CHAPTER FIVE

Materials and Methods

Quality control measures were implemented to ensure the validity of the experimental results obtained from the methods followed as discussed in this chapter. This included adhering to good laboratory practice procedures at all times when working in the laboratory. As this investigation entailed the amplification of DNA, measures were taken to prevent and 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. Such measures include keeping work areas and laboratory instrumentation free from contaminated DNA, using sterile and analar quality reagents, storing reagents correctly and minimising the risk of contaminating reagents by using working stocks made from the original reagent containers. Master mixtures were prepared for batches of reactions to avoid unnecessary pipetting of very small quantities. Laboratory consumables and pipettes used in this investigation were dedicated and not shared with other students.

5.1 ETHICAL APPROVAL OF THE STUDY

This investigation forms part of a research programme that has been approved by the Ethics Committee of the North-West University (Potchefstroom Campus) under the title "The role of the mitochondrial genome in human migration and health" with approval number NWU-00038-07-S8. Samples were only obtained after informed consent had been given by all participants or legal guardians of participants in the study.

5.2 SAMPLE DESIGN AND METHODS

To analyse the genetic diversity and phylogenetic relationships of an African Tswana population, 50 blood samples from individuals of Tswana origin were collected for analysis. These samples were collected under the Profiles of Resistance to Insulin in Multiple Ethnicities and Regions (PRIMER) study conducted by the North-West University (Potchefstroom Campus) of South Africa. The study was conducted in the Ikageng and Sonderwater urban areas located near the town of Potchefstroom, which is situated in the

southern part of the North-West province of South Africa, and in the rural areas of Ganyesa and Tklagameng in the western region of the North-West Province. The rural areas, where the study was conducted, are situated close to the border between South Africa and Botswana.

Blood samples were taken from “apparently healthy” individuals according to the following exclusion criteria: participants were non-pregnant, non-lactating, at least 35 years old and did not have any chronic diseases or acute infections (oral temperature > 37°C), nor did they use chronic medicine.

The individuals all completed socio-demographic history questionnaires in which their age, height, weight, body mass index, diabetes risk score and cardiovascular health were recorded, as well as the ethnicity of their mother and father. The individuals chosen for this investigation were all of self-reported Tswana descent according to the ethnicity of their parents, as recorded in the questionnaire mentioned. Blood samples for DNA isolation were collected in ethylene diamine tetra acetic acid (EDTA) tubes, which were aliquoted and stored in the field at -20°C for a week, after which they were stored at -80°C in a laboratory environment. The sample identification numbers of the individuals chosen for this investigation are given in Table 5.1.

Table 5.1 Sample identification numbers

#	Patient number	#	Patient number	#	Patient number	#	Patient number	#	Patient number
1	2074	11	3002	21	3471	31	4051	41	5044
2	2075	12	3027	22	3486	32	4056	42	5060
3	2077	13	3066	23	3015	33	3461	43	5062
4	2082	14	3075	24	3505	34	4075	44	5063
5	2091	15	3085	25	4111	35	4080	45	5066
6	2093	16	3107	26	4013	36	4083	46	4032
7	2095	17	3117	27	4027	37	4089	47	5083
8	2097	18	3236	28	4034	38	3506	48	5085
9	4063	19	3459	29	4037	39	4106	49	5086
10	2103	20	3466	30	3495	40	4117	50	5091

The numbers in red represent the numbers allocated to the samples for this investigation. The numbers in black represent the numbers allocated to the samples collected as part of the PRIMER field study as discussed in this section.

5.3 DNA ISOLATION

Isolation of DNA from the blood samples was undertaken by an alumnus of the Centre for Genome Research (CGR) at the North-West University. The protocol for isolation of DNA from whole blood as given in the FlexiGene®[1] DNA Handbook was followed.

DNA isolations were performed using the Qiagen[2] FlexiGene® Kit by lysing the blood cells and recovering the cell nuclei and mitochondria via centrifuging. Contaminants and proteins were in turn removed by using a denaturing buffer. The DNA was recovered by isopropanol precipitation, washed with 70% ethanol and resuspended in a hydration buffer (TrisCl buffer) ready for downstream applications. The FlexiGene® kit produces DNA yields of up to 150 kb in size, which was suitable for the later PCR and Automated Sequencing applications and protocols to be followed for the sequencing of the full mitochondrial genome.

The concentration of the DNA in the samples was measured by absorbance at 260 nanometre (nm) on an Eppendorf®[3] BioPhotometer 6131. The purity of the DNA in the samples was determined by calculating the ratio of the absorbance at 260 nm and the absorbance at 280 nm. A ratio of more than 1.7 was regarded as an indication of DNA sufficiently purified from proteins and other contaminants.

5.4 POLYMERASE CHAIN REACTION

The PCR methodology of Mullis *et al*. (1986) was followed in this investigation. PCR was used to amplify the full mitochondrial genomes of all 50 Tswana individuals in eight separate segments in order to ensure a sufficient quantity of the DNA for direct sequencing. The segments are referred to as DNA regions one to eight.

5.4.1 PCR primers

The full mitochondrial genome of all 50 samples of this investigation was amplified in eight segments along the length of the mitochondrial DNA.

1 FlexiGene® is a registered trademark of Qiagen GmbH, Hilden, Germany.
2 Qiagen® is a registered trademark of Qiagen GmbH, Hilden, Germany.
3 Eppendorf® is a trademark of Eppendorf AG, Hamburg, Germany.

0 lists the primer names, sequences and melting temperatures of the eight pairs of primers as obtained from Maca-Meyer *et al*. (2001).

Table 5.2 Eight primer pairs that were used to amplify the full mitochondrial genomes of the Tswana speaking cohort of this investigation

Primer region	Primer name	Primer sequence	T_m	Mean T_m	Product size (bp)
1	F32:mtL15996	5'-ctc cac cat tag cac cca aag c-3'	57	57	2,103
	R3:mtH1487	5'-gta tac ttg agg agg gtg acg g-3'	57		
2	F3:mtL923	5'-gtc aca cga tta acc caa gtc a-3'	53	52.5	2,789
	R7:mtH3670	5'-ggc gta gtt tga gtt tga tgc-3'	52		
3	F8:mtL3644	5'-gcc acc tct agc cta gcc gt-3'	58	56	2,227
	R11:mtH5832	5'-gac agg ggt tag gcc tct tt-3'	54		
4	F11:mtL5278	5'-tgg gcc att atc gaa gaa tt-3'	48	51	2,679
	R15:mtH7918	5'-aga tta gtc cgc cgt agt cg-3'	54		
5	F16:mtL7882	5'-tcc ctc cct tac cat caa atc a-3'	53	52.5	2,089
	R19:mtH9928	5'-aac cac atc tac aaa atg cca gt-3'	52		
6	F20:mtL9886	5'-tcc gcc aac taa tat ttc act t-3'	49	51.5	2,231
	R23:mtH12076	5'-gga gaa tgg ggg ata ggt gt-3'	54		
7	F23:mtL11486	5'-aaa act agg cgg cta tgg ta-3'	50	49.5	2,740
	R27:mtH14186	5'-tgg ttg aac att gtt tgt tgg-3'	49		
8	F27:mtL13612	5'-aag cgc cta tag cac tcg aa-3'	52	52	2,828
	R32:mtH16401	5'-tga ttt cac gga gga tgg tg-3'	52		

Adapted from M Koekemoer, PhD thesis, 2010 and Maca-Meyer *et al*., 2001. Primer regions 1 - 8 refer to amplified segments 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 rCRS where amplification is started while product size indicates the length between the starting points of the two primers in a pair. bp = base pairs. T_m = melting temperature of the primer and the mean T_m refers to the mean of the T_m values for both primers.

The primer pairs were optimised for PCR by adjusting the annealing temperature (T_a) and evaluating the quality of the PCR product by electrophoresing the DNA product on an agarose gel and inspecting the end product for sufficient quantity and presence of secondary product. This process was repeated until the optimal T_a had been determined for each primer pair. 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). The melting temperatures used in this investigation were calculated by an alumnus of the CGR of the North-West University (Koekemoer, 2010) and the T_m of each of the primer pairs were averaged and used as a starting point for the optimisation of the primer pairs.

5.4.2 PCR reaction

A standard PCR protocol was used for the PCR reactions similar to that used by the students at the CGR. A reaction mixture was prepared that served as a master mix for 12 PCR reactions and consisted of double distilled water (ddH_2O), PCR buffer, magnesium chloride ($MgCl_2$), dNTPs (deoxynucleotide triphosphate), F primer (forward primer) and R primer (reverse primer) and *Taq* enzyme. The dNTPs consisted of a combination of 2'-deoxyadenosine-5'-triphosphate (dATP), 2'-deoxycytidine-5'-triphosphate (dCTP), 2'-deoxyguanosine-5'-triphosphate (dGTP) and 2'-deoxythymidine-5'-triphosphate (dTTP) at a concentration of 10 millimolar each; this was used as a working solution and was stored at -20°C until use. 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 enzyme was supplied in a formulation that contained 50% glycerol and buffers designed to improve amplification in a concentration of 5 units/μL. The forward and reverse primers used are described in Section 5.4.1. From this master mix of reagents, 10.5 μL was aliquoted into a microcentrifuge tube for the respective single PCR reactions. To this, 50 nanogram (ng) of the sample genomic DNA was added and the reaction overlaid with a drop of mineral oil to prevent evaporation. A negative control consisting of the master mix and ddH_2O was added as a control measure to detect DNA contamination.

5.4.3 PCR conditions

The Thermo Hybaid®[1] Multiblock System 0.5 Satellite (MBS 0.5S) thermocycler was used for the amplification of DNA product. Polymerase chain reaction thermal conditions consisted of an initial denaturation cycle at 94°C for 10 minutes, followed by 30 cycles consisting of a denaturation phase at 94°C for 30 seconds, an annealing phase at the temperature determined during PCR optimisation for 30 seconds and an elongation phase at 72°C for 30 seconds. On completion of the 30 cycles, a final elongation cycle at 72°C for 7 minutes was performed. Samples were held at 4°C indefinitely.

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

5.5 AGAROSE GEL ELECTROPHORESIS

The PCR product was electrophoresed on an agarose gel to visualise the quality of the DNA amplification product. Gel mobility of DNA fragments is affected by the fragments' secondary structure and fragment length. Molecular grade Bioline®[2] agarose was used for the preparation of 0.9% agarose gels. The Bioline® agarose is DNase and RNase free and has gel strength that is high enough for feasibility of gels as low as 0.5%. This agarose product allows high resolution of DNA fragments between 500 and 1000 bp in length, depending on the gel concentration. The agarose was made up into a solution by adding ddH_2O and 10 X Tris-borate-EDTA (TBE) buffer to achieve the 0.9% weight-to-volume concentration. The solution was heated in a microwave until all of the agarose was sufficiently melted. Two (2) μL of a 0.5 $\mu g.mL^{-1}$ ethidium bromide working solution was added to the gel to intercalate with the double-stranded DNA of the loaded DNA product for purposes of visualisation by ultraviolet (UV light) (254 nm). After the solution had cooled to about 60°C, the gel was poured into a casting tray containing a sample comb. Mini casting trays were used, made of UV-transparent plastic. The gel was placed into a buffer chamber (SCIGEN) and covered with TBE buffer. The DNA product was mixed with loading buffer. The loading dye solution consisted of Orange G gel loading dye as tracking agent in the agarose and a high molecular weight Ficoll component to weigh down the sample in the sample wells. A FastRuler™[3] High Range DNA Ladder (Fermentas) of range 500-10,000 bp was included in the first and last lanes of the gel. The DNA product and DNA Ladder were electrophoresed at 100 volts (V) for 30 minutes. After electrophoresis, the DNA fragments were visualised by placing the gel on an UVIvue UV transilluminator and recorded directly by photography.

