; Torroni *et al.*, 2006). During these dispersals, the human populations became divided and isolated from one another and differentiated on a genetic level because of limited gene flow with other populations, genetic drift and selection (Salas *et al.*, 2002). The genetic structure of populations is determined by these evolutionary events and therefore provides information about the behaviour of a population in the context of other populations. The genetic structure of the Tswana population in this investigation was investigated in the context of the Global African, All African, Western, Eastern and Southern African datasets, with the aim of determining how far genetically removed the Tswana individuals were from the other African populations and to determine in which way this population was genetically differentiated, in order to interpret the genetic signals of evolutionary processes correctly.

Population substructure has an impact on the patterns of genetic diversity by causing an increased number of segregating sites in comparison to the average number of pairwise nucleotide differences between pairs of sequences. The genetic diversity is drastically reduced when inbreeding is frequent (Tajima, 1993; Fu, 1997). In addition, signals of deviation from neutrality can be present in populations where selection is in actual fact absent (Tajima, 1993; Simonsen *et al.*, 1995). Population parameters of genetic diversity, population size and selection were discussed in Section 6.9.2, Section 6.9.3 and Section 6.9.4. To verify that the interpretations made from the statistical measures obtained for those parameters were true, it was important to analyse the datasets for population differentiation to rule out any possibility that the previous results had been influenced by substructure in any of the populations investigated.

The genetic structure within and among populations was determined by AMOVA and the *F* statistics (Excoffier *et al.*, 1992) using Arlequin software version 3.5.1.2 (Excoffier and Lischer, 2010). Wright (1965) developed the concept of a fixation index (*F*) as a measure of the amount of fixation due to genetic drift present in a population. *F* was determined among all the populations of this investigation, as discussed in Section 5.14.5, by calculating the F_{ST} indices between populations from haplotype frequencies. Statistical significance was determined by using a non-parametric permutation procedure, as discussed in Section 5.14.5. The F_{ST} estimate was regarded as the correlation of the covariance component of random haplotypes between two populations, with the covariance components of the total molecular variance of random haplotypes in all the populations within the respective groups, as displayed in Table 6.51 (Excoffier *et al.*, 1992). Therefore, low F_{ST} values were interpreted as an indication of a small amount of variation between the haplotypes of populations within a group and therefore that population differentiation was small. Small population differentiation was further expected to be reflected by a low percentage of variation among populations. The results are presented in Table 6.51.

Table 6.51 Analysis of molecular variance (AMOVA) between populations of this investigation

Groups	Percentage variation WP	Percentage variation AP	AMOVA F_{ST} (*P*)
Global African vs Tswana	85.01	14.99	0.150 (0.000±0.000)
All African vs Tswana	86.23	13.77	0.138 (0.000±0.000)
Western African vs Tswana	78.64	21.36	0.214 (0.000±0.000)
Eastern African vs Tswana	82.58	17.42	0.174 (0.000±0.000)
Southern African vs Tswana	94.91	5.09	0.051 (0.00094±0.00023)
Eastern African vs Western African	97.78	2.22	0.022 (0.00294±0.00043)
Eastern African vs Southern African	80.90	19.10	0.191 (0.00000±0.00000)
Western African vs Southern African	75.74	24.26	0.243 (0.00000±0.00000)

AMOVA = analysis of molecular variation; F_{ST} = fixation index; groups = populations within groups are compared with one another and groups are analysed separately from one another; *P* = statistically significant < 0.05 after 16,000 permutations, WP = within populations; AP = among populations; variation presented as percentages; Global African = dataset 1a; All African = dataset 2a; Tswana = dataset 3b; Western African = dataset 4; Eastern African = dataset 5; Southern African = dataset 6.

In agreement with previous studies, the results indicate that the most genetic variation is observed within populations rather than between populations (Excoffier *et al.*, 1992; Salas *et al.*, 2002; Coudray *et al.*, 2009). It has been reported that 85% - 90% of the genetic variance in the human species occurs within populations and that only 10% - 15% of the genetic variation can be ascribed to differences between populations. Levels of genetic variance of up to 21% between populations of different regions in Africa have been

