DNA CHARACTERIZATION OF THE FGA LOCUS IN THE HUMAN GENOME

Estifanos Kebede Asfaw

B. Sc. Med. Lab. Technology

Jimma University

Submitted in fulfillment of the requirements for the

Master of Medical Sciences (M.Med.Sc) degree

In the

Department of Haematology and Cell Biology

Faculty of Health Sciences

University of the Free State

Bloemfontein

Sourth Africa

Supervisor: Dr. André de Kock

Co-Supervisor: Prof. G.H.J. Pretorious

Bloemfontein

November 2002

DECLARATION

Hereby I declare that this script *"DNA CHARACTERIZATION OF THE FGA STR LOCUS IN THE HUMAN GENOME"* **submitted towards a M.Med.Sc degree at the University of the Free State is my original and independent work and has never been submitted to any other university or faculty for degree purposes.**

All the sources I have made use of or quoted have been acknowledged by complete references.

Estifanos Kebede
November 2002

This thesis is dedicated to:

My parents Kebede and Wolela,

My Wife Wubitu,

My children Ruth, Michael and Bereket,

and all those who were willing to help.

Acknowledgements

I would like to thank the following institutions and persons:

- Jimma University for allowing me to pursue my study.
- Irish Aid Ethiopia for sponsoring my study.
- Prof. PN Badenhorst & Department of Haematology and Cell Biology, Faculty of Health Sciences, University of the Free State for the manifold support.
- My supervisor, Dr. André de Kock, for his unrestrained guidance, supervision, understanding and helpful recommendations throughout the study period.
- My co-supervisor Prof. G.H.J. Pretorius for his encouragements.
- All members of the department for their acceptance, concern, kindness encouragement and smiles, especially Wendy McKay, Dr. M.J, Coetzee, Almarie du Plessis, Marelie Kelderman and Dr Lindi Coetzee.
- The International Office, University of Free State.
- Dr. Chris Viljoen and Dr. Nerina van der Merwe.

Without their help this would not have materialized.

Above all, to the King eternal, immortal, invisible, the only wise loving God who stuck with me closer than a brother throughout my studies.

TABLE OF CONTENTS

CHAPTER 1

LITERATURE REVIEW

CHAPTER 2

MATERIALS AND METHODS

CHAPTER 3

CHAPTER 4

LIST OF TABLES AND FIGURES

LIST OF ABBREVATIONS

BACs:	bacterial artificial chromosomes
bp:	base pairs
CE:	capillary electrophoresis
CODIS:	combined DNA index system
DNA:	deoxyribonucleic acid
dNTP:	deoxynucleoside triphosphate
FGA:	alpha fibrinogen gene/fibrinogen alpha
GDB:	genomic database
HLA:	human Leukocyte Antigen
HPCE:	high performance capillary electrophoresis
HWE/P:	Hardy-Weinberg equilibrium
ISFH:	International Society of Forensic Heamogenetics
kb:	kilo bases
LIF:	laser induced florescence
LOH:	loss of heterzygosity
LR:	likelihood ratio
MEC:	mean chance of exclusion
MOGs:	maternal obligatory genes
MSI:	microsatellite instability
OH/$H_{o:}$	observed heterozygosity
PCE/PPE:	prior chance of exclusion/prior probability of exclusion
PCR:	polymerase chain reaction
PD:	power of discrimination
PE:	power of exclusion
PI:	probability of identity
PIC:	polymorphic information content
POGs:	paternal obligatory genes
POP:	performance optimized polymer
RFLP:	restriction fragment length polymorphism

RNA:	ribonucleic acid
SGM:	second generation multiplex
SLP:	single locus polymorphism
STR:	short tandem repeats
TGM:	third generation multiplex
VNTRs:	variable number of tandem repeats

CHAPTER 1

LITERATURE REVIEW

1.1 Introduction

Today's complex modern society gives rise to many problems that require individual identification or biological relationship determination (Brooks MA 1994). Discrete genetic markers are being used increasingly to identify individuals. Genetic marker use is varied, with paternity testing being the most established. One of the earliest documented cases of personal identification is found in the Bible where King Solomon by divine wisdom gave resolution to a maternity dispute (Bible 1 kings 3: 16-18, de Kock A 1991, Silver H 1989). According to Chinese folklore (12th to 13th century), unique blood tests were employed when attempting to settle genealogical disputes. One method required dripping blood from a claimed relative on to the skeleton of the deceased (Silver H 1989). Today many levels of society have to contend with the increasing number of children born out of wedlock (Brooks MA 1994) and genetic profiles of mother, child, and alleged father are examined and can be used to determine paternity. Using this technique, mothers can also be identified, or families separated by war can be reunited (Weir BS 1996). The use of genetic markers to resolve paternity disputes can be traced back to 1902 when Karl Landsteiner discovered the ABO blood group system (Jeffreys AJ 1993, Silver H 1989). In 1924 Bernstein clarified the ABO blood system genetics and thus human blood markers (which were assumed to be transmitted in a clear-cut way) could be used in paternity disputes (Mayr WR 1991). Genetic profile examination need not be confined to the living and are often used in inheritance disputes or identification of remains from war or other disasters. The profiles of the remains are compared to living family members (Weir BS 1996, Helminen P *et al* 1991, Lee JW *et al* 2001).

Individual identification and determination of biological relationship is used in investigating the following: newborn infants kidnapped from hospital nurseries, children abducted by non custodial parents or strangers, applicant immigrants and their familial sponsors, participants in surrogate parenting contracts, heirs to disputed estates, and cases of disputed parentage (Brooks MA 1994). Another major use of genetic profiles is in forensic case studies where the deoxyribonucleic acid (DNA) of biological samples (e.g. blood or semen) collected from a crime scene or victim, is compared to the DNA profile of the suspect. Matching sample and suspect profiles does not prove a common source or guilt, but is a major contribution to the evidence (Shiono H *et al* 1985, Weir BS 1996, Schlaphoff TE *et al* 1993). Any biological sample containing a nucleated cell can be used as a source of DNA. These include flakes of skin, hair, drops of blood, cells in faeces and urine, skeletal bone, mummified tissue, menstrual blood stains, formalin fixed tissues, and decomposed human tissue (Lassen C *et al* 1994, Sasaki M *et al* 1997, Legrand B *et al* 2002, Schneider PM 1997, Hoff-Olsen P *et al* 1999).

Many other human genetic markers have since been developed and applied. The method used to establish paternity has been based on the analysis of gene products i.e. blood group antigens, polymorphic serum proteins, red cell enzymes and the human leukocyte antigens (HLA) (Shiono H *et al* 1985, Schlaphoff TE *et al* 1993, Helminen P *et al* 1991). These classical typing systems are relatively simple, inexpensive, and can provide valuable evidence in establishing non-paternity or excluding a criminal suspect (Jeffreys AJ 1993). Although most paternity cases are solved with these markers (falsely accused men being excluded with more than 99% accuracy) there are several drawbacks (Jeffreys AJ 1993, Helminen P *et al 1991*). Firstly, most of the markers are based on blood group substances that are not present in other body tissues and can only be used to type blood. Secondly, the markers are complex biochemical substances that are unstable and frequently deteriorate in specimens. Thirdly, apart from the HLA system, they show only modest levels of individual variation (Jeffreys AJ 1993).

The use of DNA markers has subsequently revolutionized the field of human genetic analysis and has a wide variety of applications(Schlaphoff TE *et al* 1993, Richards M 2001). Since 1980, DNA markers that distinguish one individual from another have been used (Schumm JW 1996). DNA profiling is the most novel technique used in family law and criminal matters where the identity or identification of an individual is in dispute (Singh D 1995). The sequence variation of DNA means that all individuals, except for identical twins, have unique DNA sequences. There may be, on average, a million DNA sequence differences between two unrelated people. For convenience, DNA sequence variation can be categorized in that of the individual, a family or kinship, and a particular population or community (Richards M 2001).

1.2 History of DNA

Deoxyribonucleic acid research began with Freidrich Miescher (Swiss Physician and physiological chemist), who in 1868 conducted the first chemical studies on cell nuclei. Miescher detected a phosphorus containing substance that he named nuclein (Wolfe SL 1993a). Late in the nineteenth century, a German biochemist, Altman, discovered that nucleic acids consist of a sugar molecule, phosphoric acid and several nitrogen-containing bases (Wolfe SL 1993a). The nucleic acid sugar molecules were subsequently found to be deoxyribose or ribose thus giving the two forms DNA and ribonucleic acid (RNA), respectively (Lewin B 1994a).

The concept that nucleic acid contained genetic information originated when Griffith discovered transformation in 1928 (Lewin B 1994a). The first direct evidence that DNA was the bearer of genetic information was, however, described by Oswald Avery in 1944. Avery and colleagues discovered that DNA taken from a virulent bacterial strain permanently transformed non-virulent forms into virulent forms (Lewin B 1994a, Wolfe SL 1993a). By the late 1940s and early 1950s DNA was largely accepted as the genetic molecule (Lewin B 1994a). In 1950 a biochemist, Erwin Chagraff, found that the arrangement of nitrogen bases in DNA varied considerably, but the amount of certain bases always occurred in a one-to-one ratio (http://www.accessexcellence.org/ ABC/ABC/search_for_DNA.html, http://www.pbs. org/wgbh/aso/databank/entries/ d053dn.html).

Despite proof that DNA carries genetic information from one generation to the next, the structure of DNA and the mechanism by which genetic information is passed on to the next generation remained unanswered until 1953. In that year Watson and Crick were able to demonstrate the double helix model of the DNA structure (Mueller RF & Young ID 2001a). Their outstanding work was immediately accepted and has proven to be the key to molecular biology and modern biotechnology (Muller RF & Young ID 2001b, Wolfe SL 1993a).

1.3 What is DNA

The human body is composed of trillions of cells, each one of which (with exception of red blood cells) contains a full set of chromosomes inside the nucleus (Jefferys AJ 1993). Every cell in the body is derived from the initial cell formed by the fusion of egg and sperm, and each contains copies of the chromosomes inherited from the mother and father (Richards M 2001). There are 46 chromosomes per cell, 23 from each parent (Mueller RF & Young ID 2001c, Brooks MA 1993). The chromosomes are made up of a tightly coiled DNA molecule and associated proteins (Mueller RF & Young ID 2001c). These are commonly referred to as the genetic material (Wolfe SL 1993c, Richards M 2001) or the genome (Fowler JCS *et al* 1988).

DNA is an enormously long thin molecule that carries the inherited information required for the development of an individual (Richards M 2001, Ross DW 1996, Lewin B 1994b). Each DNA strand consists of a chain of four different chemical building blocks or bases, with the genetic information being stored in the precise chemical sequence of bases along the DNA strand. The human genome contains approximately 6 x 10^9 nucleotides per diploid genome. This forms the book of life, a complete set of instructions for a human being (Richards M 2001, Fowler JSC *et al* 1988). The nucleotides are distributed unequally between the 23 pairs of chromosomes. Each chromosome consists of 2 long linear polynucleotides bonded via specific hydrogen bonds and coiled as a double helix (Fowler JSC *et al* 1988). The entire complement of chromosomes in a human cell comprises about less than half a millimeter, however, if fully extended the total length of DNA contained in the nucleus of each cell would be several meters long (Mueller RF & Young ID RF 2001c).

The DNA molecule consists of the so-called Watson – Crick double helix with two complementary strands, and it can replicate by separation of these strands and synthesis of the complementary strand to produce two identical copies of the double helix, thereby ensuring that the genetic material can be inherited from cell to cell and generation to generation (Mueller RF & Young ID 2001c, Fowler JCS *et al* 1988).

The four different bases that form DNA are the purines, adenine (A) and guanine (G) and the pyrimidines, thymine (T) and cytosine (C) (Richards M 2001, Ross DW 1996). During DNA replication, special enzymes move up along the DNA ladder, unzipping the molecule as it moves along. New nucleotides move into each side of the unzipped ladder. The bases on these nucleotides are very particular and cytosine will only bind to guanine, and adenine to thymine (Ross DW 1996). The sequence of the bases in the DNA is what determines the genetic code (Mueller RF & Young ID 2001c). The genetic material of all known organisms and many viruses is DNA (Lewin B 1994a).

In the 1960s it was shown that large proportions of eukaryotic DNA are composed of repeated sequences that do not encode proteins. Long non-coding sequences or intergenic regions separate relatively infrequent islands of genes (Mueller RF & Young ID 2001c, Fowler JCS *et al* 1988). A gene is organized into segments called exons, which are separated by introns. The exons contain the

DNA sequences, which are transcribed into messenger RNA that are translated into proteins (Ross DW 1996). The numerous non-coding sequences, introns, are also found within genes, interrupting the protein-coding regions, or exons. The structure and/or enzymatic activity of each protein follows from its primary sequence of amino acids. By determining the sequence of amino acids in each protein, the gene carries all the information needed to specify an active polypeptide chain. In this way, a single type of structure, the gene, is able to represent itself in innumerable polypeptide forms (Lewin B 1994b).

The DNA of genes and all other functional and non-functional sequence elements make up the genome of an organism (Wolfe SL 1993c). It is estimated that there are up to 80000 genes in the human genome. The distribution of these genes varies greatly between different chromosomes and certain parts of the chromosomes, with the majority being located in subtelomeric regions (Mueller RF & Young ID 2001c). Most of these genes are unique single copy genes that specify the sequence of amino acids in the synthesis of proteins that are involved in a variety of cellular functions (Mueller RF & Young ID 2001c, Richards M 2001). Among the genes are those encoding mRNA, rRNA and tRNA. Other functional sequences occur as regulatory, spacing or recognition elements and as replication origins. Many genes known as multigene families have similar functions. Some are found close together in clusters, while others are widely dispersed throughout the genome occurring on different chromosomes. The remaining human genes are classical gene families and gene super families. All these genes make up about one-quarter to one-third of the human genome (Wolfe SL 1993c, Mueller RF & Young ID 2001b).

About 40% of the human genome is made up of repetitive DNA sequences, which are predominantly transcriptionally inactive. Much of this nonfunctional DNA consists of repetitive sequences; relatively short elements repeated thousands or even a million times. Repetitive sequences inflate the genomes of many eukaryotes well beyond the amount of DNA needed for coding, regulation and replication (Wolfe SL 1993c, Mueller RF & Young ID 2001c). Most are not transcribed, and a few that are, are not translated. Such sequences may exist either as a single copy, acting as spacer DNA between coding regions of the genome or exist in multiple copies, hence called repetitive DNA (Fowler JCS *et al* 1988).

Repetitive DNA sequences are further classified as highly repeated interspersed repetitive DNA sequences and tandemly repeated DNA sequences. The latter can be divided into three sub-groups: satellite, minisatellite and microsatellite DNA (Mueller RF & Young ID 2001c, Lewin B 1994b).

1. 4 Short tandem repeats

Satellite DNA occurs in all animals and plants, except lower fungi, and consists of short (from several base pairs (bp) to several thousand bp in length), tandemly arranged repeats of simple DNA located in hetrochromatic chromosome regions, which are usually not transcribed (Rieger R *et al* 1991). They neither direct functional RNA nor protein products (Holt LC *et al* 2000). The existence of repetitive non-coding sequences in eukaryotic genomes first came to light during the 1960s, when Britten and Kohne developed a method; now know as reassociation kinetics. Their research showed that all eukaryotes have three classes of DNA sequence elements: unique sequences occurring in only one copy, moderately repetitive sequences in copies from a few to 100 000, and highly repetitive sequences in hundreds of thousands to millions of copies. The presence of these repetitive elements was later confirmed by DNA sequence studies (Wolfe SL 1993a). Tandemly repeated sequences from 2 bp to kilo bases in length have been observed to exhibit high variability in the number of tandem copies of the repeated motif and have been given different names (Edwards AL *et al* 1992). Variation in the number of repeats within a block of tandem repeats appears to be a universal feature of eukaryotic DNA, regardless of the length of the repeat unit (Weber LJ & May PE 1989).

Alpha satellite DNA tandem repeats of 171 bp sequences that extend to several million bp or more in length is found near the chromosome's centromere (Jorde LB *et al* 2000). One class of satellite DNA in the vertebrate genome, the minisatellite sequence, represents many dispersed arrays of short (10-50 bp) tandem direct repeat motifs that contain variants of a common core sequence. They exhibit a high degree of length variability, probably due to changes in the copy number of tandem repeats (Rieger R *et al* 1991, Jorde LB *et al* 2000). These motifs are also referred to as variable number of tandem repeats (VNTRs) (Edwards AL *et al* 1992).

1.4.1 Definition of short tandem repeats

Short tandem repeat (STR) loci are polymorphic loci found throughout all eukaryotic genomes (Thomson JA *et al* 1999, Klintschare M *et al* 1998). They are also referred to as microsatellites or simple sequence repeats (SSRs) (Butler JM & Becker CH 2001, Rieger R *et al* 1991, Edwards AL *et al* 1992). Characteristically STRs consist of tandem arrays of short repeated sequences with core repeats of 2 to 6 bps in length (Thomson JA *et al* 1999, Butler JM & Becker CH 2001, Klintschare M *et al* 1998, Jorde LB *et al* 2000).

Short tandem repeat loci are widely distributed throughout the human genome, occurring with a frequency of 1 locus every 6-10Kb (Barber MD *et al* 1996, Amar A *et al* 1999). Many have been shown to be polymorphic; with alleles differing in the number of repeat units and in some cases their base sequence (Barber MD *et al* 1996).

The number of repeats can vary from 3 or 4 to more than 50 repeats with extremely polymorphic markers. The number of repeats and hence the size of the polymerase chain reaction (PCR) product, may vary among samples in a population, making STR markers useful in identity testing or genetic mapping studies (Butler JM & Becker CH 2001). Short tandem repeat alleles are small in size, generally less than 350 base pairs (Klintschare M *et al* 1998, Jorde LB *et al*

2000, Amar A *et al* 1999). Their polymorphic nature and accessibility to amplification using PCR, by making use of flanking sequence primers, has led to their introduction into forensic identity testing (Amar A *et al* 1999, Barber MD *et al* 1996).

1.4.2 Di-, tri- and tetranucleotide tandem repeats

In the human genome there are 50 000 – 100 000 interspersed (CA)n blocks with n ranging roughly between 10 and 60. They are referred to as hypervariable microsatellites (Litt M & Luty JA 1989, Weber LJ & May PE 1989). The dinucleotide tandem repeat blocks are uniformly spaced throughout the genome at every 30-60Kb.

The functions of the blocks are unknown, but may serve as hot spots for recombination or participate in gene regulation. Co-dominant Mendelian inheritance of these fragments has been observed (Litt M & Luty JA 1989). Dinucleotide tandem repeats are located within protein-coding regions; most are found within introns or between genes (Weber LJ & May PE 1989). These repeats have been found in several sequenced regions, including the β-globin gene cluster, the cardiac actin gene and the somatostatin gene (Litt M & Luty JA 1989). However, because of problems caused by shadow bands when analyzing dinucleotide repeats, the less common tri-, tetra- and penta-nucleotide repeats are preferred for personal identification (Urquhart A *et al* 1994, Weber LJ & May PE 1989).

It was hypothesized that there are approximately 400 million trimeric and tetrameric STR loci interspersed throughout the genome of which a high proportion are polymorphic (Van Oorschot RAH *et al* 1994). Examples of trinucleotides are HUMFABP intestinal fatty acid binding protein 4q31 and HUMARA androgen receptor Xcen q13 (Edwards AL *et al* 1992). The tetranucleotide STRs include HUMTHO1, HUMRENA, HUMHPRTB (Edwards AL *et al* 1992), TOPOX and CSF1PO (Amar A *et al* 1999).

1.4.3 Origin/Formation of STRs

For reasons that are not yet understood, the number of repeats can increase dramatically during meiosis or possibly during early fetal development (expanded repeat) (Jorde LB *et al* 2000). Very little is known about the mutation mechanism, but mutational behavior is probably locus dependent. As observed in RFLP and other STR systems, repeat mutations are often of paternal origin, correlating with the fact that at least 10 cell divisions or more occur between the zygote and sperm than between the zygote and ovum. This also illustrates that mutations tend to generate larger alleles. However, mutations of maternal origin and reduction in length have been reported. Mutation mechanisms can be sex-dependent as observed during the formation of disease-related deletions and duplications. The sequence of the repeat unit does not seem to be the primary factor of the polymorphism and thus of the mutation mechanism. Tandem reiteration, regardless of the repeat sequence, probably induces variation but is not the exclusive factor. Another factor could be the sequence surrounding the repeat (Mertens B *et al* 1999).

Duplication of entire repeats is important in the origin and early evolution of microsatellites. The rarity of repeat length polymorphism in microsatellites with few repeats does not refute slippage; it only shows that the rate is lower than the high rates that characterize longer microsatellites (Zhu Y *et al* 2000).

In an approximate state of linkage equilibrium, alleles at different loci segregate independently. Principles of gene behavior predict such inheritance of STR loci that are physically separated on different chromosomes or spatially separated along a single chromosome (Holt CL *et al* 2000).

Studies of microsatellite mutation and evolution have focused on established microsatellites with multiple repeats. The number of repeats usually increases or decreases by a single repeat unit. The mechanism appears to involve slippage during DNA replication. Most insertions are duplication of adjacent bases, but some are of already existing short repeat sequence of 2-4 units. Insertions are generally copies of adjacent sequences, and generate short microsatellites. New proto-microsatellites are also generated by substitutions. Though insertion occurs less frequently than substitutions, the relative importance in generating new repeats rapidly increases with the length of the repeat. (Zhu Y *et al* 2000).

The process that leads to expansion and polymorphism at established microsatellite loci also occurs in areas with few or no repeats. The mechanism is not clear. Slippage is generally thought to require repeats, with repeats in the new strand mispairing with other repeats on the template during DNA replication, but this is not possible in the absence of repeats or symmetric elements. It has been suggested that there might be a minimum number of repeats that must be generated by substitution before expansion by slippage can occur (Zhu Y *et al* 2000).

1.4.4 Intermediate alleles and/or microvariants

Sequencing revealed that intermediate alleles were due to a deletion some 50 nucleotides away from the repeat sequence. This observation raises a question whether the generation of intermediate alleles involves a dinucleotide in the imperfect repeat region reflecting instability of this region (Mertens B et *al* 1999).

1.4.5 Disease Association

Short tandem repeats have long been considered neutral elements devoid of biological effect (Albanese V *et al* 2001, Holt CL *et al* 2000). However, several studies suggest that repeated sequences might have a function in recombination, in generating nucleosome positioning signals and in transcription (Albanese V *et al* 2001). It was ducumented that some genetic diseases are caused when mutation increases the number of tandem repeats occurring within or near the disease genes. More than a dozen genetic diseases caused by expanded repeats are known (Jorde LB *et al* 2000). For example, abnormal expansions of trinucleotide repeats in non-coding sequences interfere with normal transcriptional activity and are responsible for several human neurological diseases. Repeated sequences may not only be associated with pathological expansions of unstable DNA stretches causing Mendelian diseases, but they may also have more subtle effects on gene expression. It was recently demonstrated that a tetranucleotide repeat, HUMTHO1 microsatellite, in the first intron of the tyrosine hydroxylase (TH) gene, acts as a transcriptional enhancer in vivo (Albanese V *et al* 2001).

Since STRs are known to be unstable in various tumor tissues, they can be used to study genetic/allelic alterations in tumors (Rubocki RJ *et al* 2000, Berger AP *et al* 2002). A partial or complete allelic deletion common to many types of cancer is referred to as the loss of heterozygosity (LOH) (Goumenou AG *et al* 2001, Rubocki RJ *et al* 2000). Numerous examples of LOH in cancer have been described and some have been mapped to areas located in close proximity to markers employed in human identity testing (Rubocki RJ *et al* 2000, Kok K *et al* 2000, Tsuneizumi M *et al* 2002, Harn H-J *et al* 2002). Despite this fact, LOH has rarely been observed for STR loci commonly employed in forensic testing. As demonstrated in other cancers, cancerous biopsies showed LOH at one STR locus (Rubocki RJ *et al* 2000). However, different STR loci exhibiting a significant mutation rate due to their different structural influences and length of the tandem repeat were reported. Alleles 17 and 18 of the vWA locus and alleles 22 to 26 of the FGA locus were found to be more susceptible to mutations/alterations (Pai C-Y *et al* 2002). The other allelic alteration is microsatellite instability, which was defined in tumor tissue that showed banding pattern alteration at two or more microsatellite loci (Harn H-J *et al* 2002). The microsatellite instability phenomenon may be caused by mutator mutations that occur in DNA mismatch repair genes (Limpaiboon T *et al* 2002). Microsatellite instability has been reported in a number of cancers (Limpaiboon T *et al* 2002, Peir G *et al* 2002, Harn H-J *et al* 2002).

1.4.6 Types of STRs/ Classification

Short tandem repeat markers are plentiful, more than two thousand STRs suitable for genetic mapping studies have been described (Murray JC *et al* 1994, The Utah Marker Development Group 1995). Of these only a limited number are used in forensic and paternity analyses (Schumm JW 1996). Simple repeats contain units of identical length and sequence (Seidl C *et al* 1999, Urquhart A *et al* 1994, Watson S *et al* 2001), and show constant basic structures and low mutation rates. Nevertheless, higher discrimination rates and exclusion probabilities can be achieved with compound or extremely complex STRs, which are much more variable and show higher mutation rates than simple polymorphic STR regions (Golck B *et al* 1997). Compound repeats comprise two or more adjacent simple repeats (Seidl C *et al* 1999, Urquhart A *et al* 1994, Watson S *et al* 2001), while complex repeats may contain several repeat blocks of variable length (Seidl C *et al* 1999, Urquhart A *et al* 1994, Watson S *et al* 2001).

The repeat structure of alleles at STR loci vary due to:

(1) length of individual repeat units

(2) number of repeat units

(3) repeat unit pattern of the individual alleles (Seidl C *et al* 1999, Urquhart A *et al* 1994).

1.4.7 Methods of detection/Test systems available

A variety of test systems have been developed that enable detection of STRs either individually (Watson S *et al* 2001), or in multiplex (Watson S *et al* 2001, Rubocki JR 2000). The polymorphic variation in allele length had previously been detected by slab gel electrophoresis with silver staining (Yoshimoto T *et al* 2001) and later with multiple color fluorescent detection. More recently, capillary electrophoresis was used to resolve and type STR alleles (Butler JM *et al* 1998). A wide range of electrophoretic systems is utilized (Gomez J & Carracedo A 2000). These include the use of 4-6% denaturing polyacrylamide gels and

Metaphor agarose gels (Gill P *et al* 1994, Gomez J & Carracedo A 2000, Watson S *et al* 2001).

Agarose gels can differ in concentration, thickness, ladders, electrophoretic and temperature conditions, and different running distances and times. A variety of detection methods are also used, the more common ones include ethidium bromide, silver staining, radio labeled primer incorporation followed by autoradiography (isotopic method), and fluorescent-labeled primer incorporation detected by laser excitation on automated sequencers (Gill P *et al* 1994, Gomez J & Carracedo A 2000, Watson S *et al* 2001). Sizing of fragments is carried out using a variety of manual and automated methods (Gomez J & Carracedo A 2000).