5.6 DNA PURIFICATION

Contamination of the PCR product with proteins, RNA, chromosomal DNA, excess PCR primers, dNTPs, enzyme, buffer components, residual salts, organic chemicals and residual detergents leads to less optimal or no sequencing results. The PCR product was therefore purified by using the Zymo Research DNA Clean & Concentrator™[4]-5 kit that consists of a DNA binding buffer, a DNA wash buffer, Zymo-Spin™[5] columns and

1 Thermo Hybaid® is a registered trademark of Hybaid Ltd., Ashford, Middlesex, UK.
2 Registered trademark of Bioline USA Inc. NJ, USA.
3 FastRuler™ is a registered trademark of Fermentas International, Inc., Ontario, Canada.
4 DNA Clean & Concentrator-5™ is a trademark of Zymo Research Corporation, Orange County, CA, USA.
5 Zymo-Spin™ is a trademark of Zymo Research Corporation, Orange County, CA, USA.

collection tubes. Ethanol was added to the DNA wash buffer prior to use according to the kit specifications. After the DNA binding buffer had been added to the amplified sample, the PCR product was transferred to the Zymo-Spin™ Column and centrifuged at 10,000 rotations per minute (rpm) for 30 seconds in an Eppendorf®[1] 5810 centrifuge using a fixed angle rotor F-45-30-11 in order to wash the buffer through from the column. The DNA that is absorbed to the Zymo-Spin™ column was washed twice with the wash buffer by centrifuging at 10,000 rpm for 30 seconds to wash the wash buffer through the column, after which the PCR product was eluted with 6-10 μL of ddH_2O by centrifuging at 10,000 rpm for 30 seconds. The eluted, purified DNA product was transferred to a 0.2 ml microcentrifuge tube and stored at 4°C indefinitely.

5.7 DNA QUANTIFICATION

The concentration of the DNA product was determined by optical density measures by using the Eppendorf® BioPhotometer 6131 instrument to determine the concentration of the DNA by spectroscopy. The purified DNA samples were diluted with ddH_2O by adding 45 μL ddH_2O to 5 μL of purified DNA sample. In cases where the concentration was high, 95 μL ddH_2O was added to a 5 μL aliquot of purified DNA sample. The sample measurements were made against a blank sample containing no DNA and prepared with 50 μL ddH_2O. The sample dilution factor in the measuring cuvette was entered when performing the measurement and was automatically included in the result calculation. Absorbance measurements were made at 260 nm, 280 nm and 320 nm and the A_{260}/A_{280} ratio was used as an indication of the DNA purity of the sample. Turbidity in solutions was indicated in a rise of all absorbency values. The absorbance measure at 320 nm represented the optical clarity of the sample and values were expected to be approximately zero for samples with low turbidity. UV absorbance at 280 nm was used as an indication of the protein concentration in the sample and values were expected to be lower than that of the A_{260}. Values of the A_{260}/A_{280} ratio less than between 1.7 and 1.9 were regarded as indicative of contamination by protein or organic chemicals.

The quantity of dsDNA dissolved in 1.0 ml in a cuvette with a 1 cm path length that gives an absorbance of 1.00 in a spectrometer equals one optical distance (OD) unit. The extinction coefficient used to convert the UV absorbance at 260 nm to double-stranded DNA (dsDNA) concentration, is 1 $A_{260} \equiv 50\ \mu g.mL^{-1}$. The instrument was set to calculate

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

the concentration of the DNA in the purified sample by using this extinction coefficient and taking the dilution factor of 10X into consideration and was programmed to provide a concentration result in ng/μL^{-1}.

5.8 AUTOMATED DNA SEQUENCING

Non-recombining markers such as mtDNA and the Y chromosome markers are popular genetic markers used in evolutionary studies because of the unparalleled resolution obtained through the use of these markers in comparison to any other genetic locus in the human genome as a result of its unique features, such as high copy number, maternal inheritance, lack of recombination and high mutation rate (Richards and Macaulay, 2001; Pakendorf and Stoneking, 2005). These markers were initially studied by using RFLPs that entailed the cleavage of the mitochondrial genome at five to six restriction enzyme sites or by using high-resolution restriction enzyme analysis, which entailed cutting the mitochondrial genome at 12 to 14 sites (Pakendorf and Stoneking, 2005). In contrast to RFLPs, sequencing strategies initially focused on the control regions only (HVS-I and HVS-II) and it is only recently that sequencing of the full mitochondrial genome has become feasible (Richards and Macaulay, 2001). Most studies have focused on studying the mitochondrial variation of human populations by using RFLPs of the whole mitochondrial genome (Cann *et al.*, 1984) or only in the hyper-variable regions I and II or HVS-I and HVS-II (Vigilant *et al.*, 1989; Chen *et al.*, 1995a; Watson *et al.*, 1996). More recently, the focus has shifted to studying variation by the sequencing of the mtDNA of especially the HVS-I and HVS-II (Soodyal *et al.*, 1996; Watson *et al.*, 1997; Alves-Silva *et al.*, 2000; Bandelt *et al.*, 2001a; Pereira *et al.*, 2001). Full genome sequencing, however, provides the highest quantity of information that can be expected on the mitochondrial tree and more information about genealogy is determined by sequencing more of the molecule. For this reason, it was decided to use the full mitochondrial genome sequencing approach in this investigation of a sample population of 50 individuals of a southern African Tswana-speaking population. Genetic diversity studies of populations are favoured by including more sequence data and giving more information about the distribution of nucleotide diversity and selection, drift and migratory forces. In population genetics, this means more accurate population parameters and the ability to select the most appropriate evolutionary hypothesis for the population under study. It also provides more accurate information about nucleotides sites, insertion and deletion events, selection, protein structure, functional conserved areas and rate and pattern of substitution (Pollock *et al.*, 2000).

5.8.1 Sequencing strategy and primers

Each of the the eight regions of the full mitochondrial genome was sequenced with four primers, as listed in Table 5.3. The eight regions amplified by PCR were all approximately 2,000 bp in length and the four sequencing primers were selected in such a way that each of these regions of 2,000 bp was sequenced in segments of approximately 500 bp in length. This was a viable length to obtain good sequence resolution and sufficient overlap between fragments and therefore a feasible strategy. Reverse primers that annealed to the heavy strand of the mitochondrial DNA were used in cases where poor resolution gave rise to ambiguous bases owing to problematic sequence compositions and were often the cause of failure to obtain overlap between the adjacent sequence segments. The primers were based on those reported by Maca-Meyer *et al.* (2001).

Table 5.3 Primers used to sequence the full mitochondrial genome

PCR region	Primer name			
1	F32:mtL15996	F1:mtL16340	F2:mtL382	F3:mtL923
2	F4:mtL1372	F5:mtL2025	F6:mtL2559	F7:mtL3073
3	F8:mtL3644	F9:mtL4210	F10:mtL4750	F11:mtL5278
4	F12:mtL5699	F13:mtL6337	F14:mtL6869	F15:mtL7379
5	F16:mtL7882	F17:mtL8299	F18:mtL8799	F19:mtL9362
6	F20:mtL9886	F21:mtL10403	F22:mtL10949	F23:mtL11486
7	F24:mtL11964	F25:mtL12572	F26:mtL13088	F27:mtL13612
8	F28:mtL14055	F29:mtL14650	F30:mtL15162	F31:mtL15676

F = forward primer, mtL = mitochondrial light strand. Number refers to the CRS position number. Numbers 1 to 8 refer to the eight PCR regions as described in Section 5.4.1.

5.8.2 Cycle-sequencing reaction protocol

A cycle-sequencing method was employed to sequence the full mitochondrial genome in this investigation. Cycle sequencing has been proven to yield reproducible results for sequencing PCR templates (Applied Biosystems, 2010). Purified PCR product was used for sequencing reactions. PCR product was verified as being pure by using agarose gels and spectroscopy in order to ensure the use of good quality DNA template in the sequencing reactions, as described in Section 5.6.

The BigDye®[1] Terminator v3.1 Cycle Sequencing Kit was used for sequencing. The kit provided the reagents required for the sequencing reactions in a ready-to-use format as described in Table 5.4.

Table 5.4 Description of reagents in BigDye® Terminator v3.1 Cycle Sequencing Kit

Reagent	Description
Ready Reaction Mix	Sequencing primer, DNA polymerase, dNTPs
pGEM®-3Zf(+)	Double-stranded DNA Control template of 1,000 bases
–21 M13	Control primer (forward) for performing the control reactions
BigDye Terminator v1.1/3.1 Sequencing Buffer (5X)	Buffer

Excess freeze-thaw cycles of the kit reagents were avoided by aliquoting the reaction mix in smaller quantities and not exceeding more than five freeze-thaw cycles per aliquot. The reagents were not thawed by heating but allowed to thaw on ice at room temperature, from which point onward the reagents were kept on ice for the duration of the set-up. Thin-walled 0.2ml micro-tubes were used for the sequencing reaction set-up.

A standard protocol was used to perform the sequencing reactions. Sequencing reactions were set up to a final total volume of 10 μL and consisted of the sequencing buffer, the reaction mix, sequencing primer and DNA template. The sequencing buffer was supplied at a 5X concentration and used at a 1X concentration according to supplier instructions. Sequencing primers were used at a concentration of 3.2 pmol from a stock solution of the primers as described in Section 5.8.1. Primer working solutions of 3.2 mM in 100 μL were prepared for daily use as a preventative measure against contamination.

Input PCR product concentration was optimised for this investigation according to the recommended quantities for optimal sequencing results of between 20 and 50 ng of PCR product of more than 2,000 bp in length (Applied Biosystems, 2010) as presented in Table 5.5.

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

Table 5.5 Amount of PCR product used for sequencing reactions in this investigation

PCR region	Primer pairs	Size of PCR product (bp)	Amount of PCR product used for sequencing (ng)
1	F32:mtL15996	2,103	50
	R3:mtH1487		
2	F3:mtL923	2,789	50
	R7:mtH3670		
3	F8:mtL3644	2,227	50
	R11:mtH5832		
4	F11:mtL5278	2,679	40
	R15:mtH7918		
5	F16:mtL7882	2,089	50
	R19:mtH9928		
6	F20:mtL9886	2,231	40
	R23:mtH12076		
7	F23:mtL11486	2,740	40
	R27:mtH14186		
8	F27:mtL13612	2,828	50
	R32:mtH16401		

F = forward primer, mtL = mitochondrial light strand, bp = base pairs, ng = nanogram. Number refers to the CRS position number. Numbers 1 to 8 refer to the eight PCR regions as described in Section 5.4.1.

The volume of sample used was determined by dividing the amount of DNA that was needed according to Table 5.5 by the DNA quantity of the individual samples, as was determined by spectroscopy (See Section 5.7). This quantity was made up to an end volume of 5 μL with ddH_2O and loaded first as preventative step to avoid contaminating other reagents with amplified DNA products.

5.8.3 Cycle-sequencing reaction conditions

The Thermo Hybaid®[1] MBS 0.5S thermocycler was used for the sequence-cycling of the DNA product. Cycle sequencing reaction thermal conditions consisted of a denaturation phase at 96°C for 10 seconds, an annealing phase at 50°C for five seconds and an elongation phase at 60°C for four minutes. On completion of the 25 cycles, the reactions were held at 4°C indefinitely.

5.8.4 Post-sequencing treatment

After completion of the sequence cycling, the sequencing reactions were treated with sodium dodecyl sulphate (SDS) by adding 1 μL of a 2.2% SDS solution to the reaction.

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

The purpose of this treatment was to remove excess dye terminators, thereby preventing dye blob artefacts from interfering with the end sequencing results. The reaction tubes were replaced in the Thermo Hybaid® MBS 0.5S thermocycler and heated to 98°C for five minutes, after which it was cooled down to 25°C for 10 minutes and held at 4°C indefinitely.

5.8.5 Purification of sequence extension product

The purification of the sequence extension products was not performed on site but by another institution on contract. The sequencing extension products were purified by using the ethanol/EDTA/sodium acetate precipitation method to rid the reaction of unincorporated dyes and dye-labelled terminators to ensure a clear signal during electrophoresis. Precipitation of the extension products was performed according to the protocol of the sequencing facility that was contracted to perform the sequencing of the samples of this investigation.

