A Molecular Study of Pakistani Populations Using Short Tandem Repeat Markers

A thesis submitted for the Degree of

Doctor of Philosophy

By

Syed Sibte Hadi

2001

Department of Forensic Medicine & Science

University of Glasgow

ACKNOWLEDGEMENT

In the name of Allah the most merciful and the most beneficient.

I would like to thank the department of Forensic Medicine and the division of Earth Sciences, Glasgow University for their co-operation and help during this work. My sincere thanks are due to my supervisor Professor Peter Vanezis for all his efforts to help me during my stay in the department. I would also like to thank Dr. William Goodwin for his interest in the project, for his efforts to obtain funding for the project and for his very thorough and helpful review of the thesis.

I sincerely thank Dr. A.M.T Linacre for allowing using the facilities in his laboratory. I also thank him for reviewing the thesis, providing enthusiasm and for a helpful critique at all stages.

My thanks are due to Mr. John Harley for his expert technical assistance and to the fellow students, Ahmed & Amani for their continuous support. I must thank all the secretaries, especially Mrs. Elizabeth Doherty in the department of Forensic Medicine for their advice and help.

I am grateful to Mr. Jeff Cockburn and Mr. Kenny Roberts for the excellent computer service provided at the department of Forensic Medicine and the division of Earth Sciences.

My thanks are due to Dr. Bruce Weir (University of North Carolina, USA) and Dr. James Curran & Dr. John Buckleton (University of Waikato, New Zealand) for helpful discussions on the much dreaded subject of forensic statistics.

I thank Dr. William Amos (Cambridge University) for helping me with a computer program for building the networks and useful discussion on STR biology.

I would like to thank Professor Angel Carracedo & Dr. Annabel Gonzalez Neira (Institute of Legal Medicine, University of Santiago de Compostela. Spain) for providing sequenced samples for the construction of allelic ladders.

I am indebted to Dr. Raymond Allchin & Dr. Bridget Allchin of Ancient India & Iran Trust, UK for providing me access to their precious library and their extreme hospitality during my visit to Cambridge.

I am grateful to Dr. Giorgia R Riboldi-Tunnicliffe and her staff Bron & Julie, at the Molecular Biology Support Unit at the Institute of Biomedical Sciences, University of Glasgow, for helping genotype the samples.

I am eternally grateful to Mr. David Martin (University of Glasgow) for lending an ear whenever I had a problem and for his support that kept me going.

Finally thanks to my beloved wife Ameena for everything she did to keep the home running. It is to her that I owe this work most. Thanks to my kids as well who braved everything including the weather.

CONTENTS

SECTION: TITLE Page

ABBREVIATIONS

AABB	American Association of Blood Banks
AMOVA	analysis of molecular variance
AmpFLPs	amplified fragment length polymorphisms
bp	base pair
DNA	deoxyribonucleic acid
dNTPs	deoxynucleotide triphosphates
EDNAP	Euorpean DNA Profiling Group
EtBr	ethidium bromide
FAM	fluorescein
FES/FPS	c-fes/fps Proto-oncogene
FGA	alpha fibrinogen gene
Gc	Group specific Component
GDA	genetic data analysis
HLA	human leucocyte antigen
Hp	Haptoglobin
HWE	Hardy Wienberg equilibrium
ISFG	International Society of Forensic Genetics
JOE	6-carboxy-2′, 7′- dimethoxy-4′, 5′-dichloroflorescein
LINES	long interspersed repeats
NRC	National Research Council
NJ	Neighbourhood joining
PCR	Polymerase Chain Reaction
PM	Probability of Match
PD	Probability of Discrimination
PE	Probability of Exclusion
PGM	Phosphoglucomutase
PI	Paternity Index
ROX	x-rhodamine
SD	standard deviation
SINES	short interspersed repeats
STRs	short tandem repeats
UPGMA	Unweighted pair group method using arithmetic averages
VNTR	Variable number of tandem repeats
vWA	von Willebrand gene
Yh1	Y chromosome haplotype 1

ABSTRACT

In order to implant DNA typing methods into the Pakistani medico legal system it was necessary to study the genetic structure of the Pakistani populations. For this purpose seven hundred and sixty fresh blood and buccal swab samples were collected from seven ethnic groups the Punjabi, Pushtoon, Sindhi, Baluchi, Makrani, Brosho & Kalash. DNA extraction from the samples was performed using the Puregene® kit and Chelex® extraction methods.

The populations were profiled for three autosomal STR loci D3S1358, vWA and FGA for which modified primers were designed and allelic ladders prepared. Some of the samples were profiled using the AmpFℓSTR® Blue kit. A database of the allele/genotype frequencies was established and the structure of the population was studied.

It was established that the allele and genotype frequencies differ significantly among the Pakistani populations. In the Punjabi and Sindhi populations the exact test p value was <0.05 at the locus FGA, in all other populations all individual loci and the combination of loci the exact test p value was >0.05. Substructure was detected in these populations and the F_{ST} values of 0.01-0.04 were determined between different populations in pairwise comparisons. Using a simulation study it was shown that incorporation of the substructure in the calculation of allele and genotype frequencies results in conservative estimates of the likelihood ratio of match. The analyses favoured the generation of separate databases for sub populations even if they are as closely related as are the Pakistani populations.

The second part of the project was to study the populations using the Y chromosome specific STR systems. For this purpose seven Y STRs (DYS19, 389I, 389II, 390, 391, 392 & 393) were selected which define the haplotype Yh1 (Y chromosome haplotype I). Sequenced allelic ladders were prepared for all the loci. The different ethnic groups of Pakistan were profiled for these loci in two multiplex reactions and 564 complete haplotypes generated. The individual locus and

haplotype frequencies were calculated for all the populations. Yh1 had a high diversity and discrimination capacity for the Pakistani populations when these populations were compared to other populations.

The diversity within and between populations was determined by AMOVA. It was established that diversity for Y STRs within the populations was greater than between populations. AMOVA also revealed that there was much less diversity between the three major populations of Pakistan pointing to closer male lineage relationship between these populations. Two smaller ethnic groups the Brosho and the Kalash generated higher diversity when compared to the major populations of Pakistan.

Phylogenetic trees (UPGMA & NJ) were constructed using the phylogenetic software PHYLIP. The phylogenetic analysis of the Pakistani populations suggested that the Punjabi and the Sindhi populations are closest to each other as are the Baluchi and the Makrani populations while the positions of the Kalash and Brosho populations were mostly distant from those of the other Pakistani populations.

CHAPTER 1: INTRODUCTION

A fundamental requirement in any criminal prosecution is that a crime be proved beyond reasonable doubt and in trying to meet this stringent requirement forensic practitioners draw upon a variety of scientific disciplines. Forensic science has therefore now entered an era in which increasingly sophisticated technology is not only desirable but also necessary in combating crime and ensuring justice. In the last two decades advances in molecular biology in particular, have created new means of resolving forensic questions and of identifying offenders with greater speed and certainty. Foremost among these is the use of analytical techniques in forensic serology.

1.1 FORENSIC GENETICS

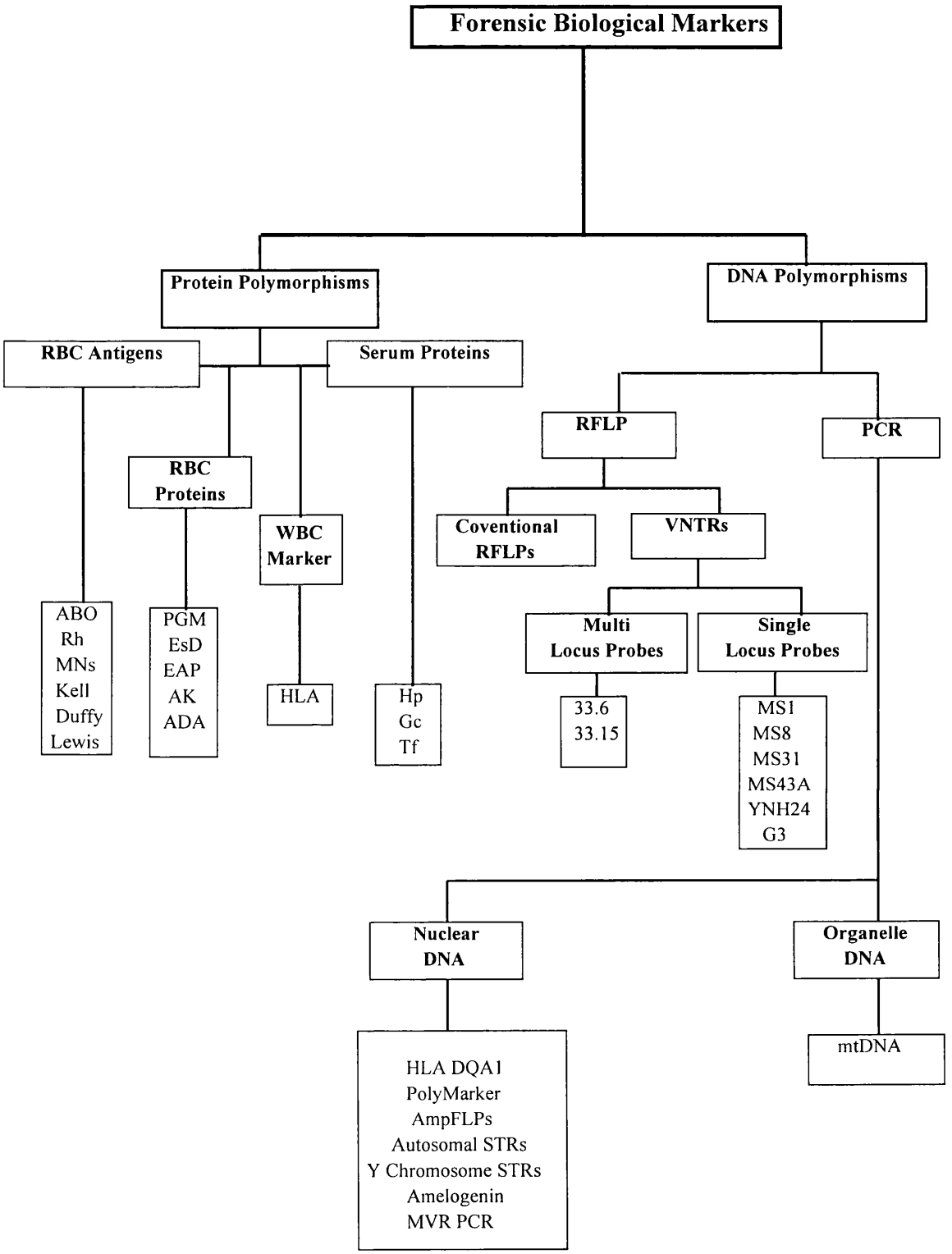
The foundations of forensic genetics were laid down when Karl Landsteiner described ABO blood group systems in 1901 (Reviewed in Wiener, A. S. 1972). Following this, the detection of other red cell antigens, serum proteins and erythrocyte enzymes made the serological analyses of blood and other body fluids possible (Reviewed in Gaensslen, R. E. *et al.*, I & II, 1985). By 1980 a battery of conventional blood grouping tests were available (Figure 1.1) which considerably improved the forensic utility specially when used in conjunction with, the white blood cell antigen system, HLA (Lincoln, P. J. 1992). However the techniques involved were laborious, technically complicated and utilised expensive sera. Further, the systems were not uniformly distributed in all the tissues and underwent rapid decomposition on drying (Denault, G. C. *et al.,* 1980).

Recent advances in deoxyribose nucleic acid (DNA) analysis presented an exciting forensic tool in forensic sciences for human identification and paternity analyses owing particularly to the high power of discrimination (Pena, S. D. J. *et al.*, 1995). Thus a number of new techniques based on DNA polymorphisms have been developed (Figure 1.1).

Figure 1.1: Serological and Genetic Polymorphisms Used for Forensic Purposes

Figure shows various systems used for forensic purposes. Due to high discriminatory powers DNA analysis has replaced the traditional serological methods used for human identification and paternity analysis in most laboratories.

RBC:	Red Blood Cell
Rh:	Rhesus Blood Group
PGM:	Phosphoglucomutase
EsD:	Esterase D
AK:	Adenylate Kinase
WBC:	White Blood Cell
HLA:	Human Leucocyte Antigen
Hp:	Haptoglobin
Gc:	Group Specific Component
Tf:	Transferrin
RFLP:	Restriction Fragment Length Polymorphism
VNTR:	Variable Number of Tandem Repeat
STR:	Short Tandem Repeat
PCR:	Polymerase Chain Reaction
mtDNA:	Mitochondrial Deoxyribose Nucleic Acid


Forensic Biological Markers
Protein Polymorphisms
DNA Polymorphisms
RBC Antigens
Serum Proteins
RFLP
PCR
RBC Proteins
WBC Marker
Coventional RFLPs
VNTRs
ABO
Rh
MNs
Kell
Duffy
Lewis
PGM
EsD
EAP
AK
ADA
HLA
Hp
Gc
Tf
Multi Locus Probes
Single Locus Probes
33.6
33.15
MS1
MS8
MS31
MS43A
YNH24
G3
Nuclear DNA
Organelle DNA
HLA DQA1
PolyMarker
AmpFLPs
Autosomal STRs
Y Chromosome STRs
Amelogenin
MVR PCR
mtDNA

1.2 DEOXYRIBOSE NUCLEIC ACID

The genetic makeup of every individual established at the time of conception is unique. It defines that individual's genetic characteristics and contains a large number of polymorphisms that can be used for human identification.

1.2.1 ORGANIZATION OF HUMAN GENOME

A small part (~2%) of the human genome consists of genes encoding for proteins (Stryer, L. 1988) while the rest appears to have little functional value (Southern, E. M. 1995) (Figure 1.2). It has been estimated that two haploid genomes chosen at random differ in one out of every 500 nucleotides which implied 6 million differences in the whole human genome (Pena, S. D. J. *et al.,* 1995) between any two individuals. Since analysing the whole genome is not routinely possible in order to differentiate between two persons, various regions of DNA were identified which were highly polymorphic.
About 25% of the extragenic DNA is organised as:

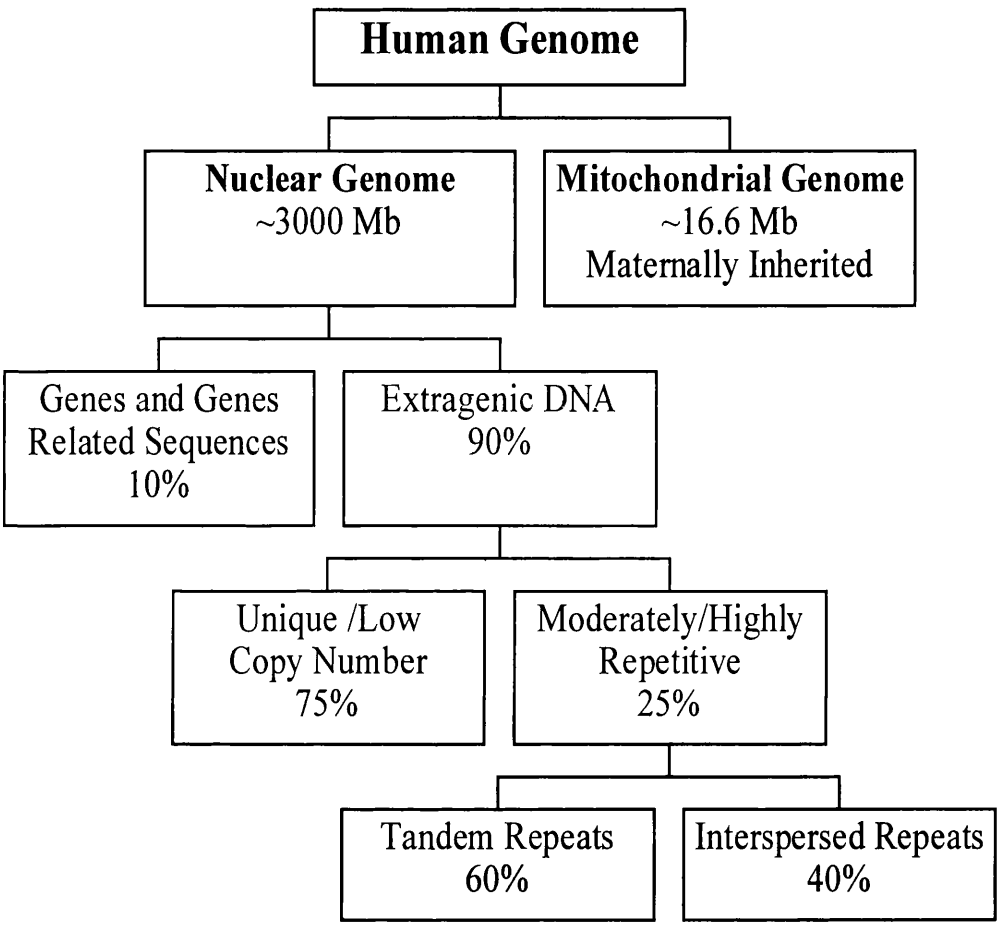
a. Interspersed Repetitive DNA

These sequences are identified as short or long repeats (SINES or LINES), both of which are scattered in the genome as units, and are either shorter than 500 bp and called SINES or longer when they are termed as LINES (Singer, M. F. 1982). SINES are more abundant and the best known are the Alu repeats which are 300 bp units and make ~5% of the genome (Batzer, M. A. *et al.*, 1990).

b. Tandemly Repeated DNA

The tandemly repeated DNA sequences are >1 basepairs (bp) in length. These have there been exploited for human identification as they exhibit a high degree of variability in the number of the core repeats (Jeffreys, A. J. *et al.*, 1985; Nakamura, Y. *et al.,* 1987; Tautz, D. 1989).

Figure 1.2: A Schematic Diagram Showing the Organisation of the Human Genome



1.2.2 TANDEMLY REPETITIVE DNA

In addition to the tandemly repeated structures in DNA satellite (I-IV, alphoid sequences), the tandemly repetitive DNA also contains minisatellites (Jeffreys, A. J. *et al.*, 1985) and microsatellites (Tautz, D. 1989).

1.2.2.1 Minisatellites

Certain minisatellite polymorphisms studied contained a large number of tandemly repeating units of a particular sequence (core repeat) and were termed 'variable number of tandem repeats' or VNTRs (Wyman, A. & White, R. 1980). Several other highly variable regions were subsequently discovered near the human insulin gene (Bell, G. I. *et al.,* 1982), *a*-related globin genes (Higgs, D.R. *et al.,* 1981), and the c-r-Ha-ras-1 oncogene (Capon, D. J. *et al.*, 1983). In each case the polymorphism resulted from differences in the number of core repeat.

In 1985 Prof. Sir Alec Jeffreys described a hypervariable minisatellite using multi locus probe technique (MLP) and suggested its utility for human identification (Jeffreys, A. J. *et al.*, 1985). The multi locus system of Jeffreys had a very high power of discrimination due to the number of alleles at a locus but MLP profiles were difficult to interpret and especially difficult to standardize. Therefore several single locus probe systems (SLPs) were developed which detect two alleles in a heterozygous individual; one from each parent (Kanter, E. *et al.*, 1986; Jeffreys, A. J. *et al.*, 1987; 1987; Wong, Z. *et al.*, 1987; Nakamura, Y. *et al.*, 1989), and this became the preferred method for DNA profiling.

1.2.2.2 AmpFLPs

In 1985 a revolutionary advance in molecular genetics came with the discovery of a simple and powerful technique, the Polymerase Chain Reaction (PCR) (Saiki, R. *et al.*, 1985). However the amplification of VNTRs was not consistent, due largely to the length of some alleles. Smaller minisatellite loci called AmpFLPs composed of short core repeat units like DIS80 and ApoB were described (Nakamura, Y. *et al.*, 1989; Boerwinkle, B. *et al.*, 1989; Charlesworth, B.

et al., 1994). Unlike VNTR loci, most of the AmpFLPs could be successfully amplified (Budowle, B. *et al.*, 1991 &1995; Cosso, S. & Reynolds, R. 1993).

1.2.2.3 Microsatellites

Microsatellites, also known as short tandem repeats (STRs), are an abundant class of DNA polymorphisms. Occurring every 300 to 500 kb in the human genome, these markers have a repeat unit of 1-6 bp in length and are highly polymorphic (Litt, M. & Lutty, J. A. 1989; Tautz, D. 1989; Polymeropoulos, M. H. *et al.*, 1991a & b; Tautz, D. 1993). These loci could be easily amplified using PCR as the size of the amplification product was 100-500bp (Edwards, A. *et al.*, 1991) and have become the most widely utilized markers for forensic DNA profiling.

1.3 SHORT TANDEM REPEAT DNA PROFILING

Early studies showed that the human STR loci selected for forensic use had no non specific reactions from micro organisms or various substrates and offered unprecedented advantages in the context of human identification (Crouse, C. & Schumn, J. W. 1995; Gill, P. 1997). It was shown that the STR loci could be sized accurately using fluorescent technology as well as by non fluorescent methods (Kimpton, C. P. *et al.*, 1993; Urqhart, A. *et al.*, 1994; Huang, N. E. *et al.*, 1995).

1.3.1 STR MUTATIONAL MECHANISMS

The forensic value of STRs is due to their high levels of polymorphisms. These polymorphisms are in part due to the process of mutation occuring in these genomic sequences which generates different alleles at a locus.

In most cases the mutations arising during DNA replication are corrected by the repair enzymatic processes. However, sometimes, 'slipped strand mispairing' leads to mutation in regions of the genome, having abundant simple repetitive sequences which has been recognised as the major mechanism involved in their generation (Tautz, D. & Renz, M. 1984; Tautz, D. *et al.*, 1986; Levinson,

G. & Gutman, G. 1987). During this process the gain or loss of a repeat region occurs depending on the looping out of the newly synthesised strand or the template strand (Bennet, P. 2000).

There is a bias towards gain type mutations (Amos, W. A, 1996; Brinkmann, B. *et al.*, 1998; Cooper, G. *et al.*, 1999) however the expansion does not continue unabated and it has been suggested that loss of multiple units might provide a mechanism to control expansion (Weber, J. L. & Wong, C. 1993; Wierdl, M. *et al.*, 1997). The number, location and sequence of the repeats has been described to affect the mutation rate (Debrauwere, H. *et al.*, 1997; Schlötterer, C. 1998). It has been shown that long STR alleles evolve faster while smaller length and interspersed irregular repeats within the alleles inhibits mutation (Brinkmann, B. *et al.*, 1998).

1.3.2 THEORIES OF MUTATIONAL MECHANICS

Two theories were forwarded to explain the mutational mechanisms of the STRs were the Infinite Allele Model (IAM) (Kimura, M & Ohta, T. 1978) and the Stepwise Mutational Model (SMM) (Kimura, M. 1993). IAM assumes the production of new alleles as unique events as there are an infinite number of states to which an existing allele could mutate, whereas under the SMM only two states are allowed, a step forwards or a step backwards resulting in gain or loss of a repeat unit (Di Rienzo, A. *et al.*, 1994). The levels of heterozygosity observed for the STR loci and the replication slippage phenomenon both favour the SMM model (Valdes, A. M. *et al.*, 1993; Edwards, A. *et al.*, 1992; Shriver, M. D. *et al.*, 1993).

Though, SMM is the favoured theory at present, all the mutational events at STR loci cannot be explained under a strict SMM, as more rare mutations of larger number of gains or losses may occur (Dauber, E. M *et al.*, 2000), therefore to explain such phenomenon an extended SMM theory was also forwarded (Di Rienzo, A. *et al.*, 1994). This two phase model allowed for single and large multistep mutations and was ascertained to fit the population data of most of the STR loci studied (Di Rienzo, A. *et al.*, 1994).

1.3.3 RATES OF MUTATION OF STRs

The actual mutation rates for the STRs are difficult to determine as the mutations are rare events and generally there is no consensus regarding them. However pedigree studies have determined the mutational rates to be in the range of $1.2x10^{-4}$ to $1.5x10^{-2}$ (Weber, J. L. & Wong, C. 1993; Bowcock, A. *et al.*, 1993). The mutation rates differ significantly between di, tri and tetra nucleotide STRs. The initial observation of higher mutational rates of the tetranucleotides than the tri or dinucleotides (Weber, J. L. & Wong, C. 1993) has recently been contested and analyses of various data has shown that the rates of mutation of STRs have an inverse relationship with the size of the repeat (Chakraborty, R. *et al.*, 1997). For tetranucleotide STRs these have been determined to be $1–1.9^{-3}$ (Dauber, E. M. *et al.*, 2000). Slightly higher mutation rates were reported in an analysis of paternity testing data (AABB Report 1999).

1.3.4 STR RESOLUTION TECHNIQUES

Electrophoresis has remained the cornerstone of STR allele detection, although the STR alleles can be separated by agarose gel electrophoresis with ethidium bromide staining, polyacrylamide gel electrophoresis (PAGE) gives better separation of the alleles (Sambrook, J. *et al.*, 1989).

1.3.4.1 STR Analysis and Automation

Instrumentation was developed for the automated analysis of DNA sequence using fluorescent dyes in 1986 (Smith, R. N. *et al.*, 1986). The automated systems such as ABI GeneScan® allowed electrophoretic information to be stored and tabulated as the alleles migrated through a gel matrix and pass a laser detection window (Ziegle, J. S. *et al.*, 1992). The availability of 4 different fluorescent dyes enabled the primers for different loci to be tagged with a distinct fluorescent dye that allowed multiplexing loci even if the size of the products overlapped and also enabled additional levels of controls, including internal size standards.

Automatic allele sizing was thus made possible by running an internal size standard with each sample (Kimpton, C. P. *et al.*, 1993). Thus electrophoretic mobility variations (from lane to lane and gel to gel), that could lead to inconsistent allele sizing were automatically normalised (Fregeau, C. J. & Fourney, R. M. 1993). These initial studies confirmed the sensitivity of the automated analysis methods over the manual staining methods and also demonstrated that the automated analysis had the requisite reproducibility, accuracy and precision for use in a forensic setting.

The casework validation studies showed that mixed samples could be detected and correctly interpreted, STR analysis yielded results where SLP analysis had failed (Lygo, J. E. *et al.*, 1994). Also the discrimination power of a multiplex STR system was much greater than the systems based on HLA DQA1 and conventional blood grouping (Lygo, J. E. *et al.*, 1994).

1.3.4.2 Use of Allelic Ladders

An allelic ladder consists of a number of alleles of a STR system and is used as a reference to designate the alleles. The use of allelic ladders was shown to size the alleles accurately and it has been a consistent recommendation that sequenced allelic ladders be used in order to designate the alleles (Puers, C. *et al.*, 1993; Olaisen, B. *et al.*, 1998; Griffths, R. A. L. *et al.*, 1998). For this purpose experiments set the windows for the allele size and these windows can be fed in computer programs like Genotyper™, the program would call the allele on the basis of the established windows, which is called the 'absolute window method' (Gill, P. *et al.*, 1996).

However, it has been shown that another method called the 'floating window method' might be better for the designation of alleles as it allows for the band shift and the correlation makes it possible to designate very rare alleles and variants which are 1-2 bp apart (Gill, P. *et al.*, 1996). The same method could be used to designate the alleles where the crime scene sample is to be compared to control samples.

a) Non Specific Amplification Peaks

The sensitive automated DNA sequencer detects all products generated during a PCR assay and displays them as peaks on the electropherogram, these can include non specific peaks and stutter peaks besides the allelic peaks.

The non specific peaks are often less than 5% in size of the allelic peak or have an atypical morphology which makes their recognition easy (Gill, P. *et al.*, 1997). The stutters however arise from enzyme slippage during the reaction, they cannot be eliminated completely from the reaction but can be recognised as a smaller (about 15%) peak than the actual allelic peak and are one repeat smaller than the actual peak (Gill, P. *et al.*, 1997). In a way these can be beneficial in the recognition of the actual peak, however in di nucleotide STR analysis, stutters pose difficulties in interpretation. Also in mixture analysis stutters might enhance the peaks of the minor component thus causing problems in the identification of the minor peak (Clayton, T. M. *et al.*, 1998).

1.3.5 ENVIRONMENTAL CONTAMINATION

The sensitivity PCR renders the microsatellite systems prone to contamination. Possible sources of contamination include, environment, contamination from a previous PCR and between samples during preparation. The latter two sources of contamination could be controlled by appropriate laboratory procedures and designated working areas. The environmental contamination has been shown to be limited (Lygo, J. E. *et al.*, 1994).

1.3.6 ACCEPTANCE OF DNA EVIDENCE AT THE LEGAL FORUM

In 1985 the forensic use of DNA began in the UK and it was applied to a civil immigration case and then to a criminal case in the UK which established the ground for the application of DNA analysis to forensic cases (Jeffreys, A. J. *et al.*, 1985; Gill, P. & Werret, D. J. 1987). The forensic applications of the technique were investigated and it was found that the technique was suitable for forensic purposes (Gill, P. *et al.*, 1985). By 1990 the systems for the DNA profiling were

standardised and most laboratories on both the sides of the Atlantic were using them. The methodology was termed as sound in the first report of National Academy of Science's National Research Council (NRC Report I, 1992).

In the UK which pioneered the DNA technology, acceptance of the technique by the legal forum was initially smooth. However challenges relating to the statistical methods used occurred in the UK in cases like R. *v* Andrew Deen (Times Law Report 1994), R. *v* Denis Adams (Cr. App. R. 1996) and R. *v* Alan James Doheny & Gary Adams (Cr. App. R. 1997). These and other cases set the guidelines for the presentation and acceptance of the DNA evidence by the courts in the UK (Lincoln, P. J. 1997).

As the DNA evidence became acceptable in the UK the FSS quickly moved to PCR based technology and developed a quadraplex system consisting of four STR loci (Figure 1.3). The Home Office, UK then commissioned the FSS for the development of a National DNA Database and using a more discriminating Second Generation Multiplex system, the database was started to be developed in 1995 (Werret, D. J. 1997; Figure 1.3). Recently the FSS has started profiling criminal cases using the 11 locus AMPFℓSTR® SGM Plus™ kit (Applied Biosystems CA, USA).

The US Congress passed the DNA Identification Act of 1994, which helped in the acceptance of the DNA evidence. It also provided the necessary powers to the FBI Director in order to improve the standard of DNA analysis by ensuring quality assurance, appointment of a DNA advisory board and to construct a national DNA databank (Shapiro, E. D. & Reifler, S. 1996). Consequently the CODIS database for STRs has been established in the USA which comprises of 13 loci including the gender marker.

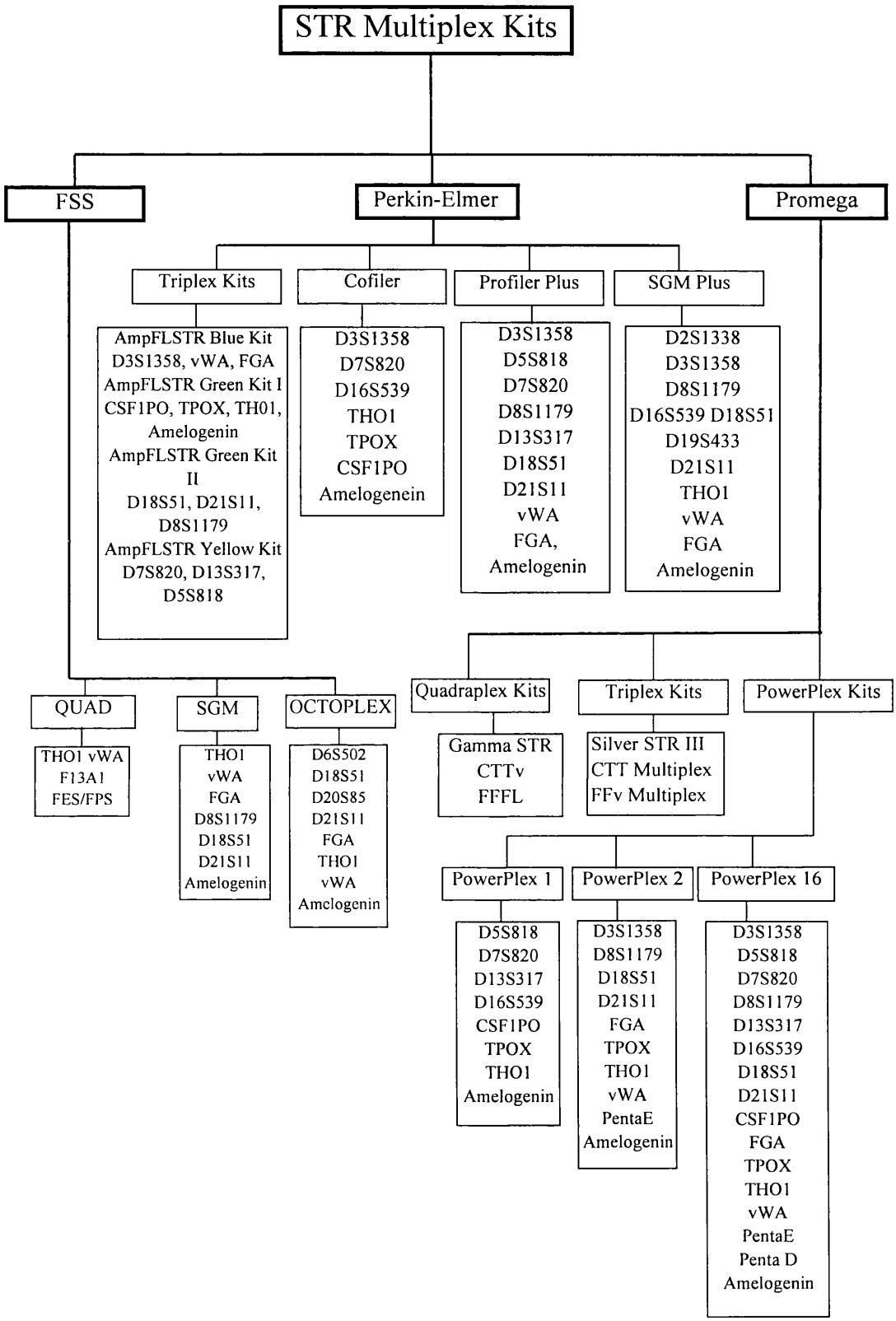
Figure 1.3: STR Multiplex Amplification Systems for Forensic Analysis

Figure showing the multiplex STR systems developed by the Forensic Science Services, UK, Perkin-Elmer CA, USA and Promega Corp, Madison, USA.

The first quadraplex system (QUAD) comprising of four short tandem repeat loci HUM VWA31/A, HUMTHO1, HUMF13A1 and HUMFES/FPS was developed by the Forensic Science Services (FSS), UK (Kimpton, C. P. *et al.*, 1994).

Later a six locus STR multiplex system along with amelogenin locus was introduced by the FSS called the second generation multiplex (SGM), for routine case work (Urquhart, A. *et al.*, 1995; Oldroyd, N. J. *et al.*, 1995).

Various commercial multiplex systems have been marketed by the two leading firms Perkin-Elmer, CA, USA and Promega Corp, Madison, USA, which have successfully multiplexed an increasing number of loci thereby enhancing the discriminatory power of the analysis.


STR Multiplex Kits
FSS
Perkin-Elmer
Promega
Triplex Kits
Cofiler
Profiler Plus
SGM Plus
AmpFLSTR Blue Kit
D3S1358, vWA, FGA
AmpFLSTR Green Kit I
CSF1PO, TPOX, TH01,
Amelogenin
AmpFLSTR Green Kit
II
D18S51, D21S11,
D8S1179
AmpFLSTR Yellow Kit
D7S820, D13S317,
D5S818
D3S1358
D7S820
D16S539
THO1
TPOX
CSF1PO
Amelogenein
D3S1358
D5S818
D7S820
D8S1179
D13S317
D18S51
D21S11
vWA
FGA,
Amelogenin
D2S1338
D3S1358
D8S1179
D16S539 D18S51
D19S433
D21S11
THO1
vWA
FGA
Amelogenin
QUAD
SGM
OCTOPLEX
Quadraplex Kits
Triplex Kits
PowerPlex Kits
THO1 vWA
F13A1
FES/FPS
THO1
vWA
FGA
D8S1179
D18S51
D21S11
Amelogenin
D6S502
D18S51
D20S85
D21S11
FGA
THO1
vWA
Amelogenin
Gamma STR
CTTv
FFFL
Silver STR III
CTT Multiplex
FFv Multiplex
PowerPlex 1
PowerPlex 2
PowerPlex 16
D5S818
D7S820
D13S317
D16S539
CSF1PO
TPOX
THO1
Amelogenin
D3S1358
D8S1179
D18S51
D21S11
FGA
TPOX
THO1
vWA
PentaE
Amelogenin
D3S1358
D5S818
D7S820
D8S1179
D13S317
D16S539
D18S51
D21S11
CSF1PO
FGA
TPOX
THO1
vWA
PentaE
Penta D
Amelogenin

1.3.7 STANDARDISATION OF STR NOMECLATURE AND TECHNIQUE

As the use of STR grew for forensic purposes there was a need to standardize each aspect of this new technology so that the laboratories conforming to the standardized methods in different countries/regions could exchange and use the data for legal and population genetic purposes.

Firstly there was a need for a common nomenclature for the widespread use of STR systems, guidelines were laid down by the International Society of Forensic Genetics (ISFG) and by one of its working groups, the European DNA Profiling Group (EDNAP) (Bar, W. *et al.*, 1997; Gill, P. *et al.*, 1997; Olaisen, B. *et al.*, 1998 a; Olaisen, B. *et al.,* 1998 b).

The basic recommendations have been that the alleles would be designated according to the number of the repeats, and if any incomplete motifs or complex repeats occur they would be expressed by the number of complete repeats followed by a decimal point and then the number of bases of the incomplete motif. Various STR classes were also described such as simple, simple with non consensus alleles, compound, complex and complex hypervariable systems according to the arrangement of the repeat region in the alleles of a particular system (Gill, P. *et al.*, 1997).

Apart from this the resolution of STRs using the flourescent technology, use of sequenced allelic ladders for the designation of alleles and use of standard multiplex sets of STRs set (1.3.4) all were a part of the efforts of the forensic community to standardize the STR technology. This has naturally lead to a boost in the confidence of the scientific and the legal communities in this technology. The International Society of Forensic Genetics as well as the European Network of Forensic Science Institutes have been performing a pivotal role in these standardisation efforts through regular interlaboratory excercises and publication of its recommendations (Martin, P.D. *et al.*, 2001).

1.3.8 POPULATION GENETICS

In 1989 some leading scientists started a debate on the statistical basis of calculating the population frequencies who argued that the use of a general racial database is incorrect and databases from the relevant ethnic population should be used in all cases (Lander, E. 1989; Cohen, J. E. 1990; Lewontin, R. C. & Hartl, D. L. 1991). It was also stated that the frequencies of the genotypes could be multiplied to generate a profile frequency if the loci only if they were independent. As a result challenges to the DNA evidence arose though DNA technology and its statistical basis stood a thorough scrutiny by the legal forum (Slyvester, J. T. 1992; Chakraborty, R & Kidd, K. K 1991; Brookfield, J. 1992; Morton, N. E. 1992; Budolwe, B. & Lander, E. 1994; Shapiro, E. D. & Reifler, S. 1996).

The National Academy of Sciences, USA in its second report (Report II, 1996a), further addressed the population genetic issues and recommended methods to calculate the population substructure and the likelihood of coincidental DNA match and laid down standards for the collection and preservation and presentation of DNA evidence.

1.3.8.1 Hardy Weinberg Principle

The Hardy-Weinburg principle (HW) states that the frequencies of alleles in a population will remain constant unless acted upon by outside agents or forces like migration, selection, genetic drift or mutation (Strickberger, M, W. 1985).

If it was shown that a population exhibits gene frequencies in accordance with the HW principle, then the calculations could be made accordingly. For forensic purposes, if the multiple loci were shown to be independent within and between the loci (Weir, B. S. 1990) then their frequencies could be multiplied together according to the 'product rule' (Chakraborty, R. *et al.,* 1993). This way the frequency of a multilocus genotype was obtained. Thus it appeared important to perform statistical tests which would show that the DNA marker in the population being studied behaves in accordance with the HW principle or otherwise.

1.3.8.2 The Exact Test

In order to study the diffrences between the observed genotype frequencies in comparison to those expected under HW principle originally Chi square test of significance was used. However this test could not be applied to the markers with a large number of alleles as unobserved genotypes or genotypes at very low frequencies were common with such systems (Evet, I. W. & Weir, B. S. 1998).

The Fisher's exact test was designed to apply the contingency table method to calculate the exact probability of getting any particular set of values by chance. This is the test currently applied most often to test if mating in a population was random. For genetic loci, the improved form of the test was described later for multiple alleles (Guo, S. W. & Thompson, E. A. 1992). In order to apply this test the starting assumption is that no significant difference exists between the observed and the expected values. If the probability turns out to be small (<0.05) significant difference is accepted, otherwise the original assumption of 'no significant difference' is held to be true.

1.3.8.3 Population Structure and F Statistics

The alleles at a locus show relationship to each other in terms of population and the respective sub populations and the quantification of this relationship is given by the F statistics F_{IT}, F_{IS} and F_{ST} (Wright, S. 1965; Weir, B. S. 1990).

The measure F_{ST} is of particular interest as it estimates the co ancestry of the alleles (Balding, D. J. & Nichols, R. A. 1994). In forensic estimates of the frequency of a profile one must take into account any kinship between the perpetrator and any other unrelated individual from the population in question. Thus it is pertinent to use a measure which can help in estimation of the co ancestry when estimating DNA profile frequency.

1.3.8.4 Effect of Substructure on Genotype Frequency Estimation

If a population shows substructure then independence of the alleles cannot be assumed and the population cannot be assumed to be mating according to the HW principle (Evett, I. W. & Weir, B. S. 1998). Thus the HW equation cannot be used to estimate the genotype frequencies. It was therefore necessary that the effects of substructure were incorporated so that the estimates were near to the true value in a population. This would allow the use of product rule for calculating the multi locus profile frequency.

It was proposed that for the allele and genotype frequencies should be calculated allowing population substructure to be incorporated into profile estimations before the estimation the multilocus profile frequency and suggested that a value of 5% might be used (Balding, N. J. & Nichols, R. A. 1994). However other workers have shown that most populations do not exhibit this level of inbreeding (Brookfield, J. 1992; Morton, N. E. 1992). In view of its forensic importance, the NRC has also recommended the use of corrections proposed by Balding & Nichols (NRC Report II, 1996a).

1.4 Y-CHROMOSOME POLYMORPHISMS

Most of the Y chromosome does not undergo recombination during meiosis thus the paternally inherited chromosome bears the genetic prints from father to the son along the whole paternal lineage (Jobling, M. A. & Smith, C. T. 1995). In all human societies the majority of violent crimes are committed by the males, thus the characteristics which could identify and/or exclude a male, have been of acute forensic interest. The use of male specific Y chromosome markers in forensic practice has been reviewed and it was held that these markers could be used as an adjunct to other markers on the autosomes and could prove to be a new and useful forensic tool on their own as well (Jobling, M. A. *et al.,* 1997).

1.4.1 INITIAL STUDIES

Initially the human Y chromosome was considered almost devoid of polymorphic markers in comparison to their autosomal counterparts, a view which resulted from systematic searches for restriction fragment length polymorphisms which met with little or no success (Leroy, P. *et al.,* 1987; Simmler M, C. *et al.,* 1987; Jacubicza, P. *et al.*, 1989; Malaspina, P. *et al.*, 1990). Sequencing of the non coding regions showed that these were monomorphic (Seielstad, M. T. *et al.,* 1994; Dorit, R. L. *et al.,* 1996). In view of the scarcity of RFLPs it was thought that the gaps in the physical map of Y chromosome might be filled by simple tandemly repeating sequences.

1.4.2 Y CHROMOSOME MINISATELLITES (MSY1 & 2)

The polymorphic minisatellite MSY1 was the first Y mini satellite to be described (Jobling, M. A. *et al.*, 1998). Minisatellite variant repeat PCR (MVR-PCR) system determines the Y-specific DNA codes, with heterozygosity of approximately 99.9%. Another minisatellite has been described recently which has shown to contain only two units, either 3 or 4 copies of 99-110 bp repeating units (Bao, W. *et al.,* 2000). It has been shown to be highly polymorphic in the Chinese population. These markers are hypervariable but due to the simplicity of the use and availability of several Y chromosome STR systems have not gained widespread use for forensic purposes. Also proper evaluation of the value of evidence was difficult with the use of minisatellites. In addition the mutation rates are high.

1.4.3 DEVELOPMENT OF Y CHROMOSOME SPECIFIC STR SYSTEMS

It was shown that 60% of the Y chromosome region on the long arm consists of interspersed tandem repeats (Cooke, H. J. 1976; Kunkel, L. M. *et al.*, 1979; Cooke, H. J. *et al.*, 1982). 27H39LR was described as the first STR on the human Y chromosome, which was found to be polymorphic and suitable for sex and paternity determination in deficiency cases (Roewer, L. *et al.,* 1992).

Soon after its description the locus DYS19 was applied to a forensic case (Roewer, L & Epplen, J. T. 1992). Later the same locus was studied again for

its forensic potential and was found very useful having 9 alleles and a high power of discrimination (Santos, F. R. *et al.,* 1993a; Muller, S. *et al.*, 1994; Trabetti, E. *et al.,* 1994; Santos, F. R. *et al.*, 1996). The locus was similar to the autosomal STRs in its allele distribution and the repeat structure.

A pentanucleotide was also described but the locus exists at the homologous X chromosome though it was suggested that the larger alleles may be present only in the males (Chen, H. *et al.,* 1994).

Three Y specific dinucleotide STRs designated as YCAI, YCAII and YCAIII, were described, (Mathias, N. *et al.,* 1994). These were moderately polymorphic and were found to be Y specific. In a major study, an additional 9 Y chromosome specific STRs were described (Kayser, M. *et al.,* 1997). Their study was supported by the population data of approximately four thousand males. Most of these loci were found to be polymorphic and the haplotypes (profile of an individual at two or more loci) defined by using all of them could achieve very high levels of individualisation of males.

In the same study two mutiplexes were described which could together amplify 7 Y STRs and the terminology to be used for the Y STR haplotypes (a panel of STRs) was also described. Thus the haplotypes were designated as Yh1 to Yh5 depending on the markers used to obtain the haplotype. They demonstrated that the power of discrimination of a haplotype consisting of a set of seven such STRs, DYS19, 389I, 389II, 390, 391, 392 & 393 (Yh1) was 74-90% in European populations and a ten marker haplotype (Yh4) could be used for human identification as well. During this study the loci DYS288, DYS388, YCAI and DXYS156 were found to have low gene diversity therefore it was recommended that these might not be used for routine forensic practice.

Most of these markers have been applied for building up population databases across the world in order to apply them to forensic casework (Kloosterman, A. D. *et al* 1998; Pestoni, C. *et al.,* 1998; Rosi, E. *et al.,* 1999; Horst, B. *et al.,* 1999; Tun, Z. *et al.,* 1999; Tagliabracci, A. *et al.,* 1999; Pawlowski, R. *et*

al., 1999; Furedi, S. *et al.,* 1999; Sasaki, M. and Dahiya, R., 2000; Gehrig, C. *et al.,* 2000 & Gonzalez-Neira, A. *et al.*, 2000).

According to an estimate the 3.5 Mb of the euchromatin of the Y chromosome should contain about 170 STRs (Jobling, M. A. *et al.*, 1997). To date 1.5 Mb has already been sequenced and this is a rich source, which can be used to search the STRs, and has been used for this purpose successfully for this purpose (Ayub, Q. *et al.,* 2000). Various new Y STRs have been described during the last two years in two studies (White, P. S. *et al.,* 1999; Ayub, Q. *et al.,* 2000) but only only two out of these (A7-1 & A7-2 described by White, P. S. *et al.*, 1999) have been validated and included in the list of accredited markers at the Leiden University, Netherlands (Table 1.1).

1.4.4 FORENSIC APPLICATIONS OF Y STRS

Y STR analysis can be useful for forensic identification in many situations where the autosomal STRs are of limited value. They can be valuable in mixture interpretation in multiple rape cases or the detection of male specific profile in azoospermic/vasectomized male suspects when spermatozoa are not available.

The Y STRs could determine the male component in the male/female mixture specimens in cases where the specimen was small or the differential lysis failed or in other body fluid mixtures where the differential lysis could not be attempted (Kayser, M. *et al.,* 1997; Jobling, M. A. *et al.,* 1997).

The value of Y chromosome STR polymorphisms has been determined as excellent for paternity testing in deficiency cases (Kayser, M. *et al.,* 1997). The Y linked loci have been proved to be better exclusionary tools in paternity cases than the autosomal markers though their value in positive identification might not have an edge over the autosomal loci (Chakraborty, R. 1985). In deficiency cases where the father was not available other paternal relatives could be tested with Y STRs, though the same valuable property is reciprocally disadvantageous in situations where suspects belong to same paternal lineage.

Table 1.1: Y Chromosome STR Markers Validated for Forensic Purposes

Locus	Repeat Motif	Allele Size Range	Alleles	Ref
DYS19	$(TAGA)_n$	174-210	9	Kayser *et al.*, 1997
DYS389I	$(TCTG)_n$	235-263	7	"
	$(TCTA)_n$			
DYS389II	$(TCTG)_n$	255-383	8	"
	$(TCTA)_n$			
DYS390	$(TCTA)_n$	191-227	9	"
	$(TCTG)_n$			
DYS391	$(GATA)_n$	271-299 *	8	"
DYS392	$(TAT)_n$	233-263	11	"
DYS393	$(AGAT)_n$	108-136	8	"
DYS385	$(GAAA)_n$	360-414	15 ◦	"
DYS388	$(ATA)_n$	126-138	5	
DXYS156Y	$(TAAAA)_n$	160-170	3	
YCAII	$(CA)_n$	144-160	31	Mathias *et al.*,1994
YCAIII	$(CA)_n$	192-204	7	"
A7 1	$(ATAG)_n$	161-181	6	White *et al.*, 1999
A7 2	$(TAGA)_n$	174-190	5	"

* Alternative primers for this locus result in a smaller product range of 139-159 bp (Gusmão. L. *et al.*, 2000)

.

1.4.5 Y STR MULTIPLEXING STRATEGIES

Multiplex amplification of several loci allows simultaneous amplification of many STR systems which conserves the usually small samples submitted for forensic analyses. Initially three multiplex systems were described which allowed multiplexing 3-4 loci (Prinz, M. *et al.,* 1997; Kayser, M. *et al.,* 1997; Redd, A. J. *et. al.*, 1997). The multiplex systems described by Prinz, M. *et al.,* and Kayser, M. *et al.*, are mostly used as both are robust and reproducible and together, allow seven Y STRs to be analysed. These systems are therefore favoured at the moment.

The multiplex system described by Thomas, M. *et al.*, 1999, allows multiplexing 6 Y STR loci, including DYS388 which is usually not favoured for forensic use owing to its low diversity and since many of the primers have been redesigned they might have to be validated before they are put to forensic use. The multiplex described by Gonzalez-Neira, A. *et al.*, allows analysing five loci, all of which are recommended for forensic use (Gonzalez-Neira, A. *et al.*, 2000; Kayser, M. *et al.*, 1997). A new multiplex capable of amplifying seven Y STR loci, the 'Y-PLEX™ 6' has been recently developed by a commercial firm Reliagene, Technologies, Inc. USA (Warren, J. E. *et al.*, 2000).

1.4.6 Y STR MUTATIONAL MECHANISMS AND RATES

In order to apply the Y STRs to forensic casework it is necessary that the mutation rates of the Y STRs and their mechanism should be known as precisely as possible. In cases where a profile at Y specific STRs is to be analysed and reported, a difference at one locus confined to single repeat may not be problematic (Jobling, M. A. *et al.,* 1997) however multiple mutations or the detection of multiple alleles may be misinterpreted and as a result incorrect conclusions may follow (Kayser, M. *et al.*, 2000; Kayser, M. & Sajntila, A. 2001). The studies on the mechanism of Y STR mutation have pointed out that the number and organization of repeats of a STR may have a significant role in the mutational events in Y STR loci. Longer repeat structure seems to favour mutation thus the two Y STRs DYS390 and 389 exhibit high rates of mutation. Similarly at the locus

DYS19 the repeat motif $(TAGA)_n$ is interrupted by a block of TAGG sequence which may have a role in mutation. The Y STR mutations are thought to favour growth in size (Cooper, G. *et al.*, 1999).

The mechanism of mutation at Y chromosome STRs relates to that in the autosomal STRs, and it has been suggested that polymerase slippage during replication might be the major source of mutation at these loci as in autosomal STRs (Forster, J. R. *et al.,* 1997; Carvalho-Silva, D. R. *et al.,* 1999; Kayser, M. *et al.,* 2000).

In the first study on the locus DYS19, 626 father/son pairs were studied and the mutation rate was determined to be 3.2×10^{-2} (Kayser, M. *et al.,* 1997). A rate of 3.09×10^{-3} was reported for seven Y STR loci after studying 40 Caucasian families (Bianchi, N. O. *et al.,* 1998).

A further attempt to estimate the mutation rates of Y STR was done in 1997 when 42 males who were known to have descended from 12 founding fathers were studied at nine Y chromosome loci (8 Y STRs and DXYS156Y) (Heyer. E. *et al.,* 1997). Loci DYS19, 391 and 389 had the highest mutation rates among these loci. The average mutation rate of all the tetranucleotide STRs was estimated as 2.1×10^{-3}, with a 95% confidence limit of $0.6–4.9 \times 10^{-3}$. Since all the pedigrees which showed more than one mutational event were excluded the estimates were described as an underestimate at that time (Heyer, E. *et al.,* 1997). The 42 male samples from this study were typed for the Y minisatellite MSY 1 and the mutational events at the Y STRs in this sample were compared to those at MSY 1 confirming that the mutational rate of 2.1×10^{-3} was conservative (Jobling, M. A. *et al.,* 1999).

Recently in a comprehensive study 5,000 father/son pairs were studied for mutations at thirteen Y STRs. The average tetranucleotide STR rates for the eight loci were estimated as 3.2×10^{-3} with a 95% confidence interval of $1.89–4.94 \times 10^{-3}$ (Kayser, M. *et al.,* 2000). These rates of mutation also matched the mutational rates of the autosomal STRs (Weber, J. L. & Wong, C. 1993).

1.4.7 PHYLOGENETIC VALUE OF Y CHROMOSOME SPECIFIC STRs

Due to the mode of inheritance Y Chromosome markers have been utilised as a tool for studying phylogenetic relationships and haplotypes based on Y markers have shown greater differentiation than autosomal or mtDNA markers (Seielstad, M. T. *et al.*, 1998).

In order to study the phylogenetic relationship of the populations for evolutionary purposes slowly mutating markers have been used as they retain ancient states, thus Y chromosome specific biallelic markers and single nucleotide polymorphisms (SNPs) both have been used for this purpose (Hammer, M. F. 1995; Seielstad, T. M. *et al.*, 1994; Whitfield, L. S. *et al.*, 1995 & Zerjal, T. *et al.*, 1997; Hammer, M. F. *et al.,* 1998). Population specific Y haplotypes have been described based on such markers (Jobling, M. & Tyler-Smith, C. 1995), which is of interest for phylogenetic as well as forensic purposes. A world wide sample was studied using biallelic markers along with a Y specific STR (DYS19), tracing early migrations not only from Africa and at least one from Asia (Hammer, M. F. *et al.*, 1997). Coalescence times of different population groups were also calculated using data based on slowly evolving markers, though different estimates were obtained (Hammer, M. F. *et al.*, 1995 & Whitfield, L. S. *et al.*, 1995).

In an early study it was shown that with only five Y STRs 15 distinct populations could be differentiated and phylogenetic trees and networks constructed in line with other data (Deka, R. *et al.*, 1996).

Subsequently it was repeatedly described how even a small number of Y STRs could differentiate between closely related populations, like Dutch and Germans or Catalans and Basques (Cooper, G. *et al.*, 1996; Perez-Lezaun, A. *et al.*, 1997; Perez-Lezaun, A. *et al.*, 1999). In a large study of the European populations marked allele frequency differences were observed between different populations, which were useful in performing phylogenetic analysis (Knijff, P. *et al.*, 1997). Thus it became apparent that Y STRs were the markers of choice when closely related populations were to be studied as the slowly mutating markers could not differentiate them. Since the mutation rates of Y STRs are high, they do not remain

linear for long times therefore Y STRs could be used for population differentiation on a historical time scale rather than an evolutionary one (Knijff, P. *et al.*, 1997; Seielstad, M. *et al.*, 1999; Forster, P. *et al.*, 2000).

A world wide sample consisting of 6 population groups has been studied for Y STRs and using genetic distances inferred the divergence times from the African populations apart from construction of phylogenetic trees (Seielstad, M. *et al.*, 1999). Recently it was shown that haplotypes consisting of higher resolution of DYS389 and 390 alleles along with other Y STRs might prove to be as powerful as biallelic markers for phylogenetic purposes (Forster, P. *et al.*, 2000).

1.4.8 Y CHROMOSOME BIALLELIC MARKERS

The Y chromosome is rich in biallelic markers and a number of such markers are available (Jobling, M.A. *et al.*, 1997; Underhill, P.A. *et al.*, 1997; Karafet, T.M. *et al.*, 1999; Su. B. *et al.*, 1999). Until recently only two single nucleotide polymorphisms were available however the SNP Consortium Ltd. (http://snp.cshl.org/data/ — 21st August 2000 release) has listed 841 SNPs on the Y, which still remain unverified.

There are a number of SNPs on the Y which have been verified and characterised (Shen, F. *et al.*, 2000 & Underhill, P.A. *et al.*, 2000). SNPs are unique events in human evolution and have low mutation rates (about 2×10-8 per base per generation) combinations of these markers thus define lineages (Thompson, R. *et al.*, 2000). The forensic utility of SNPs is limited due to low mutation rates and the biallelic nature of polymorphism therefore a large number of such markers have to be combined in order to have a meaningful assay. However if high throughput systems of analysis of SNPs are developed and standardised, they could prove to be a valuable adjunct to the present STR technology (Jobling, M. A. 2001).

1.5 BACKGROUND TO THE PROJECT

Pakistan is a large country inhabited by populations that are ethnically diverse (Figure 1.4). As in any other jurisdiction the number of medico legal cases relating to personal identification constitute a major workload of the government's Chemical Examiners Laboratories all over the country. These laboratories are currently using blood group antigens, serum proteins and serum enzyme systems for the analysis of evidentiary material submitted for forensic serological examination and there is an acute need of a forensic tool with a high probative value. In this regard DNA analysis is the latest technique which could be employed for human identification and paternity analysis.

Autosomal STRs are the cutting edge technology and forensic laboratories are using multilocus analysis in a manner depending on their need and resources.

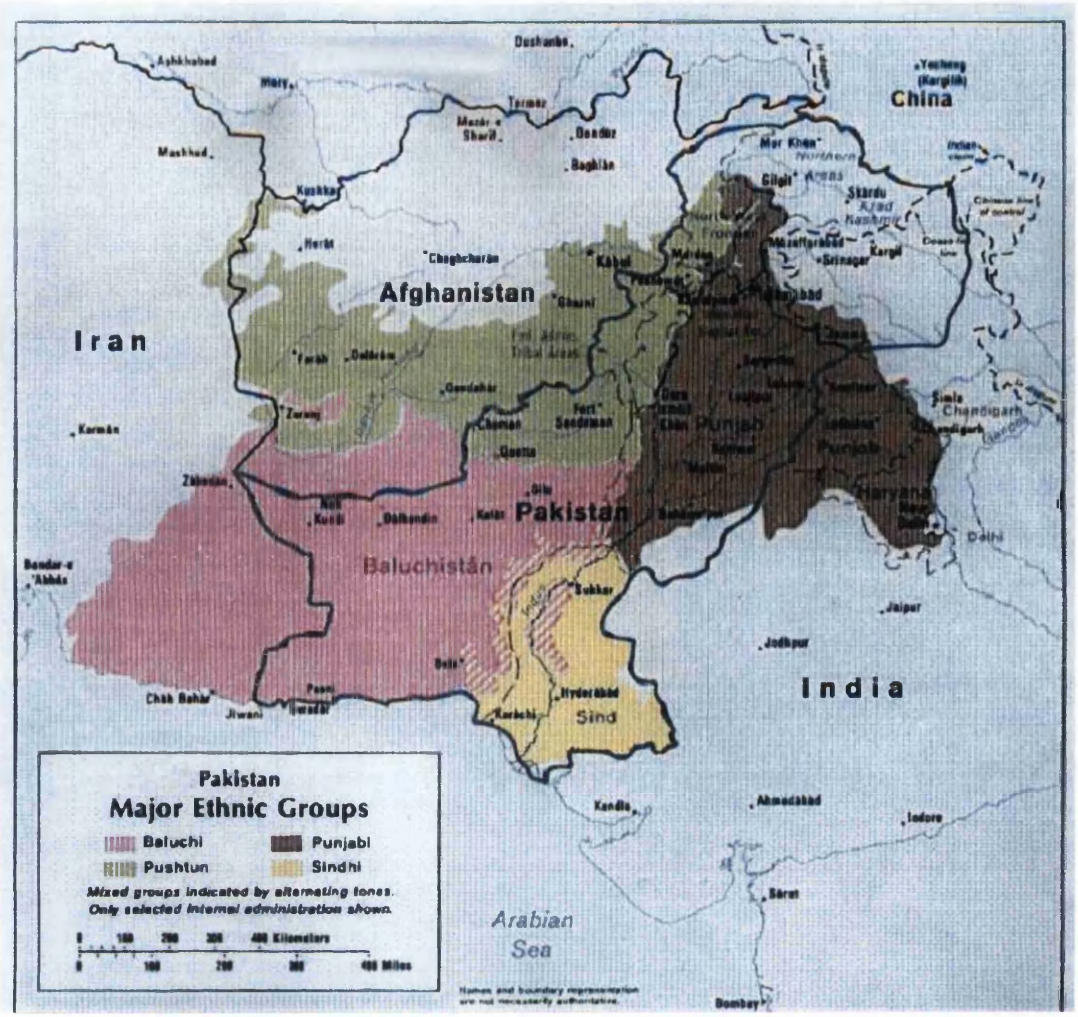
Recently Y chromosome STRs and mtDNA analysis have also gained importance in forensic work. In order to use DNA analytical methods, the STR markers have to be chosen and the structure of the populations studied in order to assess the degree of substructure. The society of Pakistan is peculiar in that it is comprised of distinct groups, which are largely endogamous due to cultural and linguistic differences. The proportion of cousin marriage is high in Pakistan among all the populations. This social structure is in contrast to for example, the European populations among which mating is more random, with lower inbreeding levels.

