Reconstruction of Major Male and Female Lineages of the Kensington Muslim Community

A thesis submitted in partial fulfilment of the requirements for the degree of Magister Scientiae in the Department of Biotechnology, University of the Western Cape

Shafieka Isaacs

September 2010

Supervisor: Prof. Mongi Benjeddou (PhD)

Keywords

Genetic Variation

Mitochondrial DNA (mtDNA)

Single Nucleotide Polymorphisms (SNPs)

Y-Chromosome

Haplogroups

Lineages

Polymerase Chain Reaction (PCR)

Restriction Fragment Length Polymorphism (RFLP)

Electrophoresis

SNaPshot Minisequencing Assay

Genotyping

Maternal Ancestry

Paternal Ancestry

Abstract

Reconstruction of Major Male and Female Lineages of the Kensington Muslim Community

S. Isaacs

MSc Thesis, Department of Biotechnology, University of the Western Cape

The earliest Cape Muslims were brought to the Cape from Africa and Asia from 1652 to 1834. They were part of an involuntary migration of slaves, political prisoners and convicts, and they contributed in the ethnic diversity of the present Cape Muslim population. In addition to migrations, this population has grown by combined and continuous multiplex processes of intermarriage, natural increase, conversion and blending with other communities. Previous studies on the Cape Muslim population and local Cape Muslim communities have taken advantage of the huge amount of archival material covering the past three hundred years of Cape Muslim history in South Africa.

The aim of this study, however was to investigate the genetic diversity and origins of one of these local Muslim communities, by taking advantage of the availability of modern molecular genetics tools. A pilot study conducted with Muslim volunteers residing all over the Cape Metropolitan area was used as a point of reference for the Kensington Muslim community study. Samples were investigated for DNA polymorphisms from both maternal (mitochondrial DNA) and paternal (Y-chromosome) lineages. Variations in these two types of DNA are grouped into continent-specific haplogroups or lineages. Six mtDNA and eight Y- chromosome SNP markers were screened using polymerase chain reaction- restriction fragment length polymorphisms (PCR-RFLP).

Maternal lineages of Cape General Muslims were 34% African, 33% Asian and 33% European. The highest frequency for African lineages was observed in Cape Coloured Muslims (47%) and Cape Malay Muslims (45%). Cape Indian Muslims however displayed a higher frequency for Asian lineages (45%), while Cape Other Muslims for European lineages (50%). Paternal lineages indicated that 5% were of African descent, 75% Asian, 13% European and 7% of lineages remained undefined. Asian lineages were the most dominant paternal lineage with the highest frequency observed in Cape Indian Muslims (97%).

In the second part of the study, contributions of major maternal African, Asian and European haplogroups to the Kensington Muslim community gene pool was investigated. This was achieved by examining mtDNA variations found within this community using twelve mtDNA coding region SNPs genotyped by the SNaPshot minisequencing assay. Maternal lineages of Kensington General Muslims indicated that 55% were of African descent, 34% Asian, 1% European and 10% were of Eurasian descent. African lineages were the most frequent lineages of Kensington Coloured Muslims (56%), Kensington Malay Muslims (67%) and Kensington Other Muslims (100%). Kensington Indian Muslims maternal lineages were mainly of Eurasian origin (43%).

The last part of the study investigated the contribution of the major male continent-specific lineages in the Y-chromosome diversity of the Kensington Muslim community by typing thirteen Y-chromosome SNP markers. Paternal lineages of Kensington General Muslims indicated that 4% were of African origin, 51% Asian and 45% European. The highest frequency for Asian lineages was observed in Kensington Malay Muslims (56%), Kensington Indian Muslims (71%) and Kensington Other Muslims

(100%). However, Kensington Coloured Muslims paternal lineages were mainly of European origin (49%).

The data obtained for Cape Coloured Muslims and Cape Malay Muslims and their Kensington Muslim counterpart's maternal and paternal gene pool were in agreement with historical findings concerning the origins of earlier Cape Muslims. Cape Indian Muslims and Kensington Indian Muslims maternal and paternal gene pool however demonstrated a strong accordance to mtDNA and Y-chromosome lineages observed in indigenous Indian populations.