5.8.6 Capillary electrophoresis

As was the case with the purification of the sequence extension products, the electrophoresis of the sequence extension products was not performed on site, but was instead contracted out to the University of Stellenbosch. On completion of the analyses by the University of Stellenbosch, the raw data files of the sequenced samples were returned to the CGR of the North-West University.

The sequencing extension products were electrophoresed on an ABI Prism®[1] 3100 Genetic Analyser and the ABI Prism® 3730 Genetic Analyser by capillary electrophoresis that separates the fluorescently labelled DNA fragments of different lengths. The extension product was loaded into a capillary by electrokinetic injection and moved through the capillary based on the net negative charge of the DNA fragment. The fluorescence of the labelled DNA fragment was detected by an optical detection device of the genetic analyser and the signal converted to digital data by the Data Collection Software. The technical specifications of the Applied Biosystems 3730 DNA Analyser and the Applied Biosystems 3100 Genetic Analyser that were used for analysing the samples of this investigation, are given in Table 5.6.

1 ABI Prism is a registered trademark of Applied Biosystems, Foster City, CA 94404, USA.

Table 5.6 Technical specifications of genetic analysers used

Type of Analyser	No of Capillaries	Capillary Array Length(cm)	Polymer type	Sample capacity
Applied Biosystems 3100 Genetic Analyser	16	36, 50, 80	POP-4™ POP-6™	96- and 384-well plates
Applied Biosystems 3730 DNA Analyser	48	36, 50	POP-7™	96- and 384-well plates

The capillary array length is the well-to-read length. Sample capacity refers to the number of samples accommodated by the autosampler. POP™[1] = Performance Optimised Polymer.

A multicomponent analysis setting was applied to the raw data generated to separate the four different fluorescent dye signals into distinct spectral components. Base calling was performed by a KB™ basecaller that provided *inter alia* per-base quality values and identification of failed samples. Electrophoretic mobility shifts were corrected for by a mobility file and a quality value that was generated for each sample and predicted the probability of a basecall error determined for each base. After the multicomponent analysis of the raw data files, the data were represented in an electropherogram format ready for further data analysis procedures, using a variety of software. For this investigation, BioEdit software was chosen as the software application to be used for manual review of the data quality.

5.9 DATA ANALYSES

Manual data review was performed by using BioEdit version 7.0.5.2 (Hall, 2001). BioEdit software is a biological sequence editor with the capability of performing basic nucleic acid sequence editing functions, alignment, manipulation and analysis and is freely available.

5.9.1 Data review

Data review was performed on the raw data .ab1 files. Each sequence was verified manually for sequencing and editing errors. The electropherograms of the four sequencing primers of each of the respective PCR regions 1 to 8 were manually evaluated for false base calling and artefacts. Ambiguously called bases were manually edited after careful evaluation of the peak morphology. Unexpected base changes within haplogoups were investigated manually by evaluation of the raw data and validated. Insertions and deletions were verified and re-sequenced to confirm their validity. All manual editing changes were recorded to ensure traceability. Sequences with artefacts such as dye blobs, poor peak

1 POP™ is a trademark of Applied Biosystems, Foster City, CA, USA

resolution or noisy data were submitted for re-sequencing. Electropherograms were only accepted if peak morphology was clear, peaks were well defined, little background noise was present and no artefacts were present. Sequences that failed were investigated for reaction failure and/or electrophoresis failure and subsequent corrective actions were employed.

5.9.2 Resequenced Cambridge Reference Sequence

The CRS was sequenced in 1981 and re-sequenced in 1999 to identify errors and rare polymorphisms (Andrews *et al.*, 1999). This has resulted in the use of the rCRS, GenBank reference number J01415.2, as the standard mitochondrial genome reference sequence to be used in evolutionary studies and phylogenetic studies (Ruiz-Pesini *et al.*, 2007). The rCRS was also used in this investigation as a reference sequence.

5.9.3 Sequence alignment

The sequenced fragments of the full mitochondrial genomes of the Tswana-speaking individuals of this investigation were aligned by using BioEdit version 7.0.5.2 (Hall, 2001) with the rCRS as the reference. The sequence alterations that were observed in the mtDNA samples of this investigation, when compared to the rCRS, were logged for traceability. Contiguous sequences were constructed by overlapping all of the sequence fragments of each sample in BioEdit to constitute the full mitochondrial genome of each of the individuals that was sequenced.

5.10 MITOCHONDRIAL GENOME REGIONS USED IN SEQUENCE DATASETS

Most evolutionary studies of the mitochondrial DNA focus on the small control region between base pair 577 and 16,023 (Gonder *et al.*, 2007) because this region has a high mutation rate and is assumed to be selectively neutral. Within this region, the hyper variable region between base pair position 16,124 – 16,383 is the most variable and therefore used to study evolution. It comprises only about 3% of the total genome.

The coding region of the mitochondria displays a less variable mutation rate as opposed to the highly variable control region of the mitochondria. The inclusion of coding region data in phylogenetic studies is advantageous because it conforms to a constant rate molecular

clock hypothesis. In addition, the slower mutation rate and extended length results in a lower occurrence of homoplasies (Non *et al.*, 2007).

RFLPs have been used extensively in the past for phylogenetic analyses and are adequate for phylogenetic tree construction, but not adequate for the estimation of mutation rate, which is used to determine TMRCA (Non, *et al.*, 2007). Accessible and cost--effective sequencing technologies allowed sequencing of the full genome of the mitochondria for the purposes of phylogenetic analysis in order to obtain more phylogenetically informative data (Non *et al.*, 2007). Therefore the full genome of the Tswana dataset was sequenced in this study. In instances where GenBank® sequences were harvested for use in phylogenetic and statistical analyses, it often happened that only the coding regions were available. Since the control region has such high variability, it was decided to use only these sequences to compile a dataset consisting of only coding regions for phylogenetic analyses. Where a full genome sequence was available for the whole dataset, it was used.

5.11 MITOCHONDRIAL GENOME SEQUENCE DATASETS

Datasets of mitochondrial sequences were compiled to be used for phylogenetic and statistical analyses in this study with the purpose to provide a representative backdrop of African maternal lineages against which the mitochondrial sequence variation, as determined by the full mitochondrial genome sequencing of a cohort of 50 Tswana-speaking individuals in this study could be positioned. To achieve the aims of this study, which included the positioning of the mitochondrial sequence variation of the Tswana cohort of this investigation phylogenetically and interpreting the sequence variation statistically with regard to population history in the context of other individuals of African origin, four datasets were constructed. These datasets consisted of individuals that were of African origin and belonged to the macrohaplogroup L. A distinction was made between African individuals that resided on the African continent and those that resided in non-African countries. A further distinction was made between the regions within Africa in which these individuals resided i.e. western, eastern and southern Africa. The number of mitochondrial sequences that were obtained for individuals of Africa that originated or resided in northern Africa were not sufficient to define a northern region of Africa.

A dataset containing GenBank® mitochondrial genome sequences that were published between January 2004 and June 2007 was obtained from a previous phylogenetic study

on individuals of African origin (Koekemoer, 2010). In this study, the GenBank® (Benson *et al.*, 2007) sequences were harvested through the GenBank® Entrez Nucleotide retrieval system by searching the database with relevant keywords and phrases that indicated mitochondrial genome sequences of African origin. Searches of the database against the first authors of publications were also performed. In total 386 mitochondrial genome sequences were retrieved, of which some contained complete mitochondrial genome sequences and some coding region sequences only. The authors of the publications in which these sequences were published are listed in Appendix B.

This pre-existing dataset was updated by searching GenBank® for further mitochondrial genome sequences that were published after June 2007 until June 2010. Keywords, phrases and first authors were used to search for additional mitochondrial genome sequences. No additional mtDNA sequences that were published after 2007 could be retrieved from GenBank® t, but 54 mitochondrial genome sequences of African origin that were studied and published by Herrnstadt *et al.* (2002) were added to the dataset. These mitochondrial genome sequences are from individuals that are of African descent and currently reside in the USA.

Further mitochondrial genome sequences of 42 Khoi-San individuals originating from Angola, Namibia and South Africa and 39 individuals from Uganda that formed part of two previous PhD theses (Koekemoer, 2010; Isabirye, 2010) about the phylogenetic relationships between and within African populations were also included in the construction of the datasets of this investigation.

5.11.1 Global African dataset

This dataset is referred to as the Global African dataset (1a) and consistes of the coding regions of all the 442 GenBank® mitochondrial genome sequences that were retrieved as discussed in the previous section, as well as 42 Khoi-San and 39 Ugandan coding regions of the mitochondrial genome sequences from previous PhD theses (Koekemoer, 2010; Isabirye, 2010) and coding regions of the mitochondrial genome sequences of 50 Tswana-speaking individuals generated as part of this study. The global dataset consists of 573 mitochondrial genome sequences and is designed to be representative of the mitochondrial genome sequences of all African individuals, that is, individuals of African origin who are no longer residing on the African continent and individuals of African origin who reside on the African continent and are referred to as indigenous Africans. A full list of

the mitochondrial sequences contained in this dataset is presented in Appendix B. Global African dataset (1b) consists of a subset of 50 individuals sampled from dataset 1a for the purpose of determination of the transition:transversion ratio.

5.11.2 All African dataset

The All African dataset (2a) consists of the coding regions of all GenBank® mitochondrial genome sequences of indigenous African origin as contained in the Global African dataset and 42 Khoi-San and 39 Ugandan coding regions of the mitochondrial genome sequences from previous PhD theses (Koekemoer, 2010; Isabirye, 2010), as well as the coding regions of the mitochondrial genome sequences of 50 Tswana-speaking individuals generated as part of this study. This dataset consists of 386 mitochondrial genome sequences and is representative of the indigenous African populations. Indigenous here refers to black African people who were born and live on the continent of Africa. The mtDNA sequences that belonged to individuals who resided in non-African countries and were removed from the Global African dataset (1a) to construct the All African dataset (2a) are listed in Appendix C. As was the case with the Global African dataset, a subset of 50 individuals was sampled for the purpose of determination of the transition:transversion ratio; it is referred to as the All African dataset (2b).

5.11.3 The Tswana dataset

A third dataset, which is referred to as the Tswana dataset (3a), consists of the complete mitochondrial genomes of 50 Tswana-speaking individuals generated as part of this study and is presented in Appendix A. The Tswana dataset (3b) consists of the coding regions only of the samples contained in Tswana dataset (3a).

5.11.4 Regional African datasets

Regional mtDNA genome subsets were assigned for purposes of statistical analysis based on the All African dataset (2a). The ethnicity of the mtDNA sequences obtained from GenBank® that was used in the datasets of this investigation was published in a study by Pereira *et al.* (2009) according to country of origin or ethnic group of origin. These ethnicities and countries were allocated to broad African regions according to the CIA, World Information, Chen *et al.* (1995a), Torroni *et al.* (2001) and Salas *et al.* (2002) as described in Section 5.11.6. Three subsets were constructed for the purpose of statistical

analysis of the genetic diversity of the western, eastern and southern African regions and were referred to as Western African dataset (4), Eastern African dataset (5) and Southern African dataset (6) respectively. A comprehensive list of datasets 4, 5 and 6 is provided in Appendix D. The Western African dataset (4) consisted of 60 mtDNA sequences, the Eastern African dataset consisted of 110 mtDNA sequences and the Southern African dataset consisted of 66 mtDNA sequences. A dataset for northern Africa was not constructed because of the insufficiently small number of individuals that belonged to countries of origin from northern Africa that were contained in the All African dataset.

5.11.5 Assignment of the subsets

Two subsets were assigned for the purpose of tree building. The first from the Global African dataset (1a) is referred to as Global African subset (1b) and the second from the All African dataset (2a) is referred to as the All African subset (2b). These subsets were constructed after haplogroup assignment of all the mitochondrial DNA sequences in the datasets had been completed.

A subset sample number of 50 was decided upon and in order to ensure that the subsets were representative of the full datasets, the same percentage of each haplogroup was included in the subsets. Numbers were rounded to the highest significant decimal. Each group of samples that belonged to each of the macrohaplogroups of the respective datasets was submitted to a simple random sampling process by using the STATISTICA data analysis software system, version 9.0. StatSoft, Inc. (2009) to select the predetermined number of samples to represent the datasets (1b) and (2b).