reported (Salas *et al.*, 2002; Coudray *et al.*, 2009), as in the case of this investigation, which detected high levels of genetic variance between the Eastern and Southern African and Western and Southern African datasets, which displayed values of 19.10% and 24.26% variance respectively. This can be explained by the genetic heterogeneity that developed within populations of Africa that became isolated from one another through dispersal into different regions of the continent (Salas *et al.*, 2002; Behar *et al.*, 2008). The observed differentiation between the western and eastern regions of Africa and the southern region of Africa is attributed to the divergent haplogroups observed in the mtDNA datasets of this investigation that most probably developed in response to the occurrence of new mutations within populations as they migrated further from one another and genetic flow became limited between them. This phenomenon was reported by several other studies that provided strong evidence that the ancient Khoi-San populations that resided in the southern African regions originated from an early split from the other ancient populations in eastern Africa at the time and that this isolated existence resulted in the development of haplogroups L0d and L0k that were mainly limited to the Khoi-San speaking populations of southern Africa (Behar *et al.*, 2008). The fixation indices between these populations were in agreement with the level of variation and indicated that the populations were significantly differentiated from each other, with F_{ST} values that were high in comparison with the other F_{ST} values observed.

The level of genetic heterogeneity between the Western and Eastern African datasets was, however, surprising. Although significant, the level of between-population variance was only 2.22%, with an F_{ST} value that was similarly low at 0.022. It has been reported that mitochondrial genome diversity is structured according to geographical region (Salas *et al.*, 2002; Poloni *et al.*, 2009; Coudray *et al.*, 2009), which is true of the differentiation displayed by the populations between the southern African and western and eastern African regions. A similar observation of low level genetic differentiation between Eastern and Western African populations was reported by Poloni *et al.* (2009) in addition to further reports of Bantu-speaking populations that displayed variance of as low as 1.2% between populations (Salas *et al.*, 2002; Coudray *et al.*, 2009). This could be explained by the many haplogroups that are shared by several different Bantu-speaking populations, such as haplogroup L2, which has been reported to be the most common haplogroup present in all regions of the African Bantu-speaking populations (Salas *et al.*, 2002). This is attributed to the migrations of the Bantu-speaking populations about 4 kya via a western and eastern route across Africa into the southern regions. Through the migration and newly acquired agricultural lifestyles in which the populations became established in environments that

promoted agricultural practices, the genetic lineages were assimilated and existing clades became enriched by the introgression of different haplogroups and their derivatives (Plaza *et al.*, 2004). This result therefore highlights the fact that although the signals of population differentiation between the Bantu-speaking populations of Africa are significantly differentiated from one another, they still display a strong genetic link between them.

In contrast to the low genetic differentiation between the Eastern and Western African regional populations of this investigation, the F_{ST} values between the Southern African regional population and the Eastern and Western African regional populations are high, implying that the Eastern and Western African regional populations are strongly differentiated from the Southern African populations. The genetic variation among the Southern African and Eastern and Western African populations are highest, with 19.10% and 24.26% respectively. This result was expected in light of the fact that the Bantu-speaking populations, that were mostly represented in the Eastern and Western African datasets of this investigation, were separated from the early hunter-gatherer populations, that were mostly represented in the Southern African dataset of this investigation, and experienced different evolutionary histories in isolation from one another (Excoffier and Schneider, 1999; Behar *et al.*, 2008; Batini *et al.*, 2011).

The Tswana population under investigation displays statistically significant signals of population differentiation from the other datasets of this investigation. When grouped with the Global African and All African datasets, it displays genetic variation of 14.99% and 13.77% respectively, accompanied by significant F_{ST} values that indicate genetic differentiation from these two large haplogroup L datasets. Similar results have been reported by Coudray *et al.* (2009) among groups of large African and non-African populations. This result is explained by the geographically linked development of the mitochondrial genome and underlines the development of a differentiated genetic structure in populations based on several evolutionary events (Salas *et al.*, 2002; Coudray *et al.*, 2009).

The Tswana population under investigation further displays statistically significant F_{ST} values when compared with the regional datasets of this investigation. In agreement with the genetic variation between the Eastern and Western African regional populations and the Southern African population, the genetic variation is similar between the Eastern and Western African regional populations and the Tswana population; 17.42% and 21.36%