Apart from forensic purposes genetic markers can be employed to study phylogenetic relationship of different populations and Y STRs are useful for studying closely related populations for paternal lineages, which can throw light on the evolution and history of the populations. Pakistani populations inhabit the area of one of the world's most ancient civilisations, i.e. the Indus Valley, thus it would be of interest to compare them with one another and populations of other regions.

Figure 1.4: Map of Pakistan

Pakistan lies 23–37 degrees north latitude and 61–76 degrees east latitude, in the Southern Asia. The country has borders with Iran, Afghanistan, India and China and has a total area of 803,904 sq. km. It has a coastline 1074 km along the Indian Ocean. Pakistan occupies an important strategic position between Southern Asia and the Middle East .

The country has four provinces Punjab, Sindh, NorthWest Frontier Province (NWFP), and Baluchistan. The northern areas consist of a federally administered division of Gilgit & Northern Areas (FANA). The tribal belt in the NWFP along the border with Afghanistan is the federally administered tribal area (FATA). The federal capital is Islamabad. The main port Karachi is located at the southern edge of Pakistan.


Afghanistan
Iran
Pakistan
India
China
Baluchistan
Punjab
Sind
Haryana
Arabian Sea
Pakistan
Major Ethnic Groups
Baluchi
Punjabi
Pushtun
Sindhi
Mixed groups indicated by alternating tones.
Only selected internal administration shown.

1.6 HISTORY OF PAKISTANI POPULATIONS

Pakistan is a large country with a total population 150 million (Official web site http://www.pak.gov.pk/public/people/index.html). The present day major Pakistani populations consist of Punjabi, Sindhi, Pushtoon and Baluchis, which inhabit the four provinces.

The Punjabi population is the largest and comrpises more than half of the total population of Pakistan. Sindhis, Pushtoons makeup most of the rest. The Baluchi population is the smallest of the four. The Punjabi population is large and has inevitably interacted with the neighbouring populations thus areas exist which have intermingled populations of Sindhis/Punjabis or Pushtoon/Punjabis. Castes and subcastes exist in the Punjabi population. The Punjabi and Sindhi populations are considered to be the indigenous populations and have common history (Dani, H. 1981). Apart from these many smaller populations exist. Out of these the Kalash and the Brosho who live in the northern highlands and the Makrani who inhabit the coastal areas were included in this study. The major ethnic groups of Pakistan extend into the neighbouring countries (Figure 1.4).

The first recorded mention of the Pushtoons has been by Herodotus who described them as four tribes inhabiting the Gandhara country which consisted of the areas of present day Pakistan (Caroe, O. 1962). The Byzantine historian Procopius described Baluchis in the 6 Century A.D in the vicinity of Caucasus Mountains (Mazari, S. S. 1999). They shifted to Kirman province in 9^{th} Century and in the 12^{th} century were expelled by the Selecuids, when they migrated to Siestan. The final Baluchi migration occurred to present day Baluchistan in the early part of 13^{th} century (Mazari, S. S. 1999).

The Makrani oral tradition is that their ancestors came from Persia to the coastal regions of Pakistan. It is known that slave trading from Karachi port brought many slaves from Ethiopia and also Siddees from other parts of Africa (Baillie, A. F. 1890), who might be considered as the ancestors of Makranis given to their morphological features. They have intermingled with the Baluchi and

Sindhi populations living in the coastal regions of Pakistan and speak the same language.

The Brosho population has indefinite origins and their language cannot be included in any linguistic groups of the region (Berger, H. 1985). Most Brosho believe that their ancestors migrated from the Caucasus to its present location. However Brosho could be considered as a population which occupied much bigger area around their present abode.

The Kalash population lives in the northern part of Pakistan. The origin of the Kalash has been an interesting subject for the linguists and anthropologists. One of the older views is based on the Kalash folklore which, describes that their great ancestor was 'most probably Alexander of Macedonia (Siiger, H. 1956; Jan, S. 1996).

The other view which is held by most now, is that the Kalash arrived in Chitral with the early waves of the Indo Aryan migration in about second millenium B.C. (Morgenstierne, G. 1932; Schomberg, R. C. F. 1938; Jettmar, K. 1975 & Parkes, P. 1973). The religion of Kalash has clear similarities to the Hindu religion and it has been suggested that Kalasha religion is a mixture of Indo Aryan and local cultural forms and their language 'Kalahsamun' is one of the oldest languages of this region (Jettmar, K. 1961; Burrow. F. 1973).

1.5.1 AIMS OF THE PROJECT

The aim of the project was to study the genetic structure and composition of the Pakistani populations using autosomal and Y chromosome STR loci in order to assess the feasibility of applying DNA profiling methods to medico-legal cases; and to develop a database of autosomal and Y STR markers in ethnically distinct populations of Pakistan for forensic and phylogenetic purposes.

1.5.2 THE DIFFERENT PHASES OF THE PROJECT

- Collection of sufficient number of relevant samples from each population group in Pakistan employing a robust method, which would ensure good yield of high quality DNA.

- To develop PCR amplification protocols for all the STR markers as single and multiplex reactions and genotyping of the samples using an automated DNA sequencer

- Construction of "in house" sequenced allelic ladders for the autosomal and Y chromosome STR markers.

- Study of the genetic structure and the extent of utility of the autosomal and Y chromosme STRs for personal identification in the various Pakistani populations in comparison with other populations.

- Investigation of phylogenetic relationship of Pakistani populations and the comparison with historical and linguistic evidence.

CHAPTER 2: MATERIALS AND METHODS

2.1 MATERIALS

2.1.1 CHEMICALS

The chemicals were obtained from Bio-Rad Laboratories Ltd UK, Fisons Scientific Equipment, UK and Sigma Chemical Co. USA, unless otherwise indicated.

2.1.2 DNA EXTRACTION KIT

The Puregene DNA extraction kit was purchased from Gentra (Flowgen, Novara group Ltd, UK), containing RBC Lysis Solution, Cell Lysis Solution, Protein Precipitation Solution and DNA Hydration Solution.

2.1.3 CHELEX RESIN

Chelex resin was obtained from Bio-Rad Laboratories, UK.

2.1.4 PCR PRIMERS

Oswel DNA Services, UK and MWG-Biotech AG, manufactured all PCR primers. The forward primer of each set was labelled with an appropriate ABI dye. Oswel DNA Service, UK, supplied the primers for the autosomal STR. All the three were labelled with 5'-FAM. It also supplied the Y chromosome STR primers for the loci DYS391, 392 and 393. The forward primers of DYS391 and DYS393 were labelled with 5'-FAM. The forward primer of DYS392 was labelled with 5'-JOE dye. MWG-Biotech AG supplied the primers for Y STR loci DYS19, 389 and 390. The sequence and dye labels of the primers used for the amplification of the autosomal and Y chrmosome STRs were as per Table 2.1.

Table 2.1: Primers Used for Autosomal & Y Chromosome STR Loci

LOCUS	SEQUENCE (5'- 3')	Size (bp)	GC %	ABI DYE
D3S1358	F*- ACTGCAGTCCAATCTGGG	18	55.6	5'-FAM
	R**- ATGAAATCAACAGAGGCTTG	20	42.0	None
vWA	F- CCCTAGTGGATGATAAGAATAAT	23	34.8	5'-FAM
	R- GGACAGATGATAAATACATAGGAT	24	33.3	None
FGA	F-CCATAGGTTTTGAACTCACAG	21	42.9	5'-FAM
	R- CTTCTCAGATCCTCTGACACTC	22	50.0	None
DYS391	F- CTATTCATTCAATCATACACCCA	23	34.8	5'-FAM
	R- CTGGGAATAAAATCTCCCTGGTGCAAG	28	46.4	None
DYS392	F- TCATTAATCTAGCTTTTAAAAACAA	25	20	5'-JOE
	R- AGACCCAGTTGATGCAATGT	20	45	None
DYS393	F- GTGGTCTTCTACTTGTGTCAATAC	24	41.7	5'-FAM
	R- AACTCAAGTCCAAAAAATGAGG	22	36.4	None
DYS19	F- CTACTGAGTTTCTGTTATAGT	21	33.3	5'-JOE
	R- ATGGCATGTAGTGAGGACA	19	47.4	None
DYS389	F- CCAACTCTCATCTGTATTATCTAT	24	33.3	5'-FAM
	R- TCTTATCTCCACCCACCAGA	20	50	None
DYS390	F- TATATTTTACACATTTTTGGGCC	23	30.4	5'-FAM
	R- TGACAGTAAAATGAACACATTGC	23	34.8	None

* F is the forward primer

** R is the reverse primer

2.1.5 AMPLI*Taq* GOLD ENZYME

Ampli*Taq* Gold was purchased from Applied Biosystems, CA, USA. It was provided with the reaction buffer at a concentration of 10 x and 25 mM Magnesium chloride solution.

2.1.6 *Taq* POLYMERASE

Taq polymerase was purchased from Promega Southampton, UK. It was provided with reaction buffer at a concentration of 10 x and 25 mM Magnesium chloride solution.

2.1.7 AMPFℓSTR® BLUE KIT

The multiplex amplification kit AmpFℓSTR® Blue used for the amplification of the loci D3S1358, vWA and FGA was purchased from Applied Biosystems CA, USA.

2.1.8 REDDYMIX™ MASTER MIX

ReddyMix™ Master mix was purchased from AB Gene®, UK.

2.1.9 CEP BUCCAL BRUSHES

CEP Buccal Brushes were procured from Life Technologies Inc. USA.

2.2 GENERAL PREPARATORY PROCEDURES

2.2.1 AUTOCLAVING

All glassware was sterilised at 15 psi for 1 h in a laboratory autoclave.

2.2.2 MICROFUGE TUBES AND PLASTIC WARE

Microfuge tubes were sterilised in a microwave oven at 100% power for 20 min.

2.2.3 ULTRAVIOLET TREATMENT OF MICROFUGE TUBES

The microfuge tubes were UV treated for 30 min exposing the inside of the tubes to UV light on transilluminator Model TC-312A, Spectroline®.

2.3 EXTRACTION OF DNA

Two methods were used to extract DNA. DNA was extracted from the liquid blood samples using the Puregene Kit. However for the stains, in addition to a modified Puregene Kit method, the chelex extraction method was also used. The buccal cell samples collected using the CEP buccal brushes were also extracted using the Puregene® Kit

2.3.1 PUREGENE® KIT BASED DNA EXTRACTION FROM BLOOD SAMPLES

UV treated microfuge tubes were placed in a rack and were labelled. To each tube 900 µl of RBC lysis solution was added. The blood samples were allowed to thaw and were mixed by inverting several times. From each sample 300 µl of blood was added to the microfuge tube containing the RBC lysis solution. This was mixed by inversion. The mixture was incubated at room temperature for 5 min, re-mixed and then again incubated for 5 min. The tubes were centrifuged for 3 min at 14000*g* in microcentrifuge (Scotlab, UK).

A white pellet was visible after centrifugation. The supernatant was removed except for ~40 µl. The pellet was resuspended by vortexing on a whirl mixer (Gallenkamp) for 1 min. Cell lysis solution (300 µl) was added to each tube and mixed by inversion until the lysate was clear. If clumps were visible then

incubation was carried on for another 4 to 6 h till no clumps were visible. Protein precipitation solution (100 µl) was added and the solution vortex on the Whirl mixer for 30 s. The tubes were centrifuged for 5 min at 14000*g* for 5 min. 300 µl of 100% isopropanol (Fisons, Co. UK) was added to labelled microfuge tubes. To these tubes, the supernatant from the previously centrifuged tubes was added.

The tubes were inverted 30 to 50 times and were then centrifuged for 1 min at 14000*g*. The supernatant was discarded. The pellet was washed in 80% absolute ethanol and again centrifuged at 14000*g* for 5 min. Ethanol was removed carefully and the pellet was air dried for 10 min. 100 µl of DNA hydration solution was added to each tube and the DNA was allowed to re-hydrate over night. The sample was then divided into two 50 µl aliquots. One was stored at –20 °C and the other at 4 °C in the refrigerator for quantitation.

2.3.2 DNA EXTRACTION FROM STAINS USING THE PURGENE® KIT

The blood sample stains were collected on 42.5 mm circles of Whatman® filter paper (Whatman International Ltd, UK). A 4-6 mm stain was cut out from the sample to be extracted and was placed in an appropriately labelled 2 ml screw cap tube. 400 µl cell lysis solution and 10 µl Proteinase-K (10mg/ml) solution were added to the tubes containing the stain. The tube was incubated at 55 °C in a rotary incubator overnight. At the end of incubation 100 µl protein precipitation solution was added to the tube and vortexed for 1 min on a whirl mixer. The tube was then incubated in a –70 °C freezer for 30 min and centrifuged at 14000*g* for 5 min.

A 1.5 ml microfuge tube was labelled and 400 µl 100% isopropanol was added to it. The supernatant from the screw cap tubes was transferred into the tube containing isopropanol. The screw cap tube was discarded. The microfuge tube containing the mixture was inverted for 50 times and again incubated in a –70 °C freezer for 30 min. The microfuge tube was centrifuged for 10 min at 14000*g*. The supernatant was discarded. 400 µl of 80% EtOH to the tube and the pellet washed by gentle inversion. The tube was again centrifuged at 14000*g* for 5 min. The

EtOH was discarded and the pellet was air dried for 10 min. 40 µl of DNA hydration solution was added to the tube. DNA was allowed to re-hydrate over night.

2.3.3 DNA EXTRACTION USING CHELEX 100

A 5% solution of chelex 100 (Bio Rad Laboratories, UK) was prepared in sterile Milli-Q. A 4 mm piece of stain was cut using a pair of scissors from the stain samples to be extracted and placed in a labelled microfuge tube. 1 ml of sterile water was added to each tube and the tubes were incubated for 30 min at room temperature with occasional mixing by inversion and then centrifuged at 14000*g* for 3 min.

The supernatant was removed without disturbing the pellet and discarded. Again 1 ml sterile water was added to the tubes and incubated on a platform shaker for 15 min. The tubes were centrifuged at 14000*g* for 3 min and the supernatant was removed. The wash step was repeated for the three times for all the samples. The supernatant was removed leaving ~20 µl behind and 200 µl of 5% chelex solution was added to each tube. The samples were incubated at 56 °C for 30 min in a water bath. The tubes were vortexed on a Whirl mixer at high speed for 10 s. The caps of the tubes were pierced with a heated needle and boiled in a water bath for 8 min, then centrifuged for 3 min at 14000*g*. These were stored at –20 °C. Before using the stored samples, they were let to thaw completely, vortexed briefly and centrifuged at 14000*g*. A volume of 2-5 µl of the supernatant was used for PCR.

2.3.4 DNA EXTRACTION FROM BUCCAL BRUSHES

The buccal brush was removed from the tube containing the sample and Proteinase-K (10mg/ml, 20 µl) solution was added to tube. The tube was incubated at 60 °C in a rotary incubator for 2 h. At the end of incubation, 100 µl protein precipitation solution was added to the tube and vortexed for 1 minute on a whirl mixer. The tube was then incubated in a –70 °C freezer for 30 min. The tube was

taken out of the freezer and allowed to thaw. Then it was centrifuged at 14000*g* for 5 min. A 1.5 ml microfuge tube was labelled and 400 µl 100% isopropanol was added to it. The supernatant from the screw cap tubes was transferred into the tube containing isopropanol. The microfuge tube containing the mixture was inverted for 50 times and again incubated in a –70 °C freezer for 30 min. The microfuge tube was centrifuged for 5 min at 14000*g*. Supernatant was discarded. EtOH (80%, 400 µl) was added to the tube and the pellet washed by gentle inversion. The tube was again centrifuged at 14000*g* for 2 min. The EtOH was discarded and the pellet was air dried for 10 min. DNA hydration solution (100 µl) was added to the tube. The DNA was allowed to re-hydrate over night.

2.4 AGAROSE GEL ELECTROPHORESIS

I.D.na® Agarose was supplied by Flowgen, Novara Group Ltd, UK. For preparing a 1% gel 0.5 g of agarose was weighed and added to a solution of 1 x TBE. This was prepared by adding 45 ml Milli-Q water to 5 ml of 10 x TBE (0.09 M Tris-borate, 2M EDTA). The agarose suspension was heated in the microwave oven for 90 s. It was allowed to cool to 60 °C and 1 µl Ethidium Bromide (10mg/ml) was added. The solution was then poured into the casting tray into which the comb was already set, and allowed to set. Enough 1 x TBE was added to just submerge the gel.

The samples to be loaded were mixed with one fifth volume of loading buffer, 0.25% bromophenol (w/v). The gels were run at 100 mA until bromophenol had migrated two-thirds of the way down and was then visualised under UV light on a transilluminator (Model TC-312A, Spectroline®). The gel was photographed using a Polaroid camera (Quick Shooter, Model QSP, International Biotechnologies Inc., New Haven, Connecticut, USA) with 667 Polaroid® film.

2.5 DETERMINATION OF NUCLEIC ACID CONCENTRATION

5 µl of the DNA solution to be quantitated was added to the microfuge tube containing 1 µl 0.25% bromophenol (w/v) loading buffer and mixed by pippetting. The mixture was loaded on a 1% agarose gel. K562 DNA Quantitation Standards (Flowgen, Novara Group Ltd, UK) 250, 125 and 62.5 ng concentration (6 µl) were also loaded besides the samples to be quantitated. The gel was photographed after the run and the bands of the extracted DNA sample were compared to those of the standards.

2.6 POLYMERASE CHAIN REACTION

2.6.1 DESIGN OF OLIGONUCLEOTIDES

2.6.1.1 Autosomal STR Loci D3S1358 and FGA

The sequences for the two loci vWAand FGA were searched in the online GenBank database. The primers reported for both the loci were modified and were assessed using the computer programs Primer Pick (Rozen. S. & Skaletsky. H. J. 1997) and Program Primer (Williamstone Corp, UK). The length, average Guanine + Cytosine content, and the Tm of the forward and reverse primers were checked.

Locus D3S1358 sequence was not available on the GenBank so the reported primers (Li, H. *et al.*, 1993) were compared to those of the other two loci and the forward primer was slightly modified.

2.6.1.2 Y Chromosome STR Loci

(i) The sequences of the forward and reverse primers for the loci DYS19, 389, 390, 392 and 393 were obtained from the Genome Database and were as follows:

DYS19 (Accession ID GDB: 455927); DYS389 (Accession ID GDB: 365241)

DYS390 (accession ID GDB: 365248); DYS392 (Accession ID GDB: 455698)

DYS393 (Accession ID GDB: 455838).

The sequence of the forward primer for the locus DYS391 was obtained from the Genome Database (Accession IDGDB: 365251). The reverse primer was as described by Gusmao. L. *et al.*, 1999.

2.6.2 LABORATORY PROCEDURES

Extreme care was necessary to avoid carryover and extraneous contamination during the work. Therefore, strict PCR laboratory procedures were used throughout the work.

2.6.2.1 Separate Areas for Pre and Post PCR Lab Work

In order to prevent carryover of the amplified DNA, all reactions were set up in a flow hood. The pipettes sets and other consumables and equipment were dedicated to the pre-PCR area. All microfuge tubes and plastic ware was subjected to UV treatment before setting up the reactions. The flowhood was cleaned with decon before and after the work. Gloves were changed frequently during the PCR work. The PCR machine was located in a separate room. Post PCR work was performed in another room, physically separated from the pre-PCR area. All equipment, laboratory coats, gloves and other consumables used in the post-PCR area were dedicated to this area and were never brought to the pre-PCR room. Also a strict uni-directional flow of personnel and equipment/consumables was maintained from the pre-PCR to post-PCR work areas.

2.6.2.2 Pippeting

Barrier tips were used from Life Sciences International, UK. Two sets of pipettes were exclusively dedicated to pre- and post- PCR areas.

2.6.2.3 Autoclaving

All milli-Q water was routinely autoclaved in laboratory autoclave. The microfuge tubes were sterilised in microwave oven and then UV treated for 30 min on the transilluminator (2.2.3).

2.6.2.4 Alkali Treatment of Plastic and Metallic Equipment

The plastic racks and other equipment were routinely placed in a strong solution of commercial bleach. Care was taken that any such equipment remained submerged overnight when it was taken out. The equipment was then washed with water and autoclaved in the microwave oven at full power and was then subjected to UV treatment for 30 min on the transilluminator (2.2.3).

2.6.2.5 Preparation of Aliquoted Reagents

The reagents to be used for PCR were prepared in the flowhood and stored in a separate area. Extreme precaution was taken during pipetting of the primers, dNTPs and enzymes. In order to avoid the risk of contamination and the pipetting errors all reagents were prepared in bulk. These were premixed and aliquoted into volumes necessary for a typical set of reactions.

2.6.2.6 Inclusion of Reaction Controls

A negative control was included for each PCR reaction. The negative control contained all the ingredients for the particular PCR reaction except the template DNA.

2.6.3 PCR REACTION PROCEDURES FOR THE AUTOSOMAL LOCI

2.6.3.1 Preparation of dNTP Stock

200 µM (400 µl) stock solution was prepared from original dNTP (Promega, UK) supplied at a concentration of 100 mM. Aliquots of 100 µl were stored at –20 °C.

2.6.3.2 Preparation of Primer Stock

The primer stocks of the autosomal loci were prepared at a concentration of 0.25 µM. Stocks of 100 µl were prepared and stored at –20 °C.

2.6.3.3 Preparation of Bulk Reaction Solutions

A bulk reaction mix was prepared using the stock solutions containing dNTP, primers and buffer solution (Table 2.2a). The table shows the volumes of the components for a single reaction. These solutions were prepared for batches of 200 reactions at a time. The batch was tested for the performance by setting up and running three reactions along with a negative control. Once the batch was tested, all other PCR reactions were performed using the same primer mix. The next reaction mix solution was then prepared using the same recipe.

2.6.3.4 Amplification of Autosomal STRs D3S1358 and FGA in Duplex Reaction

A 25 µl reaction volume was used for the amplification of the duplex. The tested batch of the primer mix was let to thaw. The *Taq* enzyme was taken out on ice and was added to the master mix solution. To the labelled 0.5 ml or 0.2 ml PCR tubes placed on ice, 23 µl of this master mix was added. DNA solution was added to each tube. To the negative control tube, sterile water was added instead of DNA.

A drop of mineral oil (Sigma Chemical Co., USA) was added to all the tubes. DNA solution was added to each PCR tube through the oil.

Table 2.2a: Reaction Mix Solution for Amplification of Loci D3S1358, vWA & FGA

Component	Singleplex (μM)	Duplex (μM)
dNTP	200	200
10x Buffer	1x	1x
Primers D3S1358	0.2	0.2
Primers vWA	0.2	0
Primers FGA	0.2	0.1
$MgCl_2$ 25mM	2.0 mM *	2.5 mM

Table 2.2b: PCR Ingredients for Amplification of Loci D3S1358, vWA & FGA

Component	Singleplex Vol per reaction (μl)	Duplex Vol per reaction (μl)
Reaction Mix	13.3	13
Enzyme	1	1
Water	8.7	9
DNA	2	2

The PCR tubes were pulse spun at 14000*g* in a micro-centrifuge and placed in the thermal cycler (PHC II, Techne, Hybaid Omnigene) and the reaction started. However when using the GeneAmp® PCR System 2400 (Applied BioSystems CA, USA) no mineral oil was added to the PCR tubes. The thermal cycler was programmed to denature the samples and to activate *Taq* Gold by the inclusion of 11 min of initial denaturation at 94 °C. This step was reduced to 3 min if the Promega *Taq* was used. Thirty cycles of PCR were used to amplify the template DNA. The samples were run on a yield gel after the completion of PCR. All the reaction tubes were stored in –20 °C freezer in a labelled scaled plastic bag, pending analysis on the automated sequencer.

2.6.3.5 PCR Amplification of Autosomal STR

The method was essentially same as for the duplex reaction (2.6.3.4). The reaction components of the singleplex amplification reaction however were slightly different (Table 2.2b).

2.6.3.6 PCR Amplification of Autosomal STRs Using AMPFℓSTR® Blue Kit

The population samples from the Baluchi, Brosho, Kalash and Makrani populations were amplified for the three autosomal loci, D3S1358, vWA and FGA, using the AmpFℓSTR® Blue Amplification kit (Applied BioSystems CA, USA). All reactions were amplified using the GeneAmp® PCR System 2400 (Applied BioSystems CA, USA), in thin walled 0.2 ml reaction tubes. The reaction volume was 10 µl. The protocol of the PCR described by the manufacturer was modified to scale down the reaction components. All the reaction components were provided with the kit and these included the AmpFℓSTR® PCR reaction mix, AmpFℓSTR® Primer Set and Ampli*Taq* Gold™ DNA Polymerase.

Template DNA was quantified (2.6) and diluted in deionised water so that 4 µl would contain ~0.2-1 ng DNA.

The master mix (6 µl) was aliquoted into each reaction tube. 4 µl of template DNA was added to each reaction tube. 4 µl of the AmpFℓSTR® Control DNA was added to the positive control reaction tube and same volume of de-ionised water was added to the negative control. PCR was performed (Table 2.3a & b) in a GeneAmp® PCR System 2400 (Applied Biosystems CA, USA).

2.6.4 PCR REACTION PROCEDURES FOR THE Y-CHROMOSOME STR LOCI

2.6.4.1 Preparation of Primer Stock

The stocks of the Y chromosome STRs were prepared at 0.25 µM except DYS19 which was prepared at 0.3 µM. Stocks were prepared in 100 µl volumes and stored at –20 °C.

2.6.4.2 Preparation of Primer Solutions

The Y-chromosome STRs were amplified using the master mix ReddyMix™, which was purchased from AB Gene®, UK. This contains the dNTPs, buffer, enzyme and Magnesium Chloride solution at the desired concentration. This solution has to be stored at –20 °C. Once the master mix was thawed it was not frozen again and was stored in the refrigerator. The solutions of the primers were prepared according to the optimized conditions for the primer concentration.

For singleplex/multiplex reactions primer volume of the forward and the reverse primer for DYS391, 392 and 393 was kept at 0.5 µl at the desired concentration. For the multiplex reactions, for a typical set of 28 reactions 14 µl of each primer (forward and reverse) of each locus was aliquoted in a 1.5 ml microfuge tube and stored at –20°C.

Table 2.3a: AmpF*ℓ* STR® Blue Kit Master Mix Ingredients

Component	Volume (µl) Per Reaction
Reaction Mix	4.1
Enzyme	0.2
Primer Set	2.1

Table 2.3b: Thermal Cycling Parameters AmpF*ℓ* STR® Blue Loci Amplification

Step	Temperature °C	Time min	No of Cycles
Initial Incubation	95	11	1
Denaturation Annealing Extension	94 59 72	1 1 1	28
Final Extension	60	45	1
Hold	25		

2.6.4.3 PCR Methods Y-Chromosome STRs

(i) Amplification of Single Locus Y STRs

The master mix for the amplification reaction consisted of the primer solutions of both the primers for the specific locus, the Reddymix™ Master mix solution and the Magnesium Chloride solution if required (Table 2.4a). The volume of each primer was kept at 0.5 µl. The DNA solution was added (~5 ng) to the reaction tube and then 23 µl of the master mix was aliquoted in the cap of each reaction tube. The tubes were closed carefully and were pulse spun. The reaction tubes were quickly transferred to the PCR machine.

(ii) Amplification of Y STR Multiplex I (DYS391, 392 & 393)

The volumes of the primer solutions and DNA were kept at a constant (Table 2.4a). $MgCl_2$ concentration was kept at 2.5 mM. If the available master mix was one at a lower concentration of $MgCl_2$, additional $MgCl_2$ solution was added to increase the concentration of magnesium chloride.

(iii) Amplification of Y STR Multiplex II (DYS19, 390, 389I & 389II)

Optimal $MgCl_2$ concentration was 2.5 mM so Reddymix™ Master mix solution at 2.5 mM was procured. The primer mix solution for all the loci was prepared in bulk. Aliquots enough for 28 reactions were prepared and were stored at –20 °C.

2.7 PREPARATION OF ALLELIC LADDERS

2.7.1 PREPARATION OF ALLELIC LADDERS FOR AUTOSOMAL LOCI

A number of samples were amplified for the loci D3S1358, vWA and FGA using the AmpFℓSTR® Blue kit (Applied Biosystems, CA, USA). The alleles of the three loci were sized using the ladder supplied with the kit and the

Table 2.4a: Reaction Components Y Chromosome STR Multiplex Amplification

Component	Multiplex I &II (µl)
ReddyMix® Mastermix	20
Primer * mix *	3
DNA	2

* Primer concentrations Multiplex I DYS391 & 392 (0.25 µM), DYS393 (0.125 µM)

* Primer concentrations Multiplex II DYS19 (0.3 µM), DYS389 &390 (0.125 µM)

Table 2.4b: Thermal Cycling Parameters for Y STR Multiplex Amplification

	PCR Cycle			
Multiplex	**Step**	**Temperature °C**	**Time min**	**No of Cycles**
Multiplex I	Denaturation	94	2	1
&	Denaturation	94	1	30
	Annealing	55/57*	1	
Multiplex II	Extension	72	1	
	Final Extension	65	10	1

* 57 °C Annealing for Multiplex I, 55 °C for Multiplex II

alleles were designated. Samples showing different genotypes were selected so that the frequent alleles were represented. Loci D3S1358, vWA and FGA were amplified as singleplex reactions using the new primers for these samples (Table 2.1). The sizes of the alleles were compared to those obtained with the Blue kit.

For preparing ladders for individual locus (D3S1358, vWA & FGA) samples amplified for each locus were selected containing different alleles. An aliquot (2 µl) of the amplified PCR products from such samples was mixed in a tube. Thus a mixture of the PCR products was prepared for each locus. The tube was vortexed and then 1 µl was pippetted out for running the samples on the automated DNA Analyser. The volume of the samples was adjusted for producing peaks of equivalent strength for all the alleles. The single locus ladders were thus prepared.

Sizes obtained for the samples of the allelic ladders with new primers and the Blue kit are shown in Table 4.8. Once prepared, 1 µl of the final ladder for each locus was used to run as an external ladder in all-subsequent gels run on the automated ABI 373 XL UPGRADE DNA sequencer..

2.7.2 PREPARATION OF ALLELIC LADDERS FOR THE Y STR LOCI

Annabel Gonzalez and Professor Angel Carracedo of the Institute of Legal Medicine, University of Santiago de Compostela kindly provided two control samples. These samples had been sequenced at various laboratories with the same results (Knijff, P. D. *et al.,* 1997 & Pestoni, C. *et al.,* 1998). The data of the control samples is attached as Appendix 3.

2.7.2.1 Amplification and Genotyping of the Control Samples for Y STRs

The control samples were provided as stains on cotton cloth. These stains were extracted with the Purgene® DNA Extraction kit (2.3.2). Approximately 10 ng of the extraction from each sample was used as template in the multiplex PCR for DYS391, 392 and 393 and a quadraplex reaction of DYS19,

390, 389I & 389II separately. The samples were run on an ABI 373 XL automated DNA sequencer and were analysed using GeneScan® software.

2.7.3 SEQUENCING OF Y CHROMOSOME STR ALLELES

Alleles for the loci DYS19, 389I, 389II, 390 & 392 were sequenced in order to confirm the allelic size.

2.7.3.1 Sample Purification

The samples selected for the preparation of the allelic ladder were amplified as a singleplex reaction for the all the loci.

(i) PCR Purification

For the loci DYS19, 390, and 392, 50 µl PCR reactions were purified using the Qiaquick PCR Purification Kit® (Qiagen Ltd, UK). The Qiaquick PCR Purification kit consists of Buffer PB, Buffer PB and the elution Buffer. Using the centrifuge method the samples were cleaned for sequencing reactions. 100% ethanol was added to the PE Buffer before starting the procedure. 50 µl PCR reaction from each sample was placed in the column and 250 µl Buffer PB was added. The lids of the collection tubes were snap closed and centrifuged at 14000*g* for 1 min. The flow through was discarded from the collection tubes by removing the column from each collection tube at a time. Buffer PE (0.75 ml) was added to all the columns, the lids of the collection were closed and the tubes were centrifuged for 1 min at 14000*g*. The flow through was discarded from the collection tubes, the columns replaced into them and all the tubes were centrifuged for 1 min again. The columns were placed in clean, labelled 1.5 ml microfuge tubes. The collection tubes were discarded. 30 µl of the elution buffer was applied to the center of the columns and the columns incubated at room temperature for 1 min. The columns were then centrifuged for 1 min. The eluate was the purified DNA, which was preserved in the refrigerator at 4 °C pending sequencing reaction.

(ii) Agarose Gel Purification

DYS389 generates two fragments with the same set of primers so the fragments had to be separated on the agarose gel and then purified. Qiaquick Gel Purification kit was used for this purpose. PCR products (75-100 µl) were run on a 1% agarose gel. The fragments were visualised under UV light and the bands were cut out with minimal exposure of the gel to the UV light. The bands were placed in labelled 1.5 ml microfuge tubes, which had been weighed previously. Three volumes of the QG buffer for 1 volume of the band were added to each tube containing the cut out band and the tubes incubated at 55 °C for 15-20 min. At the end of incubation it was ensured that the gel has completely dissolved. The colour of the mixture was compared to that of the QG buffer, if it was same the procedure was continued. If the colour of the mixture did not match that of the QG buffer then 5-10 µl of 3 M Sodium Acetate solution was added to the tube till the colour changed matched.

100% isopropanol was added to all the tubes in 1:1 w/v of the gel slice. Same number of the collection tubes was arranged on the rack as the number of samples to be purified and was labelled. Columns were placed in all the collection tubes. From each 1.5 ml microfuge tube, the mixture of the gel slice, QG Buffer and isopropanolol were transferred to the columns. The columns were spun at maximum speed for 1 min. The flow through was discarded. 0.2 ml Buffer QG was again added to all the columns and the columns centrifuged for 1 min. The flow through was discarded. 0.75 ml Buffer PE was added to the columns. The columns were incubated for 5 min and then centrifuged for 1 min. The flow through was discarded. 50 µl of Buffer EB was applied to the center of each column changing tips between each application. The columns were incubated for 1 min and then centrifuged for 1 min. The eluate was the purified DNA. This was stored in the refrigerator at 4 °C pending sequencing reaction.

(iii) An aliquot of the purified samples (2 µl) were run on 1% agarose gel with 1 Kb Ladder (Life Technologies, UK), in order to quantitate the purified DNA.

2.7.3.2 Sequencing Reactions

The Molecular Biology Sequencing Unit, Institute of Molecular Biology, University of Glasgow sequenced the purified samples with the forward primers (3.2 pmol/reaction) using ABI Prism® Big Dye™ Terminator Cycle Sequencing Kit. Unlabelled forward primer was recommended by the manufacturer however labelled forward primers were used for PCR.

2.7.4 PREPARATION OF ALLELIC LADDERS

Allelic ladders for the Y chromosome STRs were prepared using the same general strategy as for the autosomal STRs.

2.7.4.1 Preparation of the Allelic Ladder for Y Chromosome Multiplex I

The samples were amplified as a singleplex reaction for each locus. Allelic data for these samples is shown in Table 2.8. The ladders for these loci were prepared separately first by mixing an equal volume of the PCR products from the samples selected for the generation of the ladders. These ladders the ABI 373A XL (Applied Biosystems CA, USA). The data was analysed using Genescan® 2.0 software. The peak height of the alleles in the ladder was assessed. The samples were run again with less DNA if non-specific amplification or stuttering was observed. The samples were mixed and the volume of each sample was adjusted according to the peak height. The individual ladders were then mixed (Ladder DYS391 2 Parts, Ladder DYS391 2 Parts & Ladder DYS393 3 Parts). 1 µl of ladder was run with all the samples, which were amplified as Multiplex I.

2.7.4.2 Preparation of the Allelic Ladder for Y Chromosome STR Multiplex II

Samples were selected on the basis of the allelic size and were amplified for a quadraplex reaction. These were genotyped on the ABI 373 XL alongwith the

two control samples. Equal volume of the PCR reaction of all the amplifications were mixed in a tube. 1 µl of this external ladder was run with all the samples.

2.8 ELECTORPHORESIS OF AMPLIFIED PCR PRODUCTS ON AN ABI 373A XL AUTOMATED DNA SEQUENCER

The PCR products were electrophoresed on the ABI 373A XL UPGRADE Automated DNA Sequencer (Applied Biosystems CA, USA at the Molecular Biology Sequencing Unit of the University of Glasgow.

2.8.1 PREPARATION AND LOADING OF THE SAMPLES

While the machine was running on pre-run module, the samples were prepared for loading on the gel. PCR product (0.5 µl) was pipetted out from each sample into labelled 0.2 ml microfuge tubes. If the samples were being pipetted out of a tube having a mineral oil overlay, care was taken that no oil was carried out. These steps were performed during the time the gel was setting.

During the pre run of the machine, a solution was prepared containing 400 µl de-ionised formamide (Sigma) and 100 µl bromophenol gel loading buffer (Applied Biosystems CA, USA). 120 µl of this solution was mixed with 8-15 µl of GeneScan™ 350 or 500 [Rox] internal standard (Applied Biosystems CA, USA). Two µl of this solution was pipetted to each tube containing the 0.5 µl of sample. The tubes were pulse spun and heated in a DNA thermal cycler 480 (Applied Biosystems CA, USA) at 96° C, for two min. The tubes were immediately placed on ice.

1.5-2 µl of the each prepared sample was loaded on the gel, using a P2 Gilson pipette with the flat duck bill tips (Life Technologies, UK) or a multi barrel syringe (Burke). Care was taken that the wells were not over loaded and that there were no leaks into the adjacent wells.

2.9 STATISTICAL AND PHYLOGENETIC COMPUTER SOFTWARE EMPLOYED FOR THE ANALYSES OF THE DATA

In addition to Microsoft® Excel various computer software packages were used to analyse the data.

2.9.1 GENETIC DATA ANALYSIS (GDA)

GDA software (Lewis, P. O. & Zaykin, D. 2000) is a free program distributed by the authors over the web from the GDA Home Page at http://alleyn.eeb.uconn.edu/gda/. GDA version 1.0 (d15) was used for the analysing the autosomal STR data. The allele/genotype frequencies, the exact test p- values and F Statistics were calculated. The program was also used to generate a phylogenetic tree using the autosomal STR data.

2.9.2 TOOLS FOR POPULATION GENETICS ANALYSIS (TFPGA)

This was used to reanalyse the data in order to confirm the statistics generated using GDA. It has been developed by Mark Miller and is available from his website: http://herb.bio.nau.edu/~miller/ as a Windows executable.

2.9.3 POWERSTATS

PowerStats (Tereba, A. 1999) is a Microsoft® Excel based program (Promega, Corp, Madison, USA) which was used to calculate important forensic statistics (5.2.4).

2.9.4 ARLEQUIN

Arlequin (Schneider, S. *et al.*, 2000) version 2000 was used to analsye the Y STR haplotype data. The allele & haplotype frequencies, analogues of F statistics, genetic distances were calculated. Analysis of Moleular Variance was also performed using this software.

2.9.5 MICROSAT

Microsat was downloaded from Luca Cavalli-Sforza laboratory web site http://human.stanford.edu/microsat/microsat.html. It was was used to calculate the genetic distances using the Y STR individual locus data.

2.9.6 PHYLIP

PHYLIP (Felsenstein, J. 1989) is phylogenetic software and the genetic distances generated using the GDA, Arlequin and Microsat were used to construct various phylogenetic trees employing PHYLIP version 3.5c was used.

2.9.7 TREEVIEW

TREEVIEW (Page, R. D. M. 1996) was used to display and print the phylogenetic trees drawn using PHYLIP software.

CHAPTER 3: COLLECTION OF SAMPLES

3.1 INTRODUCTION

When utilising DNA analysis for forensic identification the DNA profiles generated from evidentiary material are compared to that of the suspect. If a match is declared between the two profiles, a statistical value is attached to the match in order to convey the evidentiary strength. This is possible if pertinent data is available from the reference population. It has therefore been recommended that there be a relevant database of the allele and haplotype frequencies for the DNA marker systems for this purpose (NRC Report II, 1996a).

In 1985 in the UK the Forensic Science Service (FSS) started building up a STR database which is one of the largest DNA databases and is envisaged eventually to contain 5 million profiles (Gill, P. *et al.,* 1996). Such national DNA databases have also been established in the USA, Netherlands, Austria, Germany, France, Switzerland and Denmark after necessary legislation was enforced in these countries and are also under preparation in various other countries like Belgium, Norway, Spain and Portugal (Schneider, P.M. & Martin. M.D. 2001). Usually the profiles generated from the samples obtained from the suspects and those convicted of a crime are stored in these databases. The large databases offer an opportunity for a positive identification of a suspect involved in a crime, if his profile matches one in the database.

These multilocus databases also serve as the allele/haplotype frequency database of a particular population. If that is the primary purpose of building up a DNA database it may then require a much smaller number of profiles. Such smaller databases have been established for many populations of the world including, Portugese, African, Chinese, Austrian, Hungarian, German, Egyptian, Spanish, Filipino, Japanese and Israeli (Santos, S. M. *et al.,* 1996; Budowle, B. *et al.,* 1997; Lee, J. C. *et al.,* 1997; Klintschhar, M. *et al.,* 1999a; Furedi, S. *et al.,*

1997; Kupferschmid, T. D. *et al.,* 1999; Klintschar, M. *et al.,* 1999b; Cabrero, C. *et al.*, 1995; Entrala, C. *et al.*, 1998; Halos, S. C. *et al.,* 1999; Yamamoto, T. *et al.,* 1999; Amar, A. *et al.,* 1999). The size of these databases ranges from one hundred to several hundred profiles.

3.2 RATIONALE OF COLLECTION OF SAMPLES

To date, the genetic structure of Pakistani population has not been studied for STR markers and there are no published databases designed for forensic identification for the Pakistani populations for autosomal or Y chromosome STRs. The FSS UK maintains an autosomal STR database of the Asian Indians, which includes the population from Pakistan, however it does not distinguish Pakistani population from Indian. Further, according to a rough estimate around 95% of the expatriate population from Pakistan is Punjabi in origin.

A large number of expatriates from Pakistan actually belong to Kashmir who settled in UK in 1950's and are a distinct sub-group. The other Punjabi expatriates come from various parts of Punjab. Among them, the majority are from the rural areas of central Punjab. The Punjabi population in Pakistan maintains a strong 'Biradari' system. A 'Biradari' being a group of people belonging to the same caste. In certain areas, the castes are further sub-divided into sub-castes.

Pakistan has an agrarian society and 85% of the population lives in the rural areas. In most of the rural areas of Punjab inter-marriages between two socially disparate, 'Biradaris' is still unacceptable. However, in the urban areas, population intermixing is much more common and the caste system has blurred.

Samples were not obtained from the Pakistani population living in the UK as it was felt that most of the population belonged to particular areas of Punjab and Kashmir. The number of other ethnic groups of Pakistan in UK was not sufficient for sampling. Thus, it was felt appropriate that sample collection be done in Pakistan so that representative samples could be collected from all the major ethnic groups of Pakistan.

3.3 GEOGRAPHY OF PAKISTAN

3.3.1 LOCATION

Figure 1.4.

3.4 ASSESSMENT OF BUCCAL SWAB AND BLOOD SAMPLES

The ease of DNA extraction from the sample and its yield and quality, were to determine the type of the samples to be obtained during the expedition to Pakistan. The quality of DNA extracted using the Puregene® kit (2.3.1) from blood and from the buccal swabs exposed to high temperatures was assessed considering the temperature and transportation of samples during the expedition.

3.4.1. DETERMINATION OF THE EFFECT OF HIGH TEMPERATURE ON DNA EXTRACTION

Three fresh blood samples (300 µl) and three CEP buccal swab samples were obtained from the volunteers in the laboratory. To the buccal swabs 400 µl cell lysis solution from the Puregene® kit was added before incubation. These were incubated at 36 °C in the oven for 5 days and were removed on the 5th day. DNA extraction was performed using the Puregene® kit (2.3.1 & 2.3.4). DNA extractions (5 µl) were electrophoresed on a 1% agarose gel with the standard DNA (2.4). The quality and the quantity of the DNA assessed (2.5).

There was no marked degradation of either the buccal swabs or the blood samples. The method was thus regarded adequate for collection of samples in Pakistan (Figure 3.1).

3.5 SAMPLING DESIGN

The major populations Punjabi, Pushtoon, Sindhi and Baluchi were the priority populations for collecting samples, as they represented the major

Figure 3.1: Agarose gel Electrophoresis of Samples Exposed to High Temperature

From left to right the figure shows

Lane 1: 500 ng Standard DNA

Lane 2: Buccal Swab Sample 1

Lane 3: Buccal Swab Sample 2

Lane 4: Buccal Swab Sample 3

Lane 5: Fresh Blood Sample 1

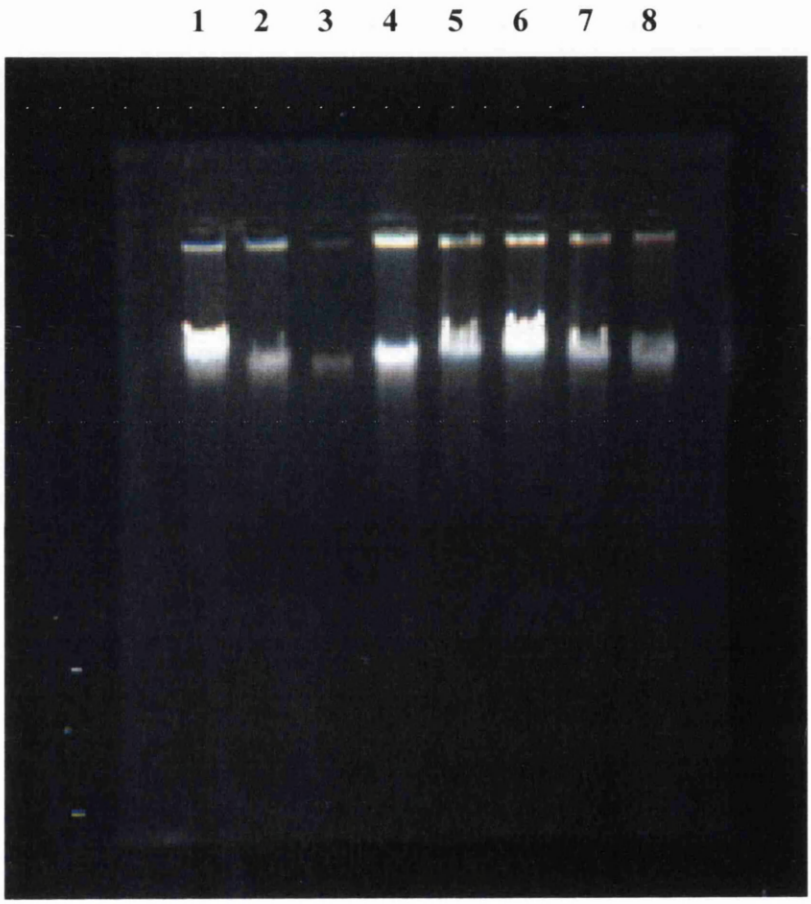
Lane 6: Fresh Blood Sample 2

Lane 7: Fresh Blood Sample 3

Lane 8: 250 ng Standard DNA

To the buccal swabs 400 µl of cell lysis solution from the Puregene® kit was added before incubation.

The buccal swabs and the blood samples were incubated at 36 °C for 5 days and agarose electrophoresis was carried out on a 1% gel.


1 2 3 4 5 6 7 8

proportion of the population. It was felt appropriate to collect samples from various clusters situated at approachable locations and having the necessary infrastructure for collection of samples. It was kept in mind that populations might have to be sampled in field environment.

3.5.1 SAMPLING THE POPULATIONS

It was planned to collect 100 samples from each population though it was kept to the time and opportunity to get as many samples as possible from all approachable populations. At the outset, none of the populations was excluded. In addition to the four major ethnic populations some two smaller populations were also sampled.

The Makrani population, which lives in close proximity with the Sindi and the Baluchi populations. was sampled as it is the largest ethnic sub-group living in Sind. The Brosho population was sampled from the Gilgit District and the Kalash population was sampled from the Chitral District from the northern areas. The last two populations were sampled, as they comprise of distinct ethnic groups living near the borders of Central Asia, China and Afghanistan.

All samples were obtained as randomly as possible. All first and second-degree relatives of the donors were excluded before taking the sample. Informed consent was obtained in writing or verbally from all the volunteers. In case the volunteer was a minor (below 16 yr.), consent was obtained from the parents. Before visiting the areas/institutions, necessary permission was taken from the relevant civil/military authorities.

3.5.2 SAMPLING OF PUNJABI POPULATION

3.5.2.1 Location

The Punjabi population is located in a vast area in the Punjab province of Pakistan (Figure 1.4). This population has further sub divisions into a large number of castes and sub-castes. The major ethnic sub-groups are however Saraiki

in the eastern part of Punjab up to its borders with Sind, and Kashmiri in the north eastern mountainous region up to the Line of Control (Figure 1.4).

3.5.2.2 Collection of Punjabi Samples

The Punjabi samples were collected from the Army Medical College students and Staff. This college is located in Punjab in Rawalpindi, which is a city adjacent to Islamabad. The college has undergraduate medical and dental students on its roll. These students come primarily from all areas of Punjab and NWFP. At the time of sampling, the college had 450 Punjabi and 60 Pushtoon students. 300 unrelated volunteers were selected from the Punjabi students.

3 ml blood was collected by venepuncture using atraumatic size 20 needle from each volunteer. The blood sample was immediately placed in a Vacutainer® EDTA tube (purple top). The tube was rolled gently, to mix the sample with EDTA for 30 s, and labelled. Samples were stored in the refrigerator at 4 °C. The collection of samples was completed in three days. The samples were collected at the Department of Haematology, Army Medical College. All the tubes containing the samples were placed on plastic racks and packed securely with cotton wool, in cardboard boxes. The samples were shipped to UK via courier service.

3.5.3 SAMPLING OF PUSHTOON POPULATION

3.5.3.1 Location

The Pushtoon live in the NWFP province (Figure 1.4). The Pushtoons are a tribal society and there are scores of tribes, some living on both sides of the borders between Pakistan and Afghanistan. The major ethnic subgroup is the Hazara, who speak Hindko language.

3.5.3.2 Collection of Samples

The Pushtoon population was sampled at three clusters. The first was the Army Medical College where 40 samples were collected (3.5.2). The second was the Department of Forensic Medicine, Khyber Medical College located at Peshawar which is the Provincial capital of NWFP. One hundred samples were collected here (3.5.2). The third cluster was the Department of Medicine, Combined Military Hospital Peshawar. Here every third patient that reported in the surgery for general medical examination was asked to volunteer, 30 volunteers donated samples. The samples were collected as stains on 42.5 mm circles of Whatman® filter paper (Whatman® International Ltd, UK) were labelled. 50 µl of the blood sample was applied to a circle of Whatman® filter paper. This was allowed to dry completely in a flow hood. Wearing disposable gloves, each stain was then taped inside a small piece of folded thick paper and was placed in a paper envelope.

3.5.4 SAMPLING OF SINDHI POPULATION

3.5.4.1 Location

Sindhi population occupies the Sind province. Most of the original Sindhis are concentrated in areas other than Karachi, where the Urdu speaking Mohajir population is concentrated along with minor numbers of other populations of Pakistan. This population though is a recognised ethnic group by most, comprises of various ethnic groups who migrated from India at the time of partition. They thus include further distinct subdivisions. Other ethnic subgroups living in Sind are Makrani, Barohi and Parsis.

3.5.4.2 Collection of Samples

The sampling of Sindhi population was done at two clusters. The first was the Sind Regimental Center, Hyderabad in Sind. Hyderabad City is located about 164 km north of Karachi. The Sind Regimental Center is a training center for the Army recruits, which are recruited from Sind. The recruits therefore belong to various areas of Sind and almost all areas of Sindh were represented in this

center. One hundered unrelated recruits volunteered for sampling. Seventy Sindhi recruits were sampled. To get adequate representation of central Sind another cluster was selected. The central Sind comprises of district Sangarh about 150 km north of Hyderabad. The sampling was done at District Hospital Sangarh. Here 50 patients reporting in the out patient department for complaints, other than infectious diseases, volunteered and 30 blood samples were collected. Blood samples were despatched to UK.

3.5.5 SAMPLING OF MAKRANI & BALUCHI POPULATIONS

3.5.5.1 Location

The Baluchi population lives in Baluchistan (Figure 1.4) which is the biggest province of Pakistan. This province has borders with Iran and Afghanistan. The major ethnic subgroups are Barohi and Makrani. The Makrani population lives along the coastal areas of Karachi and Baluchistan, which is also known as the coast of Makran. The city of Karachi is the largest in Pakistan with an estimated population of 12 million. All ethnic sub groups live here. The Makrani and Baluch populations also live here as an ethnic minority due to employment and business concerns. However both the groups maintain a strict tribal system and are largely endogamous. In view of the available time frame time and an obvious reduction in costs, it was decided that instead of travelling to Baluchistan, the two populations might be sampled in Karachi.

3.5.5.2 Collection of Samples

Most of the Baluch are concentrated alongwith Makranis in the Southern District of Karachi City. The Makrani population was sampled in the Qasba Colony, Karachi. All the Baluch and the Makrani households in the area were requested through their chieftains to donate a sample. Only one male from each household was required to donate the sample. Almost all agreed to donate a sample. Samples were obtained from 60 Makrani volunteers. Similarly 40 Baluchi volunteers were sampled.

One hundred 2 ml screw cap tubes were labelled. 400 µl of cell lysis solution included in the Puregene Kit® (Flowgen, Novara Ltd) was aliquoted in all the tubes. The samples were collected as buccal brush scrapping. CEP buccal brush (Life Technologies Inc. USA) was used to scrap the buccal epithelium (2.1.9). Gentle scrapping was done for 30 s on each side of the oral cavity. The brush is designed in a way that the swab itself can be pushed away and detached by pushing the plunger in the handle. The swab was thus placed in a labelled tube. The tube was then securely capped and placed in a plastic box.

3.5.6 SAMPLING OF KALASH POPULATION

3.5.6.1 Location

Kalash inhabit the remote Kalash valleys of the Chitral District bordering Afgahanistan and Central Asia (Figure 3.2). Kalash are a tribal society therefore the chieftains of Kalash were approached through the Assistant Commissioner Chitral, who in turn got the consent from the Kalash household seniors for getting the samples from the people.

3.5.6.2 Collection of Samples in Bumboret Valley

The total population of this valley is about 2500. There are 200 households in Bumboret. In consideration of the field conditions of sampling, donor co-operation, conservation of time and expenses buccal cell samples were obtained from the Kalash population. In Bumboret sampling from unrelated Kalash was done in the villages of Karakal and Brun. Sampling of the other villages (Batrik, Anish and Pehelwanande) was done at the schools of Karakal and Brun, and Anish since these villages were quite spread out (Figure 3.2). Samples (60 males and 10 females) were collected from unrelated persons.

Figure 3.2: A Map of the Kalash Valleys

The Kalash live in the distant valley called Kafirstan in the district Chitral of NWFP located at an elevation of 1128 meters.

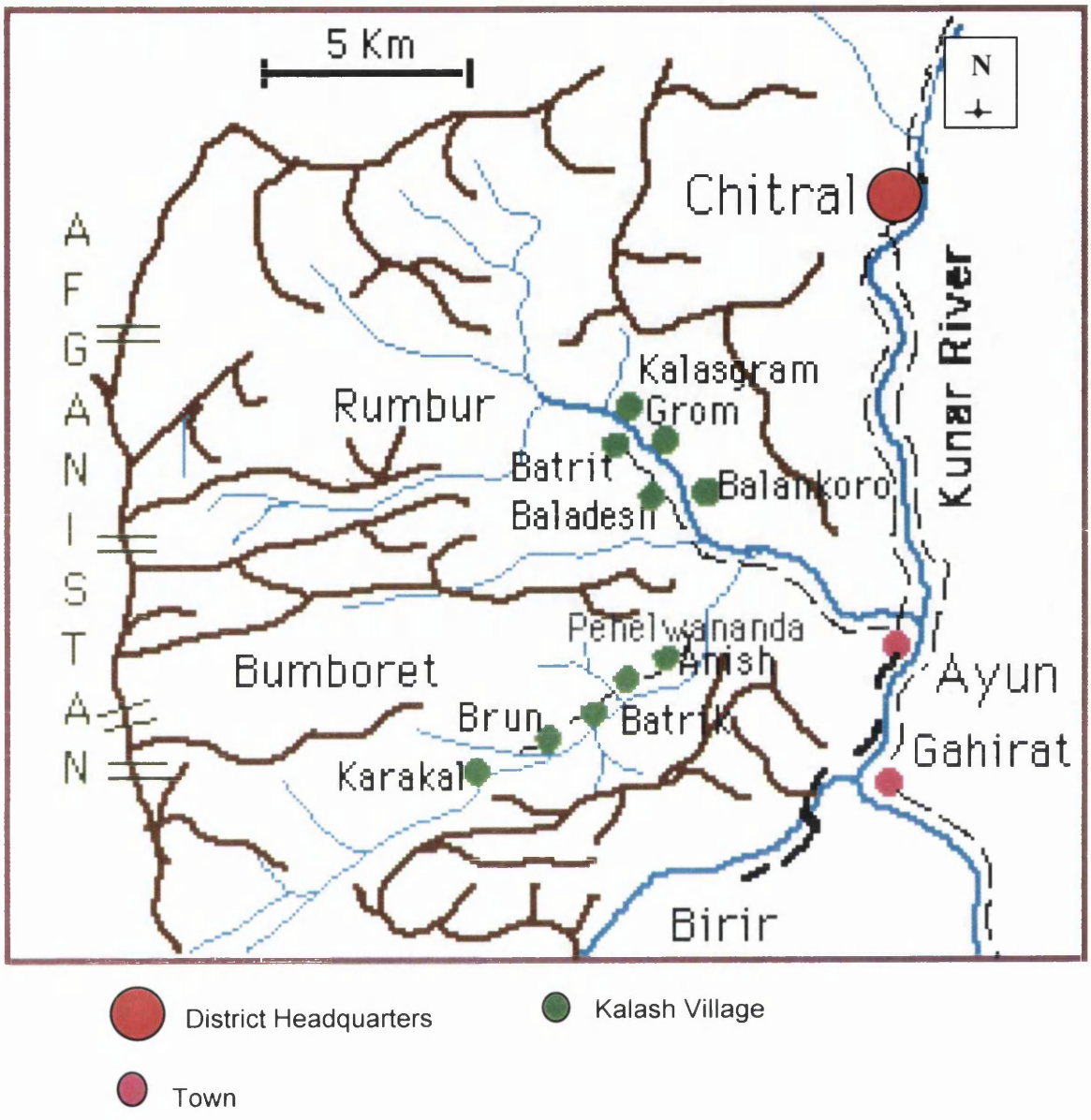
The district is bound by Afganistan on its three sides and joins Pakistan on its east. A narrow strip of Wakhan separates it from the central Asian State of Tajikistan. The Kalash valleys are located right at the border of Afghanistan adjacent to the Wakhan strip. The Kalash live in three valleys, Bumboret, Rumbur and Birir (Figure 3.3). Bumboret is about 40 km, Rumbur is 32 km and Birir is 34 km, from Chitral City. The road to the valleys is metalled as far as the small town of Ayun, about 15 km from Chitral. A non metalled narrow road from Ayun onwards connects the three valleys however the villages in the depths of the valley are approachable on foot only.

Bumboret is the largest of the three valleys. From the entry point of the valley, northwards the villages are Pehalwanande, Anish, Brun, Batrik and Karakal. All villages are situated about 2-5 km apart on the mountainsides.

Rumbur is a deep and narrow valley. The villages in this valley are Kalashgram, Batrit, Baladesh, Balankoro and Grom.

Grom is situated on the other side of the river all the other villages are on the southern edge. Balankoro is the biggest village.

Birir is the most distant valley. The population is the smallest, about 1000 only. This valley was not visited during the expedition due to problems in approaching the valley and time constraints.


5 Km
N
Chitral
Kunar River
AFGANISTAN
Rumbur
Kalasgram
Grom
Batrit
Baladesh
Balamkoro
Penelwananda
Amish
Bumboret
Ayun
Brun
Batrik
Gahirat
Karakal
Birir
District Headquarters
Kalash Village
Town

3.5.6.3 Collection of Samples in Rumbur Valley

The total population of this valley is about 1300. There are 125 households in Rumbur. The villages in this valley are far apart and are situated high up on the mountains, except for the villages of Grom and Balankoro. Buccal cell samples were collected from the unrelated Kalash students at the Secondary School in Grom and the Primary School in Balankoro. Samples from 40 males samples and 10 females were collected.

3.5.7 SAMPLING OF BROSHO POPULATION

3.5.7.1 Location

Brosho inhabit the valley of Hunza about 117 km from Gilgit, at an elevation of 2400 meters. Gilgit is a major transit and trade hub for the Northern Areas and is about 350 km northwest of Islamabad. It is situated in the middle of a large valley, with mountain passes to the west to Chitral and Afghanistan, north to China and Central Asia, east to Skardu, Tibet, and Kashmir (Figure 3.3).

3.5.7.2 Collection of Samples in the Hunza Valley

In the central part of the valley are two secondary Schools and a Health Centre situated at Aliabad and Karimabad (Figure 3.3). The area is totally inhabited by the Brosho people therefore it was convenient to take the samples at these three clusters. The Brosho population was sampled at Karimabad High School, Aliabad High School and Aliabad Rural Health Center. Verbal consent was obtained from the donors and buccal cell samples were collected. Sixty five male samples were collected from the Brosho population.

Table 3.1 & 3.2 summarise the populations, locations, number of samples from each population and the total number of the samples collected during the expedition.

Figure 3.3: A Map of Hunza Valley

Hunza valley is drained by thc Hunza river connecting the Gilgit valley to the south. Settlements range from the base of the valley to many thousand feet up on the mountains.

The red line across the map shows the Karakorum Highway which connects Pakistan with China. The border between China and Pakistan is 85 km from Aliabad. The Brosho population is spread out in a large area. The length of the valley is about 100 km.

The major towns of Hunza Valley are Karimabad and Aliabad, which are about 10 km apart. The major villages are Ganesh, Altit, Wakhi, Chaprot, Atabad and Gulmit. All these villages are located within the central part of the valley.

