Forensic DNA Profiling of a Population Sample from Upper (South) Egypt

Thesis submitted for the degree of
Doctor of Philosophy
at the University of Leicester

by

Ghada Ali Farghali Omran, MSc
East Midlands Forensic Pathology Unit
University of Leicester

April 2008

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

To my dear Husband, lovely Sons; Mohannad and Mahmoud and my dear Mother and Father

CONTENTS

Acknowledgments

The work presented in this thesis was funded through a scholarship from the Egyptian Ministry of Higher Education. I would like to express my deep thanks and gratitude for all who supported me in obtaining such an enjoyable and valuable chance to study in the United Kingdom. Throughout the duration of this PhD, I have been helped by many people to whom I would like to convey my sincere thanks and acknowledgments.

The majority of my thanks are directed towards my supervisors, **Prof. Guy N Rutty** and **Prof. Mark A Jobling** for giving me the opportunity to undertake a project under their guidance and also for providing continuous support, both academically and personally throughout the duration of my PhD studentship. I'm extremely grateful to Mark for his valuable advice in all aspects of my thesis, being patient in revising it though under continuous rush from my side. His help in contacting several authors for obtaining comparative molecular data and facilitating the use of the Genetics Department's equipment, especially the 3100 Genetic Analyzer, before arrival of that of the Forensic Pathology Unit, are much appreciated. I also thank the Genetic Department's staff namely **Dr. Emma Parkin** for technical support and **Dr. Patricia Balaresque** for advice on statistical analyses.

I'm also indebted to **Dr. Martin Evison** for his valuable supervision during the first year of this PhD which was carried out in the University of Sheffield. I also gratefully acknowledge and thank **Prof. John Buckleton** for his help in conducting the statistical analyses of Chapter 4. Many thanks I'd like to express to **Dr. Peter de Knijff** for supplying unpublished comparative data on the CEPH-HGDP diversity panel along with all other scientists who also provided similar data. All DNA donors from Upper Egypt and Sheffield's colleagues are highly acknowledged as without their donation, this research wouldn't have been possible.

I gratefully express many thanks to my colleagues and friends in Leicester; especially **Eleanor Graham**, my senior colleague, for giving me useful technical advice and taking the time to read and comment on my thesis. Special thanks to my dear friend and housemate, **Hind Abdulmajeed** for her continuous encouragement and support during the writing up period of this thesis.

Finally, I am deeply indebted to my dear husband and lovely sons for the constant support, understanding, patience and love that I received throughout my studentship period. Much of the burden of this work has fallen on them as well as me and they withstood it bravely, many thanks to them. I am also very grateful to my parents for their endless love and valuable support.

Abstract

Ghada Ali Farghali Omran

Title: Forensic DNA Profiling of a Population Sample from Upper (South) Egypt

The work presented in this thesis was carried out to assess the genetic diversity and establish a forensic population database for a southern population sample inhabiting Upper Egypt. The Egyptian population encompasses several ethnic groups that need to be genetically characterised and evaluated for possible substructure before subsequent forensic use. Upper Egyptians are one of the major ethnic groups, exhibiting cultural and linguistic differences with the Northern population inhabiting Lower Egypt.

A sample of 265 unrelated individuals from five of Upper Egypt's governorates was collected with informed consent, and profiled using fifteen autosomal STR loci contained in the AmpFℓSTR®Identifiler™PCR Amplification Kit. Statistical analyses indicated neither departure from expectations of Hardy-Weinberg equilibrium in most of the loci nor dependence of alleles between loci, allowing multiple locus profile frequency estimation using the product rule. All loci were polymorphic; the most discriminating is D18S51 while the least is TPOX. The combined power of exclusion was 0.9999986824 and the combined match probability was 1.93×10^{-18}.

Additionally, seventeen Y-STR loci included in the AmpFℓSTR®Yfiler™PCR Amplification Kit were typed in 208 males from the same population. Of 204 observed haplotypes, 200 were unique (96.6%) and four were found twice each. The 17-locus discriminating power was 0.9998. DYS385a/b and DYS458 showed the highest diversities with a high frequency of microvariant and new alleles (22% of the sample) in DYS458 locus. Other loci revealed duplicated, novel and null alleles such as DYS437, DYS635 and DYS448 respectively.