To reduce analysis cost and sample consumption, and to meet the demands of higher sample outputs, PCR amplification and detection of multiple markers (multiplex STR analysis) has become the standard technique in most forensic DNA laboratories. Short tandem repeat multiplexing is most commonly performed using spectrally distinguishable fluorescent tags and/or non-overlapping PCR product sizes. When using commercial kits, the STR alleles from multiplexed PCR products typically range from 100 - 350 bp (Butler JM & Becker CH 2001).

Multiplex amplification and automated, objective genotyping of 13 core STR loci is used in the CODIS system. Allele designation, even within overlapping locus size ranges, can easily be accomplished by exploitation of simultaneous multicolor fluorescence detection in a single gel lane or capillary injection and comparison to an allelic ladder designed for each kit (Holt CL *et al* 2000, Watson S *et al* 2001). The separation, detection and analysis of STR products can be semi-automated by the use of automated DNA sequencers and specialized software (Watson S *et al* 2001). Internal size standards are included with every sample to allow automatic sizing of alleles and to normalize differences in electrophoretic mobility between gel lanes or capillary injections (Holt CL *et al* 2000).

1.4.8 Uses of STR

Short tandem repeats have been studied extensively and applied in different areas including basic genetic research (Rubocki JR 2000), physical and genetic mapping of the human genome (Edwards AL *et al* 1992, Rubocki RJ *et al* 2000), personal identification in medical and forensic sciences (Edwards AL *et al* 1992), and to study genetic variation in distinct ethnic groups (Amar A *et al* 1999). Differences in allele proportions between ethnic groups were analyzed to form the basis of an ethnic inference system. Information of an offender's ethnicity may assist an investigation and also priorities for population mass screening are set (Lowe AL *et al* 2001). It is also applied in disease diagnosis (Edwards AL *et al* 1992), and the study of genetic alterations in tumors (Rubocki RJ *et al* 2000).

1.4.9 Advantages of using STRs

Short tandem repeat analysis is dependent on PCR, which is a very sensitive technique. As little as 1ng of genomic DNA will yield a full STR profile, whereas single locus polymorphism (SLP) analysis requires at least 100ng for reliable profiling. A testing method only requiring blood drops from finger or heel pricks or buccal swabs taken onto paper stain cards can offer significant benefits in terms of sample taking ease, transportation and storage (Thomson JA *et al* 1999, Wiegand P *et al* 2000).

Short tandem repeat analysis is less time consuming (Thomson JA *et al* 1999, Klintschare M *et al* 1998), allows simultaneous analysis of several STR loci (Thomson JA *et al* 1999), is more amenable to automation (Thomson JA *et al* 1999) and robust amplification (Rubocki JR 2000), lowers the amount of stutter produced during PCR, and does not complicate DNA mixture interpretation (Butler JM & Becker CH *et al* 2001).

Short tandem repeats display high levels of heterozygosity and polymorphism (Rubocki JR 2000), and exhibit fewer variants (Van Oorschot RAH *et al* 1994). Due to their small fragment length (usually shorter than 300 bp) small amounts of possibly degraded template DNA can be amplified by PCR (Golck B *et al* 1997, Klintschare M *et al* 1998, Van Oorschot RAH *et al* 1994). The PCR products do not have the problem of unequal amplification among alleles (i.e. dropout of large alleles) (Van Oorschot RAH *et al* 1994). Detected alleles may differ in length by a single base pair. These can be accurately identified and assigned allelic designation, allowing results to be easily compared among laboratories (Rubocki RJ *et al* 2000).

In general, PCR based STR multiplex analysis offers the advantage of increased sensitivity, improved speed of analysis and lower cost compared with conventional SLP DNA profiling techniques (Van Oorschot RAH *et al* 1994).

1.4.10 Disadvantages of using STRs

Most STRs are distinctly less polymorphic than VNTRs, with only 3-6 common alleles; therefore a large number of systems have to be typed for comparable results (Klintschare M *et al* 1998). Mutation also affects tandem repeated DNA sequences that occur within or near certain disease genes (Jorde LB *et al* 2000, Han G-R *et al* 2001*)*. There have been reports of failure to amplify STR loci, i.e. a report on the amelogenin locus in male individuals, which could not be amplified. This was ascribed to a deletion of the locus itself (Steinlechner M *et al* 2002, Thangaraj K *et al* 2002).

1.4.11 Characteristics for human identification

The continuing development and validation of STR systems in identity testing have resulted in 20 or more suitable STR systems that are available commercially or as published primer sequences (Thomson JA *et al* 1999). These STR systems have a high degree of polymorphism/variability within human populations (Lazaruk K *et al* 2001). The intermediate number of alleles (5-15 common alleles per locus) keeps the locus size range small enough for multi locus PCR amplification, and also minimizes preferential amplification. Multiple alleles per locus can also translate into a relatively higher power of discrimination at a single locus, making interpretation of mixed DNA samples practical (Lazaruk K *et al* 2001).

Other characteristics required of a good STR system are that the amplification products must be easily distinguished from one another (Schumm JW 1996), they should be amenable to PCR analysis which allows for minute and/or degraded DNA sample analysis (Lazaruk K *et al* 2001), and have a low prevalence of stutter bands (Schumm JW 1996).

1.4.12 Allele designation

Allele designation of STR PCR products depends on accurate sizing. A DNA digest labeled with the fluorescent dyes ROX or LIZ, sizes alleles precisely but not accurately. Sizing of alleles that differ by only 1 bp cannot be performed without the use of an allelic ladder (Urquhart A *et al* 1994). The quantitative nucleotide length differences in the amplified DNA fragments are the basis for allele designation. Genotyping is accomplished by comparing the unknown samples to an allelic ladder and the use of software that allows for accurate and efficient typing of the samples. Sequence variation within a tetranucleotide repeat is not detected by the methods used for fluorescent STR genotyping and, therefore, not categorized in the allele frequency estimates used for forensic testing. Even when sequence variation exists in the core STR repeat unit or the flanking sequence, genotypes can still be assigned (Lazaruk K *et al* 2001). Accurate typing of STRs requires a precise knowledge of the structural variation of alleles (Dauber EM *et al* 2000).

Allele designation for complex repeats is more problematic as each allele contains a mixture of di-, tri-, tetra-, penta- and hexanucleotides. Three options are considered: naming alleles by length in base pairs, using an arbitrary system of allele designation or naming alleles by the number of TV dinucleotide (tetranucleotide excluding invariant di-and/or trinucleotide) repeats (Urquhart A *et al* 1994).

1.4.13 Short tandem repeat nomenclature

A gene is a DNA segment that contributes to a phenotype or function. In the absence of demonstrable function, sequence, transcription or homology may characterize a gene. A locus is not a synonym for gene, but it is a specific place in the genome, identified by a marker, which can be mapped by some means. It could be an anonymous non-coding DNA fragment or a cytogenetic feature. A single gene may have several loci within it and these markers may be separated in genetic or physical mapping experiments. In such cases, it is useful to define these as different loci, but normally the gene name should be used to designate the gene itself, as this usually will convey the most information (Wain HM *et al* 2002, Blake JA *et al* 1997).

Based upon this recommendation, almost all currently used STR loci are either named according to the gene name or the DNA segment in which they are located (White JA *et al* 1997). Short tandem repeat systems that are located within a gene (intronic loci) retain the gene name: e.g. vWA (von Willebrand factor gene), FGA (alpha fibrinogen gene) and TPOX (thyroid peroxidase gene). The STR loci in non-geneic DNA segments are designated differently. Examples are D8S1179, D18S51 and D21S11 (Gill P *et al* 1997c, Schumm JW 1996). These symbols can be obtained from the Genome Database (GDB) and are assigned automatically to arbitrary DNA fragments and loci. These symbols comprise five parts described by the following guidelines:

(1) D for DNA

(2) 0,1,2...22, X, Y, XY for chromosomal assignment, where XY is for segments homologous on the X and Y-chromosomes, and 0 is for unknown chromosomal assignment.

(3) S, Z or F indicating the complexity of the DNA segments detected by the probe; with S for unique DNA segment, Z for repetitive DNA segment found at a single chromosome site and F for small, undefined families of homologous sequence found on multiple chromosomes.

(4) 1,2,3,... a sequential number to give uniqueness to the above concatenated characters.

(5) When the DNA segment is known to be an expressed sequence, the suffix E can be added to indicate this (Wain HM 2002, White JA *et al* 1997).

Whether the STR is intronic or a segment in a non-geneic area, allele nomenclature is according to the recommendations of the International Society of Forensic Haemogenetics (IFSH) (Report, DNA recommendations 1991 & 1994, Gill P *et al* 1994, Wain HM 2002, White JA *et al* 1997). However, allele designation for some complex repeats is more problematic (Urquhart A *et al* 1994). The most widely applicable allele designation and nomenclature would be to call each allele by its length in base pairs. This method would be suitable for VNTRs, normal STR and hypervariable STRs. The allele size is dependent on the primers used, and requires a precise and accurate sizing method. An alternative is to call alleles by the number of repeat units they contain. This is easy for simple repeats and some VNTRs, and can be applied to compound repeats with the use of ambiguity codes, but it is too cumbersome for complex repeats. A problem also occurs when intermediate alleles have to be described (Urquhart A *et al* 1994). Intermediate alleles are those alleles that do not have en exact number of units and thus consist of a certain number of units with the addition of one, two or three bases.

Nomenclature of simple repeats is straightforward. The notation is based upon the number of tandem repeats in the STR. The same principle applies for simple with non-consensus repeats, but if there is a variant, then its notation is based upon the number of complete repeats followed by a decimal point and the number of bases comprising the partial repeat (Holt CL *et al* 2000). The above notation also works well for compound repeat sequences. In the complex repeats two different nomenclatures have been proposed. The Moller notation is based on the number of complete tetramers, ignoring the invariant non-tetramers, and the Urquhart notation that is based on the number of dimers present and includes the invariant trimer as one repeat. These allele designations are directly comparable and can easily be inter-converted. The method of Moller is suggested for general use, since it is closer to the ISFH DNA Commission recommendations (Gill P *et al* 1997a, Urquhart A *et al* 1994).

In line with the recommendations of the IFSH DNA Commission (DNA recommendations 1991), alleles at all simple and compound repeat loci are called by their repeat number, using redundancy codes for compound repeats (M⇒ A or C, Y⇒ C or T, K⇒G or T, R⇒A or G, V⇒ A, C or G). For intermediate alleles and other alleles that fail to align with the incremental ladder of each locus, digits after a decimal are used to indicate the number of base pairs by which the allele exceeded the previous rung of the ladder. The use of the number after the decimal point does not necessarily imply the presence of a partial repeat, but may indicate variation outside the repeat region (Urquhart A *et al* 1994). Alleles with the same assignment may actually vary in their sequences and in the actual number of repeats due to insertions or deletions of repeats or the flanking regions, but this should not present a problem either for data base use or identity testing.

In modern instruments an internal lane size standard is included with every sample to allow automatic sizing of alleles and to normalize differences in electrophoretic mobility between gel lanes or capillary injection. Data is collected and analyzed using different software versions. Allelic designation is automatically assigned by comparison between sample alleles and allelic ladder alleles run on the same gel or set of injections (Holt CL *et al* 2000). Allelic ladders should be used for all STR systems detected by manual electrophoretic systems and should be based on the predominant simple repeat motif of the system in question. All the commonly occurring alleles should be present in the ladder. All alleles in an allelic ladder should be sequenced to establish the sequence of the repeat unit(s), the number of repeats present, and the actual size of the allelic fragment (DNA recommendations 1994).

Some analytical systems do not require an allelic ladder as a reference for allele typing, but internal standards within the same electrophoretic lane as the sample being tested. The alleles are characterized by their fragment size in base pairs but should be converted to the aforementioned allele designation protocol. If an allelic ladder is labeled, it should be consistent with the labeled primer used to amplify the STR alleles (DNA recommendations 1994).

1.5 The FGA locus

The human alpha fibrinogen locus (FGA) is widely used in forensic DNA testing, for individualization of biological stains as well as in paternity investigations (Neuhber F *et al* 1998, Dauber EM *et al* 2000, Gill P *et al* 1997b). This locus is also known as HUMFIBRA (Gill P *et al* 1997b) and HUMFGA (Dauber EM *et al* 2000). The FGA locus is found on the long arm of chromosome 4 and is located in the third intron of the human alpha fibrinogen gene that contains repeats beginning at nucleotide 2912 (Mills K *et al* 1992, Dauber EM et al 2000, Barber MD *et al* 1996), and it is inherited co-dominantly (Millis KA *et al* 1992). It is a complex tetranucleotide repeat with the common alleles differing in length by 4 bp, but also containing interalleles differing by 2 bp from the main alleles. Additionally, alleles that differ in 1bp from the common alleles have been reported (Neuhuber F *et al* 1998, Lazaruk K *et al* 2001). The GeneBank strand is $[TTTC]_3$ TTTT TTCT $[CTTT]_n$ CTCC $[TTCC]_2$ (http://www.cstl.nist.gov/viotech/strbase/str_fga.html, Holt CL *et al* 2000). The reported mutation rate for the FGA locus is $6x10^{-3}$ (1 in 162 meiosis) (Thomson JA *et al* 1999).

The FGA locus is among one of the loci selected for the US Combined DNA Index System (CODIS) and typed in crime laboratories throughout the world

(Klintschar M *et al* 1999, Butler J-M & Becker CH 2001). This locus has been analyzed in a number of systems; either in a single reaction (Gill P *et al* 1997b, Neuhuber F *et al* 1998, Dauber E M *et al* 2000), where it is separately amplified with specific primers in a single tube or in multiplex. Diverse multiplex systems that contain FGA were developed and are used in different laboratories for various purposes. Some of these are: AmpFlSTR profiler (Lazaruk K *et al* 1998, Pu C-E *et al* 1999, Trivedi R *et al* 2000), AmpFlSTR blue (Budowle B *et al* 1997, Holt CL *et al* 2000), AmpFlSTR profiler plus (Gamero JJ *et al* 2000, Geada H *et al* 2000, Tahir MA *et al* 2000), AmpFlSTR cofiler (Bosch E *et al* 2001, Budowle B *et al* 2002), Powerplex (Thomson JA *et al* 1999, Ashma R & Kashyap VK 2002) and AmpFlSTR SGM plus (Thomson JA *et al* 1999, Walsh SJ *et al* 2001) of the PE Applied Biosystems and Promega companies. A SGM system of the forensic science service in the UK (Walsh SJ *et al* 2001, Thomson JA *et al* 1999), as well as Genetrace has been developed for mass spectrometry (Butler JM & Becker CH 2001). These multiplex systems amplify from 3 up to 16 STR loci simultaneously in a single or very few PCR reactions.

We have compiled 29 published and unpublished population frequency reports. The smallest size of a study sub-population group was 33 from the Himalayan Ladakh Dropka population in India (Trivedi R *et al* 2002), and the largest was 6037 from the Eastern Polynesian population of New Zealand (Walsh SJ *et al* 2001). Of these, only two studies had less than 100 individuals in the study population (Trivedi R *et al* 2002, Bosch E *et al* 2001). From these published and unpublished FGA population frequency reports representing 53 population groups and sub-groups, the number of alleles and interalleles reported ranged from 8 in the Ladakh Argon Himalayan Indian population (Trivedi R *et al* 2002) to 32 in the black population residing in the Free State, South Africa (de Kock A 2002, personal communication). Furthermore, 16 alleles and inter alleles each were reported from 11 population groups (table1.1.).

Table 1.1. Distribution of FGA alleles among population groups and sub-groups compiled from published and unpublished data.

Ser no.	No. of Alleles(n)	Allele range	Population Groups or sub-groups	Reference
1	8 (2)	18-27; 19-25	Ladakh, Argon Himalayan Indian Ladakh Dropka, Himalayan Indian	Trivedi R *et al* 2002
2	9(2)	19-27	Kurmi, Bihar, India; Yupik, Native Alaska	Ashma R *et al* 2002; Budowle B *et al* 2002

3	10(4)	17-26;
		19-28;
		19-29;
		17-27
			Henan, Chinese;
			Moroccan Arabs;
			Mozabites, Algeria;
			Baniya, Bihar, India;
				Si Y *et al* 2002
				Bosch E *et al* 2001
				Ashma R & Kashyap VK 2002

4	11(4)	18-28;
		17-29;
		18-27;
		<18-28
			Canary Islands, Spanish; Southern Moroccan Barbers;
			Athabaskan , Native Alaska ;
			Inupiat, Native Alaska
				Gamero JJ *et al* 2000
				Bosch E *et al* 2001
				Budowle B *et al* 2002

5	12(3)	18-29;
		16-27;
		18-27
			Saharawis,NorthWestern Africa;
			Central Moroccan Barbers;
			US Caucasian
				Bosch E *et al* 2001
				Holt CL *et al* 2000

6	13(4)	18-28:
		19-28;
		10-27;
		17-29
			Portuguese;
			Ladakh Balti. Himalayan Indian;
			Tuscany, Central Italy;
			Asian, South Africa
				Geada E *et al* 2000
				Trivedi R *et al* 2002
				Ricci U *et al* 2002
				Police Service, unpublished doc.

7	14(3)	17-27;
		18-27;
		18-28
			Flemish population;
			Yadar, Bihar India;
			Spanish
				Van Hoofstat DEO *et al* 2002
				Ashma R & Kashyap VK 2002
				Entrala C *et al* 1998

8	15(7)	18-27;
		18-31;
		16-28;
		18-27;
		17-28;
		18-28
			Thailand;
			Egypt;
			Western Polynesian, New Zealand;
			Central Poland;
			Austrian;
			Chinese; Italians
				Pu C-E *et al* 1999;
				Klintschar M *et al* 1999;
				Walsh SJ *et al* 2001;
				Kuzniar P *et al* 2002;
				Neuhuber F *et al* 1998;
				Fung WK *et al* 2001; Garofano L *et al* 1998

9	16(11)	17-28;
		<18-30;
		18-28;
		16-28;
		17-29;
		17-27;
		18-32
			Taiwan; Philippine; Omani, Taiwani (Chinese);
			Black African, Zimbabwe;
			Buddhist Himalayan Indian; South East Asian descent, New Zealnd;
			Caucasian;
			Asian; Italians;
			Caucasian; South Africa;
			Brazilian
				Pu C-E *et al* 1998; Tahir MA *et al* 2000; Klintschar M *et al* 1999;
				Budowle B *et al* 1997;
				Trivedi R *et al* 2002; Walsh SJ *et al* 2001;
				Thomson JA *et al* 1999;
				Thomson JA *et al* 1999; Biondo R *et al* 2001;
				Police Service, unpublished doc;
				Grattapaglia D *et al* 2001

10	17(3)	16-28;
		16.2-29;
		16-27
			Eastern Polynesian;
			Afro-American;
			Austrian
				Walsh SJ *et al* 2001;
				Holt CL *et al* 2000;
				Dauber EM *et al* 2000

11	18(2)	18-32; 17-46.2	Black immigrant Spanish; Afro-Caribean	Gamero JJ *et al* 2000; Thomson JA *et al* 1999

12	19(3)	17.2-32.2
		17-41;
		17-28
			Whites, Free State, South Africa;
			Coloured, Free State, South Africa;
			Thai
				De Kock A, personal communication
				Rerkamnuaychoke B *et al* 2001

13	21(1)	15-29	Caucasian descent, New Zealand	Walsh SJ *et al* 2001
14	22(1)	16.2-46.2	Black, South Africa	Police Service, unpublished doc.
15	25(1)	16-46.2	Coloured, South Africa	Police Service, unpublished doc.
16	32(1)	16-47.1	Black, Free State, South Africa	De Kock A, personal communication

(n)= Number of population group with the same number of alleles reported in the population.

Various published and unpublished population FGA frequency reports, representing different population groups of the world, reported a total of 86 different alleles. Of these 24 were complete tetranucleotide repeats, while the remaining 62 were interalleles that vary in 1, 2 or 3 nucleotides from the complete tetranucleotide repeat. The size of the complete tetranucleotide alleles range from 10 (Ricci U *et al* 2002) to 44 (http://www.cstl.nist.gov/viotech/ strbase/var_fga.html). The size of the reported interalleles range from 12.2 (http://www.cstl.nist.gov/viotech/strbase/ str_fga.html) to 51.2 (Lazaruk K *et al* 2001). Of all of these alleles, alleles 22, 23, and 24 were the most common. From 29 studies representing 53 population groups and sub groups, these alleles were reported at a frequency of > 0.1000. Alleles 19, 20, 21 and 25 were also reported at frequencies of ≥ 0.05 in the majority of the reported groups. The interalleles most often reported, were 19.2, 20.2, 21.2, 22.2, 23.2, 24.2, 25.2 and 26.2. Interalleles 21.2, 22.2, 23.2 and 24.2 were reported with a frequency of ≥ 0.01 in some of the populations.

Of the 86 reported alleles and interalleles, the sequence of 44 alleles was described. Sixteen of these were complete tetranucleotides, while the remaining 28 were interalleles. Dauber EM *et al* (2000) reported 17 different alleles at the FGA locus and Barber MD *et al* (1996) reported 22 alleles ranging in size from 168 to 249 bp. Lazaruk K *et al* (2001) reported 36 alleles and 4 sequence variants at this locus. Additionally, a STR fact sheet documented 42 alleles and 1 sequence variant (http://www.cstl.nist.gov/viotech/strbase/ str_fga.html).

Eleven of the 44 alleles, that were investigated, displayed sequence variants. Alleles and interalleles in which sequence variations were found were 24, 26, 27, 28, 30, 42.2, 43.2, 44.2, 46.2, 47.2 and 50.2. All of these have two sequence variants each except allele 27, which had three reported sequence variants. Of the interalleles with 1 or 3 bp difference from the complete tetranucleotides the sequence of alleles 16.1 and 23.3 was described (Barber MD *et al* 1996, Dauber EM *et al* 2000, Lazaruk K *et al* 2001, http://www.cstl.nist.gov/biotech/strbase/ str_fga.html).

Dauber E.M *et al* (2001) reported that the larger FGA alleles in their study were exclusively found in the Afro-Caribbean population. Barber MD *et al* (1995) also showed that some alleles were exclusive to some ethnic groups.

For the measurement of the usefulness of a locus or group of loci, different statistical parameters are applicable. Observed heterozygosity was reported in 33 different population groups. Reported observed heterozygosity in the FGA locus ranged between 0.578 (Ashma R *et al* 2002) and 0.948 (Trivedi R *et al* 2002). The majority of these studies reported an observed heterozygosity of > 0.800. The power of discrimination was also reported by various studies representing 39 sub-population groups. The majority reported a high power of discrimination (0.900). The highest power of discrimination (0.9709) was reported in Zimbabwean black Africans (Budowle B *et al* 1997) and the smallest power of discrimination (0.791) in the Dropka, Himalayan Indian population (Trivedi R *et al* 2002). Probability of exclusion or prior chance of exclusion was reported in 25 different sub-populations. The highest probability of exclusion (0.772) was reported in Hungarian Caucasians (Egyed B *et al* 2000), while the smallest (0.5809) was in the Thai population (Rerkamnuaychoke B *et al* 2001). Hardy-Weinbergh equilibrium was also documented in 25 sub-population groups. The reported P values vary greatly; from 0.000 in the Canary Islands (Gamero JJ. *et al* 2000), to 0.999 in the Philippine population residing in Taiwan (Pu C-E *et al* 1999). Of the documented 10 reports of mean exclusion chance of the FGA locus, the highest value of 0.737 was reported in Egypt (Klintschar M *et al* 1999), while the smallest value of 0.701 came from Austria (Dauber EM *et al* 2000). Other statistical parameters such as polymorphic information content, typical paternity index, observed homozygosity, exact test, matching probability, and probability of identity were also reported by a few of the studies. According to all the above paremeters the FGA locus is a very useful tool in individual identification.

Primate FGA sequences were also studied by Lazaruk K *et al* (2001). According to this study the Chimpanzee, Orangutan, and Gorilla FGA alleles are all homologous to the human sequence before and after the repeat structure, but differ significantly from human and from each other in their core repeat structure. The chimpanzee FGA allele structure is the least complex and closest in structure to those of humans (Lazaruk K *et al* 2001, Levedakou EN *et al* 2002).

1.6 The Polymerase Chain Reaction

1.6.1 Introduction

The polymerase chain reaction offers a powerful approach to distinguishing individual alleles in a genome, and thus to diagnose diseases that are defined at the sequence level. If a disease is associated with a particular sequence change, PCR can be used to examine the sequence of a particular individual to determine whether the alleles are wild type or mutant. Amplification by PCR is so sensitive that the target sequence in an individual cell can be characterized, thus allowing the distribution of alleles in a population to be examined directly. It also allows DNA to be amplified from very small tissue samples, which is useful for diagnostic and forensic purposes (Lewin B 1994a).

1.6.2 Historical perspective

Cloning, DNA sequencing and PCR underlies almost all of modern molecular biology (Sambrook J & Russel DW 2001a). With the aid of computers, PCR revolutionized the study and manipulation of entire genomes (Wolfe SL 1993b).

The PCR method of DNA amplification was developed in 1983 by Dr. Kary Mullis (Carleton SM 1995, Mullis KB 1990, Rabinow P 1996, Ross DW 1996). Mullis and Faloona determined the basic characteristics of exponential amplification using a set of primers specific for a 118 bp region of the beta globin gene (Wolfe SL 1993a). The first medical application was the prenatal test for sickle cell anemia and beta thalassemia. Polymerase chain reaction technology has impacted on human genomic analysis, especially the genetic and physical chromosome mapping and gene expression analysis (Carleton SM 1995). Polymerase chain reaction technology is an essential part of every molecular biology laboratory, and is perhaps the single most important technique used in recombinant DNA analysis (Ross DW 1996).

1.6.3 The principle of DNA amplification using PCR

DNA amplification by PCR is an enzymatic reaction using DNA polymerase (Carleton SM 1995). It involves selection of a fragment of DNA, and amplifying this fragment by repetitive cycles of DNA synthesis. Thus, a particular DNA sequence of interest, among the background of the entire human genome can be amplified so that the small fragment becomes the majority of the DNA in the sample (Ross DW 1996).

In the reaction, two small oligonucleotide primers (complementary to each end of the DNA sequence of interest), an excess of free nucleotides along with DNA polymerase and buffer are added to the target DNA sample (Ross DW 1996). The length of the target sequence depends on the distance between the two primer binding sites (Lewin B 1994a).

1.6.4 Limitations of PCR

Polymerase chain reactions will continue to a certain point and then seem to stop. Like all enzymatic reactions, PCR is not an unlimited process. In most applications, after about 20 to 40 cycles, the reaction enters a linear phase where exponential accumulation of the product is attenuated. The so-called plateau effect, occurs when the product reaches about 10^{-8} M (about 10^{12} molecules in a 100 μl reaction) (Carleton SM 1995). A restriction on the sensitivity of the technique for examining individual sequences is that the replication event has an error rate of ~ 2×10^{-4}, which means that an error occurring in a very early cycle could become prominent throughout the amplification (Lewin B 1994a). A minute amount of DNA carried from previous samples is the most common contaminant that affects the sensitivity of a PCR reaction (Ross DW 1996).