5.11.6 Ethnicity of the individuals in the datasets

The GenBank® dataset represented mitochondrial genome sequences of individuals of many different African ethnicities with the purpose of being representative of major geographic areas of Africa. Table 5.7 lists the different ethnic groups that were included in all the datasets and the countries they originated from, as well as the regions of Africa in which the countries are located.

Table 5.7 Ethnicity, country and region of origin of the ethnic groups included in this investigation

Region	Country	Ethnic group
Northern Africa	Mauritania	Mauritania
	Morocco	---
Eastern Africa	Kenya	Kikuyu
	Tanzania	Akie, Burunge, DatogGogo, Hadza, Iraqw, Maasai, Mbugwe, Rangi, Sandawe, Turu, Wafiome
	Ethiopia	Berta
	Eritrea	---
	Sudan	Nuba
	Uganda	Acholi, Baganda, Lugbara
Western Africa	Cameroon	Bakola, Bamileke, Effik, Ewondo, Ibo (Igbo)
	Nigeria	Effik, Hausa, Yoruba, Ibo (Igbo)
	Burkina Faso	Foulbe, Rimaibe, Mossi
	Congo	Mbuti Pygmies
	Zaire	Pygmy
	---	Mbenzele Pygmies
	Ghana	---
Central Africa	Central African Republic	Biaka
Southern Africa	South Africa	Pedi, San / Khoi-San / !Kung, Sotho, Tswana, Zulu
	Namibia	Herero, San / !Kung
	Botswana	Herero
	Angola	Ambo, Herero
Middle Eastern	Jordan	---
Caribbean island	Dominican Republic	Population consists of 15% of Taino ancestry, Haiti population groups and North Africans like Lebanese, Syrians and Palestinians

Not all ethnic group identities were published in GenBank® or in the literature and this is denoted by ---. Mitochondrial genome sequence origin was only indicated by country or region and in cases where no information was published, it was denoted by ---. Zaire refers to the current Democratic Republic of Congo. Information about country of origin was compiled from CIA, World Information, Chen *et al.*, 1995a; Torroni *et al.*, 2001; Salas *et al.*, 2002.

5.12 DETERMINATION OF HAPLOGROUPS

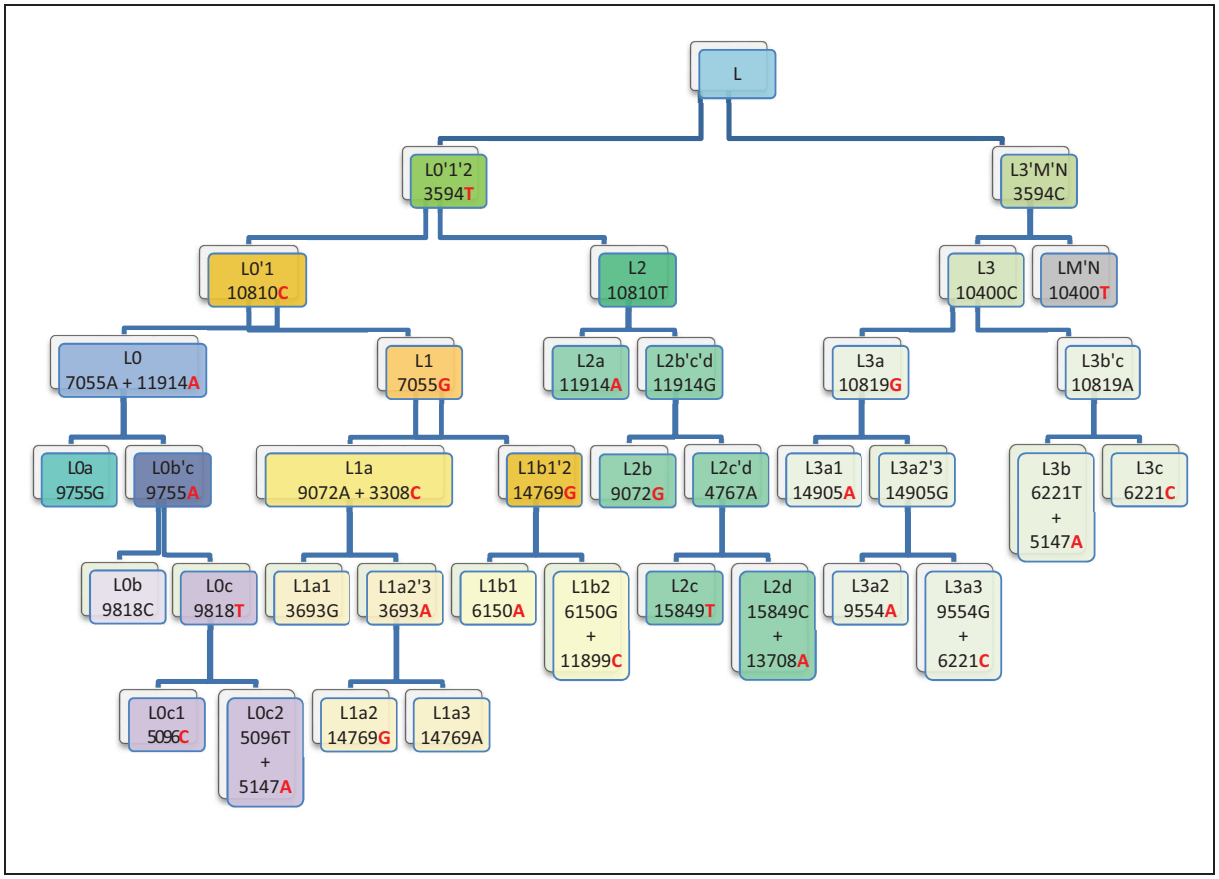
Mitochondrial haplogroup assignment developed over the years and became more detailed as technological advances provided methods through which the mitochondrial genome could be investigated at a higher resolution, as was discussed in Section 3.6.2. Many haplogroup hierarchical schemes were developed in which the shared mutations of the human mitochondrial genome, as determined by the different methods of sequence variation determination, were organised according to the occurrence of mutations of maternal lineages over time (Chen *et al.*, 1995a; Watson *et al.*, 1997; Macaulay *et al.*,

1999; Chen *et al.*, 2000; Torroni *et al.*, 2006; Quintana-Murci *et al.*, 2008). These hierarchical schemes differed in terms of the level of resolution based on the type of method that was used to investigate the sequence variation and the size and origin of the cohort that was under investigation.

Two haplogroup classification schemes were used in combination with each other to assign the haplogroups of the mitochondrial genomes of this investigation. The purpose of this approach was to ensure a broad and representative haplogroup assignment of the mitochondrial sequences under investigation. Haplogroups were determined for the Global African and All African datasets, which contained 442 and 253 mitochondrial genome sequences respectively from GenBank in addition to 42 Khoi-San individuals and 39 Bantu-speaking individuals from Uganda that were obtained from previous phylogenetic studies on individuals of African origin (Isabirye, 2010; Koekemoer, 2010) as well as determined for the 50 Tswana-speaking individuals of this investigation. The haplogroup assignments for both classification systems used in this investigation for these sequences are listed in Appendix B.

The first haplogroup classification scheme used in this investigation was based on the popular earlier high-resolution RFLP method, which determined informative sequence variants that defined mitochondrial haplogroups based on the nucleotide positions at which restriction enzymes cut the mtDNA. In this case it was developed for macrohaplogroup L. The Wallace classification scheme (Wallace, 2004) was based on 103 informative mitochondrial SNPs on the light strand of the mitochondrial DNA that either directly or in combination defined mitochondrial haplogroups of macrohaplogroup L. These SNPs are present in the coding region of the mitochondrial DNA, i.e. located in the *ATP6*, *ATP8*, *COI*, *COII*, *Cytb*, *ND1*, *ND2*, *ND3*, *ND4* and *ND5* genes of the mitochondria and presented in the hierarchical scheme in Figure 5.1.

Figure 5.1 Wallace classification system of informative SNPs used to define macrohaplogroup L

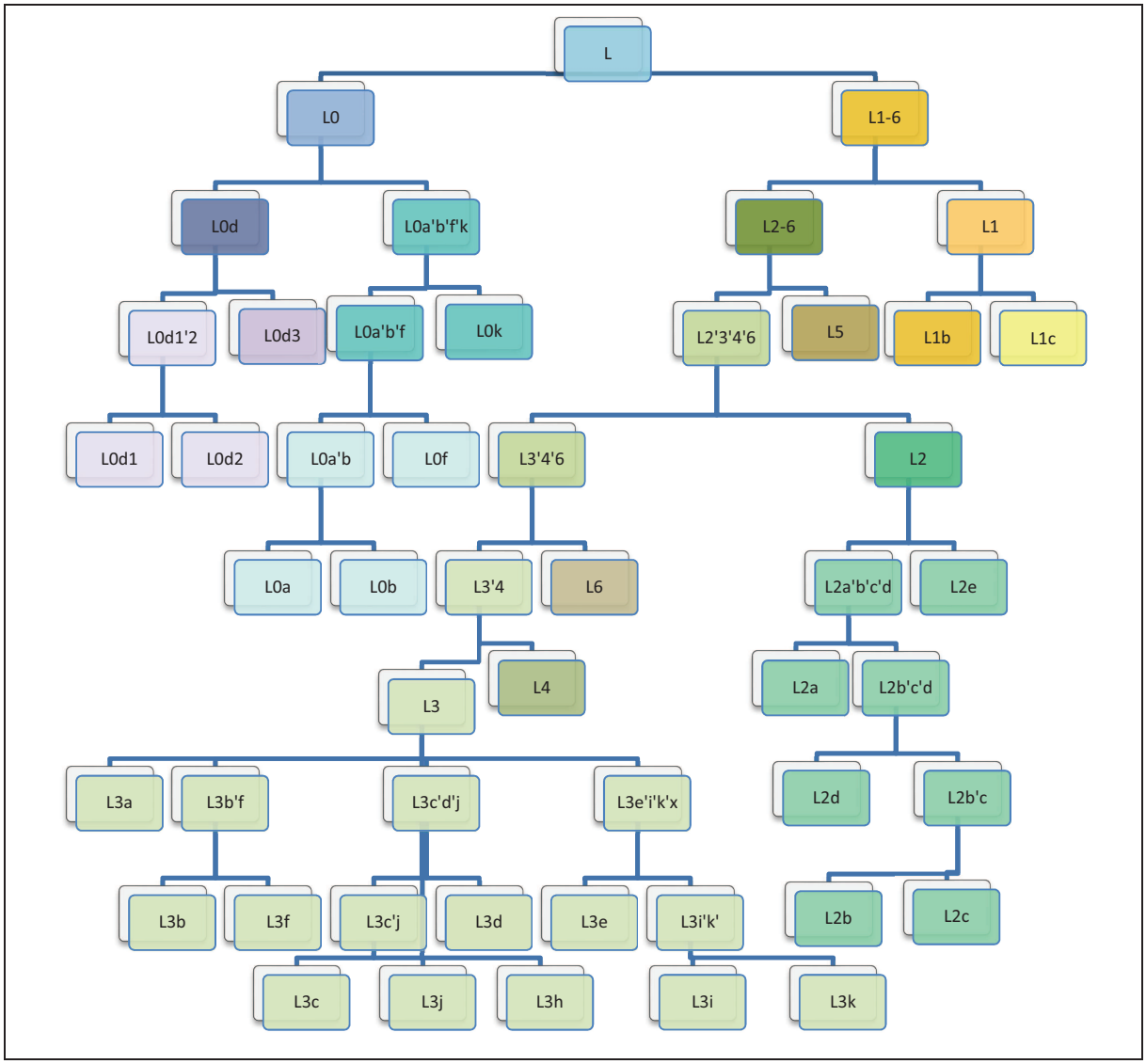


Wallace haplogroup classification scheme for macrohaplogroup L. The numbering in the text boxes indicates the informative haplogroup defining nucleotide positions. The letters that follow the nucleotide positions indicate the type of nucleotide that will define the haplogroup at a specific nucleotide position. The **bold, red** letters indicate that the type of nucleotide is different from what is 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. From Wallace, 2004.