The generated data on Upper Egyptians from both STR systems were compared with data on other global, regional and local populations. Locally, some genetic differentiation was observed with other ethnic groups including the Northern population, suggesting likely population substructure among Egyptians. Global and regional comparisons demonstrated a regional genetic continuity among populations of the Nile valley, Middle East and Arabian Peninsula.

List of Abbreviations

A	Adenine
ABI	Applied Biosystems
AD	Anno Domini
AMELY	*Amelogenin Y* gene
AMOVA	Analysis of molecular variance
AZF	Azoospermia factor
BC	Before Christ
bp	base pairs
C	Cytosine
CEPH	Centre d'Etude du Polymorphisme Humain
CODIS	Combined DNA Index System
DNA	Deoxyribonucleic acid
ESWG	English Speaking Working Group
ExHt	Extended haplotype
FBI	Federal Bureau of Investigation
FSS	Forensic Science Service
FW	Formula weight
g	gram/ times gravity
G	Guanine
h	hour
HERV	Human endogenous retrovirus
Hg	Haplogroup
HGDP	Human Genome Diversity Project
HLA	Human leukocyte antigen
HVS I	Hypervariable segment I
HVS II	Hypervariable segment II
HWE	Hardy-Weinberg equilibrium
ISFG	International Society for Forensic Genetics
kb	kilobases/ kilobase-pairs
L	Litre
LE	Llinkage equilibrium

LINE	Long interspersed nuclear element
M	Molar
Mb	Megabases/ megabase-pairs
MDS	Multidimensional scaling
mg	milligram
min	minutes
MinHt	Minimal haplotype
mM	millimolar
mtDNA	Mitochondrial DNA
NDNAD	National DNA database
ng	nanogram
nm	nanometre
PAR 1	Pseudoautosomal region 1
PAR 2	Pseudoautosomal region 2
PCR	Polymerase chain reaction
pg	picogram
pH	Potential of Hydrogen
RFLP	Restriction fragment length polymorphism
rpm	revolutions per minute
sec	seconds
SGM	Second generation multiplex
SINE	Short interspersed nuclear element
SLP	Single locus probe
SNP	Single nucleotide polymorphism
SRY	Sex-determining region, Y
STR	Short tandem repeat
SWGDAM	Scientific Working Group on DNA Analysis Methods
T	Thymine
Taq	*Thermus aquaticus*
TMRCA	Time to most recent common ancestor
UEP	Unique event polymorphism
UP	Ultra pure
USA	United States of America
UV	Ultra violet

VNTR	Variable number tandem repeat
YHRD	Y-Haplotype Reference Database
µl	microlitre
µM	micromolar

1. Introduction

1.1 General background

Forensic science is a broad speciality aimed at helping justice either in civil or criminal cases. The field comprises many interacting areas such as chemistry, biology, physics and mathematics. It also includes various disciplines that target human identification, the most pronounced of which was the use of latent fingerprints first described by Henry Faulds in 1880 and classified later by Francis Galton in 1892 (Jobling and Gill, 2004). More than two decades have elapsed since Professor Sir Alec Jeffreys of the University of Leicester discovered 'DNA fingerprinting' which has revolutionised forensic investigations (Jeffreys *et al.*, 1985c). Since then the use of genetics for resolution of legal cases and the general field of forensic genetics has been continuously expanding. As DNA is present in all nucleated cells (e.g. blood, semen, and hair), any such source can provide information that links a suspect with a crime scene. Therefore, DNA analysis today has a key role for conviction or exoneration of suspects, and identification of victims of mass disaster and crimes. The process combines the methods of molecular biology, standard techniques of population genetics, and statistics in addition to widely available population and intelligence databases. All these methods are implemented for overall assessment of significance of a match and interpreting the results of DNA typing (Jobling and Gill, 2004).