The green dotted line shows the international boundaries

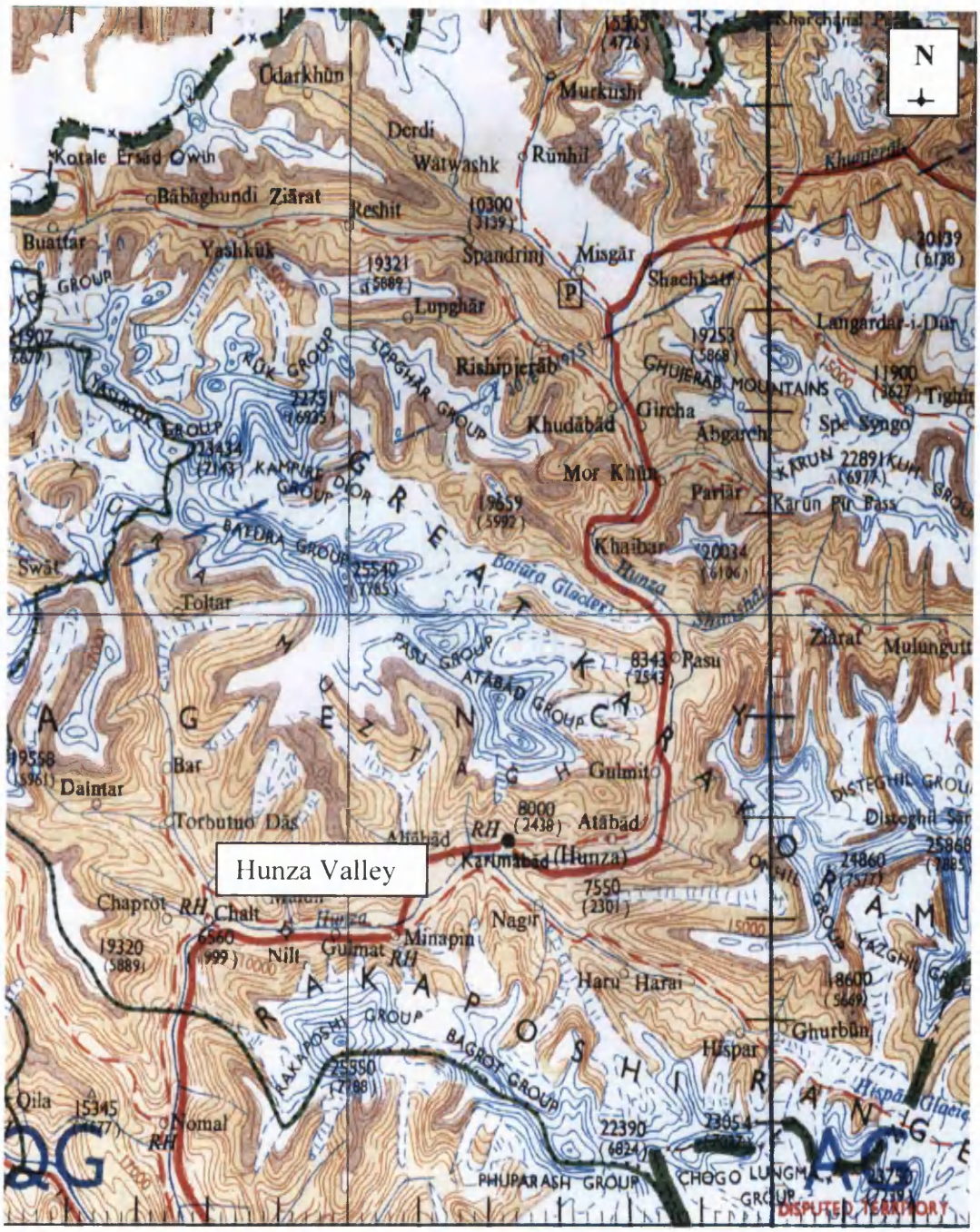

N
Hunza Valley
Udarkhūn
Kotale Ersād Owīn
Bābāghundi Ziārat
Buattar
Yashkūk
Derdi
Wātwashk
Rūnhil
Murkushi
Reshit
Spandrinj
Misgār
Shachkati
Lupghār
Rishipjerāb
Khudābād
Girdha
Ābgarch
Mor Khūn
Pariar
Khaibar
Hunza
Baltāra Glacier
Swāt
Toltar
Bar
Daimar
Torbutuo Dās
Altābād
Karimābād (Hunza)
Atābād
Gulmit
Rasu
Ziārat
Mulungutt
Chaprot
Chalt
Nilt
Gulmat
Minapin
Nagir
Haru Harai
Ghurbūn
Hūspar
Oila
Nomal
Phuparash Group
Chogo Lungma Group
Disputed Territory
Scale: 1 cm = 10 km

Table 3.1: Samples collected during the expedition to Pakistan

Population	Male Samples	Female Samples	Total
Punjabi	130	70	200
Pushtoon	120	50	170
Sindhi	100	None	100
Baluchi	40	None	40
Makrani	65	None	65
Kalash	100	20	120
Brosho	65	None	65
Total	620	120	760

Table 3.2: Method and location of collection of samples

Population	Methodof Collection	Location of Collection
Punjabi	Blood	Rawalpindi
Pushtoon	Blood + Stain	Rawalpindi & Peshawar
Sindhi	Blood	Hyderabad & Sangarh
Baluchi	Buccal Swab	Karachi
Makrani	Buccal Swab	Karachi
Kalash	Buccal Swab	Kalash Valleys, Chitral
Brosho	Buccal Swab	Hunza Valley, Gilgit

3.6 DISCUSSION

3.6.1 SAMPLING PROCEDURES

In order to perform genetic analysis of populations individuals from that population have to be profiled. The precise estimation of the frequency of a profile is possible only if all the individuals making up the population are profiled and used as a reference. Since that is usually impossible task samples are obtained from the populations aimed at developing the profiles.

Thus, collection of representative samples is usually the first step in building up DNA database. The samples for such a database should be obtained from unrelated members of the population taken at random. Random sampling implies that the selection of one sample from the population in no way influences the selection of the next sample. Though simple random sampling is the ideal way to obtain the samples, the prohibitive cost and the time involved for such an exercise make it a difficult choice. Therefore other methods of sampling the populations have been employed including stratified random sampling, cluster sampling, systematic sampling (Aitken C.G.G & Stoney D.A. 1991) and convenience sampling (NRC Report II, 1996b).

Out of these methods convenience sampling is the easiest and most feasible economically. Further, the DNA makers used for forensic and population evolution are located on the non-functional areas of DNA and selection pressures are not detectable thus convenience sampling might be done for the purposes of building up DNA databases (NRC Report II, 1996b). This has now becomes almost a standard practice and as such samples have been obtained from blood banks, hospitals, prisons and random donors for building up DNA database (Alford, R. L. *et al.*, 1994; Ovington, A. *et al.*, 1997; Evett, I. *et al.*, 1997).

3.6.2 METHODS OF COLLECTION

Generally for DNA extraction, the ideal sample had been fresh whole blood as the yield of DNA yield was good (30–50 μg/ml). The initial methods of

DNA extraction from white blood cells were continuously refined (Schmidtke J. *et al.*, 1976; Potter, A. A. *et al.*, 1985; Buffone, G. J. & Darlington, G. J. 1985; Ciulla, T. A. *et al.*, 1988; Mullenbach, R. *et al.*, 1989; Walsh, P. S. *et al.*, 1991). A number of commercial kits have been developed and validated (QIAamp® System, Qiagen Ltd; DNA Stat ™ Blood Kit, Stratagene Ltd; Puregene™ Kit, Gentra Ltd; Wizard® Genomic DNA Purification Kit, Promega Corp). Different laboratories have thus come to settle down for one or the other of these methods.

It was obvious to the scientific community in general and the forensic community in particular that extraction of DNA would be required from a wide variety of samples retrieved from particular case and/or crime scene. Thus very soon methods were developed for the extraction of DNA from other body fluids and excreta containing nucleated cells. The first of these cells targeted for their potential of DNA yield were hair and buccal epithelial cells. Hair and buccal epithelial cells were shown to yield a good quality DNA for PCR (Lench, N. *et al.*, 1988). However, the yield of DNA was low. Improved extraction methods were soon described which could isolate abundant DNA from buccal epithelial cells (Tobal, K. *et al.*, 1989).

These methods depended on mouth washes and subsequent sample centrifugation. The buccal epithelial cell collection was non invasive and the need for medical supervision was obviated. The consent procedures were easier and the donor compliance was greater than one would expect in obtaining a blood sample. Further, this also was an easy option in field environment. However, this still involved liquid sample handling and an inconsistent DNA yield. Collection of buccal epithelial cells with cytology brushes and swabs was described and the technique was immediately adopted for obtaining samples (Richards, B. *et al.*, 1993). The buccal cells were shown to be stable in extreme temperatures. The DNA yield was abundant and the PCR failure rate was lower than the DNA extracted from blood (Richards, B. *et al.*, 1993). Buccal cell samples are now collected routinely for DNA extraction.

Before undertaking the sampling expedition an assessment was performed on the method of extraction of DNA from anticoagulated blood and

buccal cells exposed to high temperature for 5 days (3.4). Both the procedures gave a good result and DNA yield and quality was satisfactory (Figure 3.1). It was thus decided that blood and buccal swabs both could be collected as samples during the expedition. Furthermore during this work in consideration of the fact that a lot of sampling was to be done in field environment it was decided that buccal epithelial cells would be an appropriate sample. The procedure showed its strength during this work. The DNA yield from these samples was of high quality.

Buccal cell sampling definitely enhanced the donor compliance, especially when working with populations in rural set up, where obtaining blood sample may have caused fear or non compliance. However wherever the infrastructure and environment permitted whole blood samples were collected.

Collection and despatch of samples was done in the light of the guidelines of the National Heart, Lung and Blood Institute, USA (Austin, A. M. 1996). The samples were despatched in special insulated boxes with cotton wool packed around the box containing the samples in order to keep the samples at an even temperature during transportation. The main advantage of taking the blood samples was greater quantity of the DNA from one sample.

3.6.3 COLLECTION OF SAMPLES

Sampling of the Baluch and the Makrani populations was done in Karachi, (3.5.5). These two groups have now been in Karachi for two or three generations but still maintain well knit communities and are largely endogamous. It was fortunate that both the populations were located in the same general area. Ancestry of each individual was established before obtaining a sample, though it took time. That was usually done by inteviewing the donor and the senior family members. The Sindhi population was sampled at two locations near Hyderabad City. The tribes from Baluchistan like the Barohis have settled in Sind. Therefore, care was taken to select donors who were Sindhis by descent, at both the locations. Sindhi people were found very well informed regarding their ancestry, which helped in sampling the population.

In a tribal society, it was found that contacting the tribal chiefs through the government officials was appropriate. Kalash were sampled in only two valleys. The samples were obtained from the donors in schools or houses. In the Hunza valley the experience gained in the Kalash valleys was applied. Samples were collected from the donors in a rural health center and two high schools.

Though sampling was done in a very intensive manner, the distances involved initial preparations to reach an area and then approaching the populations and took a lot of time. A total of 760 samples were collected during this expedition whic is the largest population sample collected from Pakistan (Table 3.1). Though the number of the samples from the Baluchi population was low still it would allow for having the genotype and haplotype frequency estimates for the Pakistani population as a whole and allow comparisons to be made with the other populations.

CHAPTER 4: AUTOSOMAL STR LOCI PCR OPTIMISATION & GENOTYPING

4.1 INTRODUCTION

Analysis of Short Tandem Repeats (STRs) has become the method of choice for forensic identification for a number of reasons. These loci are amenable to PCR allowing the analysis of small and degraded samples, allelic designation is discreet, they are highly polymorphic and they coamplify with relative ease (Urqhart, A. *et al.*, 1994; Lygo, J. E. *et al.*, 1994; Fregeau, C. J. *et al.*, 1993; Alford, R. L. *et al.*, 1994).

The indigenous Pakistani populations have a high level of consanguineous marriages and the populations are socially structured therefore it was important to study the genetic structure of these populations before any STR markers could be utilised for forensic purposes.

4.1.1 SELECTION OF LOCI

It was prudent to study the populations for robust and well characterised systems for which sufficient data was available from other populations to make comparisons. A large number of STRs have been described and validated for the forensic identification purposes (Polymeropoulos, M. H., 1991 a & b; Hearne, C. M. & Todd, J. A. 1991; Anker, R. 1992; Nishimura, D. Y. 1992; Mills, K. A. 1992; Kimpton, C. P. 1992; Kimpton, C. P. *et al.*, 1993; Hammond, H. A. *et al.*, 1994; Urqhart, A. *et al.*, 1995; Sprecher, C. J. *et al.*, 1996).

Loci D3S1358, vWA and FGA were selected for this study. A multiplex kit AmpFℓSTR® Blue Kit developed by Applied Biosystems CA, USA had been validated for forensic use (Wallin, J. M. *et al.,* 1998), and it was evident that the loci had a high power of discrimination. The combined Probability of

Identity (PI_{COMB}) of these three loci was found to be 1/5479 for Caucasians, 1/4830 for African-Americans and 1/3443 for Hispanics. The markers were thus adjudged to be extremely useful for forensic identification purposes (Wallin, J. M. *et al.,* 1998). All the three loci were included in the Combined DNA Index System (CODIS), FBI, USA database as well in a set of 13 STR loci (Budowle, B. *et al.*, 1999; Scherzinger, C. A. *et al.*, 2000).

Out of these three loci, vWA and FGA had been included in the Second Generation Multiplex (SGM) (Urqhart, A. *et al.*, 1995; Oldroyd, N. J. *et al.*, 1995; Sparkes, R. *et al.*, 1996 a & b), which was used by the Forensic Science Service (FSS), UK for building up a population database. The three loci had been thoroughly studied for somatic stability, species specificity, analysis of old/degraded samples and mixture interpretation (Wallin, J. M. *et al.,* 1998).

4.1.2 DESCRIPTION OF SELECTED LOCI

4.1.2.1 Locus D3S1358

D3S1358 is a compound tetranucleotide locus on the short arm of chromosome 3 (Li. H. *et al.* 1993). It has two repeat motifs $(TCTG)_{2-3}$ and $(TCTA)_n$. This locus can be amplified in highly degraded samples because of its relatively small size (Santos, S. M. *et al.*, 1996). During the validation studies of the AmpFℓSTR® Blue kit (Applied Biosystems CA, USA), it was shown that D3S1358 can amplify when vWA and FGA fail to amplify in severely degraded samples (Wallin, J. M. *et al.,* 1997).

Twelve alleles have been described for this STR and it has a high heterozygosity. The locus was included in the commercial multiplex kit AmpFℓSTR® Blue and remained a part of AmpFℓSTR® CoFiler™, AmpFℓSTR® Profiler Plus™, and AmpFℓSTR® SGM Plus™ (Applied Biosystems CA, USA). The sequence of the locus is shown in Figure 4.1a.

Figure 4.1a : Sequence of Locus D3S1358

The sequence has been obtained from the STRbase- STR DNA Internet Database

Figure 4.1b: Sequence of Locus vWA (Genome Bank Database)

LOCUS: HUMvWF A31.
DEFINITION: Human von Willebrand factor gene, inton 40
GenBank ACCESSION No: M25858 M25716

Figure 4.1c: Sequence of Locus FGA (Genome Bank Database)

LOCUS: HUMFIBRA
DEFINITION: Human fibrinogen alpha chain gene
GenBank ACCESSION No: M64982

The figures show the two primers used during this study in bold.
The numbers besides the sequence show the nucleotide count.

Locus D3S1358

```
  5'------------------►3'
  1 actgcagtcc aatctgggtg acagagcaag accctgtctc atagatagat
 51 agatagatag atagatagat agatagatag atagatagac agacagatag
          5'◄------------------3 '
101 atacatg caagcctctg ttgatttcat
```

Locus vWA

```
                          5'----------------------►3'
1621 cttggctgag atgtgaaagc cctagtggat gataagaata atcagtatgt
1672 gacttggatt gatctatctg tctgtctgtc tgtctatcta tctatctatc
1713 tatctatcta tctatctatc tatctatcta tctatccatc tatccatcc
  5'◄--------------------------3'
1764 atcctatgtat ttatcatctg tcctatctct
```

Locus FGA

```
          5'----------------------►3'
2821 aaatgcccca taggttttga actcacagat taaactgtaa ccaaaataaa
2872 attaggcata tttacaagct agtttctttc tttctttttt ctctttcttt
2923 ctttctttct ttctttcttt ctttctttct ttctttcttt ctttctcctt
                                                  5'◄----
2974 ccttcctttc ttcctttctt ttttgctggc aattacagac aaatcactca
     ------------------3'
3025 gcagctactt caataaccat attttcgatt tcagaccgtg
```

4.1.2.2 Locus Hum vWA F31A (vWA)

This STR locus is located in the intron 40 between nucleotide (nt) 1640–1750 of the von Willebrand gene, on chromosome 12 (Kimpton, C. *et al.,* 1992). vWA was one of the first STR loci to be used for forensic identification purposes and the locus was validated and included in the quadraplex kit developed by the FSS (Lygo, J. E. *et al.*, 1994).

Various European laboratories during the second EDNAP collaborative exercise tested this locus. The locus was shown to be robust and was reproducibly typed on different electrophoretic detection systems (Kimptom, C. *et al.,* 1995). The locus was included in the SGM kit by the FSS and co-amplified with seven other loci, had a high heterozygosity and a low match probability (Oldroyd, N. J. *et al.,* 1995). It has thus been included in almost all the commercial human identity kits, including the AmpFℓSTR® Cofiler™, AmpFℓSTR® Blue, AmpFℓSTR® SGM Plus™ (Applied Biosystems CA, USA), PowerPlex™ and PowerPlex™ 2.1(Promega Corp, Madison, USA). vWA is a compound repeat with non-consensus repeat alleles (Urqhart, A., 1994). The repeat motif is $(TCTR)_n$, where R is either A or G (Moller , A. *et al.,*1994).

There is micro heterogeneity in the structure of these alleles and due to mutational events there can be a sequence variation in two alleles of the same size (Barber, M. D. *et al.,* 1995). This does not have any effect on the designation of the alleles themselves, however if the alleles are sequenced any two samples having same alleles (in terms of size), may differ, thus in certain cases it may be worthwhile to sequence the PCR products. This then serves to add to the utility of the locus for forensic identification purposes and at the same time this calls for a cautious approach while interpreting the vWA genotypes in case work.

Two reported sets of primers have been used to amplify the locus (Urqhart, A. *et al.,* 1995 & Moller, A. *et al.,* 1994). Commercial firms have also developed their own primer sets (Applied Biosystems CA, USA & Promega Corp, Madison, USA). 16 alleles and few variants have been described, ranging from 122 to 182 bp with primer set of Urqhart, A. *et al.* 1995 and 152 to 212 bp with primer set

of the Applied Biosystems. The sequence of the locus is shown in Figure 4.1b.

4.1.2.3 Locus Hum FIBRA (FGA)

FGA or Human Alpha Fibrinogen locus is located in the third intron of the alpha fibrinogen gene on chromosome 4q28 (Mills, K. A. *et al.*, 1992). A number of subsequent studies also described FGA as a highly polymorphic locus (Urqhart, A. *et al.*, 1994 & Ovington, A. *et al.*, 1997).

The locus has been described as a complex repeat with several different types of the repeat sequences. Besides the complete tetranucleotide repeats in the alleles, the alleles have variants as well which are 2 bp apart. It has been studied in a number of populations and has shown to possess a high heterozygosity and a high power of discrimination (Rolf, B. *et al.*, 1998). FGA also exhibits a high exclusion probability (70-75%) and has thus become a powerful tool for forensic identification and paternity testing (Rolf, B. *et al.*, 1998).

The structure of the repeat region in different alleles is $[TTTC]_3$ TTTTTTCT $[CTTT]_nCTCC$ $[TTCC]_2$. The structure of the alleles shows micro heterogeneity (Barber, M. D., 1996). The locus possesses 30 known alleles including the variants. Inclusion of this locus in any profiling system has the advantage of significant enhancement in the discrimination power of the system.

It was first included in the Second Generation Multiplex (SGM) (Urqhart, A. *et al.*, 1995) and has been a part of the commercial STR kits like AmpFℓSTR® Blue, and AmpFℓSTR® SGM Plus™ (Applied Biosystems CA, USA), PowerPlex™ (Promega Corp, Madison, USA). The sequence of the locus (GenBank) is shown in Figure 4.1c.

4.2 OPTIMISATION OF PCR REACTIONS

The amplification efficiency of any STR system has to be determined empirically in order to reduce the amplification of non-specific products, primer

dimers and to avoid weak allelic signal (Kimpton, C. P. *et al.*, 1996). In addition for any PCR it is essential to achieve a high specificity of the reaction and a high yield of the product. During this work new primers were used for all the three loci therefore optimal conditions had to be established before genotyping the samples.

During the optimisation process experiments were done in order to see the effect of various PCR reagent concentrations and thermal cycling parameters on the efficiency of PCR. Reactions were run on Techne PHC II DNA thermal cycler (Techne, Hybaid Omnigene) and were analysed initially using 1% agarose gel electrophoresis. After initial experiments the reactions were run on ABI 373 DNA analyser in order to assess the optimised conditions.

4.2.1 OPTIMISATION OF PCR IN SINGLEPLEX REACTIONS

Optimisation of PCR was performed in a 25 µl reaction. The template DNA was quantitated and diluted to ~5 ng /µl (2.5). ~10 ng template DNA was used in the reaction.

4.2.1.1 Optimisation of Magnesium Chloride Concentration

$MgCl_2$ solution concentration was tested at three different concentrations 1.5 mM, 2.0 mM & 2.5 mM.

4.2.1.2 Optimisation of Annealing Temperature

The annealing temperature was varied between 56 °C to 64 °C in 2 °C increments keeping the $MgCl_2$ at 2.0 mM.

4.2.1.3 Optimisation of *Taq* Polymerase Enzyme Amount

The *Taq* polymerase enzyme was tested at 0.5, 0.625 and 1 U. All the other components of the reaction were kept at their default values.

4.2.1.4 DNA Template Amount

The DNA quantity that gave best results was assessed by varying the DNA amount from ~5 ng to ~50 ng.

4.2.1.5 Optimised Conditions for Single locus Reactions

The optimal primer and $MgCl_2$ solution concentrations as well as the annealing temperatures of the three loci were similar when amplified as single locus (Table 4.1a & b). The *Taq* amount of 0.625 U was found to be optimal.

4.2.2 OPTIMISATION OF D3S1358 AND FGA DUPLEX REACTION

As the concentration of various components and the annealing temperature of the PCR appeared to be the same for these two loci it was decided to co amplify them using same conditions.

4.2.2.1 Optimisation of Primer concentrations

The concentrations for the primers of D3S1358 locus were kept at 0.1 μM and those of FGA at 0.2 μM and then the reverse was tested.

4.2.2.2 $MgCl_2$ concentration and Annealing Temperature

PCRs were set up using $MgCl_2$ solution at 1.5, 2 and 2.5 mM concentrations. These were tested at various annealing temperatures from 55-64 °C in 2 °C increments.

4.2.2.3 Optimised Conditions for D3S1358 and FGA Duplex Reaction

The optimisation of the duplex reaction was guided by the singleplex reaction conditions. Initially the singleplex primer concentrations and the PCR thermal cycling protocol were used during the optimisation of the duplex.

Table 4.1a: Optimised PCR Ingredients for Single Locus D3S1358, vWA & FGA Amplification

Component	Final Concentration
dNTP	200 μM each
Primer (F)	0.2 μM
Primer (R)	0.2 μM
Buffer	10 x
$MgCl_2$ Sol 25mM	2.0 mM
Taq Polymerase Enzyme	0.625 U
DNA Sol	5-10 ng

Table 4.1b: Optimised PCR Thermal Cycling Parameters for Single Locus D3S1358, vWA & FGA Amplification

Cycle	Temp °C	Time min	No of Cycles
Initial Denaturation	95	5/11 *	1
Denaturation	94	1	30
Annealing	60	1	
Extension	72	1	
Final Extension	65	10	1

* 5 min for Promega *Taq*, 11 min for Ampli*Taq*® Gold

The primer concentrations of both loci were reduced to 0.1 mM in order to obtain a band of similar intensity for both D3S1358 and FGA. Lower primer concentrations of D3S1358 or FGA did not give a good result. The primer concentration of D3S1358 was then increased to 0.2 mM and that of FGA kept at 0.1 mM. This resulted in a good yield for both the loci.

2.5 mM $MgCl_2$ gave better results during the optimisation than lower concentrations for the duplex (Figure 4.2a). The dNTPs concentration was kept at 200 μM each. For the duplex the DNA input was also adjusted (Figure 4.2b). The duplex required a smaller amount to DNA than the singleplex reactions to give similar intensity of amplification of the products and bands could be visualised in the post PCR agarose gel with ~4 ng of template DNA. The optimal PCR ingredients and PCR protocols for the duplex reactions were thus established (Table 4.2a & b).

4.2.3 AMPLIFICATION OF THE THREE LOCI IN MULTIPLEX REACTION

4.2.3.1 Triplex PCR

0.2 μM primers for the locus D3S1358 and 0.1 μM primers of vWA and FGA loci were used for the multiplex PCR. The enzyme concentration was 0.625 U per reaction and the concentration of dNTPs was 200 μM each. Reactions were set up at three $MgCl_2$ concentrations at 1.5 mM, 2.0 mM and 2.5 mM. These were tested for a range of annealing temperatures from 52–58 °C.

4.2.3.2 Triplex PCR Results

It was determined that though all the three loci coamplify, the allelic range for vWA overlaps that of D3S1358. Thus triplex reactions were not feasible and the loci D3S1358 and FGA in a duplex reaction amplified. The locus vWA was amplified in singleplex reactions.

Table 4.2a: Optimal Ingredients for Locus D3S1358 & FGA Duplex PCR

Component	Concentration
dNTP	200 μM each
D3S1358 Primer (F)	0.2 μM
D3S1358 Primer (R)	0.2 μM
FGA Primer (F)	0.1 μM
FGA Primer (R)	0.1 μM
Buffer	10 x
$MgCl_2$ Sol 25mM	2.5 mM
Taq Polymerase Enzyme	0.625 U
DNA Sol	5-10 ng

Table 4.2b: Optimised Thermal Cycling Protocol for Locus D3S1358 & FGA Duplex PCR

Cycle	Temp °C	Time (min)	No of Cycles
Initial Denaturation	95	5/11 *	1
Denaturation	95	1	30
Annealing	58	1	
Extension	72	1	
Final Extension	65	10	1

* 5 min for Promega *Taq*, 11 min for Ampli*Taq*® Gold

Figure 4.2a: Optimisation of the $MgCl_2$ Concentration for Duplex PCR of the Loci D3S1358 and FGA

Annealing Temperature 58 °C

1-3: $MgCl_2$ 1.5 mM

4-6: $MgCl_2$ 2.0 mM

7-9: $MgCl_2$ 2.5 mM

Figure 4.2b :Optimisation of DNA template Amount for Duplex PCR of the Loci D3S1358 and FGA

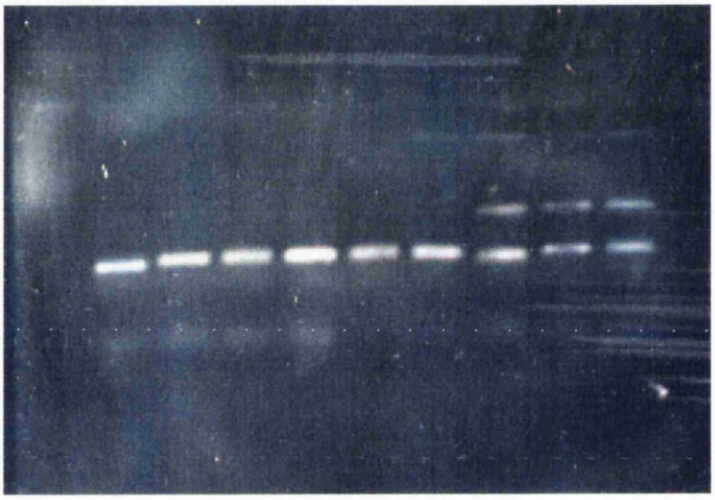
Annealing Temperature 58 °C, $MgCl_2$ 2.5

1-2 : Template 5 ng

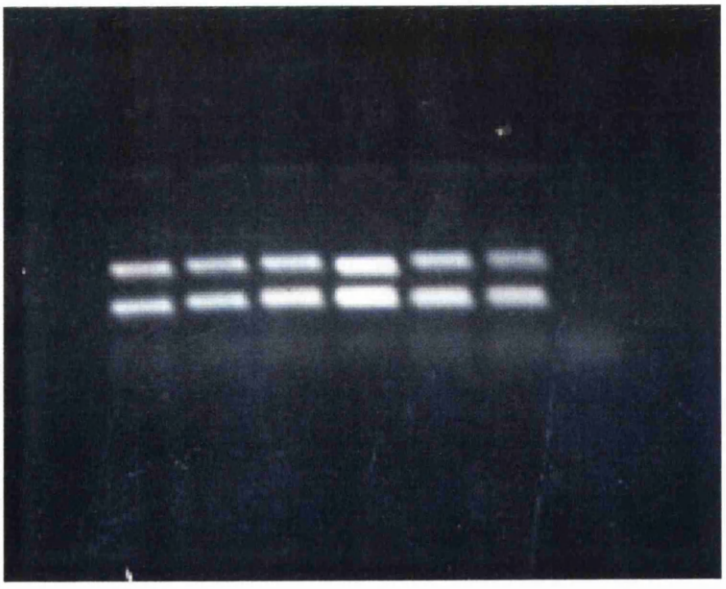
3-4: Template 10 ng

5-6: Template 20 ng

1 2 3 4 5 6 7 8 9



1 2 3 4 5 6



4.3 GENERATION OF ALLELIC SIZE DATA

The GenScan® software automatically sizes the alleles in comparison to an internal lane standard, which is included in each sample before loading the sample on the ABI 373 Automated DNA Sequencer gel, eliminating the inter gel variation in the allelic sizes to a great extent (Kimpton, C. P. *et al.*, 1993; Watts, D. 1998).

The software generates a calibration curve by analysing the internal lane standard, which results in a very precise size for each allele (Freageu, C. F. & Fourney, M. 1993).

4.3.1 PREPARATION OF ALLELEIC LADDER FOR AUTOSOMAL LOCI

It was recognised at an early stage in the development of this technology that for accurate sizing of an allele, direct comparison of the allele with a sequenced fragment would be desirable (Freageu, C. F. & Fourney, M. 1993).

This could be achieved by preparing an allelic ladder which is a cocktail containing the common alleles of a STR system (Peurs. C. *et al.*, 1994). All the alleles in the samples could thus be compared and sized against a known allele.

For this project two strategies were available to size the alleles. One was isolating and sequencing the different alleles, which would have entailed additional resources and time.

The other option was to calibrate the allele size by amplifying some samples with a commercial kit and preparing an allelic ladder using them. This method was used to prepare allelic ladders` and the alleles amplified with new primers were calibrated using the AmpF*ℓ*STR® Blue Kit (Applied Biosystems CA, USA) (Table 4.3 & Figure 4.3a, b).

Table 4.3: Calibration of Allelic Size of the Ladder Alleles for Loci D3S1358, vWA & FGA

Sample	Allelic Size (New Primer)		Allelic Size (Blue Kit)		Allelic Designation
	Locus D3S1358 Allelic Data				
1	127.88	127.86	123.35	123.57	15,15
2	135.98	139.90	131.58	135.88	17,18
3	131.92	135.87	127.50	131.91	16,17
4	131.86	131.92	127.55	127.55	16,16
5	135.88	135.97	131.69	131.72	17,17
6	119.93	127.92	115.50	123.36	13,15
7	131.89	139.91	127.52	135.86	16,17
	Locus vWA Allelic Data				
1	139.97	147.97	167.46	175.55	14,16
2	148.01	165.72	175.54	191.34	16,20
3	143.92	155.89	171.34	183.46	15,18
4	151.92	179.95	179.38	187.12	17,19
5	147.93	152.05	175.58	179.42	16,17
6	143.89	159.88	171.40	187.17	15,19
	Locus FGA Allelic Data				
1	263.97	272.05	224.49	228.56	20,21
2	272.12	288.19	228.54	244.78	21,25
3	263.76	284.01	220.56	240.71	19,24
4	268.05	283.96	224.45	240.67	20,24
5	279.90	295.82	236.65	252.70	23,27
6	260.10	299.77	216.45	256.90	18,28
7	288.09	291.81	244.72	248.80	25,26
8	263.87	283.92	220.52	240.67	19,24

Figure 4.3a: Allelic Ladder for the locus D3S1358

The allelic ladder for the locus D3S1358 contained 6 alleles from 14 to 19 (Table 4.3).

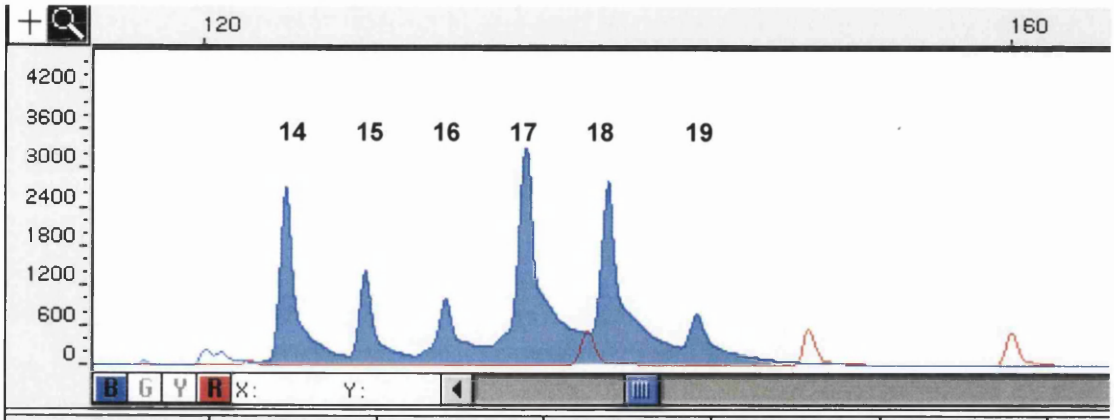
Figure 4.3b: Allelic Ladder for the Locus vWA

The ladder for the locus vWA contained nine common alleles, 14 to 22 (Table 4.3).

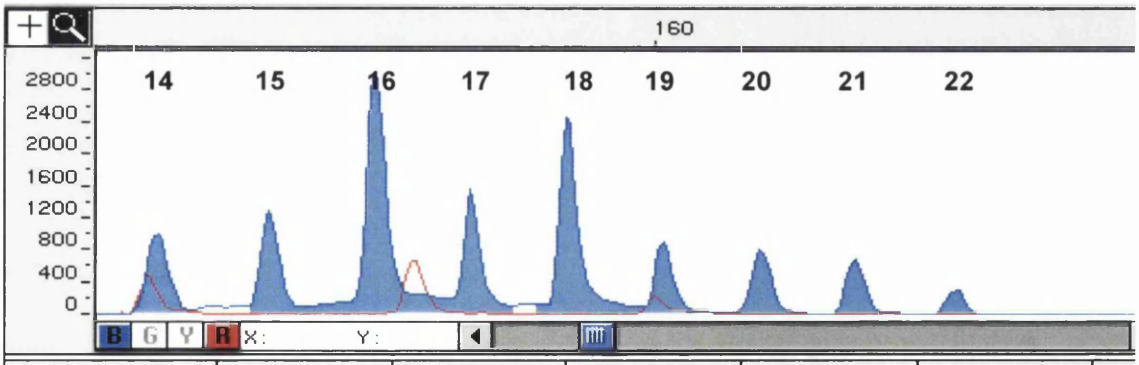
Figure 4.3c: Allelic Ladder for the Locus FGA

The FGA ladder included the common alleles 18 to 29 (Table 4.3).

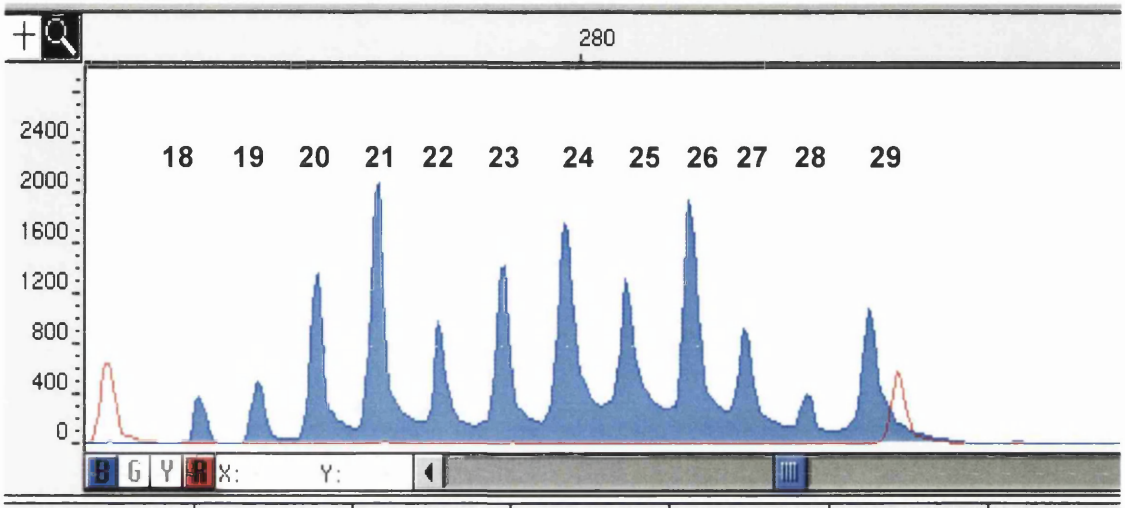
The allelic peaks in all figures are blue as the primers were labelled with 5′ FAM flourescent dye. Peaks red in colour show the GeneScan® 500 [Rox] internal standard peaks.



Dye/Sample Peak	Minutes	Size	Peak Height	Peak Area	Data Point
B, 4	100.90	123.85	2739	25065	1009
B, 5	103.10	127.58	1456	13931	1031
B, 6	105.50	131.65	1018	13825	1055
B, 7	107.70	136.41	3330	48437	1077
B, 8	110.20	139.77	2820	34645	1102
B, 9	112.20	143.95	787	13660	1200



Dye/Sample Peak	Minutes	Size	Peak Height	Peak Area	Data Point
B, 6	109.40	139.39	1045	5868	1094
B, 9	111.80	144.00	1378	6691	1118
B, 10	114.10	148.33	3164	19969	1141
B, 11	116.40	152.37	1593	8455	1164
B, 12	118.80	156.32	2626	15405	1188
B, 13	121.30	160.32	924	5802	1213
B, 14	123.80	164.31	838	5318	1238
B, 15	126.20	168.13	697	4352	1262
B, 16	128.60	172.25	321	1782	1288



Dye/Sample Peak	Minutes	Size	Peak Height	Peak Area	Data Point
3B, 16	256.30	256.01	439	3547	2563
3B, 19	260.00	259.91	511	4149	2600
3B, 20	263.60	263.70	960	7913	2636
3B, 21	267.30	267.59	820	6969	2673
3B, 23	271.48	271.48	1150	10044	2710
3B, 24	274.70	275.37	811	6979	2747
3B, 25	278.40	279.25	1442	13297	2784
3B, 26	282.10	283.14	367	2725	2821
3B, 27	285.10	287.02	1198	10442	2858
3B, 28	289.50	290.89	492	4176	2895
3B, 29	293.20	294.77	313	2368	2832
3B, 30	296.80	298.54	977	9063	2968

4.3.2 ALLELE DESIGNATION OF THE THREE LOCI

It has been suggested that a sample allele should be called within ±3 Standard deviation (SD) or 0.5 bp of the nearest ladder allele so that the allele designation is precise (Gill. P. *et al.,* 1996). SD of an allele is the deviation of the size generated by the automated sequencer on different gels for ABI 373/377 or different injections for ABI 310. In order to estimate the SD the amplified samples for the ladder alleles were analysed on three gels and the deviation of the size for each allele noted across the gels. The SD for all the alleles remained < 0.16 bp, which meant that the sizing of the alleles was sufficiently precise (Table 4.4). The size of the ladder alleles detected using the GeneScan® for the samples run on the automated sequencer was found to be sufficiently precise for accurate allelic designation (Figure 4.4).

4.3.3 PROFILING OF THE SAMPLES

All the samples for the Punjabi, Pushtoon and Sindhi populations were amplified using the 'inlab' primers according to the optimal PCR conditions (Table 4.1 & 4.2) and were analysed on ABI 373 XL UPGRADE under optimised conditions (2.9). AmpFℓSTR® Blue kit was being used for case work at the Human Identification Laboratory of the University of Glasgow and the kit was available so the Baluchi, Makrani, Brosho and Kalash population samples were amplified using the AmpFℓSTR® Blue Kit and were genotyped on ABI 310.

An allelic peak was identified as a peak within the allelic range of the locus, which was distinct and rose above the background noise. Using the allelic ladder for the individual loci the alleles were designated precisely (Figure 4.4). On analysis of the sample, if the signal was weak and the peak size was small, the PCR was repeated with increased template. On the other hand non specific amplification peaks observed in a small number of samples were always the result of too much template in the reaction. Other peaks observed within the allelic range of each locus which were one repeat smaller in size than the main peak and were about $1/10^{th}$ (or smaller), than the allelic peak.

Table 4.4: Allelic Size Data of Ladder Alleles for the Loci D3S1358, vWA and FGA

Locus	**D3S1358**		**vWA**		**FGA**	
Allele	Mean Size[1]	SD[2]	Mean Size	SD	Mean Size	SD
14	123.87	0.06	139.98	0.09		
15	127.9	0.08	144.00	0.15		
16	131.83	0.11	148.03	0.11		
17	135.98	0.09	152.07	0.15		
18	139.95	0.11	155.87	0.06	255.98	0.09
19	140.10	0.15	160.01	0.11	260.02	0.14
20			163.86	0.05	263.92	0.11
21			168.02	0.14	267.96	0.08
22			172.20	0.08	272.01	0.11
23					275.99	0.09
24					279.55	0.11
25					283.74	0.06
26					287.15	0.06
27					291.06	0.05
28					295.12	0.07
29					298.79	0.15

1. Mean Size is the mean of the size of an allele on three gels
2. SD is the standard deviation of the size of an allele across three gels

Figure 4.4: Allelic Designation of Locus FGA Alleles Using the Allelic Ladder

The figure shows the designation of alleles using the allelic ladder. The samples were amplified according to the optimised PCR conditions (Table 4.2 a, b).

The allelic peaks are in blue colour and the GeneScan® 500 [Rox] internal standard peaks are red in colour.

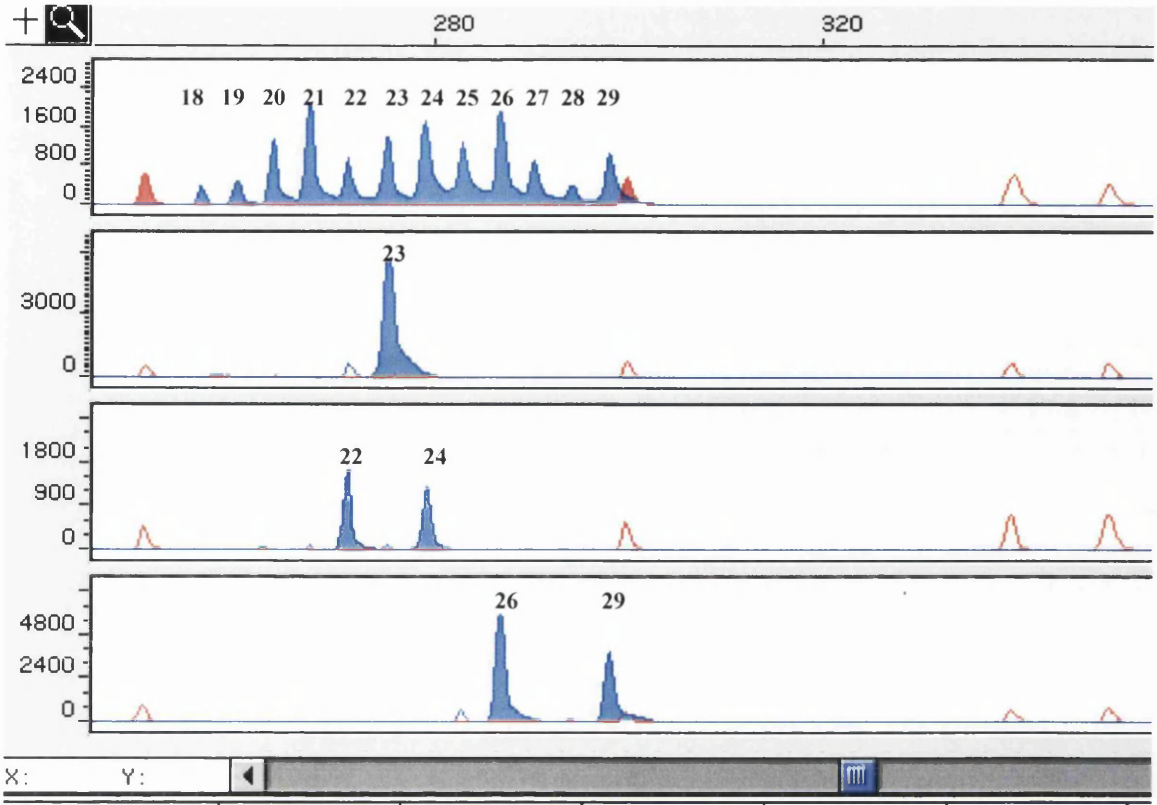
The first panel shows the allelic aldder for the locus FGA containing 12 alleles from allele 18-29 (Peaks 31B, 5-16).

The second panel shows a single allelic peak from a homozygote individual with a size of 275.45 bp (Peak 9B, 20). The nearest peak in the allelic ladder was of allele 23.

The second panel shows two allelic peaks sized as 271.14 bp & 279.32 bp (Peak 27B, 4 & 27B, 9). The nearest peaks in the allelic ladder were those of alleles 22 and 24.

The third panel shows two allelic peaks of 287.12 bp & 298.35 bp (Peak 29B, 3 & 29B, 5). The nearest peaks in the ladder were for allele 26 & 29.

All peaks showed a size difference of <0.05 bp to the size of the nearest allele in the ladder and were therefore designated as same allele.



Dye/Sample Peak	Minutes	Size	Peak Height	Peak Area	Data Point
31B, 5	181.40	255.81	399	2464	1814
31B, 6	183.90	259.54	506	3981	1839
31B, 7	186.40	263.27	1399	11231	1864
31B, 8	189.00	267.14	2134	17958	1890
31B, 9	191.60	271.02	1000	8882	1916
31B, 10	194.30	275.05	1442	13668	1943
31B, 11	197.00	279.08	1785	18861	1970
31B, 12	199.50	282.81	1340	16004	1995
31B, 13	202.20	286.84	1966	19094	2022
31B, 14	204.50	290.28	942	10582	2045
31B, 15	207.20	294.32	418	3672	2072
31B, 16	209.80	298.20	1121	13459	2098
31R, 8	177.50	250.00	674	5608	1775
31R, 9	211.00	300.00	582	4850	2110
9B, 20	196.00	275.45	5841	71969	1960
27B, 4	192.00	271.14	1738	12108	1920
27B, 9	197.50	279.32	1353	10350	1975
29B, 3	202.40	287.12	6004	58128	2024
29B, 5	209.90	298.35	4007	38808	2099

These were recognised as stutter peaks and were observed at all the loci but more commonly at locus vWA. The phenomenon of stuttering has been described at the vWA locus (Walsh, P. S. *et al.*, 1996) and has been largely characterised. It was easy to recognise the stutter in the single locus amplification. Decreasing the template in such cases reduced the incidence of stutters.

In a few samples where multiple peaks were observed most probably due to the sample leak from the adjacent well, all such samples were reanalysed on the automated sequencer. The allele calling remained unambiguous with the use of allelic ladders. The genotype data for D3S1358, vWA & FGA loci, for each population is attached as Appendix 1.

4.4 DISCUSSION

Short tandem repeat polymorphisms have become standard genetic markers for forensic identification and population studies (Urqhart, A. *et al.,* 1995; Sparkes, R. *et al.*, 1996 a, b). Characterisation of most of the populations of the world has been carried out for various STR markers and the databases thus generated are being used in DNA profiling (Brinkmannn, B. *et al.,* 1996; Pestoni, C. *et al.*, 1995; Kupferschmid, T.D. *et al.*, 1999 & Klintschar, M. *et al.,* 1999).

4.4.1 AMPLIFICATION REACTIONS

Development of an optimal PCR assay is required in order to get specific and sensitive results so that the interpretation of the results of the profiling are reliable (Wallin, J. M. *et al.*, 1998).

The PCR optimisation process results in an optimal value for each critical parameter that allows for the small variations, which are inevitable in daily laboratory practice. Hot start PCR technique using Ampli*Taq*® Gold was found particularly helpful during the optimisation of the PCR. The reaction could be performed with relative ease. However Promega *Taq* also gave comparable results though the reactions had to be prepared quickly on ice to avoid non specific

amplification. Considering the number of samples, Promega *Taq* was used for amplification reactions with a significant saving of resources.

The primers for all the three loci proved to be highly specific. In the multiplex reaction, an unbalanced primer concentration resulted in non amplification of the locus FGA, though D3 continued to amplify over longer range of primer concentration. It has been shown earlier that for an efficient PCR a molar excess of the primers and dNTPs are required (Cha, R.S & Thilly, W. G. 1993). The KCl based buffer supplied with the *Taq* DNA Polymerase was used during this work and the dNTP concentration was kept at 200 μmol each, which produced the desired results. The allele size for D3S1358 was reported as 97–126 bp for the alleles 13–19 respectively (Li, H. *et al.,* 1993; Szibor, R. *et al.,* 1998). As sequence information was not available at the time this project began, amplification with the new primer (which was slightly different from that of Li, H. *et al.*) generated a much bigger size than anticipated (Table 4.4). This prevented multiplexing all the three loci. D3S1358 forward primer could have been labelled with another fluorescent dye for multiplexing the three loci together, however, since that was not the primary aim of this study, same primer was used to amplify the samples in duplex reactions.

4.4.2 DETERMINATION OF ALLELIC FRAGMENT SIZE ON THE ABI 373 XL UPGRADE DNA SEQUENCER

It was known that the migration of the alleles through the gel depends not only on the conditions of electrophoresis but also on the sequence of the fragment (Frank, R & Koster, H. 1979). It was thus necessary that an allelic ladder which acts as a locus specific size control be used for accurate designation of an allele, thus the technique was applied to AmpFLP locus D1S80 (Budowle, B. *et al.*, 1991) and STR loci (Edwards, A. *et al.*, 1991). It was advocated that allelic ladders must be used for precise sizing of the alleles in forensic casework (Fregeau, C. J. & Fourney, R. M. 1993). The size of each allele can be expressed confidently by using an allelic ladder, and the data can be compared between different laboratories using different electrophoretic systems. The ladders were also helpful in the recognition of microheterogeneity within the system (Peurs, C. *et al.,* 1993) thus

the use of sequenced allelic ladders has been recommended by the DNA commission of the International Society of Forensic Genetics (ISFG) (Olaisen, B. *et al.,* 1997).

The standard laid down by the ISFG was that the typing errors of any system should be less than half of the repeat size (Olaisen, B. *et al.,* 1997) though for the majority of loci the typing errors are smaller than this standard and reliable sizing can be achieved by using the automated sequencer along with internal standards. Since some of the tetranucleotide STRs may show 2 bp variants the standard deviation of 1 bp would be acceptable. However variants that differ by 1 bp have been detected and thus a smaller standard was set at 0.5 bp (Gill. P. *et al.,* 1996). Following the same rule the alleles were designated within ±0.5 bp floating window with reference to the nearest ladder fragment (Figure 4.4).

All the three loci had discrete alleles for which allelic designation could be done unambiguously. The alleles, which were not present in the ladder, were assigned by extrapolation. The electrophoresis and subsequent sizing of the fragments on automated equipment is governed by several factors. These include the sizing ladder, the matrix composition, the composition of the gel, the strength of the current of the equipment during the run and the length of the run (Smith, R. N., 1995; Worley, J. M. *et al.*, 1996; Kline, M. C. *et al.,* 1996). While using the automated sequencer, care was taken to keep all these crucial factors constant, in order to have consistent and reproducible results.

During this study 702 samples from seven populations of Pakistan were profiled for three autosomal STRs. This is the first study in which various sub populations from Pakistan were profiled for autosomal STRs.

CHAPTER 5: STATISTICAL ANALYSES AUTOSOMAL STR LOCI

5.1 INTRODUCTION

Genotyping of the 702 samples from seven Pakistani populations generated data for the three STR loci D3S1358, vWA and FGA. There was a need to analyse this data in order to study the forensic and population genetics aspects. For this purpose apart from the allele and genotype frequencies, their variances and standard deviations were calculated. The data was tested for Hardy Weinberg principle applying different tests. The structure of the populations was studied from the viewpoint of forensic applications of the data.

5.2 STATISTICAL METHODS EMPLOYED

5.2.1 ALLELE AND GENOTYPE FREQUENCIES

The allele/genotype frequencies were calculated by dividing the number of a specific allele/genotype observed by the number of total alleles/genotypes. The standard deviations (SD) of the allele frequencies were calculated using the formula ($\mathbf{1.96\sqrt{}\ pu(1-pu)/n}$), where *pu* is the allele frequency and n is the number of alleles. This was done in order to determine the confidence limits of the allele and genotype frequencies within 95% probability (Weir, B. S., 1996). Standard deviations (SD) of the genotypes were calculated using the same formula with the substitution of allele frequency with genotype frequency and counts (Weir, B. S., 1996).

5.2.2 VARIANCE ESTIMATES OF THE ALLELE FREQUENCIES

Variances of the allele frequencies and the genotypes were calculated for all the populations using two formulae (Weir, B.S. 1996):

(a) $\mathbf{Var} = pu(1 - pu)/2n$

where, *pu* is the allelic frequency of allele p and n is the number of individuals profiled.

(b) Var $= (pu + puu - 2pu^2)\ 2n$

where, *pu* is the allelic frequency, *puu* is the homozygote frequency for an allele and n is the number of individuals profiled.

The first equation was used to calculate the variance of the allele frequency not assuming equilibrium. In using the second equation it was assumed that the samples have been drawn at random and the population was in equilibrium. If the populations were in equilibrium the results of both equations would be similar, however any differences would point out the deficiency or excess of the homozygote frequencies (Weir, B. S. 1996). Thus the variance estimates using these two methods becomes a test of independence.

5.2.3 EXACT TEST & F STATISTICS

The exact test (Fisher, R. A. 1934) calculates the probability of getting a particular set of values by chance. All the populations were tested for independence at the three loci by performing the Fisher's exact test. p-values for the genotype data for all the populations were calculated using the computer program GDA (Lewis, P. O. & Zaykin, D. 2000. Genetic Data Analysis: Computer program for the analysis of allelic data. Version 1.0 (d15c). Free program distributed by the authors over the internet from http://lewis.eeb.uconn.edu/lewishome/software.html). The F statistics (1.3.8) were also calculated using the same program.

5.2.4 FORENSIC PARAMETERS FOR EACH LOCUS AND THE COMBINED PARAMETERS FOR THE THREE LOCI D3S1358, vWA & FGA

The forensic parameters of the utility of a marker to be used for identification purposes are calculated before putting a marker in actual use. The forensic utility of the STRs is shown by the parameters including match probability

(pM) (Jones, D. A. 1972), probability of discrimination (PD) (Jones, D. A. 1972), probabiltiy of exclusion (PE) (Chakraborty, R. *et al.*, 1974), paternity index (PI) (Chakraborty, R. *et al.*, 1982), and the polymorphic information content (PIC) (Botstein, D. *et al.*, 1990). All the above parameters were calculated for the three loci for each population using the Promega Corp. computer software PowerStats (Tereba, A. 1999).

In addition the combined probability of discrimination (PD_{comb}) (Jones, D. A. 1972) combined probability of exclusion (PE_{comb}) (Chakraborty, R. and Jin, L. 1993) and combined Probability of Match (PM_{comb}) (Jones, D. A. 1972) for the three autosomal loci were calculated for each population separately. The definitions and formulae for all these parameters are attached as Appendix 2.

5.3 RESULTS

5.3.1 ALLELE FREQUENCIES

The alelle frequencies differed among Pakistani populations at most allele classes. The results are shown in Figures 5.1a, b & c and Tables 5.1a, b & c.

5.3.2 GENOTYPE FREQUENCIES

5.3.2.1 Locus D3S1358 Genotype Frequencies

The most frequent genotypes in all the populations were 15/16,15/17 and 16/17. The Baluchi and Makrani populations resembled each other more than the other Pakistani populations. In these two populations the frequency of genotypes 16/18 was higher than other populations. The homozygotes for allele 17 were at a high frequency in the Makranis. Highest frequency (0.24), of genotype 15/16 was observed in the Brosho population. Homozygote frequency for the allele 15 was the highest frequency (0.15) of homozygotes among the populations studied. The genotype frequencies of the locus D3S1358 differed among the various Pakistani populations to a significant extent. (Table 5.2a).

Figure 5.1a: Allele Frequencies of the Locus D3S1358 in Pakistani Populations

The allelic frequencies of the locus D3S1358 differed among the Pakistani populations (Table 5.1). Allele 13 & 14 were present at a high frequency in the Punjabi population. Frequency of allele 18 was predominantly high in the Baluchi population. On the whole allele 15 was the most common, whereas allele 11 was the most rare at this locus (Table 5.1).
In the Kalash alleles were concentrated in the three most frequent alleles and other alleles were rare.

Figure 5.1b: Allele Frequencies of the Locus vWA in Pakistani Populations

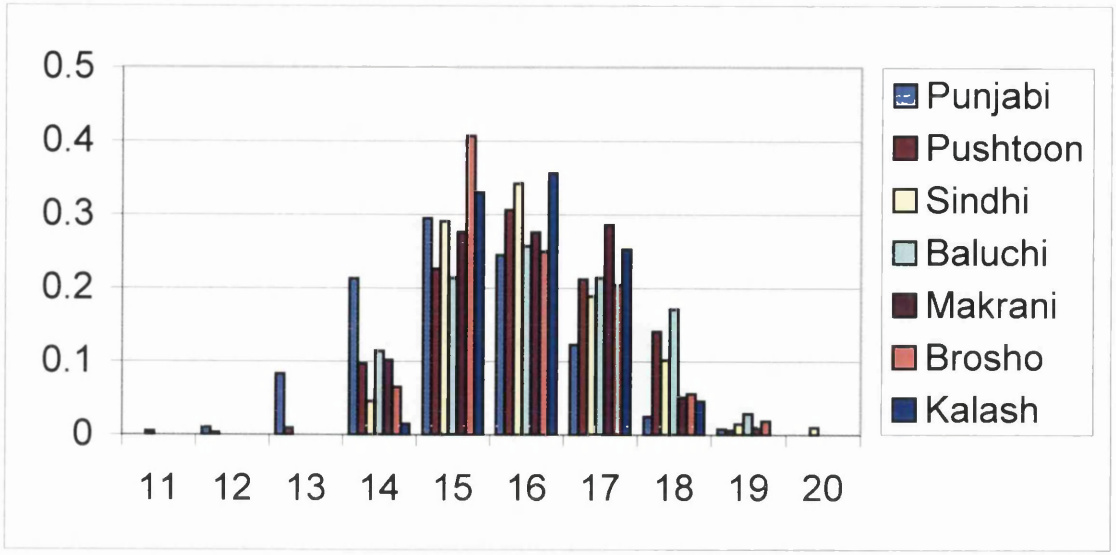
The allelic frequencies of vWA also differed among populations. A significantly high frequency of allele 17 was observed in the Makrani population at 0.46.
In the Kalash allele 15 was at twice the frequency of other populations.

Figure 5.1c: Allele Frequencies of the Locus FGA in Pakistani Populations

The allele frequency of FGA locus appeared to differ more than the other two loci among Pakistani populations. Among the mainland populations the Makrani population showed frequency of allele 25 ~ double that of the other populations. In the Pushtoon population this allele was rare. Allele 32 was observed only in the Baluchis while allele 29 was only observed in the Pushtoons.

Only 10 alleles were detected in the Kalash sample at this locus. The Kalash show a peculiar spread of FGA alleles. In contrast to the mainlnad populations alleles 22 and 20 were at a higher frequency in the Kalash. A variant allele 24.2 was detected in a high frequency in the Kalash which was not observed in any other population during this study.