Declaration

I, Shafieka Isaacs, hereby declare that this dissertation is my own work. This thesis is submitted in partial fulfilment of the requirement for the degree of Magister Scientiae in the Department of Biotechnology at the University of the Western Cape. It has not been submitted before for any degree or examination at this or any other University and that all the sources I have used or quoted have been indicated and acknowledged by complete references.

Full Name: Shafieka Isaacs

Signed:

Date: September 2010

Acknowledgements

I would like to thank all the donors particularly the Kensington Muslim Community for their participation in the study and for making this research possible. I especially would like to thank Prof. Yusuf Da Costa for making available his publications and his daughter Galiema, for her assistance in arranging collections of biological samples from the Kensington Muslim Community. I would also like to express my gratitude to Mrs M. Oliver and Mr D. Essack for their assistance in the collection of samples.

I thank my supervisor Prof. Mongi Benjeddou for his patience and support throughout the years in the preparation of this dissertation.

Special thanks to Prof. Jasper Rees for making resources available in his laboratory to conduct part of my research experiments.

I also wish to acknowledge and thank the National Research Foundation (NRF) for financial assistance.

I am truly grateful to my family and friends for their love, patience and encouragement throughout my years of studying and research.

List of Abbreviations

A, G, C, T	Adenine, Guanine, Cytacine, Thymine
BP, KBP, MBP	Base pairs, Kilo base pairs, Million base pairs
COI	Cytochrome Oxidase Subunit
D-LOOP	Displacement loop (control region) of mtDNA
DNA	Deoxyribonucleic Acid
dNTPs	Deoxynucleotide Triphosphates
ddNTPs	Dideoxynucleotide Triphosphates
EXO I	Exonuclease I
HVS-I/HVS II	Hypervariable segment one and two
$MgCl_2$	Magnesium Chloride
MtDNA	Mitochondrial DNA
NP, NT	Nucleotide position, Nucleotide
NEB	New England Biolabs
NRY	Non-Recombining Y-chromosome
PCR	Polymerase Chain Reaction
PNG	Papua New Guinea
RFLP	Restriction Fragment Length Polymorphism
rCRS	Revised Cambridge Reference Sequence
RFU	Relative Fluorescent Units
SAP	Shrimp Alkaline Phosphatase
SNPs	Single Nucleotide Polymorphisms
STRs	Short Tandem Repeats
$tRNA^{Lys}$	Lysyl transfer RNA
YCC	Y Chromosome Consortium

Definitions of Basic Terms used in Thesis

CLADE	A group of species (or, sometimes, individuals) that share a closer common ancestry with one another than with any other such group; a monophyletic assemblage (Avise 2006).
HAPLOTYPE	A sequence type that comprises all identical sequences (Serk 2004).
HAPLOGROUP	A group of haplotypes that share a common ancestor defined by an array of synapomorphic substitutions (Serk 2004).
LINEAGE	Any array of characters/mutations shared by more than one haplotype (Serk 2004).
MONOPHYLETIC	Groups which include all the descendants of the most common ancestor (Avise 2006)
PARAPHYLETIC	An artificial assemblage that includes a common ancestor and some but not all of its evolutionary descendents (Avise 2006)

List of Figures

List of Tables

Table of Contents

Chapter 1- Review of the Literature

Chapter 2 - Reconstruction of Major Maternal and Paternal Lineages of the Cape Muslim population

Chapter 1: Review of the Literature

1.1. Introduction

Genetic variation among the human population have significantly impacted studies of medical genetics, DNA forensics, human evolution and molecular genealogy (Hammer et al. 2001; Hutchinson et al. 2004; Tishkoff and Kidd 2004; Lind et al. 2007). Protein polymorphisms and ABO blood groups were among the first genetic markers initially screened for variations (Jorde et al. 1998; Hammer et al. 2001; Wells et al. 2001; Cavalli-Sforza and Felman 2003). These markers provided valuable insights into population's genetic structure however had limitations, as these systems were few in number, uniform among populations and were occasionally affected by natural selection (Jorde et al. 1998). However, advancements in molecular technologies during the 80's and 90's ushered in a new era allowing genetic variation to be examined on a DNA level. This has allowed scientist for the first time to characterise different aspects of human evolution, migration and genetic divergence among populations (Jorde et al. 1998; Cavalli-Sforza and Felman 2003, Hutchinson et al. 2004). These molecular technologies have also provided modern molecular genetics with the necessary tools to reconstruct unknown genealogies of human populations (Hutchinson et al. 2004).