1.1.1 A brief history of DNA marker development

More than a century ago, the human ABO blood group polymorphism was the first haemogenetic system discovered to differentiate individuals (Landsteiner, 1900). This simple system can unambiguously exclude a suspect but cannot prove with certainty an inclusion, as many subjects carry the same blood group. Until the 1980s, serological and protein electrophoretic methods were used to assess the diversity in several systems of blood group antigens (e.g. Rh, MNSs, Kell) and other serum proteins (e.g. Hp, Gc, Tf). The main drawbacks for such analyses were the high amount of biological material required, liability for bacterial contamination and rapid degradation in addition to low variability and informativeness (the match probability was 0.01-0.001 for eight systems used for blood stain analysis) (Jobling and Gill, 2004; Goodwin *et al.*, 2007).

Concurrent advances in molecular biology such as restriction enzymes, Southern blotting and sequencing techniques in 1960s and 1970s paved the way for scientists to examine DNA sequences. By 1978, Southern blotting was used to detect DNA polymorphism (Kan and Dozy, 1978) which was followed by description of the first highly polymorphic locus in 1980 (Wyman and White, 1980). The true DNA fingerprinting revolution started with the discovery of hypervariable loci known as 'minisatellites' or variable number tandem repeats (VNTRs) (Jeffreys *et al.*, 1985b) and demonstration of their potential use in human identification (Jeffreys *et al.*, 1985c). First, these minisatellites were detected by hybridisation of probes to Southern blots of restriction-digested DNA fragments. Shared core sequences between different minisatellite loci allowed the probe to detect many of them simultaneously producing the hypervariable multi-band patterns known as DNA fingerprints. The resulting match probability was low ($< 3x10^{-11}$ for single probe) (Gill *et al.*, 1985; Jeffreys *et al.*, 1985c). A combination of single-locus probes (based on cloned individual minisatellites) was then used for some time in paternity and criminal case investigations. Although analysis of such markers was a powerful tool, it also has many disadvantages as will be discussed later (Jobling and Gill, 2004).

The invention of PCR by Kary Mullis in 1983 (Saiki *et al.*, 1985; Mullis *et al.*, 1986) transformed the whole field of molecular biology including forensic genetics. This method allowed amplification of minute amounts of DNA even if degraded, as commonly encountered in forensic casework. Early uses of the technique involved a small number of Single Nucleotide Polymorphisms (SNPs) in the HLA-DQA1 gene (Helmuth *et al.*, 1990). The low discriminating power of this system and the difficulty of interpreting mixtures limited its use to cases where SLP technology failed (Jobling and Gill, 2004).

The discovery of 'Short Tandem Repeats' (STR) and automated sequencing technology led to the current use of the most powerful system for human identification that has now completely replaced the old systems. STR markers have the advantage of high discriminating power, sensitivity and the potential ability to resolve simple DNA mixtures. Further advances in DNA extraction and quantification techniques, the development of multiple commercial STR multiplexes

and high levels of standardisation and quality controls, have resulted in the creation of several national STR databases. Consequently, the DNA evidence is now widely recognised as a reliable forensic tool globally (Jobling and Gill, 2004; Goodwin *et al.*, 2007).

1.2 Human nuclear genome arrangement

The DNA is often referred to as 'the blueprint' for every individual, carrying all the information required for maintaining all essential processes needed for any organism to function and reproduce. This information is coded by the sequences of only four different nitrogenous bases, adenine, guanine, thymine and cytosine on the sugar phosphate backbone. In each nucleated human somatic cell there are two copies of the genome, one inherited from each parent. Each copy is packaged into 23 chromosomes and contains 3.2 billion base pairs (bp) of DNA. The Human Genome Project, initiated in 1990, has led to remarkable advances in sequencing and the ultimate understanding of the structure of the human genome. In 2004, the sequence of 99% of the euchromatic DNA was described with high accuracy (99.9%) (Collins *et al.*, 2004). The genome can be divided into different sequence classes based on structure and function as shown in Figure 1-1. The coding and associated regulatory regions, that encompass genes and other sequences for gene function regulation, constitute only a small proportion. On the other hand, the extragenic non-coding regions comprise the largest and most important classes to the forensic geneticist - and in particular repetitive DNA sequences (55% of the genomic total). Repetitive DNA can be divided into two main sub-groups: interspersed repeats (45%) and tandem repeats (10%) including satellites, minisatellites and microsatellites (Collins *et al.*, 2004; Jasinska and Krzyzosiak, 2004).