The second haplogroup classification system that was used to assign haplogroups to the mtDNA sequences of this investigation was based on more recent full mitochondrial genome sequence data. PhyloTree (Van Oven and Kayser, 2009) is a phylogenetic tree that represents global human mitochondrial DNA variation with the purpose of providing a comprehensive phylogenetic scheme for scientists interested in studying human mitochondrial DNA variation. It was constructed by using complete mitochondrial DNA sequences of 55 publications and other published complete mitochondrial DNA sequences that were not incorporated into phylogenetic schemes before and displayed referenced sequence variation and haplogroup classification based on current phylogenetic data available in the literature (Van Oven and Kayser, 2009). It was based on a broad collection of mtDNA sequences of populations from all over the world, which included mtDNA sequences that were used in studies about populations that originated in Africa and belonged to haplogroup L (Herrnstadt *et al.*, 2002; Mishmar *et al.*, 2003; Kivisild *et al.,* 2006; Gonder *et al.*, 2007; Behar *et al.*, 2008; Quintana-Murci *et al.*, 2008; Batini *et al.*,

2011). The study by Behar *et al.* (2008) contributed 309 complete African mtDNA sequences, in addition to the 315 previously published mtDNA sequences that belonged to individuals of African origin belonging to haplogroup L, to the development of a haplogroup L hierarchy in the PhyloTree classification system (Van Oven and Kayser, 2009). This resulted in a highly resolved haplogroup structure that was important for this investigation, in which it was aimed to position the Tswana-speaking individuals of this investigation in the context of other African populations. A broad outline of the PhyloTree classification system (Van Oven and Kayser, 2009) is presented in Figure 5.2.

Figure 5.2 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 was defined by several nucleotide positions and could 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.

5.13 PHYLOGENETIC ANALYSES

Phylogenetics is a field of study that is fundamentally based on mathematical evolutionary models that incorporate biological, biochemical and evolutionary information to model the genetic relatedness of the genetic variation observed within a set of individuals or between populations of individuals (Whelan *et al*., 2001). Evolutionary theory is based on the concept that shared characteristics between individuals are attributable to a common ancestor and therefore that these genetic similarities or differences can be modelled in relationship to each other. The phylogenetic relationships of a set of DNA sequences will provide valuable information about the relatedness of the individuals and therefore about their evolutionary past.

Phylogenetic tree analysis is a popular method to analyse genetic variation within or among human populations and represents the genetic diversity between individuals, as well as the fissures and behaviour of a human population over a long period of time (Jobling *et al*., 2004). This investigation aimed at providing information about the evolutionary past and therefore relatedness of the Tswana-speaking individuals of this investigation to other African populations and individuals. For this purpose, the coding regions of the mtDNA sequences of the Tswana cohort of this investigation were used in conjunction with the other mtDNA datasets to construct phylogenetic trees.

5.13.1 Step 1: Data acquisition

A detailed description of the datasets that were used in this study is listed in Section 5.11. The control region of the mitochondrial genome has a high mutation rate, leading to reverse mutations and high variation in substitution rates between base pairs. This phenomenon can make phylogenetic inferences from this region questionable and it is acceptable practice to use only the coding region for phylogenetic analysis (Ingman *et al.*, 2000). Therefore only the coding regions of the mtDNA sequences of the Global African dataset (1a), All Africa dataset (2a) and Tswana dataset (3b) were used for the phylogenetic tree constructions of this investigation.

5.13.2 Step 2: Sequence alignment

Multiple sequence alignment was performed by using the CLUSTAL X Multiple Sequence Alignment Program version 2.0.12 (Larkin *et al.*, 2007). Clustal X has a user interface that

is more user-friendly than the Clustal W (Thompson *et al.*, 1994) version and was therefore used in this study. The algorithm of this software 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) It is, however, computationally very demanding to construct a multiple alignment of many different sequences using only this approach. For this reason, a global alignment was achieved by the software using a heuristic approach and constructing a rough NJ tree with midpoint rooting or also UPGMA guide trees (Larkin *et al.*, 2007). The tree was used as a guide to the sequences that were closely related and those that were more distant, giving direction to the alignment of the more closely related ones. When the more distant sequences were aligned, there was already a good sample of related aligned sequences that provided information about the variability of the nucleotide positions. The gap positions that were appointed during the initial alignments were not changed as new sequences were added.

As Clustal X aligns sequences only in FASTA format, the mitochondrial genome sequences were converted to FASTA file format and loaded into Clustal for alignment. On completion of the alignment, the multiple alignment files were saved in usable formats for phylogenetic and statistical analyses.

Since alignment methods are fallible (Brocchieri, 2001) and to conform to quality standards and ensure valid and reliable multiple alignment data to use in the phylogenetic analysis (Levasseur *et al.*, 2008), the multiple alignments of all sequences in the respective datasets were evaluated visually in Clustal X (Larkin *et al.*, 2007) and subsequently by using BioEdit version 7.0.5.2 (Hall, 2001). Both software applications provided suitable tools to view alignments and identify misalignment. The purpose of these verifications was to remove false deletions and insertions and verify the accuracy of the alignments.

5.13.3 Step 3: Phylogenetic analyses

The phylogenetic analyses of the Global African (1a), All African (2a) and Tswana (3b) datasets of this investigation were performed by constructing phylogenetic trees in which the maternal relatedness of the mtDNA sequences under investigation were displayed in a hierarchical tree format. In order to produce these phylogenetic trees, it was essential to dictate the assumptions of an evolutionary model under which the genetic relatedness of variation of these sets of individuals, as observed in the coding region of the mtDNA sequences of the respective datasets of this investigation, could have evolved (Whelan

et al., 2001). The assumptions, models and software that were used for the purposes of the construction of the phylogenetic trees are discussed in the following sections.

5.13.3.1 Transition:Transversion ratio calculation

It is widely accepted that transitions are more prevalent than transversions in the mitochondrial genomes of humans (Wakeley, 1993) and it was therefore important to incorporate an expected transition:transversion ratio in the software algorithms for tree building to ensure accurate phylogenetic trees. The two subsets, i.e. the Global African subset (1b) and the All African subset (2b) were used to determine the transition:transversion ratios for the Global African dataset and the All African dataset respectively. The Tswana dataset was used as is for this calculation.

PHYLIP version 3.6 (Felsenstein, 1989) was used to calculate the transition:transversion ratio for all the datasets by using the DNA parsimony (Dnapars) program. This program calculated parsimonious trees on DNA sequences and used the Fitch method (Fitch, 1971) to count the number of changes of bases needed for the most parsimonious trees. The program offered the options of either counting all the base changes or only counting the transversional changes. By using this function it was possible to obtain a single average estimate for the changes from all the most parsimonious trees for both options and by using these values to calculate the transition:transversion ratios.

The search for the most parsiminious trees was randomised by setting a random number generator to choose the input order of the samples. A random seed number was used to start a process of choosing numbers at random. The seed for the random number generator was chosen to be an integer between 1 and 2^{32}-3 (Felsenstein, 2005). The seed number was selected to adhere to the requirement of giving a remainder of one after being divided by four or in other terms, of the form 4n+1 (Felsenstein, 2005). The different seeds resulted in a random search of the trees and the software program was set to jumble the different orders of samples 100 times while constructing the tree.

5.13.3.2 Gamma-shaped parameter calculation

It has been reported that it is critical to match the variation of evolution at different sites of the sequences with the observed data when constructing phylogenetic trees of the genetic relatedness of a group of mtDNA sequences, as was discussed in more depth in Section 4.3.4. Rate variation can be described by using random values from a continuous gamma distribution by incorporating a gamma-shaped parameter in the software algorithms used to construct the phylogenetic trees. The Global African (1a), All African (2a) and Tswana (3a) datasets were used to determine the gamma-shaped parameters (alpha value) for the respective datasets. The alpha value was calculated by using the software program GZ-Gamma: Estimation of the Expected Number of Substitutions at each Amino Acid (Nucleotide) Site and the Parameter for Rate Variation among Sites under copyright of Jianzhi Zhang, Xun Gu and the Pennsylvania State University.

The datasets were analysed with PHYLIP version 3.6 (Felsenstein, 1989) using the Dnadist and NEIGHBOR executable programs to draw a guide tree. Ancestral sequences were determined by using the Kimura 2 parameter model (Kimura, 1980) and the expected number of substitutions for each site was determined by using the maximum likelihood approach under the JC model (Jukes and Cantor, 1969). The alpha value of the gamma-shaped parameter was estimated from the distribution of the expected number of substitutions by using the method of Gu and Zhang (Gu and Zhang, 1997), which is based on a likelihood approach, as was discussed in more depth in Section 4.3.4. This approach has been reported to be more accurate than the parsimony-based approaches (Gu and Zhang, 1997). Further reasons for using this approach included the fast computational times for large datasets in comparison to other software algorithms and its suitability for investigating many homologous sequences at once (Gu and Zhang, 1997).

A gamma-shaped parameter (α) was used to accommodate the different rates of substitution among sites. The α value was calculated by using the method described in Section 5.13.3.2 and the GZ-Gamma software (Gu and Zhang, 1997), which determines the α value according to a distribution of the expected number of substitutions based on the alignment of sequences in a phylogenetic tree. The Gamma value was determined by using Equation 5.1.

Equation 5.1 Equation for determining the gamma value

$$\text{Gamma value} = {}^{1}/_{\sqrt{\alpha}}$$

α value = gamma shaped parameter.

5.13.3.3 Rooting the phylogenetic trees

Phylogenetic trees are rooted with an outgroup that serves as a reference against which the phylogenetic relationships of a set of individuals are determined. The outgroup represents a starting point from where the first and deepest roots of the phylogenetic tree will be positioned and the order of the rest of the branching of the tree will be determined (Baldauf, 2003). An outgroup should, therefore, be related to the groups of sequences of the phylogenetic tree but not more closely related than any of the sequences of the phylogenetic tree are to each other (Graham *et al.*, 2001). The mtDNA sequence of the chimpanzee, *Pan troglodytes*, which is distantly related to *Homo sapiens*, was chosen as an outgroup and retrieved from GenBank® under accession number D38113 and was specified to root the trees.

5.13.3.4 Tree-building methods used in this investigation

Tree-building methods are based on different principles in structuring a vast amount of data into an evolutionary history and can for several reasons fail to build these evolutionary histories accurately. The strengths and weaknesses of the available phylogenetic tree-building methods were discussed in Section 4.4. For reasons based on the amount of data, the timeline, the hardware and software that were available and the strong and weak points of each of the existing treebuilding methods, as discussed in Section 4.4.5, it was decided to combine a genetic distance method and a discreet data method to provide different tree outcomes that could be compared to ensure valid phylogenetic results. Phylogenetic trees were, therefore, constructed in this investigation by using the NJ method developed by Saitou and Nei (1987) and the MP method (Fitch, 1971).

5.13.3.5 Neighbour-joining tree

The PHYLIP version 3.6 (Felsenstein, 1989) was used to construct the NJ trees of this investigation. The Seqboot program in PHYLIP was used to create 1,000 bootstrap

datasets of the original datasets for further phylogeny analyses. Bootstrapping is performed by sampling *N* characters randomly with replacement and with the assumption that the characters evolve independently, resulting in a dataset of the same size that was started with. The characters in the bootstrapped datasets are, therefore, randomised. A random seed number was used to start the randomised sampling and was estimated in the same way as described in Section 5.13.3.1.

PHYLIP version 3.6 (Felsenstein, 1989) provided the JC (Jukes and Cantor, 1969), Kimura-2-parameter (Kimura, 1980), the F84 (Felsenstein and Churchill, 1996) and LogDet distance model (Steel, 1994) as evolutionary models. The F84 model was chosen because it relaxed the assumption that the four bases were at equilibrium and incorporated the prevalence of unequal nucleotide frequencies based on a bias towards transition substitutions (Felsenstein and Churchill, 1996).