1.6.5 Polymerase chain reaction set-up

In a PCR based application, success is determined by two important factors, the quality of the amplification reaction, and the accuracy of the method used to analyze the reaction products. An efficient, specific, uncontaminated, and reliable PCR amplification requires attention to a relatively large, but discrete set of important factors. Although PCR is a complex reaction, with at least 13 different components, the reaction parameters that influence its yield and efficiency can be adjusted systematically. The major controllable variables include the concentration of primers and templates, the Mg^{++} ion concentration, the concentration of dNTPs, and the annealing temperature and thermal cycling conditions (Carleton SM 1995).

1.6.6 Method of PCR product detection

Methods to detect PCR reaction products vary greatly in terms of sensitivity, specificity, difficulty, and cost. The ideal detection method should allow an accurate determination of size and purity, and when necessary, provide information about the DNA sequence of the amplified product. These techniques range from simple agarose gels to DNA sequencing (Carleton SM 1995). See section 1.4.9 for detailed detection methods.

1.7 DNA sequencing

The power of DNA sequencing is in its ability to reduce genes and genomes to chemical entities of defined structure (Sambrook J & Rusell DW 2001a). The information obtained from DNA sequencing is one of the primary sources of the molecular revolution (Wolfe SL 1993b). In molecular cloning laboratories, DNA sequencing is used chiefly to characterize newly cloned cDNAs, to confirm the identity of a clone or mutation, to check the fidelity of a newly created mutation, ligation junction, or PCR products, and in some cases, as a screening tool to identify polymorphisms and mutations in genes of particular interest (Sambrook J & Rusell DW 2001b).

Sequence data provides insights into gene functions and the mechanism by which genes are regulated. In some cases comparisons of normal and mutant gene sequences have revealed the molecular basis of hereditary diseases (Wolfe SL 1993b). Sequencing of alleles at STR loci used in forensic identity as well as paternity testing has also been undertaken. Off-ladder alleles are studied to establish a consistent nomenclature for new loci. The species specificity of STR allele sequences is examined and percent stutter correlation with allele length investigated. Validations of the chosen STR loci are also conducted (Lazaruk K *et al* 2001). Sequencing of a STR loci yields an abundance of information about the specific locus and about tetranucleotide repeats in general as a class of length polymorphism (Buscemi L *et al* 1998).

The best-known DNA sequencing techniques are the enzymatic method of Sanger *et al* and the chemical degradation method of Maxam and Gilbert (Wolfe SL 1993b, Sambrook J & Rusell DW 2001a). Although very different in principle, these two methods generate populations of oligonucleotides that begin from a fixed point and terminate at a particular type of residue (Sambrook J & Rusell DW 2001a, Lewin B 1994a).

Sanger and his colleagues developed the chain termination or dideoxy method. This technique is similar to the chemical method except that it uses DNA replication to provide the consecutive sequence lengths for gel electrophoresis (Sambrook J & Rusell DW 2001a). In this enzymatic method of DNA replication, priming of DNA synthesis is achieved by the use of a primer that is complementary to a specific sequence on the template strand. Additionally, modified forms of the four DNA nucleotides are used in which a single H is bound to the 3'-carbon of the deoxyribose sugar instead of an OH. During DNA replication a new nucleotide is normally added to the 3'-OH group of the most recently added nucleotide in the copy. Because the dideoxynucleotides have no 3'-OH available for addition of the next base, DNA replication stops wherever one of the modified nucleotides is inserted instead of an unmodified nucleotide (Sambrook J & Rusell DW 2001a, Wolfe SL 1993b, Rieger R *et al* 1991). Because the stopping points are random, extended replication produces a series of sequence fragments in which each fragment starts at the same place but ends at a different place in the sequence. Running the fragments on an electrophoretic gel separates them by length with the shortest at the bottom (Wolfe SL 1993b, Rieger R *et al* 1991). They can be separated by electrophoresis on acrylamide gels (Lewin B 1994a), and/or capillary electrophoresis on automated instruments (Tagliaro F *et al* 1998).

In fluorescent sequencing using PE Applied Biosystems automated sequencers, fluorescent dye labels are incorporated into DNA extension products using 5' dye labeled primers (dye primers) or 3' dye labeled dideoxy nucleotides triphosphates (dye terminators). Both methods employ cycle sequencing, whereby successive rounds of denaturation, annealing and extension in a thermal cycler result in linear amplification of extension products. The most appropriate method to use depends on the sequencing objectives, the performance characteristics of each method, and personal preference.

The new high sensitivity dye (BigDye terminators) structures contain a fluorescein donor dye, linked to one of four dichlororhodamine (dRhodamine) acceptor dyes. The excitation maximum of each dye label is that of the fluorescein donor, and the emission spectrum is that of the dRhodamine acceptor. The donor dye is optimized to absorb the excitation energy of the argon ion laser in the DNA genetic analyzer.

With the BigDye Terminator cycle Sequencing Ready Reaction format, the dye terminators, deoxynucleoside triphosphates, AmpliTaq, DNA Polymerase, FS, rTth pyrophosphatase, magnesium chloride, and buffer are premixed into a single tube of Ready Reaction Mix. These reagents are suitable for performing fluorescence-based cycle sequencing reactions on single-stranded or double-stranded DNA templates, on PCR products and on large templates.

Cycle sequencing offers a number of advantages: protocols are robust and easy to perform, much less template DNA than single temperature extension methods is required, it is more convenient than traditional single-temperature labeling that requires a chemical denaturation step of double-stranded DNA, the high temperature reduces secondary primer-to-template annealing, the same protocol is used for double and single-stranded DNA, the protocol works well for direct sequencing of PCR products and difficult templates such as bacterial artificial chromosomes (BACs) can be sequenced (Automated DNA Sequencing Guide, Applied Biosystems).

PE Applied Biosystems DNA genetic analyzers detect fluorescence from four different dyes that are used to identify the A, C, G, and T extension reactions. Each dye emits light at a different wavelength when exited by an argon ion laser. The four different colors and therefore the four bases can be detected and distinguished in a single gel lane or capillary injection (Automated DNA Sequencing Guide, Applied Biosystems).

1.8 Capillary electrophoresis

Capillary electrophoresis (CE) or high performance capillary electrophoresis (HPCE), is an instrumental evolution of the traditional slab gel electrophoretic technique (Tagliaro F *et al* 1998, Buscemi L *et al* 1998). This technique is used to analyze DNA restriction fragments, PCR products and for DNA sequencing (Manetto G *et al* 1998). The separation, detection and analysis of polymorphic DNA loci have benefited from the use of semi-automated sequencers, introduction of CE and specialist software (Buscemi L *et al* 1998, Watson S *et al* 2001). In addition, the availability of multiple fluorescent dyes enables the co-amplification of multiple loci with overlapping size ranges, and the co-running of an internal size standard with each sample. This results in extremely precise sizing of products and facilitates reliable allele designation (Watson S *et al* 2001).

The basic set up of CE equipment is simple. Briefly, a CE instrument consists of an injection system, a separation capillary, a high voltage source, electrodes, electrode jars and detectors. The integration of laser induced fluorescence detection with CE is the ideal complement to PCR using either fluorescent dye labeled or cold primers (Monetto G *et al* 1998). The capillary electrophoresis instruments are capable of simultaneous multicolor detection and high resolution of PCR fragments (Lazaruk K *et al* 1998).

In CE the migration of size standards and alleles during capillary electrophoresis is controlled by various factors. Some prominent factors are current, pH and conductivity of the buffer, temperature, partial replacement of the polymer after each electrophoresis run, and the charge of the migrating species (Shewale JG *et al* 2000). When using automated instruments, optimizing electro-kinetic injection parameters can greatly improve data quality, run-to-run precision in sizing, and reproducibility in the amount of sample loaded. The goal is to inject sufficient DNA to yield peaks of adequate height while maintaining the resolution and precision required by the application (Gene Scan Reference Guide, Applied Biosystems).

When the DNA fragments reach a detector window in the capillary, a laser excites the florescent dye labels. Emitted fluorescence from the dyes is collected once per second by a CCD camera, which simultaneously detects all wavelengths from 525 to 680 nm. The fragments are stored as digital signals on a computer for processing (Automated DNA sequencing, Lazaruk K *et al* 1998).

Compared to slab gels, CE separation with non cross-linked gels may offer considerable advantages in terms of sensitivity, reproducibility, versatility and productivity (Tagliaro F *et al* 1998). Furthermore, it enables faster separation and higher resolution, the possibility of automating sample loading, semi quantitative analysis of results (Buscemi L *et al* 1998), and minimal sample requirements (Righetti PG & Gelfi C 1998).

Lazaruk *et al* (1998) determined the sizing precision of a capillary electrophoresis instrument by two different modes using population data base samples where sequence variation may exist between some STR alleles, and by running allelic ladders from ten loci in 105 consecutive injections on PCR products of known DNA sequences. The sizing precision using the performance-optimized polymer 4 (POP 4) with various size standards showed a standard deviation $\leq$ 0.13 bp (Lazaruk K *et al* 1998).

Using the ABI Prism genetic analyzer and POP-4 polymer, Buscemi L *et al* (1998) reported that PCR fragment sizes obtained by comparison with a commercial internal size standard provides very good accuracy and precision.

1.9 Paternity testing

1.9.1 Paternity testing systems

As the number of paternity tests being performed increases, so has the variety of genetic markers used in the resolution of parentage. Excluding DNA tests, as many as 62 immunological and biologic systems are potentially applicable. Genetic markers are recognizable characteristics inherited from parents and controlled by alleles on a pair of chromosomes. Recognizable characteristics can be gross physical qualities such as hair and eye color, or the molecular, serologically and biochemically detectable properties of blood and other body tissue components (Brooks MA 1996). Paternity testing is most often practiced in forensic medicine (Shiono H *et al* 1985). Conventional testing in the investigation of disputed paternity consists of typing alleles in a number of polymorphic systems (Schlaphoff TE *et al* 1993).

In latter years, not only were the classical blood groups exploited, but serum/plasma proteins, red cell enzymes and the HLA system, which exists in alternative forms as polymorphisms, were included (Shiono H *et al* 1985 129, Schlaphoff TE *et al* 1993).

Although the methodology for the determination of the above-mentioned systems is well established and relatively cheap, many of the systems are not polymorphic and hence not very informative (du Tiot E 1993). As a result, their use in parentage testing is often substituted by the more polymorphic systems such as HLA and other DNA markers. In recent years, short repetitive sequence DNA markers have been employed in most laboratories. The DNA markers in common use are RFLPs/VNTRs and STRs (Zhang X-W *et al* 1991, Mangin PD & Ludes B-P 1991, Yokoi T *et al* 1990, Sullivan KM *et al* 1993, Peake IR *et al* 1990).

The continuing development and validation of STR systems for the use in identity testing has now resulted in 20 or more suitable STR systems being available either commercially or via published primer sequences (Thompson JA *et al* 1999). For the determination of STRs, a number of methods including those used for RFLPs / VNTRs, mentioned earlier, have been employed. In all the techniques, PCR followed by electrophoresis is common, with a variety of detection methods available. The detection methods are: silver staining (Yamamoto T *et al* 2001, Minaguchi K *et al* 2000, Ricci U *et al* 1999), ethidium bromide staining (Haddad AP & Sparrow RL 2001), auto radiographic hybridization (Litt M & Luty JA 1989), and fluorescent staining (Schumm JW 1996, Seidl C *et al* 1999, Shewale JG *et al* 2000, Watson S *et al* 2001). Cloning (Yoshimoto T *et al* 1999, 2001), and sequencing using different techniques have also been reported (Seidl C *et al* 1999, Minaguchi K *et al* 2000, Watson S *et al* 2001). In addition, STR detection can also be accomplished by time-of-flight-mass spectrometry (Butler JM *et al* 1998).

1.9.2 Assumptions & calculations in Paternity Testing

To be suitable for paternity testing, the genetic markers must follow Mendel's basic rules of inheritance. Furthermore, systems used for paternity testing must fulfill the following conditions: the mode of inheritance must be known with certainty (Schneider PM 1997, Mayr WR *et al* 1991, Silver H 1989), the markers must be developed and detected at birth, they must not change throughout life or be affected by external agents, the markers should only reflect the genotype, the characteristics should be defined unambiguously and the techniques for the detection of the markers must be simple and reproducible (Mayr WR *et al* 1991, Silver H 1989, du Tiot E 1993). In order to be able to make biostatistical calculations, the allele frequencies should be available (Mayr WR *et al* 1991). Additional selection criteria for test systems are genetic stability and low mutation rates, which are desirable to avoid wrongful exclusions (Schumm JW 1996, Schneider PM 1997). To increase exclusion chances, many alleles with uniform allele frequencies should be tested. The physical stability of the marker should also be taken into consideration (Schneider PM 1997). The reagents used in testing must be reliable, consistent in reactivity and easily attainable (Silver H 1989, Tipett P 1978).

The evaluation of DNA evidence for paternity is usually concerned with three observations: the genotype of the mother, child and alleged father (Lee JW *et al*

2001). Interpretation of paternity investigations relies on the power of exclusion, which is the comparison between the genotypes of the trio to the genotypes of the population. The power of exclusion (PE) depends on; the number of possible alleles per locus, the number of loci in a haplotype, the frequency with which those alleles occur in the population and the capability of detecting the genetic markers (Walker RH 1978, Brooks MA 1994). This is calculated by listing phenotypes of all possible trios, calculating the frequencies of these possible matings, assuming random matings, using phenotype and genotype frequency charts for a given racial group, combining the results for all possible trios in a system, and combining the results for all systems used (Brooks MA 1994, Weir BS 1996).

The efficiency of a genetic system in paternity testing is evaluated by the mean exclusion chance and it usually indicates the degree of polymorphism at the locus under consideration (Lee JW *et al* 2001). The greater the power of the combined test systems, the less likely that a falsely accused man will not be excluded by the tests. This is referred to as the prior probability of exclusion (Brooks MA 1994, Weir BS 1996).

The interpretation of paternity investigations is based on a careful review of all data followed by application of the laws and principles of genetics. During the analysis a number of assumptions are made in the determination of the phenotypes to aid in the conclusion of the investigation (Walker RH 1978). The attempt to determine paternity begins with the assumption that the mother is the biological mother of the child. Consequently, the child's phenotype is, in part, the result of the mother's genetic make-up. The biological mother is obliged to pass one set of the genes in her genotype to her child and it is referred to as the maternal obligatory genes (MOGs). The identification of possible MOGs enables the determination of possible paternal obligatory genes (POGs) (Brooks MA 1994). Additional assumptions include: that mutations do not significantly affect the markers; positive identification of each party in the trio was achieved; no clerical errors occurred in the labeling of tubes, aliquoting of the specimens and recording of results; the reagents employed were potent and specific as indicated by the results obtained with control specimens and all the tests were performed with attention to detail and correct procedure (Walker RH 1978).

Finding exclusion at two or more loci when testing a biological father has a probability of less than 10^{-7}. This is true for the average locus specific exclusion rates as well as for the most likely genotypes for the mother child pairs. The low mutation rate at the STR loci, furthermore, suggests that even a one-locus exclusion is likely to be due to non-paternity and may not be caused by mutation at the locus where the exclusion is observed (Chakraborty R & Stivers DN 1996). In practice, however, the opinion of non-paternity should only be given on the basis of exclusions at two or more loci that are located on two different chromosomes. This guards against the chance of a mutation falsely indicating exclusion at one locus (Weir BS 1996, Bein G *et al* 1998).

The probability with which a particular allele is passed from a genotype depends on whether the genotype is heterozygous or homozygous. In the event of homozygosity, the probability of an allele appearing in the genotype of the offspring is 100 %. On the other hand, in the heterozygous state, each allele of the genotype has a 50 % chance of being transmitted. The probability that a child is the result of the mating of the biological mother and a random man can thus be computed (Brooks MA 1994).

Before any locus can be used in paternity testing, it needs to be tested for Hardy-Weinberg equilibrium in the specific population (Weir BS 1996, Sudbury AW & Marinopouls J 1993). An equilibrium state is one in which properties of the population do not change over successive generations. Testing for Hardy-Weinberg equilibrium establishes whether the observed genotype frequencies are close enough to the expected genotype frequencies, thus ensuring that the same relationship is true for the population frequencies (Weir BS 1996). Hardy-Weinberg equilibrium follows from the assumption of random mating within a population, but random mating very rarely occur. Further, it is well known that allele frequencies differ from one human sub-population to another. Thus, even if a population is in strict Hardy-Weinberg equilibrium, a mixture of sub-populations may still occur (Sudbury AW & Marinopouls J 1993).

The process of estimating the probability of a match involves multiplying the frequencies of each allele. The assumption is that each system is genetically independent of the others used in the investigation i.e. having a particular marker in one system is not related to, or influenced by markers which the person possesses for other DNA systems (Sudbury AW & Marinopouls J 1993, Lincoln PJ 1997). However, it is almost certain that this assumption is not strictly true (Sudbury AW & Marinopouls J 1993).

To calculate a probability of paternity, it is necessary to know the racial origin of the putative father and to have the allele frequencies for the racial group or groups to which he belongs. Allele frequencies may be determined by a direct count if the system has co-dominant alleles and it is based on the Hardy-Weinberg principle (Brooks MA 1994, Weir BS 1996).

According to the basic rules of inheritance, a genetic marker in a child must be present in one or both of the biological parents, a homozygous marker in the biological parent must be transmitted to the offspring, and a homozygous marker found in a child must be present in both biological parents (Silver H 1989). Therefore, the probability that a child was the result of the mating of the mother and alleged father is the summation of the probabilities that the following two events occurred simultaneously: (1) the mother passed the MOG to her child, and (2) the alleged father passed the POG to his child (Brooks MA 1994).

In the event of the alleged father not being excluded, the probability of paternity can be calculated, either by the paternity index or the probability of paternity according to Essen-Moller. These calculations are based on two competing hypotheses:

(1) the probability that the child resulted from the mating of the mother and alleged father

(2) the probability that the child resulted from the mating of the mother and a random man (Brooks .A 1994, Weir BS 1996, Morling N *et al* 2002).

The ability of genetic systems to define the non-paternity of an alleged father and, if there is no exclusion, the possibility of to what extent the results can show evidence for or against the paternity of the man, determines the usefulness of the systems (Mayr WR *et al* 1991).

An alleged man can be excluded from paternity in two situations:

(1) the child may posses a marker that is absent from the mother and the putative father ("first order" or "direct" or "class I" exclusion) (Mayr WR *et al* 1991, Silver H 1989).

(2) the alleged father has a monomorphic marker, which cannot be demonstrated in the child ("second order" or "indirect" or "class II" exclusion) (Mayr WR *et al* 1991, Silver H 1989).

With proper testing procedures, it is rare that a first-order exclusion can be false, and one first-order exclusion has been considered sufficient for a determination of non-paternity. Although the possibility of mutation is rare, in recent years it has become practice in most laboratories to require at least two exclusions (Brooks MA 1994).

The probability of paternity will be high when there are specific obligatory genes for all of the systems studied and when the obligatory genes are not commonly found in the random population. Low probabilities of paternity usually result when obligatory genes are not rare (occur often in the random population), and/or multiple obligatory genes exist for some or most of the systems tested (Brooks MA 1994). This scenario may present when the putative father is not available for testing and profiles of extended family are used to compile the profile of the putative father.

1.10 Aim of the study

The aim of this study can be summarized as:

a) The selection of samples in such a manner as to cover all the possible FGA alleles in the local population groups.

b) The selection of samples that contain off-ladder and variant FGA alleles.

c) The determination of the DNA sequence of all the normal-, off-ladder- and variant FGA alleles.

d) Comparison of the FGA allele sizes obtained by DNA sequencing with the sizes assigned to the same alleles by genotyping methods (especially the off-ladder alleles).

e) Comparison of the DNA sequences of the FGA alleles with published data.

f) Describing new- or variant FGA alleles in the local populations.

g) Determination of possible ethnic restriction of certain FGA alleles.

CHAPTER 2

MATERIALS AND METHODS

2.1 SAMPLES

All DNA samples included in this study were extracted from EDTA anticoagulated blood samples of paternity test clients from the Free State population. Ethnic affiliation of each individual was documented at the time of phlebotomy by self-declaration. Before analysis, the link of these samples to the donors thereof was broken to guarantee anonymity. The ethics committee of the University of the Free State (ETOVS NR 87/01) cleared this project.

2.2 SAMPLE SELECTION

DNA samples, previously isolated and used in paternity testing that showed off ladder, interalleles were selected. Additional samples were also selected to represent all possible alleles at the FGA locus. For the majority of the selected specimens, extracted leftover DNA, stored at -80° C, from routine STR paternity analysis was used.

2.3 DNA EXTRACTION

Re-extraction of DNA from some of the specimens, with lower DNA concentration, was done from the blood specimens, which were available.

The Promega Wizard Genomic DNA Purification Kit was used to extract DNA from 300 µl EDTA blood according to the protocol.

1. Add 900µl of Cell Lysis Solution to a sterile 1.5ml microcentrifuge tube.
2. Mix the blood thoroughly by gentle rocking and add 300µl of blood to the tube containing the cell lysis solution. Invert the mixture 5-6 times to mix the blood and Red Cell Lysis Solution.
3. Incubate the mixture for 10 minutes at room temperature, with occasional inversion, to lyse the red blood cells.

4. Centrifuge at 13,000g for 20 seconds at room temperature.
5. Remove as much as possible of the supernatant and discard without disturbing the visible white cell pellet.
6. Repeat the red sell lysis, if necessary, until the pellet appears white and contains no red cells.
7. Vortex the tube for 15 -20 sec until the white cells are resuspended.
8. Add 300µl of Nuclei Lysis Solution and mix by pipetting 6 times to lyse the white blood cells.
9. Add 100µl of Protein Precipitation Solution to the nuclear lysate.
10. Vortex the mixture vigorously for 20 seconds.
11. Centrifuge at 13,500 for 10 minutes at room temperature.
12. Transfer the supernatant to a clean 1.5ml microcentrifuge tube containing 300µl of room temperature isopropanol.
13. Mix the solution gently by inversion until the white thread-like strands of DNA form a visible mass,
14. Centrifuge at 13,500g for 1 minute to pellet the DNA.
15. Remove the supernatant carefully and discard.
16. Invert the tube on clean absorbent paper to air-dry.
17. Add 100µl of DNA Rehydration Solution.
18. Incubate the mixture at 65° C for 1 hour or overnight at room temperature.
19. Store the extracted DNA at 4° C.

2.4 DNA CONCENTRATION DETERMINATION

The DNA concentration of samples was determined spectrophotometrically using GeneQuant pro RNA/DNA calculator.

1. Make a 1:20 dilution of DNA in sterile distilled water.
2. Use the DNA Rehydration Solution from the Promega Kit as a blank.
3. Determine the optical density at 260 nm for DNA concentration and the relative purity at 260 and 280 nm (A260/A280 should be greater that 1.8).

4. A ratio of greater than or equal to 1.7 was considered to be pure DNA. However, DNA samples with less purity were tested in the absence of alternative sample.
5. To calculate DNA concentration from the O.D., the dilution factor and the constant 50 for double stranded DNA is applied. Thus:

   [DNA]µg/ml = OD_{260} X dilution factor X 50
   (A260 = 1 is equivalent to 50µl/ml of ds DNA)

2.5 DILUTION OF SAMPLES

1. Dilute samples with sterile distilled water to give a concentration of 2ng/µl DNA.
2. Samples with O.D reading of less than 0.02 are used undiluted in the PCR reaction.

2.6 PCR

1) Segments of the FGA locus are amplified from genomic DNA using the following oligonucleotide primers (expected product size: 210 - 355 bp).
   a) primer 1: 5'-ATT ATC CAA AAG TCA AAT GCC CCA TAG G-3'
   b) primer 2: 5'-ATC GAA AAT ATG GTT ATT GAA GTA GCT G-3'.
      (Urquhart A *et al* 1995)
   c) PCR amplification primers are reconstituted in TE (10 mM of Tris-HCL (pH 8.3), 0.1 mM of EDTA) buffer.
   d) The primers are left overnight at room temperature, vortexed, centrifuged and aliquoted into 25µl, followed by long term storage at –80° C freezer (Carleton SM 1995).
   e) Those for immediate use are stored at -20°C.
2) With every reaction a positive control, DNA specimen with known FGA alleles, and a blank control that contained all the reagents, but the DNA is replaced by sterile dH_2O, is included.
3) Amplification is performed in a 25µl mixture containing:

a) 10ng of genomic DNA

b) 200µM of each dNTP (Promega)

c) 25pmol of each primer (MWG-Biotech AG)

d) 1 unit of Amplitaq (Roche) in Thermophilic DNA polymerase buffer (10mM Tris-HCl, 50 mM KCl, 0.1% Triton X100. pH 8.3) and 1.5mM $MgCl_2$.

4) In most cases a master mix is prepared from the reagents and mixed with each specimen, while in a number of specimens each reagent is pipetted separately.

5) Samples are amplified in a GeneAmp PCR 2400 thermocycler (Applied Biosystems) using the following parameters:

a) initial denaturation at 94°C for 5 min

b) denaturation at 94°C for 30s

c) annealing at 60°C for 30s } 35 cycles

d) extension at 72°C for 30s.

e) a final extension followed at 72°C for 5 min.

6) PCR products are immediately run on a gel or stored overnight at 4° C.

2.7 GEL ELECTROPHORESIS

1) 2.5 % Agarose gels (Agarose DILE) in 1x TBE buffer is used throughout.

2) A 5X TBE stock buffer is prepared by mixing:

a) 54g of Tris base

b) 27.5g of boric acid

c) 20ml of 0.5 M EDTA (pH 8.0)

d) the pH is adjusted to 7.3

e) make up to 1 liter using ddH_2O

3) The 5X TBE stock buffer is diluted 1:5 for daily use.

4) Add 1.25 g of agarose to 50 ml of TBE buffer.

5) Melt the mixture in a microwave for two minutes, and cool under running tap water.

6) Add 2.5µl of 1% ethidium bromide (0.5µg/ml) and mix thoroughly.

7) Pour the gel in the gel plate and leave for 30 minutes to solidify.

8) Mix the amplified PCR product (10-15 ul) with a 1/5 volume of loading buffer (0.25% bromophenol blue in 40% Sucrose solution) on para-film.
9) Load the mixtures of each sample, control samples and a molecular-weight marker on the solidified gel in separate slots and electrophorese at 6.5 V/cm for 2 hrs in 1x TBE buffer.
10) Visualize the DNA bands visualized on a UV transilluminator.
11) Excise the bands of interest from the gel with a sterile surgical blade and purify the DNA from the gel using GenELute minus EtBr spin columns (Sigma).
    a) Pre-wash the spin column pre-washed by adding 100µl of 1x TE, capping and centrifuging at 13.500g for 10 seconds.
    b) Discard the elute and transfer the spin column to a fresh collection tube.
    c) Load the excised bands of interest, cut into small pieces, into each of the spin columns.
    d) Centrifuge the columns at 3,500g for 10 minutes.
    e) Store the purified DNA in the collection tube at 4° C until further testing.
12) Amplify 5 µl aliquots of the purified PCR product and re-purify in a 2nd to 4th round of PCR (parameters as for 1st PCR) until a single band is obtained.
13) Check the purity of PCR products by running the aliquots on a 2.5% agarose gel as mentioned above.