respectively. These observations are further supported by the relatively high F_{ST} values of 0.214 and 0.174. On investigation of the genetic differentiation between the Southern African population of this investigation, which consisted mostly of Khoi-San populations of southern Africa, it was observed that the genetic variation was lower (5.09%) than among the other regional populations and that the F_{ST} value was also lower (0.051). These results indicate that the Tswana population of this investigation is significantly differentiated from the other populations considered in this investigation and that the genetic signals of Tswana-speaking individuals of this investigation are therefore not biased by it being a subpopulation of another African population. Although the Tswana population contains haplogroups that are connected to the Bantu-speaking populations that migrated to the southern regions of Africa, it harbours a much closer genetic relationship with the Khoi-San speaking populations on a maternal level. The phenomenon of high levels of differentiation between the Tswana population and the Bantu-speaking populations of different regions of Africa and the much lower levels of genetic differentiation between the Tswana population of this investigation and the Southern African dataset, that were mostly comprised of Khoi-San speaking individuals, was also observed when the Y chromosome and autosomal short tandem repeat polymorphism data of Bantu-speaking populations of southeastern Africa were compared with each other (Lane *et al.*, 2002). This suggested that these populations had not been isolated for long enough to differentiate into populations that harboured strong genetic differences. The main difference between the Y chromosome data of southeastern African Bantu-speaking populations and the mtDNA data of the Tswana population observed in this investigation is the fact that the paternal lineages are strongly linked to Bantu-speaking ancestors and that the maternal lineages are strongly linked to the Khoi-San speaking populations of southern Africa. Unfortunately the Southern African dataset of this investigation does not contain a sufficient number of Bantu-speaking individuals to investigate the differentiation of the Tswana population from the other Bantu-speaking populations of southern Africa on a maternal level, which would be an important aspect to investigate. It is clear, however, that the genetic differentiation of the Tswana population under investigation from the ancestral eastern and western African Bantu-speaking populations is most likely due to the combined effect of the assimilation of Khoi-San females into these populations and genetic drift over the past two thousand years.

6.9.6 Coalescence-time estimation

The coalescence-times of each of the major lineages observed in the Tswana population under investigation were determined to verify the ancestral origins of the individuals of this population through the alignment of the time depths of the genetic dates with archaeological, climatic and other reported genetic evidence. The major lineages of the Tswana population of this investigation that were dated are presented in Table 6.52. The most recent ancestor of these lineages was also dated to provide information about the time of divergence of the lineages.

Table 6.52 Maternal lineages of the Tswana population of this investigation

Haplogroup classification[a]	Tswana dataset sequence name	Total number of individuals/ lineage
L0a'b'f	TS_44_5063	1
L0a1b	TS_09_4063	3
	TS_21_3471	
	TS_46_4032	
L0a2a	TS_33_3461	4
	TS_01_2074	
	TS_14_3075	
	TS_38_3506	
L0k	TS_28_4034	1
L0d1	TS_40_4117	1
L0d1a	TS_24_3505	4
	TS_32_4056	
	TS_34_4075	
	TS_41_5044	
L0d1b	TS_02_2075	6
	TS_09_4063	
	TS_12_3027	
	TS_27_4027	
	TS_36_4083	
	TS_48_5085	
L0d1c	TS_29_4037	1
L0d2a	TS_11_3002	1

Table 6.52 Continued…

Haplogroup classification[a]	Tswana dataset sequence name	Total number of individuals/ lineage
L0d2a1	TS_05_2091	9
	TS_07_2095	
	TS_17_3117	
	TS_18_3236	
	TS_19_3459	
	TS_25_4111	
	TS_31_4051	
	TS_37_4089	
	TS_47_5083	
L0d3	TS_04_2082	5
	TS_13_3066	
	TS_22_3486	
	TS_35_4080	
	TS_50_5091	
L1c2a	TS_06_2093	2
	TS_23_3015	
L2a1a	TS_10_2103	2
	TS_45_5066	
L2a1b	TS_16_3107	2
	TS_20_3466	
L2a1c	TS_26_4013	3
	TS_39_4106	
	TS_43_5062	
L2a1d	TS_42_5060	1
L2a1f	TS_03_2077	2
	TS_49_5086	
L3d1a	TS_15_3085	1
L3e1	TS_30_3495	1

a = classification according to PhyloTree classification system (Van Oven and Kayser, 2009)

A model-free approach, in which no assumptions were made about the prehistory of the Tswana population under investigation, was used in this investigation to determine the time depth of coalescence times of the different lineages in the Tswana population. This approach was used because no other population genetic parameters with regard to the demographic processes and population structure were needed, as these analyses were performed in other sections of this chapter. Model-free approaches are reportedly robust and useful for the purpose of determination of the time depth of ancestral types (Bandelt *et al.*, 2006a) and therefore suitable for use in this investigation.