The Dnadist program in PHYLIP was used to compute the distance matrices for purposes of NJ tree building in this investigation. The genetic distances were calculated under the F84 model, which incorporated calculated transition:transversion rates and allowed different nucleotide frequency rates. This distance method incorporated site-specific rates of change and allowed for inequalities of the rate of transitions and transversions and base composition (Felsenstein and Churchill, 1996). The transition:transversion ratio and input to the program of the coefficient of variation of the rate of substitution among sites ($1/\sqrt{\alpha}$) were provided, as described in Sections 5.13.3.1 and 5.13.3.2.

The Neighbour program in PHYLIP was used to generate the NJ tree from the output files from the Dnadist program. This program used the NJ method of Saitou and Nei (1987) and the UPGMA method of clustering.

5.13.3.6 Consensus trees

On completion of the NJ tree calculation, the output file of the NJ tree was used to determine the consensus tree by using the Consense program in PHYLIP version 3.6 (Felsenstein, 1989). This program was designed to construct a consensus tree from computer readable trees in the same way it was constructed by the Neighbour program. This approach used methods that included the strict consensus, majority rule and majority rule extended methods. The majority rule, which included sets of sequences that were present in more than 50% of the input trees, and the majority rule extended, which

included the sequences that were present in more than 50% of the input trees and also added any other lower frequency sequences that were compatible to the tree until the tree was fully resolved, were selected for use. These methods were discussed in more detail in Section 4.6.3.

5.13.3.7 Maximum Parsimony tree

The Molecular Evolutionary Genetics Analysis (MEGA) software Version 5.0 (Tamura *et al.*, 2011) was used to generate the maximum parsimony trees. The Clustal X aligned sequence files, as described in Section 5.13.2, with gaps included, were used for MP tree drawing. The Clustal file format was converted to a MEGA file by using the functionality within the MEGA software. On completion, a final consensus tree was produced from the many different parsimonious trees that were produced following the majority-rule option, as discussed in Section 5.13.3.6, i.e. including branches that were present in more than 50% of trees.

In order to assess the reliability of the MP trees, the bootstrap test was selected under the phylogeny test options with 1,000 replicates under a random seed number of 64,238. The close-neighbour-interchange (CNI) heuristic search method was selected above the branch-and-bound search method, which was too time consuming when working with a dataset of more than 15 sequences (Tamura *et al.*, 2011). The CNI search method used a branch-swapping method that reduced time spent searching for the optimal tree by producing a temporary tree through the random addition of sequences. It then examined all the topologies of the trees that were produced that were different from the initial temporary tree by a topological distance of two to four to obtain a final optimal tree.

5.14 STATISTICAL ANALYSES

The evolutionary history of the Tswana population under investigation and the African population as represented by the different datasets, that contained a broad range of African individuals, were investigated by studying the evolutionary signature left in the nucleotide changes of the human mitochondrial genome. Although mutation is regarded as the fundamental source of genetic diversity it is not the only factor that contributes to observed patterns of genetic diversity in human populations. The evolutionary forces that determine genetic drift and natural selection in a complex interplay with the mutational effects on the human genome were determined in this investigation in addition to the

sequence variance in order to determine the forces that were ultimately responsible for the observed genetic diversity of the Tswana-speaking individuals of this investigation.

Mathematical models exist to describe the processes of evolutionary change in populations and form the basis on which statistical methods extract information to determine the role that each of these forces play in shaping genetic variation. Evolutionary history was inferred from the genetic sequence variation data of the datasets of this investigation by using statistical methods to measure genetic diversity in the context of the evolutionary history of the populations studied.

Statistical analyses in this investigation were performed on the Global African dataset (1a), the All African dataset (2a), the Tswana dataset (3a and 3b), the Western African dataset (4), the Eastern African dataset (5) and the Southern African dataset (6), as described in Section 5.11 and Appendix B, Appendix C and Appendix D. Using the control and coding regions of the mtDNA genome in the statistical determinations can be problematic. The high frequency of convergent mutations in the control region of the mtDNA genome can obscure the evolutionary signal, thus influencing the estimation of nucleotide diversity and TMRCA (Tamura and Nei, 1993). Rate heterogeneity between the control and the coding regions of the mtDNA genome is another factor that is problematic when calculating divergence time estimates (Excoffier and Schneider, 1999) and therefore to avoid these problems, only the coding regions of the mtDNA sequences of the Global African, All African, Western, Eastern and Southern African regional datasets and Tswana datasets were used for purposes of genetic diversity estimations. To determine the amount of genetic diversity, basic statistical methods to measure nucleotide diversity within populations were employed by using different computer program applications, as listed in Table 5.8.

Table 5.8 Software programs used for statistical analyses

Software program	Version	Type of analyses	Short description of functionalities used	Software reference
MEGA	5.0	Descriptive statistics Population subdivision Neutrality tests	Nucleotide diversity Evolutionary distances between sequences and populations Selective neutrality	Tamura *et al.*, 2011

Table 5.8 Continued…

Software program	Version	Type of analyses	Short description of functionalities used	Software reference
Arlequin	3.5.1.2	Descriptive statistics Population subdivision Neutrality tests Demographic expansion Inter-population comparisons Divergence times	Genetic diversity Nucleotide diversity Population parameter Population differentiation Selective neutrality within populations by using Tajima's and Fu's FS neutrality tests Mismatch distribution AMOVA Genetic distances between populations (F-statistics)	Excoffier and Lischer, 2010
DnaSP	5.10.01	Descriptive statistics Population subdivision Neutrality tests Demographic expansion Divergence times Migration rates	Pairwise distances between populations Demographic parameters from mismatch distributions Selective neutrality	Librado and Rozas, 2009

Descriptive statistics refer to standard indices and molecular diversity indices.

The number of segregating sites (*S*), the average number of nucleotide differences (*k*), the nucleotide diversity (π) and population parameters (θ) were determined as measures of intra-population genetic variation within the datasets of this investigation. The Tswana dataset was subjected to tests for population subdivision as an indicator of the effect of genetic drift within the population. Mismatch distribution tests were performed on all the datasets as a descriptive measure of the genetic diversity and as an inferential measure of population expansion.

To determine the evolutionary force that selection played on the observed genetic diversity, Tajima's and Fu and Li's neutrality tests for selective pressure under the infinite site model were performed on all the datasets. These methods were discussed in more detail in Section 5.14.4. Inter-population analyses were used to determine the genetic distances between the populations datasets used in this investigation by using Wright's F statistic (F_{ST}) value and as discussed in Section 5.14.5.

5.14.1 Nucleotide composition

Basic diversity statistical analyses were performed on the Tswana population under investigation as a descriptive measure of the sequence variation observed within the Tswana population. MEGA version 5 (Tamura *et al.*, 2011) was used to determine the

nucleotide composition of the complete mtDNA genomes of the Tswana dataset (3a). The G + C concentration of the same dataset was determined by using the DnaSP version 5 (Librado and Rozas, 2009) software program. MEGA version 5 (Tamura *et al.*, 2011) was furthermore used to determine the nucleotide composition and G+C content of the respective codon positions across the mitochondrial genome. The G+C content of the codon positions 1 and 2 were also determined by using DnaSP version 5 (Librado and Rozas, 2009) and used to verify the results obtained by MEGA version 5.

5.14.2 Nucleotide diversity

All of the calculations for nucleotide diversity in this investigation were performed by using DnaSP version 5 (Librado and Rozas, 2009). Nucleotide diversity was determined for the coding regions of the mtDNA genomes of the Global African, All African, Western, Eastern and Southern African, and Tswana datasets respectively as described in Section 5.11 and Appendix D and was measured by the number of *S* and the average number of nucleotide differences (*k*). The *k* (Tajima, 1983), the stochastic variance ($V_{st}(k)$) and sampling variance ($\hat{V}_s(k)$) and the total variance of *k* ($\hat{V}(k)$) were determined by using Equation 5.2 (Tajima, 1993).

Equation 5.2 Average number of nucleotide differences (*k*) and nucleotide diversity (π)

1. Average number of nucleotide differences (*k*):

A. Average number of nucleotide differences:

$$\hat{k} = \sum_{i<j} \sum k_{ij} / \binom{n}{2}$$

B. Sampling variance of *k*:

$$\hat{V}_s(k) = \frac{2(3n-1)k + 2(2n+3)k^2}{11n^2 - 7n + 6}$$

C. Stochastic variance of *k*:

$$V_{st}(k) = \frac{(3n^2 - 3\widehat{n+2})k + 2n(n-1)k^2}{11n^2 - 7n + 6}$$

D. Total variance of *k*:

$$\hat{V}(k) = \frac{3n(n+1)k + 2(n^2+n+3)k^2}{11n^2 - 7n + 6}$$

2. **Nucleotide diversity (π)**

$$\hat{\pi} = \sum_{i \neq j} \hat{d}_{ij} / [n(n-1)]$$

Equation 1: $\hat{k}$ = average number of nucleotide differences, k_{ij} = the number of nucleotide differences between sequence i and j, n = number of sequences samples from a population, $\binom{n}{2} = n(n-1)/2$ is the total number of sequence comparisons. From Tajima (1983) and Tajima (1993). Equation 2: $\hat{\pi}$ = nucleotide diversity in a population, $\hat{d}_{ij}$ = estimate of the number of nucleotide substitutions per site between sequence i and j, n = number of sequences, $\sum_{i \neq j}$ = all pairwise comparison. From Nei and Jin (1989).

The nucleotide diversity (π) was determined for each of the datasets by determination of the average number of nucleotide differences per site between two sequences by using Equation 5.2 (Nei and Jin, 1989). In addition, the sampling variance (V) of the nucleotide diversity (π) was calculated by using Equation 5.3 and the standard deviation (SD) determined by the square root of the variance (Tajima, 1983).

Equation 5.3 Sampling variance (V) of nucleotide diversity (π)

$$V(\hat{\pi}) = \frac{1}{[n(n-1)]^2}\left[\sum_{i \neq j} V(\hat{d}_{ij}) + \sum_{i \neq j}\sum_{k \neq l} Cov(\hat{d}_{ij}, \hat{d}_{kl})\right]$$

$V(\hat{\pi})$= variance of nucleotide diversity, n= number of DNA sequences, $\hat{d}_{ij}$= number of nucleotide substitutions between sequences, $\sum_{i \neq j}$ = all pairwise comparison, $Cov(\hat{d}_{ij}, \hat{d}_{kl})$= based on phylogenetic relationship among sequences. From Nei and Jin (1989).

5.14.3 Population size

Population expansion was determined by using different statistical approaches to identify traces of population expansion in the Tswana population under investigation as well as in the Global African, All African and regional African datasets of this investigation with the purpose to identify signals of past population growth of the Tswana population under investigation in context of the population behaviours of the other African populations. The datasets used in this investigation were described in Section 5.11 and consisted of two broadly sampled African populations, the Global African and All African datasets, consisting of the mtDNA sequences of individuals of African origin residing in African and non-African countries, as well as smaller regional datasets, i.e. the Western, Eastern and Southern African datasets, that contained the mtDNA sequences of individuals of specific African regions.

Firstly, Fu's F_S statistic, which is based on the assumption that a population that has expanded will display a large number of rare alleles or singletons, was used to determine past population growth within the datasets of this investigation. The F_S test has been described as the most powerful statistical test to indicate population growth (Fu, 1997) and was applied by using DnaSP version 5 (Librado and Rozas, 2009) software for this purpose, by performing calculations for the different datasets using Equation 5.4. The F_S test was performed on the assumptions that no recombination was present and that the mutation parameter was equal to the genetic diversity of the population (θ_π), by equating it

to the average number of observed pairwise differences of the sequences (π) of the population sample.

Equation 5.4 Fu's F_s statistic

$$F_S = ln\left(\frac{S'}{1-S'}\right)$$

S' is calculated from θ_π (Tajima's estimate) where $\theta = 2Nu$ where N = the size of the haploid population and u is the mutation rate per generation; S = number of segregating sites. From Fu (1997).

The significance of F_S was determined by using the coalescence simulation functionality of the DnaSP version 5 software by re-computing the F_S statistic for random samples from a stationary population. Five thousand simulations were performed to determine the null distribution of the F_S statistic, based on a critical point at the lower second percentile of the empirical distribution (Fu, 1997).