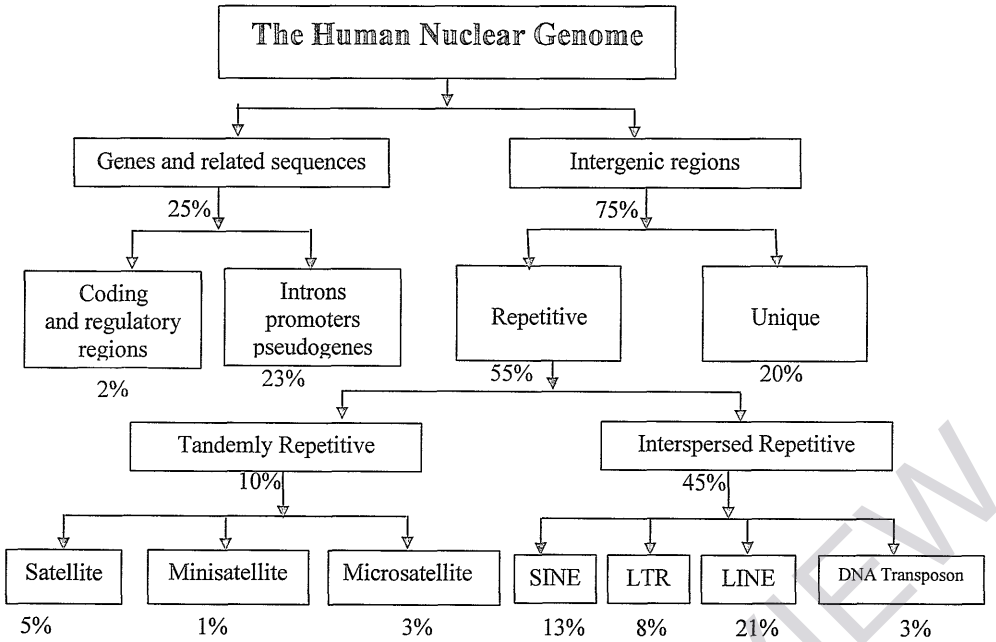

Figure 1-1: Human nuclear genome structure classified into classes of different structure and function. The percentage shares of various functional and non-functional sequences are shown. SINE are short interspersed nuclear elements, LINE are long interspersed nuclear elements and LTR are long terminal repeats (Jasinska and Krzyzosiak, 2004).

1.2.1 Tandem repeats

The field of forensic genetics has focused upon tandemly repeated sequences. These sequences come in all scales of both repeat unit length and array length. 'Satellites' are typically made up of repeat units of a few hundred bp to a few kilobases in length, in arrays that can extend for several megabases. They include the centromeric satellite arrays that play a vital function in chromosomal segregation. The classes that are extensively used in forensic practice are on a smaller scale: 'minisatellites' or variable number tandem repeats (VNTRs) and 'microsatellites', otherwise known as simple sequence repeats (SSR) or short tandem repeats (STRs). Each class has a short DNA sequence, repeated tandemly many times. The number of repeats determines the allele length which is the source of variation that is measured (length polymorphism). These kinds of loci are usually found between genes or within gene introns (Ellegren, 2004).

VNTRs were the first markers used in forensic casework, in the resolution of an immigration case (Jeffreys *et al.*, 1985a). These markers are located mostly in the sub-telomeric regions of chromosomes and consist of a core repeat unit ranging in size from 6 to 100bp and the total array length of a few hundred base pairs to over 20 kb (Nakamura *et al.*, 1987). Polymorphism at these loci is based on the allelic differences in the number of repeats which can result in as many as 2000 alleles at a single locus such as MS1 (Berg *et al.*, 2003). This is why they are extremely variable and have very high discriminating capacity (the most common five-locus VNTR pattern has a frequency no greater than about 10^{-6} in USA populations) (Risch and Devlin, 1992). However, there are also a number of limitations. These are the need for high molecular weight DNA which may be unavailable, unsuitability for typing a degraded sample, and time-consuming laborious techniques along with difficulties in the sizing of alleles and subsequent statistical interpretation (Jobling and Gill, 2004). These factors have led to the replacement of VNTRs by short tandem repeats (STRs).