Only the substitutions (i.e. indels) were excluded, of the coding regions of the mtDNA sequences used for the determination of the coalescence-time estimates owing to the mtDNA evolving in a clock-like manner and therefore affording a more reliable measurement of the substitutions that evolved over time. The control region of the mitochondrial genome did not adhere to a clock-like evolutionary rate because of the high mutation rate heterogeneity between sites and mutational hotspots and thus some regard it as unsuitable as a genetic marker for coalescence-time estimates (Ingman *et al.*, 2000). Although it has recently been proven that the coding region does exhibit reverse mutations and homoplasy, it is generally accepted that the control region of the mtDNA genome displays much higher rates of homoplasy and incidences of reverse mutations than the coding region because of the high mutation rate (Bandelt *et al.*, 2006a). This could further distort the true number of substitutions that occurred over time and therefore lead to an underestimation of coalescence-times (Tamura and Nei, 1993; Excoffier and Schneider, 1999). Although the rate of mutation in the coding region of the mtDNA is generally regarded as clock-like, it has been reported that the accumulation of mutations in the human mtDNA coding region is nonlinear as a result of purifying selection, as has been proven to be at play in the mtDNA sequences of this investigation, albeit at low levels, as discussed in Section 6.9.4. As a result, the coalescence-time estimates for the lineages of the terminal branches, where the effects of deleterious substitutions would not have been removed by selection yet, could lead to the overestimation of coalescence-times by methods that assume a linear coding-region clock. This phenomenon has not been reported to bias coalescence-time estimates to such an extent that use of coding region sequence data is not advisable under a linear coding-region clock (Macaulay *et al.*, 2005) and it has therefore been used in this investigation. The results of this investigation were interpreted by taking the possibility of overestimation of coalescence-time estimates into account. To account further for the different frequencies of the nucleotides and the bias of transitions to transversions, an evolutionary rate determined by using the HKY85 substitution model for sequence evolution was used (Hasegawa *et al.*, 1985; Mishmar *et al.*, 2003).

The coalescence-time estimates were determined by calculation of the unbiased estimator of the average genetic distance to the root of a node, rho (ρ), based on the All African MP tree lineage groupings as discussed in Section 6.8.4. The method of Forster *et al.* (1996) was used, as described in Section 5.14.6, after which the ρ values were converted to absolute time by using an evolution rate of 1.26×10^{-8} (Mishmar *et al.*, 2003). The SD (σ)

was determined according to the method of Saillard *et al.* (2000) and also converted to absolute time. Results are presented in Table 6.53.

Table 6.53 Coalescence-time estimates of the All African dataset of this investigation

Lineage[a]	Number of sequences	Coalescence-time estimates (± SD)
L0	136	141,295 ± 26,717
L0a'b'f'k	64	100,704 ± 22,607
L0a'b'f	49	77,070 ± 19,524
L0a'b	41	68,335 ± 18,496
L0a1	15	25,690 ± 11,303
L0a1b	10	21,579 ± 10,276
L0a2	25	34,938 ± 13,358
L0a2a	10	12,331 ± 7,707
L0k	15	25,690 ± 11,303
L0d	72	83,235 ± 20,654
L0d1	48	54,976 ± 16,441
L0d1a	9	21,065 ± 10,276
L0d1b	17	40,590 ± 14,386
L0d1c	22	27,745 ± 11,817
L0d2	15	57,203 ± 16,955
L0d2a	12	48,811 ± 15,414
L0d3	9	53,092 ± 16,489
L1c	27	105,842 ± 23,121
L1c2	5	53,435 ± 16,441
L2a	61	51,380 ± 15,927
L2a1	46	25,176 ± 11,303
L2a1a	10	11,817 ± 7,707
L2a1b	5	8,220 ± 6,165
L2a1c	9	23,121 ± 10,789
L2a1d	5	28,772 ± 11,817
L2a1f	8	8,991 ± 6,679
L3d	19	47,783 ± 15,414
L3d1	11	49,324 ± 15,414
L3e	35	50,352 ± 15,927
L3e1	7	15,414 ± 8,734

Coalescence-time estimates presented as years before present (ybp); number of sequences obtained from All African dataset (2a) includes the Tswana dataset (3b); a = lineages according to the PhyloTree classification system (Van Oven and Kayser, 2009); SD = standard deviation presented here by conversion of σ values to absolute time estimates.

The coalescence date of 141,295 ± 26,717 ybp displayed by haplogroup L0 of the All African dataset of this investigation is in agreement with archaeological and fossil evidence that suggest an origin of modern humans around 115,000 – 130,000 ybp ago (Grün *et al.*, 1990) in eastern Africa (McDougall *et al.*, 2005). Furthermore, in addition to

the archaeological and paleobiological evidence, it is also in agreement with the genetic coalescence-time estimates of the divergence of the L0 haplogroup from the L1'5 haplogroup lineages, which reportedly occurred within a time frame of 103,600 ybp up to 210,000 ybp according to different studies (Ingman *et al.*, 2000; Salas *et al.*, 2002; Gonder *et al.*, 2007; Behar *et al*., 2008; Batini *et al.*, 2011). The wide range of time estimates may be attributed to the different regions of the mtDNA genome, types of data, methods and mutation rates that were applied in the determination of the TMRCA by the different studies, as summarised in Table 6.54. The size and origin of the populations used in these time estimations, not presented here, also contributed greatly to the wide range of time estimates reported.