Locus D3S1358



Locus vWA

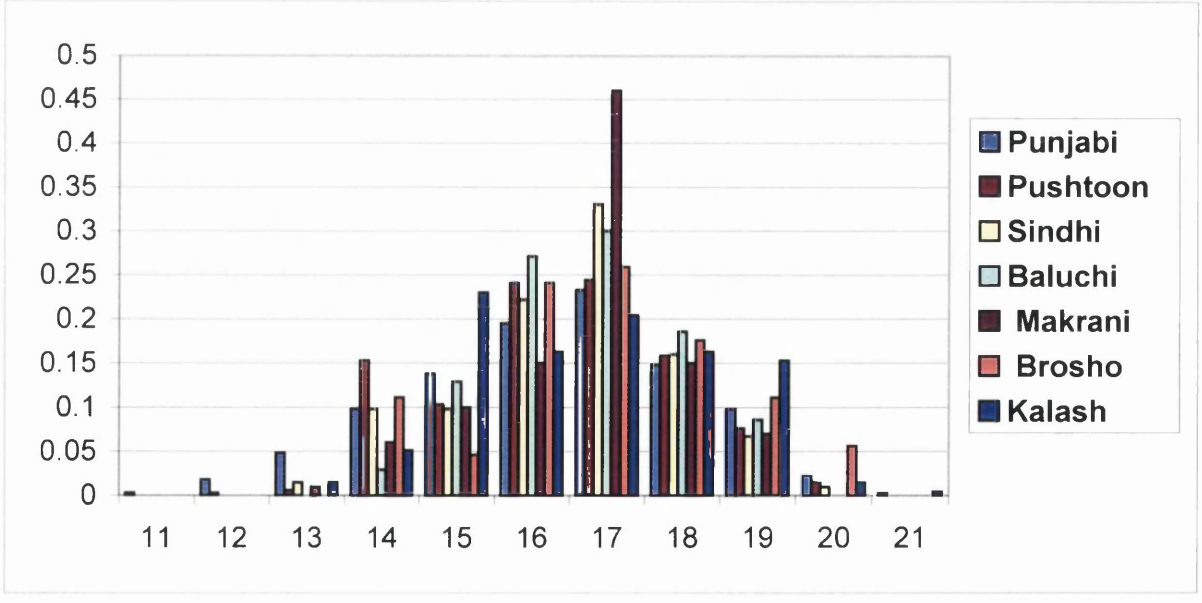


Figure 5.1 c: Allelic Frequencies of Locus FGA in Pakistani Populations

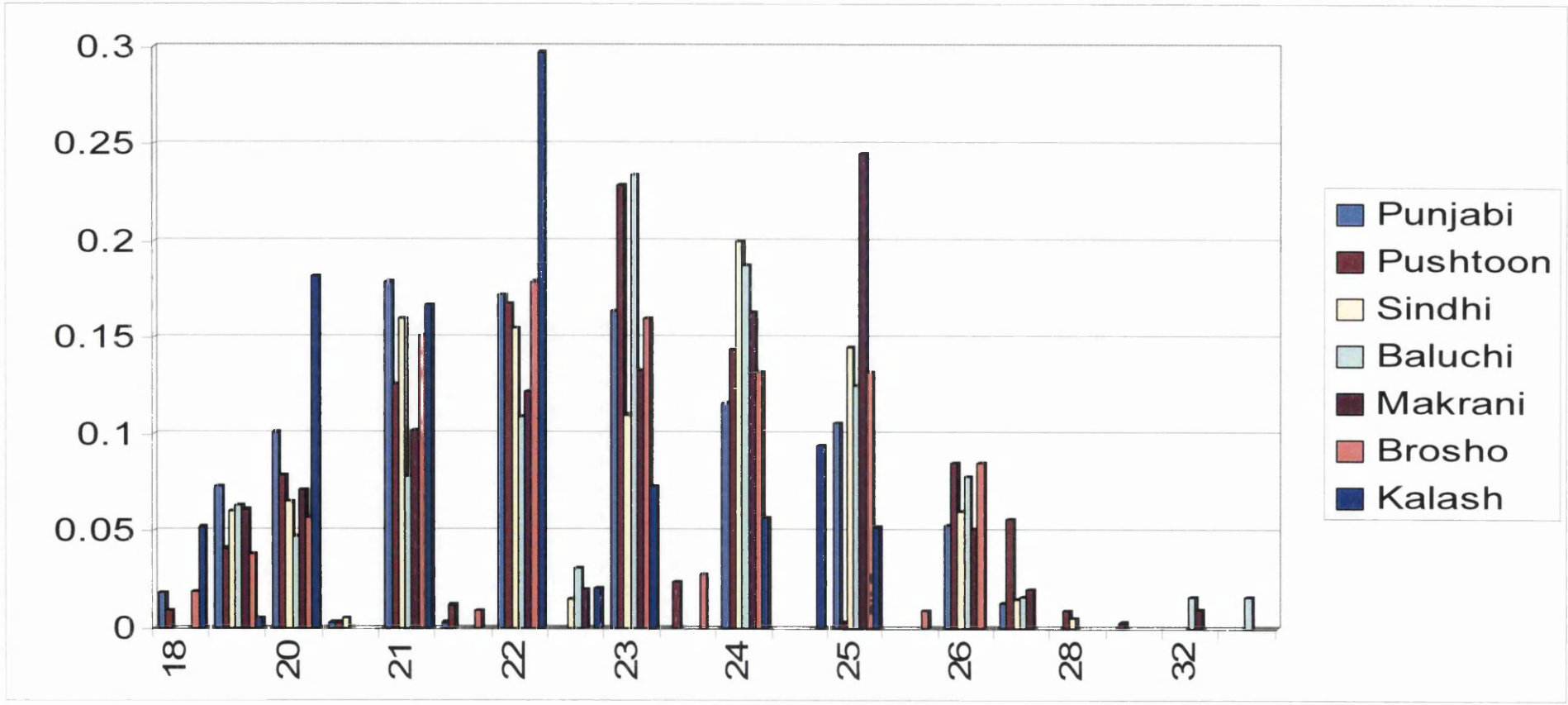


Table 5.1a: Allele Frequencies of the Locus D3S1358 in Pakistani Populations

Allele	Punjabi (200)[1]	Pushtoon (170)	Sindhi (98)	Baluchi (35)	Makrani (48)	Brosho (54)	Kalash (97)
11			0.005 (0.009)				
12	0.010 (0.009)[2]	0.003 (0.006)					
13	0.083 (0.027)	0.009 (0.010)					
14	0.213 (0.040)	0.097 (0.031)	0.046 (0.030)	0.114 (0.070)	0.102 (0.060)	0.065 (0.046)	0.015 (0.017)
15	0.295 (0.045)	0.226 (0.044)	0.291 (0.060)	0.214 (0.096)	0.276 (0.089)	0.407 (0.093)	0.33 (0.066)
16	0.245 (0.042)	0.306 (0.049)	0.342 (0.066)	0.257 (0.102)	0.276 (0.089)	0.25 (0.082)	0.356 (0.067)
17	0.123 (0.032)	0.212 (0.043)	0.189 (0.055)	0.214 (0.096)	0.286 (0.090)	0.204 (0.076)	0.253 (0.061)
18	0.025 (0.015)	0.141 (0.037)	0.102 (0.043)	0.171 (0.088)	0.051 (0.044)	0.056 (0.043)	0.046 (0.290)
19	0.008 (0.009)	0.006 (0.008)	0.015 (0.017)	0.029 (0.040)	0.01 (0.020)	0.019 (0.026)	
20			0.010 (0.014)				

1. Number of the individuals profiled

2. Values in parentheses show the 95% confidence limit of standard deviation

Table 5.1b: Allele Frequencies of the Locus vWA in Pakistani Populations

Allele	Punjabi (200)[1]	Pushtoon (170)	Sindhi (98)	Baluchi (35)	Makrani (48)	Brosho (54)	Kalash (97)
11	0.003 (0.005)[2]						
12	0.018 (0.013)	0.003 (0.006)					
13	0.048 (0.021)	0.006 (0.008)	0.015 (0.017)		0.010 (0.020)		0.015 (0.017)
14	0.098 (0.029)	0.153 (0.038)	0.098 (0.042)	0.029 (0.039)	0.060 (0.046)	0.111 (0.059)	0.051 (0.031)
15	0.138 (0.033)	0.103 (0.032)	0.098 (0.042)	0.129 (0.079)	0.100 (0.060)	0.046 (0.040)	0.230 (0.059)
16	0.195 (0.039)	0.241 (0.045)	0.222 (0.058)	0.271 (0.104)	0.150 (0.071)	0.241 (0.081)	0.163 (0.052)
17	0.233 (0.041)	0.244 (0.045)	0.330 (0.066)	0.300 (0.107)	0.460 (0.100)	0.259 (0.083)	0.204 (0.057)
18	0.148 (0.035)	0.159 (0.039)	0.160 (0.052)	0.186 (0.091)	0.150 (0.070)	0.176 (0.072)	0.163 (0.052)
19	0.098 (0.029)	0.076 (0.028)	0.067 (0.047)	0.086 (0.066)	0.070 (0.050)	0.111 (0.059)	0.153 (0.051)
20	0.023 (0.015)	0.015 (0.013)	0.010 (0.014)			0.056 (0.043)	0.015 (0.017)
21	0.003 (0.005)						0.005 (0.010)

1. Number of the individuals profiled

2. Values in parentheses show the 95% confidence limit of standard deviation beneath the frequency

Table 5.1c: Allelic Frequencies of the Locus FGA in Pakistani populations

Pop	18	19	20	20.2	21	21.2	22	22.2	23	23.2	24	25	26	27	28	29
Punjabi (200)*	0.018	0.073	0.1	0.003	0.179	0.003	0.172		0.164		0.116	0.106	0.053	0.013		
SD	0.011	0.025	0.031	0.0052	0.038	0.005	0.04		0.04		0.03	0.03	0.02	0.01		
Pushtoon (170)	0.009	0.041	0.079	0.003	0.126	0.012	0.168		0.229	0.024	0.144	0.003	0.085	0.056	0.009	0.003
SD	0.01	0.02	0.03	0.006	0.035	0.012	0.04		0.045	0.016	0.037	0.006	0.03	0.024	0.01	0.006
Sindhi (98)		0.06	0.065	0.005	0.16		0.155	0.015	0.11		0.2	0.15	0.06	0.015	0.005	
SD		0.033	0.034	0.01	0.05		0.05	0.02	0.043		0.055	0.049	0.033	0.017	0.005	
Baluchi1(32)		0.063	0.047	0.01	0.078		0.109	0.031	0.234		0.188	0.125	0.078	0.016		
SD		0.06	0.052		0.06		0.08	0.04	0.1		0.096	0.08	0.066	0.031		
Makrani[2] (48)		0.061	0.071		0.102		0.122	0.02	0.133		0.163	0.245	0.051	0.02		
SD		0.05	0.05		0.06		0.07	0.03	0.07		0.07	0.08	0.046	0.03		
Brosho[3](54)	0.019	0.038	0.057		0.151	0.009	0.179	-	0.16	0.028	0.132	0.132	0.085			
SD	0.026	0.036	0.06		0.043	0.018	0.072		0.069	0.031	0.064	0.064	0.052			
Kalash[4] (97)	0.052	0.005	0.182		0.167		0.297	0.021	0.073		0.057	0.052				
SD[5]	0.031	0.01	0.054		0.052		0.064	0.02	0.037		0.032	0.032				

1. Rare alleles found in this population were 32 (0.016) and 33 (0.016)
2. Rare allele found in this population was 32 (0.01)
3. Variant allele found in this population only was 25.2 (0.009)
4. Variant allele found in this population only was 24.2 (0.094
5. Standard deviation

* Number of individuals profiled in parentheses

5.3.2.2 Locus vWA Genotype Frequencies

The Punjabi population showed the most genotypes at this locus (36), whereas the Baluchis showed the least (13). The Brosho population showed similarity to the Pushtoons in the distribution of the vWA genotypes.

The genotype frequencies of the locus vWA showed a similar distribution of the most frequent genotypes among the Punjabi and Pushtoon populations. Thus, the frequency of genotype 17/18, 16/17, 16/18 and 15/16 were similar in both populations. The Sindhi population had a high proportion of genotype 16/17 (0.2) and 17/18 (0.124). Other genotype frequencies were similar to Punjabis and Pushtoons. Baluchis had genotype 17/18 at the highest frequency (0.23) whereas Makranis had 15/17 as the most frequent genotype (0.12). The homozygote frequency for the allele 17 was high in all Pakistani populations except Punjabis and Pushtoons. The frequency of this homozygote was extremely high in the Makrani population (0.28) (Table 5.2 b).

5.3.2.3 Locus FGA Genotype Frequencies

70 different genotypes were observed in all the samples at locus FGA. Frequencies of FGA genotypes in the Punjabi population had a wide spread of alleles. This population showed 43 different genotypes and the genotype frequencies were evenly distributed over the whole range. The same pattern was observed in the Pushtoon population in which 37 genotypes were observed. The variant allele 23.2 was observed only in Pushtoon population at a low frequency. Other than this Pushtoons and Punjabis do not differ significantly in the distribution of genotypes at locus FGA. The Sindhi population exhibited 33 genotypes. The highest frequency was that of genotype 22/23 (as in Punjabis). The other frequent genotypes were 24/25 and 19/24. In the Punjabi, Pushtoon and Sindhi population none of the genotypes was at a frequency higher than 0.086 (Table 5.2 c).

5.3.2.4 Genotypes of Loci D3S1358, vWA and FGA in Kalash Population

In the Kalash population the most frequent genotypes for the locus D3S1358, were 16/17 and 15/17. Three homozygote classes were observed for the alleles 15, 16 and 17. The frequency for the homozygote 15/15 was quite high. For the locus vWA genotype 18/19, 15/18 and 15/16 were at almost similar frequency of 0.1. Other genotypes were observed at lower frequency. For the locus, FGA 32 genotypes were detected in the Kalash. The genotypes 21/22 and 22/24.2 were the most frequent in the Kalash. Only three homozygotes were detected for the alleles 20, 21 and 22. The frequency of the homozygotes for allele 22 was higher (0.094) in the Kalash than other Pakistani populations (Table 5.2 c & 5.3).

5.3.3 TESTS FOR HARDY WEINBERG PRINCIPLE

5.3.3.1 The Exact Test

For most populations the analysis of the genotype data for exact test resulted in a p-value of >0.05 for the three loci. The only exceptions were the Punjabi and the Sindhi populations for the locus FGA. In these populations a lower p-value was obtained. Further, in order to test the independence of the loci from each other the exact test was employed to test the combination of loci. With the exception of Sindhi and Brosho, no significant difference from independence was found in other populations (Table 5.4).

5.3.3.2 Variance Estimates of the Allelic Frequencies

If repeated samples are taken from the same population the allele/ genotype frequencies would differ as sample to sample variation is a rule. The extent of this varation between the samples can be estimated as the variance of the frequency (Weir, B. S. 1996).

The results showed that differences existed in the two variance estimates for the alleles in the Punjabi and the Sindhi populations. However, these were largely similar as in other populations. In the Sindhi population allele 17 had a significant difference among the two variance estimates.

Table 5.2a: Genotype Frequencies for the LocusD3S1358 in Pakistani Populations

	Punjabi (200)		**Pushtoon** (170)		**Sindhi** (98)		**Baluchi** (35)		**Makrani** (48)		**Brosho** (54)	
G	P	SD	P	SD	P	SD	P	SD	P	SD	P	SD
11,17					0.006	0.02						
12,13	0.01	0.001	0.006	0.01								
13,13	0.01	0.01										
13,14	0.02	0.02	0.006	0.01								
13,15	0.06	0.03										
13,16			0.012	0.02								
13,17	0.03	0.02										
14,14	0.05	0.028	0.006	0.01	0.006	0.02	0.028	0.06	0.02	0.04		
14,15	0.15	0.05	0.047	0.03	0.012	0.02	0.028	0.06	0.061	0.07	0.09	0.08
14,16	0.11	0.04	0.059	0.03	0.023	0.03	0.086	0.01	0.102	0.09	0.02	0.04
14,17	0.04	0.02	0.047	0.03	0.006	0.01	0.057	0.08			0.02	0.04
14,18	0.02	0.017	0.024	0.02								
15,15	0.07	0.03	0.035	0.03	0.059	0.05	0.086	0.09	0.081	0.08	0.15	0.1
15,16	**0.16**	0.05	**0.165**	0.05	**0.112**	0.06	0.057	0.08	0.122	0.09	**0.24**	0.11
15,17	0.07	0.03	0.1	0.04	0.047	0.04	0.143	0.13	0.163	0.1	0.15	0.1
15,18	0.01	0.01	0.071	0.04	0.035	0.04	0.028	0.06	0.02	0.04	0.02	0.04
15,19					0.006	0.01			0.02	0.04	0.02	0.04
15,20					0.006	0.01						
16,16	0.05	0.03	0.082	0.04	0.059	0.05			0.02	0.04	0.05	0.06
16,17	0.06	0.016	0.118	0.05	**0.112**	0.06	0.143	0.12	**0.224**	0.12	0.09	0.08
16,18	0.02	0.01	0.088	0.04	0.023	0.03	**0.171**	0.13	0.061	0.07	0.04	0.05
16,19							0.057	0.08				
16,20							0.028	0.06				
17,17	0.03	0.02	0.035	0.03	0.006	0.01	0.028	0.06	0.081	0.08	0.04	0.05
17,18			0.076	0.04	0.041	0.04			0.02	0.04	0.05	0.06
18,18			0.012	0.01	0.006	0.01	0.057	0.08			0.02	0.04
18,19	0.01	0.01	0.012	0.01	0.006	0.01						
18,20					0.006	0.01						

G, Genotype

P, Frequency of the genotype

SD, Standard deviation of the frequency

Highest Frequency of a genotype detetcted in a population is shown in bold

Table 5.2b: Genotype Frequencies for the Locus vWA in Pakistani Populations

	Punjabi		Pushtoon		Sindhi		Baluchi		Makrani		Brosho	
	(200)		(170)		(100)		(35)		(48)		(54)	
G	P	SD	P	SD	P	SD	P	SD	P	SD	P	SD
11,12	0.005	0.010	0.006	0.006								
12,12	0.005	0.010										
12,15	0.005	0.010										
12,16	0.010	0.014										
12,17	0.005	0.010										
13,13	0.010	0.014										
13,14	0.015	0.017			0.010	0.020						
13,15	0.025	0.022			0.010	0.020						
13,16	0.005	0.010	0.006	0.012	0.010	0.020						
13,17	0.015	0.017							0.020	0.038		
13,18	0.015	0.017										
14,14	0.005	0.010	0.023	0.023	0.030	0.034					0.020	0.037
14,15	0.045	0.029	0.035	0.028			0.060	0.080				
14,16	0.055	0.032	0.080	0.041	0.021	0.029			0.020	0.038	0.090	0.076
14,17	0.040	0.027	0.080	0.041	0.062	0.048			0.080	0.075	0.050	0.058
14,18	0.025	0.022	0.035	0.028	0.021	0.029			0.020	0.038		
14,19	0.025	0.022	0.010	0.015	0.010	0.020					0.040	0.052
14,20	0.010	0.014	0.006	0.012	0.010	0.020						
15,15	0.010	0.014	0.006	0.012	0.010	0.020					0.020	0.037
15,16	0.065	0.034	0.040	0.029	0.051	0.044	0.080	0.090	0.020	0.038	0.020	0.037
15,17	0.055	0.032	0.053	0.034	0.051	0.044	0.030	0.057	0.120	0.090	0.020	0.037
15,18	0.030	0.024	0.047	0.032	0.041	0.039	0.030	0.057	0.040	0.050	0.020	0.037
15,19	0.020	0.020	0.020	0.021	0.010	0.020	0.060	0.079	0.020	0.038		
15,20	0.035	0.025			0.010	0.020						
16,16	0.010	0.010	0.060	0.036	0.041	0.039	0.100	0.099	0.060	0.060	0.050	0.058
16,17	**0.100**	0.042	**0.118**	0.048	**0.200**	0.080	0.080	0.090	0.040	0.054	**0.110**	0.083
16,18	0.065	0.034	0.080	0.041	0.062	0.048	0.080	0.090	0.080	0.075	0.070	0.068
16,19	0.045	0.029	0.030	0.026	0.021	0.029	0.060	0.079	0.020	0.038	0.040	0.052
16,20	0.005	0.010	0.006	0.012							0.040	0.052
17,17	0.045	0.029	0.053	0.034	0.082	0.055	0.114	0.105	**0.280**	0.124	**0.110**	0.084
17,18	**0.100**	0.042	0.082	0.041	0.124	0.066	**0.230**	0.139	0.060	0.065	0.050	0.058
17,19	0.045	0.029	0.035	0.028	0.062	0.048	0.030	0.057	0.040	0.054	0.040	0.052
17,20	0.015	0.017	0.012	0.016							0.020	0.037
18,18			0.018	0.020	0.031	0.020	0.030	0.057	0.040	0.054	0.050	0.061
18,19			0.035	0.028	0.010	0.010			0.040	0.054	0.070	0.068
18,20			0.006	0.012							0.020	0.037
19,19	0.010	0.014	0.012	0.016	0.010	0.010						
19,20	0.005	0.010			0.010	0.010					0.040	0.052
19,21	0.005	0.010										

G, Genotype P, Frequency of Genotype SD, Standard Deviation
Highest Frequency of a genotype detected in a population is shown in bold
Number of Individuals profiled are in paraentheses

Table 5.2c : Genotype Frequencies for the Locus FGA in Pakistani Populations

	Punjabi		Pushtoon		Sindhi		Baluchi		Makrani		Brosho	
G	P	SD	P	SD	P	SD	P	SD	P	SD	P	SD
18,19	0.005	0.005	0.006	0.006								
18,20	0.01	0.007										
18,21			0.006	0.006							0.019	0.019
18,22	0.01	0.007										
18,23												
18,24	0.01	0.007										
18,25											0.019	0.019
19,19	0.015	0.009					0.031	0.031				
19,20	0.01	0.007										
19,20.2			0.006	0.006								
19,21	0.015	0.009	0.012	0.008	0.02	0.014			0.02	0.020	0.019	0.019
19,22	0.015	0.009	0.018	0.010								
19,23	0.01	0.007	0.018	0.010	0.01	0.010						
19,24	0.01	0.007			0.06	0.024	0.062	0.043	0.041	0.029	0.019	0.019
19,25	0.025	0.011			0.03	0.017			0.062	0.035		
19,26	0.02	0.010									0.038	0.026
19,27	0.005	0.005										
20,20	0.01	0.007	0.023	0.011								
20,20.2	0.005	0.005			0.01	0.010						
20,21	0.025	0.011	0.023	0.011	0.07	0.026	0.031	0.031	0.041	0.029		
20,22	0.01	0.007	0.035	0.014	0.01	0.010			0.02	0.020	0.038	0.026
20,23	0.02	0.010	0.029	0.013	0.01	0.010			0.02	0.020	0.038	0.026
20,23.2											0.019	0.019
20,24	0.045	0.015	0.012	0.008	0.02	0.014	0.062	0.043	0.02	0.020		
20,25	0.04	0.014							0.02	0.020		
20,26	0.015	0.009	0.012	0.008	0.01	0.010			0.02	0.020	0.019	0.019
21,21	0.025	0.011	0.012	0.008	0.03	0.033						
21,22	0.086	0.020	0.035	0.014	0.04	0.038					0.076	0.036
21,21.2			0.006	0.006					0.041	0.029	0.019	0.019
21,23	0.056	0.016	0.065	0.019	0.02	0.014	0.031	0.031	0.041	0.029	0.057	0.032
21,24	0.056	0.016	0.029	0.013	0.04	0.020	0.062	0.043	0.02	0.020	0.019	0.019
21,25	0.04	0.014			0.04	0.020	0.031	0.031			0.076	0.036
21,26	0.025	0.011	0.035	0.014	0.03	0.017			0.02	0.020		
21,27	0.005	0.005	0.018	0.010					0.02	0.020		
22,22	0.03	0.012	0.023	0.011	0.05	0.022			0.041	0.029	0.038	0.026
22,22.2							0.031	0.031				
22,23	0.05	0.015	0.09	0.022	0.07	0.026	0.125	0.058	0.062	0.035	0.057	0.032
22,23.2			0.012	0.008								
22,24	0.035	0.013	0.05	0.017	0.04	0.020	0.031	0.031	0.041	0.029	0.057	0.032
22,25	0.04	0.014	0.006	0.006	0.03	0.017	0.031	0.031	0.02	0.020	0.057	0.032
22,26	0.025	0.011	0.023	0.011					0.02	0.020		
22,27	0.005	0.005	0.018	0.010	0.01	0.010						
22,28					0.01	0.010						
22.2.23							0.031	0.031				
22.2,24			0.012	0.008								
22.2.24					0.01	0.010						
22.2,25					0.02	0.014						
23,23	0.045	0.015	0.06	0.018	0.01	0.010	0.094	0.052	0.02	0.020	0.019	0.019
23,23.2											0.019	0.019
23,24	0.05	0.015	0.065	0.019	0.05	0.022	0.031	0.031	0.062	0.035	0.057	0.032
23,25	0.035	0.013			0.01	0.010			0.041	0.029	0.057	0.032
23,26	0.01	0.007	0.05	0.017	0.02	0.014	0.031	0.031				
23,27	0.005	0.005	0.05	0.017			0.125	0.058				
23.2,24											0.019	0.019
23.2,26			0.006	0.006								
23.2,27			0.006	0.006								
23.2,28			0.018	0.010								
24,24	0.01	0.007			0.05	0.022	0.031	0.031	0.02	0.020	0.038	0.026
24,25					0.07	0.026			0.102	0.044	0.019	0.019
24,27	0.005	0.005			0.01	0.010						
25,25	0.01	0.007			0.04	0.020			0.082	0.040	0.019	0.019
25,26	0.01	0.007			0.01	0.010	0.062	0.043	0.041	0.029		
25,27									0.02	0.020		
25.2,26											0.019	0.019
26,26			0.006	0.006	0.02	0.014					0.038	0.026
26,27			0.018	0.010								
26,29			0.006	0.006								
26,32							0.031	0.031	0.02	0.020		
26,33							0.031	0.031				
27,27					0.01	0.010						

Table 5.3: Genotype Frequencies of the Loci D3S1358, VWA & FGA in Kalash Population

	D3S1358		vWA		FGA		
Genotype	**P**	**SD**	**P**	**SD**	**Genotype**	**P**	**SD**
13,16			0.01	0.020	18,20	0.04	0.039
13,17			0.01	0.020	18,21	0.01	0.020
13,18			0.01	0.020	18,22	0.02	0.028
14,14			0.01	0.020	18,23	0.02	0.028
14,15	0.01	0.020	0.02	0.027	18,24	0.02	0.028
14,17	0.020	0.028	0.01	0.020	18,24.2	0.02	0.028
14,18			0.02	0.027	19,23	0.02	0.028
14,19			0.02	0.027	20,20	0.042	0.040
14,21			0.01	0.020	20,21	0.052	0.044
15,15	0.134	0.068	0.06	0.047	20,22	0.07	0.051
15,16	0.165	0.074	0.09	0.056	20,22.2	0.01	0.020
15,17	0.185	0.077	0.08	0.053	20,23	0.04	0.039
15,18			0.09	0.056	20,24	0.03	0.034
15,19			0.05	0.043	20,24.2	0.02	0.028
16,16	0.144	0.034	0.05	0.043	20,25	0.02	0.028
16,17	**0.216**	0.070	0.05	0.043	20,25	0.01	0.020
16,18	0.040	0.082	0.02	0.027	21,21	0.01	0.020
16,19			0.04	0.038	21,22	**0.134**	0.068
16,20			0.01	0.020	21,22.2	0.01	0.020
17,17	0.031	0.039	0.06	0.047	21,23	0.04	0.039
17,18	0.02	0.034	0.05	0.043	21,24	0.01	0.020
17,19			0.07	0.050	21,24.2	0.02	0.028
17,20			0.01	0.020	21,25	0.03	0.034
18,18			0.01	0.020	22,22	0.094	0.058
18,19			**0.11**	0.061	22,23	0.01	0.020
19,20			0.01	0.020	22,23.2	0.01	0.020
					22,24	0.03	0.034
					22,24.2	0.09	0.057
					22,25	0.03	0.034
					24,24.2	0.01	0.020
					24,25	0.01	0.020
					24.2,25	0.02	0.028

P, Genotype Frequency

SD, Standard Deviation

Higest Frequency of a genotype detected is shown in bold

The Sindhi and Baluchi populations showed the highly significant difference in the variance estimates of alleles 27 & allele 19 respectively, showing that there were more homozygotes for this allele than expected under the HW principle. The Kalash also showed differences at two alleles 24 & 25 showing that there was a deficiency for homozygotes for these alleles. Baluchi and Kalash both showed an exact test p-value of >0.05 while the Sindhi population showed a lower value. These results show the low power of the exact test to detect the extent of deviation of a population from HW principle (Table 5.5).

5.3.3.3 Heterozygosity Test

The HW principle explains the relationship of the frequency of heterozygotes and homozygotes in a population observing the principle (Strickberger, M. W. 1985). It shows how near or far are the frequencies of the homozygotes and the heterozygotes in the population being tested from the expected frequencies. The observed number of heterozygotes /homozygotes in a particular data set can be used to show if their number is in accordance with the HW principle or not.

The value of expected number heterozygotes were calculated for each population using the value of the expected heterozygosity using the statistical package GDA (Table 5.6). These values were calculated for each population at three loci. Only for Sindhi and Pushtoon populations the expected and the observed number of heterozygotes were slightly different. Standard errors for the value of the expected heterozygotes were also calculated in order to see the fitness of the observed number of heterozygotes.

The results showed that in the Sindhi population the observed number of heterozygotes for the locus FGA were 76 whereas the expected number was 85. Differences were also observed in Pushtoon population at the D3S1358 locus and the Makrani and Brosho populations at vWA locus though these were not as significant.

Table 5.4: Exact Test *p-values* For Individual Loci and Combination of Loci

Population	D3S1358	vWA	FGA	D3/ vWA	D3/ FGA	vWA/ FGA	3 Locus[1]
Punjabi	0.341	0.216	**0.01**	0.52	0.33	0.66	0.53
Pushtoon	0.905	0.79	0.07	0.57	0.56	0.16	0.54
Sindhi	0.224	0.278	**0.02**	0.21	**0.031**	**0.02**	0.15
Baluchi	0.093	0.106	0.64	0.08	0.18	0.48	0.41
Makrani	0.124	0.234	0.367	0.165	0.433	0.06	0.26
Brosho	0.88	0.424	0.239	0.41	0.73	**0.032**	0.22
Kalash	0.385	0.139	0.757	0.288	0.46	0.75	0.501

1. Three Locus Exact Test

2. Exact test p-values < 0.05 are shown in bold script

Table 5.5: Variance Estimates for the Locus FGA Allele Frequencies

Alleles	19	20	21	22	23	24	25	26	27	p-value
Punjabi										0.01
Variance1	0.0002	0.0002	0.0004	0.0004	0.0003	0.0003	0.0002			
Variance2	0.0002	0.0002	0.0003	0.0003	0.0004	0.0002	0.0002			
Pushtoon										0.07
Variance1	0.0003	0.0002	0.0003	0.0004	0.0005	0.0005				
Variance2	0.0002	0.0002	0.0003	0.0004	0.0005	0.0005				
Sindhi										0.013
Variance1				0.0007	0.0005	0.0008	0.0065	0.0003	**0.00005**	
Variance2				0.0008	0.0005	0.0009	0.0074	0.0004	**0.00010**	
Baluchi										0.064
Variance1	0.0009				0.0028	0.0024				
Variance2	0.0013				0.0034	0.0023				
Makrani										0.367
Variance1				0.0010	0.0012	0.0014	0.0019			
Variance2				0.0014	0.0012	0.0013	0.0021			
Brosho										0.239
Variance1				0.0014	0.0012	0.0011	0.0011	0.0007		
Variance2				0.0014	0.0012	0.0012	0.0011	0.0008		
Kalash										0.757
Variance1			0.0008	0.0007	0.0011	**0.0006**	**0.0005**			
Variance2			0.0008	0.0006	0.0011	**0.0003**	**0.0002**			

1. *p*-value is for the exact test 2. Empty cells show undetected homozygote for the allele

3. Variance1 = pu(1 – pu)/2n , whereas pu is the allelic frequency of allele p and n is the number of individuals profiled

4. Variance 2 = (pu + puu - 2 pu^2)/2n , whereas pu is the allelic frequency, puu is the homozygote frequency for 'allele p' and 'n' is the number of individuals profiled 5. A difference of two or > two times between the two variance estimates is shown in bold

Table 5.6: Heterozyghosity Test for the Loci D3S1358, vWA &FGA in Pakistani Populations

Locus	Statistic	Punjabi	Pushtoon	Sindhi	Baluchi	Makrani	Brosho	Kalash
D3S1358	N	200	170	98	35	48	54	97
	Ho	159	141	75	28	38	41	67
	He	157	133	74	28	36	39	68
	SE of He	5.8	5.4	4.3	2.3	3	3.3	4.5
vWA	Ho	173	139	77	27	31	40	78
	He	169	139	77	27	35	44	80
	SE of He	5.1	5	4	2.4	3.1	2.8	3.7
FGA	Ho	171	142	76	31	40	46	82
	He	174	147	85	28	42	47	80
	SE of He	4.7	4.4	3.3	1.86	2.34	2.39	3.66

N is the number of profiles for the locus

Ho is the Observed Heterozygotes

He is the Expected Heterozygotes

SE of He is the Standard Error of the expected number of heterozygotes

5.3.4 FORENSIC PARAMETERS FOR THE LOCI D3S1358, vWA & FGA

The results of the forensic parameters calculated (5.2.4) are shown in Table 5.7a & b.

For estimating these parameters for the combined population of Pakistan, the genotype data for Punjabi, Sindhi, Pushtoon, Baluchi and Makrani populations was pooled as these populations inhabit the same geographical region whereas the Brosho and the Kalash are isolated populations. Probability of discrimination (PD_{comb}) and the probability of exclusion (PE_{comb}) for the three loci were lower for the Baluchi, Makrani and Kalash populations as compared to other populations. The results (Table 5.7a & b) show how these important forensic parameters can vary between the sub populations of a large population.

5.3.5 POPULATION STRUCTURE AND F- STATISTICS

The alleles at a locus show relationship to each other in terms of the population and its respective subpopulations. The quantification of this relationship is given by the F statistics. These are F_{IT}, F_{IS} and F_{ST}. These provide with a correlation between the alleles relative to the sub populations and the whole population (Wright, S. 1965).

F_{IT} is the extent to which the alleles in an individual are related to each other in comparison to the alleles in other individuals of the total population. This is given by the average ratio of variance over all sub populations.

F_{IS} is the extent to which the alleles in an individual are related to each other in comparison to those of the other individuals in his own sub population. It is given by the ratio of the variance of the gene frequencies in a randomly mating sub population.

Table 5.7a: Forensic Parameters for the Loci D3S1358, VWA and FGA

Parameter	Punjabi	Pushtoon	Sindhi	Baluchi	Makrani	Brosho	Kalash	Pak[1]
D3S1358								
Homo [2]	0.196	0.171	0.232	0.2	0.204	0.241	0.309	0.196
Hetero [3]	0.804	0.829	0.768	0.8	0.796	0.759	0.691	0.804
PM [4]	0.087	0.088	0.114	0.102	0.125	0.13	0.152	0.079
PD [5]	0.913	0.912	0.886	0.898	0.875	0.87	0.848	0.921
PE [6]	0.607	0.655	0.54	0.599	0.591	0.526	0.414	0.606
PI [7]	2.55	2.93	2.15	2.5	2.45	2.08	1.62	2.55
PIC [8]	0.75	0.75	0.71	0.77	0.71	0.68	0.64	0.76
VWA								
Homo	0.13	0.171	0.206	0.229	0.38	0.259	0.194	0.185
Hetero	0.87	0.829	0.794	0.771	0.62	0.741	0.806	0.815
PM	0.05	0.061	0.085	0.113	0.123	0.066	0.064	0.058
PD	0.95	0.939	0.915	0.887	0.837	0.934	0.936	0.942
PE	0.735	0.655	0.588	0.547	0.316	0.494	0.61	0.628
PI	3.85	2.93	2.43	2.19	1.32	1.93	2.58	2.71
PIC	0.83	0.79	0.76	0.74	0.69	0.79	0.8	0.8
FGA								
Homo	0.146	0.171	0.21	0.125	0.163	0.151	0.146	0.167
Hetero	0.854	0.829	0.79	0.875	0.837	0.849	0.854	0.833
PM	0.037	0.037	0.041	0.068	0.048	0.044	0.058	0.03
PD	0.963	0.963	0.959	0.932	0.952	0.956	0.942	0.97
PE	0.702	0.655	0.581	0.745	0.669	0.693	0.703	0.661
PI	3.41	2.93	2.38	4	3.06	3.31	3.43	2.99
PIC	0.85	0.85	0.85	0.85	0.84	0.86	0.81	0.86

1, Populations Combined were Punjabi, Pushtoon, Sindhi, Baluchi and Makrani

Homozygosity 3, Heterozygosity

4, Probability of Match 5, Probability of Discrimination,

6, Probability of Exclusion 7, Paternity Index 8, Polymorphic Information Index

Table 5.7b: Combined Forensic Parameters for the Loci D3S1358, vWA & FGA in Pakistani Populations

Parameter/population	PD_{comb}	PE_{comb}	PM_{comb}
Punjabi	0.9998	0.9690	0.0002
Pushtoon	0.9998	0.9590	0.0002
Sindhi	0.9996	0.9206	0.0004
Baluchi	0.9992	0.9239	0.0008
Makrani	0.9993	0.9074	0.0007
Brosho	0.9996	0.9264	0.0004
Kalash	0.9994	0.9211	0.0006
Pak Combined *	0.99986	0.9503	0.00014

PD_{comb}, combined probabiltiy of discrimination (Loci D3S1358, vWA &FGA)

PE_{comb}, combined probabiltiy of exclusion (Loci D3S1358, vWA &FGA)

PM_{comb}, combined probabiltiy of match (Loci D3S1358, vWA &FGA)

Pak Combined refers to the pooled data for the mainland populations

Formulae for these parameters are as per Appendix 2

F_{ST} is the third F statistics, which compares the relatedness of the alleles in different sub populations relative to those of the total population. This is also called the co ancestry coefficient (Wright, S. 1965; Evett, I. W., Weir, B. S. 1998). The measure F_{ST} is of particular interest as it estimates the co ancestry of the alleles (Balding, D. J. & Nichols, R. A. 1994)

5.3.5.1 Estimation of F Statistics in Pakistani Population

The genotype data from the various sub populations of the Pakistani population was used to study the population structure. (Table 5.8a & b). The three F statistics were calculated using the software GDA. In order to measure the F_{ST} in the Pakistani populations and to have realistic estimates all the population pairs were tested (Table 5.8b). The F_{ST} was found similar at all the three loci, which was expected as the loci share common population histories. The overall value of F_{ST} over the three loci was 0.01. Higher values were obtained for the Makrani and Kalash. The Kalash exhibited very high F_{ST} levels against all the populations whereas the Brosho had low levels of F_{ST} in all the pairwise comparisons.

5.3.5.2 Effect of Substructure on Genotype Frequency Estimation

The estimation of the F-statistics and the tests for HW principle showed that substructure exists in the Pakistani population. This can have an effect on the estimation of the genotype frequencies in forensic cases as in that case the genotype frequencies calculated from the observed allele frequencies would not give the true values. In this connection formulae were described so that the F_{ST} could be incorporated in the estimation of the allele/genotype frequencies (Balding, D. J. & Nichols, R. A. 1994).

The effect of the proposed formulae by Balding & Nichols (BN) (Appendix 2) was studied by simulating cases. For this the genotype frequencies for the locus FGA in Sindhi population were recalculated by three different methods.

Table 5.8a: F statistics For Combined Pakistani Population across Three Autosomal STR Loci D3S1358, vWA & FGA

Locus	F_{IS}	F_{IT}	F_{ST}
D3S1358	-0.03	-0.01	0.019
vWA	0.01	0.018	0.007
FGA	0.04	0.05	0.01
Composite	0.008	0.02	0.011

Table 5.8b: Pairwise F_{ST} Values for Pakistani Populations across Three Autosomal STR Loci D3S1358, vWA & FGA

	Baluchi	Brosho	Kalash	Makrani	Punjabi	Pushtoon
Baluchi						
Brosho	0.007					
Kalash	**0.034**	**0.027**				
Makrani	0.009	0.014	**0.04**			
Punjabi	0.012	0.007	**0.025**	**0.022**		
Pushtoon	0.002	0.009	**0.027**	**0.026**	0.013	
Sindhi	0.002	0.003	**0.023**	0.006	0.012	0.009

Values in bold show significantly high levels of F_{ST} for a population pair.

The first was the HW rule (2pq for heterozygote and p^2 homozygotes) using the observed allele frequencies. In the second method, the BN formulae for genotype estimation were applied using the observed allelic frequencies of the Sindhi data and an F_{ST} value of 0.01. This value of F_{ST} was calculated before (Table 5.8a).

The third method was to recalculate the allele frequencies using the formula $(p + 2)(n + 4)$, where 'p' is the observed allele frequency and 'n' is the number of persons profiled and then estimating the genotype frequencies using BN formulae for genotype estimation and an F_{ST} value of 0.01.

The results showed that there were major differences between the observed and the expected genotype frequencies when the genotype frequencies were calculated under the assumption of HW principle. The genotype frequencies conformed to those expected under HW principle by applying the BN formulae. For most of the genotypes, the frequency was higher than would be expected under HW principle. The genotype frequencies calculated by the third method give higher estimates than by only applying the genotype corrections, which is the result of an added effect of higher frequency and genotype estimation by the BN formulae (Table 5.9).

5.3.5.3 Correlation of HW & Corrected Genotype Frequency Estimate

A peculiar correlation exists between the expected genotype frequency under HWE and that obtained by applying the BN formulae. The ratio between the two frequencies increases as the rarity of the genotype frequency increases (Table 5.10). Thus the application of BN formulae yield conservative estimates of the genotype frequency particularly with rare alleles.

Table 5.9: Estimation of FGA Genotype Frequency in Sindhi Population by Three Methods

Genotype	Observed		HW	BN [1]	BN [2]
	Freq	Std Dev	Expected		
19,21	0.020	0.027	0.019	0.023	0.024
19,23	0.010	0.020	0.013	0.016	0.017
19,24	0.060	0.047	0.024	0.028	0.029
19,25	0.030	0.033	0.018	0.021	0.022
20,20.2	0.010	0.020	0.0006	0.002	0.001
20,21	0.070	0.050	0.021	0.024	0.024
20,22	0.010	0.020	0.020	0.024	0.024
20,23	0.010	0.020	0.020	0.017	0.017
20,24	0.020	0.027	0.026	0.003	0.03
20,26	0.010	0.020	0.008	0.010	0.01
21,21	0.030	0.033	0.026	0.033	0.028
21,22	0.040	0.038	0.050	0.053	0.054
21,23	0.020	0.027	0.035	0.039	0.041
21,24	0.040	0.038	0.064	0.068	0.07
21,25	0.040	0.038	0.048	0.050	0.054
21,26	0.030	0.033	0.0192	0.023	0.024
22,22	0.050	0.043	0.024	0.031	0.026
22,23	0.070	0.050	0.034	0.038	0.038
22,24	0.040	0.038	0.062	0.066	0.067
22,25	0.030	0.033	0.046	0.049	0.05
22,27	0.010	0.020	0.005	0.008	0.006
22,28	0.010	0.020	0.001	0.005	0.003
22.2,24	0.010	0.020	0.062	0.010	0.008
22.2,25	0.020	0.027	0.006	0.007	0.006
23,23	0.010	0.020	0.012	0.017	0.014
23,24	0.050	0.043	0.044	0.048	0.05
23,25	0.010	0.020	0.033	0.037	0.038
23,26	0.02	0.027	0.013	0.016	0.017
24,24	0.050	0.043	0.04	0.048	0.044
24,25	0.070	0.050	0.060	0.064	0.067
24,27	0.010	0.020	0.006	0.010	0.008
25,25	0.040	0.038	0.022	0.029	0.026
25,26	0.010	0.020	0.018	0.02	0.022
26,26	0.020	0.027	0.004	0.007	0.005
27,27	0.010	0.020	0.0002	0.0015	0.0004

Table 5.10: Relationship of the Expected and Recalculated Genotype Frequencies

Genotype	Expected Frequency [1]	BN Frequency[2]	Ratio 1& 2
22/24	0.062	0.066	1.06
22/23	0.034	0.038	1.11
19/24	0.024	0.028	1.16
20/26	0.008	0.01	1.25
22/27	0.0046	0.008	1.73
27/27	0.00022	0.0015	6.8

1. Frequency Expected as per HW principle
2. Frequency Calculated using the formulae proposed (Balding, D. J. & Nichols, R. A. 1994).

The allele/genotype frequencies were obtained from the Sindhi population FGA locus data (Table 5.1c & 5.2c) for studying these statistics.

5.3.5.4 Effect of Substructure on Multilocus Profile Estimation

In order to see the effect of the substructure on the estimation of multilocus genotype frequency a simulation study was conducted. Four multiplocus genotypes from the Sindhi populations were selected and the genotype frequencies calculated with and without taking into consideration the substructure effects. Likelihood ratio (LR) is a good indicator of the evidential value and is defined as the probability of match if the DNA from the crime scene and that from the defendant are same (Gill, P. & Evett, I. 1995). Therefore for each multilocus profile selected likelihood ratio was calculated and lograthmic graphs plotted for the likelihood ratios in each case (Figure 5.2).

5.4 DISCUSSION

5.4.1 ANALYSES OF THE DATA

In genetic profiling of the samples the calculation of the genotype and allele frequencies is the basis of all the subsequent statistical analyses. This includes the parameters of forensic usefulness of the loci and the structure of the population. The number of the samples of each population varied from 35 to 200. This resulted in an opportunity to examine the effects of the different sizes on the frequency estimates and other statistical variations.

There were significant differences in the allelic frequencies of the three loci among the populations studied. Some alleles were not observed in the Baluchi population which most probably was the result of the small size of the sample (35). In the Sindhis the frequencies of allele 11, 12 and 13 at the locus D3S1358 were the higher than other mainland populations. Allele frequencies at loci D3S1358 and vWA in the Pakistani sub populations differed however at locus vWA the differences were not as significant as at loci D3S1358 and FGA (Table 5.1).

Figure 5.2: Effect of Substructure on the Likelihood Ratio

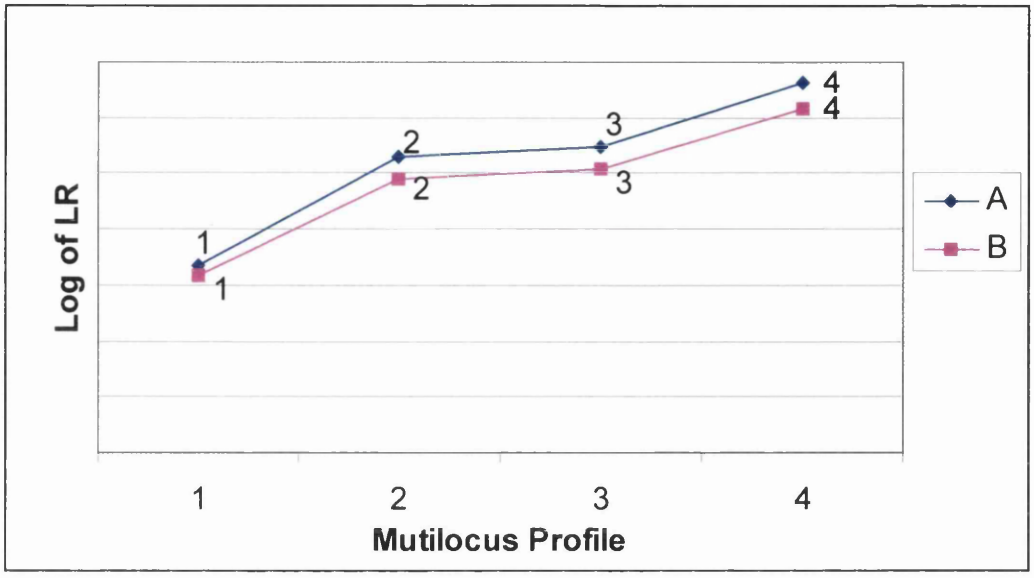
In order to gauge the effect of the substructure on the estimation of multilocus genotype frequency a simulation study was conducted. Four genotypes from the Sindhi populations were selected. Two of the profiles chosen were rarer than the other two based on allele frequency and genotype frequency observed for each allele. The allele/genotype frequencies were obtained from the Sindhi population FGA locus data (Table 5.1c & 5.2c) for studying these statistics. The genotypes selected were as follows

1. D3 15/16, VWA 17/18, FGA 20/21
2. D3 16/16, VWA 15/16, FGA 22/23
3. D3 14/15, VWA 13/16, FGA 19/21
4. D3 17/18, VWA 18/18, FGA 24/27

First, the multilocus genotype frequencies were calculated by using the HW formulae 2pq (heterozygote frequency) or p^2 (homozygote frequency) and the multilocus frequency by applying the product rule. The allele/genotype frequencies were recalculated by applying the BN formulae (Balding, D. J. & Nicholas, R. A. 1994) and then the product rule was applied to calculate the multilocus profile frequency. These are represented as HW frequency and BN Recalculated Frequency in Figure 5.2.

Since in case of being true the probability is 1 thus the likelihood ratios (LRs) for any frequencies are calculated as the inverse of that frequency (NRC Report, 1996). As interpretation becomes easier in these terms so the LRs of the profile frequencies were estimated (Table 5.11). The LRs for the four multilocus genotypes were plotted as their log values for profiles before (A) and after (B) the application of the BN formulae.

The figure shows that the LRs decrease when the substructure effects were incorporated in the estimation of profile frequency resulting in conservative estimates.



Genotype	HW Frequency (A)		BN Recalculated Frequency (B)	
	Frequency	LR	Frequency	LR
1	0.0004	2500	0.0006	1666
2	0.000005	200000	0.000012	83333
3	0.000003	333333	0.000008	125000
4	0.0000002	5000000	0.0000007	1428571

At the locus FGA the frequencies varied more between the populations studied, than the other two loci. In the Baluchi population allele 23 was detected at a high frequency (0.234). All other alleles were at a higher frequency in the Punjabi and Pushtoon populations. Allele 25 was at high frequency (0.245) in the Makrani population. Makrani and the Baluchi populations are statistically similar at all the three loci however imporatnt differences existed like high frequency of allele 23 in the Baluchi population and allele 25 in the Makrani population.

In the Brosho population high frequency of allele 15 at D3 was significant as this allele was at a high frequency (0.407) only in the Chinese population (Pu, C. *et al.*, 1998) and Taiwanese population (Pu, C. *et al.,* 1999). In all other populations, the frequency of this allele was much lower. In comparison to the Pakistani population the pattern of the allelic frequencies at these three loci in the Brosho was much more similar to the Chinese population. Statistically the Brosho were similar to Chinese population for the loci D3 (exact test *p-value* 0.045), vWA (*p-value* (0.05) and FGA (*p-value* 0.4), when the data was tested for the frequent alleles using the R x C contingency test (Miller, M. 1998).

The Kalash population showed a different pattern of the allelic frequency at the three loci. At the locus D3S1358 the frequencies of the common alleles was much higher than any other population from Pakistan, whereas the less frequent alleles were at a lower frequency. Some of the alleles were at a much higher frequency than the other Pakistani populations at the locus vWA like allele 15 and 19. Other alleles were either lower or similar frequency. At locus FGA also, the frequencies of complete alleles 18, 20, 22 were much higher than any other Pakistani population. The variant allele 24.2 was present at a particularly high frequency in the Kalash. This has been detected in very low frequency the Asian Indians (Evett, I. W. *et al.,* 1996), Caucasians (Schroer, P *et al.,* 2000), Japanese (Yamamoto, T. 1999), Chinese (Pu, C. *et al.,* 1998), and Turks (Rolf, B. *et al.,* 1997). This variant was detected in higher frequency (0.013) in the Australian Aborigines (Rolf, B. *et al.,* 1997). This may be of significance when the high frequency of this variant in the Kalash (0.094) is compared

to these populations. Though the high frequency might have been the result of founder effect. As expected populations of Pakistan resemble each other more than the other racial groups studied for the same autosomal STR loci (Budowle, B. & Monson, L. 1994).

When the populations of Pakistan (excluding the Brosho & Kalash) were pooled together and compared with, other racial groups, interesting features were detected. The allele frequencies of the three loci were different from the African/ Caucasian groups (Figure 5.3a, b & c).

5.4.2 SIZE OF THE DATABASE

The National Research Council (NRC) and the Scientific Working Group on DNA Analysis Methods (SWGDAM), recommend that databases should be established for the populations for different racial groups. The NRC recommends that the database may consist of a few hundred individuals. However, the simple statistical rule is that bigger the database more accurate would be the estimated frequency from the database and generally a size of 100 has been recommended as sufficient for the statistical analysis in casework when such a database is to be used (Lander, E. S. 1994; Weir, B. 1992).

During this work, the size of each population was different. The effect of the size of the population on the variance and consequently the standard deviation was that the smaller the size of the database greater was the variance. When the Baluchi population (35) was compared to Punjabis (200) or Pushtoons (170), the variance estimates were found to be at least a magnitude lower for the populations having bigger size of the database. This fact calls for bigger but realistic databases for forensic use. The comparison of the variance estimates for the databases of the seven Pakistani populations showed that the databases of >100 profiles show low variance hence smaller standard deviation. This size of the database was therefore considered sufficient for estimating the allelic frequencies of STR markers in a given population.

Figure 5.3a: Locus D3S1358 Allele Frequencies in Different Populations

Allele frequencies of most of the common alleles in Pakistani population were different from other groups. Allele 14 was at a higher frequency in Pakistani and African populations. There was a similarity to the Caucasian population in the frequency of allele 18. Allele 19, which was infrequent in African and Chinese, was at higher frequency in Pakistani as well as in Caucasian populations.

Figure 5.3b: Locus vWA Allele Frequencies in Different Populations

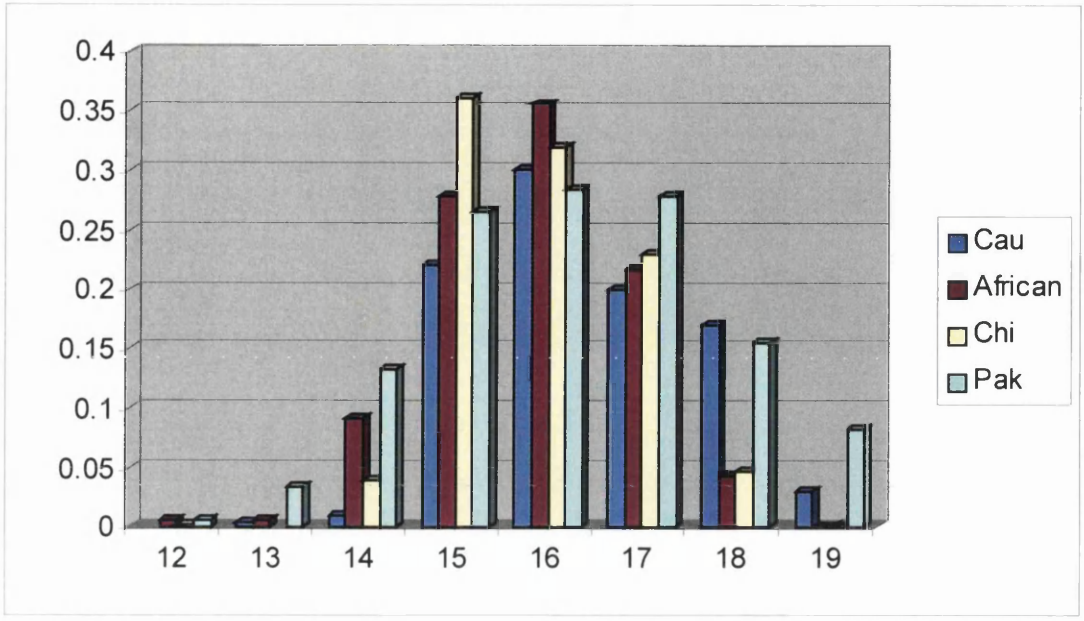
The allele frequencies in Pakistani population were distinctly different from other population groups. Allele 13 was detected at a higher frequency while allele 18 and 19 were at significantly lower frequencies than in the other populations

Figure 5.3c: Locus FGA Allele Frequency in Different Populations

Distribution of FGA alleles was as diverse in the Pakistani population as in the African population. The frequencies of most of the alleles were similar to the Caucasian population. The variant allele 23.2 was detected at a significantly higher frequency in the Pakistani population.

The data for this comparison was obtained for Zimbabwe Africans, Southern Spanish and Chinese populations (Budowle, B. *et al.*, 1997; Entrala, C. *et al.*, 1998; & Pu, C. *et al.*, 1999).

Locus D3S1358



Locus vWA Allele

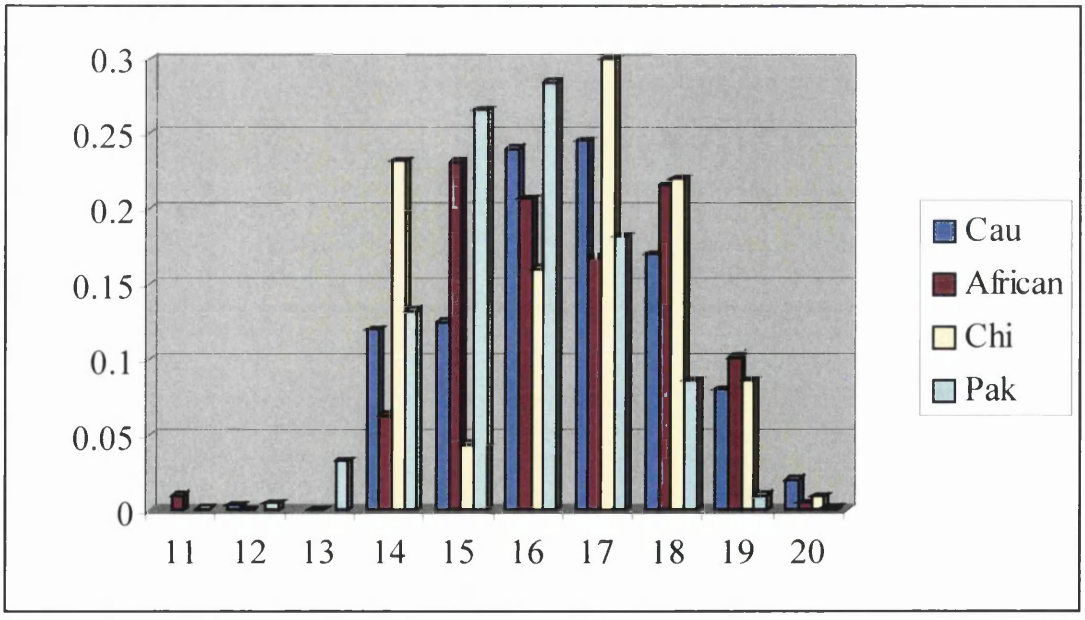
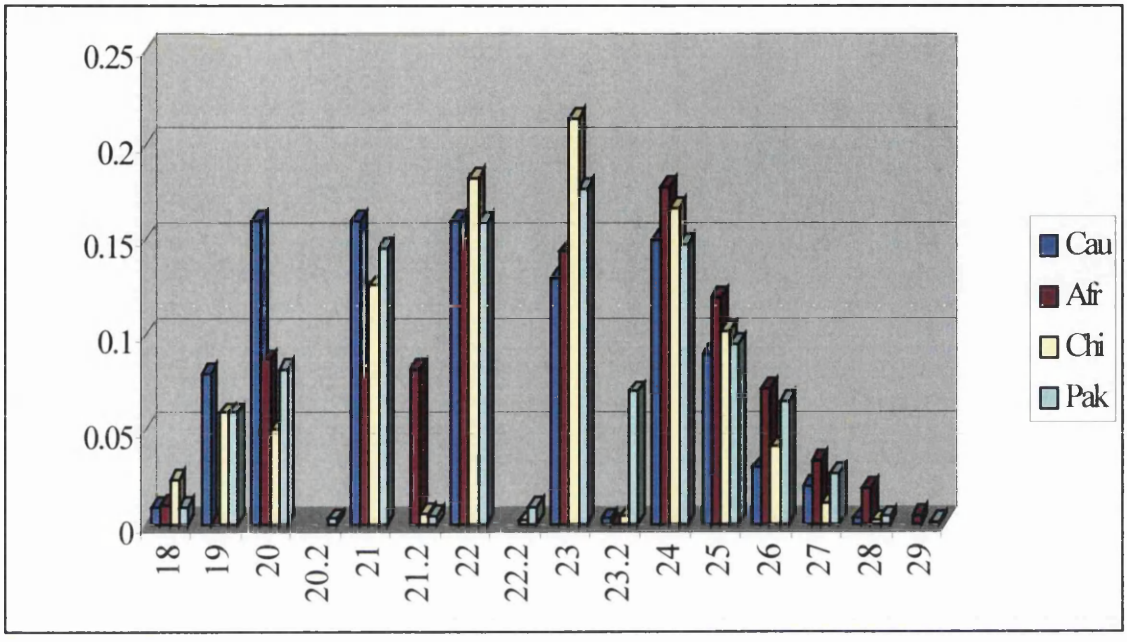


Figure 5.3c: Locus FGA Allele Frequencies in Different Populations



Most of the statisticians agree that the profile may be estimated along with its variance and standard deviation from the database used and do not specify any particular size of the database (Weir, B. S. & Brenner, C. pers comm). Thus the frequency of the profile might be calculated alongwith its standard deviation. In casework it would be possible to give a range of the frequency instead of one value. Also the likelihood ratios could be calculated as a range which might be a better way of presenting evidence in legal forums.

5.4.3 INDEPENDENCE WITHIN AND BETWEEN THE LOCI

In most studies for the autosomal STRs, the populations have been shown to follow the HW principle or the departures detected were insignificant. These departures or disqeuilibria refer to the deviation from the ideal genotype probabilities estimated from the allele frequencies (Haribson, S. A. & Buckleton, J. S. 1998). The fact is that there is additional information available from the databases like substructure effects and that if more individuals are profiled from the population it can lead to better estimates about the population frequencies of the loci tested. The application of the simple HW rule leads to the negation of these facts. In practice, therefore it is better to make corrections for the subpopulation effects if such effects are demonstrated in the populations as this removes the need for the assumption of independence.

There is a real need to reassess the power and value of the exact tests, which is employed to test the data to be used for forensic purposes is as a rule at present. This test does give an estimate but the fact that it gives only a probability cannot be ignored. The probability level of 0.05 is taken as a standard now because the statistics assume that if an event occurs at this probability then this is unlikely to occur by chance alone. During this study, the exact test gave a high value for the Brosho population but the population failed the test of variance for independence. That meant that the population might have departures from the HW principle as if the reverse was true, the two formulae for variance would give same result (Weir, B. S. 1996).

In the exact test, the observed values are tested against the assumed HWE frequencies, a procedure, which might result in p-values of >0.05 for a certain sample of a population and at other times it may give lower p-values for another sample and *vice versa.* Thus low values of the exact test might also result from mere sampling errors as described for other populations (Evett, I. W. *et al.,* 1996; Budowle, B. *et al* 1999).

It is also important that strict assumption about a population following the HW principle is not a requirement for forensic computations and it is not necessary that the loci must follow the principle, enabling their use in forensic identification (Chakraborty, R. & Stivers, D. N. 1998). However the HW principle is the basis for the computation of the genotype frequencies, given the allele frequencies. While it may detect departures from HW it appears better to estimate and incorporate the effects of substructure while computing the allele/genotype frequencies, which can be done in the light of the recommendations of the NRC (NRC Report II, 1996a).

5.4.4 POPULATION STRUCTURE

Subpopulations of a large population exhibit differences in allele frequency due to three major factors (i) the number of sub populations, (ii) the time of divergence and (iii) the rate of gene flow between the sub populations (Wall, W. J. 1993). In any related individuals, there is a chance that an allele carried by both is identical by descent (ibd). This phenomenon increases the chance of homozygosity in the individuals of an inbred population.

The tests of independence give only a probability for the degree of the departure of allele frequencies from the HW principle. This degree of departure from independence in a population can be quantified with the help of F-statistics. In forensic calculations, it has been recommended that the realistic estimates of F_{ST} should be calculated in order to incorporate real time estimates of the effects of sub structuring in profile estimation (Gill, P. & Evett, I. 1995). Furthermore it is a matter

of high concern that when the DNA evidence is presented to the courts the frequency of the profile estimates should not be an under estimate.