1.2.2 Other forensic markers

There are a number of alternative markers that can be used in forensic identification and paternity testing. These include single nucleotide polymorphisms (SNPs) which are changes in a single base of DNA through base substitution or insertion/deletion events. On average, there are millions of SNPs differentiating any two copies of the human genome, so the opportunities for further exploitation are almost unlimited (Thorisson and Stein, 2003). They are widely used in the study of medical genetics and human evolution. SNPs usually have only two alleles and they are formed because of misincorporation of nucleotides during DNA replication or as a result of mutagenic effects caused by hazardous chemical or physical agents (Jobling *et al.*, 2004a). In the forensic context, the major advantage of these markers is their possible usage with highly degraded DNA as in disaster victim identification projects (Dixon *et al.*, 2005). However, having two alleles only for a SNP locus severely compromises its discriminating capacity. Approximately 50 SNPs have to be investigated to give a match probability equal to that obtained from 10 STRs (Gill, 2001). Therefore, it is difficult to incorporate SNPs in forensic casework with the current technology, since it would be financially demanding compared to the lower number of STRs that can be efficiently used instead (Gill *et al.*, 2004). However, several studies have been conducted to test various combinations of autosomal SNPs, the latest of which examined a panel of 52 SNPs that are polymorphic in major population groups. It yielded an extremely low match probabilities ranging from $5x\ 10^{-19}$ to $5x\ 10^{-21}$ (Sanchez *et al.*, 2006). SNP typing of the uniparentally inherited mitochondrial DNA (mtDNA) and Y-chromosome have potentially useful applications such as ethnic identification of major population groups through female and male lineages (Brion *et al.*, 2005; Wetton *et al.*, 2005). Some autosomal SNPs can be more reliable in inferring geographical ancestry, being highly polymorphic and escaping the sex bias of lineage-based SNPs (Frudakis *et al.*, 2003). Some of the latter groups of SNPs lie in the coding regions of pigmentation genes and those involved in metabolism of xenobiotics, and hence look promising in providing extra knowledge regarding the phenotype of a crime perpetrator such as eye and hair colour (Duffy *et al.*, 2007).

1.3 Autosomal STRs

STRs have been the most commonly used markers in forensic genetics since the mid-1990s. They have smaller repeat units (2-6bp) than VNTRs, and a total array length usually less than 300bp (Butler, 2005). This means that they are amenable to PCR amplification so can be used in the analysis of degraded samples containing less than 1ng of DNA (Pizzamiglio *et al.*, 2006; Ballantyne *et al.*, 2007). STR loci that have been selected for forensic use usually have 7-30 different alleles and a per-locus population heterozygosity of nearly 80% compared to 90% or more for VNTRs. The relatively small number of alleles and single repeat unit resolution leads to unambiguous results, but limits the amount of statistical information gained from a single locus (Butler, 2005, 2006).

STRs can be used in a wide variety of forensic cases such as missing persons investigations (Hagelberg *et al.*, 1991), mass disaster victim identification (Clayton *et al.,* 1995), paternity testing (Alford *et al.,* 1994) and linking suspects to crime scenes (Butler, 2006). There are thousands of polymorphic STRs scattered throughout the human genome, but those currently used in forensic practice and population databases are around 20 core markers commonly included in different multiplexes within commercial kits. An enormous amount of information describing all aspects of STRs used in human identity testing is incorporated in one of the most comprehensive widely used internet databases, the National Institute of Standards and Technology Short Tandem Repeat Internet Database (www.cstl.nist.gov/biotech/strbase/) (Ruitberg *et al.*, 2001).