1.2. Types of Genetic Data

Autosomal DNA, X Chromosomal DNA, Mitochondrial (mtDNA) and Y-chromosome DNA (Y-DNA) represent four classes of DNA that are routinely used to study population genetics (Jorde et al. 1998; Cavalli-Sforza and Felman 2003, Hutchinson et al. 2004). Their differing mode of inheritance has made it possible to either trace recent or ancient genealogies of any population. This provides knowledge into a population's genetic structure indicating founder events, sex-biased gene flow and the ongoing history of a population. However, only two DNA types namely mitochondrial DNA

and the Y-chromosome are extensively used by molecular genetists to reconstruct ancestral genealogies of a population (Zerjal et al. 1997; Underhill 2000; Behar et al. 2007; Lind et al. 2007; Hutchinson et al. 2004). Since only one variant of mtDNA and Y-chromosome DNA can be transmitted in contrast to autosomal and X chromosomal DNA making these loci ideal candidates to determine the origins of a contemporary population's diversity (Jorde et al. 1998; Underhill 2003; Cavalli-Sforza and Feldman 2003; Hutchison et al. 2004).

1.2.1. Mitochondrial DNA (MtDNA)

The human mitochondrial DNA (mtDNA) genome is a closed double stranded circular molecule, 16,569 base pairs (bp) in length (Fig 1.1) (Wallace et al. 1999; Ballard and Whitlock 2004; Parkendorf and Stoneking 2005). The genome is located in the mitochondria, which is involved in both cellular respiration and oxidative phosphorylation. There are hundreds of mitochondria in cells each containing an abundance of mtDNA molecules. This feature alongside a lack of recombination, a high mutation rate and the mtDNA mode of inheritance has made it popular in tracing a population's maternal line (Sunnucks 2000; Budowle et al. 2003; Ballard and Whitlock 2004).

Mother's pass their mitochondrial DNA directly to both male and female children. Males however, are unable to transmit their mitochondria to their offspring. MtDNA is therefore transmitted unaltered to each female and carriers with it information regarding their ancestral maternal line (Wallace et al. 1999; Sunnucks 2000; Lell and Wallace 2000; Hutchinson et al. 2004). This maternal line is grouped into diverse mtDNA haplogroups according to distinct polymorphisms. These polymorphisms are mainly identified by detecting variations occurring within the mtDNA coding region,

and sequencing the two-hypervariable regions in the displacement loop (D-loop) (Wallace et al. 1999; Van Oven and Kayser 2008).

The D-loop is a 1.1kbp segment situated in the control region containing elements responsible for replication and transcription but is unable to code for DNA (Hutchinson et al. 2004). Yet it represents the most variable region of the mtDNA with an estimated mutation rate of 0.075-0.165 $x10^{-6}$ substitution/site/per year according to pedigree studies. The D-loop has been influenced by genetic drift and natural selection and has acquired mutations at an appreciable frequency (Lell and Wallace 2000; Hutchinson et al. 2004; Parkendorf and Stoneking 2005). Nevertheless, it offers limited discriminating powers as haplogroups identified in this region are poorly resolved. Analysis of single nucleotide polymorphisms (SNPs) in the coding region however has resolved this major drawback and subsequently increased mtDNA haplogroup discrimination (Quintáns et al. 2004; Álvarez-Iglesias et al. 2007).