The ethnic and geographical subdivisions within a major population (Caucasian) have been shown to have little effect on the forensic estimates of the likelihood of the occurrence of the profile (Monson, K. L & Budowle, B. 1998). Still the NRC has recommended that whenever possible forensic databases be used from the populations to which the perpetrator belongs (NRC Report II, 1996c).

If the database of the population to which the perpetarator belongs is not available or the population of origin of the perpetarator is not known then the NRC recommendation is to compensate the effects of substructure while using the general population database (Balding, D.J. & Nichols, R.A. 1994). In essence, the proposal describes a statistical method for using the allele frequency data without assuming independence, taking into account the F_{ST}. Thus, this method has the capability for the data to be used without making an inference of the probability of random mating through exact or other tests. In this regard, the recommendation of NRC is that an F_{ST} upper bound value of 0.01 may be used or the maximum value of 0.03 may be used, if limited data is available. Still it would be best to use the databases from the relevant population and realistic estimates of F_{ST} calculated so that estimation of the profile frequency is accurate.

The Pakistani population was thought to have a higher degree of inbreeding as cousin marriages are a common occurrence and the caste/tribal system is still in place. However, as the results show there is only a moderate level of inbreeding in the populations. The four main populations, the Baluchis, Pushtoons, Punjabis and Sindhis all display a low F_{ST} value ranging between 0.002-0.013 (Table 5.8 b).

The over all value of F_{ST} obtained for the Pakistani population is however, much higher (0.01) when compared to that (0.002) reported in a previous study for the Asian Indian population (Gill, P. & Evett, I. W. 1995). The fact to be taken into account in this regard is that the high F_{ST} value obtained during this work is for all Pakistani sub populations together. In contrast, the Birmingham population of Asian Indians profiled by Evett, I. W. *et al.,* has a predominant majority of Punjabi population from Hindus, Sikhs and Pakistani Muslims. These results also show the value of having realistic estimates of the F_{ST} as has been advised earlier (Gill, P. & Evett, I. W. 1995).

5.4.5 ESTIMATION OF ALLELE AND GENOTYPE FREQUENCIES UNDER THE ASSUMPTION OF DEPENDENCE

The effect of the estimation of the allelic frequencies and genotype frequencies using the formulae proposed by Balding & Nichols was gauged by recalculating the allele frequencies and genotype frequencies for the Sindhi population (Balding, D. J. & Nichols, R. A. 1994). The assumption in using these formulae was of dependence rather than independence. The effect was that the frequencies of the profiles became higher (Table 5.11). The representation of a genotype from a population thus became more realistic as the substructure effects were incorporated. In this simulation study in three out of four profiles, the incorporation of F_{ST} reduced the LRs to more than half as against simple product estimation.

The results have shown that the use of these formulae the genotype frequencies come to lie within the bounds of those expected if HWE was true or rather more conservative. This corrective effect increased with the rarity of the allele frequency or the genotype frequency. If an allele was absent from the database the application of the formula gave an estimate of 0.01, as recommended by the NRC (NRC Report II, 1996a).

In conclusion a database of three autosomal STR loci has been established for the seven Pakistani subpopulations having 702 profiles for three loci D3S1358, vWA & FGA which has shown these loci together are an extremely powerful tool in the context of human identification. The analysis of the allele frequencies revealed that they resemble each other more than the other populations. However differences existed between the allele frequencies of Pakistani populations which underscores the importance of the development of regional databases. Substructure was detected in Pakistani population. The effects of incorporation of realistic estimates of substructure effects in the estimation of genotype and multilocus profile estimation have shown that it results in conservative estimates, and therefore this method is viewed as the way forward in forensic haemogenetics.

CHAPTER 6: Y CHROMOSOME STR ANALYSES

6.1 INTRODUCTION

Y chromosome STRs have been analysed in various populations of the world (Kloosterman, A. D. *et al.,* 1998; Pestoni, C. *et al.,* 1998; Rosi, E. *et al.,* 1999; Horst, B. *et al.,* 1999; Tun, Z. *et al.,* 1999; Pawlowski, R. *et al.,* 1999; Furedi, S. *et al.,* 1999; Sasaki, M. and Dahiya, R., 2000; Gehrig, C. *et al.,* 2000; Gonzalez-Neira, A. *et al.*, 2000; Tagliabracci, A. *et al.,* 2000; Umesta, K. *et al.,* 2000) and many other Y databases exist besides these. However indigenous Pakistani populations have not been studied for Y chromosome STRs that have been validated for forensic use.

The seven Y STR loci which included 6 tetrameric STRs DYS19, 389I, 389II, 390,391, 393 and the trimeric STR DYS392 were chosen which define the haplotype Yh1 (Y chromosome haplotype1) (Kayser, M. *et al.,* 1997). These loci, which can be amplified in two separate multiplex reactions were chosen as they are widely used in the forensic community and the data are available from most population groups for comparison therefore the forensic utility could be gauged in Pakistani populations *vis*-à-*vis* other populations (Kayser, M. *et al.,* 1997).

6.1.1 PRIMER SETS USED FOR AMPLIFICATION

The primers used were the same as reported by Kayser. M. *et. al.*, 1997 except for one locus DYS391 (Table 2.2). This locus was previously reported to occasionally produce a non specific product of approximately 256, 260 or 264 bp and the primers also amplified female DNA with products in the same range (Redd. A. J. *et al.*, 1997; Kayser, M. *et. al.*, 1997). Further in the presence of higher concentration of female DNA in a mixture the primers failed to detect the male component (Gusmao, L. *et al.,* 2000). An alternative reverse primer designed 23 bp upstream of the old primer

was found to be strictly male specific, robust and the product size was smaller than the previously reported set (Gusmao, L. *et al.,* 2000). The primer had been tested for its efficiency in multiplexes (pers comm Gonzalez-Neira. A). In view of obvious advantages it was decided to use this primer.

6.2 Y STR PCR OPTIMISATION

Before amplifying the seven Y STRs from a large number of samples the amplification reactions had to be optimised. Two commercial multimix kits the Gibco PCR Supermix (Life Technologies, UK) and Reddymix™ Mastermix (AB Gene®, UK) were tested. The results of singleplex amplification with the Reddymix™ Mastermix were much better than the Gibco PCR Supermix. Further, the Reddymix™ had a crimson dye in it, which enabled direct loading of the PCR product on the agarose gel without having to add the loading buffer which decreased the analysis time. Reddymix™ was also stable at 4 °C for over 8 weeks.

6.2.1 OPTIMISATION OF SINGLEPLEX REACTIONS

It was important to have an optimised and reproducible amplification reaction for each locus. All optimisation reaction amplifications were performed using the GeneAmp® PCR System 2400 (Applied BioSystems CA, USA) in thin walled 0.2 ml reaction tubes in 25 µl reactions. All PCRs included two controls during optimisation process. To one control ~10 ng of female DNA was added and to the other water was added. The input template DNA quantity was initially kept at ~10 ng and the concentrations of all primers were 0.25 µM.

The annealing temperature was increased in 2 °C increments from 55-63 °C. The $MgCl_2$ concentration was tested at 1.5, 2 and 2.5 mM at all annealing temperatures.

The concentration of the primers for the locus DYS393 were decreased to 0.125 µM in order to have bands of similar intensity in the product gel. Once the PCR

amplifications had been analysed on the agarose gel the reactions were run on the automated sequencer in order to assess the optimisation results (Table 6.1a & b, Figure 6.1).

6.2.2 OPTIMISATION OF MULTIPLEX PCR

The optimisation of the multiplex reactions were performed in two multiplex reactions. The three loci DYS391, 392 & 393 were amplified together (Multiplex I) and loci DYS19, 389I & II and 390 were amplified in a separate reation (Multiplex II).

The optimisation of both multiplex reactions was performed in light of the amplification conditions determined previously (Kayser, M. *et al.*, 1997) and the single locus PCR optimisation (6.2.1). Initially for 25 µl reactions the primer concentrations for the all the loci were 0.25 µM. The template DNA input was approximately 10 ng. Volumes of the ingredients were kept at a scale where that of the ReddyMix™ did not decrease below 18 µl.

The same PCR cycling conditions were used for both multiplex reactions and the annealing temperature was varied from 50-60 °C in 2 °C increments. The $MgCl_2$ concentration was tested at 1.5, 2 and 2.5 mM at all annealing temperatures. Final extension temperature and time were optimised as for singleplex reactions (6.2.1) (Table 6.1a & b).

6.3 PREPARATION OF ALLELIC LADDERS

The use of sequenced ladders is recommended for the designation of an unknown allele against a fragment of known sequence (Bär, W. *et al.*, 1997). Two sequenced samples (provided by Annabel Gonzalez and Professor Angel Carracedo, Compostella, Santiago, Spain) were available and were used to size the alleles during the preparation of allelic ladders.

Table 6.1a: Optimised PCR Cycling Conditions for Single & Multiplex Reactions

Locus	Initial Incubation	Denaturation	Annealing	Extension
DYS391	95 °C 2 min	94 °C 2 min	58 °C 1 min	72 °C 1 min
DYS392	95 °C 2 min	94 °C 1 min	58 °C 1 min	72 °C 1 min
DYS393	95 °C 2 min	94 °C 15 s	58 °C 20 s	72 °C 20 s
DYS19		94 °C 1 min	54 °C 1 min	72 °C 1 min
DYS390	95 °C 2 min	94 °C 1 min	58 °C 1 min	72 °C 1 min
DYS389I &II		94 °C 1 min	56 °C 1 min	72 °C 1 min
Multiplex I	95 °C 2 min	94 °C 1 min	57 °C 1 min	72 °C 1 min
Multiplex II	95 °C 2 min	94 °C 1 min	55 °C 1 min	72 °C 1 min

* A final extension temperature was 65 °C was carried out for 10 min and the number of cycles were 30 for all reactions.

Table 6.1b: Optimised PCR' Reaction Ingredients for Single & Multiplex Reactions

Locus	Singleplex		Multiplex	
	Primer Conc. µM	$MgCl_2$ Conc. mM	Primer Conc. µM	$MgCl_2$ Conc. mM
DYS391	0.25	2	0.25	
DYS392	0.25	2	0.25	2.5
DYS393	0.125	1.5	0.125	
DYS19	0.3	2.5	0.3	
DYS390	0.25	2	0.125	2.5
DYS389I & II	0.25	2	0.125	

Figure 6.1a: Optimisation of $MgCl_2$ Concentration for the Amplification of the Locus DYS19

Figure shows the effect of variation $MgCl_2$ concentrations on the amplification of DYS19 at 54 °C.

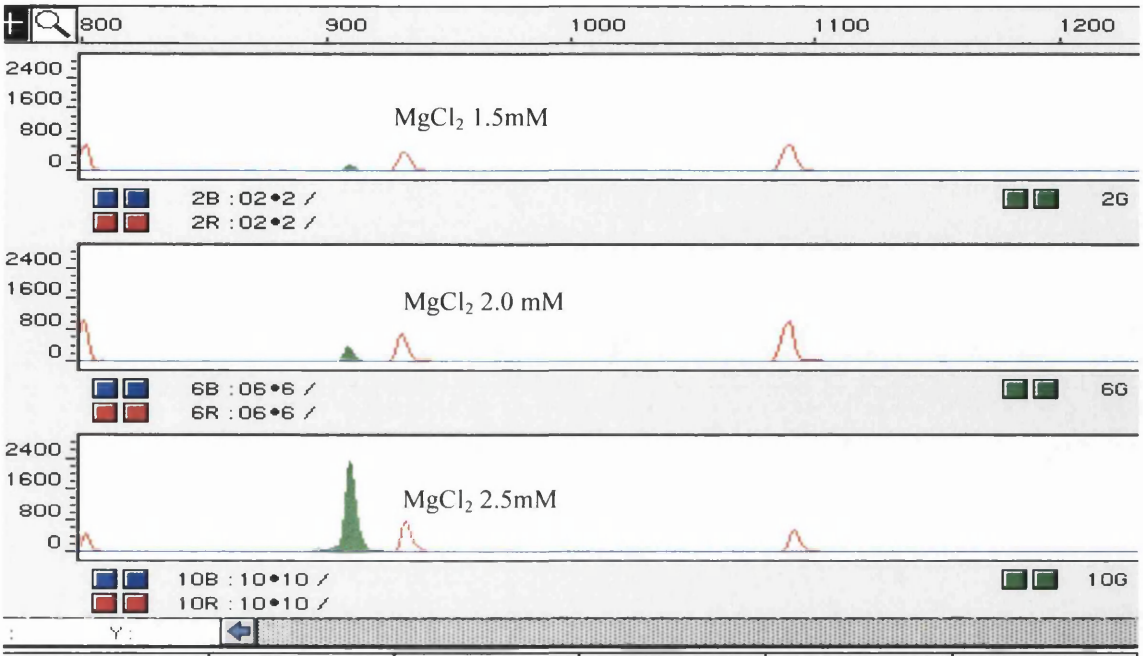
The locus was tested at three $MgCl_2$ concentrations 1.5 mM, 2.0 mM & 2.5 mM. At 2.5 mM $MgCl_2$ the allelic peak was of higher amplitude and the non specific amplification was minimal. 2.5 mM $MgCl_2$ concentration was determined as optimal. This was tested at different annealing temperatures.

The red peaks show the internal standard GeneScan™ 500 [Rox]. The locus peaks are green.

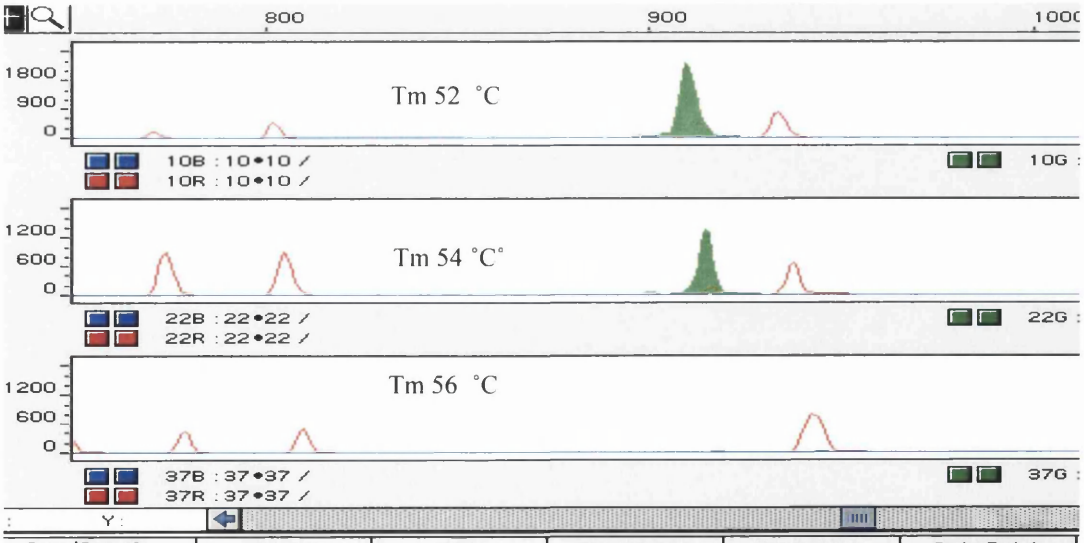
Figure 6.1b: Optimisation of Annealing Temperature for the Amplification of Locus DYS19

Figure shows the amplification of locus DYS19 in a singleplex PCR at different annealing temperatures. The $MgCl_2$ concentration was kept at 2.5 mM.

The locus amplified efficiently between 52-54 °C. For this locus the optimal annealing was found to be at 54 °C as with further increase of the annealing temperature the peak amplitude was very low as shown in the figure.



Dye/Sample Peak	Minutes	Size	Peak Height	Peak Area	Data Point
2G,3	91.00	192.98	173	1080	910
6G,3	90.90	193.80	398	2382	909
10G,4	91.00	192.82	2345	14296	910

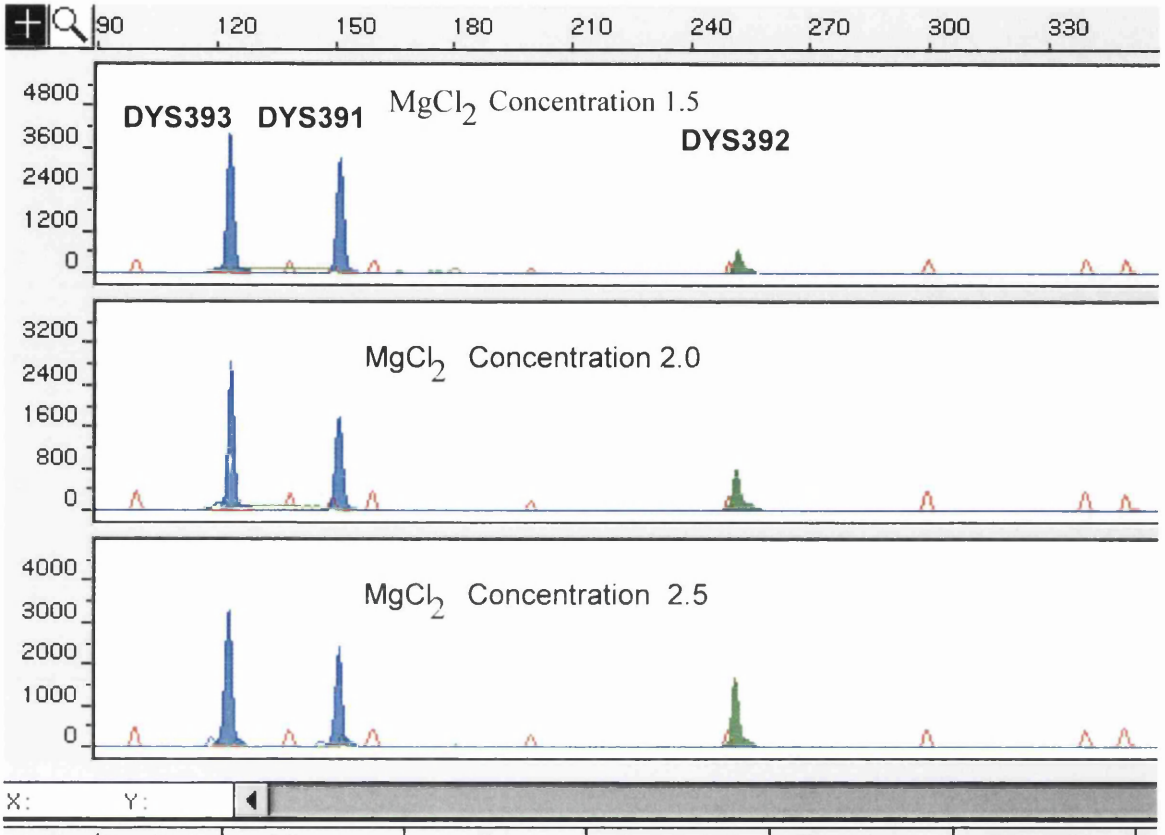


Dye/Sample Peak	Minutes	Size	Peak Height	Peak Area	Data Point
10G,4	91.00	192.82	2345	14296	910
22G,2	91.50	193.18	1379	7505	915
37G,3	91.60	192.60	61	373	918

Figure 6.2: Optimisation of $MgCl_2$ concentration for Multiplex I

Effect of variation of $MgCl_2$ concentrations is shown for the multiplex amplification of DYS391, 392 &393 at 57 °C (Table 6.1). Increasing the $MgCl_2$ concentration improved the yield of the product of DYS391 and 392 and 2.5 mM $MgCl_2$ concentration resulted in more balanced peaks than the peaks at lower concentration. At 1.5 mM $MgCl_2$ concentration the non specific amplification of the locus 392 was common.

The thermal cycling protocol was as per Table 6.1a. The red peaks show the internal standard GeneScan™ 500 [Rox]. The locus peaks for the loci DYS391 & 393 are blue and those for the locus DYS392 are green.



Dye/Sample Peak	Minutes	Size	Peak Height	Peak Area	Data Point
27B, 7	61.70	123.47	4181	19215	617
27B, 9	69.10	151.55	3419	16230	691
27G, 11	98.40	252.35	1100	6404	984
28B, 5	61.70	123.97	2936	13369	617
28B, 6	69.10	151.54	1870	9455	691
28G, 17	98.30	252.06	858	7280	620
30B, 5	62.00	123.77	3350	16895	692
30B, 7	69.30	151.55	2442	11434	693
30G, 12	98.50	251.74	1746	11980	985

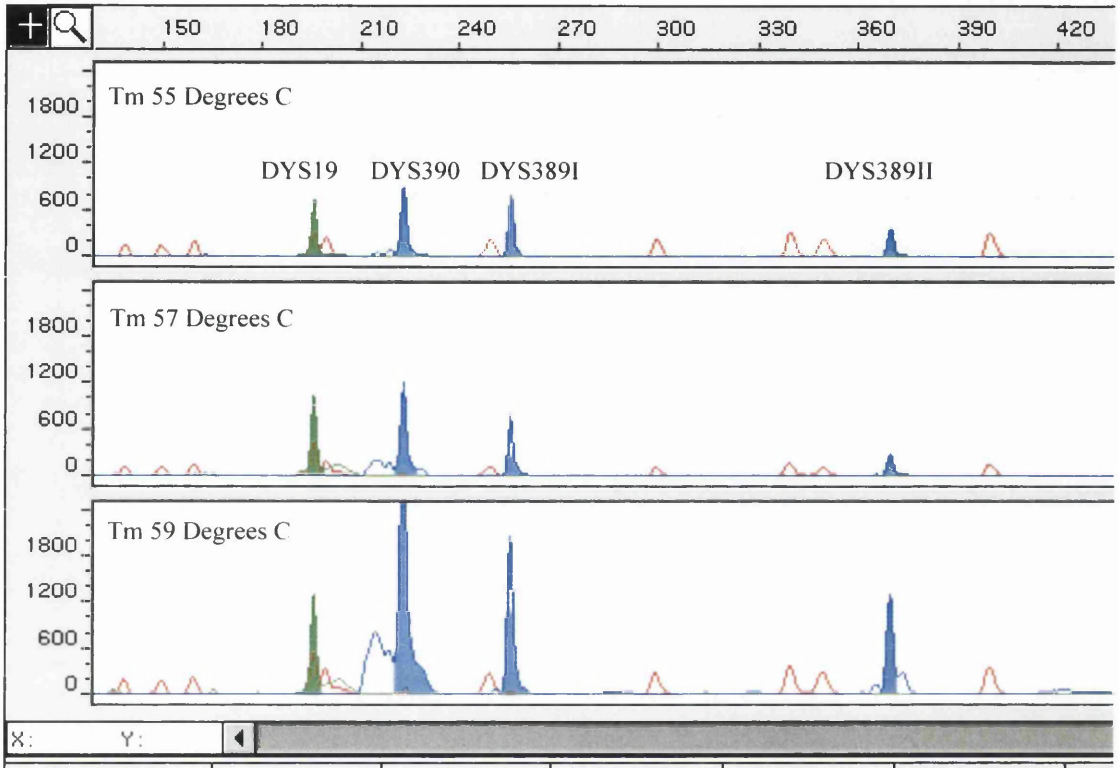
Figure 6.3: Optimisation of Annealing Temperature Multiplex II

The figure shows the optimisation of annealing temperature of multiplex amplification of loci DYS19, 390, 389I & 389II.

Higher annealing temperatures increased the peak amplitude but non specific amplification increased for the loci DYS389II & 390, further at 55 °C more balanced peaks resulted for all the loci therefore this multiplex was optimised at lower annealing temperature of 55 °C.

Multiplex II was found to be more sensitive than Multiplex I requiring lower amounts of template DNA input.

The thermal cycling protocol was as per Table 6.1 a and only the annealing temperature was varied. The red peaks show the internal standard GeneScan™ 500 [Rox]. Peaks for the locus DYS19 are green while those of the other loci are blue in colour.



Dye/Sample Peak	Minutes	Size	Peak Height	Peak Area	Data Point
3B,4	152.90	223.65	909	8825	1529
3B,6	169.40	256.19	799	7409	1694
3B,7	231.20	370.14	392	3940	2312
3G,2	139.70	196.38	769	6880	1397
7B,4	152.90	223.76	1252	12744	1529
7B,6	169.40	256.21	791	7734	1694
7B,7	231.40	370.02	305	3070	2314
7G,3	139.60	196.35	1069	9349	1396
11B,4	153.20	223.88	3433	49996	1532
11B,6	169.70	256.17	2057	23449	1697
11B,8	231.90	370.20	1334	15095	2319
11G,4	139.90	196.56	1296	12600	1399

6.3.1 SELECTION OF SAMPLES

Two sequenced samples and a number of samples selected randomly were amplified under the optimised conditions in two multiplex reactions (6.2.2). An aliquot of these amplifications was run on an ABI 373 automated DNA sequencer and the samples exhibiting different alleles at various loci with reference to the two sequenced samples were selected and reamplified for each locus in singleplex reactions using the optimised conditions (6.2.1). This generated stocks of amplification of various alleles for each Y STR locus which were then used for sequencing reactions and subsequent preparation of ladders.

6.3.2 SEQUENCING OF LADDER SAMPLES FOR THE LOCI

The amplified product was cleaned up using the Qiaquick™ spin purification kit (2.7.3). 30 µl elution buffer was used to elute the DNA. After quantitation the purified samples were sequenced at the Molecular Biology Sequencing Unit of the University of Glasgow using the forward primer of each locus (Table 6.2 & Figure 6.4).

6.3.3 PREPARATION OF ALLELIC LADDERS

The two composite ladders were run on an ABI 373 automated DNA sequencer and assessed. The volume of the individual ladders was adjusted to obtain balanced peaks. An aliquot of the prepared ladder for each multiplex was run on three gels in an ABI 373 automated DNA sequencer, using same conditions.

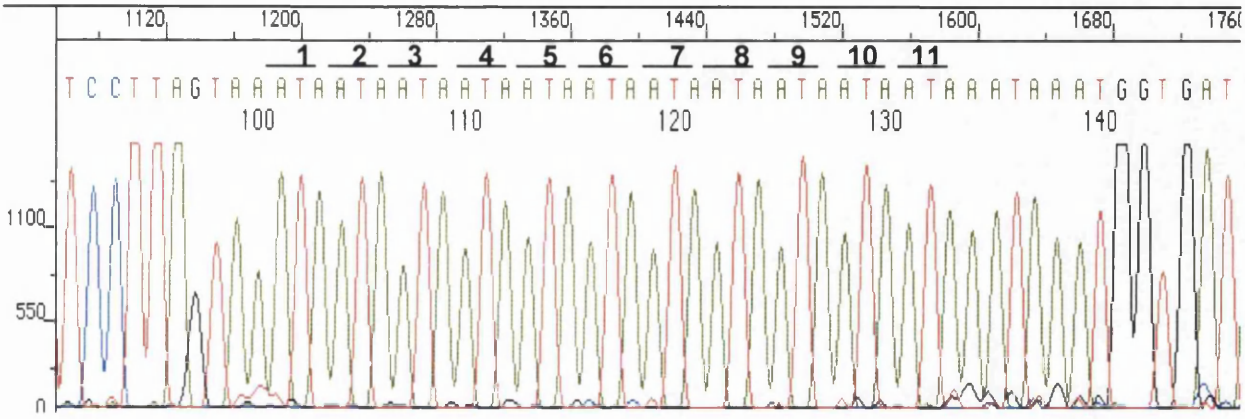
All alleles showed a standard deviation of less than 0.16 (Table 6.3). Same ladders were then used to run with all the gels used for electrophoresing the amplified samples for Multiplex I &II (Figure 6.5 & 6.6).

Figure 6.4: Electropherograms Showing the Sequence of Various Alleles of the Locus DYS392

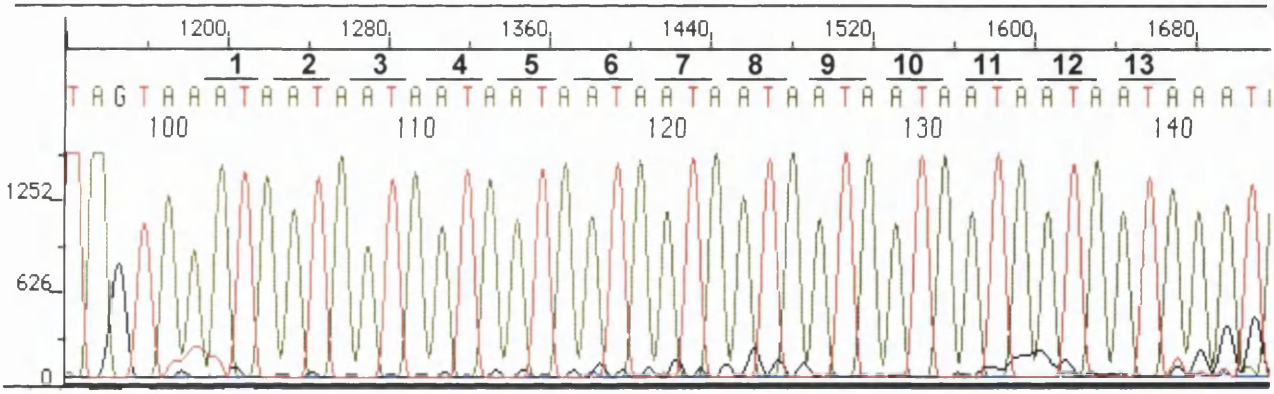
DYS392 is a simple trinucleotide STR with a repeat motif of **TAT** which is regularly spaced.

The sequence of three alleles 11, 13 & 14 are shown. The repeat motif is repeated for 11, 13 and 14 times for each allele respectively.

Allele 11



Allele 13



Allele 14

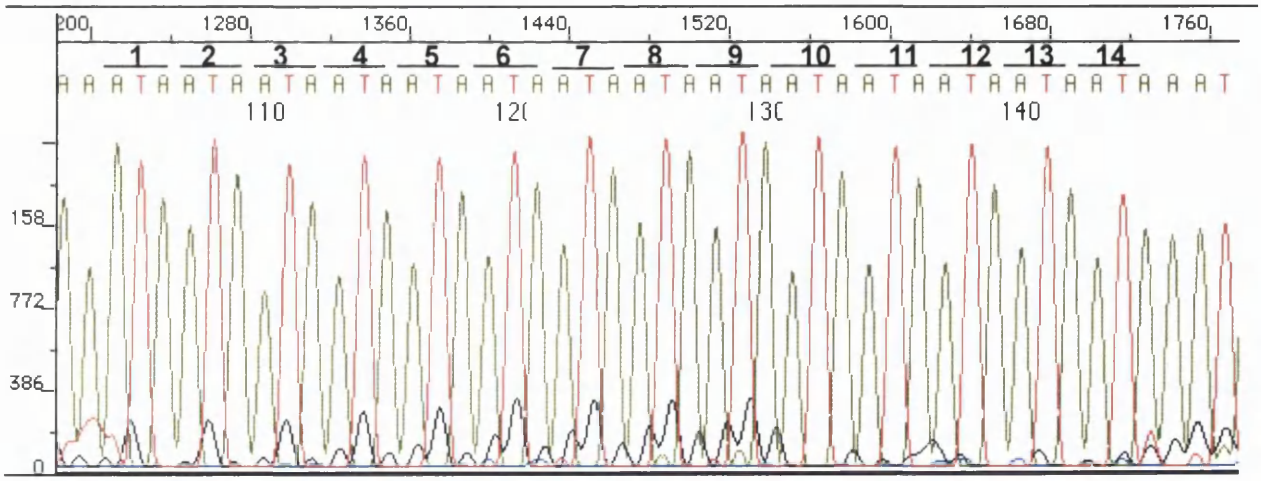


Table 6.2: Sequenced Structure Of Alleles of Y Chromosome STR Loci

Locus	Allele	Structure
DYS391	10	F (23bp)- ta$(TCTG)_3$**$(TCTA)_{10}$**$(TCTG)(CCTA)$tctgcctg$(CCTA)_2$ 17 bp- R (28bp)
DYS392	13	F (25bp)- 48bp (ATT)tttttcttgtatcacc(ATT)(TAT)t(**TAT)$_{13}$** ttactaaggaatggg(ATT) 77bp- R (20bp)
DYS393	14	F (24bp)- **(AGAT)$_{15}$**$(ATGT)_2$ 18bp- R (22bp)
DYS19	14	F (21bp)- 8bp$(AT)_5$agtgtt$(AT)_5$agtgtt$(TAGA)_3$tagg**$(TAGA)_{14}$**tata 31bp- R (19bp)
DYS390	23	F (23bp)- 26bp$(TCTA)_2(TCTG)_8$**$(TCTA)_{10}$**$(TCTG)(TCTA)_4TCA(TCTA)_2$ 29bp- R (23bp)
DYS389I	10	F (25bp)TATC$(TGTC)_3$**$(TATC)_{10}$**cctccctc(TATC)tatt(TATC)tagc 71 bp TGTC 43 bp- R (20bp)
DYS389II	27	F (25bp)- 4bp $(TGTC)_5$**$(TATC)_{11}$** 44bp (TATC)$(TGTC)_3$**$(TATC)_8$** 150bp- R (20bp)

F is forward primer, R is reverse primer
The sequences in bold are the hypervariable segements of the alleles
DYS 389I & II are two products of amplification with the same primer set
Modified from Pestoni, C. *et al*., 1998

Table 6.3: Allelic Ladder Fragment Sizes

Locus	DYS391		DYS392		DYS393		DYS19	
Allele	Mean Size*	SD**	Mean Size	SD	Mean Size	SD	Mean Size	SD
9	143.17	0.09						
10	147.56	0.04						
11	151.75	0.06	251.66	0.04	115.67	0.06		
12	155.67	0.13	254.37	0.11	119.55	0.11		
13			257.48	0.12	123.29	0.15	188.53	0.15
14			260.26	0.15	127.56	0.18	192.21	0.12
15					131.62	0.11	196.39	0.15
16							200.45	0.17
17							204.73	0.13

Locus	DYS389I		DYS389II		DYS390	
Allele	Mean Size	SD	Mean Size	SD	Mean Size	SD
9	252.41	0.13				
10	255.5	0.16				
11	258.4	0.17				
22					210.34	0.16
23					214.54	0.11
24					218.46	0.13
25			366.04	0.17	222.51	0.11
26			370.26	0.12	227.62	0.15
27			374.38	0.07	231.04	0.14
28			377.85	0.11		
29			381.89	0.21		

* The mean size & SD were obtained by running the same sample on three gels

** Standard deviation

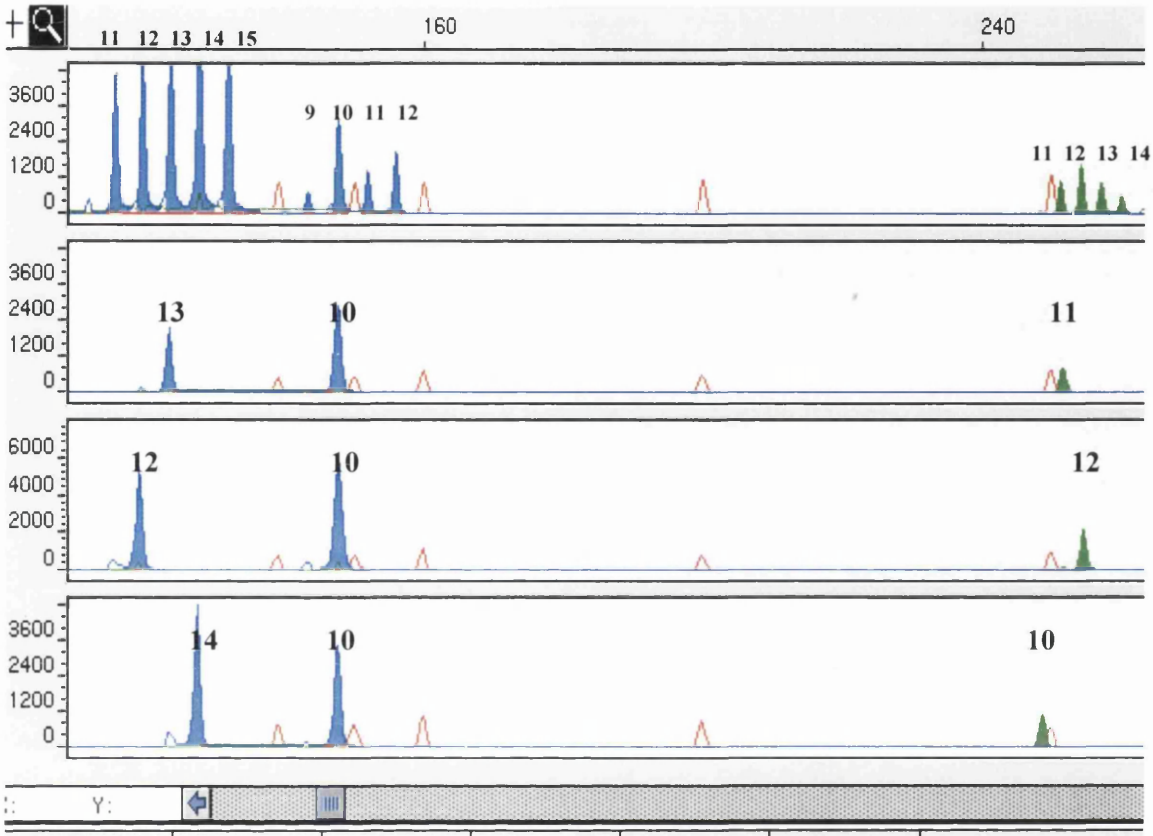
Figure 6.5a: Allelic Ladder for Multiplex I (DYS391, 392 & 393)

The composite ladder for Multiplex I was run on the sequencer and the ladder assessed. Locus DYS392 (JOE labelled) which is a trinucleotide appeared to be diluted and the peaks were less intense than the other two loci so the ladder for this locus was added at twice the volume of the other two ladders.

The first panel shows the ladder fragments comprising of (from left to right) Allele 11, 12, 13, 14 &15 for the locus DYS393 (Blue peaks); Allele 9, 10, 11 & 12 for the locus DYS391 (Blue peaks) and Allele 11, 12, 13, & 14 for the locus DYS392 (Green peaks). The red peaks show the internal standard GeneScan™ 500 [Rox].

Panels 2- 4 show the designation of alleles in comparison to the allelic ladder.

DYS393 **DYS391** **DYS392**



Dye/Sample Peak	Minutes	Size	Peak Height	Peak Area	Data Point
17 B, 12	140.40	115.61	4954	31691	1404
17 B, 14	143.60	119.52	5542	36459	1436
17 B, 16	146.60	123.60	6021	42127	1469
17 B, 17	150.20	127.73	5969	49638	1502
17 B, 19	153.50	131.91	5860	41525	1535
17 B, 21	162.20	143.29	805	4587	1622
17 B, 23	165.60	147.78	3238	21925	1556
17 B, 24	168.90	151.95	1535	9769	1689
17 B, 26	172.30	156.02	2167	13364	1723
17 G, 50	253.50	251.46	1160	9063	2535
17 G, 51	256.00	254.28	1735	14178	2560
17 G, 52	258.60	257.21	1094	9543	2586
17 G, 53	261.20	260.14	649	5929	2612
8 B, 7	109.20	123.45	2258	11455	1092
8 B, 9	121.20	147.58	2959	15796	1212
8 G, 43	170.10	251.72	887	5592	1701
10 B, 14	107.10	119.24	5564	32384	1071
10 B, 18	121.20	147.77	5927	37418	1212
10 G, 20	171.70	254.78	2181	13967	1717
13 B, 13	111.40	127.56	4926	25303	1114
13 B, 15	121.30	147.56	3541	19069	1213
13 G, 45	168.90	248.15	1209	7455	1689

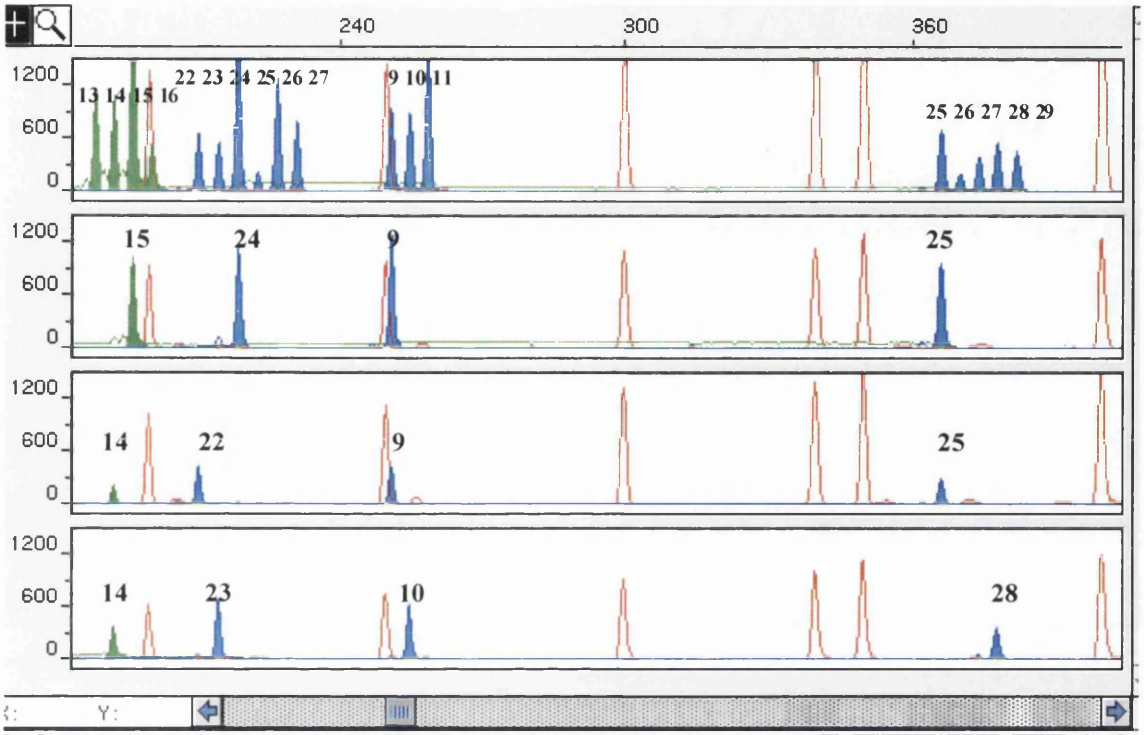
Figure 6.5b: Allelic Ladder for Multiplex II (DYS19, 389I, 389II & 390)

The ladder was run on a 6% gel at 40 W for 12 h (2.8) with GeneScan™ 500 [Rox] internal lane standard three times so that the standard deviation of each allele could be established.

The top panel shows the allelic ladder.

The lower three panels show three multiplex amplifications for Multiplex II. The alleles were designated in accordance with the ±0.5 bp rule with reference to the allelic fragments in the ladder.

DYS19 DYS390 DYS389I DYS389II



Dye/Sample Peak	Minutes	Size	Peak Height	Peak Area	Data Point
48B, 6	193.20	210.38	687	4431	1932
48B, 7	196.00	214.66	565	3682	1960
48B, 8	198.70	218.77	2041	13417	1987
48B, 9	201.40	222.89	241	1498	2014
48B, 10	204.10	227.00	1326	8982	2041
48B, 11	206.80	231.12	825	5570	2068
48B, 12	220.00	251.12	969	7432	2200
48B, 13	222.70	254.88	915	7303	2227
48B, 14	225.50	258.79	1642	12792	2255
48B, 16	298.00	366.21	736	5677	2980
48B, 17	300.50	370.07	218	1631	3005
48B, 18	303.10	374.10	411	3375	3031
48B, 19	305.60	377.97	575	4598	3056
48B, 20	308.20	381.99	466	3745	3082
48G, 24	178.50	188.78	1158	9197	1785
48G, 26	181.20	192.61	1106	7693	1812
48G, 28	184.00	196.59	2382	17626	1840
48G, 29	186.60	200.61	559	4930	1868
32B, 15	200.00	218.85	1145	8029	2000
32B, 17	221.50	251.10	1330	10255	2215
32B, 19	300.80	366.21	991	8217	3008
32G, 104	185.10	196.67	1094	9014	1851
30B, 17	194.70	210.57	459	3021	1947
30B, 18	221.80	251.23	445	3338	2218
30B, 19	301.20	366.32	303	2335	3012
30G, 28	182.50	192.74	246	1628	1825
11B, 11	197.20	214.69	718	5080	1972
11B, 12	224.20	254.95	654	5284	2242
11B, 14	308.10	377.83	375	3393	3081
11G, 89	182.30	192.68	407	4097	1823

6.4 PROFILING OF SAMPLES

Male samples (564) from seven populations were amplified using the optimised PCR conditions (Table 6.1a & b) with a high success rate.

Product size of the amplified alleles were determined on an ABI 373 XL UPGRADE automated DNA sequencer (2.8). For Multiplex I; 4.5% polyacrilamide gel was used and samples electrophoresed for 3-4 h while for multiplex II; 6% gel was used and samples were run for 8-12 h.

6.5 STATISTICAL ANALYSES

The data was analysed on Microsoft® Excel and the DNA analysis software ARLEQUIN version 2.000 (Excoffier, L. *et al.*, 2000).

6.5.1 ALLELE FREQUENCY AND GENETIC DIVERSITY

The allele frequencies were calculated for each locus. Unbiased genetic diversity and standard deviation were calculated using the software package ARLEQUIN. The genetic diversity (h) of each locus (Nei, M. *et al.*, 1987) was calculated using the equation:

$$h = \frac{\left(1 - \sum x^2\right)n}{n-1}$$

where x is the allele frequency & n is the number of individuals profiled

As Y Chromosome STRs are haploid markers the genetic diversity is also a measure of power of discrimination and exclusion (Kayer, M. *et al.*, 1997).

The standard deviation (Nei, M. *et al.*, 1987) was calculated by taking square root of the variance employing the equation:

$$V(h) = \frac{2\left[\sum x^3 - \left(\sum x^2\right)^2\right]}{n}$$

where x is the allele frequency & n is the number of individuals profiled

6.5.2 HAPLOTYPE FREQUENCY AND DIVERSITY

Haplotypes were organised as Haplotype I & II. Haplotype I comprised of loci DYS391, 392 & 393 and haplotype II comprised of loci DYS19, 390, 389I & 389II. The allelic data for each sample was also arranged as a haplotype for all the seven loci (Yh1) and unbiased haplotype diversity was calculated for each population.

6.5.3 DISCRIMINATION CAPACITY

The number of haplotypes observed only once in a population, which were not shared with other populations, were designated as 'unique haplotypes'. The discrimination capacity of the haplotype Yh1 consisting of the seven loci studied was calculated for each population as a percentage of the unique haplotype (Kayser, M. *et al.,* 1997; Rossi, E. *et al.,* 1998).

6.5.4 PROBABILITY OF IDENTITY

Haplotypes observed in each population were compared with those of all the other populations, and shared haplotypes were documented as a list (Appendix 5). The probability of identity between different Pakistani populations was calculated using the equation:

$p = \sum_{i,j}^{n} x_i x_j$ where, x_i and x_j are the frequencies of the shared haplotype in two populations (Melton, T. *et al.,* 1995).

6.6 RESULTS

6.6.1 ALLELE FREQUENCIES

564 male samples from different populations were profiled and complete haplotypes (Yh1) determined for all the samples (Appendix 4). The allele distribution at most loci was unimodal (Figure 6.6).

Figure 6.6: Comparison of Allele Frequencies of Y STR Loci

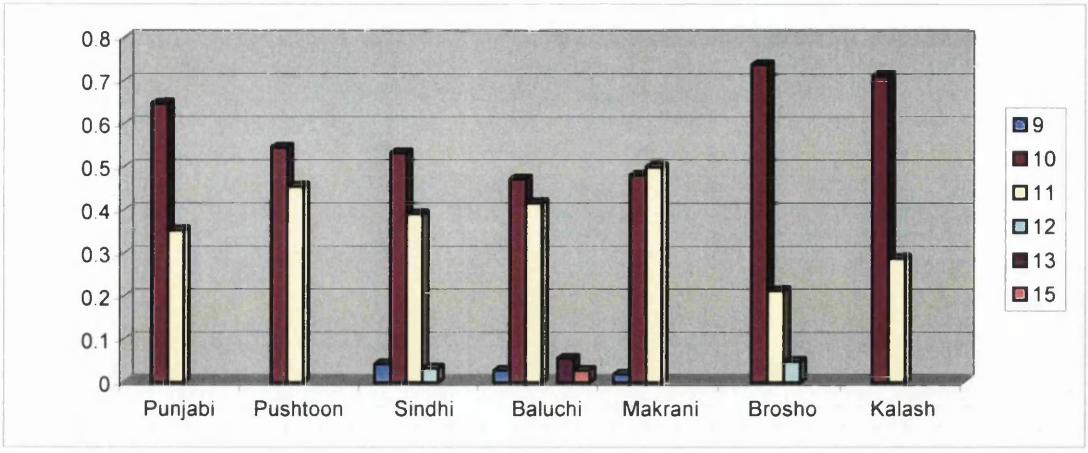
Allele frequencies were similar among the mainland Pakistani populations at the loci DYS391 and 392. The loci DYS19, 389I, 389II, 390 and 393 showed significantly different allele frequencies among all populations. However at there were similarities as well like allele 13 at the locus DYS393 and allele 15 at the locus DYS19 were at high frequency in all populations.

Sindhi, Baluchi and Makrani populations were more polymorphic for locus DYS391 than other populations. Baluchi and Makrani populations had fewer differences in the allele frequencies at all the loci as compared to other populations.

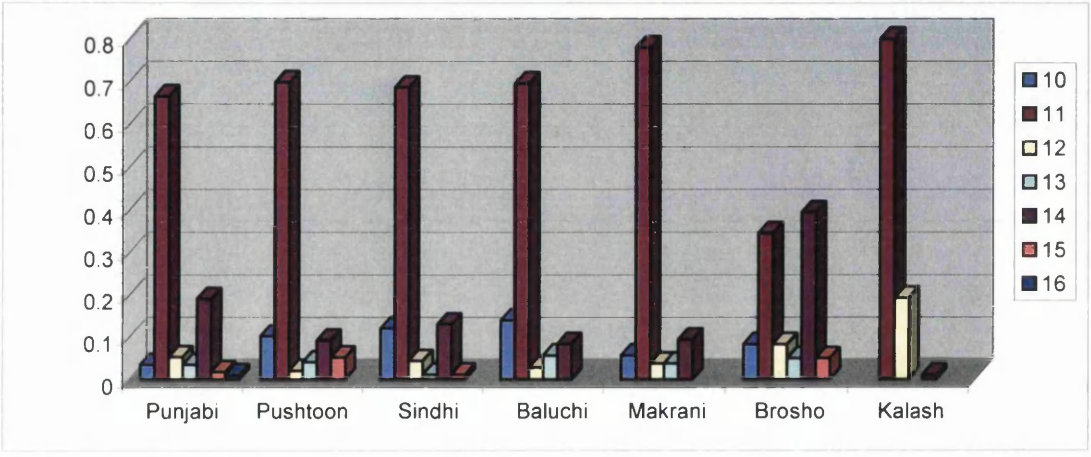
Baluchi population despite its small size had a greater number of alleles than other populations.

Kalash and Brosho showed a characteristic pattern of allele distribution at most loci.

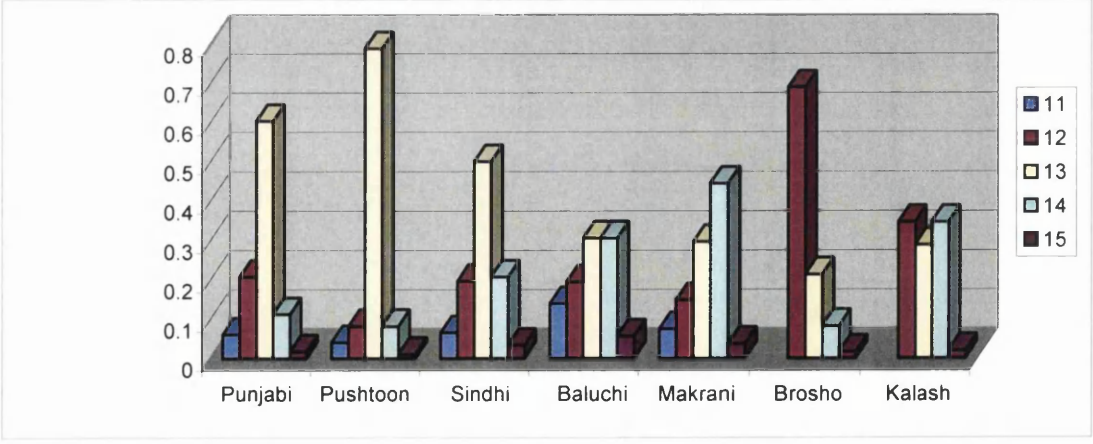
DYS391



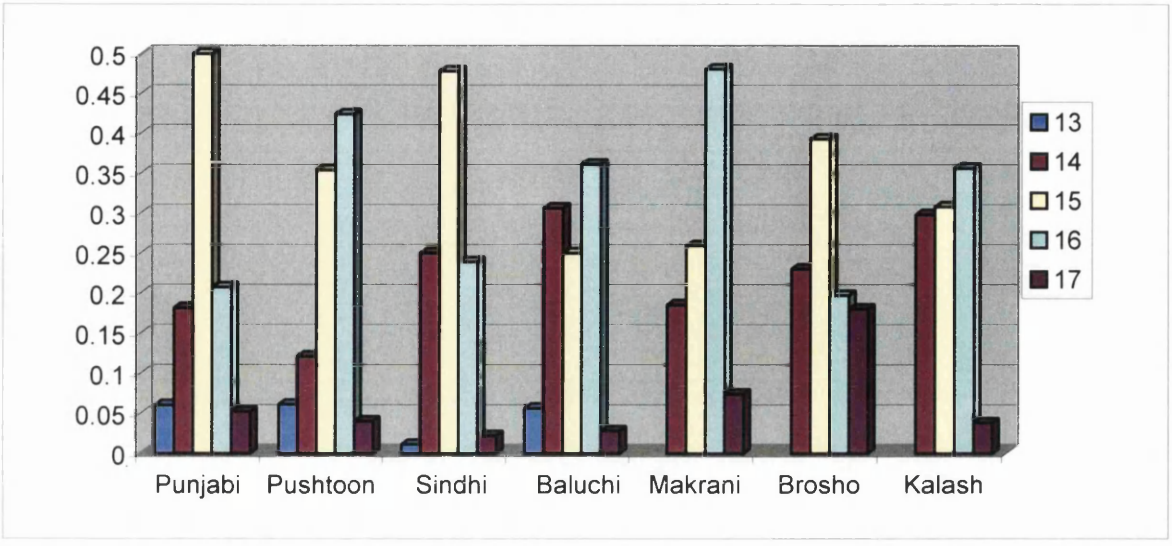
DYS392



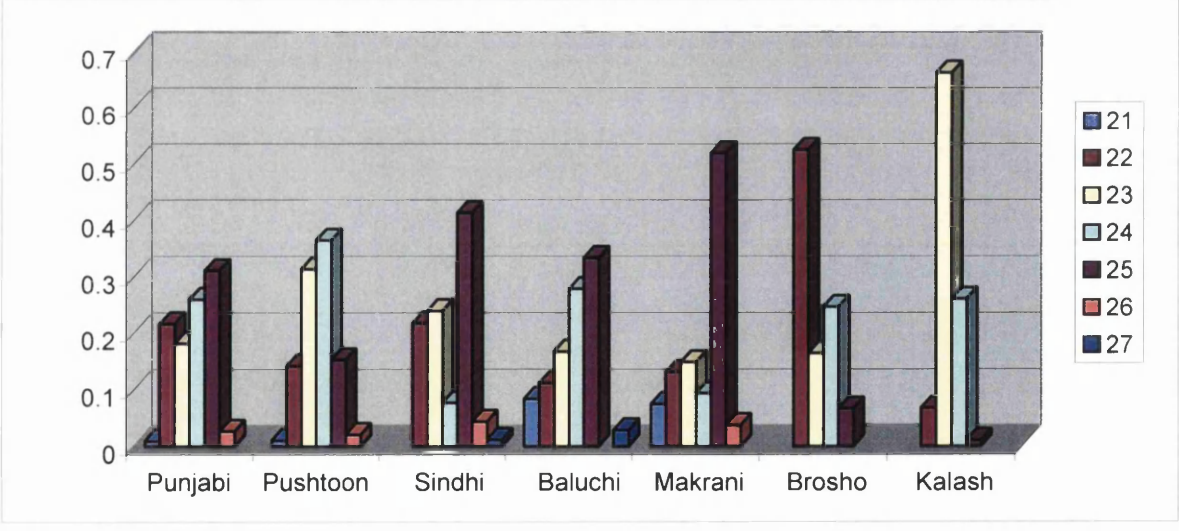
DYS393



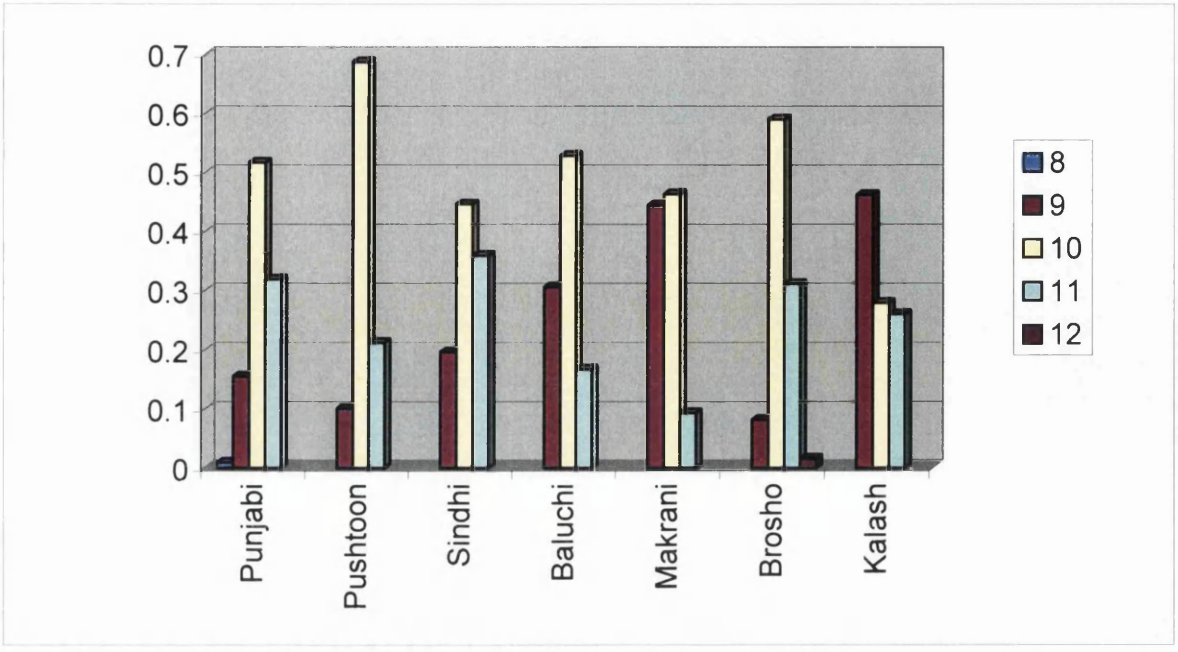
DYS19



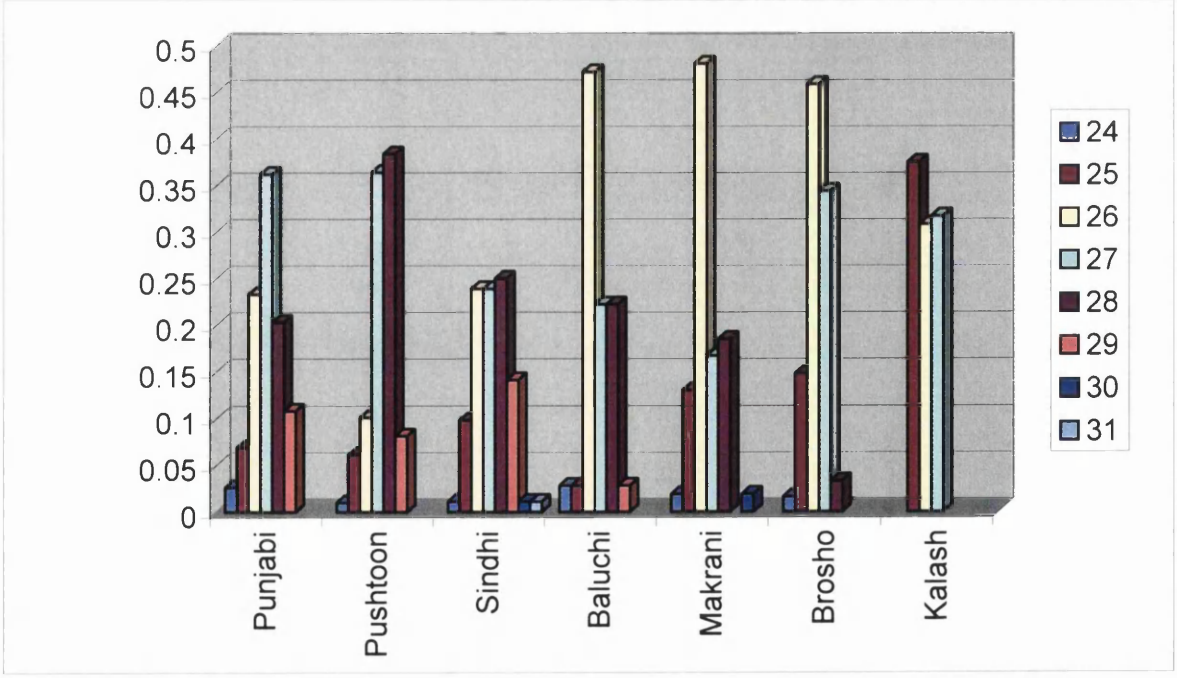
DYS390



DYS389I



DYS389II



6.6.2 COMPARATIVE DATA FOR THE LOCI AMPLIFIED AS MULTIPLEX

Haplotype Multiplex I (DYS391, 392 & 393) had a low diversity as compared to the Haplotype Multiplex II (DYS19, 390, 389I & 389II) which was due to the high polymorphism of DYS390 and 389II (Table 6.4 & Figure 6.7).

6.6.3 Yh1 HAPLOTYPE FREQUENCIES AND DIVERSITY

When the data were arranged as a 7 loci haplotype (Yh1) for all the populations, the highest haplotype diversity exhibited by the Punjabi population and the lowest by the Kalash population. All other populations were diverse when compared to each other (Table 6.5).