Table 6.54 Coalescence-time estimates published for haplogroup L0

Published study	MtDNA region	Type of marker	Coalescence-time estimation
Ingman *et al.*, 2000	Coding region	MtDNA sequence data	150,000-170,000
Chen *et al.*, 2000	Complete genome	HR-RFLP	122,800-161,800
Gonder *et al.*, 2007	Coding region	MtDNA sequence data	121,300-171,500
Behar *et al*., 2008	Coding region	MtDNA sequence data	140,000-210,000
Batini *et al*., 2011	Coding region	MtDNA sequence data	103,600–142,400

HR-RFLP = high resolution restriction fragment length polymorphism; coalescence-time estimates are presented as years before present (ybp) and include the standard deviations by presenting these as a time range.

The coalescence-time estimate of the L0a'b'f'k lineage was reported by Behar *et al.* (2008) as between 133,000 ybp and 155,000 ybp and was ascribed to a time of sporadic settlements of early populations of modern humans in response to unfavourable climate fluctuations. The slightly earlier time of divergence of the L0a'b'f'k lineages from the L1'5 lineages, as displayed by the results of this investigation, was possibly due to the smaller sample size used in this investigation as opposed to the broad population samples that were included in the Behar *et al.* (2008) study. The smaller population sample was most probably less representative of the L haplogroups than the study conducted by Behar *et al.* (2008) and would therefore not have exhibited a wide range of substitutions.

The early divergence date estimated for the L0a'b'f lineage of the All African dataset of this investigation, from the branch that would become the L0k haplogroup, supports the hypothesis that the L0a'b'f' lineages split from the L0d and L0k lineages early in the history of modern humans (Salas *et al.*, 2002; Behar *et al.*, 2008). Different hypotheses exist to explain this event. The most accepted hypothesis is that the populations of early modern humans underwent a split somewhere in eastern Africa and that the L0d and L0k lineages developed because of the small size of the population that migrated south. It is assumed

that the effects of genetic drift owing to the small population size would have been severe and would have caused the distinctly separate lineages to develop subsequent to the population split and isolated existence (Gonder *et al.*, 2007; Tishkoff *et al.*, 2007). Another hypothesis, however, suggests that the L0 lineages were localised in the southern regions of Africa after a population split from L1'5 and that there was a dispersal of the L0a'b'f lineage subset back to the eastern regions of Africa where it merged with the localised L1'5 lineages. The L0 lineages that were not part of the back migration developed into the L0d and L0k lineages (Behar *et al.*, 2008). Based on the latter theory, it was proposed that the L0a'b'f and L1'5 dispersed later at about 70,000 ybp to the western and central regions of Africa where the L1 and later the L2 and L3 lineages developed (Behar *et al.*, 2008). The results of this investigation support this hypothesis. The coalescence-time depths of the L0a'b'f'k, L0a'b'f, L0a'b and L0d lineages of the All African MP tree of this investigation are deeper than the time depths of the L1, L2 and L3 lineages, which suggests that the ancestors of the L0a and L0d sub-groups diverged earlier in human history. This could, however, not be concluded with absolute certainty, because the coalescence-times of the roots of the L1, L2 and L3 haplogroups of the All African MP tree were not determined.

The L0a1 lineage of the All African MP tree of this investigation displays a coalescence-time depth of 25,690 ± 11,303 ybp, which is in agreement with the reported coalescence date reported by Salas *et al.* (2002) but slightly lower than that reported by Chen *et al.* (2000), as presented in Table 6.55. The All African MP tree L0a1b sub-clade, which contains three (3) Tswana-speaking individuals of this investigation, displays a similar coalescence-time of 21,579 ± 10,276 ybp for the L0a1 lineage, which suggests that this lineage developed in quick succession to the L0a1 lineage.

Table 6.55 Coalescence-time estimates published for haplogroup L0a and sub-haplogroups

Published study	MtDNA region	Type of marker	Lineage	Coalescence-time estimation
Chen *et al.*, 2000	Complete genome	HR-RFLP	L0a	91,400–120,500
Gonder *et al.*, 2007	Coding region	MtDNA sequence data	L0a	48,900-60,300
Batini *et al.*, 2010	Coding region	MtDNA sequence data	L0a	34,900-61,100
Salas *et al.*, 2002	HVI	RFLP	L0a1	16,750-49,950
Chen *et al.*, 2000	Complete genome	HR-RFLP	L0a1	57,200–75,500
Salas *et al.*, 2002	HVI	RFLP	L0a2	4,650-11,950

HR-RFLP = high resolution restriction fragment length polymorphism; HVI = hyper variable region 1; coalescence-time estimates are presented as years before present (ybp) and include the standard deviations by presenting these as a time range.