2.8 PURIFICATION OF FRAGMENTS FOR SEQUENCING

When the presence of a single band is confirmed, the remaining PCR product is purified using the High Pure PCR Product Purification Kit (Roche) to remove primers, dNTP's and Taq before used in the cycle sequencing reaction.

1) Mix 40µl of the PCR product with 400µl of binding buffer.
2) Transfer this mixture to a High Pure filter tube connected to a collection tube.
3) Centrifuge for 1min at 13,000g at room temperature, and discard the flow-through solution.
4) Reconnect the filter tube to the collection tube, add 400µl wash buffer to the upper reservoir and centrifuge the assembly centrifuged as above.

5) Discard the flow-through solution, recombine the tubes, add 200µl of wash solution and centrifuge as above. This ensures optimal purity and complete removal of wash buffer from the glass fibers.
6) Discard the flow-through solution and collection tube and connect the filter tube to a clean prelabeled 1.5µl microcentrifuge tube.
7) Add 50µl of Elution Buffer and centrifuge as above.
8) Store the elute, containing the purified DNA, at -15° C for later analysis.

2.9 SEQUENCING

Sequencing of the purified PCR products are performed using the ABI Prism Big Dye Terminator Ready reaction kit vs.3 (Applied Biosystems).

1. The forward sequencing reaction is performed using primer 1 and the following parameters:
   a) denaturing at 96°C for 10s
   b) annealing at 60°C for 5s } 25 cycles
   c) extension at 72°C for 30s.
2. The reverse sequencing is performed using primer 2 and the same parameters except for annealing at 55°C for 5s.
3. Since determination of DNA concentration is not possible, 8 µl of the purified sequence product is used in a 20µl total reaction volume.
4. The concentration of primers used is 3.2 pmol (4µl of 0.8pmol/µl primer).
5. Because of variations in the annealing temperatures, the forward, reverse and control primer reactions are run separately for each set of reactions.

2.10 PURIFICATION OF CYCLE SEQUENCE PRODUCTS

Purification of sequence reaction products is performed using Ethanol/Sodium acetate precipitation in micro centrifuge tubes according to ABI PRISM BigDye Terminator v3.0 Ready Reaction Cycle Sequencing kit protocol with a slight modification.

The procedure is:

1. Place 3µl of 3M sodium acetate (pH5 and 4° C), 62.5µl of 95% ethanol (-20° C) and 14.5µl sterile distilled water in a 1.5 ml Eppendorf tube.
2. Add 20µl of the sequence product, mix, vortex for 30s and centrifuge for 10 seconds.
3. Keep the mixture at room temperature for 30 min and centrifuge for a further 30 min at 13,500g. The tubes are carefully orientated, as the precipitate is not visible.
4. Aspirate the supernatants carefully with separate tips for each sample.
5. Add 250µl of 70% ethanol (-20° C) to the pellets. The samples are vortexed for 2 min and centrifuged again for 10 min with proper orientation.
6. Aspirate the supernatants, and leave the tubes at 90° C for 1min to dry.

2.11 RESUSPENDING THE SAMPLES FOR SEQUENCING WITH POP-6 POLYMER

Since specific times are not given for some of the steps in the protocol, we use a uniform procedure for all specimens.

1. Add 25 µl template suppression reagent (TSR, supplied with the polymer), Vortex the samples for 1 minute and centrifuge at high speed for 30 seconds.
2. Denature the samples for 2min at 95° C and chill on ice for 5 minutes.
3. Vortex the samples for 10 seconds and centrifuge at high speed for additional 30 seconds.
4. Place the resuspended product on ice in the dark until loaded on the sequencer.

2.12 CAPILLARY ELECTROPHORESIS

1. After denaturation, samples are run on an ABI Prism 310 Genetic Analyzer (Applied Biosystems, Foster City, CA, USA).

2. Separation is performed in a coated capillary, 61 cm by 50µm internal diameter, filled with separation polymer POP-6 (performance optimized polymer 6).
3. Samples are electro-kinetically injected in for 30sec at 12.2KV and electrophoresed for 70min at 50volts/cm, 5µA and 50° C.
4. The run module used is the seqPOP6 (1.0 ml) and the mobility file is DT POP6 (BD set any primer). The base caller used is CE-1.

2.13 DATA ANALYSIS

1. After capillary electrophoresis, the sequencing results are analyzed using Sequencing Analysis Software version 2.1 (PE Applied Biosystems).
2. The settings for analysis are the BigDye VS3 filter set and DT (BD set any-primer) mobility file.
3. Data obtained from the analysis is checked, printed and compared with STR analysis results and reported sequence data.
4. The variation observed from STR analysis is presented in the number of nucleotides.

CHAPTER 3

RESULTS

The study was conducted in two phases. In both phases a total of 62 specimens were tested and characterized for either one or both alleles. The racial origin of the specimens was 52 from the Negroid, 5 from the Mixed Ancestry, 4 from the Caucasian and 1 from the SAN group. The first phase of the study was conducted only on samples that were believed to be inter-alleles and off-ladder alleles. In the second phase of the study all the alleles that are found in this study group was included to fill in the total information of the FGA locus in our population groups.

In this study a total of 76 alleles, representing 27 different alleles, were sequenced both by forward and reverse priming. Table 3.1 is a summation of all the alleles that were sequenced in this study. Since the selection of alleles for sequence analysis was not random, the number of sequence variants found is not a reflection of the proportion of variants expected in a random sampling of the population groups studied. Before sequencing the alleles were amplified and separated on agarose gels. Repeated amplification, separation and purification were done until a pure single band was observed. See figs. 3.1 to 3.6 for gelpictures illustrating agarose gel separation of DNA fragments. Fig. 3.1 and 3.2 are examples of PCR products of heterozygous FGA alleles; being the results of the first round of PCR. Each lane of the gel represents one sample with two alleles at the FGA locus. All the samples on these gels are heterozygous at the FGA locus. Fig.3.3 and 3.4 represent PCR products of both homozygous (only one band) and heterozygous (two bands) FGA alleles. At this stage the band of interest is excised from the gel, the DNA purified and used in the second round of PCR. Fig.3.5 and 3.6 are PCR products after the second round PCR. The fact that only one band per sample is visible is proof of an effective isolation and purification of the specific FGA allele. All these photographs are shown as examples to illustrate the general method used for separation and purification of single alleles and do not refer to specific samples that were analysed.

In the first phase of the study samples were selected to include all the possible interalleles and off-ladder alleles as was observed in routine paternity investigations. The sequence of these alleles had not been described at the time of launching the study. A total of 37 samples were included: 32 Negroid, 3 Mixed Ancestry and 2 Caucasian. Of these samples 25, were off-ladder alleles (alleles greater than the largest allele (i.e. 30) in the AmpFlSTR Profiler Plus Kit allelic ladder from Apllied Biosystems). A total of 38 DNA fragments, representing 9 different alleles, were selected for investigation. According to designation by routine STR analysis, the within-ladder alleles in this group that were sequenced were 16.1(n=9), 24.2(n=1) and 29.2(n=3). The off-ladder alleles were 31.2(n=2), 32.2(n=1), 41(n=6), 42(n=1), 44(n=14) and 45(1) respectively. After sequence characterization 8 different alleles were distinguished. These were 16.1(n=9), 24.2(n=1), 29.2(n=3), 31.2(n=1), 40.2(n=5), 41.2(n=1), 43.2(n=17) and 44.2(n=1). The 13 within-ladder alleles (16.1, 24.2 and 29.2) were correctly assigned by routine STR analysis while all (i.e.. 25) of the off-ladder alleles were wrongly assigned. From the 25 wrongly assigned off-ladder alleles 80%(20/25) had a 2bp difference from the sequence confirmed allele length. The remaining 20%(5/25) had a difference of four or more base pairs. The wrong assignment of off-ladder alleles can have serious implications in paternity and forensic investigations, especially if results from different laboratories are used in an investigation. It could lead to incorrect results and thus a false deduction from the results, thus falsely excluding or not-excluding an individual in a paternity or forensic investigation.

Three new alleles, of which the sequence had not been described, were found; 29.2(fig. 3.29) was found in 3 of the negroid specimens, 40.2(fig. 3.33) was found in 5(2 negroid and 3 mixed ancestry) of the specimens and 41.2(fig. 3.34) was found in only one negroid specimen. Two sequence variants, i.e. alleles of which the repeat motifs differs from the published, were also found in one negroid sample each (43.2', 44.2)(figs. 3.36 and 3.37).

In the second phase of the study samples were selected to cover all the possible alleles other than those included in the first phase. A total of 26 specimens (22 Negroid, 2 Mixed Ancestry, 1 Caucasian and 1 SAN) representing 38 alleles were sequenced. In this phase of the study 24 different alleles were sequenced. Only two specimens contained off-ladder alleles i.e. 31.2 and 40.2(figs. 3.32 and 3.33). These alleles were found to be 32 and 40 respectively by routine STR analysis. This implies a difference of 2bp from the routine STR allele designation. Of the remaining 36 within-ladder alleles. 11.11%(4/36) were wrongly assigned. Two of these differed only by 1bp, one by two bp's and the other one by more than four base pairs. One previously sequence-undescribed allele was found in this group of samples i.e. allele 40.2 which was found in only one negroid sample (fig. 3.33). Both alleles 26 and 28 showed two sequence variants each of which was found in only one negroid sample (figs. 3.23, 3.24, 3.26 and 3.27).

Again here, the difference in size between the actual allele size and the size assigned by the genotyping could have adverse effects on investigations in both paternity and forensic applications. If data is generated in two different laboratories one would not be able to compare such data; an example would be for data in the Interpol DNA database. Different laboratories contribute to this database and a difference in size of alleles could result in serious mismatches and could lead to the false exclusion of criminals.

In both studies a total of 27 different FGA alleles were sequence characterized. Of these, three alleles 26(263bp)(fig. 3.23 and 3.24), 28(271bp)(fig. 3.26 and 3.27) and 43.2(333bp)(fig. 3.35 and 3.36) had 2 sequence variants each. Allele 43.2 was found in 13 negroid, 2 caucasian and 1 mixed ancestry samples. Three new, previously undescribed, alleles were found in this study, i.e. allele 29.2 in three negroid samples (fig. 3.29), 40.2 in three negroid and three mixed ancestry samples (fig.3.33) and 41.2 in one negroid sample (fig. 3.34).

Based on the fragment size, two distinct groups of FGA alleles were observed: those in the allele size range from 16.1 - 31.2 and those in the size range 40.2 - 44.2. No alleles were found in the allele size range between 31.2 and 40.2. According to sequence similarities however, the FGA alleles can be divided into three groups. The first group is the shorter alleles (16.1-26) that have a [CTCC (TTCC)$_2$] motif in common at the 3' end. An exception is the 16.1 allele that contains a T insertion interrupting the long CTTT repeat sequence. The second group is the alleles between 26 and 31.2 that contain an additional (CTTT)$_{3\text{-}5}$ repeat motif before the CTCC (TTCC)$_2$ motif at the 3' end. Exceptions are the 29.2, 30.2 and 31.2 alleles that contain a [CTCC (TTCC)$_4$] repeat motif at the 3' end. The third group is the larger alleles (40.2 - 44.2) that contain a (CTGT)$_{3\text{-}6}$ repeat motif interrupting the long CTTT repeat sequence. An additional insertion of a (CTTC)$_{3\text{-}4}$ (CTTT)$_3$ repeat motif is found before the CTCC (TTCC)$_4$ motif at the 3' end. In addition, all the sequenced alleles had a TTTC TTCC TTTC TTTTTT motif at the 3' end that was not included in the allele designation. No mutation was observed either in the flanking or the primer binding regions of this short tandem repeat locus. A summary of all the allele sequence formulas is given in Table.3.1. The electropherograms of the allele sequences are shown in figures 3.7 to 3.38. All the alleles were sequenced by forward and reverse priming to confirm the sequences. In most of the specimens, the beginning of the forward sequencing was unclear, this could however, be clarified by the reverse sequence reaction. As an example, the forward and reverse sequences of alleles 16.1 and 44.2 are given. For the remaining alleles, only the forward sequence is given due to a restriction on the number of pages allowed for a thesis.

Table 3.1. Summary of all the FGA allele formulas encountered in this study.

Allele	*n*	**Repeat motif**										**bp**
16.1	11	(TTTC)$_3$	TTTT	TTCT	(CTTT)$_5$	T	(CTTT)$_3$			CTCC	(TTCC)$_2$	224
18	1	(TTTC)$_3$	TTTT	TTCT	(CTTT)$_{10}$					CTCC	(TTCC)$_2$	231
18.2	1	(TTTC)$_3$	TTTT	TT	(CTTT)$_{11}$					CTCC	(TTCC)$_2$	233
19	4	(TTTC)$_3$	TTTT	TTCT	(CTTT)$_{11}$					CTCC	(TTCC)$_2$	235
19.2	1	(TTTC)$_3$	TTTT	TT	(CTTT)$_{12}$					CTCC	(TTCC)$_2$	237
20	2	(TTTC)$_3$	TTTT	TTCT	(CTTT)$_{12}$					CTCC	(TTCC)$_2$	239
21	2	(TTTC)$_3$	TTTT	TTCT	(CTTT)$_{13}$					CTCC	(TTCC)$_2$	243
21.2	1	(TTTC)$_3$	TTTT	TT	(CTTT)$_{14}$					CTCC	(TTCC)$_2$	245
22	4	(TTTC)$_3$	TTTT	TTCT	(CTTT)$_{14}$					CTCC	(TTCC)$_2$	247
22.2	2	(TTTC)$_3$	TTTT	TT	(CTTT)$_{15}$					CTCC	(TTCC)$_2$	249
23	2	(TTTC)$_3$	TTTT	TTCT	(CTTT)$_{15}$					CTCC	(TTCC)$_2$	251
23.2	1	(TTTC)$_3$	TTTT	TT	(CTTT)$_{16}$					CTCC	(TTCC)$_2$	253
24	1	(TTTC)$_3$	TTTT	TTCT	(CTTT)$_{16}$					CTCC	(TTCC)$_2$	255
24.2	1	(TTTC)$_3$	TTTT	TT	(CTTT)$_{17}$					CTCC	(TTCC)$_2$	257
25	2	(TTTC)$_3$	TTTT	TTCT	(CTTT)$_{17}$					CTCC	(TTCC)$_2$	259
26	3	(TTTC)$_3$	TTTT	TTCT	(CTTT)$_{18}$					CTCC	(TTCC)$_2$	263
26'	1	(TTTC)$_3$	TTTT	TTCT	(CTTT)$_{12}$	CCTT	(CTTT)$_5$			CTCC	(TTCC)$_2$	263
27	1	(TTTC)$_3$	TTTT	TTCT	(CTTT)$_{13}$	CCTT	(CTTT)$_5$			CTCC	(TTCC)$_2$	267
28	1	(TTTC)$_3$	TTTT	TTCT	(CTTT)$_{20}$					CTCC	(TTCC)$_2$	271
28'	1	(TTTC)$_3$	TTTT	TTCT	(CTTT)$_{14}$	CCTT	(CTTT)$_5$			CTCC	(TTCC)$_2$	271
29	1	(TTTC)$_3$	TTTT	TTCT	(CTTT)$_{15}$	CCTT	(CTTT)$_5$			CTCC	(TTCC)$_2$	275
29.2	3	(TTTC)$_4$	TTTT	TT	(CTTT)$_{13}$	(CTTC)$_3$	(CTTT)$_3$			CTCC	(TTCC)$_4$	277
30	1	(TTTC)$_3$	TTTT	TTCT	(CTTT)$_{16}$	CCTT	(CTTT)$_5$			CTCC	(TTCC)$_2$	279
30.2	1	(TTTC)$_4$	TTTT	TT	(CTTT)$_{14}$	(CTTC)$_3$	(CTTT)$_3$			CTCC	(TTCC)$_4$	281
31.2	2	(TTTC)$_4$	TTTT	TT	(CTTT)$_{15}$	(CTTC)$_3$	(CTTT)$_3$			CTCC	(TTCC)$_4$	285
40.2	6	(TTTC)$_4$	TTTT	TT	(CTTT)$_9$	(CTGT)$_3$	(CTTT)$_1$	(CTTC)$_3$	(CTTT)$_3$	CTCC	(TTCC)$_4$	321
41.2	1	(TTTC)$_4$	TTTT	TT	(CTTT)$_9$	(CTGT)$_3$	(CTTT)$_1$	(CTTC)$_3$	(CTTT)$_3$	CTCC	(TTCC)$_4$	325
43.2	16	(TTTC)$_4$	TTTT	TT	(CTTT)$_8$	(CTGT)$_5$	(CTTT)$_1$	(CTTC)$_4$	(CTTT)$_3$	CTCC	(TTCC)$_4$	333
43.2'	1	(TTTC)$_4$	TTTT	TT	(CTTT)$_{12}$	(CTGT)$_6$	(CTTT)$_9$	(CTTC)$_3$	(CTTT)$_3$	CTCC	(TTCC)$_4$	333
44.2	1	(TTTC)$_4$	TTTT	TT	(CTTT)$_9$	(CTGT)$_5$	(CTTT)$_1$	(CTTC)$_4$	(CTTT)$_3$	CTCC	(TTCC)$_4$	337

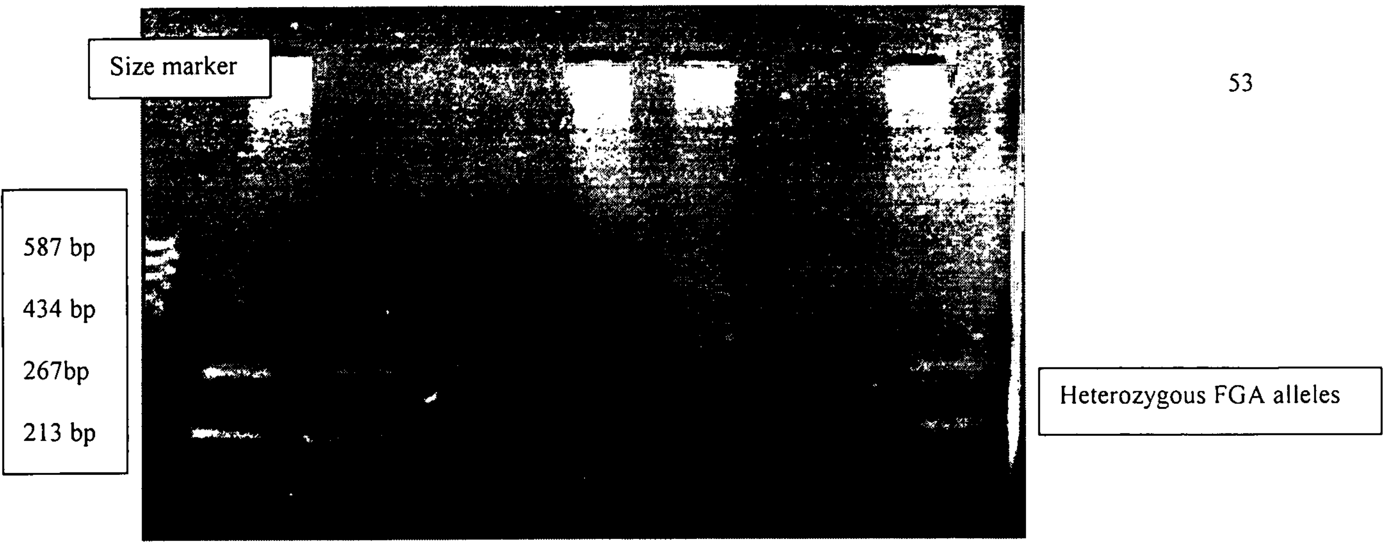


Fig.3.1. PCR products of heterozygous FGA alleles after the 1st round of PCR.

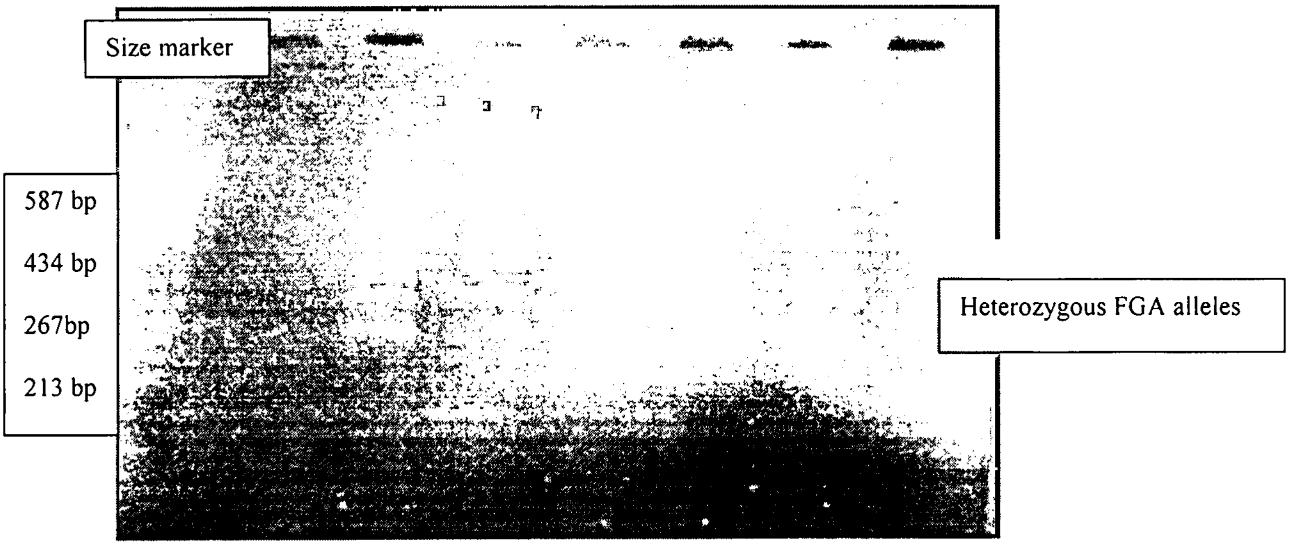


Fig.3.2. PCR products of heterozygous FGA alleles after the 1st round of PCR.

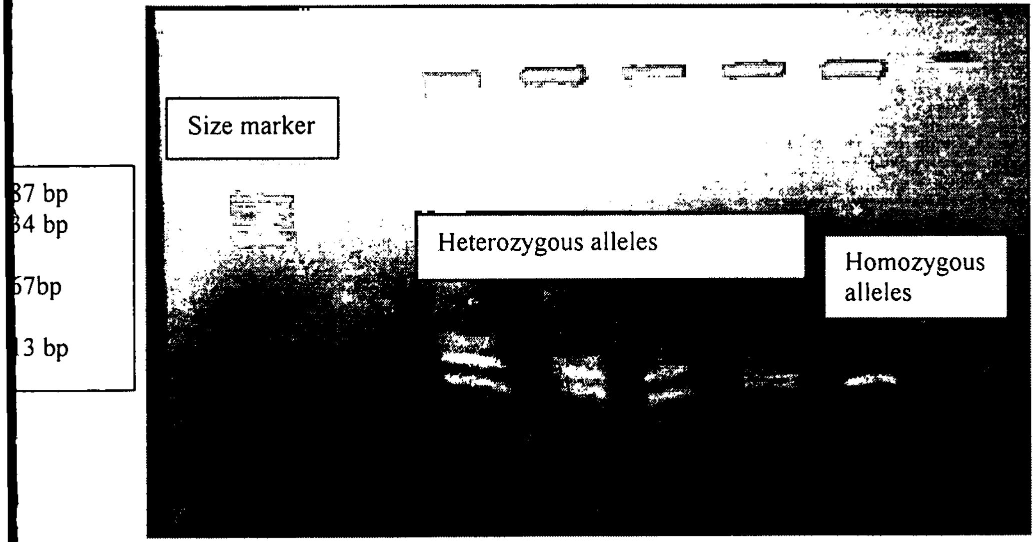
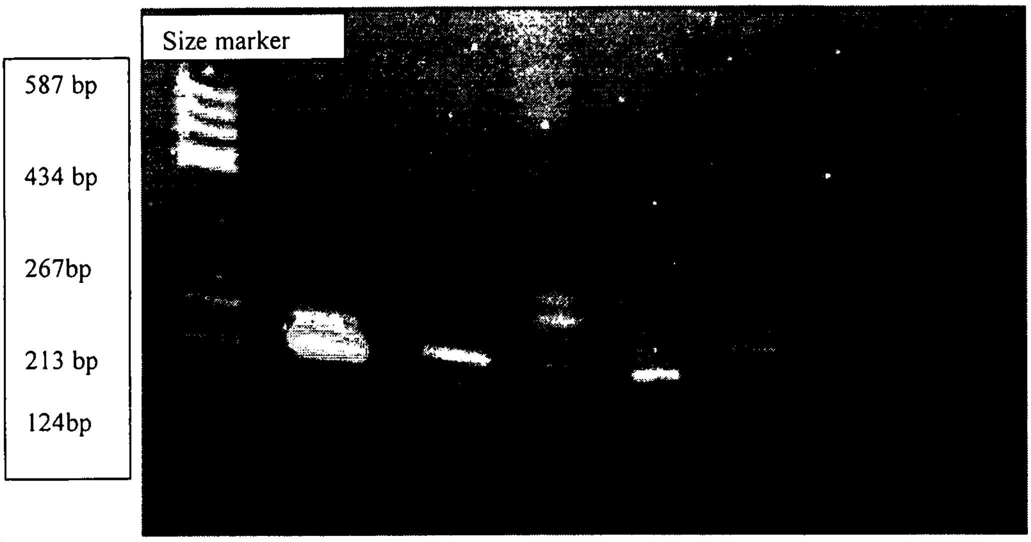
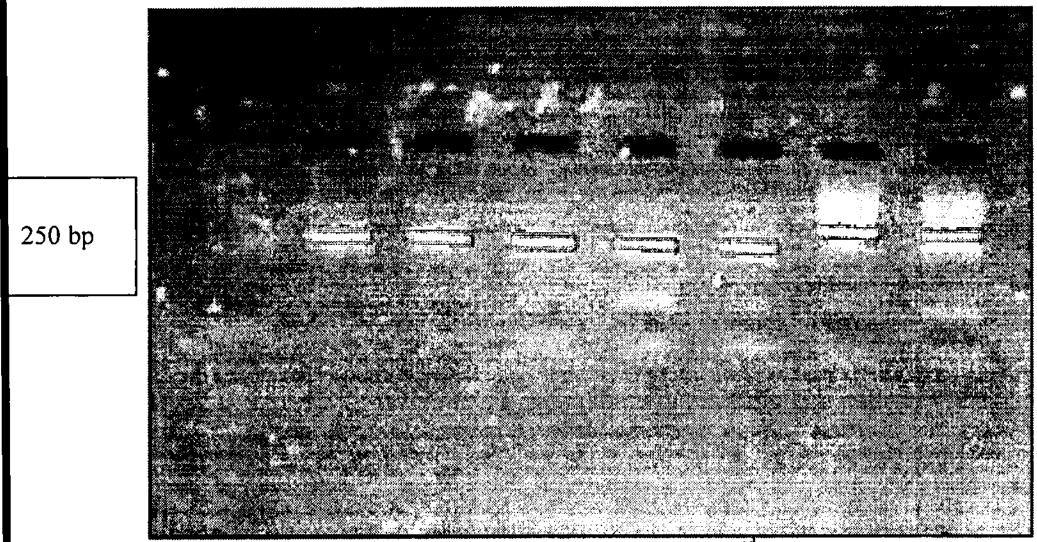


Fig.3.3. PCR products of hetero- and homozygous FGA alleles after the 1st round of PCR.