The mismatch distribution is regarded as a statistic that describes genetic diversity of a population as well as an inferential statistic that is used to infer population size and growth. Population size and growth are inferred based on the frequencies of the number of sequence differences between pairs of mtDNA genomes within a population as an indication of whether sequence diversity was retained within the population, which would be an indication of population growth, or whether the sequence diversity was lost, indicating that the population under investigation was of constant size, and that sequence variation was lost in response to genetic drift (Harpending, 1994). Therefore the mismatch distribution was used to investigate population size and growth of the Tswana population under investigation, as well as of the other African datasets of this investigation as used in the determination of Fu's F_S tests. The mismatch distributions were determined by using the DnaSP version 5 (Librado and Rozas, 2009) software under the hypothesis of constant population size and no recombination. The distribution was displayed as the observed pairwise nucleotide differences in context of the expected number of pairwise differences under the assumptions previously mentioned and by using Equation 5.5.

Equation 5.5 Mismatch distribution under constant population size and no recombination

$$Q(i) = \frac{1}{1+\theta}\left(\frac{\theta}{1+\theta}\right)^i$$

$Q(i)$ = the number of differences between a pair of genes where i is the number of different genes; $\theta = 2N\mu$ where N = the size of the haploid population and u is the mutation rate per generation. From Slatkin and Hudson (1991).

The mismatch distribution of a stationary population would be ragged and that of an expanding population would be smooth with a peak (Harpending, 1994). The smoothness of the mismatch distribution is, therefore, an indication of whether the population under investigation was stationary or expanding and was determined by using the raggedness statistic or *rg* (Harpending, 1994) as determined by DnaSP version 5 (Librado and Rozas, 2009) and Equation 5.6.

Equation 5.6 Raggedness statistic (*rg*)

$$r = \sum_{i=1}^{d+1} (x_i - x_{i-1})^2$$

rg= the raggedness statistic; *d*= the number of differences in the mismatch distribution; *x*= the mismatch distribution, *i* = the number of sequences. From Harpending (1994).

The confidence intervals and statistical significance (*P*) of the raggedness statistic were determined by using computer simulations. These simulations were based on the coalescence algorithm in DnaSP version 5 (Librado and Rozas, 2009) software.

In addition, the Ramos-Onsins and Rozas' R_2 test was also used to indicate population expansion. This test is based on the principle that a growing population would display a large number of private mutations or singletons in contrast to a stationary population that would display smaller numbers of singletons. The R_2 statistic was determined by using DnaSP version 5 (Librado and Rozas, 2009) software using Equation 5.7. As with the other statistical measures determined, the confidence intervals and statistical significance (*P*) of the R_2 statistic were determined by using computer simulations that were based on the coalescence algorithm in DnaSP version 5 (Librado and Rozas, 2009) software.

Equation 5.7 Ramos-Onsins and Rozas R_2 statistic

$$R_2 = \frac{\left(\sum_{i=1}^{n} \left(U_i - \frac{k}{2} \right)^2 \Big/ n \right)^{1/2}}{S}$$

R_2= Ramos-Onsins and Rozas R_2 statistic; *n*=sample size; *S*=total number of segregating sites; *k*=average number of nucleotide differences between two (2) DNA sequences; U_i=the number of singletons in sequence *i*. From Ramos-Onsins and Rozas (1994).

5.14.4 Selection

Quantification of the effects of natural selection and random genetic drift is critical to the determination of the levels of genetic diversity within species. Many of the applications of mtDNA as a genetic marker in population genetics are based on a model of selective neutrality within the mtDNA genome, as was the case in the interpretation of the genetic signals of population growth in Section 5.14.3. In order to verify the genetic signals of population growth determined within the datasets of this investigation, it is important to determine deviations from neutrality and the effects of selection on the mtDNA data used in these determinations.

The statistical methods used in this investigation for testing the effects of selection within a population compared models of sequence evolution against the Kimura (1971) neutral theory of molecular evolution. Neutral sequence evolution was rejected if the observed data disagreed extensively with the expectations of the neutral model, which predicted a Poisson distribution of nucleotide differences between individuals and an equal ratio of nonsynonymous (NS) to synonymous (S) substitutions within the coding regions of the mtDNA genome. This was determined by using site frequency spectrum methods, which are based on the distribution of nucleotide substitutions within the mtDNA genome and by comparative methods that investigate the rate of amino-acid substitutions within the protein coding regions of the mtDNA, i.e. the frequency of observed nonsynonymous substitutions or NS (Nielsen *et al.*, 2007). The effects of selection on the mtDNA genome of participants in this study were further determined by using a whole-genome approach to detect overall violations of the neutral model and a gene-by-gene approach to detect violations of selection within specific regions of the mtDNA genome.

Tajima's *D* test of neutrality (Tajima, 1989) and the D* and F* tests for neutrality as proposed by Fu and Li (1993) were used to detect an excess of substitutions of intermediate frequency relative to the high and low frequency substitutions observed in the mtDNA genome. These tests were performed on the Global African, All African, Western, Eastern and Southern African datasets and on the Tswana dataset (3b) of this investigation. In addition, the major L haplogroups of the All African dataset were also investigated to compensate for the possibility that the regional datasets were not sufficiently representative of all the haplogroup polymorphisms that had developed within the African populations.

Tajima's *D* test was reported to be the most powerful statistical test for neutrality under the condition that sample sizes of at least 50 individual sequences were used (Simonsen *et.al.*, 1995). It compares *S* to π within the mtDNA sequences of the dataset. A skewed frequency spectrum towards intermediate-frequency substitutions or low-frequency substitutions indicates positive or negative *D* values respectively (Tajima, 1989). Significant *D* values were interpreted as evidence of deviation from neutrality within a sample of mtDNA sequences through selection, population growth or population substructure. Tajima's *D* statistic was determined by Equation 5.88. as used in the DnaSP version 5 (Librado and Rozas, 2009).

Equation 5.8 Tajima's D statistic

Tajima's D:

$$D = \frac{d}{\sqrt{\hat{V}(d)}}$$

d is defined as:

$$d = \hat{k} - \frac{S}{a_1}$$

a is defined as:

$$a_1 = \sum_{i=1}^{n-1} \frac{1}{i}$$

D = Tajima's D statistic; $\hat{V}$ = variance; *k* = average number of nucleotide differences between pairs of sequences; *n* = number of sequences; *S* = total number of segregating sites; *i* = 1 to n-1, From Tajima, 1989.

The confidence limits for Tajima's *D* statistic were determined by using the assumption that *D* followed the beta distribution as described by Tajima (1989) and by using the confidence limits of *D,* also described by Tajima, 1989. In addition to Tajima's D statistic, Fu and Li's D* and F* statistics were also determined as a measure of deviation from neutrality. Fu and Li's statistical tests are based on the principle that the presence of negative selection would cause an excess of private mutations to be present in the terminal branches of a phylogenetic tree as opposed to the number of mutations present in the internal branches of the phylogenetic tree. The opposite holds true for positive selection. The total number of mutations among the sequences, as opposed to the single mutations among the sequences, provide an estimate for the D* test statistic. Neutrality was rejected if the observed value did not fall within the expected differences under an assumption of neutrality (Fu and Li, 1993). DnaSP version 5 (Librado and Rozas, 2009) was used to determine the D* test statistic according to 0. An excess of private mutations was evident from negative D* values and a deficiency of private mutations in terminal branches was evident from positive D* values.

Equation 5.9 Fu and Li's D* test statistic

Fu and Li's D:

$$D^* = \frac{\left(\frac{n}{n-1}\right)\eta - a_n\eta_s}{\sqrt{u_{D*}\eta + v_{D*}\eta^2}}$$

v_{D^*} is defined as:

$$v_{D*} = \left[\left(\frac{n}{n-1}\right)^2 b_n + a_n^2 d_n - 2\frac{na_n(a_n+1)}{(n-1)}\right] / (a_n^2 + b_n)$$

$u_{D^*}$is defined as:

$$v_{D*} = \frac{n}{n-1}\left(a_n - \frac{n}{n-1}\right) - v_{D*}$$

a_n is defined as:

$$a_n = \sum_{k=1}^{n-1} \frac{1}{k}$$

b_n is defined as:

$$b_n = \sum_{k=1}^{n-1} \frac{1}{k^2}$$

D^* = Fu and Li's D test statistic; n = number of sequences; η =total number of mutations ; η_s = number of singleton mutations; k = average number of nucleotide differences between pairs of sequences, a = ∑(1/i) from i=1 to n-1; From Fu and Li (1993).

The difference in the occurrence of single mutations and the mean number of pairwise nucleotide substitutions was used to estimate the Fu and Li's F* test statistic when using DnaSP version 5 (Librado and Rozas, 2009) according to Equation 5.10. An excess of private mutations was evident from negative F* values and a deficiency of private mutations in terminal branches was evident from positive F* values.

Equation 5.10 Fu and Li's F* test statistic

Fu and Li's F*:

$$F^* = \frac{\Pi_n - \left(\frac{n-1}{n}\right)\eta_s}{\sqrt{u_{F*}\eta + v_{F*}\eta^2}}$$

v_{D^*} is defined as:

$$v_{F*} = \left[d_n + \frac{2(n^2+n+3)}{9n(n-1)} - 2\frac{1}{n-1}\left(4b_n - 6 + \frac{8}{n}\right)\right] / (a_n^2 + b_n)$$

$u_{D^*}$is defined as:

$$v_{F*} = \frac{\left[\frac{n}{n-1} + \frac{n+1}{3(n-1)} - 2\frac{2}{n(n-1)} + 2\frac{n+1}{(n-1)^2}\cdot\left(a_{n+1} - \frac{2n}{n+1}\right)\right]}{a_n - v_{F*}}$$

a_n is defined as:

$$a_n = \sum_{k=1}^{n-1}\frac{1}{k}$$

b_n is defined as:

$$b_n = \sum_{k=1}^{n-1}\frac{1}{k^2}$$

*F** = Fu and Li's F test statistic;π_n = the mean number of pairwise differences for n sequences; *n* = number of sequences; η =total number of mutations ; η_s = number of singleton mutations; *a* = ∑(1/i) from i=1 to n-1; From Fu and Li (1993).

The statistical significance for Fu and Li's D* and F* test statistics was determined in DnaSP version 5 (Librado and Rozas, 2009) by using percentage points for distributions of D* and F*. The critical points were determined from empirical distributions, which were computed by using a large number of sample simulations and determined conservatively by using a θ value of between two and 20 (Fu and Li, 1993).

To investigate the role of selection within the different regions of the mtDNA, the number of synonymous (S) and nonsynonymous (NS) substitutions was investigated for each of the haplogroups L0, L1, L2 and L3 of the All African dataset of this investigation. The regional datasets were not included because of the under-representation of all the major L haplogroup types per region. This would have skewed the NS/S ratios to such an extent that the true signals of selection would have been obscured. The Global African dataset was excluded based on the fact that the non-African component was too small to contribute critically to the outcome of the analysis for the same reasons as used for the exclusion of the regional datasets. The All African dataset was included based on the fact that it consisted of the pooled major L haplogroups and constituted a large representative dataset of African individuals of different regions and all major L haplogroups.