The L0a2 lineage of the All African MP tree of this investigation displaysd a coalescence-time of 34,938 ± 13,358 ybp, which indicates an earlier divergence than the L0a1 lineage. In contrast, Salas *et al.* (2002) reported that this lineage was more recent than the L0a1 lineage, as presented in Table 6.55. The deep time estimate for the L0a2 lineage of the All African MP tree of this investigation may be attributed to the large component of ancient lineages that belonged to the Pygmy and Tanzanian individuals that were present in this clade of the All African MP tree and that will be reflected in the time estimates. To support this viewpoint further, the coalescence-time estimate of this investigation is in agreement with the reported mean coalescence-time estimate of the L0a2 lineages within the Western Pygmy population, which indicates coalescence at 20,000 to 34,000 ybp (Batini *et al.*, 2011). The L0a2a lineage of the All African MP tree, which contains four Tswana-speaking individuals of this investigation, displays a coalescence-time of 12,331 ± 7,707 ybp, which is more recent than the coalescence of the L0a1b lineage. This is in agreement with evidence that suggestsd a more recent emergence of the L0a2 lineage in central Africa (Salas *et al.*, 2002). In summary, the coalescence-times estimated for the L0a lineages of this investigation support the hypothesis that the L0a lineage originated in eastern Africa at about 40,000 ybp to 55,000 ybp, that the sub-lineage L0a1 coalesced about 30,000 ybp and that the L0a2 lineages are likely to be connected to the more recent population expansions during the dispersal of the Bantu-speaking populations to the southern regions of Africa (Salas *et al.*, 2002; Behar *et al.*, 2008; Soares *et al.*, 2009).

The L0d lineages are the largest haplogroup component of the Tswana population of this investigation (54%). The coalescence-time estimate for the L0d lineage of the All African MP tree is 83,235 ± 20,654 ybp, which supports the theory that the L0d lineages are ancient and developed long before the arrival of the Bantu speakers in the southern regions of Africa (Gonder *et al.*, 2007; Tishkoff *et al.*, 2007; Behar *et al.*, 2008). It further agrees with the archaeological evidence of the existence of modern humans in southern Africa about 70,000 ybp (Henshilwood *et al.*, 2002). The reported coalescence-time estimates for the L0d lineages, which appear to be restricted to the Khoi-San speaking populations of southern Africa and Angola and individuals from Tanzania in eastern Africa, have been dated at about 100,000 ybp (Chen *et al.*, 2000; Salas *et al.*, 2002; Gonder *et al.*, 2007; Tishkoff *et al.*, 2007; Behar *et al.*, 2008). A detailed summary of the coalescence-time estimates of the L0d lineages by other studies is presented in Table 6.56.

Table 6.56 Coalescence-time estimates published for haplogroup L0d and sub-haplogroups

Published study	MtDNA region	Type of marker	Lineage	Coalescence-time estimation
Salas *et al.*, 2002	HVI	HR-RFLP	L0d	36,150-63,050
Gonder *et al.*, 2007	Coding region	MtDNA sequence data	L0d (Tanzanian and SAK)	85,800-126,200
			L0d (Tanzanian)	12,800-48,400
			L0d SAK	71,500-109,300
Tishkoff *et al.*, 2007	HVI and HVII	MtDNA sequence data	L0d (Tanzanian and SAK)	33,700-90,100
			L0d Tanzanian	9,400-50,700
			L0d SAK	18,500-145,000
Batini *et al.*, 2011	Coding region	MtDNA sequence data	L0d	61,750–97,260

HR-RFLP = high resolution restriction fragment length polymorphism; HVI and HVII = hyper variable region 1 and 2; coalescence-time estimates are presented as years before present (ybp) and include the standard deviations by presenting these as a time range; SAK = South African Khoi-San.