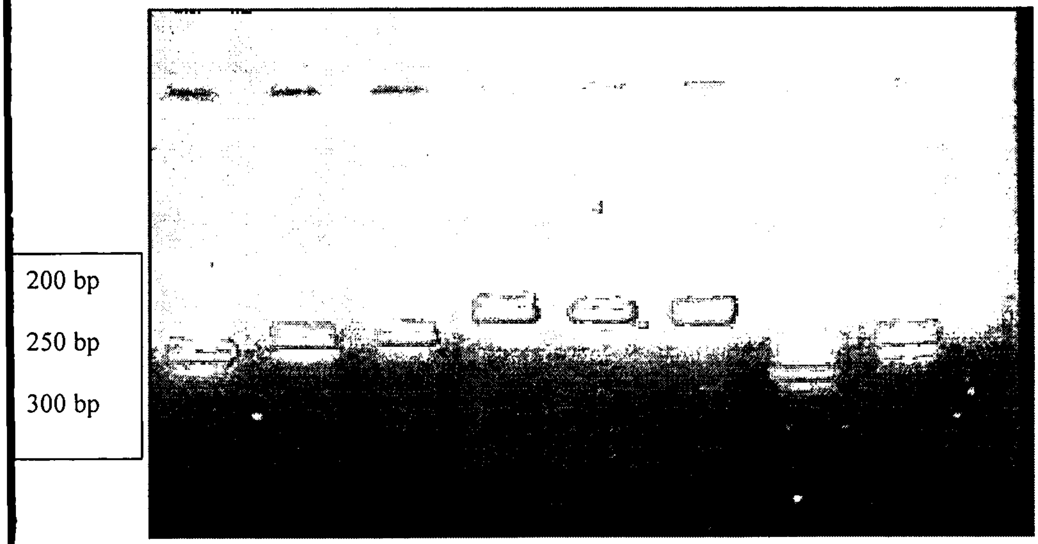
Single-, double- and multiple bands resulting from poor excision of bands after the 1st round of PCR

Fig.3.4. PCR products of FGA alleles exhibiting single, double and multiple bands after the 1st round of PCR.



Single FGA alleles after a good excision of single alleles after the 1st round of PCR

Fig.3.5. PCR products of single FGA alleles after the 2nd round of PCR.



Single FGA alleles after a good excision of single alleles after the 1st round of PCR

Fig.3.6. PCR products of single FGA alleles after the 2nd round of PCR.

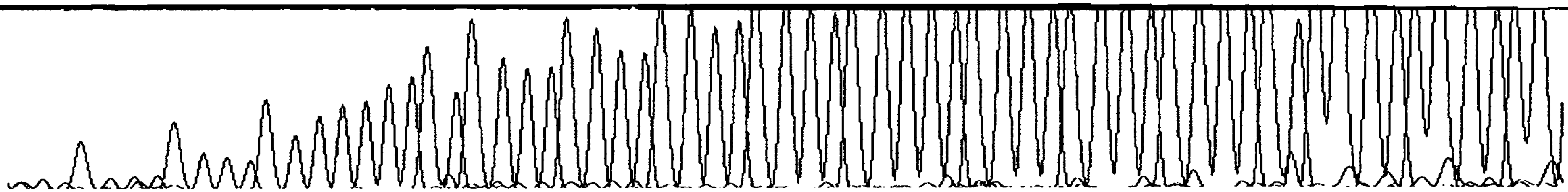


Fig. 3.7. FGA-16.1 forward sequence.

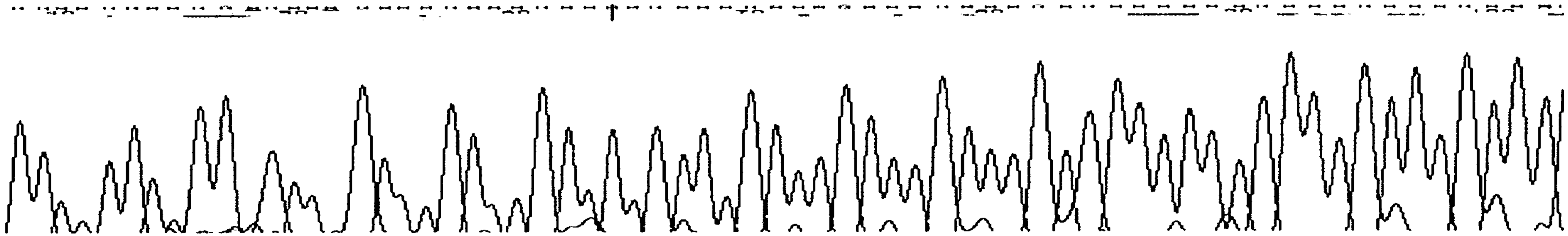


Fig.3.8. FGA-16.1 reverse sequence.
Repeat motif: (TCCC)$_3$ TTTT TTCT (CTTT)$_5$ T (CTTT)$_3$ CTCC (TTCC)$_2$

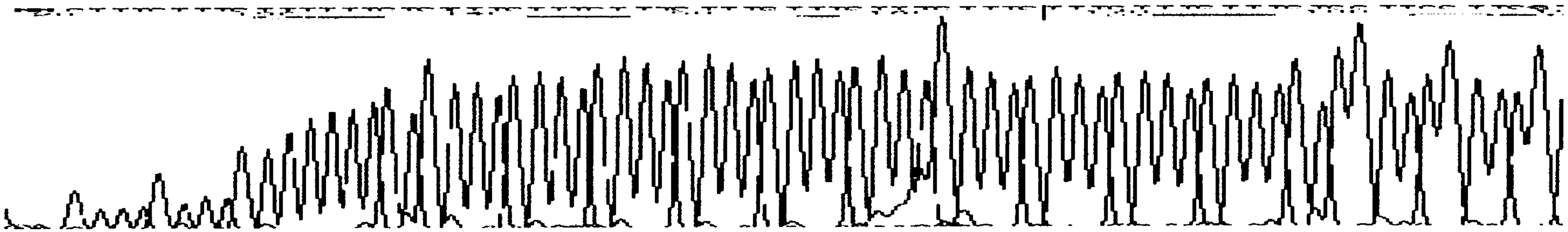


Fig.3.9. FGA-18 forward sequence.
Repeat motif: (TCCC)$_3$ TTTT TTCT (CTTT)$_{10}$ CTCC (TTCC)$_2$

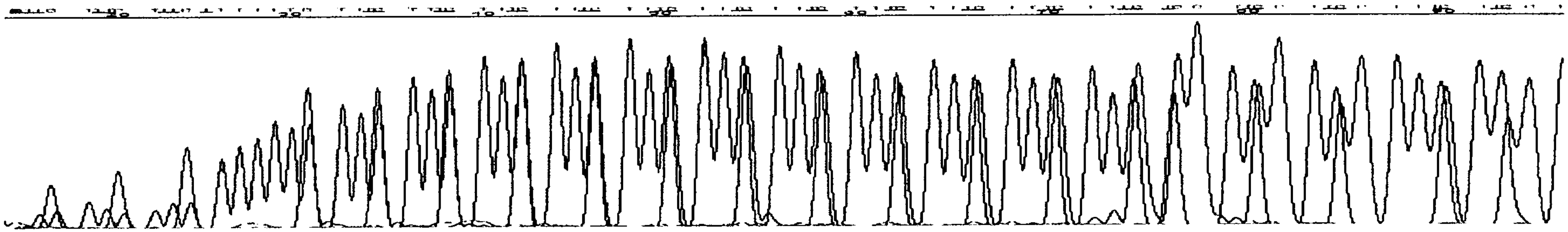


Fig 3.10. FGA-18.2 forward sequence.
Repeat motif: (TCCC)$_3$ TTTT TT(CTTT)$_{11}$ CTCC (TTCC)$_2$

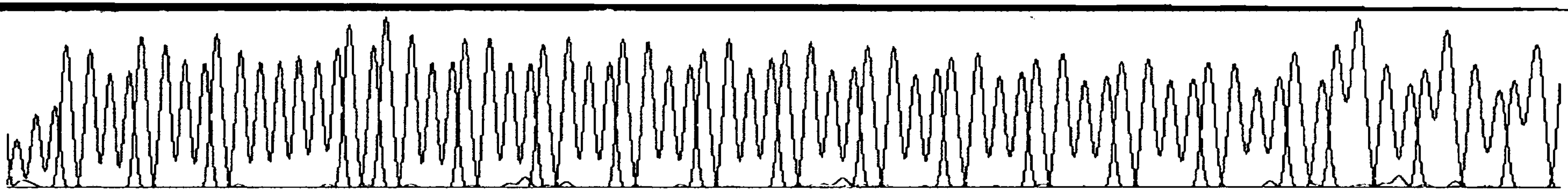


Fig. 3.11. FGA-19 forward sequence.
Repeat motif: (TCCC)$_3$ TTTT TTCT (CTTT)$_{11}$ CTCC (TTCC)$_2$

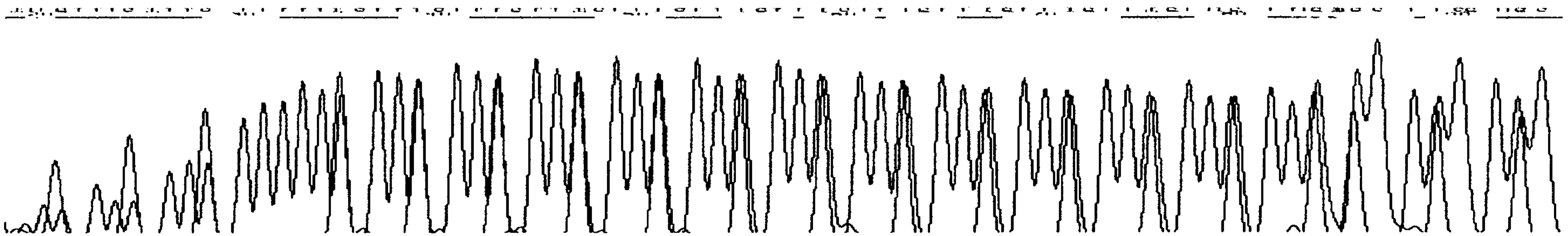


Fig 3.12. FGA-19.2 forward sequence.
Repeat motif: (TCCC)$_3$ TTTT TT(CTTT)$_{12}$CTCC (TTCC)$_2$

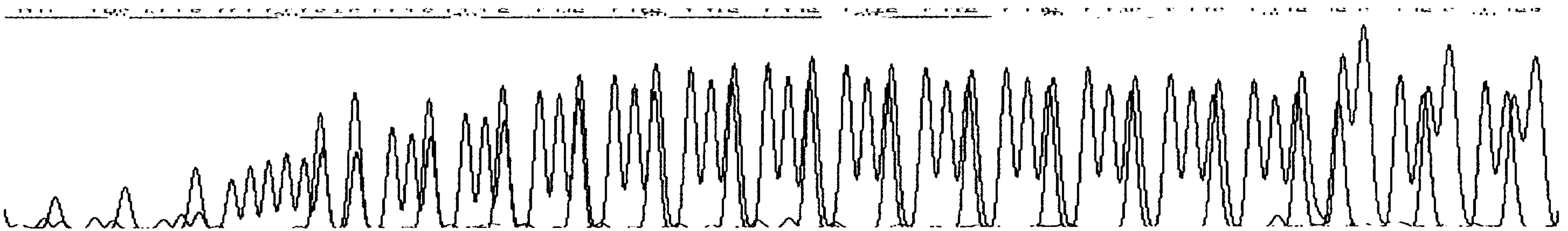


Fig.3.13. FGA-20 forward sequence.
Repeat motif: (TCCC)$_3$ TTTT TTCT (CTTT)$_{12}$CTCC (TTCC)$_2$

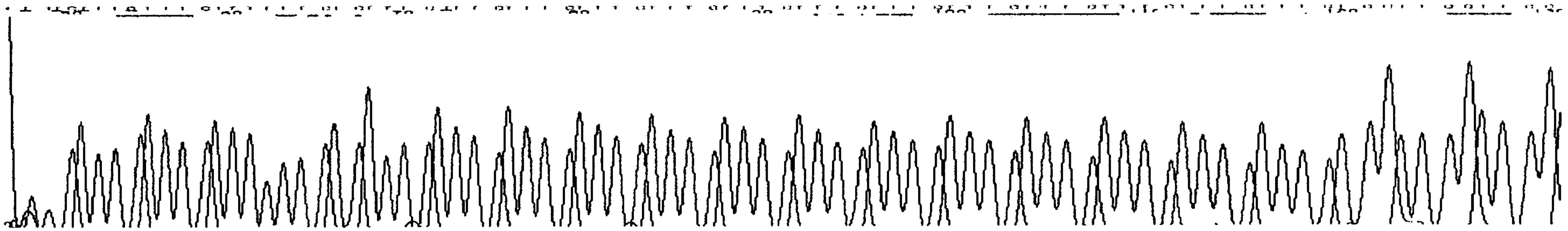


Fig. 3.14. FGA-21 forward sequence.
Repeat motif: (TCCC)$_3$TTTT TTCT (CTTT)$_{13}$ CTCC (TTCC)$_2$

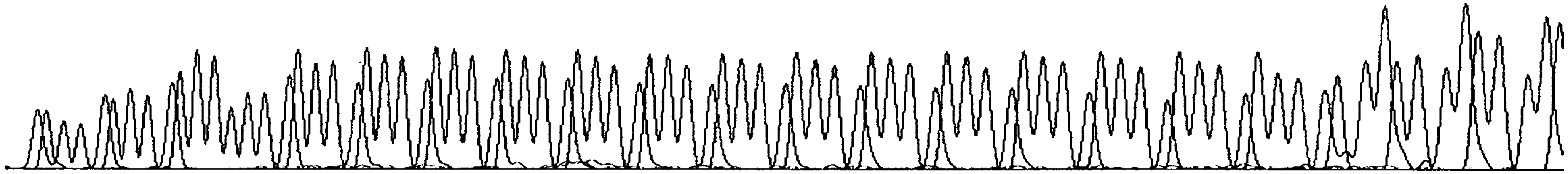


Fig. 3.15. FGA-21.2 forward sequence.
Repeat motif: (TCCC)$_3$ TTTT TT (CTTT)$_{14}$ CTCC (TTCC)$_2$

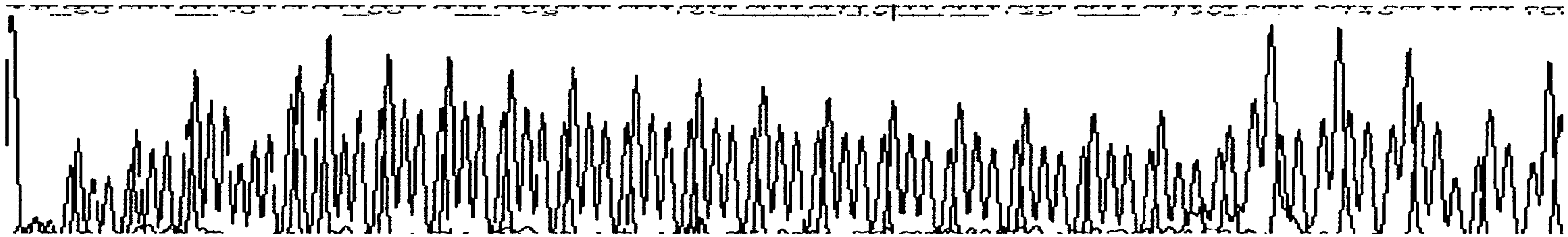


Fig. 3.16. FGA-22 forward sequence.
Repeat motif: (TCCC)$_3$ TTTT TTCT (CTTT)$_{14}$ CTCC (TTCC)$_2$

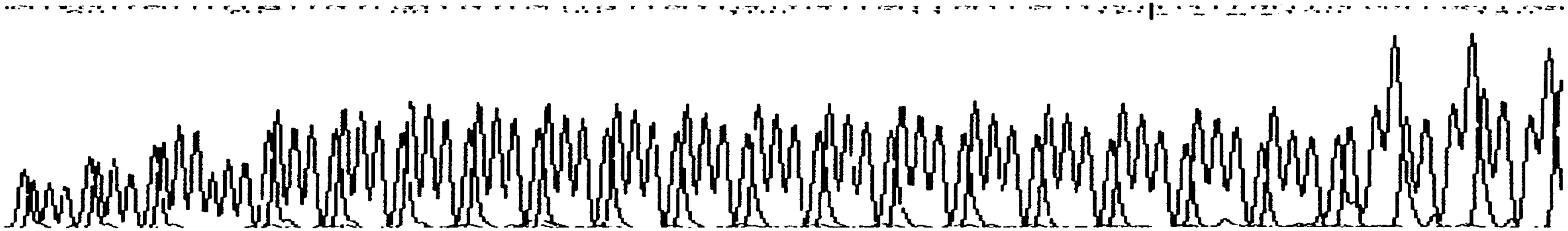


Fig. 3.17. FGA-22.2 forward sequence.
Repeat motif: (TCCC)$_3$ TTTT TT(CTTT)$_{15}$ CTCC (TTCC)$_2$

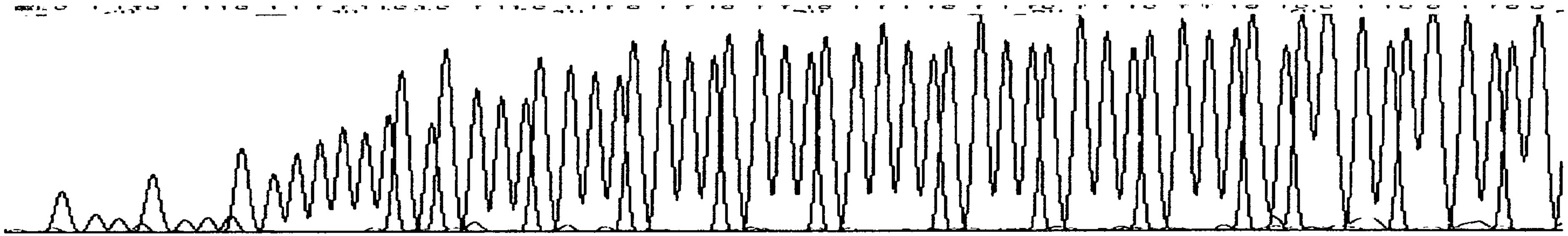


Fig. 3.18. FGA-23 forward sequence.
Repeat motif: (TCCC)$_3$ TTTT TT(CTTT)$_{11}$ CTCC (TTCC)$_2$

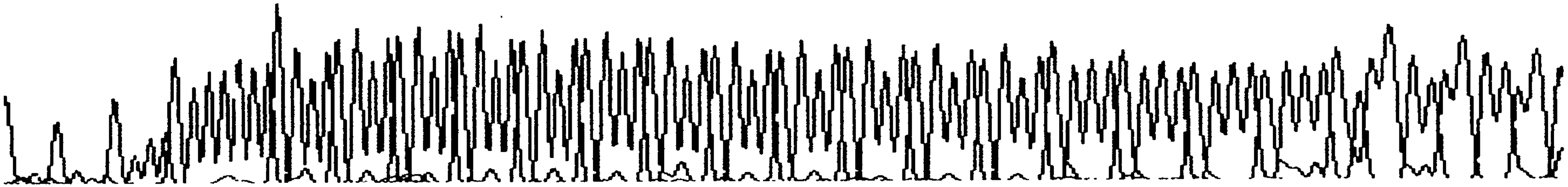


Fig 3.19. FGA-23.2 forward sequence.
Repeat motif: (TCCC)$_3$ TTTT TT (CTTT)$_{16}$ CTCC (TTCC)$_2$

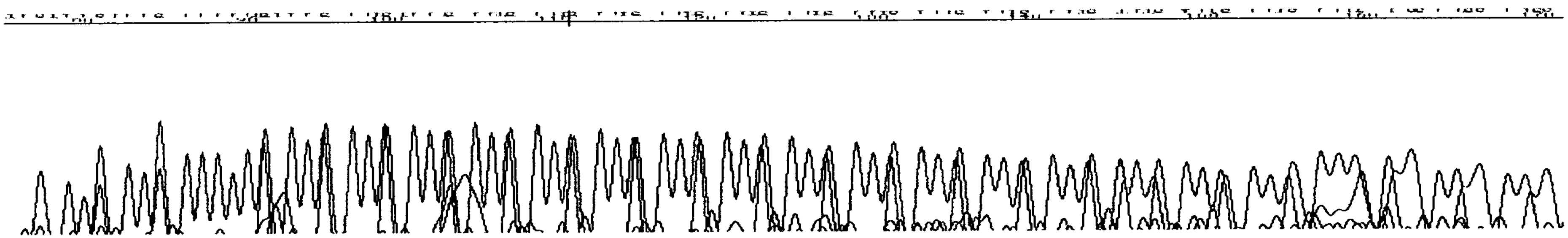


Fig. 3.20. FGA-24 forward sequence.
Repeat motif: (TCCC)$_3$ TTTT TTCT (CTTT)$_{15}$ CTCC (TTCC)$_2$

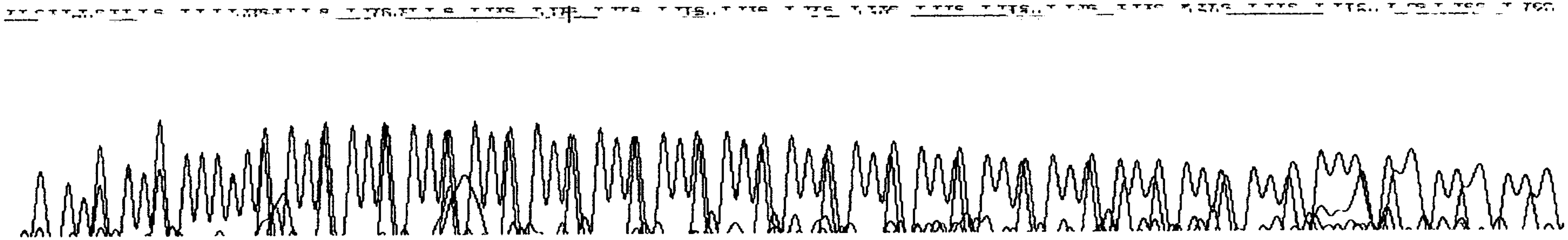


Fig. 3.21. FGA-24.2 forward sequence.
Repeat motif: (TCCC)$_3$ TTTT TT (CTTT)$_{17}$ CTCC (TTCC)$_2$

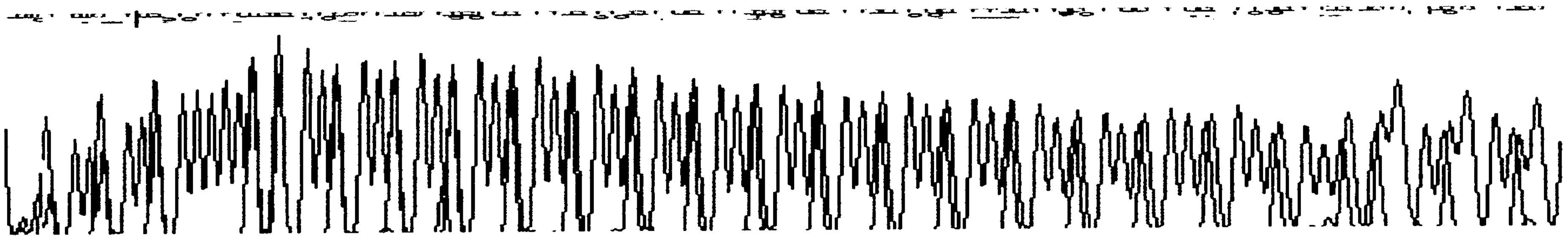


Fig. 3.22. FGA-25 forward sequence.
Repeat motif: (TCCC)$_3$ TTTT TTCT (CTTT)$_{17}$ CTCC (TTCC)$_2$

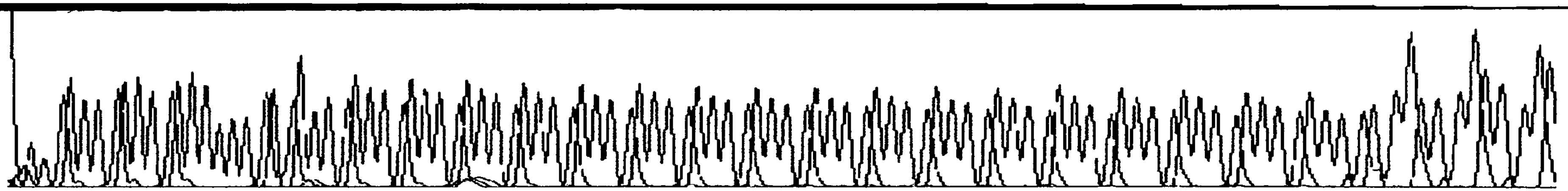


Fig. 3.23. FGA-26 forward sequence.
Repeat motif: (TCCC)$_3$ TTTT TTCT (CTTT)$_{18}$ CTCC (TTCC)$_2$

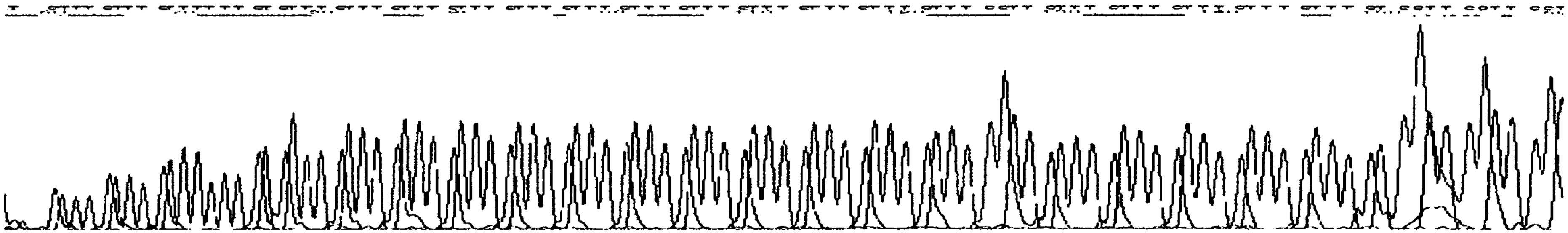


Fig. 3.24. FGA-26' forward sequence.
Repeat motif: (TCCC)$_3$ TTTT TTCT (CTTT)$_{12}$ CTCC (CTTT)$_5$ CTCC (TTCC)$_2$