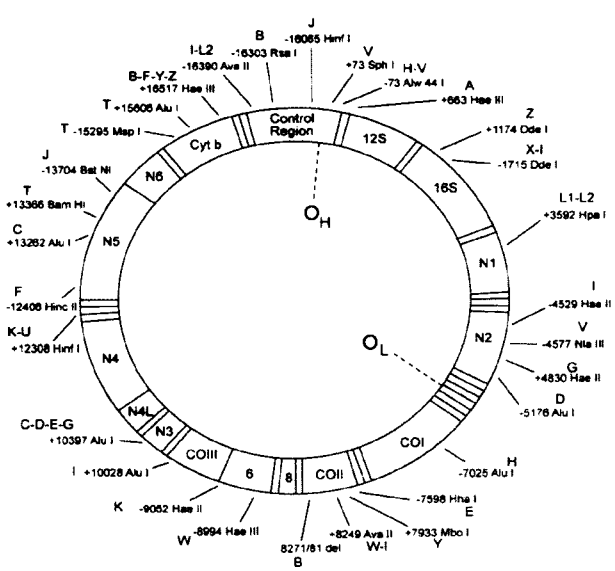

Figure 1.1: The human mitochondrial DNA genome with major mtDNA haplogroups identified in the control and coding region using RFLP analysis (adapted from Rubicz et al. 2007).

1.2.2. Y-Chromosome DNA

The Y-chromosome constitutes the largest non-recombining block of nucleotide (NRY), or haploid (non-paired) DNA in the human genome (Fig 1.2). The region consist of 60 million base pairs (Mbp) of DNA representing 95% of the entire genome (Jobling 2001; Butler 2003; Hammer and Zegura 2002; Rootsi 2004). Only two segments known as the pseudoautosomal regions undergoes recombination with the X chromosome which represents less than 3Mb of the genomes 60Mbp length (Jobling and Tyler Smith 2003; Rootsi 2004).

A large portion of the genome is unaltered by meiotic recombination and therefore preserves the allelic state of the Y-chromosome DNA (Underhill et al. 2001; Hammer and Zegura 2002; Underhill and Kivisild 2007). This will be inherited only by males from their fathers due to the sex-determining role of the Y-chromosome. Thus like its maternal counterpart from the mitochondria, the Y-chromosome DNA may be used to trace the paternal line of a population (Jobling and Tyler Smith 2003; Hammer et al. 2006; Onofri et al. 2006; Kayser et al. 2005).

Single nucleotide polymorphisms (SNPs) and short tandem repeats (STRs) markers are the main types of markers usually used in male genealogy studies. Both of these markers offer insights into the paternal ancestry of a population. However, they reflect different time scales of a population's history. Y-STRs indicate recent events while Y-chromosome SNPs reveal more ancient events (Hutchison et al. 2004; Onofri et al. 2006; Kaiser et al. 2005 ; Karafet et al. 2008).

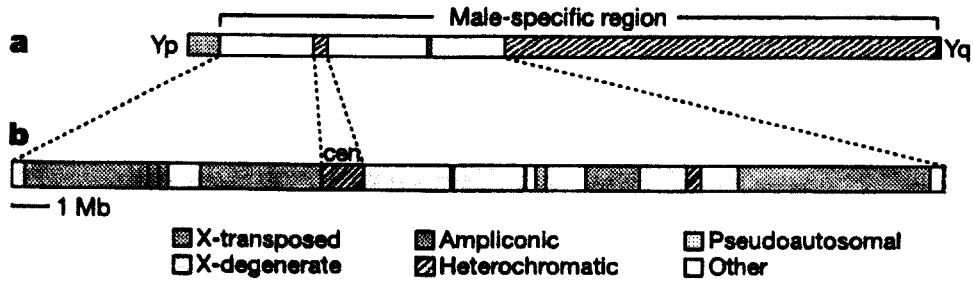

Figure 1.2: The male-specific region of the Y-chromosome. A) Is a representation of the entire Y-chromosome. B) Is an enlarged view of the euchromatic region with various euchromatic sequence classes being shown (adapted from Skaletsky et al. 2003).