The sub-haplogroups L0d1, L0d2 and L0d3 of the All African MP tree, of which 25% consist of Tswana individuals from this investigation, display similar estimated coalescence-times, suggesting that the sub-haplogroups diverged at about the same time. The L0d1b lineage of the All African MP tree of this investigation displays a deeper coalescence-time estimate of 40,590 ± 14,386 ybp than the L0d1a and L0d1c lineages, which display coalescence-times of 21,065 ± 10,276 ybp and 27,745 ± 11,817 ybp respectively. The L0d2a lineage of the All African MP tree of this investigation also displays a deeper coalescence-time, which is similar to the L0d1b lineage, providing evidence that the Tswana population under investigation most likely introgressed with the more ancient Khoi-San speakers rather than with individuals from Tanzania, who display L0d sub-clades of more recent origin, as presented in Table 6.56 and as reported by other studies (Gonder *et al.*, 2007; Tishkoff *et al.*, 2007).

The L0k lineage has been reported to be present mainly in the southern African Khoi-San-speaking populations and to a more limited extent in Tanzanian populations (Tishkoff *et al.*, 2007). It is a sister clade of L0a'b'f and has been reported to have diverged about 35,000 ybp (Behar *et al.*, 2008; Batini *et al.*, 2011). The coalescence-time estimate for the L0k lineage of the All African MP tree of this investigation is 25,690 ± 11,303 ybp and therefore in agreement with these studies. Based on these results it is hypothesised that

the L0k lineage present in one of the Tswana individuals of this investigation, has most probably been inherited from a Khoi-San-speaking maternal ancestor, who carried the L0k lineage that diverged from the L0a'b'f'k ancestor.

Only two (2) Tswana-speaking individuals of this investigation belong to haplogroup L1 and more specifically display the sub-haplogroup L1c2. The coalescence-time estimate for the L1c lineage of the All African MP tree is 105,842 ± 23,121 ybp and therefore supports the hypothesis that haplogroup L1c originated from a maternal ancestor from central Africa about 70,000 ybp and evolved into the different lineages that were commonly observed in the Bantu-speaking agriculturalist populations and Pygmy populations later on (Batini *et al.*, 2010). The coalescence-time estimate for the L1c2 lineage of the All African MP tree of this investigation is 53,435 ± 16,441 ybp and this suggests that the L1c haplogroup diversified and matured into the L1c2 derivative over a long period of time, as reported by Salas *et al.* (2002).

The Tswana-speaking individuals of this investigation comprise 33% of the L2a clade of the All African MP tree and display a coalescence-time estimate of 51,380 ± 15,927 ybp. This result supports the theory that haplogroup L2a sub-clades coalesced about 45,000 ybp to 55,000 ybp, as reported in other studies presented in Table 6.57 (Salas *et al.*, 2002; Kivisild *et al.*, 2004; Behar *et al.*, 2008).

Table 6.57 Coalescence-time estimates published for haplogroup L2a and sub-haplogroups

Published study	MtDNA region	Type of marker	Lineage	Coalescence-time estimation
Chen *et al.*, 2000	Complete genome	HR-RFLP	L2a	39,000–51,400
Salas *et al.*, 2002	HVI	RFLP	L2a	35,800-74,500
Behar *et al.*, 2008	Coding region	MtDNA sequence data	L2a	38,096-53,970
Batini *et al.*, 2011	Coding region	MtDNA sequence data	L2a	34,930–59,430
			L2a1	27,580–47,085

HR-RFLP = high resolution restriction fragment length polymorphism; HVI = hyper variable region 1; coalescence-time estimates are presented as years before present (ybp) and include the standard deviations by presenting these as a time range.

The coalescence-time estimates of the L2a1 sub-lineages of the All African MP tree of this investigation range from 2,055 ybp to 40,589 ybp with a mean coalescence-time of 16,184 ybp, which is in agreement with the theory that the L2a1 lineage was commonly observed in the Bantu-speaking populations that dispersed to the eastern and western regions of Africa from the central Sahel region about 14,000 ybp (Salas *et al.*, 2002). The L2a1a, L2a1b and L2a1f lineages of the All African MP tree further display

coalescence-time estimates consistent with lineage development during the earliest Bantu dispersals to the southern regions of Africa about 4,000 ybp.

Only two Tswana-speaking individuals of this investigation belong to the major haplogroup L3. One individual belongs to sub-haplogroup L3d1 and the other to sub-haplogroup L3e1, which display coalescence-times of 49,324 ± 15,414 ybp and 15,414 ± 8,734 ybp respectively. The root to L3d1 in the All African MP tree coalesced at 47,783 ± 15,414 ybp and this supports the hypothesis that the sub-haplogroup L3d coalesced between 30,000 ybp and 40,000 ybp and was prevalent in the western region of Africa before the dispersal to the southern regions of Africa, after which it was commonly observed in the maternal genetic pool of the populations that resided in southeastern Africa (Salas *et al.*, 2002; Behar *et al.*, 2008). It is hypothesised that the root to haplogroup L3e1 developed in central Africa about 40,000 to 50,000 ybp, after which the sub-haplogroup L3e1 developed about 16,000 ybp (Salas *et al.*, 2002; Behar *et al.*, 2008). The coalescence-time estimate for haplogroups L3e and L3e1 of the All African MP tree of this investigation are 50,352 ± 15,927 ybp and 15,414 ± 8,734 ybp respectively and the result therefore supports this theory. It has been postulated that the sub-haplogroup L3e1 arose at that time and became frequent among the Bantu-speaking populations that migrated to the southern regions of Africa (Salas *et al.*, 2002).