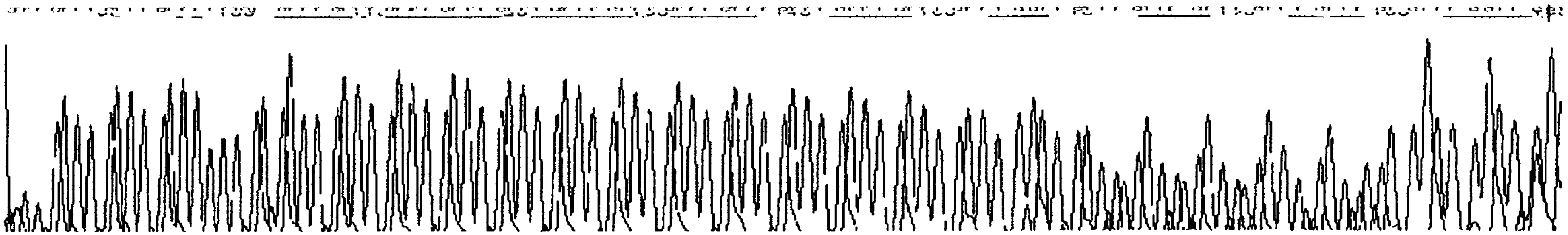


Fig. 3.25. FGA-27 forward sequence.
Repeat motif: (TCCC)$_3$ TTTT TTCT (CTTT)$_{13}$ CTCC (CTTT)$_5$ CTCC (TTCC)$_2$

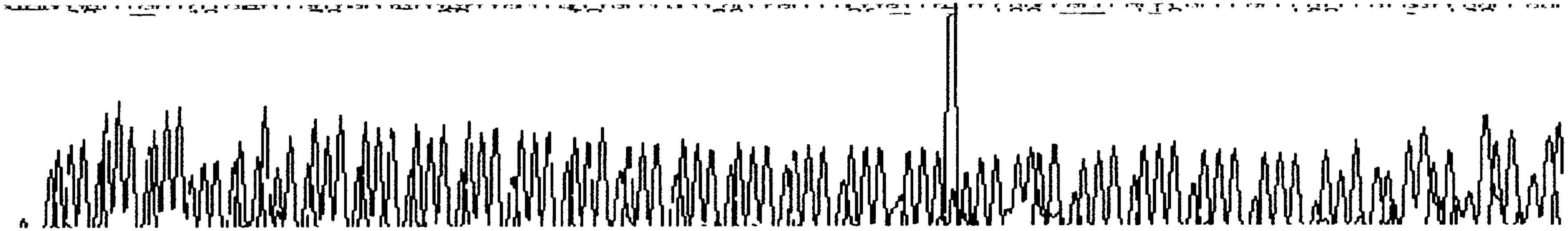


Fig. 3.26. FGA-28 forward sequence.
Repeat motif: (TCCC)$_3$ TTTT TTCT (CTTT)$_{20}$ CTCC (TTCC)$_2$

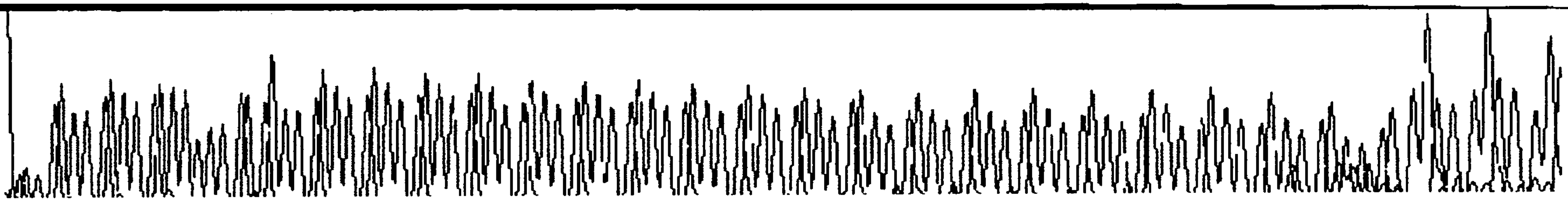


Fig. 3.27. FGA-28' forward sequence.
Repeat motif: (TCCC)$_3$ TTTT TTCT (CTTT)$_{14}$ CCTT (CTTT)$_5$ CTCC (TTCC)$_2$

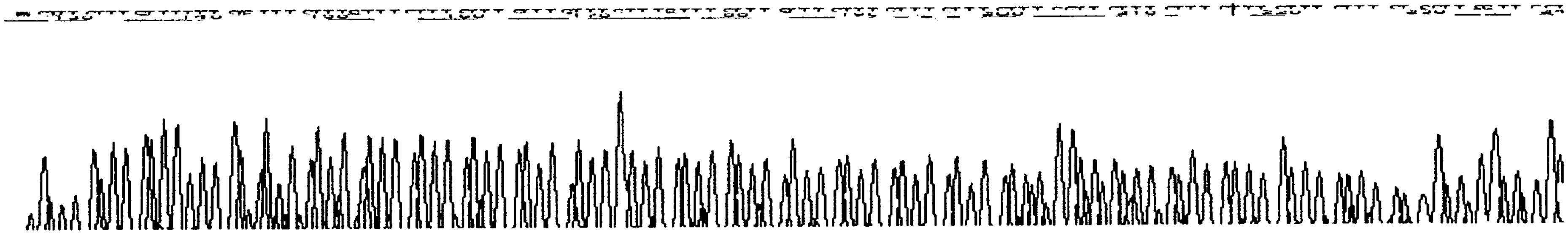


Fig. 3.28. FGA-29 forward sequence.
Repeat motif: (TCCC)$_3$ TTTT TTCT (CTTT)$_{15}$ CCTT (CTTT)$_5$ CTCC (TTCC)$_2$

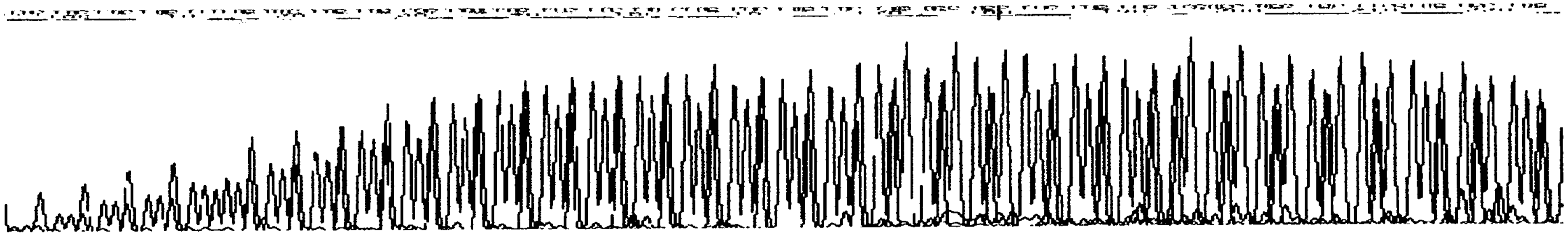


Fig. 3.29. FGA-29.2 forward sequence.
Repeat motif: (TTTTC)$_4$ TTTT TT (CTTT)$_{13}$ (CTTC)$_3$ (CTTT)$_3$ CTCC (TTCC)$_4$

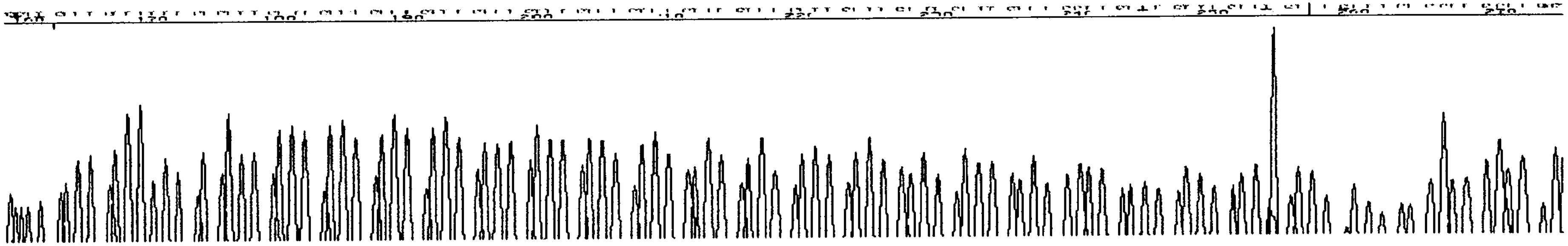


Fig. 3.30. FGA-30 forward sequence.
Repeat motif: (TCCC)$_3$ TTTT TTCT (CTTT)$_{16}$ CCTT (CTTT)$_5$ CTCC (TTCC)$_5$

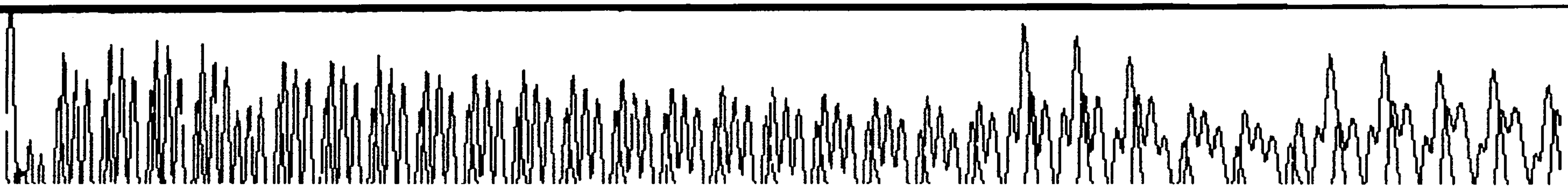


Fig. 3.31. FGA-30.2 forward sequence.

Repeat motif: (TTTC)$_4$ TTTT TT (CTTT)$_{14}$ (CTCC)$_3$ (CTTT)$_5$ CTCC (TTCC)$_2$

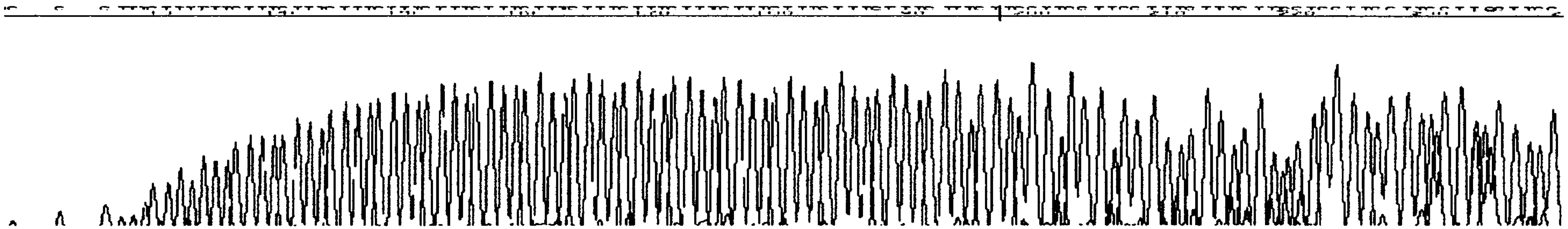


Fig. 3.32. FGA-31.2 forward sequence.

Repeat motif: (TTTC)$_4$ TTTT TT (CTTT)$_{15}$ (CTCC)$_3$ (CTTT)$_3$ CTCC (TTCC)$_2$

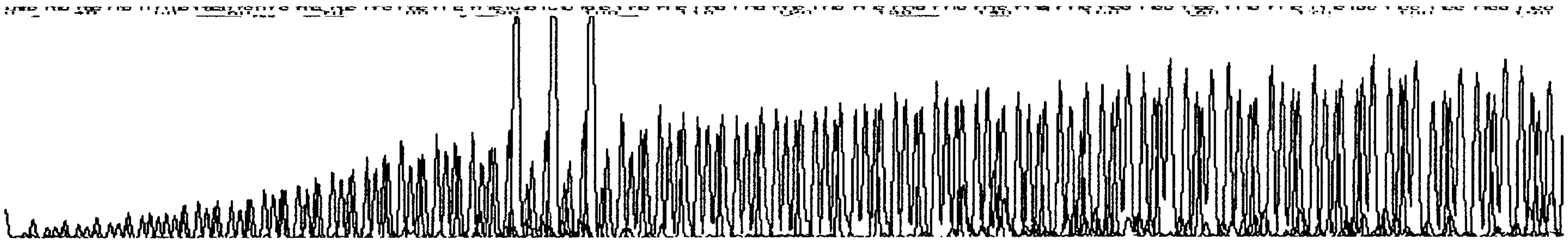


Fig. 3.33. FGA-40.2 forward sequence.

Repeat motif: (TTTC)$_4$ TTTT TT (CTTT)$_9$ (CTGT)$_3$ (CTTT)$_{12}$ (CTTC)$_3$ (CTTT)$_3$ CTCC (TTCC)$_4$

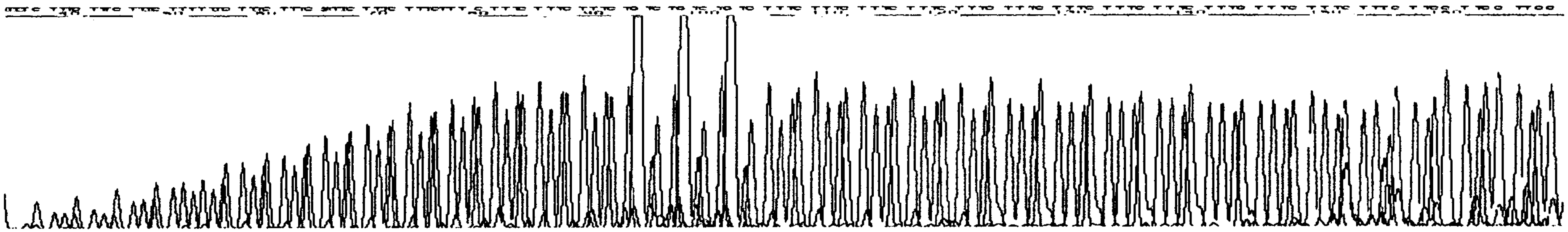


Fig. 3.34. FGA-41.2 forward sequence.

Repeat motif: (TTTC)$_4$ TTTT TT (CTTT)$_9$ (CTGT)$_3$ (CTTT)$_{13}$ (CTTC)$_3$ (CTTT)$_3$ CTCC (TTCC)$_4$

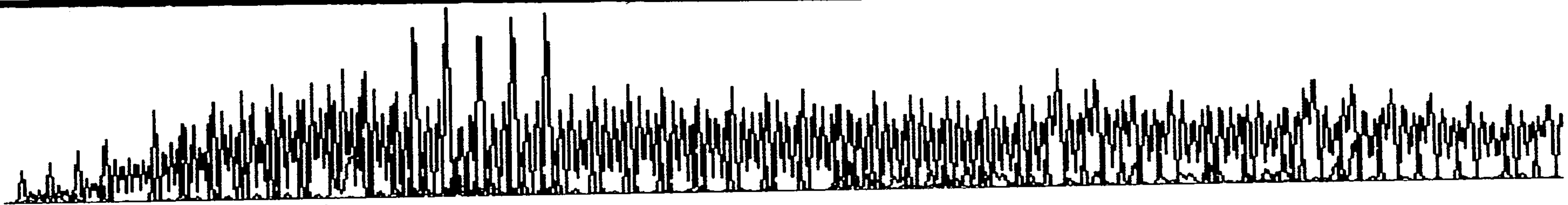


Fig. 3.35. FGA-43.2 forward sequence.
Repeat motif: (TTTC)$_4$ TTTT TT (CTTT)$_8$ (CTGT)$_5$ (CTTT)$_{13}$ (CTTC)$_4$ (CTTT)$_3$ CTCC (TTCC)$_4$

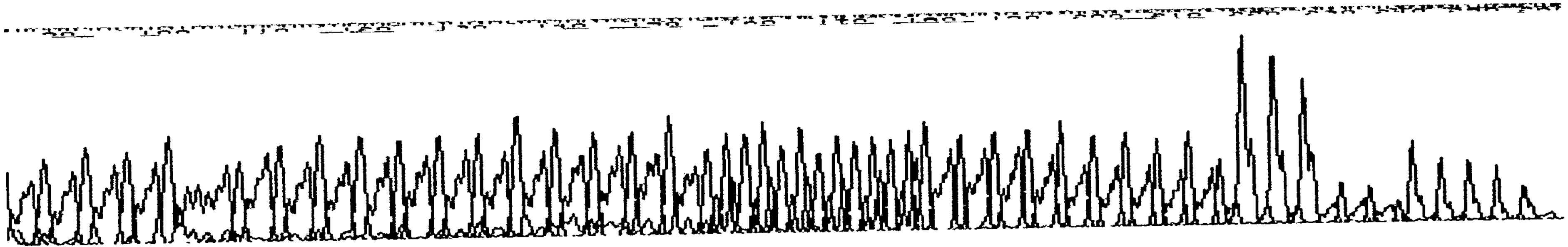


Fig. 3.36. FGA-43.2' forward sequence.
Repeat motif: (TTTC)$_4$ TTTT TT (CTTT)$_{12}$ (CTGT)$_6$ (CTTT)$_9$ (CTTC)$_3$ (CTTT)$_3$ CTCC (TTCC)$_4$

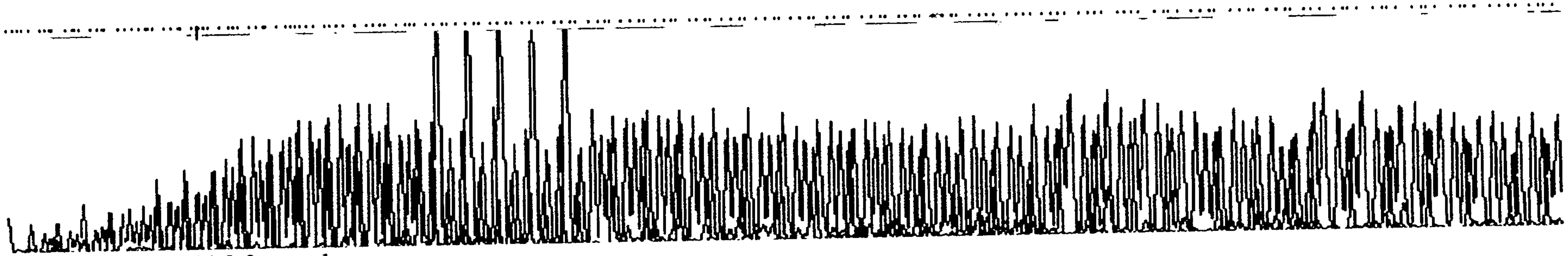


Fig. 3.37. FGA-44.2 forward sequence.

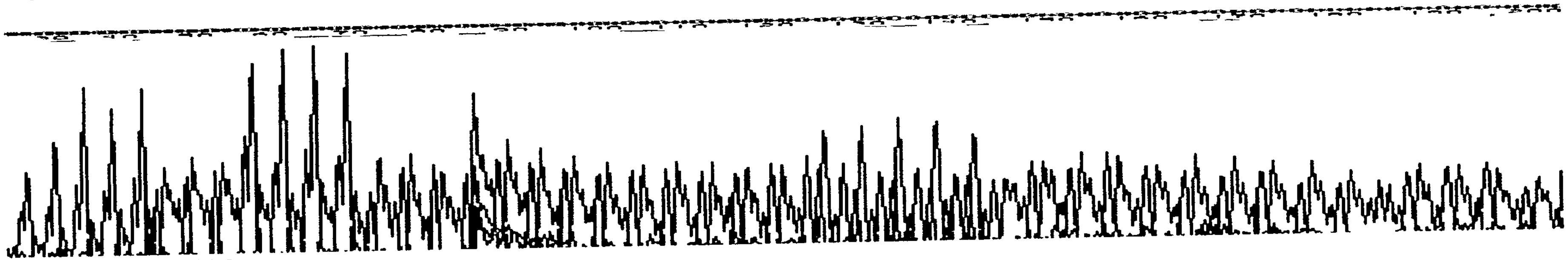


Fig. 3.38. FGA-44.2 reverse sequence.
Repeat motif: (TTTC)$_4$ TTTT TT (CTTT)$_9$ (CTGT)$_5$ (CTTT)$_{13}$ (CTTC)$_4$ (CTTT)$_3$ CTCC (TTCC)$_4$

CHAPTER 4

DISCUSSION AND CONCLUSION

In brief, the findings of this study revealed the following observations. Two new alleles, (40.2 and 41.2) which had not been described before, were found and characterised in our study. Allele 29.2 had been described before, but it's sequence structure was not known (Dr A de Kock personal communication, South African Forensic Science Lab, http://www.cstl.nist.gov/biotech/strbase/str_fga.html). Three alleles, having two sequence variations each (26', 28' and 43.2'), was also described; these alleles had been found in other studies although the 43.2' in our study was a new sequence variant. (Lazaruk K 2001, Griffiths *et al* 1998, Barber *et al* 1995, http://www.cstl.nist.gov/biotech/strbase/ str_fga.html). Another allele, 44.2, had been reported by other authors, but we describe a different sequence variant (http://www.cstl.nist.gov/biotech/strbase/str_fga.html, Lazaruk K 2001). Despite the fact that our study was not a population investigation, it agreed to other published reports where some alleles were only found in certain specific ethnic groups (Dauber EM *et al* 2001, Barber MD *et al* 1995).

The FGA locus is one of the many short tandem repeat loci that are extensively used in personal identification. Many population studies have been conducted on this locus, but only a few sequence studies have been reported. According to these studies, FGA is a highly polymorphic locus with 86 described alleles and inter-alleles known to date. The polymorphism at this locus causes it to have a high heterozygosity score as well as a high discriminatory power. The FGA thus is a very useful locus to include in paternity testing and forensic investigations.

In this laboratory the FGA locus, together with other STR loci, have been extensively employed in paternity testing. During these routine STR analyses, off-ladder and inter-alleles were observed. The off-ladder alleles are those alleles that are either larger or smaller than the alleles found in the commercial allelic ladders and the inter-alleles are the alleles that are not entirely made up of complete tetranucleotide repeat motifs.

In this study 27 normal FGA alleles and 3 sequence variants were characterized. Dauber EM *et al* (2000) reported 17 different alleles at the FGA locus, while Barber MD *et al* (1996) found 22 different alleles in their study. In a large study performed by Lazaruk K (2001) 36 normal alleles were described as well as 4 sequence variants. According to the STRBase Internet web page, several different authors have described 42 normal FGA alleles and 1 sequence variant. Combining the data from published and unpublished population studies, 86 different alleles have been reported for the FGA locus. Among these alleles, 24 are complete tetranucleotides with the remaining 62 being inter-alleles (or microvariants).

It is important to note that sequence variants are not only found in the FGA locus. Variants are also encountered in the following STR loci: D3S1358, D5S230, D8S1132, D11S488, D12S391, D14S229, D21S11, vWF2 and vWA (Zhou H-G *et al* 1997, Buscemi L *et al* 1998, Seidl C *et al* 1999, Yoshimoto T *et al* 1999, Henke L *et al* 2001, Hering S & Muller E 2001, Haddad AP & Sparrow RL 2001, Lazaruk K 2001, Momhinweg E *et al* 1998, Xiao FX *et al* 1998).

The aim of this study was, in the first instance, to characterise the sequence of the off-ladder and inter-alleles. Secondly, as many as possible of the observed alleles, found in our study population groups, were to be included in the study. These investigations were undertaken to ascertain and confirm the allelic designation and nucleotide sequence structure of the alleles at the FGA locus.

The smallest FGA allele with a known sequence structure is the FGA-15 allele (Barber MD *et al* 1996, http://www.cstl.nist.gov/biotech/strbase/str_fga.html) and the largest allele is the FGA-51.2 allele (Lazaruk K 2001). As mentioned before, the alleles of the FGA locus, based on their sequence structure, can be divided into three separate groups i.e. small, intermediate and large alleles.

In our study the small FGA alleles (between alleles 16.1 and 26) all have a similar sequence structure. They have a long stretch of CTTT repeats, followed by a CTCC $[CCTT]_2$ repeat motif at the 3' end. The general structure of these alleles is:

$[TTTC]_3$ TTTT TTCT $[CTTT]_n$ CTCC $[TTCC]_2$

and that of the inter-alleles in this group being:

$[TTTC]_3$ TTTT TT $[CTTT]_n$ CTCC $[TTCC]_2$

The number of CTTT repeats in the complete tetranucleotide alleles range from 8 in the 16.1 allele to 18 in the 26 allele. The number of CTTT repeats in the inter-alleles is usually one more than the previous complete allele with the loss of CT before the CTTT repeat motif.

The smallest allele that was sequenced in our study was the FGA-16.1 and its sequence structure was:

$[TTTC]_3$ TTTT TTCT $[CTTT]_5$ T $[CTTT]_3$ CTCC $[TTCC]_2$.

All of the FGA-16.1 alleles (n=11) that were sequenced in this study had this exact sequence with the T insertion causing it to be an inter-allele. This inter-allele was also described by Griffiths RAL *et al* (1998) (http://www/cstl/gov/biotech/strbase/ str_fga.html). It is interesting to note that during routine STR typing, using a commercial kit, two of the eleven 16.1 alleles were actually mistyped as FGA-16 alleles. The problem of mistyping occurs as the inter-alleles are not represented on the allelic ladder of the kit and it is thus not always possible to assign them correctly. The allelic ladder of the commercial kit we used had a range of alleles 18 through 30. The 16.1 allele is unique in the sense that it was only found in our Negroid people. This corresponds to the STR database, which is used by the Department of Haematology for routine paternity investigations (personal communication Dr A de Kock). Several other authors (Thompson JA *et al* 1999, Holt CL *et al* 2000, Dauber EM *et al* 2000, Lazaruk K *et al* 2001, Walsh SJ *et al* 2001) reported on this allele.

No other inter-alleles with a single base insertion were noted by this study. The only known FGA allele in this group of small alleles that we did not find in our study groups was the FGA-17 allele.

The intermediate FGA alleles range from allele 26' to allele 31.2. As before, these alleles all have a similar sequence structure. The structure of the complete alleles is:

[TTTC]$_{3-4}$ TTTT TT(CT)* [CTTT]$_{n1}$ CCTT [CTTT]$_{n2}$ CTCC [TTCC]$_{2-4}$

n1 = 12 to 16 and 20 in the 26' allele
n2 = 3 or 5
*   the CT is not found in the 29.2, 30.2 and 31.2 alleles.

The 26' allele is a variant of the 26 allele in the sense that it has a CCTT motif splitting a [CTTT]$_{17}$ repeat motif in two unequal parts. This variant of allele 26 was also described by Lazaruk K *et al* (2001).

26   [TTTC]$_3$ TTTT TTCT [CTTT]$_{18}$ CTCC [TTCC]$_2$
26'   [TTTC]$_3$ TTTT TTCT [CTTT]$_{12}$ CTCC [CTTT]$_5$ CTCC [TTCC]$_2$

The 28 and 28' alleles are similar to the 26 alleles and a CTCC unit also splits the long CTTT repeat.

28   [TTTC]$_3$ TTTT TTCT [CTTT]$_{20}$ CTCC [TTCC]$_2$
28'   [TTTC]$_3$ TTTT TTCT [CTTT]$_{14}$ CTCC [CTTT]$_5$ CTCC [TTCC]$_2$

In a sense the 28 allele fits the sequence pattern of the smaller FGA alleles, but is placed in the intermediate group solely on the basis of it's number of basepairs.