1.3. Genetic Genealogy Markers

1.3.1. Short tandem repeats (STR)

Short tandem repeats (STRs) are short DNA sequences consisting of repeating units ranging from 1-6 base pairs (Hammer and Zegura 2002; Hutchinson et al. 2004). The number and sequence of these individual repeat units may vary, and are abundant in its distribution throughout the euchromatic parts of genomes (Hutchinson et al. 2004). Pedigree studies estimate STR's mutation rate at 2×10^{-3} per generation which is higher than SNP's thereby allowing for high levels of allelic variation (Hammer and Zegura 2002; Hutchinson et al. 2004). STRs can therefore have a dozen or more alleles whereas SNPs will only have two allelic states. This feature has therefore made STR's a useful tool for the elucidation of human population history and in forensic analysis (Hammer and Zegura 2002; Hutchinson et al. 2004).

1.3.2. Single Nucleotide Polymorphisms (SNPs)

Single Nucleotide polymorphisms (SNPs) are the simplest and most frequent class of genetic polymorphisms in the human genome (Sobrino et al. 2005). These polymorphisms occur as insertions, deletions or base substitutions at single positions in

the genome (Butler 2003; Budowle 2004). The latter represent the most common form of the polymorphism occurring either between purines (A>G) and pyrimidines (C>T) or between these two groups (A>T) (Budowle 2004; Hutchinson et al. 2004).

Single nucleotide polymorphisms in both the mitochondrial and Y-chromosome DNA provide insight into the history and evolution of their respective lineages. This is mainly attributed to the low mutation rate of SNPs, which enables a low-resolution view of maternal and paternal lineages (Butler 2003). Maternal (MtDNA) and paternal (Y-chromosome) lineages are also highly geographic and population specific, and are therefore routinely employed when examining the genealogical origins of populations (Sanchez et al. 2003; Hutchinson et al. 2004; Hammer et al. 2006).

A number of methods are currently available to genotype mtDNA and Y-chromosome SNPs this includes conventional techniques such as restriction fragment length polymorphism (RFLP) or one of the more automated techniques. The newly developed automated typing technologies allow for the rapid screening of multiple SNPs using multiplex amplification and detection formats. This provides a more cost effective and less time-consuming alternative to conventional techniques (Budowle 2004; Vallone and Butler 2004; Álvarez-Iglesias et al. 2007).

1.4. SNP Typing Methodologies

The growing interest in SNPs and their utilization in a variety of applications such as medical diagnostics, human identity testing and population genetics have lead to the development of a number of new typing technologies and assays (Vallone and Butler 2004; Budowle 2004; Sobrino et al. 2005). These were developed according to allelic discrimination and detection formats. Products of allelic discrimination can be detected

with various methods, which analyze reactions performed with different assay formats (Budowle 2004; Sobrino et al. 2005).

The main technologies used in SNP typing include allele specific hybridization, primer extension, oligonucleotide ligation and invasive cleavage. Each of these technologies can use several detection methods such as electrophoresis fluorescence, fluorescence resonance energy transfer (FRET), luminescence and mass measurement (Budowle 2004; Sobrino et al. 2005).

This review will focus mainly on two of these technologies namely Luminex and SNaPshot due to their simultaneous evaluation of multiple SNP markers (Butler 2003; Vallone and Butler 2004). This review will also cover some of the conventional methods as a number of laboratories still frequently utilize these techniques.

1.4.1. Conventional Typing Methods

1.4.1.1. Restriction Fragment Length Polymorphism (RFLP)

Restriction Fragment Length Polymorphisms (RFLP) analysis identifies variations in the mtDNA and Y-chromosome using restriction endonucleases (Parker et al. 1998; Schurr 2000). The region containing the variation or SNP is first amplified using polymerase chain reaction (PCR) (Navajas and Fenton 2000; Schlotterer 2004). This product is then analysed for the polymorphism using a restriction endonuclease that cleaves at a specific nucleotide sequence known as a recognition site. Since each polymorphism is unique, it will therefore have a unique recognition site. An individual will therefore be assigned to a haplogroup based on the RFLP pattern generated from restriction digestion reactions after agarose gel electrophoresis (Fig1.3.) (Parker et al. 1998; Navajas and Fenton 2000; Schlotterer 2004). A number of polymorphic restriction sites are screened prior to the haplogroup assignment using either low