The Kalash show the highest frequency (31%) of a most common haplotype (MCH). In Makrani population MCH occurred at high frequency (17%), in all other populations MCH frequency was low (Table 6.5b)

6.6.4 PROBABILITY OF IDENTITY

Probability of identity between different populations was low in all population pairs except Baluchi/Makrani pair for which it was the highest (Table 6.6). This showed that there was a lower probabilityof finding a match between populations than within a population. However, in case of Baluchi and Makrani populations this probability was relatively high.

6.6.5 DATABASE CONTENT FOR COMBINED PAKISTANI POPULATION

Y STR haplotype data of all the populations was combined to generate the allele and haplotype frequencies for Pakistani populations which exhibited a high haplotype diversity of 0.9952. 62% of haplotypes were detected in the database only once. Nine haplotypes of the mainland Pakistani population were shared with Brosho and Kalash populations which decreased the discrimination capacity of haplotype Yh1 in Pakistani population.

Figure 6.7: Graphical Representation of Haplotype Diversity

The Baluchi population was as diverse for both haplotypes. The Makrani and Brosho were less diverse for haplotype II and showed less number of haplotypes than Punjabi, Pushtoon and Sindhi populations.

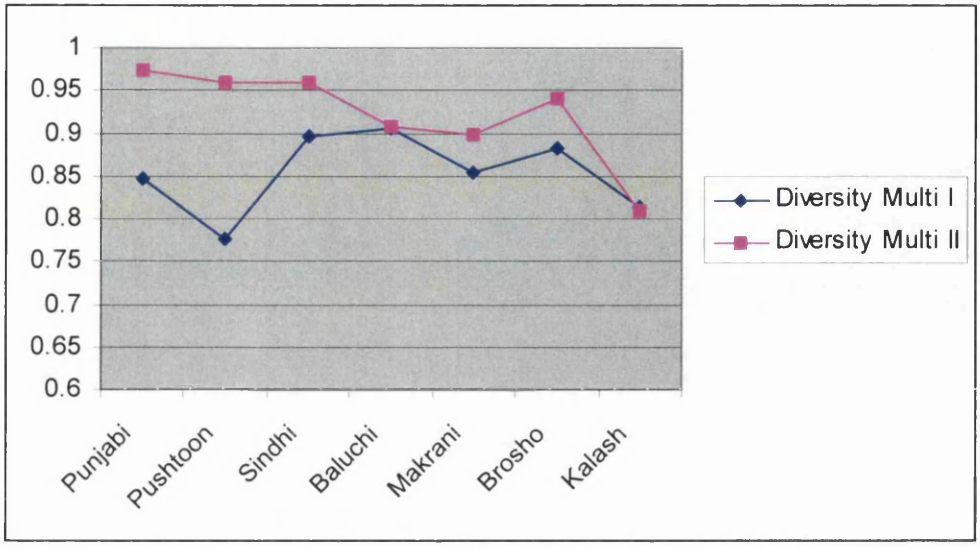


Table 6.4: Comparison of Diversity of Multiplex I &II

Population	Genetic Diversity Haplotype Multiplex I	Genetic Diversity Haplotype Multiplex II
Punjabi	0.848	0.974
Pushtoon	0.777	0.961
Sindhi	0.896	0.959
Baluchi	0.905	0.908
Makrani	0.854	0.898
Brosho	0.882	0.942
Kalash	0.813	0.809

Table 6.5a: Comparison of Individual Locus & Haplotype Diversities for Different Pakistani Populations

Locus DYS	391	392	393	19	390	389I	389II	Mean	Yh1 *
Punjabi	0.457	0.518	0.577	0.668	0.757	0.607	0.756	0.62	0.9924
Pushtoon	0.495	0.491	0.364	0.675	0.726	0.473	0.700	0.56	0.9781
Sindhi	0.56	0.497	0.663	0.651	0.717	0.634	0.793	0.65	0.9872
Baluchi	0.599	0.488	0.753	0.710	0.764	0.600	0.676	0.66	0.9828
Makrani	0.750	0.380	0.687	0.662	0.676	0.580	0.689	0.63	0.9692
Brosho	0.407	0.71	0.473	0.721	0.568	0.548	0.648	0.58	0.9825
Kalash	0.41	0.326	0.677	0.688	0.488	0.641	0.664	0.56	0.8833

* Yh1= DYS391, 392, 393, 19 390, 389I & 389II Haplotype

Table 6.5b: Haplotype Results & Analysis in Different Populations

Population	N	Yh1	Unique	MCH	DC
Punjabi	116	99	77	0.03	0.66
Pushtoon	99	76	50	0.06	0.5
Sindhi	92	72	56	0.03	0.61
Baluchi	36	34	28	0.02	0.82
Makrani	54	45	37	0.15	0.68
Brosho	61	49	39	0.05	0.64
Kalash	104	37	21	0.31	0.2

N: The total number of males profiled

Unique: Oberved only in one individual among populations

MCH: Frequency of Most Common Haplotype observed

DC: Discrimination Capacity = No of Unique Haplotypes / Total Haplotypes

Table 6.6: Probability of Identity & Number of Shared Haplotypes

	Punjabi	**Pushtoon**	**Sindhi**	**Baluchi**	**Makrani**	**Brosho**	**Kalash**
Punjabi		13*	5	1	0	1	1
Pushtoon	0.003**		4	2	4	0	2
Sindhi	0.001	0.001		3	7	3	4
Baluchi	0.0006	0.001	0.002	0	5	0	1
Makrani	0	0.001	0.005	0.015	0	0	1
Brosho	0.0002	0	0.0009	0	0		0.001
Kalash	0.0009	0.0004	0.0004	0.0003	0	0.0008	

* Number of shared haplotypes between the population pairs above the diagonal.

** Probability of identity between the population pairs below the diagonal.

6.6.6 COMPARISON OF PAKISTANI WITH OTHER POPULATIONS

Combined Pakistani data were also analysed with a view to compare the Y chromosome haplotype data with other populations so as to assess the value of the Y chromosome analysis in the Pakistani population as a whole. Comparisons were made to the data for Japanese, Chinese (Horst, B. *et al.*, 1999), Germans, Turks (Brinkmann, C. *et al.,* 1999), Italians (Tagliabracci, A. *et al.,* 2000) and also for combined European population for the extended haplotype [Y User Group Web Site: http://ystr.charite.de/index_mkl.html].

The Pakistani population exhibited a higher diversity than all these population for Yh1 and slightly lower diversity than the extended haplotype (Yh1+DYS385+YCAII) (Figure 6.9).

Comparison of the discrimination capacity was also interesting, which for Yh1 is only 28% for most of the European populations (Y User Group Web Site: http://ystr.charite.de/index_mkl.html). It has much greater discrimination capacity in Pakistani population (62%) which is more than that of the minimal haplotype (Yh1+DYS385) for the European populations (50%). This phenomenon might partially be due to the fact that the number of samples in the European database is much higher (~5000) than the present Pakistani sample (564).

6.7 DISCUSSION

Pakistan is a diverse country on linguistic and anthropological basis. Two studies conducted for determining the genetic variation have shown similar results (Seielstad, M. *et al.,* 1999; Ayub, Q. *et al.,* 2000). The populations residing in the same general area might show ethnic differences, which is why micro geographic sampling has been recommended for forensic purpose (Brinkmann, C. *et al.,* 1999).

6.7.1 PCR OPTIMISATION PROCEDURES AND PROFILING

The Y STR loci were amplified as two multiplex reactions. In Multiplex I new primers were used to amplify the locus DYS391. Using Reddymix® Mastermix eased the optimisation process.

In most of the earlier reports *Taq* Gold™ was recommended for efficient multiplex amplification (Kayser, M. *et al.,* 1997). During this work Reddymix™ was found to be easier to use and gave reproducible results. The dye in the multimix does not affect the amplification or the analysis on the automated sequencer and can be stored in the refrigerator for more than 8 weeks avoiding freeze, thaw cycles.

Multiplexing the Y STR loci has been attempted in different combination of loci and many multiplex systems have been reported for the Y STR loci (Prinz, M. *et al.,* 1997; Kayser, M. *et al.,* 1997; Redd, A. J. *et. al.*, 1997; Thomas, M. *et al.,* 2000; Gonzalez-Neira, A. *et al.*, 2000; Anslinger, K. *et al.,* 2000).

During this study besides the two multiplex systems used, a new multiplex was also tried comprising of loci DYS19, 391, 392 & 393, which worked very well and the same was reported recently with amelognein locus added to it (Anslinger, K. *et al.,* 2000). However it would be of much value if a standard multiplex kit allowing simultaneous amplification of a greater number of Y STR loci would be developed for the Y STRs as has been achieved in the case of autosomal STRs. In this connection efforts are underway and recently a commercial firm developed a multiplex (Y-Plex™ 6) for 6 Y STR loci (Reliagene Technologies, Inc. USA) (Warren, J. 2000).

6.7.2 ALLELE FREQUENCY OF Y STRs IN PAKISTANI POPULATIONS

Common ancestry results in similar, common most frequent allele in different populations (Deka, R. *et al.,* 1996). The observation of the same most common allele in all the Pakistani population means that the ancestral alleles in different systems are largely the same among them.

In the Kalash two population specific differences were observed, a high frequency of allele 23 at locus DYS390 (0.66) detected at lower frequency in the Pakistani populations where the highest frequency was found in the Sindhi population (0.24). The second was the absence of allele 13 at locus DYS392 the frequency of which ranged from 0.056-0.012 in other Pakistani populations.

6.7.3 GENETIC DIVERSITY

All the loci had genetic diversities comparable to other populations, like German and UK Caucasians (Lessig, R. & Edelmann, J. 1998; Philip, C. P. *et al;* 2000). When the loci DYS391, 392 and 393 were considered as a haplotype, the genetic diversity of this haplotype was lower in all Pakistani populations in comparison to the quadraplex DYS19, 390, 389I & 389II. In the Punjabi population for which the sample size was bigger than other populations, the genetic diversity for DYS391, 392 and 393 was 0.848 whereas the haplotype diversity for the quadraplex was 0.974. Loci DYS19, 390 and 389II were particularly polymorphic and the diversity for these was higher in Pakistani populations as compared to the Caucasian population (Kloosterman, A. D. *et al;* 1997).

The haplotype diversity of Yh1 was ~0.98 in individual Pakistani populations. The highest diversity for the haplotype Yh1 was observed in the Punjabi population (0.99) which inhabits the area where most migrating populations have come to settle down over many centuries thus creating a diverse gene pool. This was also reflected in the STR data for the Punjabi population (Figure 5.1 a, b, c & Table 5.1 a, b, c). The haplotype diversity for the combined mainland population was 0.9952.

6.7.4 DISCRIMINATION CAPACITY

For the Y chromosome another measure of diversity is the discrimination capacity which is the frequency of the unique haplotypes in a populations (Kayser. M. *et al.,* 1997 & Rossi, E. *et al.,* 1997). This was highest in the Baluchi population and for 36 individuals 34 haplotypes were observed. Since haplotypes present in the

Baluchi population were also detected in other populations, the discrimination capacity of Yh1 haplotype was 80%, which was much higher than other Pakistani populations. When the mainland populations (populations excluding the Brosho and Kalash) were treated as a group the discrimination capacity of Yh1 was 62% which was quite high when compared to combined European population in which Yh1 has a discrimination capacity of only 27% (Y-STR Haplotype Reference Database. Available at website: http://ystr.charite.de/index_kl.html). This phenomenon might partly be due to the greater number of samples in the European database.

6.7.5 SHARED HAPLOTYPES AMONG PAKISTANI POPULATIONS

Common haplotypes among closely located populations may imply common paternal ancestry (Furedi, S. *et al.,* 2000) and in forensic cases when the ethnic origin of the perpetrator is in question, this might have an important bearing on the probability of identity between different populations. Therefore the database of each population was compared for shared haplotypes. The Kalash shared 9 haplotypes with other populations (Table 6.7). Out of these shared haplotypes two were detected at a high frequency in the Kalash, which might point to common paternal lineage. 13 haplotypes were shared among Punjabi and Pushtoon populations and less among Punjabi/Sindhi and Punjabi/Baluch populations.

6.7.6 COMPARISON OF PAKISTANI AND OTHER POPULATIONS

In view of low diversity of the Y chromosome STRs in the Caucasian and other populations more Y STR loci were sought in order to enlarge the panel of Y STRs available to the forensic community. It was shown that inclusion of a tetranucleotide DYS385 increased the discrimination power of the Y STR panel therefore it was included for forensic casework and the same was recommended for court use (Caglia, A. *et al.*, 1998 a; Pascali, V. L. *et al.,* 1998). In order to close the gap in the powers of discrimination and exclusion of the Y chormosome specific markers, it was later recommended that an extended haplotype be used for forensic

casework that includes Yh1+DYS385+YCAII (Caglia, A. *et al.,* 1998 b; Kayser, M. *et al.,* 1998).

Currently databases are being developed for the extended haplotype (Roewer, L. *et al.,* 2000). The haplotype diversity of various Caucasian populations for the extended haplotype is ~0.99 (Furedi, S. *et al.,* 1999; Ansingler, K. *et al*; 2000; Caglia, A. *et al*; 1998).

When the Pakistani populations of the mainland were treated as a group, the haplotype diversity of Yh1 was found to be higher as compared to Chinese, Italians, Germans, Japanese, Turks and the Eurpopean combined population database (Figure 6.8).

It was interesting that Yh1 achieves only slightly lower level of diversity as the extended haplotype (Yh1+DYS385+YCAII) in the European populations (Figure 6.8). It follows that a database of the extended haplotype would achieve a much higher discrimination power in Pakistani populations than the European populations. However for casework purposes even the Yh1 would perform better in Pakistani than the European Caucasian population.

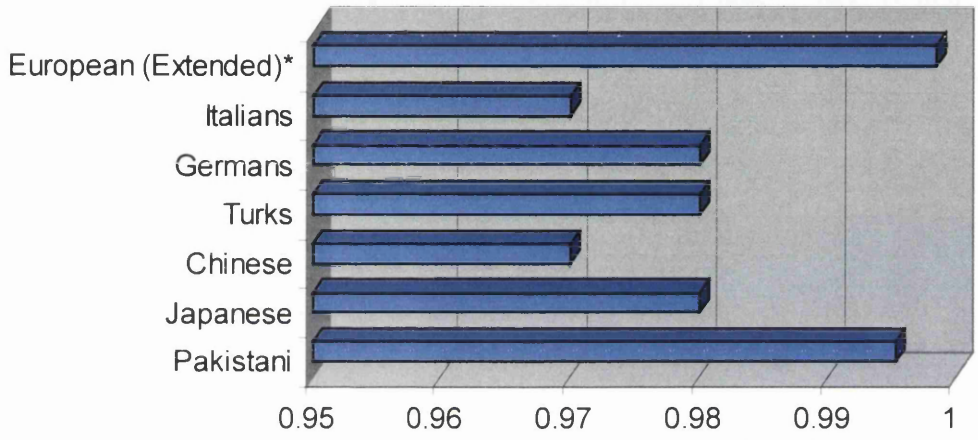
The comparison of the frequency of most common haplotype (MCH) also yielded similar information and the frequencies of MCH in the European database for the minimal (Yh1+DYS385) and extended haplotypes are 3.05% & 1.59% respectively. In the Pakistani database in the individual populations the frequency of MCH for Yh1 was not more than 6% in the mainland populations while it was only 3% in Sindhi and Punjabis which are the two major populations.

In conclusion a large database has been established for the Pakistani populations, for seven loci Y STR haplotype, which has a high diversity in Pakistani populations and would serve as an efficient forensic identification tool.

Figure 6.8: Haplotype Diversity Comparison between Different Populations of the World

Pakistani population exhibits higher diversity than all other populations for Yh1 and a slightly lower diversity than the extended haplotype for the combined European population.

Japanese & Chinese data: Horst, B. *et al.*, 1999
German & Turk data: C. Brinkman. *et al.,* 1999
Italian data: Tagliabracci, A. *et al.*, 2000
European combined data: Y-STR Haplotype Reference Database. Available at website: http://ystr.charite.de/index_kl.html
* Extended Haplotype: Yh1+DYS385+YCAII



	Pakistani	Japanese	Chinese	Turks	Germans	Italians	European (Extended)
Series1	0.9952	0.98	0.97	0.98	0.98	0.97	0.9984

CHAPTER 7: THE ANALYSIS OF MOLECULAR VARIANCE & PHYLOGENETIC RELATIONSHIP OF PAKISTANI POPULATIONS

7.1 INTRODUCTION

mtDNA and other studies point to a Caucasoid gene flow to the populations of the subcontinent, which is more marked in the populations inhabiting the north west of present India (Passarino, G. *et al.*, 1996 & Underhill, P. A. *et al.*, 1997). This has been suggested due to admixture between the Indo Aryans entering the north of the subcontinent and the indigenous Dravidians 3,000-4,000 years ago (Cavavlli-Sforza. *et al.*, 1993).

The effects of this migration of the Indo Aryans had a major impact in the north and centre of the subcontinent, which includes the present day Pakistan. The genetic distinctiveness of the Northwest and Southeast Asia has been described (Nei, M & Roychoudry, A. K. 1993). There has been no study to date to compare the different populations of Pakistan in order to study their phylogenetic relationship with each other, or with the Indian populations.

In this chapter the molecular relationship between the populations in Pakistan are examined using phylogenetic techniques. The Kalash population is analysed further to examine the possibility of exploring the pre history using the STR markers.

7.1.1 Y CHROMOSOME STRs & PHYLOGENETICS

The Y STRs are relatively fast evolving markers and the effective size of Y chromosome is less than the autosomes thus they are subject to more drift and founder

effects than the autosomes. The early Y chromosome studies showed that the Y STRs can be valuable for studying differences between closely related populations (Jobling, M. A. & Tyler-Smith, C. 1995; Roewer, L. *et al.*, 1996 & deKnijff, P. *et al.*, 1997). It was therefore thought that Y STRs would be valuable in studying the population structure, history and degree of relatedness of the Pakistani populations.

7.1.2 PHYLOGENETIC TREE CONSTRUCTION USING MOLECULAR DATA

Molecular phylogenetics enables conversion of molecular data into branching diagrams or trees, which show the relationship of the populations. Various tree building approaches have been used, like discreet character method or distance matrix method (Page, R. D. M & Holmes, E. C. 1998). The use of distance matrix method involves, conversion of data to a distance matrix, which can then be used to contruct the trees. Puzzle, Parsimony, the "NJ"(Neighbourhood Joining Method) and the "UPGMA" method (Unweighted Pair Group of Arithmetic means) etc. are used to build the trees

7.2 HISTORICAL PERSPECTIVES

The traces of early farming in the neolithic period in Pakistan (Baluchistan) show that human habitation was wide spread and the "Indus Valley" Civilisation emerged on the banks of the river Indus as an urban culture later on. This region which now constitutes Pakistan is important as a melting pot of various human groups, as a result of prehistoric and the historic time migrations to this region from Iran and the Caucasus (Chattopadhyaya, S. 1955). The present day Pakistan is inhabited by populations, which are predominantly a Caucasoid stock (Table 7.1).

Table 7.1: A Timeline Chart Showing Important Events in the Pre history & History of Pakistan

Pre history	Reference	Early History	Reference
400 - 500,000 Years Before Present		**300- 331 B. C.**	
Evidence of Hominoid Inhabitation	Rendell, H. M. *et al.*, 1989	Pakistan part of Achaemenid Empire	Caroe, O. 1958
		326 B. C.	
30000 Years Before Present		Alexander invaded the subcontinent	*ibd*
Migration of humans from Central Asia	Sankiala, H. D. 1967		
		327- 320 B. C.	
		Pakistan part of Greek Empire	*ibd*
6500 B.C.			
Evidence of farming in Baluchistan	Costani, L. 1986	**323 - 227 B. C.**	
		Chandargupta founded Mauryan Empire	*ibd*
3300-2800 B.C.		Asoka next Mauryan king	
Pre Harrapan Phase of Indus Valley	Jarridge, J. F. 1981	Buddhism flourished	
2300 -2000 B.C.		**227 B.C - 75 A.D.**	
Emergence of Indus Valley Civilisation	Marshall, J. H. 1951	Bactrians, Scyths, Parthians, Sakas successively ruled the Indus Valley	*ibd*
1500 B. C.			
Indo Aryan migration from Central Asia	Cavalli- Sforza. *et al.*, 1993	**75 - 225 A.D.**	
		Kushan empire was established	*ibd*
		230 A.D - 642 A.D.	
		Sassanids ruled Pakistani territory	*ibd*
		712 - 1947 A.D.	
		Arab & Central Asian Muslim rulers invaded Subcontinent. Mughal Empire established in 1526 A.D. The British occupied the subcontinent in 1857, ruled for a century and left in 1947 dividing it into India and Pakistan.	

7.3 STATISTICAL ANALYSES

In order to study the phylogenetic relationship of different Pakistani populations the AMOVA and the genetic distances were used.

7.3.3 ANALYSES OF MOLECULAR VARIANCE (AMOVA)

In order to look into the phylogenetic relationships, the mutational events between the Y STR haplotypes have to be considered. These are best determined by the Analysis of Molecular Variance (AMOVA) developed for genetic analysis (Excoffier, L. *et al.*, 1992). This generates estimates of variance components and F-statistics analogues that give the measure of haplotypic diversity at different levels of hierarchical subdivision within and between population components for the cases in which multiple populations have been surveyed.

The AMOVA was performed using the Arlequin software (Schneider. S. *et al.,* 1996) which has been used previously for comparison of populations for Y STRs (Roewer, L. *et al.,* 1996 & deKnijff, P. *et al.,* 1997).

7.3.4 PHYLOGENETIC ANALYSES

The pairwise F_{ST} generated by ARLEQUIN 2.000, was used as a genetic distance. Different genetic distances were also calculated using MICROSAT (Minch, E. *et al*; 2000), which has been specially developed for microsatellites.

Distance matrices were used for the phylogenetic analysis of the Pakistani populations. For this purpose PHYLIP software version 3.5c (Felsenstein, J. 1993) and its embedded programs 'Neighbour' and 'Fitch' & 'CONTML', were used. 'UPGMA' and 'Neighbourhood Joining' trees generated by PHYLIP were viewed with TREEVIEW program (Page, R. D. M. 1996). Data from African populations was used to root the trees (Seielstad, M. *et al.*, 1999).

7.3.5 NETWORK ANALYSES OF THE KALASH POPULATION

A network of Y STR haplotypes (Cooper, G. *et al.*, 1996) was drawn for Kalash populations in order to show the relationship of the haplotypes within this population using standard method of connection of haplotypes, which are one step mutation different from each other.

7.3.6 ESTIMATION OF THE TIME OF ACCUMULATION OF DIVERSITY IN KALASH POPULATION

The time the Kalash population took to generate the diversity was estimated with the method based on estimating the number of mutations in an ancestral haplotype per locus per individual (λ) (Perez Lezaun, A. *et al.*, 1997). If a Poisson distribution of the mutation accumulation is assumed then the relationship of mutation rate and the time for the accumulation of diversity is given by the equation

$$\lambda = \mu t$$

where μ is the mutation rate and t is the time in number of generations required to accumulate the diversity.

7.4 RESULTS

7.4.1 AMOVA ANALYSES OF PAKISTANI POPULATIONS

The genetic homogeneity of the Pakistani populations was tested through AMOVA. Genetic variance was estimated using the different allele option, which takes into account the difference among the alleles of two haplotypes. Mean pairwise differences were calculated for all population pairs.

For AMOVA the populations were grouped as below:

A. Group I Kalash, Group II All other populations

B. Group I Kalash, Group II All other populations except Brosho

C. Group I Brosho, Group II All other populations except Kalash

D. Group I Brosho, Pushtoon & Punjabi, Group II Makrani, Baluchi & Sindhi

E. Group I Pushtoon & Punjabi, Group II Makrani, Baluchi & Sindhi

The grouping was done to explore the diversity between the Kalash and Brosho *vis a vis* the Pakistani population as a group and then to examine the levels of diversity between the southern (Baluchi & Sindhi) and northern (Punjabi & Pushtoon) Pakistani populations.

7.4.1.1 AMOVA Results for Different Populations and Groups

AMOVA showed that the variation between the individuals of populations was much greater than that between the populations or to the arbitrary groups, formed for the analysis.

Variation between Kalash and the Pakistani populations was slightly less than that between the Brosho and Pakistani population. Variation among the Pakistani main land populations was the lowest when the Brosho population was included than when it was not included. The variation between the major Pakistani population was very low, when the northern and southern populations were tested as groups (Table 7.2).

7.4.1.2 Variation Between the Pooled Pakistani Populations & Kalash /Brosho

Since the Kalash and the Brosho showed differences in AMOVA analysis between the various mainland populations, the mainland populations were pooled together and then tested against Kalash and Brosho in order to perform the AMOVA.

Table 7.2: AMOVA Results For Pakistani Populations

A.

Group I Kalash and Group 2 All Other Pakistani Populations

Source of Variation	d.f	Sum of Squares	Variance	% Variation
Among Groups	1	69.171	0.10701 Va	4.51
Among Populations Within Groups	5	126.806	0.14212 Vb	5.99
Among Individuals Within Populations	555	2356.724	2.12317 Vc	89.5

B.

Group I Kalash and Group 2 All Other Pakistani Populations Excluding Brosho

A.

Among Groups	1	71.12	0.14703 Va	6.22
Among Populations Within Groups	4	72.846	0.09077 Vb	3.84
Among Individuals Within Populations	496	2110.058	2.12707 Vc	89.94

C.

Group I Brosho and Group 2 All Other Pakistani Populations Excluding Kalash

Among Groups	1	54.933	0.18378 Va	7.55
Among Populations Within Groups	4	72.846	0.09033 Vb	3.71
Among Individuals Within Populations	453	1957.467	2.16056 Vc	88.74

D.

Group I Brosho ,Pushtoon, Punjabi, Group II Makrani, Baluchi, Sindhi

Among Groups	1	34.764	0.022Va	0.95
Among Populations Within Groups	4	92.042	0.12865 Vb	5.57
Among Individuals Within Populations	452	1951.69	2.15886Vc	93.48

E.

Group 1 Pushtoon, Punjabi, Group II Makrani, Baluchi, Sindhi

Among Groups	1	33.036	0.0467Va	2.05
Among Populations Within Groups	3	38.832	0.05892 Vb	2.59
Among Individuals Within Populations	392	1699.084	2.1672Vc	95.35

This showed that the variation between the Pakistani population and Brosho was 10%, while it was lower (7%) between them and Kalash.

7.4.1.3 Locus By Locus AMOVA

The diversity among the groups and among the populations within the groups was a measure of the diversity of the different loci. When each locus was studied for its share in the diversity a relationship emerged between the role of different loci for differentiating the groups and the populations.

The results showed that for all the groups major contribution of diversity was by the locus DYS393, which accounted for 7-15% variation among the populations.

For Kalash and Pakistani populations (Group B) DYS390 was responsible for the major share of variation among the two groups. For Brosho and Pakistani populations (Group C) DYS392 & 393 contributed 36% of variation between the two groups (Table 7.3)

7.4.1.4 Pairwise F_{ST} Calculation for Pakistani Populations

Inter population diversity can be measured by F_{ST} (Deka, R. *et al.*, 1995). The pairwise F_{ST} also showed the same measure of inter population diversity as shown by AMOVA (Table 7.2). The average F_{ST} between the mainland populations and Brosho (0.122) was found to be higher than that between the mainland populations and the Kalash (0.11). On the population pair basis the Punjabi population was more related to the Brosho, than the Kalash.

Makrani population also showed higher variation to other populations. The highest value F_{ST} was between Makrani/Pushtoon pair among the mainland Pakistani populations.

Table 7.3: Locus By Locus AMOVA Results For Groups of Pakistani Populations

	Group A		Group B		Group C		Group D		Group E	
Locus	Group	Population	Group	Population	Group	Population	Group	Population	Group	Population
DYS 391	0.7	2.38	2.84	0.72	5.63	0.67	1.04	1.64	0.47	0.23
DYS392	3.67	5.07	5.21	0.24	**15.02**	0.05	-0.22	4.97	-0.06	0.11
DYS393	-2.44	**15.04**	4.3	**7.99**	**20.89**	**7.02**	**0.37**	**14.94**	7.54	3.48
DYS19	-0.91	3.32	-0.99	3.35	-0.34	3.3	-0.88	3.84	-1.32	4.34
DYS390	**13.69**	6.4	**15.22**	4.1	8.36	4.15	3.66	4.66	3.26	2.29
DYS389I	7.98	3.9	6.44	4.53	-2.24	5.12	2.34	2.9	1.22	4.08
DYS389II	5.37	4.93	7.12	4.01	3.28	4.15	0.02	5.1	2.32	2.55
Total Diversity	**28.06**	**41.04**	**40.14**	**24.94**	**50.6**	**24.46**	**6.33**	**38.05**	**13.43**	**17.08**

The table shows the %age of diversity within groups and within the populations

The significant finding was that for the paternal lineages Punjabis were more closely related to Sindhis (as were Baluchis to Makranis) than to any other population. On the basis of pairwise differences between the haplotypes the Kalash showed a lower value than all other populations indicating a closer relationsship of haplotypes within the Kalash than within other populations (Table 7.4).

7.4.1.5 Comparison of Autosomal & Y STR F_{ST}

Since the data for F_{ST} for autosomal STR was available a comparison was possible. It seemed that the values for this statistic were much higher for Y STRs than that for the autosomal STRs (Table 5.8a & b). The value for autosomes and the Y STR can be related through the formula:

$$\frac{Fst_{au}}{(1-Fst_{au})} = \frac{Fst_{y}}{4(1-Fst_{y})}$$ (Perez-Lezaun, A. *et al.*, 1997)

The correction for the small size of the Y chromosome was thus incorporated and higher F_{ST} values obtained than those generated using the autosomal STRs (Table 7.4).

7.4.1.6 Mean Pairwise Difference Between Populations

The mean pairwise differences showed that these were different between different pairs and the highest values obtained were for the Baluchi population, while the lowest were for the Kalash (Table 7.5).

7.4.2 PHYLOGENETIC ANALYSES

The basic structure of the trees was similar. The Punjabi and Sindhi population cluster together showing common origin of the Y chromosome of these two populations, though the Pushtoons also seem to have contributed to their gene pool but they branch off quite early. A similar cluster was seen for the Baluchi and Makrani

Table 7.4: Comparison of Y chromsome & Autosomal STR F_{ST}

a. F_{ST} Generated Using the Y Chromosome STR Data

	Baluchi	Brosho	Kalash	Makrani	Punjabi	Pushtoon	Sindhi
Baluchi		0.030	0.024	0.002	0.010	0.020	0.008
Brosho	0.110		0.044	0.048	0.025	0.048	0.028
Kalash	0.090	0.150		0.028	0.028	0.041	0.025
Makrani	0.010	0.160	0.100		0.020	0.031	0.013
Punjabi	0.0400	0.090	0.100	0.070		0.008	0.002
Pushtoon	0.07	0.160	0.140	0.110	0.030		0.013
Sindhi	0.030	0.100	0.090	0.050	0.010	0.050	

Above the diagonal are the corrected values of F_{ST}

Below the diagonal are the actual F_{ST} values

b. F_{ST} Generated Using the Autosomal STR Data

	Baluchi	Brosho	Kalash	Makrani	Punjabi	Pushtoon
Brosho	0.007					
Kalash	0.034	0.027				
Makrani	0.009	0.014	0.040			
Punjabi	0.012	0.007	0.025	0.022		
Pushtoon	0.002	0.009	0.027	0.026	0.013	
Sindhi	0.002	0.003	0.023	0.006	0.012	0.009

Table 7.5: Mean Pairwise Differences Between Populations for Y STRs

	Kalash	Makrani	Baluchi	Sindhi	Pushtoon	Punjabi	Brosho
Kalash	5.04	6.54	7	7	6.46	6.75	6.94
Makrani	0.694	6.66	7.3	7.37	6.97	7.28	8.11
Baluchi	0.609	0.13	7.7	7.66	7.22	7.53	8.29
Sindhi	0.861	0.424	0.188	7.24	6.76	7.09	8.04
Pushtoon	1.023	0.733	0.453	0.232	5.82	6.5	7.48
Punjabi	0.823	0.546	0.267	0.063	0.176	6.82	7.48
Brosho	1.158	1.52	1.167	1.158	1.3	0.804	6.53

Above the diagonal average number of pairwise difference

Below the diagonal are the genetic distances

Genetic Distance has been calculated by the formula $D_{12} - [(D_1 + D_2)/2]$, whereas,

D_{12} is the mean pairwise difference between pop1 and pop 2,

D1 and D2 are the pairwise differences within Pop 1 & Pop 2

populations, though historically these are two distinct populations but Makranis have remained on the coastal areas of the Baluchistan and Sindh for a long time and intermingled with the Baluch.

The position of Brosho was significantly different in the trees generated by the pairwise F_{ST} and the other distances. In the tree generated using pairwise F_{ST} as distance the Brosho appear as distant as Kalash from other Pakistani populations, however the positions changes with other distances specially the allele sharing distance (D*as*) the Brosho join the Punjabi and Sindhi cluster (Figure 7.1)

7.5. ORIGINS OF KALASH

Since the origins of Kalash is a hotly debated subject among the linguists and anthropologists, and samples from two different areas were available from the Kalash valley, it was decided to examine the Y STR data for the Kalash in detail.

In Kalash the male lineages are remembered and memorized by locals (Asad, M. pers comm). Information gathered in this way has indicated that a first cousin of Rajawai, the last Kalash king, colonised Rumbur. In Bumboret few related lineages from Rumbur exist in the shape of Rajawai ancestry while at least three other lineages have been described (Cacopardo, A. 1996).

The methods applied here include the network approach for connecting the haplotypes which are one or two step different from each other, which has been applied for the Kalash population.

Figure 7.1a: UPGMA Tree Showing the Phylogenetic Relationship of Pakistani Populations Using Pairwise F_{ST} as Distance Calculated for Y STR Haplotype

The tree was generated using programs Neighbour and CONTML embedded in "Phylip" using pairwise F_{ST} as genetic distance for Y STR haplotype.
The African data was used for rooting the tree.

Figure 7.1b: Neighbourhood Joining Tree Showing the Phylogenetic Relationship of Pakistani Populations Using Pairwise F_{ST} as Distance Calculated for Y STR Haplotype

The tree was generated using programs Neighbour and CONTML embedded in "Phylip" using pairwise F_{ST} as genetic distance calculated for Y STR haplotype.
Position of the Makrani and Baluch populations has changed with reference to Sindhi and Punjabi population.

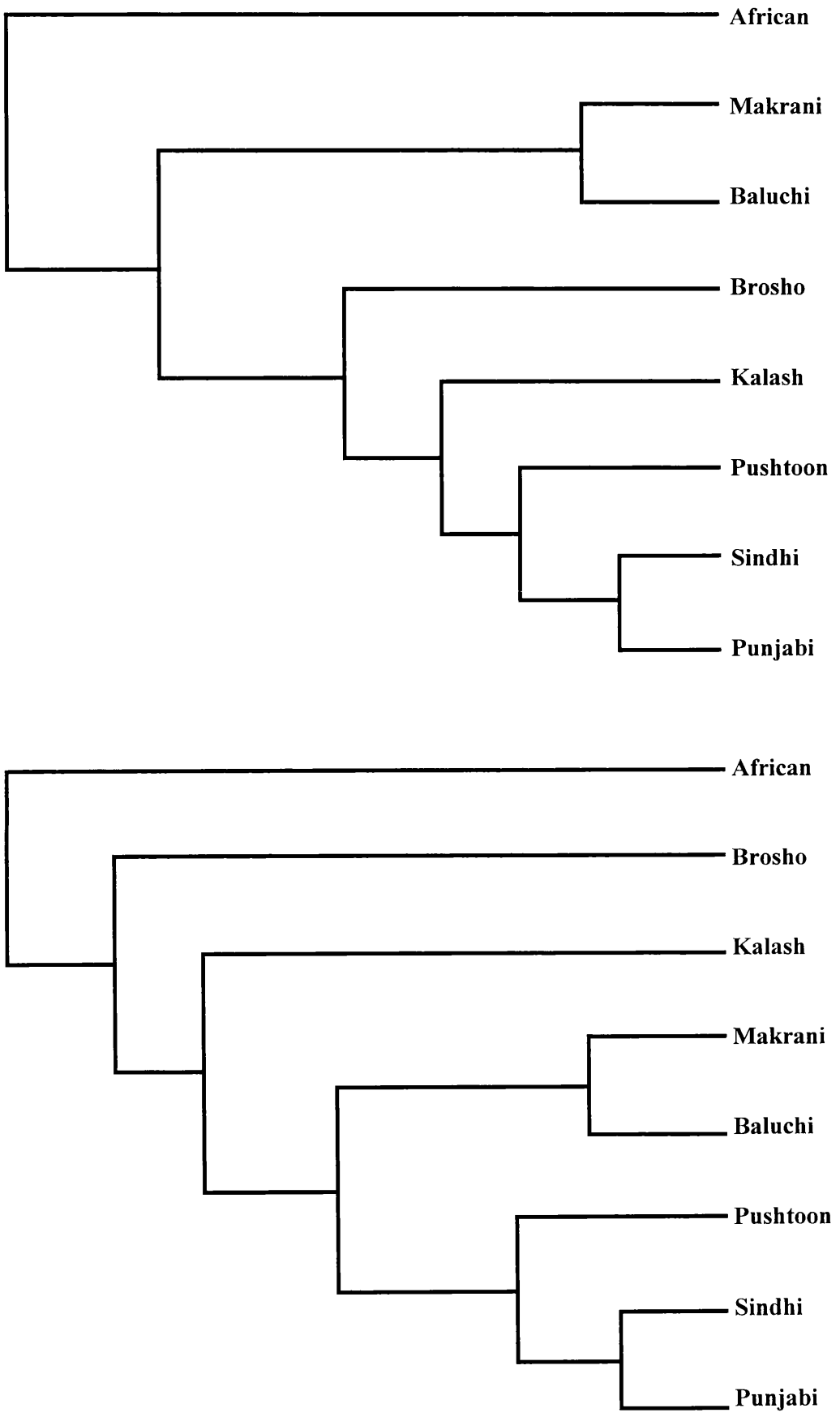

African
Makrani
Baluchi
Brosho
Kalash
Pushtoon
Sindhi
Punjabi
African
Brosho
Kalash
Makrani
Baluchi
Pushtoon
Sindhi
Punjabi
Scale:
0.01

Figure 7.1c: UPGMA Tree Showing the Phylogenetic Relationship of Pakistani Populations Using Allele Shared Distance (D*as*) Calculated for Seven Y STR Loci

The UPGMA tree was generated using pairwise allele sharing distance D*as*, which was calculated using the Program "Microsat" using the seven Y STR loci DYS391, 392, 393, 19, 389I, 389II & 390.

The tree shows that the Brosho comes to lie within the mainland populations when this distance measure was used instead of the F_{ST}.

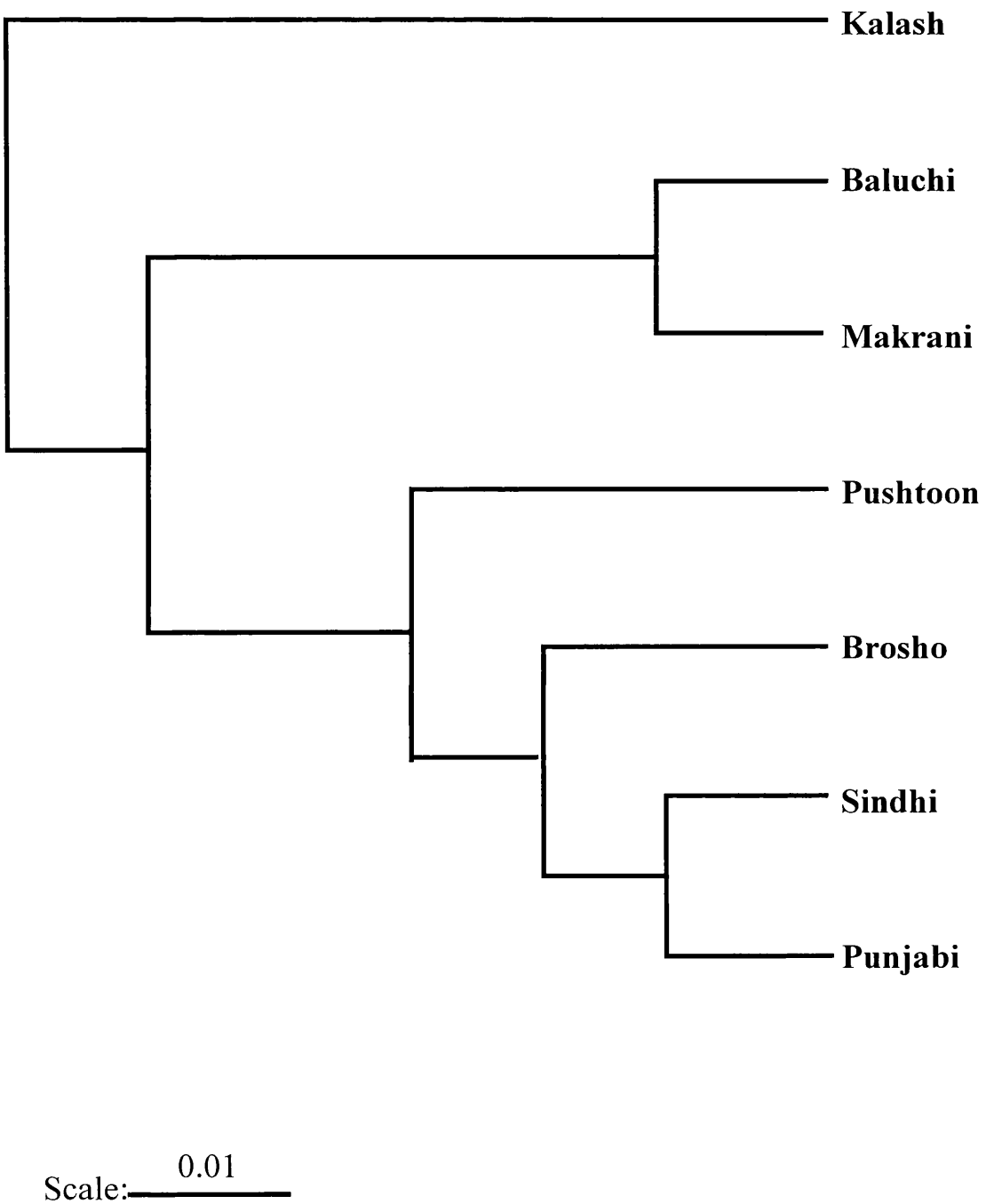
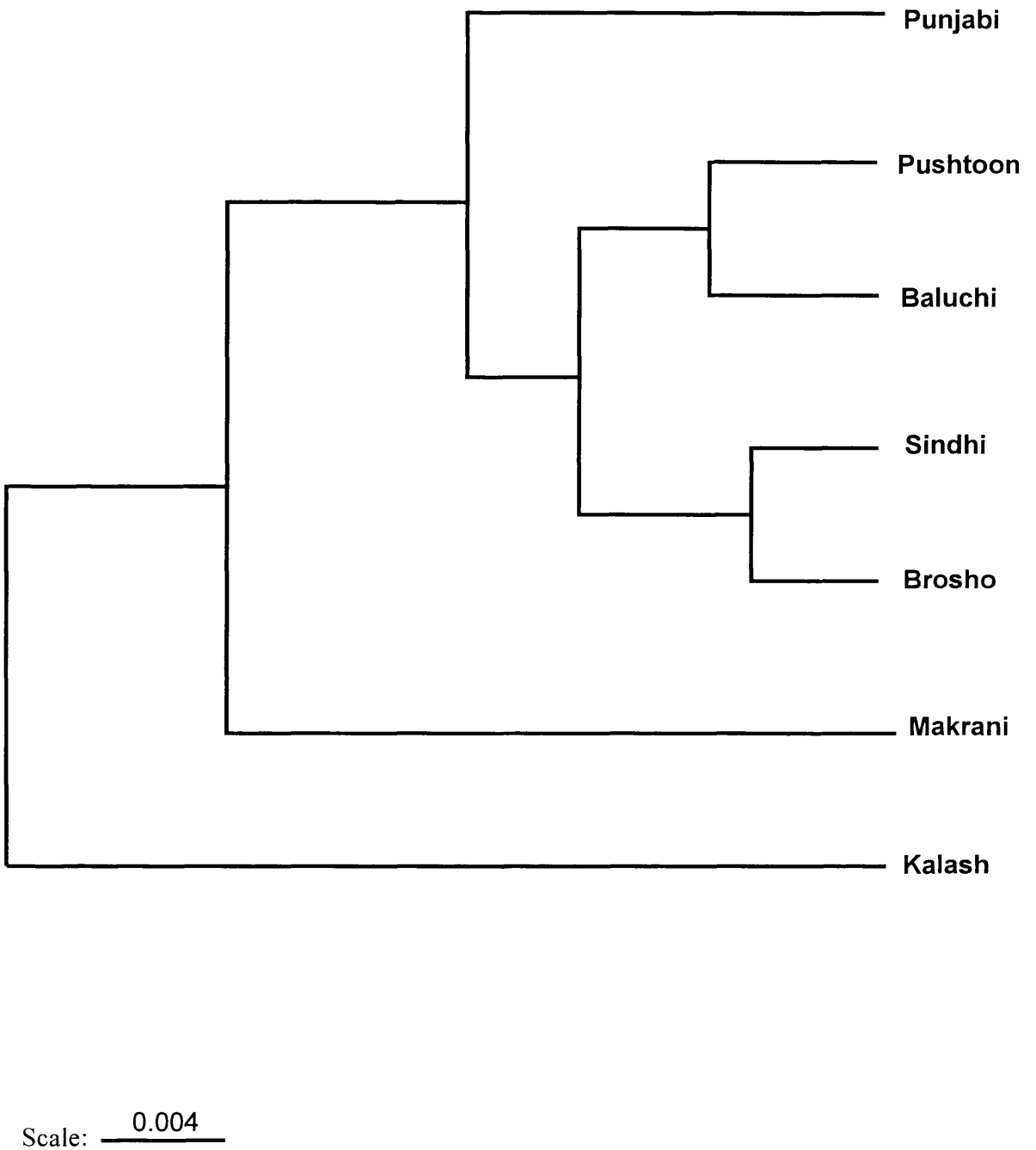

Kalash
Baluchi
Makrani
Pushtoon
Brosho
Sindhi
Punjabi
Scale: 0.01

Figure 7.1d: UPGMA Tree Showing the Relationship of Pakistani Populations Using F_{ST} Calculated using Autosomal STR Loci Genotype Data

The phylogenetic relationship of Pakistani populations changed significantly when autosomal STR data for the loci D3S1358, vWA & FGA was used. The cluster of Punjabi and Sindhi populations shown in the trees generated by the Y STR data breaks down. The Pushtoon and Baluch form a cluster. The Makranis separate from all populations much earlier and Brosho come to occupy a position within the general group of Pakistani populations. The Kalash remain the most distant population.


Punjabi
Pushtoon
Baluchi
Sindhi
Brosho
Makrani
Kalash
Scale: 0.004

7.5.1. NETWORK ANALYSIS OF KALASH HAPLOTYPES

A detailed analysis of correlation of the Kalash haplotypes and the relationship of the haplotypes had to be generated in a more obvious and simple way. It was therefore decided that the approach of network construction would be useful in this regard. All alleles were redesignated, with the smallest allele (in size) at a locus being 1 and increasing the number for each successive allele. This generated a haplotype in a format of 1.2.3.4.5, which could be easily added to give a particular 'weight' to a haplotype (Cooper, G. *et al.*, 1996). Haplotypes having same weight were placed on the same horizontal plane. Each haplotype, which was one or two step distance (mutation), from the other, was joined to the nearest such haplotype. This eventually worked out into a network (Figure 7.2).

7.5.1.1 Interpretation of the Kalash Network

The network showed two separate groups for the Kalash, one from Bumboret and the other from Rumbur. Though one haplotype from Rumbur did embed with those of Bumboret, except that there was an obvious grouping. Most haplotypes consisted of shorter alleles and were connected to the haplotypes of lower weight following the mutational pattern of STRs.

The group of haplotypes from the Bumboret area is clearly divided into three portions, a lower portion containing two common haplotypes, a middle portion which has a group of closely related hapolotypes and an upper portion which contains another group of related haplotypes. It seems that two ancestral lineages exist in Rumbur while three exist in Bumboret.

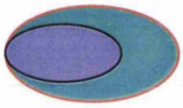
A significant inference from the network is the location of those haplotypes which are connected if two steps of mutation are allowed. These are peripheral and also consist of longer alleles which means that these are more recent than the other haplotypes.

Figure 7.5: Network of 7 Locus Y STR Haplotypes Drawn for the Kalash Population

7 Locus Y STR Haplotype in Kalash from Bumboret Valley

7 Locus Y STR Haplotype in Kalash from Rumbur Valley

Haplotype Shared between the Bumboret and Rumbur Valleys



7 Locus Y STR Haplotype from Bumboret Valley which can be connected only if >2 mutations are assumed

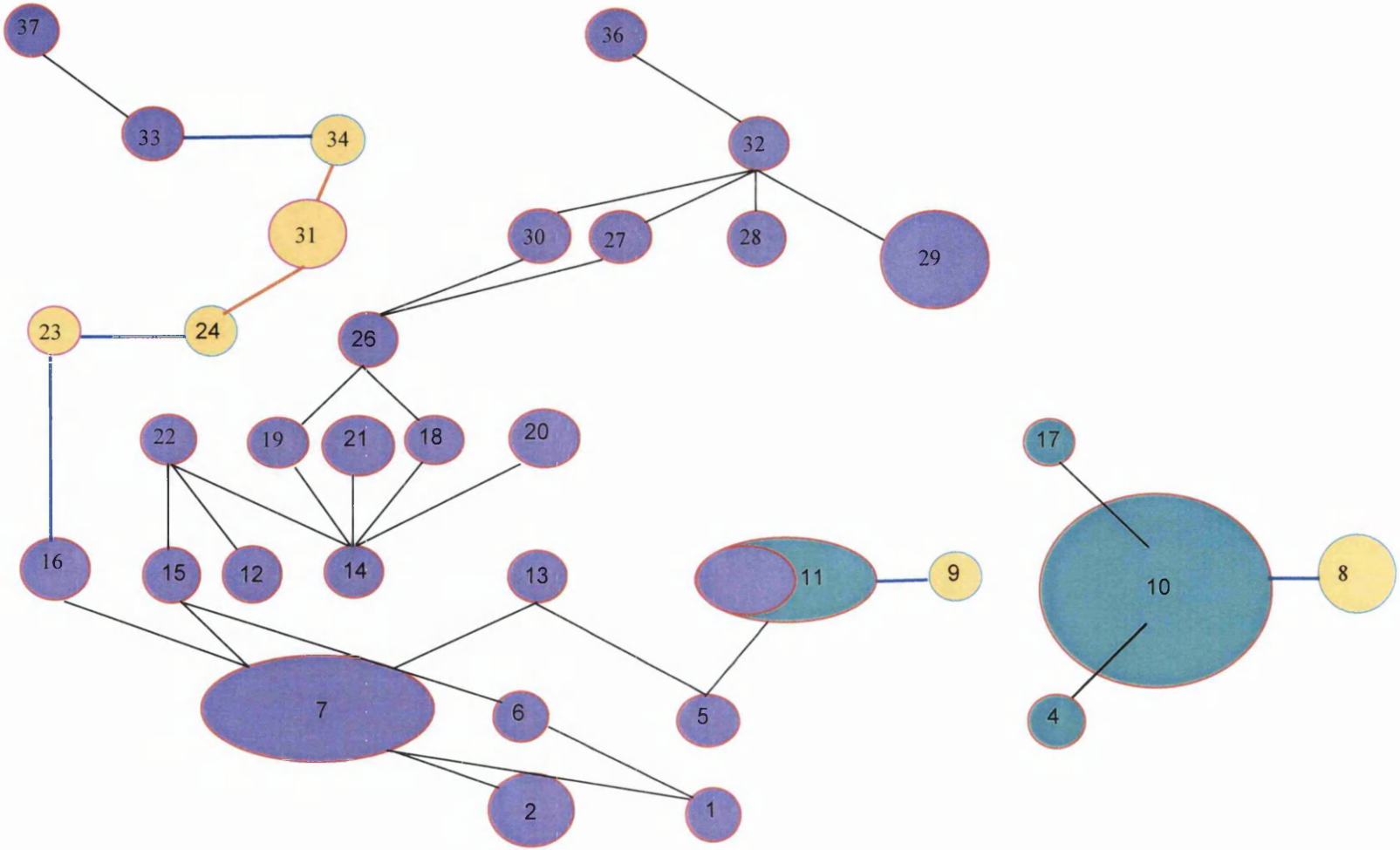
7 Locus Y STR Haplotype from Rumbur Valley which can be connected only if >2 mutations are assumed

Connection lines showing one mutational (one repeat over 7 loci) difference between two haplotypes

two mutational differences difference between two haplotypes

three mutational differences difference between two haplotypes

The size of spheres/ovals shows the relative frequency of the haplotype in the Kalash population. Though the network was drawn by hand, it was later improved by comparing to the one generated by a computer program kindly provided by Dr.William Amos, Cambridge University, UK.


37
33
34
31
23
24
36
32
30
27
28
29
26
22
19
21
18
20
16
15
12
14
13
11
9
17
10
8
7
6
5
4
2
1

7.5.2. ESTIMATION OF THE TIME OF ACCUMULATION OF DIVERSITY IN KALASH POPULATION

In order to estimate the time for accumulation of diversity an ancestral haplotype has to be selected so that all other haplotypes could be compared with it and the number of mutations counted. Among the two samples of Kalash from Bumboret and Rumbur only one haplotype was common as it was at low frequency, so other haplotypes were considered as ancestral. The first was the most common haplotype, but the mutations per locus per individual were too high with this haplotype, further this haplotype had only two connections on the network. A haplotype (15/10/26/24/11/11/12) detected at the lowest horizontal plane of the network which also occurred at moderate frequency was selected as ancestral.

The λ was 0.2 with this haplotype as ancestral (Table 7.6). The mutation rate of 1.2×10^{-3} with a confidence limit of 4.6×10^{-4} to 2.8×10^{-3} was used for the estimation of haplotype age (Bianchi, N. O. *et al.*, 1998). The age of Kalash haploytpe diversity was thus estimated to be 4000 years before present with a confidence limit of 1,800 to 10,800 years before present.

7.6 DISCUSSION

7.6.1 THE AMOVA ANALYSES

Pakistani populations were studied and the variability of the populations compared using AMOVA analysis. The genetic diversity among human groups, is only a fraction of the global genetic diversity even when distant racial or geographical groups are considered (Barbujani, G. *et al.,* 1997), thus AMOVA yields smaller values for the diversity within the groups of populations than within the populations from same area, the same was shown for the data analysed.

Table: 7.6: Calculation of Number of Mutations From Ancestral Haplotype

Bumboret Haplotypes From Four Locations							No [1]	Mutations [2]	Mean [3]
DYS19	389I	389II	390	391	392	393			
14	10	26	23	10	11	12	1	1	0.143
14	10	27	23	10	11	12	1	1	0.143
14	10	27	23	10	11	13	2	3	0.428
14	11	27	23	11	11	12	1	4	0.511
14	11	27	23	10	12	13	3	5	0.714
14	11	27	23	10	11	12	1	3	0.428
14	11	27	23	11	12	13	1	6	0.857
14	11	27	23	10	11	13	1	4	0.511
14	11	27	23	10	12	12	1	4	0.511
14	11	27	24	11	12	12	1	6	0.857
14	11	27	24	11	11	13	3	6	0.857
14	11	27	24	11	12	13	8	7	1
14	11	27	24	10	12	13	1	6	0.857
14	11	27	25	11	12	13	1	8	1.143
15	9	26	24	11	11	12	1	3	0.428
15	9	26	24	10	11	12	1	2	0.286
15	9	26	24	11	11	12	2	3	0.428
15	**10**	**26**	**23**	**10**	**11**	**12**	**9**	**0**	
15	10	26	23	10	11	13	1	1	0.143
15	10	26	22	10	11	12	5	1	0.143
15	10	27	23	10	11	12	1	1	0.143
15	11	27	23	10	11	12	1	2	0.286
15	10	26	23	11	12	13	2	3	0.428
15	10	26	23	10	11	13	1	1	0.143
16	10	26	24	10	12	13	1	4	0.511
17	10	27	24	11	11	13	1	6	0.857
17	11	27	24	10	11	12	1	5	0.714
Total							53		
Rumbur Haplotypes From Three Locations									
14	9	26	23	10	12	13	1	4	0.511
14	10	26	23	10	11	12	1	1	0.143
14	11	27	22	10	14	12	1	7	1
15	9	25	23	10	11	14	3	4	0.511
15	9	26	24	11	11	12	4	3	0.428
15	9	27	23	11	11	12	1	3	0.428
15	10	26	24	11	11	13	1	3	0.428
15	10	26	24	10	11	12	1	1	0.143
16	9	25	23	10	11	14	32	5	0.714
16	9	25	23	10	11	15	2	6	0.857
16	9	25	23	11	11	13	2	5	0.714
17	10	26	24	11	11	13	1	5	0.714
17	10	26	24	10	11	14	1	5	0.714
Total							51		**20.775**

1 The number of haplotypes observed

It had been suggested that if a correction is applied for the smaller number of Y chromosomes the F_{ST} values for autosomal and Y STRs correspond (Perez-Lezaun, A. *et al.*, 1997). This was not found to be the case with the Pakistani population. An average across all the pairwise F_{ST} for the autosomal STRs was significantly lower than the same for Y chromosome STRs (0.015, 0.023 respectively).

When the pairwise F_{ST} values for Y STRs were compared to the corrected autosomal F_{ST}, 16 out of 21 values were higher for Y STRs. These were almost 2 to 7 times the autosomal F_{ST} in case of Brosho population. The reason for higher interpopulation Y STR F_{ST} might be higher mutation rate of Y STRs or higher female migration than the males, as it has been shown that the Y STRs mutation rates are comparable to those of the autosomal STRs (Bianchi, N. O. *et al.*, 1998).

It follows that in Pakistani populations the female migration rates might have been higher than the males as elucidated in mtDNA studies in other populations of the world (Seielstad, M. T. *et al.*, 1998).

The high F_{ST} value for the Baluchi and Pushtoon populations appears significant. After the application of the correction for smaller number of Y chromosome the value was 0.02, which was higher than that from the analysis of the autosomal data (-0.005-0.006). This fact can be attributed to the strict patrilocal tribal society of both the Pushtoons and Baluchis. Within the major Pakistani linguistic groups, Baluchi appear to be the most diverse on the basis of pairwise difference between the haplotypes (7.7), which if compared to Kalash (5.04) or the Pushtoon (5.82), is significantly high.

When the Pakistani data was compared to that of the Central Asia (Perez-Lezaun, A. *et al.*, 1999), the diversity was higher in the population of Central Asia than the Pakistani populations. Further the Pakistani populations are nearer to each other than are the central Asian populations are, to each other, where F_{ST} values and the pairwise differences are much higher than have been detected within the Pakistani populations (Table 7.5).

AMOVA of the Pakistani population with Kalash and Brosho revealed that the Pakistani/Brosho pair had higher diversity between them than the Kalash/Pakistani pair. Further the corrected F_{ST} for Pak/Brosho pair was 0.028 whereas that for Pak/Kalash pair was 0.024 which was in contrast to the value between Kalash and Brosho which even after the application of correction of small Y size was very high (0.042).

7.6.2 ORIGINS OF KALASH

More haplotypes were shared between Kalash and Pakistani populations than the Broshos point to comparative nearness of Kalash to the Pakistani population. On the other hand if these haplotypes were shared because of common ancestry then the most should have been between Kalash and Pushtoons, which are in geographical proximity. However Kalash and Pushtoons are the farthest to each other on the basis of shared haplotypes and F_{ST}. This might also be due to the fact that the Pushtoons spread to their present location much later than other populations having originated most probably from the Bactria (Caroe, O. 1958). This also indicates that the Kalash have been in their present location for a long time.

The haplotype data shows that there are many haplotypes that are shared between the Kalash and Pakistani populations while there are others that though not shared are just one or two steps away from those in other populations. The time for the accumulation for the diversity in Kalash, which is equivalent to the time of divergence between the Bumboret and Rumbur populations, has been estimated to be 4000 years before present with a range of 1,800-10,800 yrs. This though a rough estimate does provide with a date which almost coincides with linguistic estimates and thus it seems plausible that the Kalash entered the subcontinent before this time.

7.6.3 PHYLOGENETIC RELATIONSHIP OF PAKISTANI POPULATIONS

Phylogenetic analysis, AMOVA and F_{ST} showed that the major Pakistani populations were much more closely related to each other than to the Kalash or the

Brosho as a group. The Y STR haplotype consisting of seven loci retained enough resolution for comparing the closely related populations of Pakistan with each other.

Among the phylogenetic trees generated the neighbourhood joining tree was most consistent with the AMOVA, and the available historic, linguistic and archaeological evidence.

The Punjabi and Sindhi populations reside in the Indus Valley and might be regarded as the two original populations with some admixture due to regular ingress and inhabitation of people from Iran and Central Asia (Dani, A. H. 1981; Caroe, O. 1958). The history of the populations of Sindh and Punjab run parallel till the languages of the two populations separated sometimes between 750-1400 A.D. (Bordie, J. G. 1981). The Pushtoons have a predominant Central Asian and Iranian admixture due to the successive invasions by the rulers of the two areas and resultant population influx and the Baluchi tribes have an Iranian origin. Phylogenetically also the Baluchi and Pushtoons appear relatively distant to Punjabi and Sindi populations.

The differences between the trees generated using Y STRs and the autosomal STRs show that the Y STRs might be better for inferring phylogeny of the closely related populations.

In conclusion the analyses have revealed that a higher degree of substructure exists in the Pakistani populations for Y STRs than that determined for the autosomal STRs, suggesting a higher male mediated diversity within and between the Pakistani populations. The AMOVA and the phylogenetic analyses have shown closer genetic relationship of Punjabi/Sindhi & Baluchi/Makrani populations, which is consistent with the recent history of these populations. The analyses for the Kalash have revealed that they are genetically distant from other Pakistani populations. The approximate date of their migration to their present location in Pakistan indicated that the Kalash have most probably arrived in the subcontinent with the early waves of the Indo Aryan migrations.