6.10 TSWANA mtDNA CONSENSUS SEQUENCE

The rCRS is used as a standard measure in mitochondrial studies and represents the reconstruction of a single European individual's mtDNA sequence (Andrews *et al.*, 1999). It does not represent the genetic diversity of African individuals that belong to major haplogroup L and does not provide a true reflection of nucleotide-by-nucleotide variation that could be expected in African or more specifically Tswana-speaking individuals of South Africa. The construction of a consensus sequence for the Tswana-speaking individuals of this investigation indicates the major sequence variants observed in the Tswana-speaking individuals and therefore also represen the major nucleotide differences between the mtDNA sequences of the Tswana-speaking individuals and the rCRS. The importance of the construction of a consensus sequence lies in the fact that it represents the most frequently observed sequence variants rather than the phylogenetic tree topology, which represents all of the rare sequence variants (Carter, 2007).

A consensus approach determines an ancestral mitochondrial model sequence that encompasses all homoplasies and rare variants. It is dependent on the sampling of existing phylogenetic branches and will under optimal sampling conditions point to a similar ancestral sequence as determined by a phylogenetic approach (Carter, 2007). Therefore, the consensus sequence displays sequence variation that does not necessarily determine the phylogenetic tree topology and more fully represents the degree of mtDNA sequence variation than observed in the Tswana-speaking individuals of this investigation. Even though the relatively small size of the sample of the Tswana-speaking cohort of this study is of some concern with regard to the representation of the full spectrum of sequence variants in the current Tswana population of South Africa, it still represents the first novel consensus of mitochondrial sequence variation that has been constructed for a Bantu-speaking population of South Africa and is presented in Appendix G.

The Tswana mtDNA consensus contains 16,569 nucleotides, of which 69 nucleotide sites differ from the rCRS. Therefore 0.4% of the genome contains variable sites and 99.6% identical to the rCRS. Ten of the sequence variants in the Tswana consensus sequence ambivalent and are indicated by using the IUPAC codes. The ambivalence refers to two possible sequence variants at a single nucleotide position that are distributed in about equal quantities across the Tswana samples of this investigation and therefore both represent major sequence variants. To ensure that both sequence variants are acknowledged as contributing to the total sequence variance of the Tswana consensus sequence, they have been indicated as ambivalent by the BioEdit version 7.0.5.2 (Hall, 2001) software. The Tswana dataset of 50 mtDNA sequences furthermore contains 459 variant positions in total as opposed to the 69 variant positions observed in the Tswana consensus sequence. Therefore only 15% of the sequence variants that have been observed in the Tswana-speaking individuals of this investigation are present in the majority of the individuals and are therefore represented in the Tswana consensus sequence.

The African (Yoruba) reference mtDNA sequences listed in MITOMAP differed from the rCRS at only 40 nucleotide positions, reflecting the lower level of sequence variance when compared with the Tswana consensus sequence. The African (Yoruba) reference sequence was, however, not a consensus sequence and represented the sequence variance observed in a single Yoruban individual, which highlights the fact that, as is also the case with the rCRS, the purpose of standard reference sequences are not to reflect the degree of variation displayed by a population but only to provide a guideline according to

which site variation could be determined, based on a standard set of sequence variants and according to standard nucleotide positions (Carter, 2007). Reference sequences may contain rare sequence variants, as was observed in the rCRS (Andrews *et al.*, 1999), which would not be reflected in the consensus sequence. Comparisons between reference mtDNA sequences and consensus sequences are therefore a comparison of sequence variances based on totally different underlying origins in terms of the sequence variation displayed. The Tswana consensus sequence constructed in this investigation represents the mtDNA sequence variation of a South African Tswana-speaking population of a specific region and should be used as such. The contribution of the Tswana consensus sequence to an understanding of the evolutionary history of the Bantu-speaking African populations lies in the benchmark it provides for sequence variation, which can be used for the measurement of mitochondrial variation in the context of African populations, as compared to different global populations.