1.3.1 Types of STRs and their forensic selection criteria

Typically, STRs comprise 2-6bp repeat units that are repeated tandemly many times. Their repeat sequence is named by the length of repeat units - for example, di-, tri-, tetra-, penta- or hexanucleotide repeats (Beckman and Weber, 1992). Tetranucleotide repeats have become more popular than di- or tri-nucleotides. This is mainly due to reduced stutter product formation during PCR amplification. Stutter products are PCR products which are one or more repeat units shorter in length than the true allele, and arise during PCR because of strand slippage (Edwards *et al.*, 1991).

STR sequences vary not only in repeat unit and number of repeats but also in repeat pattern. They have thus been classified into loci made up of 'simple repeats' of identical length and sequence (e.g. TPOX), 'compound' repeats that contain two or more adjacent simple repeats (e.g. vWA) and 'complex repeats' of multiple simple repeat blocks with some interspersion of invariable dinucleotide, trinucleotide and hexanucleotide sequences e.g. D21S11 (Urquhart *et al.*, 1994). A less commonly used type of sequence repeats is 'complex hypervariable repeats' which consists of tetranucleotide repeats with different mononucleotide, dinucleotide, trinucleotide and hexanucleotide invariants scattered throughout the locus. This leads to formation of numerous non-consensus alleles that differ in size and sequence and hence are incompatible with genotyping reproducibility and simple allele nomenclature e.g. SE33 (Gill *et al.*, 1994).

In the choice of STR markers in human identification, certain criteria are particularly important in selection of candidate STR loci (Carracedo and Lareu, 1998; Butler, 2006). These are the following:

- High degree of polymorphism (>70%) and high discriminating power with a wide allele frequency distribution. Loci such as TH01 and TPOX possess low number of alleles with frequencies <60% so, they are not highly polymorphic but can be included with other markers to be more discriminating.
- Distinct chromosomal locations so linkage equilibrium is achieved, which is favourable for using the product rule in match probability calculations. Two widely used loci (D5S818 and CSF1PO) are both located on chromosome 5, but are separated by several megabases, and thus are frequently separated by recombination.
- Low stutter characteristics, as in tetranucleotide repeats, which are particularly desirable when analysing mixtures (Urquhart *et al.*, 1994).
- Robustness and reproducibility when multiplexed with other markers.
- Low mutation rate i.e., variability within a locus has be stable enough to accurately pass the allele to the next generation. This is especially important in paternity analysis to avoid false exclusion of suspected fathers (Gill *et al.*, 1995).

- Small sized alleles (90-500bp), and compact allele range with discrete alleles, which are advantageous for degraded forensic DNA samples (Gill *et al.*, 1995).

1.3.2 Nomenclature of STR alleles

In an effort to facilitate the inter-laboratory reproducibility and comparison of data, a well established nomenclature has been developed for forensic use through the recommendations of the International Society for Forensic Genetics; ISFG (Bär *et al.*, 1997). They recommended that the choice of strand should be based on the coding strand for those STRs included within a genic region e.g. TPOX, TH01. STRs not associated with genes should be referred to as D#S### (e.g. D18S51) according to the Genebank entry, where D stands for DNA and S for single copy sequence. Regarding recommendations for choosing motif and allele designation, it is recommended to designate the repeat sequence motif (e.g. TCA) to be the first possible repeat nearest to the 5′ end. In addition, microvariant alleles containing partial repeats should be designated as the complete repeat first separated by a decimal point then the number of nucleotides in the partial repeat e.g. TH01 allele 9.3 (Puers *et al.*, 1993). Another important recommendation which is crucial for multiplex commercial kits is that a reference allelic ladder should be used as a reference for unknown sample allele designation; such ladders should contain sequenced alleles which are named according to the above recommendations, and contain all common alleles (Bär *et al.*, 1997).

1.3.3 Commonly used autosomal STRs in forensic practice

Since STR markers became an effective tool for human identity testing in the last decade, several attempts have been made to search for new convenient loci. Several multiplexes have been developed commercially containing various combinations of STR loci. The idea of multiplexing relies on the same criteria described above for the choice of individual forensic markers. Fortunately, PCR permits more than one region to be amplified simultaneously by simply adding more than one primer pair to the reaction mixture. Simultaneous amplification decreases the time, cost, and DNA consumption, and increases the number of loci that could be amplified in a single tube (Edwards and Gibbs, 1994).