The sequence of the 29.2 allele has not been sequenced before, although its sequence pattern is no surprise.

29.2 $[TTTC]_4$ TTTT TT $[CTTT]_{13}$ $[CTTC]_3$ $[CTTT]_3$ CTCC $[TTCC]_4$

It fits well into the group of longer intermediate alleles and also follows the pattern of the inter-alleles with a loss of the CT before the longest CTTT repeat. In this allele we also, for the first time, encounter a $[TTTC]_4$ at the 5' end of the alleles and a $[CTTC]_3$ in stead of a CCTT repeat unit that splits the long CTTT repeat unit into two unequal parts. At the 3' end the motif also changes to a CCTT $[TTCC]_4$, whereas all the previous alleles had a CCTT $[TTCC]_2$ motif at this end of the allele. The 29.2 allele described in this study was only found in Negroid samples. The 30.2 and 31.2 alleles are similar to the 29 allele in their sequence and the 30 allele similar to the 26', 27, 28'and 29 alleles.

In this study, no alleles were found in the range between 31.2 and 40.2. Several authors, however, described alleles 32.2, 33.2 and 34.2 (Griffiths RAL *et al* 1998, Lazaruk K *et al* 2001, Barber MD *et al* 1996, http://www.cstl.nist.gov/biotech/ strbase/str_fga.html). These authors had also sequenced these alleles. No alleles between 34.2 and 42.2 were reported before, although we characterized alleles 40.2 and 41.2.

The last group of alleles is referred to as the large FGA alleles and again all of the alleles in this group share sequence similarity. The alleles in this group are between 40.2 and 51.2 (Lazaruk K *et al* 2001), although the only alleles encountered in this study was 40.2, 41.2, 43.2, 43.2' and 44.2. The general sequence of these alleles is:

$[TTTC]_4$TTTTTT $[CTTT]_{n1}[CTGT]_{n2}[CTTT]_{n3}[CTTC]_{n3}[CTTT]_3$CTCC$[TTCC]_4$

n1 = 8, 9 or 12
n2 = 3, 5 or 6
n3 =9, 12 or 13
n4 = 3 or 4

All of the alleles in this group are inter-alleles and thus do not possess the CT before the long CTTT repeat unit. A new motif (CTGT) is found in this group and the number of these repeats are 3 in the 40.2 and 41.2 alleles, 5 in the 43.2 and 44.2 alleles and 6 in the 43.2' allele. The 43.2 and 44.2 alleles that were described in our study had been sequenced before, although one of the 43.2 alleles and the 44.2 allele from our study were new sequence variants (Lazaruk K *et al*, 2001, http://www.cstl.nist.gov/biotech/strbase/str_fga.html). The variant allele 43.2 that was found in this study differs from other described 43.2 alleles. This is the sequence of the normal 43.2 allele that was found in our population groups and was also described by Griffiths RAL *et al* (1998).

$[TTTC]_4$ TTTT TT $[CTTT]_8$ $[CTGT]_5$ $[CTTT]_{13}$ $[CTTC]_4$ $[CTTT]_3$ CTCC $[TTCC]_4$

The 43.2' variant allele that was observed in one sample of this study had a different sequence structure and it has not been seen in any of the other population studies that we investigated.

$[TTTC]_4$ TTTT TT $[CTTT]_{12}$ $[CTGT]_6$ $[CTTT]_9$ $[CTTC]_3$ $[CTTT]_3$ CTCC $[TTCC]_4$

Lazaruk K *et al* (2001) also described a variant of the 43.2 allele and it had the following sequence.

$[TTTC]_4$ TTTT TT $[CTTT]_{13}$ $[CTGT]_3$ $[CTTT]_{11}$ $[CTTC]_3$ $[CTTT]_3$ CTCC $[TTCC]_4$

The 44.2 allele encountered in this study also differed in sequence from the published sequences.
This study:

$[TTTC]_4$ TTTT TT $[CTTT]_9$ $[CTGT]_5$ $[CTTT]_{13}$ $[CTTC]_4$ $[CTTT]_3$ CTCC $[TTCC]_4$

Lazaruk K *et al* 2001

$[TTTC]_4$ TTTT TT $[CTTT]_{13}$ $[CTGT]_4$ $[CTTT]_{10}$ $[CTTC]_4$ $[CTTT]_3$ CTCC $[TTCC]_4$

Griffiths RAL *et al* 1998, http://www.cstl.nist.gov/biotech/strbase/str_fga.html

$[TTTC]_4$ TTTT TT $[CTTT]_{11}$ $[CTGT]_3$ $[CTTT]_{14}$ $[CTTC]_3$ $[CTTT]_3$ CTCC $[TTCC]_4$

According to a few population studies, an interesting fact about these large FGA alleles (>40) is that they are predominantly found in individuals that are of recent African decent (Barber MD *et al* 1996, Thomson JA *et al* 1999, Lazaruk K *et al* 2001). Dauber EM *et al* (2001) reported that the large FGA alleles in their study were exclusively found in the Afro-Caribbean population and Barber MD *et al* (1995) also showed that some alleles were specific for certain ethnic groups. In this study, however, FGA alleles that was larger than the 30 allele were found in Negroid, Mixed Ancestry (Coloured) and Caucasian samples. The 16.1 allele was exclusively found in the Negroid samples in this study, indicating a possible ethnic restriction of this specific allele.

The occurrence of this so-called ethnic restriction of certain alleles has a possible application in forensic science. It would aid tremendously and save money and time if the ethnic affiliation of a crime sample could be deduced from the allelic profile. The search and screening strategies of crime investigators could accordingly be altered to include only individuals from a specific racial origin. To achieve this goal, large population sequencing studies would have to be conducted on all the STR loci currently in use. The difference in the allelic distribution would have to be used in conjunction with population frequency data to form the basis of such an ethnic inference system (Lowe AL *et al* 2001).

The most frequently observed alleles at the FGA are those between alleles 19 and 25. The occurrence of large FGA alleles are thus probably due to the insertion of different sequence repeats into the long CTTT repeat motif. In this regard Zhu Y *et al* (2000) reported that slippage mutations and repeat number polymorphism are more common at high repeat numbers. This is in accordance with other published data on the FGA locus where most of the reported sequence variants are those found in alleles larger than allele 30 (Lazaruk K *et al* 2001, Barber MD *et al* 1996, http://www.cstl.nist.gov/biotech/strbase/str_fga.html). In these reports 7 out of the 11 reported sequence variants were also ≥ allele 30.

Several authors also report on the investigation of accuracy and precision of FGA fragment sizes. Buscemi L *et al* (1998) reported that PCR fragment sizes, obtained by using an ABI Genetic Analyser, POP-4 polymer and comparison to a commercial internal size standard, provided excellent accuracy and precision. Lazaruk K *et al* (1998) reported a similar observation and their precision in sizing showed a standard deviation of less that 0.15 bp. In their study 99.7% of alleles that were of equal size fell within ± 0.45 bp of the allelic ladder allele size. As confirmed by sequence analysis, 4 out of 49 samples that were within-ladder were miss assigned by 1 or 2 bp. In contrast to this, all of the off-ladder alleles (27/27) in the large FGA allele range (> 30) were miss assigned by more than 2 bp. This emphasises the necessity of using an allelic ladder that spans the entire allelic range that one encounters in specific population groups.

In contrast to the above observations, Griffiths RAL *et al* (1998) reported that allele sizes determined by a specific kit were one bp larger than what was determined by sequencing. They suggest that this is a result of the ability of Taq polymerase to catalyse a non-template mediated addition of adenine to the 3' hydroxyl group of PCR products. They described this phenomenon as a (n + 1 or "+A") form. Applied Biosystems actually encourages the formation of the "+A" form by optimisation of primer sequences to encourage it and by suggesting a final extension step at 60° C for 60 minutes to allow the enzyme to complete the A addition to all double-stranded DNA product. If this step is not allowed, alleles will be represented by two peaks each, one bp apart.

According to the results of this study, however, the majority of sequenced large off-ladder alleles are smaller by two bp than those designated by the AmpFISTR Profiler Plus Kit (Applied Biosystems). The short off-ladder allele (16.1) was, however, typed as 16. Therefore it is not possible to ascribe both increment and loss of bases to the polymerase enzyme. Most probable explanations for this phenomenon would be insufficient and/or poor resolution of the typing system as reported by Gill P *et al* (1997c) and the typing efficiency affected by size fragments at both extremes, which are not represented on the allelic ladder. This again emphasises the necessity of using an allelic ladder that spans the entire allelic range that one encounters in specific population groups.

The presence of inter-alleles, off-ladder alleles and sequence variants could possibly pose problems both in paternity and forensic investigations. A problem might arise in the instance where an exclusion at only one locus is found. This fact highlights the importance of caution that has to be taken when interpreting DNA data. The general rule in paternity testing is that an exclusion is only considered when two or more loci show such an exclusion (Chakraborty R *et al* 1996, Weir BS 1996, Bein G *et al* 1998).

Pawlowski R (1999) compared the usefulness of the FGA locus with nine other commonly used STR loci. Using observed heterozygosity, power of discrimination and polymorphic information content, he reported that the FGA locus is only exceeded by the ACTBP2 locus based on these parameters. Other population studies reported a similar observation and this is also true in the population groups used in this study (personal communication Dr A de Kock) (See literature review on FGA). The increased polymorphism of the FGA locus obviously increases the discriminatory capability of this locus, although the presence of increased allelic sequence variation in a population could compromise the usefulness of a locus. It is thus advisable to set specific criteria as to when sequencing would be required to supplement routine STR typing.

Studies such as the one that was performed here, has already resulted in new allelic ladders that contain extended FGA alleles (42.2 to 51.2). Allelic ladders should thus be adjusted regularly to include new alleles. In this study, alleles 40.2 and 41.2 were reported and these alleles are not represented on any FGA allelic ladders in commercial kits that are currently available. Although these extended FGA alleles occur at low frequencies, one should include them in the allelic ladder for correct assignment when they are found.

In general, the nomenclature of FGA alleles proved to be problematic. At the 3' end of all the alleles that was sequenced in this study, a repeat motif of TTTC TTCC TTTC TTTT TT was observed. According to the rules of nomenclature of STR alleles, these repeats are not considered to be part of the allele (Gill P *et al* 1997c). This could become confusing when one attempts to assign alleles after sequencing. The possibility exists that because no variation is found in this area, it was decided not to include it as part of the allele itself.

The results of this study indicate the following:

1. Three new alleles and three sequence variants are reported (two of these sequence variants had not been reported in any previous studies).
2. Sequencing of STR loci yields an abundance of information about the specific loci being sequenced as well as general information on short tandem repeats. New alleles and variants can be found by sequencing of loci.
3. STR sequencing data could aid in the setting up of better and more representative allelic ladders.
4. Population groups differ in the STR alleles that are found and certain alleles are specific to certain population groups only. This observation needs to be confirmed by extensive population studies.
5. Ethnically restricted STR alleles can aid investigators in mass screening and search strategies.
6. It is not possible to assign off-ladder alleles correctly, and one could only type alleles accurately if they are represented in the allelic ladder.

7. The FGA locus is highly polymorphic and population studies must be performed to establish new and variant alleles as well as their distribution.
8. Once the distribution and frequency of variants at the FGA, and other loci, is known, one could consider to use this information in the calculations of paternity and match probabilities.

References

Albanese V, Biguet MF, Kiefer H, Bayard E, Mallet J, Meloni R. Quantitative effects on gene silencing by allelic variation at a tetranuclotide microsatellite. *Hum Mol Genet 2001; 10(17): 1785-1792.*

Amar A, Brautbar C, Motro U, Fisher T, Bonne-Tamir B, Israel S. Genetic variation of three tetrameric tandem repeats in four distinct Israeli ethnic groups. *J For Sci 1999; 44(5):983-986.*

Ashma R, Kashyap VK. Genetic polymorphism at 15 STR loci among three important subpopulation of Bihar, India. *For Sci Int. 2002, 130: 58-62..*

Automated DNA, PE Applied Biosystems, Dye Terminator Cycle Sequencing Kits. In Automated DNA sequencing-chemistry guide 1998 pp 2.2 - 2.11.

Barber MD, McKeown BJ, Parkin BH. Structural variation in the alleles of a short tandem repeat system at the human alpha fibrinogen locus. *Int J Leg Med. 1996; 108: 180-185.*

Bein G, Driller B, Schurmann M, Schneider PM. Pseudo –exclusion from paternity due to maternal unuparental disomy 16. *Int J Leg Med 1998; 111: 328-330.*

Berger AP, Parson W, Stenzl A, Steiner H, Bartsch G, Klocker H. Microsatellite alterations in human bladder cancer: Detection of tumor cells in urine sediment and tumor tissue. *Eur Urol 2002; 41(5): 532-539.*

Biondo R, Spinella A, Montagna P, Walsh PS, Holt C, Budowle B. Regional Italian Allele frequencies at nine short tandem repeat loci. *For Sci Int 2001; 115: 95-98.*

Blake JA, Davisson MT, Eppig JT, Maltais LJ, Povey S, White JA, Womack JE. A report on the international Nomenclature Workshop Held May 1997 at The Jackson Laboratory, Bar harbor, Maine, U.S.A. *Genomics 1997; 45: 464-468.*

Bosch E, Clarimon J, Perez-Lezaun A, Calafell F. STR data for 21 loci in northwestern Africa. *For Sci Int 2001; 116: 41-51.*

Brooks MA. Paternity Testing in Modern Blood Banking and Transfusion Practices. In Harmening D.M *1994 3rd Ed. F.A Davis company pp 465-475.*

Budowle B, Chidambaram A, Strickland L, Beheim CW, Taft GM, Chakraborty R. Population studies on three Native Alaska population groups using STR loci. *For Sci Int 2002; 129: 51-57.*

Budowle B, Nhari LT, Moretti TR, Kanoyangwa SB, Musuka E, Defenbaugh DA, Smerick JB. Zimbabwe black population data on the six short tandem repeat loci –CSF1PO, TPOX, THO1, D31358, VWA and FGA. *For Sci Int 1997; 90: 215-221.*

Buscemi L, Tagliabracci A, Sassaroli C, Bianchi F, Canestrari S, Rodriguez D. Polymerase chain reaction typing of D21S11 short tandem repeat polymorphism by capillary electrophoresis. Allele frequencies and sequencing data in a population sample from central Italy. *For Sci Int 1998; 92: 251-258.*

Butler JM, Li J, Shaler TA, Monforte JA. Reliable genotyping of short tandem repeat loci without an allelic ladder using tome-of flighyt mass spectrometry. *Int J Leg Med 1998; 112:45-49.*

Butler JM, Becker CH. Improved analysis of DNA short tandem repeats with Time-of –Flight Mass Spectrometry. *Science and technology report 2001.*

Carleton SM. The Polymerase Chain Reaction: Applications in Genomic Analysis. In Verma RS, Babu A.s' Human Chromosomes Principle and Techniques. 1995; 2nd edn. McGraw Hill inc. pp312-345.

Chakraborty R, Stivers DN. Paternty Exclusion by DNA markers: Effects of Paternal Mutations. *J For Sci 1996; 41(4): 671-677.*

Dauber EM, Glock B, Schwartz DWM, Mayer WR. Further sequence and length variation at the STR loci HumFES/FPS, HumVWA, HumFGA and D12S391. *Int J Leg Med. 2000; 113:76-80.*

De Kock A. HLA frequencies in blacks resident of Orange Free State. *Ph.D. thesis 1991*

DNA recommendation: 1991 report concerning recommendations of the DNA commission of the international society for forensic haemogenetics relating to the use of DNA polymorphisms. *For Sci Int 1992; 52: 125-130.*

DNA recommendations: 1994 report concerning further recommendations of the DNA commission of the ISFH regarding PCR-based polymorphisms in STR (short tandem repeat) systems. *Int J Leg Med 1994; 107: 159-160.*

Du Toit E. Paternity testing, In Histocompatibility testing. A Practical Approach. P Dyer and D Middleton, Oxford University Press 1993, pp 211-226.

Edwards AL, Hammond HA, Jin L, Caskey CT, Chakraborty R. Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups.*Genomics 1992; 12 :241-253.*

Egyed B, Furedi S, Angal M, Boutrand L, Vandenberghe A, Woller J, Padar ZS. Analysis of eight STR loci in two Hungarian populations. *For Sci Int 2000; 113: 25-27.*

Entrala C, Lorente M, Lorente JA, Alvarez JC, Moretti T, Budowle B, Villanueva E. Flourescent multiplex analysis of nine STR loci: Spanish population data. *For Sci Int 1998; 98: 179-183.*

Fowler JCS, Brogoyne LA, Scott AC, Harding HWJ. Repetitive Deoxyribonucleic Acid (DNA) and Human Genom Variation - A concise review relevant to forensic biology. *J For Sci 1988; 33(5): 1111-1126.*

Fung WK, Ye J, Hu L, Zhao X, Liu B, Wong DM, Law MY. Allele frequencies for nine loci in Beijing Chinese. *For Sci Int 2001; 121: 207-209.*

Gamero JJ, Romero JL, Gonzalez JL, Arufe MI, Cuesta MI, Corte-Real F, Carvalho M, Anjos MJ, Vieira DN, Vide MC. A study of ten short tandem repeat systems: African immigrant and Spanish population data. *For Sci Int 2000; 110: 167-177.*

Garofano L, Pizzamiglio M, Vecchio C, Lago G, Floris T, D'Errico G, Brembilla G, Romano A, Budowle B. Italian population data on thirteen short tandem repeat loci: HUMTH01, D2S11, D18S51, HUMVWFA31, HUMFIBRA, D8S1179, HUMTOPX, HUMCSF1PO, D16S539, D7S820, D13S317, D5S818, D3S1358. *For Sci Int 1998; 97: 53-60.*

Geada H, Birto RM, Ribeiro T, Espinheira R. Portuguese population and paternity investigation studies with a multiplex PCR –the AmpFISTR^{R} Profiler Plus^{TM}. *For Sci Int 2000; 108: 31-37.*

Gene Scan Reference Guide ABI PRISM 310 Genetic Analyzer 1997; pp 2.9-2.13.

Gill P, Kimpton C, D'Aloja E, Andersen JF, Bar W, Brinkmann B, Holgersson S, Johnsson V, Kloosterman AD, Lareu MV, Nellemann L, Pfitzinger H, Phillips CP, Schmitter H, Schneider PM, Stenersen M. Report of the European DNA profiling group (EDNAP)-towards standardization of short tandem repeat (STR) loci. *For Sci Int 1994; 65: 51-59.*

Gill P, Brinkmann B, d'Aloja E, Andersen J, Bar W, Carracedo A, Dupuy B, Eriksen B, Jangbald M, Johnsson V, Kloosterman AD, Lincoln P, Morling N, Rand S, Sabatier M, Scheithauer R, Schneider P, Vide MC. Consideration from the European DNA profiling group (EDNAP) concerning STR nomenclature. *For Sci Int 1997a; 87: 185-192.*

Gill P, d'Aloja E, Andersen J, Dupuy B, Jangblad M, Johnsson V, Kloosterman AD, Kratzer A, Lareu MV, Meldegaard M, Phillips C, Pfitzinger H, Rand S, Sabatier M, Scheithauer R, Schmitter H, Schneider P, Vide MC. Report of the European DNA profiling group (EDNAP): an investigation of the complex STR loci D21S11 and HUMFIBRA (FGA). *For Sci Int 1997b; 86: 25-33.*

Gill P, Sparkes R, Kimpton C. Development of guidelines to designate alleles using an STR multiplex system. *For Sci Int 1997c; 89: 185-197.*

Golck B, Dauber E-M, Schwartz WM, Mayr WR. Additional variability at the d12S391 STR locus in an Austrian population sample: sequencing data and allele distribution. *For Sci Int 1997; 90:197-203.*

Gomez J, Carracedo A. The 1998-1999 collaborative exercises and proficiency testing program on DNA typing of the Spanish and Portuguese Working Group of the International Society for Forensic Genetics (GEP-ISFG). *For Sci Int 2000; 114: 21-30.*

Goumenou AG, Arvanitis DA, Matalliotakis LM, Koumantakis EE, Spandidos DA. Microsatellite DNA assays reveal an allelic imbalance in p16^{ink4}, GALT, P53, and APOA2 loci in patient with endometriosis. *Fert & Ster 2001; 75(1): 160-165.*

Grattapaglia D, Schmidt AB, Silva CC, Stringher C, Fernandes AP, Ferreira ME. Brazilian population database for the 13 STR loci of the AmpFlSTR^{R} Profiler Plus^{TM} and Cofiler^{TM} multiplex kits. *For Sci Int 2001; 118: 91-94.*

Griffiths RAL, Barber MD, Johnson PE, Giibard SM, Haywood MD, Smith CD, Arnold J, Bruke T, Urquhart AJ, Gill P. New reference allelic ladders to improve allelic designation in a multiple STR system. *Int. Leg Med. 1998; 111:267-272*

Haddad AP, Sparrow RL. The short tandem repeat locus VWF2 in Intron 40 of the von Willebrand factor gene consists of two polymorphic sub-loci. *For Sci Int 2001; 119: 299-304.*

Han G-R, Song E-S, Hwang J-J. Non-amplification of an allele of the D8S1179 locus due to a point mutation. *Int J Leg Med 2001; 115:45-47.*

Harn H-J, Fan H-C, Chen C-J, Tsai N-M, Yen C-Y, Huang S-C. Microsatellite alteration at chromosome 11 in primary human nasophryngeal carcinoma in Taiwan. *Oral onc 2002; 38: 23-29.*

Helminen P, Johnsson V, Ehnholm C, Peltonen L. Proving paternity of children with deceased fathers. *Hum Genet 1991; 87:657-660.*

Henke L, Fimmers R, Josephi E, Cleef S, Dulmer M, Kenke J. Usefulness of conventional blood groups, DNA-minisatellites, and short tandem tepeat polymorphisms in paternity testing: a comparison. *For Sci Int 1999; 103:133-142.*

Henke L, Fimmers R, Reinhold J, Dulmer M, Cleef S, Arnold J, Henke J. Sequence analusis and population data on the 'new' short tandem repeat locus D5S2360. *For Sci Int 2001; 116:55-58.*

Hering S, Muller E. New alleles and mutational events in D12S391 and D8S1132 sequence data from an eastern German population. *For Sci Int 2001; 124:187-191.*

Hoff-Olsen P, Mevag B, Staalstrom E, Hovde B, Egeland, Olaisen B. Extraction of DNA from decomposed tissue: An evaluation of five extraction methods for short tandem repeat typing. *For Sci Int 1999 (105): 171-183.*

Holt CL, Stauffer C, Wallin JM, Lazaruk KD, Nguyen T, Budowle B, Walsh S. Practical application of genotypic surveys for forensic STR testing. *For Sci Int 2000; 112: 91-109.*

http://www.accessexcellence. Org/ABC/ABC/search_for_DNA.html

http://www.cstl,nist.gov/biotech/strbase/var_fga.htm

http://www.cstl.nist.gov/biotech/strbase/str_fga.htm

http://www.pbs.org/wgbh/aso/databank/entries/d053dn.html

Jeffreys AJ. DNA typing: approaches and applications *JFSS 1993; 33(4):204-211.*

Jorde LB, Carey TC, Bamshad MJ, White RC. Genetic variation and its origin and detection; In Medical genetics. 2000, 2nd edition, Mosby, pp 31-50.

Klintschar M, Al-Hammadi N, Teichenpfader B. Population genetic studies on the tetrameric short tandem repeat loci D3S1358, VWA, FGA, D8S1179, D21S11, D18S51, D5S818, D13S317, and D7S820 in Egypt. *For Sci Int 1999; 104: 23-31.*

Klintschar M, Glock B, Dauber EM, WR Mayr. Genetic variation and sequence studies of a highly variable short tandem repeat at the D17S976 locus. *Int. J Leg Med. 1998; 112: 50-54.*

Kok K, Draaijers TG, Mosselaar A, de Jong D, Buys CHCM. Inclusion of new microsatellite repeats in allelic loss analysis excludes retention of heterozygosity in renal cell carcinoma critical region in 3p21. *Cancer Genetics and Cytogenetics 2000; 116(1): 40-43.*

Kuzniar P, Soltyszsewski I, Ploski R. Population genetics of 10 STR loci in a population of Central Poland. *For Sci Int. 2002, 130 : 55-57.*

Lassen C, Hummel S, Herrmann G. Comparison of DNA extraction and amplification from ancient human bone and mummified soft tissue. *Int. J Leg Med. 1994; 107: 152-155.*

Lazaruk K, Wallin J, Holt C, Nguyen T, Sean Walsh P. Sequence variation in humans and other primates at six short tandem repeat loci used in forensic identity testing. *For Sci Int 2001; 119: 1-10.*

Lazaruk K, Walsh PS, Oaks F, Gilbert D, Rosenblum BB, Menchen S, Scheibler D, Wenz HM, Holt C, Wallin J. Genotyping of forensic short tandem repeat (STR) systems based on sizing precision in a capillary electrophoresis instrument. *Electrophoresis 1998; 19: 86-93.*

Lee JW, Lee H-S, Park M, Hwang J-J. Paternity determination when the alleged father's genotypes are unavailable. *For Sci Int 2001; 127: 150-155.*

Legrand B, De Mazancourt P, Durigon M, Khalifat V, Crainic K. DNA genotyping of unbuffered formalin fixed tissues. *For Sci Int 2002; 125: 205-211.*

Levedakou EN, Freeman DA, Budzynski MJ, Ealrly B, Damaso RC, Pollard AM, Townley AJ, Gombos JL, Lewis JL, Kist FG, Hockensmith ME, Terwilliger ML, Amiott E, McElfresh KC, Schumm JW, Ulery SR, Konotop F, Sessa TL, Sailus JS, Crouse CIA, Tosey CS, Ban JD, Nelson MS. Characterization and Validation Studies of PowerPlex $^{\text{TM}}$ 2.1, a Nine-Locus Sort Tandem Repeat (STR) Multiplex and penta D Monoplex. *J For Sci 2002; 47(4): 757-772.*

Lewin B. DNA is the genetic material in Genes V *(1994a) Oxford University Press pp81-109.*

Lewin B. Organization of the eukaryotic genome. In Genes V (*1994b) Oxford University Press pp631-797.*

Limpaiboon T, Krissadark K, Sripa B, Jearanaikoon P, Bhuhisawasdi V, Chau-in S, Romphruk A, Pairojkul C. Microsatellite alterations in liver fluke related cholangiocarcinoma are associated with poor prognosis. *Cancer Letters 2002; 181: 215-222.*

Lincoln PJ. Criticisms and concerns regarding DNA profiling. *For Sci Int 1997; 88:23-31.*

Litt M, Luty JA. A Hypervariabvle microsatellite Revealed by In vitro Amplification of a Dinucleòtide repeat within the Cardiac muscle Actin Gene. *Am. J Hum Genet 1989; 44: 397-401.*

Lowe AL, Urquhart A, Foreman A, Evett IW. Inferring ethnic origin by means of an STR profiles. *For Sci Int 2001; 119:17-22.*

Manetto G, Crivellente F, Tagliaro F, Turrina S, Pascali VL. A simplified approach to capillary electrophoretic separation of polymerase chain fragments of forensic interest. . *For Sci Int 1998; 922:59-268.*

Mangin PD, Ludes B-P. A forensic application of DNA typing. Paternity determination in a putrefied fetus. *Am. J. For-Med-Pathol 1991; 12(2): 161-163.*

Mayr WR, Brinkmann B, Rand S. Paternity testing-Quo vadis. *Blood Rev 1991; 5:51-54.*

Mertens B, Gielis M, Mommers N, Mularoni A, Lamaratine J, Heylen H, Muylle, Vandenbergh A. Mutaton of the repeat number of the HPRTB locus and structure of rare intermediate alleles. *Int J Leg Med 1999; 112: 192-194.*

Mills KA, Even DA, Murray JC. Tetranucleotide repeat polymorphism at the human alpha fibrinogen locus (FGA). *Hum Mol Genet 1992; 1: 779.*

Minaguchi K, Takenaka O. Structural variation of the VWA locus in human and comparison with non-human primates. *For Sci Int 2000; 113:9-165.*

Momhinweg E, Luckenbach C, Fimmers R, Ritter H. D3S1358: Sequence analysis and gene frequency in a German population. *For Sci Int 1998; 95: 173-178.*

Morling N, Allen RW, Carracedo A, Geada H, Uidet F, Hallengerg C, Martin W, Mayr WR, Olaisen B, Pascali VL, Schneider PM. Paternity testing commission of the International Society of Forensic Genetics: Recommendations on genetic investigations in paternity cases. *For Sci Int 2002; 129:148-157.*

Mueller RF, Young ID. Chromosomes and cell division. In EMERY'S Elements of medical genetics. 2001a; 11th edition, Churchill Livingston pp 29-53.