5.14.4.1 Gene-specific effects of selection

To investigate the regional effects of selection on the mitochondrial genome, the 13 protein-coding genes of the mtDNA coding region were investigated individually by following the same approach as for the investigation of the whole genome through the determination of the ratio of NS and S substitutions. An increase in the number of NS to S substitutions within a gene suggests an increase in amino acid substitutions and can, therefore, be interpreted as evidence of positive or adaptive selection, whereas the opposite may be interpreted as evidence of negative or purifying selection (Nei and Gojoborit, 1986). To determine a more valid signal of selection, it was decided to distinguish between the evolutionary ages of the substitutions observed in the mtDNA data of this investigation by identifying two different classes of substitutions. The substitutions were classified as haplogroup-associated nonsynonymous (NS_H) and synonymous (S_H) substitutions that were present in the internal branches of the phylogenetic tree and a second group, which was classified as private nonsynonymous (NS_P) and synonymous (S_P) substitutions; these were present in the mtDNA sequences located on the terminal branches of the phylogenetic tree. The haplogroup-associated substitutions were determined by the presence of substitutions in the mtDNA of at least two individuals that belonged to the same haplogroup clade. Identical substitutions that arose independently in different haplogroups were counted for each haplogroup. In contrast to this class of substitutions, the private substitutions were counted as substitutions that could only be observed in a single individual within a haplogroup. A conservative counting approach was followed, in which a substitution was counted only once, even if the same substitution occurred in another clade of the same haplogroup. Homoplastic substitutions were excluded by this approach to counteract the possibility that some of the haplogroup-associated substitutions that were shared early in the lineages were counted more than once. Elson *et al.* (2004) reported that this approach was valid and did not bias the analysis towards negative selection.

The model of neutrality predicted that the NS and S substitutions would be subjected to the same evolutionary forces and would therefore be expected to be present in equal distributions. Based on this principle, the frequency of the NS substitutions could be normalised by dividing it by the number of S substitutions (NS/S), which was determined for both observed classes of substitutions. Determination of these two values allowed for an estimate of the ratio between the frequencies of these substitutions as an indication of whether selection was at play within the population. An index of neutrality (NI) (Rand and

Kann, 1996) was determined for the substitutions that were observed by using Equation 5.11.

Equation 5.11 Neutrality Index

$$NI = \frac{NS/S_P}{NS/S_H}$$

NI = neutrality index; *NS* = nonsynonymous substitution; *S* = synonymous substitution; *P* = private substitutions; *H* = haplogroup associated substitutions. From Rand and Kann, 1996.

The NI estimated the degree of deviation from neutrality with an NI value of one that reflected absolute neutrality. NI values > 1 indicated purifying selection due to an excess of NS_P and NI values < 1 indicated adaptive selection due to an excess of NS_H. Statistical significance was determined by an exact parametric randomisation test of the actual observed values of the two classes of substitutions. DnaSP version 5 was used to perform the one-tailed Fischer's exact test of independence (Templeton, 1996).

5.14.5 Population genetic structure

When new populations are formed through the subdivision of existing populations, the subpopulations generally display allele frequencies that are different from the parent population because of the founder effect on the demes. In addition, the allele frequencies in the newly formed demes would be a sample of the parent population and therefore the genetic diversity would be smaller, which would cause genetic drift to push the fixation of alleles more quickly. It would be expected that the haplotype diversity and the number of segregating sites between demes would increase (Excoffier and Lischer, 2010). Wright's *F* statistic (Wright, 1965), as an estimate of the genetic structure of populations as described by Weir and Cockerham (1984), was used in this investigation to determine the covariances of sequence diversity between the populations under investigation.

The genetic structure of the Global African, All African, Western, Eastern and Southern African and Tswana datasets of this investigation were investigated by using an analysis of molecular variance (AMOVA) approach as defined and used by Excoffier *et al.* (1992) in the Arlequin software version 3.5.1.2 (Excoffier and Lischer, 2010). The genetic structures that were investigated were defined within single groups between two populations for analyses. The Tswana population was compared with the Global African and All African datasets respectively to determine the genetic distances from the African and non-African populations and then analysed with the respective regional datasets to determine the

genetic distances between the Tswana population and the populations that resided in the different regions of Africa. The regional datasets were also compared to each other to determine the genetic distances between them. This was used to interpret the genetic distances obtained for the Tswana population under investigation in the context of the other African regions.

Arlequin version 3.5.1.2 (Excoffier and Lischer, 2010) uses a hierarchical approach to the analysis of molecular variance by the construction of a matrix that consists of the squared differences between all pairs of haplotypes. In a multidimensional distance matrix consisting of different levels of subdivision as used in Arlequin version 3.5.1.2, the conventional sum of squares becomes a sum of squared deviations (SSD) that are converted to the mean squared deviations (MSD) by dividing by the degrees of freedom. Each of the SSD values was accompanied by covariance components ($\sigma_a^2, \sigma_b^2, \sigma_c^2$) that were extracted by equating the MSDs to their expected values (Excoffier *et al.*, 1992). The F_{ST} estimate was regarded as the correlation of random haplotypes within a population with random haplotypes within all the populations together and calculated by Equation 5.12.

Equation 5.12 Fixation Index (F_{ST})

$$F_{ST} = \frac{f_0 - f_1}{1 - f_1} = \frac{\bar{t}_1 - \bar{t}_0}{\bar{t}_1}$$

F_{ST} = fixation index; f_0 = probability of identity by descent of two different genes drawn from the same population; f_1 = probability of identity by descent of two genes drawn from two different populations; $\bar{t}_0$ = is the mean coalescence time of two genes drawn from the same population; $\bar{t}_1$ = the mean coalescence times of two genes drawn from two different populations. From Slatkin and Hudson (1991).

The significance of the F_{ST} statistics and covariant components was determined under the null hypothesis that samples are obtained from a global population without any population substructure and that variation was due to random sampling in the determination of the populations (Excoffier *et al.*, 1992). This was achieved by a non-parametric random permutation of the haplotypes among populations and subsequent determination of the covariant estimates of each permutation. The total sum of squares was partitioned and F-statistics and covariance components were determined according to Equation 5.13.

Equation 5.13 Determination of total sum of squares, F-statistics and covariance components for haplotype data within one group

Hierarchical groups	Degrees of freedom	SSD	Expected MSD
Among populations (*AP*)	*P* - 1	SSD (*AP*)	$n\sigma_a^2 + \sigma_b^2$
Within populations (*WP*)	*N* - *P*	SSD (*WP*)	σ_b^2
Total	*N* - 1	SSD (*T*)	σ_T^2

P = total number of populations; *N* = total number of gene copies; MSD = mean squared deviation; SSD (*T*) = total sum of squared deviations; SSD (*AP*) = sum of squared deviations among populations; SSD (*WP*) = sum of squared deviations within populations; σ_a^2 = covariance component due to differences among populations; σ_b^2 = covariance component due to differences among haplotypes in different populations within a group; σ_T^2 = total molecular variance. From Excoffier and Lischer (2010).

The estimates for *n* and F_{ST} were defined by Equation 5.14.

Equation 5.14 Estimates for *n* and fixation index (F_{ST}) defined

n was defined by:

$$n = \frac{N - \sum_p \frac{N_p^2}{N}}{P - 1}$$

F_{ST} was defined by:

$$F_{ST} = \frac{\sigma_a^2}{\sigma_T^2}$$

F_{ST} = fixation index; *N* = total number of gene copies; *p* = population; *P* = total number of populations; σ_a^2 = covariance component due to differences among populations; σ_T^2 = total molecular variance, From Excoffier and Lischer (2010).

5.14.6 Coalescence-time estimation

The TMRCA 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. Only the coding regions of the mtDNA sequences were used for this purpose because of the high incidence of reverse mutations and mutational hotspots in the control region that could affect the outcome of the time estimations (Ingman *et al.*, 2000; Gonder *et al.*, 2007; Batini *et al.*, 2011). The indels were also not taken into account because of their hypervariability. The lineages were classified according to the PhyloTree classification system (Van Oven and Kayser, 2009)

Genetic dating was performed by following the model-free lineage approach as described by Forster (2004) by equating the genetic distances between the phylogenetic clusters with elapsed time through the measurement of the number of substitutions observed in the respective phylogenetic lineages under investigation. This measurement was referred to as the demographically unbiased parameter rho, *ρ*, and was determined by counting the number of substitutions to the nearest ancestral nodes of the founding major lineages of

the Tswana individuals of this investigation and of the All African MP phylogenetic tree as discussed in Section 6.9.6 (Forster *et al.*, 1996). The statistical significance of the measurement was determined by determining the standard deviation (SD) or σ of ρ that reflected the number and type of lineages that had developed over time within a specific haplogroup cluster, as depicted in the sample under investigation (Saillard *et al.*, 2000). The genetic dating was determined by multiplying the number of substitutions with the mutation rate of the coding region of the mtDNA (Forster *et al.*, 1996).

The time estimates were determined by using the MP phylogenetic tree of the All African dataset (2a) of this investigation as described in Section 6.9.6 to count the number of observed substitutions (l_i) along the branches of individuals that belonged to the i^{th} haplogroup or sub-haplogroup or lineage, which was in turn divided by the number of individuals of the i^{th} haplogroup or sub-haplogroup or lineage. The p_i's of the respective haplogroups or sub-haplogroups or lineages were determined as described in Equation 5.15, which reflected the average distance to the most recent common ancestor per haplogroup or sub-haplogroup or lineage.

Equation 5.15 Estimator of the genetic distance to the ancestral node of a haplogroup, sub-haplogroup or lineage

$$\rho_i = {}^{l_i}/_{n_i}$$

p_i = the demographically unbiased estimator of the average genetic distance to the root of a node; n_i= the number of individuals present in the i^{th} haplogroup, sub-haplogroup or lineage; l_i = the number of observed substitutions in the i^{th} haplogroup, sub-haplogroup or lineage. From Saillard *et al.* (2000).

The standard deviation (σ^2) was determined as described in Equation 5.16. This measure was an estimate of the efficiency of the coalescence-time estimation of the sample (Saillard *et al.*, 2000).

Equation 5.16 Efficiency of coalescence-time estimation of a sample

$$\sigma^2 = \frac{\rho}{n}$$

Thus:

$$\sigma = \sqrt{{}^{\rho}/_{n}}$$

ρ = average genetic distance to the root of a node; σ = standard deviation also known as estimator of variation of ρ; n = number of individuals. From Saillard *et al.* (2000).

A molecular clock with an average rate of 1.26 X 10^{-8} nucleotide substitutions per site per year or in other terms, a rate of 5,138 years per coding region nucleotide substitution, was

used to determine the coalescence time (Mishmar *et al.*, 2003; Behar *et al.*, 2008). This substitution rate was assumed to be representative of the whole mitochondrial genome and based on a separate treatment of transitions and transversion substitution rates as described in the HKY85 substitution model (Hasegawa *et al.*, 1985).

5.15 CONSTRUCTION OF A CONSENSUS SEQUENCE FOR THE TSWANA-SPEAKING COHORT OF THIS INVESTIGATION

Ancestral heritage was determined in this investigation according to a phylogenetic approach in which ancestral states were estimated based on evolutionary models. The methodology of phylogenetic analysis is based on the key assumptions of a clock-like model for evolution, lack of recombination and lack of selection in human mitochondrial genomes. Several studies have, however, cast some doubt on these assumptions, as the clock-like properties of evolution in the mtDNA were questioned (Howell *et al.*, 2003), as well as whether the mtDNA fits the neutral model of evolution (Excoffier, 1990; Mishmar *et al.*, 2003; Kivisild *et al.*, 2006) and whether recombination might be the underlying reason for certain homoplasies observed in the mtDNA of humans (Zsurka *et al.*, 2007). Based on these uncertainties supplemental methodologies, such as the construction of a consensus sequence, are used to verify the ancestral sequence of a sample of mtDNA sequences. The purpose of the construction of a consensus sequence is to reflect on the ancestry of the mitochondrial genomes of a population and represent the degree of human mitochondrial variation that is typical of that population of mtDNA sequences (Carter, 2007).

A consensus sequence was constructed for the full genome mtDNA sequences of the Tswana-speaking cohort of 50 individuals of this investigation. The same multiple sequence alignment of mtDNA sequences that was used for the phylogenetic analyses by using the CLUSTAL X Multiple Sequence Alignment Program version 2.0.12 (Larkin *et al.*, 2007) as described in Section 5.13.2 was used for the construction of the consensus sequence. The rCRS was used as a standard for nucleotide numbering purposes. The mitochondrial consensus sequence was constructed by selecting the majority sequence variant at each of the variable positions of the full mitochondrial genomes of the 50 Tswana-speaking individuals of this investigation by using BioEdit version 7.0.5.2 (Hall, 2001). The gaps and missing data were included in the determination of the consensus sequence and manually verified on completion of the construction of the consensus sequence.