CHAPTER 8: SUMMARY AND FUTURE WORK

The genetic structure of Pakistani population had not been studied before therefore it was necessary to carry out such work before the establishment of a database and implementation of forensic DNA analysis in Pakistan. For this study appropriate samples from all the major populations and three smaller ethnic groups were collected in Pakistan.

8.1 AUTOSOMAL STR ANALYSIS

Three autosomal loci D3S1358, vWA and FGA were selected mainly because they could be compared to other populations and also owing to their high discrimination power and validated status for forensic use. New primers were designed for the three loci. 'In house' allelic ladders were prepared for the loci in order to designate the alleles unambiguously. Amplification of the three loci for the samples of Punjabi, Pushtoon and Sindhi populations was performed using the new primers. Other samples were profiled using the AmpFℓSTR® Blue kit (Applied Biosystems CA, USA).

The statistical analysis revealed that there were significant levels of substructure within the Pakistani populations. A simulation study was conducted which showed that incorporation of F_{ST} in the calculation of allele and genotype frequencies leads to lowering of the likelihood ratio of a match. Thus it is a conservative approach and does not undermine the strength of DNA evidence. The allele frequencies were significantly different between some of the Pakistani populations which demonstrated the importance of generating databases for the sub population groups before applying the data to casework.

8.2 Y CHROMOSOME STR ANALYSIS

The male samples available from the Pakistani ethnic groups were profiled for seven Y STR loci (DYS19, 389I, 389II, 390, 391, 392 & 393) in two multiplex PCRs. 'In house' sequenced allelic ladders were prepared for both multiplex systems in order to have a precise allelic designation.

The Y STR analysis revealed higher haplotype diversity for the haplotype defined by the loci for Pakistani population in comparison to Caucasian populations. The discrimination capacity of this haplotype was only slightly lower than that of the extended haplotype which includes two more loci DYS385 & YCAII to the seven locus Yh1, in European populations. The Y STR analysis in Pakistani populations would be thus very valuable even when the Yh1 is used for forensic casework.

Higher F_{ST} values were obtained for Y STRs than the autosomal STRs and the populations remained well differentiated for the paternal lineage that might reflect greater female migration rates contributing to lower F_{ST} values determined by autosomal STR analysis. AMOVA was performed to analyse the data, which revealed different patterns of diversity between the populations and the various groups of populations.

8.3 PHYLOGENETIC ANALYSIS

Y STR and autosomal STR data were used for phylogenetic analysis. Genetic distances were calculated using STR data and phylogenetic trees were developed using the software PHYLIP. The phylogenetic trees showed that the Punjabi and Sindhi populations which have throughout their known histories inhabited the Indus valley, were distinctly grouped together. The Pushtoons were also grouped with these populations with an earlier branching from the Sindhi/Punjabi cluster. The Baluch and Makrani populations form another cluster. Position of the Kalash and Brosho remained distant to the other populations.

The Kalash population was investigated in detail for determining their origins. The levels of diversity and the estimation of time for the accumulation of diversity in the Kalash pointed towards an early Indo Aryan origin of the Kalash rather than a European one as has been hypothesized before.

8.4 FUTURE WORK

The database of three autosomal STR loci has a high power of discrimination however for higher exclusion power particularly in paternity analysis it would be necessary to develop a DNA database for Pakistani populations using more loci.

In this connection a panel of loci can be selected and can be either the AmpFℓSTR® SGMPlus™ Kit loci (Applied Biosystems CA, USA) or the thirteen loci included in the Combined DNA Index System (CODIS). Since Pakistan has closer ties with the UK and the Pakistani medicolegal system is nearer to that of the UK it would be better to use the AmpFℓSTR® SGMPlus™ kit as the same is being used by the FSS, UK in criminal cases. This would allow exchange of data between the UK and Pakistan and also other European countries which might be generating databases using the same loci.

The present database of Y STRs would benefit by the addition of data for other loci like DYS385, YCAII, DXY156Y and the new Y STRs recently described, in order to achieve an increased discriminatory power of the Y chromosome STR analysis. Such a database would be a powerful adjunct to the autosomal analysis in appropriate cases.

An exciting area of future studies is the Y chromosome SNPs which would allow to study the population structure and history more definitively.

The other important future work is the establishment of an accredited forensic DNA laboratory in Pakistan in order to ensure reliable results in cases of personal identification and disputed paternity.

APPENDICES

APPENDIX 1

Appendix 1

AUTOSOMAL STR GENOTYPE DATA FOR THE LOCI D3S1358, vWA & FGA

Population: Punjabi

S. No's	Sample	D3S1358		vWA		FGA	
1	J1	15	17	15	17	19	25
2	J2	14	14	17	17	21	25
3	J3	16	16	16	17	18	22
4	J4	16	17	16	17	20	25
5	J5	15	16	13	18	22	25
6	J6	15	16	16	17	21	25
7	J7	16	16	17	19	19	19
8	J8	14	14	14	15	22	26
9	J9	13	15	15	19	22	25
10	J10	15	16	15	16	22	26
11	J11	14	16	16	18	22	26
12	J12	14	14	14	17	21	25
13	J13	16	16	16	17	22	23
14	J14	14	15	16	19	22	22
15	J15	13	16	15	17	24	27
16	J16	14	15	17	19	20	20.2
17	J17	16	16	16	16	21	21
18	J18	14	15	14	18	19	20
19	J19	14	16	17	19	23	25
20	J20	15	15	16	17	23	24
21	J21	14	14	17	17	21	22
22	J22	14	16	16	17	23	23
23	J23	15	16	15	16	23	23
24	J24	13	15	13	15	20	25
25	J25	15	16	16	18	24	24
26	J26	18	19	17	18	23	23
27	J27	15	16	17	18	23	23
28	J28	13	17	17	18	23	23
29	J29	14	16	17	18	22	23
30	J30	14	17	16	19	18	19
31	J31	14	15	16	19	22	24
32	J32	14	17	16	16	22	22
33	J33	17	17	13	13	18	24
34	J34	13	15	15	16	18	24
35	J35	13	14	15	17	21.2	22
36	J36	14	15	17	18	18	20
37	J37	14	15	16	18	23	27
38	J38	14	15	14	17	20	25
39	J39	14	16	17	17	20	24
40	J40	13	17	16	20	23	24
41	J41	15	16	16	17	24	24
42	J42	14	18	16	17	19	26
43	J43	13	15	18	19	20	26
44	J44	17	17	14	16	21	26
45	J45	16	17	18	19	21	22
46	J46	17	17	17	18	20	21
47	J47	14	15	18	19	21	24
48	J48	15	17	15	19	21	24
49	J49	14	16	17	18	21	23
50	J50	14	16	14	18	22	26
51	J51	14	16	15	16	21	26
52	J52	16	16	17	17	23	25
53	J53	15	17	14	18	20	21
54	J54	14	15	15	17	21	22
55	J55	14	17	15	16	21	22
56	J56	12	14	15	17	21	22
57	J57	14	15	17	18	19	24

Population: Punjabi

S. No's	Sample	D3S1358		vWA		FGA	
58	J58	14	17	14	19	21	24
59	J59	14	16	17	18	23	24
60	J60	16	16	18	18	20	26
61	J61	14	14	16	17	18	22
62	J62	14	15	15	17	21	23
63	J63	14	15	15	18	21	23
64	J64	13	17	14	15	23	24
65	J65	12	13	13	18	21	23
66	J66	14	17	15	18	18	20
67	J67	13	16	14	18	22	23
68	J68	14	16	13	18	20	24
69	J69	15	16	17	18	20	24
70	J70	14	15	17	19	21	25
71	J71	15	16	11	12	21	22
72	J72	15	16	16	17	21	22
73	J73	15	15	17	18	21	23
74	J74	14	15	13	17	21	23
75	J75	14	14	13	14	21	25
76	J76	15	16	12	15	20	25
77	J77	14	16	14	16	20	25
78	J78	16	17	16	17	20	24
79	J79	15	16	15	19	21	27
80	J80	14	15	15	16	20	25
81	J81	14	15	16	16	23	25
82	J82	14	15	15	18	21	26
83	J83	16	16	17	18	20	22
84	J84	14	16	17	18	22	24
85	J85	14	15	16	16	21	25
86	J86	14	15	17	18	19	21
87	J87	16	16	19	19	21	26
88	J88	14	16	16	17	19	21
89	J89	15	15	15	19	21	22
90	J90	15	15	17	18	23	23
91	J91	15	16	17	19	23	23
92	J92	15	16	15	16	20	22
93	J93	14	16	13	15	21	24
94	J94	13	15	16	19	19	26
95	J95	16	17	15	16	21	25
96	J96	13	14	17	18	21	21
97	J97	13	14	17	18	19	25
98	J98	14	16	17	17	21	22
99	J99	16	18	16	18	23	25
100	J100	13	17	15	18	21	23
101	J101	16	17	15	16		
102	J102	13	16	14	19	22	24
103	J103	13	17	14	16	19	24
104	J104	14	15	13	15	22	27
105	J105	13	15	13	17	20	24
106	J106	15	18	13	14	21	24
107	J107	16	17	14	15	22	24
108	J108	15	16	13	15	20	24
109	J109	12	13	15	16	21	24
110	J110	13	16	14	16	20	23
111	J111	13	14	13	14	21	21
112	J112	15	16	16	16	23	24
113	J113	16	16	16	17	22	25
114	J114	15	17	14	15	21	24

Population: Punjabi

S. No's	Sample	D3S1358		vWA		FGA	
115	J115	14	16	13	16	20	24
116	J116	14	17	15	20	21	25
117	J117	14	16	14	17	23	25
118	J118	13	15	15	16	22	24
119	J119	14	16	14	15	23	25
120	J120	12	17	14	20	21	22
121	J121	14	18	15	15	23	26
122	J123	15	16	14	15	19	21
123	J124	13	16	17	18	19	22
124	J125	14	14	17	20	21	24
125	J126	15	17	17	18	22	23
126	J127	15	17	18	19	25	25
127	J128	15	17	15	15	19	22
128	J129	15	16	16	16	19	22
129	J130	14	15	14	20	21	22
130	J131	17	17	14	15	22	22
131	J132	13	17	16	18	21	21
132	J133	15	17	14	15	19	25
133	J134	14	15	16	18	22	23
134	J135	14	15	16	18	20	20
135	J136	15	16	17	19	20	24
136	J137	15	16	16	18	22	23
137	J138	13	15	15	17	19	23
138	J139	15	17	17	19	22	22
139	J140	16	17	15	18	20	23
140	J141	15	15	14	15	25	26
141	J142	14	15	17	17	22	25
142	J143	15	16	16	18	21	24
143	J144	13	16	16	19	22	23
144	J145	14	16	17	18	21	22
145	J146	16	17	16	19	21	24
146	J147	15	16	16	18	22	22
147	J148	15	15	17	19	21	23
148	J149	15	16	17	17	22	25
149	J150	13	16	14	19	20	21
150	J151	16	17	14	19	25	25
151	J152	15	16	16	18	23	24
152	J153	14	18	17	20	21	22
153	J154	17	17	15	17	20	25
154	J155	13	15	16	17	23	26
155	J156	15	16	16	19	23	24
156	J157	15	15	16	16	23	24
157	J158	14	15	16	19	22	25
158	J159	14	15	16	17	23	24
159	J160	15	16	17	20	20	20
160	J161	15	15	14	17	22	23
161	J162	15	17	16	19	21	22
162	J163	14	17	17	17	21	23
163	J164	15	15	14	19	21	23
164	J165	15	17	19	19	21	22
165	J166	15	16	14	17	21	24
166	J167	15	17	14	17	23	24
167	J168	16	17	14	17	23	25
168	J169	15	17	16	17	22	24
169	J170	15	16	15	17	22	24
170	J171	15	15	15	17	20	26

Population: Punjabi

S. No's	Sample	D3S1358		vWA		FGA	
172	J173	16	18	16	18	21	22
173	J174	16	17	12	12	19	19
174	J175	14	16	14	18	25	26
175	J176	15	16	16	18	22	22
176	J177	15	17	14	16	23	23
177	J178	14	15	14	17	19	26
178	J179	13	15	16	17	20	23
179	J180	13	16	18	19	21	23
180	J181	16	16	17	19	22	25
181	J182	14	16	18	18	19	19
182	J183	14	16	17	17	19	25
183	J184	15	14	18	19	20	25
184	J185	15	15	15	18	19	25
185	J186	15	16	16	17	21	22
186	J187	15	15	17	18	23	23
187	J188	15	16	15	17	19	27
188	J189	14	15	16	17	20	23
189	J190	15	15	18	20	22	23
190	J191	14	14	15	16	19	23
191	J192	13	13	15	16	20	21
192	J193	16	18	19	20	19	26
193	J194	14	14	19	21	20	24
194	J195	13	15	14	14	23	23
195	J196	16	19	12	16	21	21
196	J197	16	18	13	13	21	26
197	J198	15	18	12	16	22	26
198	J199	16	17	13	17	22	23
199	J200	15	19	12	17	22	25
200	J201	15	15	13	15	19	20

Population: Pushtoon

S. No's	Sample	D3S1358		vWA		FGA	
1	P1	14	16	17	19	22	24
2	P2	16	16	15	16	21	23
3	P3	14	17	18	19	22	23
4	P4	17	18	18	19	22	23
5	P5	14	15	17	19	22	23
6	P6	17	18	17	20	19	23
7	P7	16	16	14	17	21	26
8	P8	16	16	16	20	21	23
9	P9	13	16	15	15	22	23
10	P10	15	16	14	17	21	23
11	P11	16	16	15	16	24	24
12	P12	15	17	13	16	26	27
13	P13	16	17	14	18	19	21
14	P14	16	18	16	17	23	24
15	P15	16	16	14	16	20	24
16	P16	15	16	17	17	21	23
17	P17	14	17	15	19	23	23
18	P18	16	16	14	18	21	21
19	P19	17	18	18	19	23	24
20	P20	15	17	16	17	23	24
21	P21	16	16	17	17	23	27
22	P22	16	17	16	17	23	24
23	P23	16	18	16	18	23	23
24	P24	16	17	17	18	23.2	28
25	P25	15	17	17	18	23.2	27
26	P26	15	16	16	18	20	21
27	P27	15	15	17	18	20	26
28	P28	15	17	17	18	21	23

Population: Pushtoon

S. No's	Sample	D3S1358		vWA		FGA	
31	P31	16	17	15	17	22	24
32	P32	15	16	16	17	22.2	24
33	P33	15	18	17	18	20	22
34	P34	15	16	15	17	22.2	24
35	P35	15	18	17	18	20	22
36	P36	15	16	16	17	22	22
37	P37	16	17	14	16	21	26
38	P38	15	16	14	17	21	22
39	P39	16	17	14	15	19	20.2
40	P40	14	17	14	16	21	26
41	P41	15	16	14	17	20	23
42	P42	16	17	14	15	22	23
43	P43	16	16	14	16	21	23
44	P44	16	17	17	20	21	23
45	P45	14	17	15	19	23	23.2
46	P46	16	17	15	17	23.2	28
47	P47	14	16	15	17	22	23.2
48	P48	16	16	16	18	18	23
49	P49	15	16	16	17	21	26
50	P50	15	16	14	17	20	22
51	P51	15	16	13	14	21	24
52	P52	15	16	15	17	21	23
53	P53	15	17	17	18	22	23
54	P54	14	17	14	15	22	23
55	P55	17	17	18	18	22	22
56	P56	17	18	16	17	22	27
57	P57	15	16	15	18	21	27
58	P58	12	16	12	16	23	26
59	P59	15	16	16	17	23	23
60	P60	14	14	16	18	23	26
61	P61	15	16	15	17	21	23
62	P62	15	17	17	19	20	24
63	P63	17	18	18	18	24	26
64	P64	16	17	15	19	22	27
65	P65	15	16	16	16	21	24
66	P66	16	17	14	20	22	23.2
67	P67	14	15	15	16	21	27
68	P68	16	17	17	17	24	24
69	P69	17	18	16	17	20	23
70	P70	14	15	17	18	26	26
71	P71	14	16	17	18	22	26
72	P72	14	16	14	19	21	22
73	P73	14	16	15	18	22	23
74	P74	17	19	17	18	21	23
75	P75	14	16	16	16	26	27
76	P76	17	18	14	16	22	23
77	P77	17	18	14	14	21.2	23
78	P78	14	17	15	17	20	20
79	P79	18	18	14	15	19	21
80	P80	15	15	14	16	23	27
81	P81	16	18	14	16	22	25
82	P82	14	15	14	17	22	24
83	P83	15	16	14	17	22	23
84	P84	14	16	16	18	21	22.2
85	P85	14	15	16	19	23	24
86	P86	16	17	16	16	24	26
87	P87	16	16	16	18	21	26
88	P88	16	18	17	17	20	20
89	P89	16	17	16	18	23	24
90	P90	16	18	16	17	19	22

Population: Pushtoon

S. No's	Sample	D3S1358		vWA		FGA	
91	P91	16	17	16	17	20	23
92	P92	17	17	17	19	21	22
93	P93	14	17	15	16	20	22
94	P94	16	16	18	19	23	23
95	P95	14	18	17	17	21	24
96	P96	15	17	14	18	23	23
97	P97	17	17	14	16	21	24
98	P98	15	18	19	19	20	21
99	P99	15	17	17	17	20	21
100	P100	16	17	14	18	26	29
101	P102	15	17	15	16	24	26
102	P103	16	18	15	16	20	21
103	P104	15	17	14	17	21.2	27
104	P105	14	17	14	17	19	19
105	P106	14	16	14	16	22	22
106	P107	15	15	15	18	23	24
107	P108	17	17	18	18	22	26
108	P109	16	18	16	16	24	26
109	P110	14	15	16	18	23	23
110	P111	17	18	16	18	23	23
111	P112	16	18	14	17	22	26
112	P113	17	18	15	17	23	23
113	P114	16	17	16	17	24	26
114	P115	16	18	18	19	22	24
115	P116	13	16	16	17	21	22
116	P117	16	18	14	18	22	26
117	P118	16	18	15	18	18	21
118	P119	15	16	14	15	22	22
119	P120	15	17	14	17	20	20
120	P121	15	16	16	16	20	23
121	P122	15	16	14	15	19	23
122	P123	15	18	14	14	19	22
123	P124	15	18	17	18	22	23
124	P125	15	17	14	16	23	27
125	P126	18	18	17	18	19	23
126	P127	15	16	16	16	22	23
127	P128	14	18	17	18	19	19
128	P129	15	16	14	16	18	19
129	P130	14	16	16	18	22	23
130	P131	13	14	16	18	23	24
131	P132	15	16	16	16	21	26
132	P133	17	17	18	20	24	27
133	P134	16	16	16	17	24	24
134	P135	17	18	16	17	23	23
135	P136	17	18	16	17	24	24
136	P138	15	15	15	16	23	23
137	P139	17	17	16	19	24	24
138	P140	17	19	16	19	24	27
139	P141	15	16	17	19	21	24
140	P142	15	16	16	17	22	27
141	P143	16	18	16	17	24	27
142	P144	16	18	18	19	22	24
143	P145	15	18	17	17	23	24
144	P146	16	17	17	17	22	23
145	P147	15	16	15	17	20	22
146	P148	15	18	16	18	19	22
147	P149	15	18	16	19	23	27

Population: Pushtoon

S. No's	Sample	D3S1358		vWA		FGA	
150	P152	14	18	15	18	22	24
151	P153	15	17	14	19	21	27
152	P154	16	18	16	19	20	23
153	P155	17	18	14	14	21	23
154	P156	14	18	16	16	21	22
155	P157	15	18	14	16	23	24
156	P158	15	18	14	16	26	27
157	P159	14	16	17	19	21.2	22
158	P160	14	15	17	17	20	22
159	P161	15	17	16	16	22	24
160	P162	16	16	14	16	23	26
161	P163	15	18	14	14	20	20
162	P164	15	17	16	18	21.2	24
163	P165	15	16	14	17	23	26
164	P166	14	15	16	17	23	26
165	P167	15	17	16	17	22	24
166	P168	15	17	14	17	21	22
167	P169	15	15	14	18	21	21
168	P170	16	17	15	18	20	26
169	P171	15	18	15	18	22	23
170	P172	15	16	15	18	23	24

Population: Sindhi

S. No's	Sample	D3S1358		vWA		FGA	
1	S1	16	17	17	17	21	23
2	S2	15	16	16	19	25	25
3	S3	15	16	16	16	21	23
4	S4	15	15	16	17	22	23
5	S5	15	17	14	14	22.2	25
6	S6	15	18	15	17	21	25
7	S7	16	17	18	19	24	26
8	S8	16	17	17	19	20	21
9	S9	16	19	15	17	24	25
10	S10	16	17	14	14	20	20.2
11	S11	15	20	16	17	25	25
12	S12	15	15	14	16	23	26
13	S13	16	18	16	17	21	26
14	S14	15	18	15	20	21	26
15	S15	16	17	16	18	21	22
16	S16	16	17	15	17	22	24
17	S17	15	16	14	18	25	25
18	S19	15	15	17	17	23	25
19	S20	15	18	17	19	19	24
20	S22	17	18	16	16	19	24
21	S23	15	16	17	18	22	23
22	S24	14	15	15	15	23	23
23	S25	16	17	14	17	21	22
24	S26	16	16	14	17	24	24
25	S27	15	16	14	16	27	27
26	S28	16	16	16	18	22	25
27	S29	17	17	14	14	25	25
28	S30	15	17	16	18	22	22
29	S31	15	17	15	18	22	22
30	S32	16	17	17	17	22	22
31	S33	16	18	17	17	19	24
32	S34	16	17	15	17	22.2	24
33	S35	14	14	17	17	21	24
34	S36	11	17	17	18	19	25

Population: Sindhi

S. No's	Sample	D3S1358		vWA		FGA	
38	S40	18	20	13	16	23	24
39	S41	15	17	15	18	20	21
40	S42	16	16	14	20	24	25
41	S43	15	18	19	19	20	24
42	S44	16	16	17	18	23	25
43	S45	15	15	16	17	23	24
44	S46	15	16	16	19	24	24
45	S47	15	16	17	19	22	22
46	S48	14	16	18	18	24	25
47	S49	17	18	16	17	19	24
48	S50	17	18	15	16	19	24
49	S51	16	16	17	18	21	21
50	S52	16	16	17	17	21	24
51	S53	14	16	14	17	23	26
52	S54	16	16	17	17	22	25
53	S55	16	16	16	18	22	23
54	S56	16	17	14	17	20	21
55	S57	15	17	15	16	21	21
56	S58	16	17	17	19	21	25
57	S59	16	17	17	18	20	21
58	S60	15	15	16	17	22.2	25
59	S61	15	15	15	18	20	21
60	S62	16	17	17	18	19	24
61	S63	14	17	17	18	22	23
62	S64	18	19	16	16	22	24
63	S65	15	16	16	16	20	22
64	S66	14	16	15	16	19	21
65	S67	18	18	15	19	19	23
66	S68	16	17	13	15	26	26
67	S69	15	17	16	17	24	25
68	S70	15	15	15	16	23	24
69	S71	17	18	16	17	22	23
70	S72	16	16	16	17	23	24
71	S73	15	15	17	18	19	21
72	S74	17	18	16	18	20	26
73	S75	15	16	17	18	21	21
74	S76	16	18	16	18	22	25
75	S77	16	17	16	17	25	26
76	S78	15	15	14	18	22	24
77	S79	15	16	17	18	20	24
78	S80	15	19	16	18	20	21
79	S81	15	17	16	17	20	21
80	S82	15	16	17	19	21	25
81	S83	14	16	16	17	24	24
82	S84	15	16	16	17	20	23
83	S85	15	18	13	14	21	25
84	S86	15	15	15	17	22	27
85	S87	16	16	16	17	21	22
86	S88	15	18	17	18	21	22
87	S89	15	16	16	17	22	24
88	S90	16	17	17	18	24	24
89	S91	15	16	16	17	21	24
90	S92	15	16	14	17	21	24
91	S93	15	16	16	17	23	24
92	S94	16	17	14	17	24	25
93	S95	17	18	16	17	22	22
94	S96	16	18	18	18	24	25
95	S97	16	17	15	16	22	28
96	S98	15	16	18	18	22	23
97	S99	15	16	14	19	24	24
98	S100	16	17	17	17	26	26

Population: Baluchi

S. No's	Sample	D3S1358		vWA		FGA	
1	B1	14	15	16	16	20	24
2	B2	15	17	15	19	22	23
3	B3	15	17	16	17	22	23
4	B4	16	17	17	17	20	24
5	B5	14	14	18	19	24	25
6	B6	16	18	15	16	19	24
7	B7	16	17	17	17	19	19
8	B8	17	18	17	17	22.2	23
9	B9	15	17	15	17	22	25
10	B10	16	18	16	19	21	24
11	B11	15	17	14	15	22	22.2
12	B13	18	18	16	16	24	25
13	B14	16	19	16	16		
14	B15	16	19	17	19	22	24
15	B16	15	15	15	19	25	26
16	B17	14	16	17	17	23	23
17	B19	16	18	15	16	22	23
18	B20	15	16	15	18		
19	B22	16	18	16	17	20	21
20	B23	16	17	17	18		
21	B24	15	15	15	16	25	26
22	B25	15	15	16	17	23	23
23	B27	17	17	16	16	23	26
24	B28	16	17	17	18	26	32
25	B30	14	16	17	18	24	25
26	B31	16	18	16	18	21	24
27	B32	14	16	14	15	23	24
28	B33	14	17	17	18	23	27
29	B34	14	17	17	18	22	23
30	B35	15	18	17	18	23	23
31	B36	16	17	17	18	26	33
32	B37	18	18	16	18	21	25
33	B38	15	16	16	18	19	24
34	B39	15	17	17	18	24	25
35	B40	16	18	16	19	21	23

Population: Makrani

S. No's	Sample	D3S1358		vWA		FGA	
1	M1	16	17	15	17	19	21
2	M2	16	17	17	18	23	20
3	M3	16	17	17	17	25	25
4	M4	16	17	17	19	24	25
5	M7	15	18	17	17	22	22
6	M8	16	17	17	17	21	23
7	M10	15	17	15	17	20	21
8	M11	14	16	17	17	25	25
9	M13	14	16	14	17	23	25
10	M14	15	16	17	17	23	24
11	M15	14	16	17	17	24	24
12	M16	16	17	16	18	22	26
13	M17	16	17	14	17	23	24
14	M18	14	16	18	19	25	26
15	M19	15	15	15	16	21	24
16	M21	15	15	16	16	22	23
17	M22	17	17	15	17	21	22.2
18	M23	15	16	14	19	20	22
19	M24	16	17	15	17	25	25
20	M25	17	17	17	17	21	22.2
21	M26	17	17	15	17	19	24
22	M28	14	14	17	18	23	24
23	M29	15	17	18	18	20	24
24	M30	15	16	17	18	21	27
25	M31	15	17	17	17	25	27
26	M32	14	15	17	17	19	25
27	M33	16	17	15	17	25	32
28	M34	15	16	17	17	24	25
29	M35	15	17	17	17	25	25
30	M37	17	18	18	18	20	25
31	M38	16	18	14	17	22	23
32	M39	14	16	17	17	24	25
33	M40	14	15	14	17	19	25
34	M41	15	17	16	17	19	25
35	M42	16	18	16	18	24	25
36	M43	15	17	17	17	23	25
37	M44	16	17	13	17	22	24
38	M45	15	15	18	19	25	26
39	M46	15	19	16	18	22	22
40	M50	15	16	16	19	24	25
41	M52	16	17	16	18	19	24
42	M53	15	15	17	17	20	21
43	M54	16	16	14	16	21	26
44	M55	15	16	16	17	22	23
45	M56	16	18	16	16	20	26
46	M57	17	17	15	19	23	23
47	M58	15	17	15	18	22	24
48	M60	14	15	17	19	22	25

Population: Kalash

S. No's	Sample	D3S1358		vWA		FGA	
1	K1	15	17	15	17	20	20
2	K3	16	17	15	17	18	20
3	K5	16	16	15	19	18	20
4	K7	15	15	16	18	21	22
5	K9	15	17	15	16	20	23
6	K11	15	16	17	17	20	22
7	K13	15	16	15	16	20	22.2
8	K15	16	17	16	17	20	22
9	K17	16	16	17	19	22	22
10	K20	15	16	15	18	21	22
11	K21	15	16	16	19	20	22
12	K23	15	18	17	19	24	25
13	K25	15	15	17	17	21	23
14	K27	16	17	16	16	21	23
15	K29	16	17	15	16	22	22
16	K31	16	17	17	19	21	24
17	K33	15	15	18	19	22	25
18	K35	15	17	16	16	20	23
19	K37	15	17	16	17	22	22
20	K39	16	16	17	18	21	22
21	K41	15	16	15	17	18	20
22	K43	15	15	14	17	21	22
23	K45	15	17	17	17	24.2	25
24	K47	16	17	16	16	20	20
25	K49	16	16	15	19	22.2	23
26	K51	15	15	17	19	22	24
27	K55	16	17	15	15	20	21
28	K57	15	16	14	19	21	24.2
29	K59	15	16	18	19	22	24
30	K62	15	16	17	18	21	22
31	K63	16	17	16	17	20	21
32	K67	16	17	15	15	21	22
33	K69	16	17	15	18	21	25
34	K71	15	17	17	20	24.2	25
35	K73	16	16	15	15	21	23
36	K77	15	16	17	19	21	22
37	K79	15	17	15	18	20	22
38	K81	15	15	15	19	20	22
39	K85	17	17	18	19	24	24.2
40	K87	16	17	15	17	22	22
41	K89	14	15	14	21	23	24
42	K91	15	17	17	19	20	21
43	K93	16	16	18	19	20	20
44	K95	17	17	15	16	22	24.2
45	K97	15	15	18	19	18	21
46	K101	16	17	18	19	22	25
47	K102	15	15	15	17	20	20
48	K103	15	16	18	19	22	22
49	K104	15	16	17	19	21	22
50	K105	16	17	14	14	21	24.2
51	K106	15	16	15	18	20	24
52	K108	16	16	16	19	21	22
53	K109	15	17	16	17	22	22.2

Population: Kalash

S. No's	Sample	D3S1358		vWA		FGA	
54	K111	15	17	14	19	21	22
55	K112	15	15			18	20
56	K113	15	17	17	18	20	24
57	K114	16	17	18	19	20	25
58	K116	16	16	18	18	22	24.2
59	K117	14	17	19	20	21	23
60	K118	15	17	15	18	21	25
61	K119	15	15	15	16	20	22
62	K120	15	15	14	15	21	21
63	K121	15	18	14	15	20	21
64	K122	16	16	17	18	22	22
65	K123	15	15	15	16		
66	K125	14	17	16	20	18	24
67	K126	15	17	16	18	20	24.2
68	K127	16	17	15	17	22	24.2
69	K128	16	17	17	18	22	23
70	K129	15	17	16	16	22	24.2
71	K130	16	16	15	15	18	24.2
72	K131	15	17	15	18	19	23
73	K133	15	17	16	19	21	22
74	K134	15	17	17	17	20	21
75	K135	16	17	15	18	22	22
76	K136	15	18	15	17	21	22.2
77	K137	16	16	15	15	22	22
78	K138	16	17	14	18	22	25
79	K139	16	18	17	17	22	24.2
80	K141	17	17	13	18	21	22
81	K142	16	18	15	16	22	24.2
82	K143	16	17	13	17	22	24.2
83	K144	15	15	18	19	18	22
84	K145	16	18	15	16	22	24.2
85	K146	17	18	14	18	18	23
86	K147	15	16	15	16	21	25
87	K148	16	17	13	16	18	22
88	K149	16	16	16	19	22	24.2
89	K150	15	16	15	19	20	24
90	K151	15	16	15	17	20	23
91	K152	15	16	15	18	20	22
92	K153	16	17	15	18	23	24.2
93	K154	16	16	16	17	22	22
94	K155	16	16	18	19	20	24.2
95	K156	17	18	15	15	21	22
96	K157	16	18	16	16	20	23
97	K158	15	17	18	19	22	24

APPENDIX 2

Appendix 2

Formulae Used for the Calculation of Forensic Parameters for the Autosomal STR Loci

Observed Heterozygosity

Definition: Observed heterozygosity is the frequency of the heterozygotes in a population.

Formula:

$$\mathbf{h} = \frac{\mathbf{n_h}}{\mathbf{n}}$$

where, $\mathbf{h}$ is the heterozygosity, $\mathbf{n_h}$ is the number of heterozygotes and $\mathbf{n}$ is the total number of individuals profiled.

Example:

If 100 persons were genotyped and 76 were heterozygotes, the frequency is 0.76.

Expected Heterozygosity:

Definition: This is the heterozygosity we expect in a particular population sample and depends on the frequency of the alleles as well as the number of the alleles.

Formula:

$$\mathbf{He = (1-\Sigma\ pi^2)(\ n/n-1)}$$

Where pi is the allele frequency of allele i and n is the number of the alleles.

Example:

In a population (Brosho), the allele frequencies at locus D3S1358 were 0.065, 0.407, 0.25, 0.204, 0.056 & 0.019 for alleles 14 to 19.

These frequencies are squared individually and then added up.
The result is .277. The number of alleles was 108
Putting these values in the formula we get the result which is 0.728

Ref: Nei, M. & Roychoudry, A. (1974). Sampling variances of heterozygosity and genetic distance. Genetics. 76: 379-390

Paternity Index

Definition: The paternity index reflects how many times likely it is, that the person being tested is the biological father than a randomly selected man. Generally a PI of less than one is indicative of non-relatedness.

Formula:

$$\mathbf{PI}_{\text{typical}} = \frac{1}{2\mathbf{H}}$$

Example:

The formula gives the relationship of the homozygosity and the paternity index Thus if homozygosity if 0.2 then the paternity index is 2.5 (1/4).

The paternity index of several loci is the product of the value for the individual loci and be represented by:

$$\mathbf{PI}_{\text{comb}} = \prod_{i=1}^{n} (\mathbf{PI}_i)$$

Ref: Brenner, C. & Morris, J. (1990). Paternity index calculations in single locus hypervariable DNA probes:validation and other studies. In Proceedings of the International Symposium on human identification. Promega Corporation: 21-53.

Match Probability

Definition: It is the probability that the two randomly selected individuals will have identical genotype.

Formula:

$$\mathbf{pM} = \sum_{k=1}^{m} \mathbf{p}_k^2$$

where, $\mathbf{pM}$ is the Match probability, p_k represents the frequency of each distinct genotype, m is the number of the distinctive genotypes

The combined probability of match over several loci, is the product of the value for all the loci.

Example:

The, frequencies of blood systems, A_1, A_2, B, A_1B, A_2B and O were determined in the British people as,0.34, 0.08, 0.09, 0.024, 0.006 and 0.46 respectively*.

If the individual frequencies are squared and then added up.

$$\sum_{k=1}^{6} p_k^2 \cong 0.342$$

Putting this value in the formula the probability of match is 0.34 or 34%.

The probabilty of match of several independent systems is the product of the match probability of individual system**.

Ref: *, ** Jones, D. A. (1972). Blood samples: Probability of discrimination. J Forensic Sci Soc. 12: 355-359.

Power of Discrimination

Definition: It is the probability that two randomly selected individuals will have different genoytpes. This is the reciprocal of the probability of match.

Formula:

$$P_d = 1 - pM$$

For several loci the formula is: $P_{dcomb} = 1 - \prod_{i=1}^{n} (1 - P_{di})$

Example:

1. If P_M is 0.34, then the ability of the system to discriminate is 0.66

The Pd increases with the number of genotypes and is maximum if the genotypes occur at roughly similar frequencies*

2. The P_d for the loci for loci D3, vWA & FGA for the Punjabi population are 0.913, 0.95 & 0.963. Subtracting each value from 1 the P_M for each locus is 0.087, 0.05 & 0.037.

Multiplying these values together (0.00016), and then putting the value in the formula the P_{dcomb} is obtained, which is 0.9998.

Ref: Fisher, R. A (1951). Standard calculations for evaluating a blood group system. Heredity. 5: 95-102.

*Jones,D, A (1972). Blood samples: Probability of discrimination. J Forensic Sci Soc. 12: 355-359.

Power of Exclusion

Definition: It is the fraction of the individuals that is different from that of a randomly selected individual. It can also be defined as the power of a locus to exclude a person being the biological father. Thus the value differs in each case. The average for a locus is the power for a single locus.

Formula:

$$\mathbf{PE} = \mathbf{h}^2(1 - 2\mathbf{h}\mathbf{H}^2) *$$

where, h is the heterozygosity and H is the homozygosity at the locus

For a population heterozygosity is 0.8 , the homozygosity is thus 0.2.

Substituting these values in the formula the result is:

0.64*[1-(2x 0.8x 0.04)]= 0.64x 0.936 = 0.599

For several loci the formula is:

$$\mathbf{PE}_{\mathbf{comb}} = 1 - \prod_{l=1}^{L} (1 - \mathbf{PE}_l) **$$

where L is the number of the loci and $\mathbf{PE}_l$ is the exclusion probability for the l_{th} locus

Example:

For a population the probability of exclusion at loci D3, vWA & FGA were, 0.606, 0.735 & 0.702.

Subtracting each value from 1 and multiplying the values together a value of 0.031 is obtained. According to the formula this value is subtracted from 1 to get 0.9689, which is the combined probability of exclusion of the three loci in this population

Ref: * Brenner, C. & Morris, J. (1990). Paternity index calculations in single locus hypervariable DNA probes:validation and other studies. In Proceedings of the International Symposium on human identification. Promega Corporation: 21-53.

**Chakraborty, R. & Jin, L. (1993). Determination of relatedness between individuals using DNA fingerprinting. Human Biology. 65, 6: 875-895.

Polymorphic Information Content

Definition: This indicates the degree of polymorphism of a locus and depends on the frequencies of the alleles.

Formula:

$$\mathbf{PIC} = 1 - \sum_{i=1}^{n} \mathbf{pi}^2 - \left(\sum_{i=1}^{n} \mathbf{pi}^2\right)^2 + \sum_{i=1}^{n} \mathbf{pi}^4$$

where p*i,* is the frequency of each distinct allele, and n is the number of distinct alleles.

Example:

In population (Kalash) the allele frequencies at locus D3S1358 for alleles14 to 19 were 0.015, 0.33, 0.356, 0.253 & 0.046 receptively.

First each frequency is squared to obtain the pi^2 and then each is squared again to get pi^4. In this case $\Sigma\ pi^2 = .302$, $(\Sigma\ pi^2)^2 = .09$ and $\Sigma\ pi^4 = 0.032$

Putting these values in the formula above the PIC for this population is 0.64.

Ref: Botstein, D., White, R. L., Skolnick, M., and Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum. Gent. 32: 182-190.

Balding Formulae For The Estimation of Genotype Frequencies, incorporating the F_{ST} (θ)

For Homozygotes the formula is:

$$= \frac{[2\theta + (1-\theta)\mathbf{p}_i][3\theta + (1-\theta)\mathbf{p}_i]}{(1+\theta)(1+2\theta)}$$

where θ is the F_{ST} & Pi is the allele frequency

For Heterozygotes the formula is:

$$= \frac{2[\theta + (1-\theta)\mathbf{p}_i][\theta + (1-\theta)\mathbf{p}_j]}{(1+\theta)(1+2\theta)}$$

Ref: Balding, D. J. & Nichols, R. A. (1994). DNA profile match probability calculation: how to allow for population stratification, relatedness, database selection and single bands. Forensic Sci Int. 64:125-140.

APPENDIX 3

UNIVERSIDADE DE
SANTIAGO DE COMPOSTELA

Instituto de Medicina Legal
Biología Forense

San Francisco. s/n.
15705 Santiago de Compostela
Galicia-España
Tfno. 981 582 327 - 981 563 100. Ext. 12216
Fax 981 580 336

Y CHROMOSOME

CONTROL SAMPLES

Locus	Control sample-T	Control sample- J
DYS19	14	16
DYS389-I	10	10
DYS389-II	27	27
DYS390	23	24
DYS393	12	16
DYS391	10	10
DYS392	11	12
DYS385	13-19	14-15

References:

De Knijff P, Kayser M, Cagliá A, Corach D, Fretwell N, Gehrig C, Graziosi G, Heidorn F, Herrmann S, Herzog B, Hidding M, Honda K, Jobling M, Krawczak M, Leim K, Meuser S, Meyer E, Oesterreich W, Pandya A, Parson W, Penacino G, Perez-Lezaun A, Piccini A, Prinz M, Schmitt, Schneider PM, Szibor R, Teifel-Greding J, Weichhold G, Rower L (1997) Chromosome Y microsatellites: population genetics and evolutionary aspects. Int J Legal Med 110:134-149.

Pestoni C, Cal ML, Lareu MV, Rodríguez-Calvo MS, Carracedo A. (1998) Y chromosome STR haplotypes: genetic and sequencing data of the Galician population (NW Spain). Int. J. Legal Med., 112 15-21.

APPENDIX 4

Appendix 4

Yh1 Haplotypes for Baluchi Population

S.No's	DYS19	DYS389I	YS389II	DYS390	DYS391	DYS392	DYS393	No Observed
1	13	10	27	24	10	13	13	1
2	13	10	28	25	10	14	11	1
3	14	9	24	22	11	11	13	1
4	14	10	26	23	10	10	14	1
5	14	10	26	24	10	10	15	1
6	14	10	26	22	10	11	13	1
7	14	10	26	24	11	11	14	1
8	14	10	27	23	10	11	12	1
9	14	10	27	24	10	11	15	1
10	14	10	27	24	11	11	14	1
11	14	11	26	23	11	11	12	1
12	14	11	28	24	10	10	14	1
13	14	11	28	22	10	11	13	1
14	15	9	26	25	11	11	14	1
15	15	9	26	25	13	10	11	1
16	15	10	26	23	11	11	14	2
17	15	10	27	23	9	11	14	1
18	15	10	p	24	11	11	12	1
19	15	10	28	25	11	11	13	1
20	15	11	26	22	11	11	13	1
21	15	11	28	24	13	11	11	1
22	16	9	25	25	15	10	11	1
23	16	9	26	25	10	11	12	1
24	16	9	26	25	10	14	11	1
25	16	9	26	25	11	11	14	2
26	16	9	26	25	11	13	12	1
27	16	9	26	25	11	12	13	1
28	16	9	28	27	11	11	13	1
29	16	10	26	21	10	11	13	1
30	16	10	27	21	10	11	12	1
31	16	10	27	21	10	11	13	1
32	16	10	28	25	10	11	12	1
33	16	10	29	24	10	14	13	1
34	17	11	28	24	10	11	14	1

Yh1 Haplotypes for Brosho Population

S.No's	DYS19	DYS389I	YS389II	DYS390	DYS391	DYS392	DYS393	No Observed
1	14	10	27	24	10	10	14	1
2	15	10	26	23	10	10	14	1
3	15	11	26	24	10	10	14	1
4	14	10	27	24	10	11	12	2
5	15	10	26	23	10	11	12	1
6	15	11	26	23	10	11	12	1
7	16	10	24	25	10	11	12	1
8	14	10	26	24	10	11	13	1
9	16	10	26	22	10	11	13	1
10	16	10	26	24	10	11	13	1
11	15	11	26	22	10	11	14	1
12	15	10	27	24	10	11	15	1
13	16	10	26	24	10	12	12	1
14	17	11	27	22	10	12	12	1
15	14	10	25	22	10	13	13	2
16	14	10	27	24	10	13	13	1
17	14	10	27	24	10	14	12	2
18	14	9	25	23	10	14	12	1
19	15	10	26	22	10	14	12	3
20	15	11	27	22	10	14	12	1
21	14	10	26	24	10	14	12	1
22	15	10	25	22	10	14	12	1
23	15	9	26	22	10	14	12	1
24	15	10	27	22	10	14	12	1
25	15	10	26	23	10	14	12	1
26	16	10	26	22	10	14	12	3
27	16	10	25	25	10	14	12	1
28	17	9	27	22	10	14	12	1
29	17	11	27	22	10	14	12	4
30	17	11	27	23	10	14	12	1
31	17	10	27	24	10	14	12	1
32	17	11	28	24	10	14	12	1
33	15	10	26	23	10	15	12	2
34	16	10	26	22	10	15	12	1
35	15	11	26	24	11	10	13	1
36	16	10	26	23	11	10	13	1
37	14	9	26	22	11	11	12	1
38	14	10	25	22	11	11	12	1
39	15	12	28	22	11	11	12	1
40	15	11	27	24	11	11	12	1
41	17	10	25	22	11	11	12	1
42	15	11	27	22	11	11	13	1
43	14	9	25	23	11	11	13	1
44	17	11	27	22	11	11	13	1
45	15	11	26	22	11	12	13	1
46	15	11	27	22	11	12	13	1
47	15	11	26	22	12	11	12	2
48	16	10	25	25	12	11	12	1
49	16	10	26	25	11	12	14	1

Yh1 Haplotypes for Makrani Population

S.No's	DYS19	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	No Observed
1	14	9	24	22	10	14	11	1
2	14	9	25	22	10	14	11	1
3	14	9	25	25	11	11	12	1
4	14	9	25	25	11	11	13	1
5	14	9	26	22	11	11	14	1
6	14	10	26	22	10	11	14	1
7	14	10	26	23	11	11	14	1
8	14	10	26	24	11	11	14	1
9	14	10	27	23	10	12	15	1
10	14	11	27	22	11	11	14	1
11	15	9	25	23	10	10	14	1
12	15	9	25	23	11	11	13	1
13	15	9	26	25	10	10	14	1
14	15	9	26	25	10	13	13	1
15	15	10	26	22	10	11	13	1
16	15	10	26	23	11	11	13	1
17	15	10	26	23	11	11	14	1
18	15	10	27	21!	10	11	12	1
19	15	10	27	21!	10	11	13	1
20	15	10	27	23	11	11	13	1
21	15	10	28	23	11	12	12	1
22	15	11	26	24	10	14	13	1
23	15	11	28	25	10	11	13	1
24	15	11	30	25	10	11	12	1
25	16	9	26	25	10	11	13	1
26	16	9	26	25	10	11	14	3
27	16	9	26	25	10	14	11	1
28	16	9	26	25	10	14	13	1
29	16	9	26	25	11	11	13	1
30	16	9	26	25	11	11	14	8
31	16	10	26	22	11	11	12	1
32	16	10	27	21!	10	10	15	1
33	16	10	27	21!	10	11	14	1
34	16	10	27	25	11	11	14	1
35	16	10	28	26	9	13	12	1
36	16	10	28	25	10	11	11	1
37	16	10	28	25	10	11	12	1
38	16	10	28	25	10	11	13	1
39	16	10	28	25	11	11	13	1
40	16	10	28	25	11	11	14	1
41	16	11	28	24	11	11	14	1
42	17	10	25	24	11	11	12	1
43	17	10	25	24	10	11	14	1
44	17	10	27	25	10	11	13	1
45	17	10	28	26	11	11	13	1

Yh1 Haplotypes for Kalash Population

S.No's	DYS19	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	No Observed
1	14	9	26	23	10	12	13	1
2	14	10	26	23	10	11	12	2
3	14	10	27	23	10	11	12	1
4	14	10	27	23	10	11	13	2
5	14	11	27	23	10	11	12	1
6	14	11	27	24	10	11	12	1
7	14	11	27	23	10	11	13	1
8	14	11	27	23	10	12	12	1
9	14	11	27	23	10	12	13	4
10	14	11	27	22	10	14	12	1
11	14	11	27	23	11	11	12	1
12	14	11	27	24	11	11	13	3
13	14	11	27	24	11	12	12	1
14	14	11	27	23	11	12	13	1
15	14	11	27	24	11	12	13	8
16	14	11	27	25	11	12	13	1
17	15	9	25	23	10	11	14	3
18	15	9	26	24	10	11	12	1
19	15	9	26	24	11	11	12	6
20	15	9	27	23	11	11	12	1
21	15	10	26	22	10	11	12	9
22	15	10	26	23	10	11	12	5
23	15	10	26	24	10	11	12	1
24	15	10	26	23	10	11	13	1
25	15	10	26	24	11	11	13	1
26	15	10	26	23	11	12	13	2
27	15	10	27	23	10	11	12	1
28	15	11	27	23	10	11	12	1
29	15	11	27	23	10	11	13	1
30	16	9	25	23	10	11	14	32
31	16	9	25	23	10	11	15	2
32	16	9	25	23	11	11	13	2
33	16	10	26	24	10	12	13	1
34	17	10	26	24	10	11	14	1
35	17	10	26	24	11	11	13	1
36	17	10	27	24	11	11	13	1
37	17	11	27	24	10	11	12	1

Yh1 Haplotypes for Pushtoon Population

S.No's	DYS19	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	o Observed
1	13	10	27	22	10	15	13	1
2	13	10	28	25	11	11	13	1
3	13	11	26	22	10	15	15	1
4	13	11	28	23	10	10	13	1
5	13	11	29	21	11	11	13	1
6	14	9	25	22	11	11	13	1
7	14	9	26	22	10	13	11	1
8	14	9	26	22	11	11	13	1
9	14	10	27	23	10	11	13	1
10	14	10	28	24	10	11	13	1
11	14	10	28	22	11	11	13	1
12	14	11	27	23	10	10	13	1
13	14	11	27	23	10	11	13	2
14	14	11	28	23	10	10	13	2
15	14	11	28	23	10	13	13	1
16	15	9	25	22	10	11	11	1
17	15	9	25	23	10	12	12	1
18	15	9	25	24	10	14	14	1
19	15	9	27	24	10	11	13	1
20	15	10	24	22	10	14	11	1
21	15	10	25	23	11	11	13	1
22	15	10	26	22	10	15	12	1
23	15	10	27	23	10	11	13	2
24	15	10	27	26	10	11	13	1
25	15	10	27	22	10	14	13	1
26	15	10	27	22	11	11	13	1
27	15	10	27	24	11	11	13	3
28	15	10	27	25	11	11	13	3
29	15	10	28	24	10	11	13	1
30	15	10	28	25	10	11	13	1
31	15	10	28	23	10	13	13	1
32	15	10	28	24	11	11	13	2
33	15	10	28	25	11	11	13	1
34	15	10	29	24	11	11	13	1
35	15	10	29	26	11	11	13	1
36	15	11	27	23	10	10	14	1
37	15	11	27	22	10	11	12	1
38	15	11	27	23	11	11	13	1
39	15	11	28	24	10	14	13	1
40	15	11	28	24	11	10	13	1
41	15	11	28	24	11	11	13	1
42	15	11	29	25	10	11	13	2
43	15	11	29	23	10	15	13	1
44	16	9	25	22	11	14	13	1
45	16	9	26	22	10	11	13	1

Yh1 Haplotypes for Pushtoon Population

S.No's	DYS19	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	No Observed
46	16	9	26	23	10	12	13	1
47	16	10	26	23	10	14	12	1
48	16	10	26	23	11	11	13	1
49	16	10	26	25	11	11	13	1
50	16	10	27	23	10	11	13	1
51	16	10	27	24	10	11	13	2
52	16	10	27	25	10	11	13	1
53	16	10	27	24	10	14	11	1
54	16	10	27	24	11	11	12	1
55	16	10	27	23	11	11	13	2
56	16	10	27	24	11	11	13	6
57	16	10	28	24	10	10	14	1
58	16	10	28	25	10	11	12	2
59	16	10	28	23	10	11	13	1
60	16	10	28	24	10	11	13	1
61	16	10	28	25	10	11	13	1
62	16	10	28	23	10	11	14	1
63	16	10	28	24	10	11	14	1
64	16	10	28	24	10	13	14	1
65	16	10	28	24	10	14	14	2
66	16	10	28	23	11	10	13	1
67	16	10	28	24	11	11	12	1
68	16	10	28	23	11	11	13	2
69	16	10	28	24	11	11	13	4
70	16	11	28	23	11	11	13	1
71	16	11	29	23	11	11	13	2
72	17	10	26	25	11	11	13	1
73	17	10	27	24	10	15	13	1
74	17	10	28	25	10	10	13	1
75	17	10	28	24	10	11	13	1

Yh1 Haplotypes for Punjabi Population

S.No's	DYS19	YS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	No Observed
1	13	10	26	23	10	10	13	1
2	13	10	26	22	10	11	12	1
3	13	10	26	22	10	11	12	1
4	13	10	26	23	10	16	13	1
5	13	10	26	22	11	11	13	1
6	13	11	27	22	10	10	14	1
7	13	11	27	24	11	11	14	1
8	14	9	24	22	10	11	12	1
9	14	9	24	22	10	13	11	1
10	14	9	25	22	10	14	11	2
11	14	9	25	22	10	14	14	1
12	14	9	25	22	11	11	13	1
13	14	9	26	23	10	11	12	1
14	14	9	26	22	11	12	13	1
15	14	10	24	23	11	11	13	1
16	14	10	26	23	11	11	13	1
17	14	10	26	22	11	14	12	1
18	14	10	27	24	10	11	13	3
19	14	11	26	23	10	11	13	1
20	14	11	27	24	10	10	14	1
21	14	11	27	24	10	11	13	1
22	14	11	27	23	10	14	11	1
23	14	11	27	23	11	11	13	1
24	14	11	28	24	10	12	12	1
25	14	11	28	23	11	11	13	1
26	15	8	27	24	11	11	13	1
27	15	9	25	24	10	11	13	1
28	15	9	25	24	11	11	13	1
29	15	9	25	24	11	14	12	1
30	15	9	26	24	10	11	13	1
31	15	9	26	24	10	14	14	1
32	15	9	26	24	11	11	13	1
33	15	9	27	24	10	13	13	1
34	15	9	27	22	11	11	13	1
35	15	10	26	22	10	11	12	1
36	15	10	26	22	10	11	13	1
37	15	10	26	24	10	11	13	2
38	15	10	26	25	10	11	13	1
39	15	10	26	22	10	14	11	1
40	15	10	26	22	10	14	13	1
41	15	10	26	24	10	14	15	1
42	15	10	26	22	11	14	12	1
43	15	10	27	23	10	11	13	1
44	15	10	27	25	10	11	13	1
45	15	10	27	22	10	11	14	1
46	15	10	27	25	10	11	14	1
47	15	10	27	22	11	11	13	2
48	15	10	27	24	11	11	13	2

Yh1 Haplotypes for Punjabi Population

S.No's	DYS19	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	No Observed
49	15	10	27	25	11	11	13	2
50	15	10	27	24	11	14	12	1
51	15	10	28	24	10	11	12	1
52	15	10	28	25	10	11	12	1
53	15	10	28	25	10	11	13	3
54	15	10	28	25	10	15	13	1
55	15	10	28	25	11	11	13	2
56	15	11	27	22	10	11	12	1
57	15	11	27	22	10	11	13	1
58	15	11	27	25	10	12	13	1
59	15	11	27	24	10	13	13	1
60	15	11	27	22	10	14	12	1
61	15	11	27	24	10	14	14	1
62	15	11	27	23	11	11	13	1
63	15	11	28	22	10	11	13	1
64	15	11	28	25	10	11	13	1
65	15	11	28	26	10	11	13	1
66	15	11	28	23	10	14	15	1
67	15	11	28	25	11	11	13	1
68	15	11	28	26	11	11	13	1
69	15	11	28	24	11	11	14	1
70	15	11	29	25	10	11	12	1
71	15	11	29	25	10	11	13	4
72	15	11	29	25	11	11	13	1
73	16	9	25	23	10	11	12	1
74	16	9	26	23	10	12	13	1
75	16	10	26	23	10	14	12	1
76	16	10	26	21	11	14	14	1
77	16	10	27	25	10	11	12	1
78	16	10	27	25	10	11	13	1
79	16	10	27	25	10	14	11	1
80	16	10	27	25	10	15	14	1
81	16	10	27	25	11	11	13	2
82	16	10	27	23	11	11	14	1
83	16	10	28	26	10	11	12	1
84	16	10	28	25	10	14	11	1
85	16	10	28	24	11	11	13	2
86	16	10	28	25	11	14	12	1
87	16	10	29	23	10	14	12	1
88	16	10	29	25	11	11	13	1
89	16	11	26	23	10	14	12	1
90	16	11	26	23	11	11	13	1
91	16	11	27	23	10	11	13	1
92	16	11	27	23	10	13	12	1
93	16	11	29	25	10	11	13	1
94	16	11	29	24	10	12	12	1
95	17	10	27	25	10	11	13	1
96	17	10	27	25	11	11	13	1
97	17	10	28	25	11	11	13	2
98	17	11	27	24	10	12	13	1
99	17	11	29	24	10	10	14	1

Yh1 Haplotypes for Sindhi Population

S.No's	DYS19	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	No Observed
1	15	9	25	24	9	11	12	1
2	15	11	27	24	9	11	13	2
3	15	9	27	22	9	11	13	1
4	14	9	26	23	9	11	14	1
5	12	11	27	25	10	10	13	1
6	15	11	29	25	10	10	13	1
7	14	10	26	23	10	10	14	1
8	14	11	26	23	10	10	14	1
9	14	10	25	24	10	10	14	1
10	15	11	28	23	10	10	14	1
11	15	10	27	25	10	10	14	1
12	15	11	27	23	10	10	15	2
13	15	11	30	25	10	10	15	1
14	14	10	27	23	10	11	12	1
15	14	11	28	23	10	11	12	1
16	14	9	25	24	10	11	12	1
17	15	10	26	23	10	11	12	1
18	15	11	28	22	10	11	12	1
19	15	10	27	25	10	11	12	1
20	14	10	26	23	10	11	13	1
21	14	11	27	23	10	11	13	1
22	15	9	27	22	10	11	13	2
23	15	11	29	25	10	11	13	2
24	15	11	31	25	10	11	13	1
25	15	11	29	26	10	11	13	1
26	16	10	28	25	10	11	13	1
27	16	10	29	25	10	11	13	1
28	16	11	29	25	10	11	13	1
29	14	10	27	23	10	11	14	1
30	15	10	26	22	10	11	14	1
31	15	10	26	23	10	11	14	2
32	15	9	27	25	10	11	14	1
33	15	11	29	25	10	11	14	1
34	16	11	29	25	10	11	14	1
35	14	9	26	22	10	12	13	1
36	14	9	27	22	10	12	12	2
37	15	9	24	22	10	12	12	2
38	17	10	28	24	10	12	13	1
39	14	11	28	27	10	13	13	1
40	14	9	25	22	10	14	11	1
41	15	11	29	26	10	14	11	1
42	15	10	26	22	10	14	12	1
43	15	11	27	22	10	14	12	1
44	15	10	26	23	10	15	13	1
45	16	10	28	25	11	10	13	1
46	14	11	27	23	11	11	12	1

Yh1 Haplotypes for Sindhi Population

S.No's	DYS19	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393	No Observed
50	15	10	26	23	11	11	13	1
51	15	11	28	23	11	11	13	1
52	15	10	27	24	11	11	13	1
53	15	10	28	24	11	11	13	1
54	15	10	25	25	11	11	13	1
55	15	10	27	25	11	11	13	1
56	15	10	27	26	11	11	13	1
57	15	11	29	26	11	11	13	1
58	16	10	28	23	11	11	13	1
59	16	9	26	25	11	11	13	2
60	16	10	27	25	11	11	13	6
61	16	10	28	25	11	11	13	2
62	17	11	27	25	11	11	13	2
63	15	10	28	25	11	11	14	3
64	15	11	29	25	11	11	14	1
65	16	9	26	25	11	11	14	2
66	16	11	26	25	11	11	14	2
67	15	10	26	22	11	11	15	2
68	14	9	25	22	11	14	11	1
69	15	10	26	22	11	14	12	1
70	15	11	28	22	11	14	12	1
71	15	10	28	25	12	11	13	1
72	16	10	28	25	12	11	13	1

APPENDIX 5

Appendix 5

List Of Shared Haplotypes

Punjabi/Pushtoon/Sindhi

No	DYS19	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393
1	15	10	27	25	11	11	13

Punjabi/Pushtoon/Baluchi

1	15	10	28	25	11	11	13

Punjabi/Pushtoon

1	14	9	25	22	11	11	13
2	15	10	27	22	11	11	13
3	15	10	27	23	10	11	13
4	15	10	27	24	11	11	13
5	15	10	27	25	11	11	13
6	15	10	28	25	10	11	13
7	15	10	28	25	11	11	13
8	15	11	27	22	10	11	12
9	15	11	27	23	10	11	13
10	16	9	26	23	10	12	13
11	16	10	26	23	10	14	12
12	16	10	27	25	10	11	13
13	16	10	28	24	11	11	13

Punjabi/Sindhi

1	14	11	28	23	11	11	12
2	15	10	27	25	11	11	13
3	15	11	27	22	10	14	12
4	15	11	29	25	10	11	13
5	16	11	29	25	10	11	13

Sindhi/Baluchi/Makrani

1	16	9	26	25	11	11	14

Sindhi/Baluchi

1	14	10	26	23	10	10	14
2	14	10	27	23	10	11	12
4	16	9	26	25	11	11	14

Sindhi/Makrani

No	DYS19	DYS389I	DYS389II	DYS390	DYS391	DYS392	DYS393
1	14	9	25	22	10	14	11
2	15	10	26	23	11	11	14
3	16	9	26	25	10	11	12
4	16	9	26	25	11	11	14
5	16	10	28	25	11	11	13

Sindhi/Pushtoon

1	15	10	28	24	11	11	13
2	16	10	28	25	10	11	13
3	14	11	27	23	10	11	13

Baluchi/Pushtoon

1	16	10	28	25	10	11	12

Kalash & Other Populations*

1	14	10	27	23	10	11	12
2	14	10	27	23	10	11	13
3	14	11	27	23	10	11	13
4	14	11	27	23	11	11	12
5	15	10	26	22	10	11	12
6	15	10	26	23	10	11	12

* Haplotype No. 1 is shared by Sindhi and Baluchi populations
Haplotype No. 2 is shared by the Pushtoon population
Haplotype No. 3 is shared by Pushtoon and Sindhi populations
Haplotype No. 4 & 6 are shared by Sindhi population
Haplotype No. 5 is shared by Punjabi population

REFERENCES

REFERENCES

AABB Report Summary for 1999, pp: 1-12, available at http://www.aabb.org

Aitken, C. G. G. & Stoney, D. A. (1991). Populations and samples. pp: 53-54. In, The use of statistics in forensic science. London: Ellis Horwood Ltd.

Alford, R. L., Hammond, H. A., Coto, I. & Caskey, C. T. (1994). Rapid and efficient resolution of parentage by amplification of short tandem repeats. Am J Hum Genet. 55: 190-195.

Amar, A., Brautbar, C., Motro, U., Fisher, T., Bonne-Tamir, B. Israel, S. (1999). Genetic variation of three tetrameric tandem repeats in four distinct Israeli ethnic groups. J Forensic Sci. 44: 983-6.

Amos, W., Sawcer, S. J., Feakes, R. W. & Rubinsztein, D. C. (1996). Microsatellites show mutational bias and heterozygote instability. Nat Genet. 13: 390-1.

Anker, R., Steinbrueck, T. 7 Donis-keller (1992). Tetranucleotide repeat polymorphism at the human thyroid peroxidase (hTPOX) locus. Hum Mol Genet. 1: 137.

Anslinger, K., Keil, W., Weichhold, G., & Eisenmenger, W. (2000). Y-chromosomal STR haplotypes in a population sample from Bavaria. Int J Legal Med. 113: 189-92.

Austin, M. A., Ordovas, J. M., Eckfeldt, J. H., Tracy, R., Boerwinkle, Lalouel, J. M., & Printz, M. (1996). Guidelines of the National Heart, Lung and Blood Institute Working Group on Blood drawing, processing and storage for genetic studies. Am J Epidem. 144: 437-41.

Ayub, Q., Mohyuddin, A., Qamar, R., Mazhar, K., Zerjal, T., Mehdi, S. Q. & Tyler-Smith, C. (2000). Identification and characterisation of novel human Y-chromosomal microsatellites from sequence database information. Nucleic Acids Res. 28: 8-9.

Baillie, A. F. (1890). In, Kurrachee: Past, present and future. pp: 35. London Publishing Co. London, UK.

Balding, D. J. & Nichols, R. A. (1994). DNA profile match probability calculations: how to allow for population stratification, relatedness, database selection and single bands. Forensic Sci Int. 64: 125-140.

Bao, W., Zhu S, Pandya, A., Zerjal, T., Xu, J., Shu, Q., Du, R., Yang, H. & Tyler-Smith, C. (2000). MSY2: a slowly evolving minisatellite on the human Y chromosome which provides a useful polymorphic marker in Chinese populations. Gene. 244:29-33.

Bär, W., Brinkmann, B., Budowle, B., Carracedo, A., Gill, P., Lincoln, P., Mayr, W.R., & Olaisen, B. (1997). DNA recommendations: Further report of the DNA Commission of the ISFH regarding the use of short tandem repeat systems. Int J Legal Med. 110:175-176.

Barber, M. D., Mckeown, B. J. & Paekin, B. H. (1996). Structural variation in the alleles of a short tandem repeat system at the human alpha fibrinogen locus. Int J Legal Med. 108: 180-185.

Barber, M. D., Piercy, R. C., Andersen, J. F. & Parkin, B. H. (1995). Structural variation of novel alleles at the Hum vWA and Hum FES/FPS short tandem repeat loci. Int J Legal Med. 108: 31-35.

Barbujani, G., Magagni, A., Minch, E. & Cavalli-Sforza, L. L. (1997). An apportionment of human DNA diversity. Proc Natl Acad Sci USA. 94:4516-19.

Batzer, M. A., Kilroy, G. E., Richard, P. E., Shaikh, T. H., Desselle, T. D., Hoppens, C. L., Deininger, P. L. (1990). Structure and variability of recently inserted Alu family members. Nucl Acids Res.18: 6793-98.

Bell, G. I., Selby, M. J. & Rutter W. J. (1982). The highly polymorphic region near the human insulin gene is composed of simple tandemly repeating sequences. Nature. 295:31-5.

Bennett, P. (2000). Microsatellites. J Clin PathMol Path. 53:177-183.

Berger, H. (1985). A survey of Brushaski studies. J Central Asia. 1: 33-38.

Bianchi, N. O., Catanesi, C. I., Bailliet, G., Martinez-Marignac, V. L., Bravi, C. M., Vidal-Rioja, L. B., Herrera, R. J., & Lopez-Camelo, J. S. (1998). Characterization of ancestral & derived Y-chromosome haplotypes of New World native populations. Am J Hum Genet. 63:1862-71.

Boerwinkle,E., Xiang,W., Fourets, E., & Chan, L. (1989). Rapid typing of tandemly repeated hypervariable loci by polymerase chain reaction. Application to the apolipoprotein B3 hypervariable region. Proc Natl Acad Sci USA. 86: 212-16.

Bordie, J. G. (1981). An inquiry into the glotto-chronology of Sindhi phonology. pp: 270-277. In, Sind through the centuries. Hameeda Khuro (Ed). Karachi Oxford Press, Pakistan.

Botstein, D., White, R. L., Skolnick, M., & Davis, R. W. (1980). Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Gent. 32: 182-90.

Bowcock, A., Osborne-Lawrence, S., Barnes, R., Chakravarti, A., Washington, S. & Dunn, C. (1993). Microsatellite polymorphism linkage map of human chromosome 13q. s. Genomics. 15: 376-86.

Brenner, C. & Morris, J. (1990). Paternity index calculations in single locus hypervariable DNA probes: validation and other studies. pp: 21-53. In: Proceedings of the International Symposium on human identification. Promega Corporation.

Brinkmann, B., Klintschar, M., Neuhuber, F., Huhne, J. & Rolf, B. (1998). Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. Am J Hum Genet. 62: 1408-15.

Brinkmann, B., Sajantila, A., Goedde, H. W., Matsumoto, H., Nishi, K. & Wiegand, P. (1996). Population genetic comparisons among eight populations using allele frequency and sequence data from three microsatellite loci. Eur J Hum Genet. 4: 175-182.

Brinkmann, C., Forster, P., Schurenkamp, M., Horst, J., Rolf, B., & Brinkmann, B. (1999). Human Y-chromosomal STR haplotypes in a Kurdish population sample. Int J Legal Med. 112: 181-83.

Brookfield, J. (1992). The effect of population subdivision on estimates of the likelihood ratio in criminal cases using single locus probes. Heredity 69: 97-100.

Budowle, B. & Lander, E. S. (1994). DNA fingerprinting dispute laid to rest. Nature. 371: 735-38.

Budowle, B. Moretti, T. R., Baumstark, A. L., Defenbaugh, D. A., Keys, K. M. (1999). Population data on the thirteen CODIS core short tandem repeat loci in African Americans, U.S. Caucasians, Hispanics, Bahamians, Jamaicans, and Trinidadians. J Forensic Sci: 44: 1277-86.

Budowle, B., Chakraborty, R., Giusti, A., Eisenberg, A. & Allen, R. (1991). Analysis of the VNTR locus D1S80 by the PCR followed by high-resolution page. Am J Hum Genet. 48: 137-44.

Budowle, B., Lindsey, J. A., DeCou, J. A., Koons, B. W., Giusti, A. M., Comey, C. T. (1995). Validation and population studies of the loci LDLR, GYPA, HBGG, D7S8, and Gc (PM loci), and HLA-DQ alpha using a multiplex amplification and typing procedure J Forensic Sci. 40: 45-54.

Budowle, B., Nhari, L.T., Moretti, T.R., Kanoyangwa, S.B., Masuka, E., Defenbaugh, D.A.,Smerick, J.B. (1997). Zimbabwe black population data on the six short tandem repeat loci -CSF1PO, TPOX, THO1, D3S1358, VWA and FGA. Forensic Sci Int. 90: 215-21.

Buffone, G. J., & Darlington, G. J. (1985). Isolation of DNA from biological specimens without extraction with phenol. Clin Chem. 31:164-65.

Burrow, F. (1973). The proto - Indo Aryans. J R A S: 123-40.

Cabrero, C., Diez, A., Valverde, E., Carracedo, A., Alemany, J. (1995). Allele frequency distribution of four PCR-amplified loci in the Spanish population. Forensic Sci Int. 71:153-64.

Cacopardo, A. & Cacopardo, A. (1996). The other Kalasha in Southern Chitral, Part 3: Jingeret Koun and Kalasha origins. pp: 299-313. In, Proceedings of 2nd International Hindukush Cultural Conference. Oxford University Press. Karachi.

Capon, D. J., Chen, E. Y., Levinson, A. D., Seeburg, P. H. & Goeddel, D.V. (1983). Complete nucleotide sequences of the T24 human bladder carcinoma oncogene and its normal homologue. Nature. 302: 33-7.

Caroe, O. (1958). The White Huns. pp: 81-93. In, The Pathans 550 B. C–A D. 1957. Macmillan & Co. London.

Carvalho-Silva, D. R., Santos, F. R., Hutz, M. H., Salzano, F. M., & Pena, S. D. J. (1999). Divergent human Y-chromosome microsatellite evolution rates. J Mol Evol. 49: 204-14.

Casanova, M., Leroy, P., Boucekkine, C., Weissenbach, J., Bishop, C., Fellous, M.,Purrello, M., Fiori, G. & Siniscalco, M. A. (1985). Human Y-linked DNA polymorphism and its potential for estimating genetic and evolutionary distance. Science. 230: 1403-06.

Cavalli-Sforza, L. L., Menozzi, P. & Piazza, A. (1993). Demic expansions and human evolution. Science. 259: 639-46.

Cha, R. S. & Thilly, R. S. (1993). Specificity, efficiency and fidelity of PCR. PCR Meth & Appl. 3: S18-29.

Chakraborty R. & Kidd, K. K. (1991). The utility of DNA typing in forensic work. Science. 254: 1735-39.

Chakraborty, R. & Ferrel, E. R. (1982). Correlation of paternity index with probability of exclusion and efficiency criteria of genetic markers for paternity testing. For Sci Int. 19: 113-124.

Chakraborty, R. & Jin, L. (1993). Determination of relatedness between individuals using DNA fingerprinting. Hum Biol. 65: 875-95.

Chakraborty, R. & Stivers, D. N. (1998). Further response to Mueller & Thompson: Considerations on the tests of independence of alleles that are relevant for forensic applications. Correspondence. J Forensic Sci. 43: 448-49.

Chakraborty, R. (1985). Paternity testing with genetic markers are Y-linked genes more efficient than autosomal ones. Am J Med Genet. 21: 297-305.

Chakraborty, R., Kimmel, M., Stivers, D. N., Davison, L. J. & Deka, R. (1997). Relative mutation rates at di-, tri-, and tetranucleotide microsatellite loci. Proc Natl Acad Sci U S A. 94: 1041-46.

Chakraborty, R., Shaw, M., & Schull, W. J. (1974). Exclusion of paternity: The current state of the art. Am J Hum Genet. 26: 477-488.

Chakraborty, R., Srinivasan, M. R., & Daiger, S. P (1993). Evaluation of Standard error and confidence interval of estimated multilocus genotype probabilities, and their implications in DNA forensics. Am J Hum Gent. 52: 60-70.

Charlesworth B., Sniegowski P. & Stephan W. (1994). The evolutionary dynamics of repetitive DNA in eukaryotes. Nature. 371: 215-220.

Chattopahyaya, S. (1955). The Sakas in India. pp: 8 & 31. In, Santiniketan, Visva Bharti. 1st Edition, Delhi Press, India.

Chen, H., Lowther, W., Avramopoulos, D., & Antonarakis, S. E. (1994). Homologous loci DXYS156X and DXYS156Y conatin a polymorphic pentanucleotide repeat (TAAAA)n and map to human X and Y chromosomes. Human Mutation. 4: 208-211.

Ciulla, T. A., Sklar, R. M., Hauser, S. L. (1988). A simple method for DNA purification from peripheral blood. Anal Biochem. 174: 485-8.

Clayton, T. M., Whitaker, J. P., Sparkes, R., and Gill, P. (1998). Analysis and interpretation of mixed forensic stains using DNA STR profiler. Forensic Sci Int. 91: 55-70.

Cohen, J. E. (1990). DNA fingerprinting for forensic identification: potential effects on data interpretation of subpopulation heterogeneity and band number variability. Am J Hum Genet. 46: 358-68.

Cooke, H. J., Fantes, J. & Green, D. (1983). Structure and evolution of human Y chromosome DNA. Differentiation. 23: 48-55.

Cooke, H. J., Schmidtke, J. & Gosden, J. R. (1982) Characterisation of a human Y chromosome repeated sequence and related sequences in higher primates. Chromosoma. 87: 491-502.

Cooper, G., Amos, W., Hoffman, D. & Rubinsztein, D. C. (1996). Network analysis of human Y microsatellite haplotypes. Hum Mol Genet. 5: 1759-66.

Cooper, G., Burroughs, N. J., Rand, D. A., Rubinsztein, D. C. & Amos, W. (1999). Markov chain Monte Carlo analysis of human Y-chromosome microsatellites provides evidence of biased mutation. Proc Natl Acad Sci. USA. 96: 11916-21.

Cosso, S. & Reynolds, R. (1993). Validation of the AmpliFLP D1S80 PCR Amplication Kit for Forensic Casework Analysis According to TWGDAM Guidelines. J Forensic Sci. 40: 424-34.

Costani, L. (1981). The beginning of agriculture in the Kachi plains: The evidence of Mehrgarh. J South Asian Archaeology. 1: 29-60.

Court of Appeal Report (England). (1996). Part 3: 467-82.

Court of Appeal Report (England). (1997). Part 3: 369-89.

Crouse, C. & Schumm, J. W. (1995). Investigation of species specificity using nine PCR based human STR systems. J Foresnic Sci. 40: 952-56.

Dani, A. H. (1981) A glimpse into the early history of Sind. pp: 35-42. In, Sind through the centuries. Hameeda Khuro (Ed). Karachi Oxford Press, Pakistan.

Dauber, E. M., Glock, B., Scwartz, D. W. M., and Mayr, W. R. (2000). Mutational events at human micro and minisatellite loci : mutation rates and new STR alleles.

pp: 21–23. In Sensabaugh, G, F., Brinkmann B., Lincoln P. (Eds), Progress in Forensic Genetics 8. Excerpta Medica, Amsterdam.

de Knijff, P., Kayser, M., Caglia, A., Corach, D., Fretwell, N., Gehrig, C., Graziosi, G., Heidorn, F., Herrmann, S., Herzog, B., Hidding, M., Honda, K., Jobling, M., Krawczak, M., Leim, K., Meuser, S., Meyer, E., Oesterreich, W., Pandya, A., Parson, W., Penacino, G., Perez-Lezaun, A., Piccinini, A., Prinz, M., Schmitt, C., Schneider, P. M., Szibor, R., Teifel-Greding, J., Weichhold, G. M. & Roewer, L. (1997). Chromosome Y microsatellites: population genetic and evolutionary aspects. Int J Legal Med. 110: 134-40.

Debrauwere, H., Gendrel, C. G., Lechat, S. & Dutreix, M. (1997). Differences and similarities between various tandem repeat sequences:minisatellites and microsatellites. Biochimie. 79: 577-86.

Deka, R., Jin, L., Shriver, M. D., Yu, L. M., Saha, N., Barrantes, R., Chakraborty, R., & Ferrell, R. E. (1996). Dispersion of human Y chromosome haplotypes based on five microsatellites in global populations. Genome Res. 6: 1177-84.

Deka, R., Shriver, M. D., Yu, L. M., Ferrell, R. E. & Chakraborty, R. (1995). Intra- and inter-population diversity at short tandem repeat loci in diverse populations of the world. Electrophoresis. 16: 1659-64.

Denault, G. C., Jakimoto, H. H., Kwan, Q. Y.& Pallos, A. (1980). Detectability of selected genetic markers in dried blood on ageing. J Forensic Sci. 25: 479.

Di Rienzo, A., Peterson, A. C., Garza, J. C., Valdes, A. M., Slatkin, M. & Freimer, N. B. (1994). Mutational processes of simple-sequence repeat loci in human populations. Proc Natl Acad Sci U S A. 91: 3166-70.

Dorit, R. L., Akashi, H. & Gilbert, W. (1995). Absence of polymorphism at the ZFY locus on the human Y chromosome. Science. 268: 1183-5.

Edwards, A., Civitello, A., Hammond, H. A. & Caskey C. T. (1991). DNA Typing and Genetic Mappinng with Trimeric and Tetrameric Tandem Repeats. Am J Hum. Genet. 49: 746-56.

Edwards, A., Hammond, H. A., Jin, L., Caskey, T. & Chakarborty, R. (1992). Genetic Variation at five trimeric and tetrameric tandem repeat loci in four human population groups. Genomics. 12: 241-53.

Entrala, C., Lorente, M., Lorente, J. A., Alvarez, J. C., Moretti, T., Budowle, B. & Villanueva, E. (1998). Fluorescent multiplex analysis of nine STR loci: Spanish population data. For Sci Int. 98: 179-83.

Evett, I, W., Lambert, J. A., Buckleton, J. S., Weir, B. S. (1996). Statistical analysis of a large file of data from STR profiles of British Caucasians to support forensic casework. Int J Legal Med. 109: 173-77.

Evett, I. W & Weir, B. S (1998). Population genetics. In Interpreting DNA Evidence. pp. 79-131. Sinauer Associates, Inc. USA.

Evett, I. W., Gill, P. D., Lambert, J. A., Oldroyd, N., Frazier, R., Watson, S., Panchal, S., Conolly, A. & Kimpton, C. (1997). Statistical analysis of data for three British ethnic groups from a new STR multiplex. Int J Legal Med. 110: 5-9.

Excoffier, L., Smouse, P. E., Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics. 131: 479-91.

Felsenstein, J. (1989). 'PHYLIP' Phylogeny Inference Package (Version 3.2). Cladistics 5: 164-166. Distributed free from web site: http://evolution.genetics.washington.edu.

Forster, P., Kayser, M., Meyer, E., Roewer, L., Pfeiffer, H., Benkmann, H., & Brinkmann, B. (1998). Phylogenetic resolution of complex mutational features at Y-STR DYS390 in Aboriginal Australians and Papuans. Mol Biol Evol. 15: 1108-14.

Forster, P., Rohl, A., Lunnemann, P., Brinkmann, C., Zerjal, T., Tyler-Smith, C., and Brinkmann, B. (2000). A Short Tandem Repeat-Based Phylogeny for the Human Y Chromosome. Am J Hum Genet. 67: 182-96.

Fowler, J. C. S., Burgoyne, L. A., Scott, A. C., & Harding, H. W. J. (1988). Repetitive DNA and human genome variation. J Forensic Sci. 33: 111-26.

Frank, R. & Koster, H. (1979). DNA chain length markers and the influence of base composition on the electrophoretic mobility of oligonucleotides in polyacrilamide gels. Nucl Acids Res. 6: 2069-87.

Fregeau, C. J. & Fourney, R. M. (1993). DNA Typing with flourescently tagged short tandem repeats: A sensitive and accurate approach to human identification. Biotechniques. 15: 100-19.

Furedi, S., Angyal, M., Kozma, Z., Setalo, J., Wolle, R, J., Padar, Z. (1997). Semi-automatic DNA profiling in a Hungarian Romany population using the STR loci HumVWFA31, HumTH01, HumTPOX, and HumCSF1PO. Int J Legal Med. 110: 184-87.

Furedi, S., Woller, J., Padar, Z., & Angyal, M. (1999). Y-STR haplotyping in two Hungarian populations. Int J Legal Med. 113: 38-42.

Gaensslen, R. E., Lee, H. C., Pagliaro, E. M., & Bremser, J. K. (1985). Evaluation of antisera for bloodstain grouping: I. ABH, MNS, Rh. J Forensic Sci. 30: 632.

Gaensslen, R. E., Lee, H. C., Pagliaro, E. M., Bremser, J. K. & Carroll. R. C. (1985). Evaluation of antisera for bloodstain grouping: II. Ss, Kell, Duffy, Kidd and Gm/Km. J Forensic Sci. 30: 655.

Gehrig, C., Hochmeister, M. & Budowle, B. (2000). Swiss allele frequencies and haplotypes of 7 Y-specific STRs. J Forensic Sci. 45: 436-39.

Gill, P. & Evett, I. (1995). Population genetics of short tandem repeat (STR) loci. Genetica. 96: 69-87.

Gill, P. & Werret, D. J. (1987). Exclusion of a man charged with murder by DNA fingerprinting. Forensic Sci Int. 35: 145-48.

Gill, P., Brinkmann, B., d'Aloja, E., Andersen, J., Bar, W., Carracedo, A., Dupuy, B., Eriksen, B., Jangblad, M., Johnsson, V., Kloosterman, A. D., Lincoln, P.,

Morling, N., Rand, S., Sabatier, M., Scheithauer, R., Schneider, P. & Vide, M. C. (1997). Considerations from the European DNA profiling group (EDNAP) concerning STR nomenclature. Forensic Sci Int. 87: 185-92.

Gill, P., Jeffreys, A. J. & Werret,D. J. (1985). Forensic applications of DNA 'fingerprinting'. Nature. 38: 577-79.

Gill, P., Sparkes, R. & Kimpton, C. (1997). Development of guidelines to designate alleles using an STR multiplex system. Forensic Sci Int. 89: 185-97.

Gill, P., Urqhart, A., Millican, E., Oldroyd, N., Watson, S., Sparkes, S., Kimpton., C. P. (1996). A New Method of STR Interpretation Using Inferential Logic-Development of a Criminal Intelligence Database. Int J Legal Med. 109: 14-22.

Gonzalez-Neira, A., Gusmao, L., Barral, S., Lareu, M. V., and Carracedo, A. (2000). Multiplexing Y chromosome STRs: Analysis of artifactual bands and PCR strategies. pp: 436-438. In Sensabaugh, G, F., Brinkmann B., Lincoln P. (Eds), Progress in Forensic Genetics 8. Excerpta Medica, Amsterdam.

Gonzalez-Neira, A., Gusmao, L., Brion, M., Lareu, M. V., Amorim, A. & Carracedo, A. (2000). Distribution of Y-chromosome STR defined haplotypes in Iberia. Forensic Sci Int. 110:117-26.

Government of Pakistan Official web site: http://www.pak.gov.pk/public/people/index.html.

Griffiths, R. A. L., Barber, M. D., Johnson, P. E., Gillbard, S. M., Haywood, M. D., Smith, C. D., Arnold, J., Burke, T., Urquhart, A., and Gill, P. (1998). New reference allelic ladders to improve allelic designation in a multiplex STR system. Int J Legal Med. 111: 267-72.

Guo, S. W & Thompson, E. A. (1992). Performing exact test of Hardy- Weinberg proportion for multiple alleles. Biometrics. 48: 361-72.

Gusmão. L., Gonzalez- Neira. A., Sanchez-Diz. P., Lareu.M. V. & Carracedo. A. (2000). Alternative primers for DYS 391 typing: advantages of their application to forensic genetics. Forensic Sci Int. 106: 163-72.

Halos, S. C., Chu, J. Y., Ferreon, A. C., Magno, M. M. (1999). Philippine population database at nine microsatellite loci for forensic and paternity applications. Forensic Sci Int.101: 27-32.

Hammer, M. F. (1995). A recent common ancestry for human Y chromosome. Nature. 379: 376-78.

Hammer, M. F., Karafet, T., Rasanayagam, A., Wood, E. T., Altheide, T. K., Jenkins, T., Griffiths, R. C., Templeton, A. R. & Zegura, S. L. (1998). Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. Mol Biol Evol. 15: 427-41.

Hammer, M. F., Spurdle, A. B., Karafet, T., Bonner, M. R., Wood, E. T., Novelletto, A., Malaspina, P,, Mitchell, R. J., Horai, S., Jenkins, T., & Zegura, S. L. (1997). The geographic distribution of human Y chromosome variation. Genetics. 145: 787-805.

Hammond, H. A., Jin, L., Zhong, Y., Caskey, C. T., & Chakorborty, R. (1994). Evaluation of 13 short tandem repeat loci for use in personal identification applications. Am J Hum Genet. 55: 175-89.

Hardman N. (1986). Structure and function of repetitive DNA in eukaryotes. J Biochem. 234:1-11.

Haribson, S. A. & Buckleton, J. S. (1998). Applications and extensions of sub population theory: a caseworkers guide. J Forensic Sci Soc. 38: 249-53.

Hearne, C. M. & Todd, J. A. (1991). Tetranucleotide repeat polymorphism at the HPRT locus. Nucl Acids Res. 19: 5450.

Heyer, E., Puymirat, J., Dieltjes, P., Bakker, E. & de Knijff, P. (1997). Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. Hum Mol Genet. 6: 799-803.

Higgs, D. R., Goodbourn, S. E., Wainscoat, J. S., Clegg, J. B. & Weatherall, D. J. (1981). Highly variable regions of DNA flank the human alpha globin genes. Nucl Acids Res. 9: 4213-24.

Horst, B., Eigel, A., Sanguansermsri, T., & Brinkmann, B. (1999). Human Y-chromosomal STR types in north Thailand. Int J Legal Med. 112: 211-12.

Huang, N. E., Schumm, J. & Budowle, B. (1995). Chinese population data on three tetrameric short tandem repeat loci--HUMTHO1, TPOX, and CSF1PO--derived using multiplex PCR and manual typing. Forensic Sci Int. 71: 131-6.

Jakubiczka. S., Arnemann, J., Cooke, H. J., Krawczak, M., & Schmidtke, J. (1989). A search for restriction fragment length polymorphisms on the human Y chromosome. Hum Genet. 84: 86-88.

Jan, S. (1996). History and development of the Kalasha. pp: 239-242. In: Proceedings of 2nd International Hindukush Cultural Conference. Oxford University Press. Karachi.

Jarridge, J. F. & Meadow, R. H. (1980). The antecedents of civilisation in the Indus valley. Sci American. 243: 122-33.

Jeffreys, A. J., Brookfield, J. F. Y. & Semeonoff, R. (1985). Positive identification of an immigration test case using human DNA fingerprints. Nature. 317: 818-19.

Jeffreys, A. J., Macleod, A., Tamaki, K., Neil, D. L., Monckton, D. G. (1991). Minisatellite repeat coding as a digital approach to DNA Typing. Nature. 354: 202-209.

Jeffreys, A. J., Wilson, V., & Thein, S. L. (1987). Individual specific fingerprints of DNA. Nature. 316: 76-79.

Jeffreys, A. J., Wilson, V., Thein, S. L. (1985). Hypervariable minisatellite regions in human DNA. Nature. 314: 67-73.

Jettmar, K. (1975). Die Religionen des Hindukush. In Collection: Die Religionen der Menschheit. 4 (1). Stuttgart, Germany.

Jin L, Chakraborty R. (1995). Population structure, stepwise mutations, heterozygote deficiency and their implications in DNA forensics. Heredity. 74: 274-85.

Jobling, M. A. & Tyler-Smith, C. (1995). Fathers and sons: the Y chromosome and human evolution. Trends Genet. 11: 449-56.

Jobling, M. A. (1994). Survey of long-range DNA polymorphisms on the human Y chromosome. Hum Mol Genet. 3: 107-14.

Jobling, M. A., Bouzekri, N. & Taylor, P. G. (1998). Hypervariable digital DNA codes for human paternal lineages: MVR-PCR at the Y-specific minisatellite, MSY1 (DYF155S1). Hum Mol Genet. 7: 643-53.

Jobling, M. A., Heyer, E., Dieltjes, P., & Knijff, P. (1999). Y chromosome specific microsatellite mutation rates re-examined using a minisatellite, MSY1 Hum Mol Genet. 8: 2117-20.

Jobling, M. A., Pandya, A., & Tyler-Smith, C. (1997). The Y chromosome in forensic analysis and paternity testing. Int J Legal Med. 110: 118-24.

Jones, D. A. (1972). Blood samples: Probability of discrimination. J Forensic Sci Soc. 12: 355-59.

Kanter, E., Baird, M., Scaler, R., Bazalas, I. & Glassberg, J. (1986). Analysis of restriction fragment length polymorphism in deoxyribonucleic acid (DNA) recovered from dried blood stains. J Forensic Sci. 31: 403-408.

Kayser, M., Caglia, A., Corach, D., Fretwell, N., Gehrig, C., Graziosi, G., Heidorn, F., Herrmann, S., Herzog, B., Hidding, M., Honda, K., Jobling, M., Krawczak, M., Leim, K., Meuser, S., Meyer, E., Oesterreich, W., Pandya, A., Parson, W.,

Penacino, G., Perez-Lezaun, A., Piccinini, A., Prinz, M., Schmitt, C., Schneider, P. M., Szibor, R., Teifel-Greding, J., Weichhold, G. M., de Knijff, P. & Roewer, L. (1997). Evaluation of Y-chromosomal STRs: a multicenter study. Int J Legal Med. 110: 125-33.

Kayser, M., Roewer, L., Hedman, M., Henke, L., Henke, J., Brauer, S., Kruger, C., Krawczak, M., Nagy, M., Dobosz, T., Szibor, R., de Knijff , P., Stoneking, M., & Sajantila, A. (2000). Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs. Am J Hum Genet. 66: 1580-88.

Kimpton, C. P., Fisher, D., Watson, S., Adams, M., Urquhart, A., Lygo, J. & Gill, P. (1994). Evaluation of an automated DNA profiling system employing multiplex amplification of four tetrameric STR loci. Int J Legal Med. 106: 302-11.

Kimpton, C. P., Gill, P., d'Aloja, E., Andersen, J. F., Bar, W., Holgersson, S., Jacobsen, S., Johnsson, V., Kloosterman, A. D., Lareu, M. V., Nellemann, L., Pfitzinger, H., Phillips, C. P., Rand, S., Schmitter, H., Schneider, P. M., Sternersen, M. & Vide, M. C. (1995) Report on the second EDNAP collaborative STR exercise. Forensic Sci Int. 71: 137-52.

Kimpton, C. P., Gill, P., Walton, A., Urqhart, A., & Millican, E. S. M. (1993) Automated DNA profiling employing multiplex amplification of short tandem repeat loci. PCR Methods and Applications. 3: 13-22.

Kimpton, C. P., Oldroyd, N. J., Watson, S. K., Frazier, R. R. E., Johnson, P. E., Millican, E. S., Urquhart, A., Sparkes, B. L. & Gill, P. (1996). Validation of highly discriminating multiplex short tandem repeat amplification systems for individual identification. Electrophoresis. 17: 1283-93.

Kimpton, C. P., Walton, A., & Gill, P. (1992). A further tetranucleotide repeat polymorphism in the vWF gene. Hum Mol Genet. 1: 287.

Kimura, M. & Ohta, T. (1978). Stepwise mutation model and distribution of allelic frequencies in a finite population.Proc Natl Acad Sci USA. 75: 2868-72.

Kimura, M. (1993). The neutral theory of molecular evolution. Cambridge University Press, Cambridge.

Kline, M. C., Duewer, D. L., Newall, P., Redman, J. W., Reeder, D. J. & Richard, M. (1997). Interlaboratory evaluation of short tandem repeat triplex CTT. J Forensic Sci. 42: 897-906.

Klintschar, M., Al-Hammadi, N., Reichenpfader, B. (1999b). Population genetic studies on the tetrameric short tandem repeat loci D3S1358,VWA, FGA, D8S1179, D21S11, D18S51, D5S818, D13S317 and D7S820 in Egypt. Forensic Sci Int. 30: 104: 23-31.

Klintschar, M., Ebner, A. & Reichenpfader, B. (1999a). Population genetic studies on nine tetrameric short tandem repeat loci using fluorescence dye-labeled primers and capillary electrophoresis in the Austrian population. Electrophoresis. 20 : 1740-42.

Kloosterman, A. D., Pouwels, M., Daselaar, P. & Janssen, H. J. T. (1998). Population genetic study of Y-chromosome specific STR-loci in Dutch Caucasians. pp: 491-493. In Olaisen, B., Brinkmann, B., Lincoln, P. (Eds), Progress in Forensic Genetics 7, Excerpta Medica, Amsterdam.

Kunkel, L. M., Smith, K. D., Boyer, S. H. (1979). Organization and heterogeneity of sequences within a repeating unit of human Y chromosome deoxyribonucleic acid. Biochemistry. 18: 3343-53.

Kupferschmid, T. D., Calicchio, T., Budowle, B. (1999). Maine Caucasian population DNA database using twelve short tandem repeat loci. J Forensic Sci. 44: 392-95.

Lander, E. S. (1989). DNA fingerprinting on trial. Nature. 339: 501-5.

Lee, J. C., Chen, C. H., Tsai, L. C., Linacre, A., Chang, J. G. (1997). The screening of 13 short tandem repeat loci in the Chinese population. Forensic Sci Int. 87: 137-44.

Lench, N., Stainer, P., & Williamson, R. (1988). Simple non-invasive method to obtain DNA for gene analysis. Lancet. 18: 1356- 58.

Leroy, P., Casanova, M., Seboun, E., Mediguies, C., Siniscalco, M., & Felous, M. (1985). DNA sequence and analysis of the human Y chromosome: presence of restriction fragment length polymorphisms. Cyto Genet Cell Genet. 40: 680.

Lessig, R. & Edelmann, J. (1998). Y chromosome polymorphisms and haplotypes in West Saxony (Germany). Int J Legal Med. 111: 215-18.

Levinson, G, Gutman, G. A. (1987). Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol. 4: 203-21.

Lewis, P. O., & Za Laurent Excoffierykin, D. 2000. Genetic Data Analysis: Computer program for the analysis of allelic data. Version 1.0 (d15). Free program distributed by the authors over the internet from the GDA Home Page at http://alleyn.eeb.uconn.edu/gda/.

Li, H., Schmidt, L., Wei, M., Hustad, T., Lerman, M. I., Zbar, B. & Tory, K. (1993). Three tetranucleotide polymerphisms for loci, D3S11352; D3S1358; D3S1359. Hum Mol Genet. 2: 1327.

Licoln,P. J. (1997). Criticisms and concerns regarding DNA profiling. Forensic Sci Int. 88: 23-31.

Lincoln, P. J. (1992). Presidential Address: 'I am in blood....' (Macbeth Act III Scene V. Med Sci Law. 32: 277-88.

Litt, M., & Luty, J. A. (1989). A Hypervariable microsatellite revealed by in vitro amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am J Hum Genet. 44: 397-401.

Lygo, J. E., Johnson, P. E., Holaway, D. J., Woodroffe, S., Whitaker, J. P., Clayton, T. M., Kimpton, C. P. & Gill, P. (1994). The validation of short tandem repeat loci for use in forensic casework. Int J Leg Med. 107: 77-89.

Malaspina, P., Persichetti, F., Novelletto, A., Iodice, C., Terrenato, L., Wolfe, J., Ferraro, M., Prantera, G. (1990). The human Y chromosome shows a low level of DNA polymorphism. Ann Hum Genet. 54: 297-305.

Marshall, J. H. (1951). Taxilla. pp 62-80. Cambridge University Press, UK.

Martin, P. D; Schmitter, H., Schneider, P. M. (2001). A brief history of formation of DNA databases in forensic science within Europe. Forensic Sci Int. 199: 225-231.

Mathias, N., Bayes, M., Tyler-Smith, C. (1994). Highly informative compound haplotypes for the human Y chromosome. Hum Mol Genet. 3: 115-23.

Mazari, S. S. (1999) Baluch History. pp: 85-88. In, Journey to Disillusionment. Mazari, S. (Ed). Oxford University Press, Karachi.

Melton, T., Peterson, R., Redd, A. J., Saha, N., Sofro, A. S., Martinson, J. & Stoneking, M. (1995). Polynesian genetic affinities with Southeast Asian populations as identified by mtDNA analysis. Am J Hum Genet. 57: 403-14.

Miller,P. M. (2000). Department of Biological Sciences Northern Arizona University. Free program distributed by the authors over the internet from the TFPGA (Ver 6.0) Home Page at http://herb.bio.nau.edu/~miller/tfpga.htm.

Mills, K. A., Even. D. & Murray, J. C. (1992). Tetranucleotide repeat polymorphism at the human alpha fibrinogen locus (FGA). Hum Mol Genet. 1: 779.

Minch, E., Ruiz-Linares, A., Goldstein, D., Feldman, M. & Cavalli-Sforza, L. L. (2000). Microsat v.1.5d: a computer program for calculating various statistics on microsatellite allele data downloaded from the Stanford University web site http://lotka.stanford.edu/microsat/microsat.html.).

Moller, A., Wiegand, P., Gruschow, C., Seuchter, S.A., Baur, M. P. & Brinkmann, B. (1994). Population data and forensic efficiency values for the STR systems HumvWA, HumMBP and HumFABP. Int J Legal Med. 106: 183-9.

Monson, K. L. & Budowle, B. (1998). Effect of reference database on frequency estimates of polymerase chain reaction (PCR)-based DNA profiles. J Forensic Sci. 43: 483-8.

Morgenstierne, G. (1973). Report on a linguistic mission to Northwestern India. CIII-1. pp: 181-238. Instituttet for Sammenligende Kultuforskining, CIII-1.

Morton, N. E. (1992). Genetic structure of forensic populations. Proc Natl Acad Sci USA. 89: 2556-60.

Mullenbach, R., Lagoda, P. J, Welter, C. (1989). An efficient salt-chloroform extraction of DNA from blood and tissues. Trends Genet. 5: 391.

Muller. S., Gomolka, M. & Walter, H. (1994) The Y-specific SSLP of the locus DYS19 in four different European samples. Hum Hered. 44: 298-300.

Nakamura, Y., Leppert, M. O., Wolff, R., Holm, T., Culver, M., Martin, C., Fujimoto, E., Hoff, M., Kumlin, E. & White, R. (1987). Variable number of tandem Repeat (VNTR) markers for human gene mapping. Science. 235:1616–22.

Nakamura,Y., Carlson, M., Krapcho, K. & White, R. (1989). Isolation and mapping of a polymorphic DNA sequence (pMCT118) on chromosome 1p (D1S80). Nucl Acids Res. 16: 9364.

National Research Council. (1992). DNA technology in forensic science. National Academy Press, Washington, DC.

National Research Council. (1996a). Population genetics. pp: 89-124. In: The Evaluation of Forensic DNA Evidence. National Academy Press, Washington, DC.

National Research Council. (1996b).DNA evidence in legal system. pp: 166-211. In: The Evaluation of Forensic DNA Evidence. National Academy Press, Washington, DC.

National Research Council. (1996c). Overview. pp: 9-46. In: The Evaluation of Forensic DNA Evidence. National Academy Press, Washington, DC.

Nei, M. & Roychoudhury, A. K. (1993). Evolutionary relationships of human populations on a global scale. Mol Biol Evol. 10: 927-43.

Nei, M. (1987). Molecular evolutionary genetics. Columbia University Press, New York, USA New York.

Nishimura, D. Y. & Murray, J. C. (1992). A tetranucleotide repeat for the F13B locus. Nucl Acids Res. 20: 1167.

Olaisen, B., Bar, W., Brinkmann, B., Budowle, B., Carracedo, A., Gill, P., Lincoln, P., Mayr, W. R. & Rand, S. (1998). DNA Recommendations of the International Society for Forensic Genetics. Vox Sanguinis. 7461-63.

Olaisen, B., Bar, W., Mayr, W.R., Lincoln, P., Carracedo, A., Brinkmann, B., Budowle, B. & Gill, P. (1998). DNA recomendations–further report of the DNA Commission of the ISFH regarding the use of short tandem repeat systems. Forensic Sci Int. 87: 179-184.

Oldroyd, N. J., Urquhart, A. J., Kimpton, C. P., Millican, E. S., Watson, S. K., Downes, T. & Gill, P. D. (1995). A highly discriminating octoplex short tandem repeat polymerase chain reaction system suitable for human individual identification. Electrophoresis. 16: 334-37.

Ovington, A., Daselaar, P., Sjerps, M. & Kloosterman, A. (1997). A Dutch population study of the STR loci D21S11 and HUMFIBRA. Int J Legal Med. 110: 14-17.

Page, R. D. M. (1996). TREEVIEW: An application to display phylogenetic trees on personal computers. Computer Applications in the Biosciences. 12: 357-358.

Parkes, P. (1983). Alliance and Elopment: Economy, Social order and sexual antagonism among the Kalash (Kalash Kafirs) of Chitral. Doctoral Thesis. University of Oxford.

Passarino, G., Semino, O., Bernini, L. F., Santachiara-Benerecetti, A. S. (1996). Pre-Caucasoid and Caucasoid genetic features of the Indian population, revealed by mtDNA polymorphisms. Am J Hum Genet. 59: 927-34.

Pawlowski, R., Branicki, W. & Kupiec, T. (1999). Y-chromosomal polymorphic loci DYS19, DYS390, DYS393 in a population sample from northern Poland. Electrophoresis. 20: 1702-06.

Pena, S. D. J., Prado, V. F., Epplen, J. T. (1995). DNA Diagnosis of human genetic individuality. J Mol Med. 73: 555-64.

Perez-Lezaun, A., Calafell, F., Comas, D., Mateu, E., Bosch, E., Martinez-Arias, R., Clarimon, J., Fiori, G., Luiselli, D., Facchini, F., Pettener, D., & Bertranpetit, J. (1999). Sex-specific migration patterns in Central Asian populations, revealed by analysis of Y chromosome short tandem repeats and mtDNA. Am J Hum Genet. 65: 208-19.

Perez-Lezaun, A., Calafell, F., Seielstad, M., Mateu, E., Comas, D., Bosch, E. & Bertranpetit, J. (1997). Population genetics of Y-chromosome short tandem repeats in humans. J Mol Evol. 45: 265-70.

Pestoni, C., Cal, M. L., Lareu, M. V., Rodriguez-Calvo, M. S., & Carracedo, A. (1998). Y chromosome STR haplotypes: genetic and sequencing data of the Galician population (NW Spain). Int J Legal Med: 112: 15-21

Pestoni, C., Lareu, M. V., Rodriguez, M. S., Munoz, I., Barros, F. & Carracedo, A. (1995). The use of the STRs HUMTH01, HUMVWA31/A, HUMF13A1, HUMFES/FPS, HUMLPL in forensic application: validation studies and population data for Galicia (NW Spain). Int J Legal Med: 107: 283-90.

Philips, C. P., Thacker, C., Sandher, S., & Syndercombe Court, D. (2000). UK Caucasian databases for six Y chromosome STRs. pp: 290-292. In Sensabaugh, G. F., Brinkmann, B., Lincoln, P. (Eds), Progress in Forensic Genetics 8. Excerpta Medica, Amsterdam.

Polymeropoulos, M. H., Rath. D, S., Xiao, H. & Merril, C. R. (1991 b). Tetranucleotide Repeat Polymorphism At the Human Tyrosine Hydroxylase Gene (TH). Nucl Acid Res. 19: 3753.

Polymeropoulos, M. H., Rath. D, S., Xiao, H. & Merril, C. R..(1991 a). Tetranucleotide Repeat Polymorphism At the Human c-fes Proto-oncogene(FES). Nucl Acid Res. 19: 4018.

Potter, A. A., Hanham, A. F., Nestmann, E. R. (1985) A rapid method for the extraction and purification of DNA from human leukocytes. Cancer Lett. 26: 335-41.

Prinz, M., Boll, K., Baum, H., & Shaler, B. (1997). Multiplexing of Y chromosome specific STRs and performance for mixed samples. For Sci Int. 85: 209-218.

Pu, C., Hsieh, C., Chen, M., Wu, F., & Sun, C. (1999). Genetic variation at nine STR loci in populations fram the Philippines and Thailand living in Taiwan. Forensic Sci Int. 106: 1-6.

Pu, C., Wu, F., C, C., Wu, K., Chao, C., Li, J. (1998). DNA short tandem repeat profiling of Chinese population in Taiwan determined by using an automated sequencer. Forensic Sci Int. 97: 47-51.

Puers, C., Hammond, H. A,, Caskey, C. T., Lins, A. M,, Sprecher, C. J., Brinkmann, B. & Schumm, J. W. (1994). Allelic ladder characterization of the short tandem repeat polymorphism located in the 5' flanking region to the human coagulation factor XIII A subunit gene. Genomics. 23: 260-64.

Puers, C., Hammond, H. A., Jin, L., Caskey, C. T. & Schumm, J. W. (1993). Identification of repeat sequence heterogeneity at the polymorphic short tandem repeat locus HUMTH01[AATG]n and reassignment of alleles in population analysis by using a locus-specific allelic ladder. Am J Hum Genet. 53: 953-58.

R. v Deen. Times Law Reports. (1994). Jan 10. pp: 11.

Redd, A. J., Clifford, S. L., & Stoneking, M. (1997). Multiplex DNA typing of short tandem repeat loci on the Y-chromosome. Biol Chem. 378: 923-27.

Rendell, H. M., Dennel, R. W. & Halim, M. A. (1989). Pleistocene and paleolithic investigations in the Soan Valley, north Pakistan. British Archeological Mission to Pakistan, Report. Series 2: 191.

Richards, B., Skoletsky, J., Shuber, A. P., Balfour, R., Stern, R. C., Dorkin, H. L., Parad, R. B., Witt, D., & Klinger, K. W. (1993). Mutiplex PCR amplification from the CFTR gene using DNA prepared from buccal brushes/swabs. Hum Mol Genet. 2: 159-63

Roewer, L. & Epplen, J. T. (1992). Rapid and sensitive typing of forensic stains by PCR amplification of polymorphic simple repeat sequences in case work. Forensic Sci.Int. 53: 163-71.

Roewer, L., Arnemann, J., Spurr, N. K., Grzeschik, K. H., & Epplen, J. T. (1992). Simple repeat sequences on the human Y chromosome are equally polymorphic as their autosomal counterparts. Hum Genet. 89: 389-94.

Roewer, L., Kayser, M., Anslinger, K., Augustin, C., Caglià, A., Corach, D., Füredi, S., Geserick, G., Henke, L., Hidding, M., Kärgel, H. J., de Knijff, P., Lessig, R., Pascali, V. L., Parson, W., Prinz, M., Rolf, B., Schmitt, C., Schneider, P. M., Szibor, R., Teifel-Greding, J. & Krawczak, M. (2000). European Y-STR Haplotype Reference Database For Forensic Application. pp: 613-615. In Sensabaugh, G. F., Brinkmann, B., Lincoln, P. (Eds), Progress in Forensic Genetics 8. Excerpta Medica, Amsterdam.

Roewer, L., Kayser, M., Dieltjes, P., Nagy, M., Bakker, E., Krawczak, M. & de Knijff, P. (1996) Analysis of molecular variance (AMOVA) of Y-chromosome-specific microsatellites in two closely related human populations. Hum Mol Genet. 5: 1029-33.

Rolf, B., Waterkamp, K. & Huhne, J. (1998). Allele frequency data for the FGA locus in eight populations. Int J Legal Med. 111: 55-56.

Rossi, E., Rolf, B., Schurenkamp, M., & Brinkmann, B. (1999). Y-chromosome STR haplotypes in an Italian population sample. Int J Legal Med. 112: 78-81.

Rozen, S. & Skaletsky, H. J. (1997) Primer3. Code available at http://www-genome.wi.mit.edu/genome_software/other/primer3.html.

Ruitberg, C. M., Reeder, D. J., Butler, J. M. (2001). STRBase: a short tandem repeat DNA database for the human identity testing Community. Nucleic Acids Res. 29: 320-22.

Saiki, R.K., Scharf, S., Faloona, F., Mullis, K. B., Horn, G.T., Erlich, H. A & Arnheim, N. (1985). Enzymatic amplification of beta-globin genomic sequences and restriction site analysis of sickle cell anemia. Science. 230: 1350-54.

Sambrook, J., Fritsch, E. F. & Maniatis, T. (1989). Gel electrophoresis of DNA. pp: 6.37. In, Molecular Cloning. A Laboratory Manual. Cold Spring Harbour Laboratory Press, New York, USA.

Sankiala, H. D. (1967). Pre and proto history in India and Pakistan: New discoveries and fresh interpreattions. J Indian History. 45: 103.

Santos, F. R., Gerelsaikhan, T., Munkhtuja, B., Oyunsuren, T., Epplen, J. T. & Pena, S. D. J. (1996). Geographic differences in the allele frequencies of the human Y-linked tetranucleotide polymorphism DYS19. Hum Genet. 97: 309-13.

Santos, F. R., Pena, S. D. J. & Epplen, J. T. (1993 a) Genetic and population study of a Y-linked tetranucleotide repeat DNA polymorphism with a simple non-isotopic technique. Hum Genet. 90: 655-56.

Santos, S. M., Budowle, B., Smerick, J. B., Keys, K. M. & Moretti, T. R. (1996). Portuguese population data on the six short tandem repeat loci, CSF1PO, TPOX,THO1, D3S1358, VWA and FGA. Forensic Sci Int: 83:229-35.

Sasaki, M. & Dahiya, R. (2000). The polymorphisms of various short tandem repeats on the Y chromosome in Japanese and German populations.Int J Legal Med. 113: 181.

Scherczinger, C. A., Hintz, J. L., Peck, B. J., Adamowicz, M. S., Bourke, M. T., Coyle, H. M., Ladd, C., Yang, N. C. S., Budowle, B., & Lee, H. C. (2000). Allele frequencies for the CODIS core STR loci in Connecticut populations. J Forensic Sci. 45: 938-40.

Schlotterer. C. (1998). Genome evolution: are microsatellites really simple sequences. Curr Biol. 12: R132-4.

Schmidtke J. A. (1976). DNA extraction procedure practicable under field work conditions. Experientia. 32: 400.

Schneider, P. M. & Martin, P. D. (2001). Criminal DNA databases: the European situation. Forensic Sci Int. 119: 232-238.

Schneider, S., Roessli, D., & Excoffier, L. (2000). Arlequin ver. 2.000: A software for population genetic data analysis. Genetics and Biometry Laboratory, University of Geneva, Switzerland. Distributed free over web: http://lgb.unige.ch/arlequin/.

Schomberg, R. C. F. (1938). Kafirs and Glaciers. pp: 26-52. In, Trvales in Chitral. London Press.

Schroer, P., Schmitt, C. & Staak, M. (2000). Western German population data for the two highly polymorphic STR loci D21S11 and FGA and sequence data for rare alleles. pp: 139-141. In Progress in Forensic Genetics 8. In Sensabaugh, G, F., Brinkmann B., Lincoln P. (Eds). Excerpta Medica, Amsterdam.

Seielstad, M. T., Hebert, J. M., Lin, A. A., Underhill, P. A., Ibrahim, M., Vollrath, D. & Cavalli-Sforza, L. L. (1994). Construction of human Y-chromosomal haplotypes using a new polymorphic A to G transition. Hum Mol Genet. 3: 2159-61.

Seielstad, M. T., Minch, E. & Cavalli-Sforza, L. L. (1998). Genetic evidence for a higher female migration rate in humans. Nat Genet. 20: 278280.

Seielstad, M., Bekele, E., Ibrahim, M., Toure, A., & Traore, M. (1999). A view of modern human origins from Y chromosome microsatellite variation. Genome Res. 9: 558-67.

Shapiro, E. D. & Reifler, S. (1996). Forensic DNA analysis and the United States Government. Med Sci Law. 36: 43-51.

Shriver, M. D., Jin, L., Chakraborty, R., Boerwinkle, E. (1993). VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach. Genetics. 134: 983-93.

Siiger, H. (1956). Ethnological field research in Chitral, Sikkim and Assam. Hist. Fil. Medd.).Koneglige Danske Viden. Selskab. 36 (20): 5-35.

Simmler, M. C., Johnson, C., Petit, C., Rouyer, F., Vergnaud, G., Wessenbach, J (1987). Two highly polymorphic minsatelltes from the pseudoautosomal region of the human sex chromosomes. EMBO J. 6: 963–69.

Singer, M. F (1982). SINES and LINES- highly repeated short and long interspersed sequences in mammalian genomes. Cell. 28: 433-34.

Slyvester, J. T. (1992). Recent developments in the DNA admissibility. pp: 61-83. In, Proceedings of 3rd International Promega Symposium on Human Identification.

Smith, L. M., Sanders, J. Z., Kaiser, R. J., Hughes, P., Dodd, C., Connell, C. R., Heiner, C., Kent, S. B. & Hood, L. E. (1986). Fluorescence detection in automated DNA sequence analysis. Nature. 321: 674-79.

Smith, R. N. (1995). Accurate size comparison of the short tandem repeat alleles by PCR. Biotechniques.18: 122-28.

Southern, E. M. (1995). DNA fingerprinting by hybridisation to oligonucleotide arrays. Electrophoresis. 16: 1539-1542.

Sparkes, R., Kimpton, C., Gilbard, S., Carne, P., Andersen, J., Oldroyd, N., Thomas, D., Urquhart, A. & Gill, P. (1996 a). The validation of a 7-locus multiplex STR test for use in forensic casework (II). Artefacts, casework studies and success rates. Int J Legal Med. 109: 195-204.

Sparkes, R., Kimpton, C., Watson, S., Oldroyd, N., Clayton, T., Barnett, L., Arnold, J., Thompson, C., Hale R, Chapman, J., Urquhart, A. & Gill, P. (1996 b). The validation of a 7-locus multiplex STR test for use in forensic casework (I). Mixtures, ageing, degradation and species studies. Int J Legal Med. 109: 186-94.

Sprecher, C. J., Puers, C., Lins, A. M., & Schumm, J. M. (1996). General appraoch to analysis of polymorphic short tandem repeat loci. Biotechniques. 20: 266-76.

Spurdle, A. B. & Jenkins T. (1993). Complex polymorphisms are revealed by Y chromosome probe 49a with BglII,HindIII, PstI and SstI. Ann Hum Genet. 57: 41-53.

Stoneking, M., Fontius, J. J., Clifford, S. L., Soodyall, H., Arcot, S. S., Saha, N., Jenkins, T., Tahir M. A., Deininger, P. L., Batzer, M. A. (1997). Alu insertion polymorphisms and human evolution: evidence for a larger population size in Africa. Genome Res.7: 1061-71.

Strickberger, M. W. (1985). Gene Frequencies and equilibrium. In: Genetics, 669-685. 3rd edition. Macmillan Publishing Co. Singapore.

Stryer, L. (1988). Eukaryotic chromosomes and gene expression. pp: 823-850. In, Biochemistry. W. H. Freeman & Co. New York.

Szibor, R., Lautsch. S., Plate, I., Bender., K. & Krase, D. (1998). Population genetics of the STR HumD3S1358 in two regions of Germany. Int J Legal Med . 111: 160-61.

Tagliabracci, A., Buscemi, L., Pesaresi, M., Rodriguez, D., Caenazzo, L., Ponzano, E., Fenato, F. A., Cortivo, P., Previdere, C., Peloso, G., Grignani, P., Pierucci, G., Polizzi, E., Nardone, M., Della Mora, C., and Domenici, R. (2000). Population study and paternity testing of seven Y-chromosome STR-loci in an Italian population sample. pp: 269-271. In Sensabaugh, G, F., Brinkmann, B., Lincoln, P. (Eds), Progress in Forensic Genetics 8. Excerpta Medica, Amsterdam.

Tautz, D. & Renz, M. (1984). Simple sequences are ubiquitous repetitive components of eukaryotic genomes. Nucl Acids Res. 12: 4127-38.

Tautz, D. (1993). Notes on definition and nomenclature of tandemly repetitive DNA sequences. pp: 21- 28. In Pena, D. J., Chakraborty, R., Epplen, J. T., Jeffreys, A. J. (Eds), DNA Fingerprinting: State of Science. Birkhauser Verlag, Basel.

Tautz, D.(1989). Hypervariability of simple sequences as a general source for poylmorphic DNA markers. Nucl Acids Res.17: 6463-71.

Tautz, D., Trick. M., Dover G. A. (1986). Cryptic simplicity in DNA is a major source of genetic variation. Nature. 322: 652-56.

Tereba, A. (1999). Tools for Analysis of Population Statistics PowerStats. Profiles in DNA. 2: 14-16.

Thomas, M. G., Bradman, N., & Flinn, H. M. (1999). High throughput analysis of 10 microsatellite and 11 diallelic polymorphisms on the human Y-chromosome. Hum Genet. 105: 577-81.

Tobal, K., Layton, D. M., & Mufti, G. J. (1989). Non-invasive isolation of constitutional DNA for genetic analysis. Lancet. 2: 1281-2.

Trabetti, E., Galavotti, R., & Piganti, P. (1993). Genetic Variation in the Italian population at five tandem repeat loci amplified in vitro: use in paternity testing. Molecular & Cellular Probes. 7: 81-7.

Tun, Z., Honda, K., Nakatome, M., Nakamura, M., Shimada, S., Ogura, Y., Kuroki, H., Yamazaki, M., Terada, M., & Matoba, R. (1999). Simultaneous detection of multiple STR loci on sex chromosomes for forensic testing of sex and identity. J Forensic Sci. 44: 772-77.

Tyler-Smith, C., Taylor, L. & Muller, U. (1998). Structure of a hypervariable tandemly repeated DNA sequence on the short arm of the human Y chromosome.J. Mol. Biol. 203: 837-48.

Umetsu, K., Watanabe, G., Yuasa, I., Ago, K., Nakayashiki, N., Miyoshi, A. & Kashimura, S. (2000). Haplotype distribution of four Y chromosomal STR loci in the East Asian populations. pp: 266-268. In Sensabaugh, G, F., Brinkmann B., Lincoln P. (Eds), Progress in Forensic Genetics 8. Excerpta Medica, Amsterdam.

Underhill, P. A, Jin, L., Lin, A. A., Mehdi, S. Q., Jenkins, T., Vollrath, D., Davis, R. W., Cavalli-Sforza, L. L. & Oefner, P. J. (1997). Detection of numerous Y

chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. Genome Res. 7: 996-1005.

Urqhart, A., Kimpton, C. P., Downes, T. J., & Gill, P. (1994). Variation in short tandem repeat sequences--a survey of twelve microsatellite loci for use as forensic identification markers. Int J Legal Med. 107: 13-20.

Urquhart, A., Oldroyd, N. J., Kimpton, C. P., & Gill, P. (1995). Highly Discriminating Heptaplex Short Tandem Repeat PCR System for forensic identification. Biotechniques 18: 116-21.

Valdes, A. M., Slatkin, M. & Freimer, N. B. (1993). Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics. 133: 737-49.

Waincoat, J. S., Pilington, S., Petro, T. E. A., Dell, J. & Higgs, D. R. (1987). Allele Specific Identity Patterns, Hum. Genet. 35: 384-87.

Wall, W. J., Williamson, R., Petrou, M., Papaioannou, D. & Parkin, B. H. (1993). Variation of short tandem repeats within and between populations. Hum Mol Genet. 2: 1023-29.

Wallin, J. M., Martin, M. P. H., Buoncristiani, M. R., Lazaruk, K. D., Fildes, N., Holt, C. L., & Walsh, P. S. (1998). TWGDAM validation of the AmpFISTR™ blue PCR amplification kit for forensic casework analysis. J Forensic Sci. 43: 854-70.

Walsh, P. S., Fildes, N. J. & Reynolds, R. (1996). Sequence analysis and characterization of stutter products at the tetranucleotide repeat locus vWA. Nucl Acids Res. 24: 2807-12.

Walsh, P. S., Metzger, D. A., Higuchi, R. (1991). Chelex 100 as a medium for simple extraction of DNA for PCR-based typing from forensic material. Biotechniques. 10: 506-13.

Warren, J. E., Planz, J. V., Richey, S., Aron, F., Arcot, S., Sinha, S., Cook, P., & Eisenburg, A. J. (2000). Forensic validation and population studied on the Y-PLEX™ kit for Y chromosome STR analysis. Proceedings of 11th International Symposium on Human Identification, Missippi, USA.

Watts, D. (1998). Genotyping STR loci using an automated DNA sequencer.pp: 193-208. In Forensic DNA profiling protocols. Humana Press Inc., Otawa, N J.

Weber, J. L. & May, P. E. (1989). Abundant class of Human DNA polymorphisms which can be typed using polymerase chain reaction. Am J Hum Genet. 44: 388-96.

Weber, J. L. & Wong, C. (1993). Mutation of human short tandem repeats. Hum Mol Genet. 2: 1123-28.

Weiner, A. S. (1972). Forensic blood groups, critical & historical review. New York State Journal of Medicine. 72: 810-16.

Weir, B. S. (1992). Population genetics in the forensic DNA debate. Proc Natl Acad Sci. USA. 89: 11654-59.

Weir, B. S. (1996). Estimating frequencies. In: Genetic data analysis II. pp: 31-90. Sinauer Associates Inc. USA.

Werret, D. J. (1997). The National DNA database. Forensic Sci Int. 88: 33-42.

White, P. S., Tatum, O. L., Deaven, L. L., & Longmire, J. L. (1999) New, male-specific microsatellite markers from the human Y chromosome. Genomics. 57: 433-437.

Whitfield, L. S., Sulston, J. E. & Goodfellow, P. N. (1995). Sequence variation of the human Y chromosome. Nature. 378: 379-80.

Wierdl, M., Dominska, M., Petes, T. D. (1997). Microsatellite instability in yeast: dependence on the length of the microsatellite.Genetics. 146: 769-79.

Wong, Z., Wilson, V., Patel, I., Povey. S. & Jeffreys, A. J. (1987). Characterisation of a Panel of Highly Varaiable Minisatellite Isolated from Human DNA. Ann Hum Genet. 51: 269-88.

Worley, J. M., Mansfiled, E. S. & Rubin, R. B. (1996). STR typing accuracy using different molecular markers. pp: 180. In Proceedings of the 6th International symposium on Human Identification Promega Corporation.

Wright, S. (1965). The interpretation of population structure by F-statistics with special regard to systems of mating. Evolution. 19: 395-420.

Wu, S., Seino, S. & Bell, G, I. (1990). Human cola gen type 11, alpha 1(COLA 2A1), VNTR polymorphism detected by gene amplification. Nucl Acid Res. 18: 3102.

Wyman, A. & White, R. (1980). A highly polymorphic locus in human DNA. Proc Natl Acad Sci USA. 77: 6754-58.

Yamamoto, T., Uchihi, R., Nozawa, H., Huang, X. L., Leong, Y. K., Tanaka, M., Mizutani, M., Tamaki, K., Katsumata, Y. (1999). Allele distribution at nine STR loci--D3S1358, vWA, FGA, TH01, TPOX, CSF1PO, D5S818, D13S317 & D7S820 in the Japanese population by multiplex PCR and capillary electrophoresis. J Forensic Sci. 44: 167-70.

Zerjal, T., Dashnyam, B., Pandya, A., Kayser, M., Roewer, L., Santos, F. R., Schiefenhovel, W., Fretwell, N., Jobling, M. A., Harihara, S., Shimizu, K., Semjidmaa, D., Sajantila, A., Salo, P., Crawford, M. H., Ginter, E. K., Evgrafov, O. V. & Tyler-Smith, C. (1997). Genetic relationships of Asians and Northern Europeans, revealed by Y-chromosomal DNA analysis. Am J Hum Genet. 60:1174-83.

Ziegle, J. S., Su, Y., Corcoran, K. P., Nie, L., Mayrand, P. E., Hoff, L. B., McBride, L. J., Kronick, M. N. & Diehl, S. R. (1992). Application of automated DNA sizing technology for genotyping microsatellite loci. Genomics. 14: 1026-31.