Mueller RF, Young ID. Mathematical and population genetics. In EMERY'S Elements of medical genetics. 2001b; 11th edition, Churchill Livingston pp 113-126.

Mueller RF, Young ID. The cellular and molecular basis of inheritance in EMERY'S Elements of medical genetics. 2001c; 11th edition, Churchill Livingston pp 11-27.

Mullis KB. The Unusual Origin of the Plymerase Chain Reaction. Sci Am 1990 (April), pp36-43.

Murray JC, Buetow KH, Weber JL, Ludwigsen S, Scherpbier-Hedemna T, Manion F, Quillen J, Sheffield VC, Sunden S, Duyk GM, Weissenbach GJ, Gyapay G, Did C, Morrissette J, Lathrop GM, Vignal A, White R, Matsunami N, Gerken S, Melis R, Albertsen H, Plaetke R, Odelberg S, Ward D, Dausset J, Cohen D, Cann HA. Coprehensive Human Linkage Map with Centimorgan density. *Science 1994; 265:2049-2054.*

Neuhuber F, Klintschar M, Radacher M. A collaborative geneic study on the STR system FGA in two Austrian population samples. *For Sci Int 1998; 91: 1-6.*

Pai C-Y, Hsieh L-L, Tsai C-W, Chiou F-S, Yang CH, Hsu B-D. Allelic alterations at the STR markers in the buccal tissue cells of oral cancer patients and the oral epithelial cells of healthy betel quid-chewers: an evaluation of forensic applicability. *For Sci Int 2002; 129:158-167.*

Pawlowski R. HUMFIBRA allele distribution in Northern Poland using capillary electrophoresis. *Int J Leg Med 1999; 112: 192-194.*

Peake IR, Bowen D, Bignell P, Liddell MB, Sadler JE, Standen G, Bloom AL. Family studies and prenatal diagnosis in severe von Willebrand Disease by polymerase chain reaction amplification of a variable number tandem repeat region of the von Willebrand factor gene. *Blood 1990; 76(3): 555-561*

Peir G, Diebold J, Lohse P, Ruebsamen H, Lohse P, Baretton GBLU. Microsatellite instability, loss of heterozygosty, and loss of hMLH1 and hMSH2 protein expression in endometrial carcinoma. *Hum path 2002; 33(3): 347-354.*

Pu C-E, Hsieh C-M, Chen M-Y, Wu F-C, Sun C-F. Genetic variation at nine STR loci in populations from the Philippines and Thailand living in Taiwan. *For Sci Int 1999; 106: 1-6.*

Pu C-E, Wu F-H, Cheng C-L, Wu K-C, Chao C-H, Li J-M. DNA short tandem repeat profiling of Chinese population in Taiwan determined by using an automated sequencer. *For Sci Int 1998; 97: 47-51.*

Rabinow P. Making PCR: A Story of Biotechnology. 1996; The University of Chicago Press pp 1-176.

Rerkamnuaychoke B, Chantratita W, Jomsawat U, Thaakitgosate J, Thanakitgosate J, Runngvithauanon T, Rojanasunan P. Database of nine tetrameric STR loci---D3S1358, vWA, FGA, D8S1179, D21S11, D18S51, D5S818, D13S317, and D7S820 in Thai population. *For Sci Int 2002; 119: 123-125.*

Ricci U, Uzielli MLG, Klintschar M. Modified primers for D12S391 and a modified silver staining technique. *Int J Leg Med 1999; 112:342-344.*

Ricci U, Sani I, Giunti L, Guarducci S, Coviello S, Uzielli MLG. Analysis of 13 tetrameric short tandem repeat loci in population of Tuscany (Central Italy) performed by means of an automated infrared sequencer. *For Sci Int 2002; 125: 83-85.*

Richards M. How distinctive is genetic information? Stud. *Hist. Phil. Biol. & Biomed .Sci. 2001; 32(4):663-687.*

Rieger R, Michaelis A, Green MM. Glossary of Genetics, Classical and Molecular. 1991; 5th edition, Springer-Verlag.

Righetti PG, Gelfi C. Recent advances in capillary zone electrophoresis of DNA. *For Sci Int 1998; 92: 239-250.*

Ross DW. Introduction to Molecular Medicine.1996 second edition, Springer pp 3-51.

Rubocki RJ, Duffy KJ, Shepard KL, McCue BJ, Shephaerd SJ, Wisecarver JL. Loss of heterozygosty detected in a Short Tandem Repeat (STR) locus commonly used for human DNA identification. *J For Sci 2000; 45(5): 1087-1089.*

Sambrook J, Russel DW. DNA Sequencing. In Molecular cloning a laboratory manual. 2001a vol 2, 3rd edn., CSHL press, pp 12.1-12.114.

Sambrook J, Russel DW. In vitro amplification of DNA by the Polymerase Chain Reaction. In Molecular cloning a laboratory manual. 2001b vol 2, 3rd edn., CSHL press, pp 8.4-8.24.

Sasaki M, Siono H, Fukushima T, Shimizu K. Human identification by genotyping of personal articles. *For Sci Int 1997; 90: 65-75.*

Schlaphoff TE, Reavis SC, Rouseau J, Creemers PC, Du Toit ED. The value of variable number of tandem-repeat polymorphisms in cases of disputed paternity nit resolved by conventional markers: two case reports. *Transfusion 1993; 33:751-753.*

Schneider PM. Basic issures in forensic DNA typing. *For Sci Int 1997; 88:17-22.*

Schumm JW. http://www.promega.com/pnotes/58/5189c.html. New approaches to DNA fingerprint analysis. *Promega corporation1996*

Siedl C, Muller S, Jager O, Seifried E. Sequence analysis and population data of short tandem repeat polymorphism at loci D8S639 and D11S488. *Int J Leg Med.1999; 112: 355-359.*

Shewale JG, Richey SL, Sinha SK. Detection and correction of a migration anomaly on a 310 Genetic Analyzer. *J For Sci 2000; 45(6): 1339-1342.*

Shiono H, Azumi J-I, Sakamoto Y, Fujiwara M, Morita M. Chromosome hetero-morphisms and paternity testing. *Am J For Med path 1985 6(3): 199-203*

Si Y, Wang J, Zaho C, Hao B, Li Y, Zhu W, Wang Y, Yu L. Allele frequencies for nine PCR –typed STR loci in a population from middle China. *For Sci Int 2002; 125: 83-85.*

Siedl C, Muller S, Jager O, Seifried E. Sequence analysis and population data of short tandem repeat polymorphism at loci D8S639 and D11S488. *Int J Leg Med.1999; 112: 355-359.*

Silver H. Paternity testing. *Critical Reviews in Clinical Loboratoty Sciences 1989; 27(5): 391-408.*

Singh D. DNA Profiling: 'Insurmountable Proof" or Exaggeration? *Med law 1995 14:445- 451*

Steinlechner M, Berger B, Niederstatter H. Rare failures in the amelogenin sex test. *Int J Leg Med 2002; 116: 117-120.*

Sudbury AW, Marinopouls J. Assessing the evidential value of DNA profiles matching without using the assumption of independent loci. *JFSS 1993; 33(2): 73-82.*

Sullivan KM, Walton A, Kimpton C, Tully G, Gill P. Flourescencce based DNA segment analysis in forensic science. *Transactions of the Biochemical Society 1993; 21(1): 116-120.*

Tagliaro F, Manetto G, Crivellente F, Smith FP. A brief introduction to capillary electrophoresis. *For Sci Int 1998; 92: 75-88.*

Tahir MA, Balamurugan K, Tahir UA, Amjad M, Awin MB, Chaudhary OR, Hamby JE, Budowle B, Herrera RJ. Allelic distribution of nine short tandem repeat (STR), HLA-DQA1, and polymarker loci in an Omani sample population. *For Sci Int 2000; 109: 81-85.*

Thangaraj K, Reddy AG, Singh L. Is the amelogenin gene reliable for gender identification in forensic casework and paternity diagnosis. *Int. J Leg Med 2002; 116: 121-123.*

The Utah marker development group. A collection of ordered tetranucleotide repeat markers from the human genome. *Am J Hum Genet 1995; 57: 619-28.*

Thomson JA, Pilotti V, Stevens P, Aures KL, Debenham PG. Validation of short tandem repeat analysis for the investigation of casts of disputed paternity. *For Sci Int1999; 100: 1-16.*

Tipett P. Blood group genetics and paternity test in Paternity testing. American Association of Blood Banks. 1978; pp 1-33.

Trivedi R, Chattopadhyay P, Maity B, Kashyap VK. Genetic polymorphism at nine microsatellite loci in four high altitude Himalayan desert human populations. *For Sci Int 2002; 127:150-155.*

Tsuneizumi M, Emi M, Hirano A, Utada Y, Tsumagari K, Takahashi K, Kasumi F, Akiyama F, Sakamoto G, Kazui T, Nakamura Y. Association of allelic loss at 8p22 with prognosis among breast cancer cases treated with high-dose adjuvant chemotherapy. *Canc Lett 2002; 180: 75-82.*

Urquhart A, Kimpton CP, Downes TJ, Gill P. Variation in short tandem repeat sequences - a survey of twelve microsatellite loci for use as forensic identification markers. *Int J Leg Med 1994; 107:13-20.*

Urquhart, A., Oldroyd, N.J., Kimpton, C.P. and Gill, P. Highly discriminating heptaplex short tandem repeat PCR system for forensic identification. *BioTechniques 1995; 18: 116-121.*

Van Hoofstat DEO, Haerinck SAA, Deforce DLD. Flemish population data using the AmpFlSTR Profiler^{TM} PCR amplification kit. *For Sci Int 2002; 127: 156-157.*

Van Oorschot RAH, Gutowski SJ, Robinson SL. HUMTHO1: Amplification, species specificity, population genetics and forensic applications. *Int J Leg Med 1994; 107:121-126.*

Wain HM, Bruford EA, Lovering RC, Lush MJ, Wright MW, Povey S. Guidelines for human gene nomenclature. *Genomics 2002; 79(4): 464-470.*

Walker RH. Probabilty. In the analysis of paternity test results in paternity testing. American Association of blood banks 1978; pp 69-114.

Walsh SJ, Cullen JR, Harbison SA. Allele frequencies for the four major sub-populations of New Zealand at the 10 AmpFlSTR SGM Plus loci. *For Sci Int 2001; 122: 189-195.*

Watson S, Allsop R, Foreman L, Kelsey Z, Gill P. Sequenced allelic ladders and population genetics of a new STR multiplex system. *For Sci Int 2001; 115: 207-217.*

Weber JL, May PE. Abundant class of human DNA polymorphisms, which can be typed using the polymerase chain reaction. *Am J Hum Genet 1989; 44:388-396.*

Weir BS. Individual identification in Genetic Data Analysis II; 1996; Sinauer Associates, Inc publishers pp 31-111 and 209-215.

White JA, McAlpine PJ, Antonarakis S, Cann H, Eppig JT, Frazer K, Frezal J, Lancet D, Nahmias J, Pearson P, Peters J, Scott A, Scott H, Spurr N, Talbot Jr C, Povey S. Guidelines for Human Gene Nomencalature. *Genomics 1997; 45:468-471.*

Wiegand P, Meyer E, Brinkmann B. Microsatellite structures in the contxt of human evolution. *Electrophoresis 2000; 21: 889-895.*

Wolfe SL. Historical Origins if Cell and Molecular Biology. In Molecular and Cellular biology. *1993a; Wadsworth: pp 29-37.*

Wolfe SL. Major investigative Methods of Cell and Molecular Biology. In Molecular and Cellular Biology. 1993b; Wadsworth: pp 140-141.

Wolfe SL. Organization of the genome and genetic rearrangements. In Molecular and Cellular Biology. *1993c; Wadsworth: pp 750-771.*

Xiao FX, Gilissen A, Cassiman J-J, Decorte R. Quadruplex fluorescent STR typing system (HUMVWA, HUMTHO1 and HPRT) with sequence-defined allelic ladders Identification of a new allele at D21S11. *For Sci Int 1998; 94:39-46.*

Yamamoto T, Tamaki K, Huang X-L, Yoshimoto T, Mizutani M, Uchihi R, Katsumata Y, Jefferys AJ. The application of minisatellite variant repeat mapping by PCR (MVR-PCR) in a paternity case showing false exclusion due to STR mutation. *J For Sci 2001; 46(2): 374-378.*

Yokoi T. Nata M, Odaira T, Sagisaka K. Hypervariable regions of DNA for parentage testing and individual identification. *Z Rechtsmed 1990; 103: 487-489.*

Yoshimoto T, Yamamoto T, Uchihi R, Tamaki K, Huang X-L, Mizutani M, Tanaka M, Armour JAL, Katsumata Y. A new triplex STR system without irregular alleles by silver staining and its potential application to forensic analysis. *J For Sci 2001; 46(3): 448-452.*

Yoshimoto T, Tamaki K, Katsumata S, Huang XL, Uchihi R, Tanaka M, Uchida H, Yamamoto T, Chen S, Armour JAL, Katsumata Y. Sequence analysis of alleles at a microsatellite locus locus D14S299 (wg1c5)and population genetic comparison. *Int J Leg Med 1999; 113: 15-18.*

Zhang X-W, Lan L, Huo Z-Y, Duan B-Z, Kobilinsky L. Restriction fragment length polymorphism analysis of forensic science casework in the People's Republic of China. *J For Sci 1991; 36 (2): 531-536.*

Zhou H-G, Sato K, Nishimaki Y, Gang L, Hasekura H. The HumD21S11system short tandem repeat DNA polymorphisms in Japanese and Chinese. *For Sci Int 1997; 86: 109-118.*

Zhu Y, Strassmann JE, Queller DC. Insertions, substitutions, and the origin of microsatellites. *Genet. Res. 2000; 76: 227-236.*

ABSTRACT

The human alpha fibrinogen (FGA) short tandem repeat locus is found in the long arm of chromosome 4. It is located in the third intron of the alpha fibrinogen gene. This complex highly polymorphic tetranucleotide repeat locus together with other STR DNA markers is extensively used in personal identification in medical and forensic sciences. STRs are also used to study genetic variation in distinct ethnic groups and in disease diagnosis. More than 80 alleles have been reported for this locus from various population frequency studies. A few sequence studies have also reported 11 sequence variants to date. The FGA locus was found to have high heterozygosity and power of discrimination.

The aim of this study was to characterise the sequence of microvariant and off-ladder complete and microvariant alleles of the FGA locus that were observed during routine paternity analyses. The characterization of the sequence of as many as possible of alleles observed in our study population would also be attempted.

A total of 62 DNA specimens were selected and sequence characterized either for one or both alleles. The DNA specimens were 52 from Negroid, 5 mixed ancestry, 4 Caucasian and 1 SAN origin. The PCR reaction was used to amplify the selected alleles. The band of interest was cut from the gel and purified in consecutive PCR and purification steps till separate single bands were obtained. The purified single bands were sequenced using a BigDye terminator ready reaction kit in both forward and reverse reactions separately. These products were precipitated with ethanol acetate and subjected to capillary electrophoresis on an ABIPrism 310 Genetic Analyser using POP6 polymer. The results were analysed using the "Sequence Analysis Software version 2.1". The data obtained were checked, printed and compared with STR analysis results and FGA sequence reports.

From the selected 62 specimens a total of 76 complete and microvariant alleles, the size of which ranged between 16.1(224bp) to 44.2 (337bp) were found. These represent 27 different alleles (13 complete and 14 interalleles). In this study 2 novel (previously undescribed) alleles (40.2 and 41.2) were found. Three sequence variants (26, 28 and 43.2) with two variants each were observed. Two alleles 43.2 and 44.2 that had reported sequence variants were found to have different sequence structures from the published sequences. Forty-nine of the 76 sequenced alleles were within ladder and the remaining 29 were off-ladder. Only 8.41% (4/49) of the within ladder alleles had been wrongly assigned allelic numbers with routine STR analysis. The difference between the routine assignment and the sequencing of these alleles was only 1 or 2 bp. In contrast, all of the 29 off-ladder alleles were wrongly assigned. In this instance the difference was 2 or more base pairs.

Although this study was conducted on conveniently selected DNA samples, it had significant results. Three sequence variants, 2 new alleles and, 1 allele, which had been reported, but sequence had not been described was found.

Additionally, two other alleles with reported sequences were found, but their sequence structure differed from the published sequences.

The samples in this investigation were not representative of the population groups that are found in the Free State province and we suggest further population-based studies of STR loci that are commonly used in paternity and forensic investigation. The information obtained from such studies will disclose the frequency of sequence variant alleles.

Key terms

Paternity testing, Forensic, Short Tandem Repeats (STRs), DNA markers, Polymerase chain reaction (PCR), DNA sequencing, Sequence variants, Electrophoresis, FGA, Microvariant alleles, Tetranucleotide repeats.

ABSTRAK

Die menslike alfa fibrinogeen (FGA) kort tandem herhalingslokus (STR) is geleë op die lang arm van chromosoom 4. Dit kom voor in die derde intron van die alfa fibrinogeen geen. Hierdie komplekse hoogs polimorfiese tetranukluotied herhalingslokus word saam met ander STR DNA merkers aangewend vir persoonlike identifikasie in die mediese en forensiese wetenskappe. STR lokusse kan ook aangewend word om genetiese variasie in etniese groepe te bestudeer asook in siekte assosiasie studies. Tydens veskeie populasie frekwesie studies is daar al meer as 80 allele in die FGA lokus gevind. Basispaaropeenvolgingstudies het ook al 11 variant allele aangetoon. Die FGA lokus is aangetoon om 'n hoë heterosigisiteit en krag van diskriminasie te vertoon.

Die doel van hierdie studie was om die basispaaropeenvolging te bepaal van die mikrovariant- en buite alleliese leer allele wat tydens roetine vaderskapsondersoeke gevind is. Die basispaaropeenvolging van so veel al moontlik van die allele van die FGA lokus sou ook bepaal word.

Twee-en-sestig DNA monsters is gekies en die basispaaropeenvolging van een of beide allele is bepaal. Die DNA monsters was afkomstig van swart (52), kleurling (5), blanke (4) en boesman (1) persone. Die geselekteerde allele is met behulp van PKR vermeerder. Die spesifieke alleelband is uit die gel gesny en gesuiwer in herhaalde PKR en suiwerings stappe totdat 'n suiwer enkel alleelband gevind is. Die suiwer allele se basispaaropeenvolging is daarna bepaal met behulp van die "BigDye terminator" reaksie. Voorwaartse- en tru-reaksies is apart uitgevoer. Die produkte van die reaksies is met etanolasetaat neergeslaan waana kapilêre elektoforese in POP-6 polimeer gevolg het op 'n ABIPrism 310 genetiese analiseerder. Die data is verwerk met "Sequence Analysis" sagteware. Die resultate is nagegaan, gedruk en vergelyk met STR analises en FGA basispaaropeenvolgings publikasies.

Uit die 62 DNA monsters is 76 volledige- en mikrovariantallele gevind wat wissel vanaf 'n 16.1 (224bp) tot 'n 44.2 (337bp) alleel. Hierdie allele verteenwoordig 27 veskillende allele (13 volledig en 14 inter-allele). In hierdie studie is twee nuwe allele (40.2 en 41.2) beskryf. Daar is ook variasie in die basispaaropeenvolging gevind in allele 26, 28 en 43.2 met twee variante elk. Twee allele, 43.2 en 44.2, se basispaaropeenvolging het ook verskil van die gepubliseerde data. Van die 76 allele waarvan die basispaaropeenvolging bepaal is, was 49 binne die alleliese leer en 29 buite die alleliese leer. Slegs 8.41% (4/49) van die allele binne die alleliese leer is verkeerd aangetoon tydens roetine STR analise. Hierdie 4 allele het verskil met 1 of 2 basispare. In teenstelling hiermee is al 29 buite alleliese leer allele verkeerd aangetoon met 'n verskil van meer as twee basispare.

Alhoewel hierdie studie uitgevoer is op geselekteerde DNA monsters is betekenisvolle resultate verkry. Drie variante in basispaaropeenvolging, twee nuwe allele sowel as een alleel waarvan die opeenvolging onbekend was, is beskryf. Daar is verder ook twee allele beskryf waarvan die struktuur van die opeenvolging verskil van dit wat in die literatuur beskryf is.

Aangesien die DNA monsters wat in hierdie studie gebruik is, nie verteenwoordigend is van die bevolkingsverspreiding van die Vrystaat Provinsie nie, stel ons verdere populasie studies voor om die STR verspreiding van lokusse, wat algemeen aangewend word in vaderskaps- en forensiese ondersoeke, deeglik te beskryf. Sodanige studies sal lig werp op die frekwensie van variasie in die alleel basispaaropeenvolging van STR lokusse.

Sleutelwoorde

Paternity testing, Forensic, Short Tandem Repeats (STRs), DNA markers, Polymerase chain reaction (PCR), DNA sequencing, Sequence variants, Electrophoresis, FGA, Microvariant alleles, Tetranucleotide repeats.

Appendix A

DNA SEQUENCE CHARACTERISATION OF THE FGA STR LOCUS IN THE FREE STATE POPULATION OF SOUTH AFRICA

André de Kock*, Estifanos Kebede

*Dept of Haematology and Cell Biology, School of Medicine, Faculty of Health Sciences, University of the Free State, PO Box 339 (G2), Bloemfontein, 9300, South Africa.
+27-51-4053283(T) +27-51-3331036(F) E-mail: gnhmadk@med.uovs.ac.za

Abstract

The FGA STR locus is a complex highly polymorphic, tetranucleotide DNA locus, that is used extensiveily, together with other markers, in personal identification. Population frequency studies reported more than 80 alleles and 11 sequence variants at this locus. The aim of this study was to characterizs the sequence of off-ladder and microvariant FGA alleles in the local population. A total of 62 DNA specimens were selected and isolated from the blood of paternity test clients from different racial origins. These were amplified by PCR, purified and sequenced. Two novel alleles (40.2 and 41.2) and one reported but sequence undescribed allele (29.2) were observed. Variants of alleles 43.2 and 44.2, that vary from reported data, were also identified. The existence of many sequence variants complicates the analysis and interpretation of DNA STR data. However, if alleles are ethnically restricted, they could aid in placing an offender in a specific ethnic group and thus limit the number of suspects that are investigated.

Key words: Paternity testing; STR; Sequence variants; FGA; off-ladder

1. Introduction

A major component of non-coding sequences in DNA are repetitive elements [1]. Most forensic and paternity typing systems are based on these tandemly repeated sequences. Since their discovery in 1989, short tandem repeats (STRs) have become widely used due to their ease of use, high polymorphism, small size (degraded DNA can be used), potentiality for multiplexing and amenability for use in a PCR reaction [1,2,3]. STRs are widely distributed throughout the human genome, occurring with frequency of 1 locus every 6 to10 Kb and are composed of tandemly repeated sequences of 2 to 5 bp in length [4]. For forensics and paternity testing only a limited number of markers from the thousands that have been generated for genetic mapping are used [5,6]. The most frequently used STR systems typically contain short 4 to 5 bp repeat motifs [7,8], and are simple, compound or complex [9,10,11]. Complex repeats may contain several repeats at blocks of variable length along with variable intervening sequences. Very often a high degree of polymorphism is associated with remarkable sequence variation [12].

Structural and sequence variations of particular STR loci among different populations have been reported [8,13,14]. This variation at STR loci is due to the length of the individual repeat units, the number of repeat units and the repeat unit pattern of individual alleles [13]. Furthermore, sequence variation occurs as a result of larger insertions or deletions, the frequency of which is estimated to be 1 in 250 to 300 nucleotides [3]. Hence, accurate typing requires a precise knowledge of the structural variation of alleles [13].

The human alpha fibrinogen locus (FGA) is widely used in forensic DNA testing as well as in paternity investigations [7,10,13]. This locus is also known as HUMFIBRA [10] and HUMFGA [13]. The FGA locus is found on the long arm of chromosome 4 and is located in the third intron of the human alpha-fibrinogen gene that contains repeat sequences beginning at nucleotide 2912 [4,13,15]. It is a complex tetranucleotide repeat with the complete alleles differing in length by 4 bp, but also containing interalleles differing by 2 bp. However, alleles that differ in 1 bp have also been reported [3,4].

Our laboratory has been performing DNA STR for paternity testing since 1999, using the AmpF*l*STR Profiler Plus and later the AmpF*l*STR Identifiler kit from Applied Biosystems. During routine investigations certain off-ladder and microvariant alleles were observed in the FGA and other loci, especially in the local Negroid population. It is difficult to size some of the larger (>30) FGA alleles, as these are not included in the Allelic Ladder of the AmpF*l*STR Profiler Plus kit. The aim of this study was to sequence the variant FGA alleles in the local population to: (1) characterize these alleles, and (2) determine the possibility of ethnic restriction of these alleles.

2. Materials and methods

2.1 DNA samples

DNA samples, previously isolated from paternity clients using the Promega Wizard kit, were selected to cover as many of the FGA alleles and variants as possible. The ethnic affiliation of each individual was documented at the time of phlebotomy by self-declaration. The study was conducted in two phases. In both phases a total of 62 specimens were sequenced for either one or both alleles. The racial origins of the specimens were 52 from the Negroid, 5 from the Mixed Ancestry, 4 from the Caucasian and 1 from SAN. In the first phase of the study samples were selected to include all previously undescribed microvaraint and off-ladder alleles observed in routine paternity testing. In the second phase of the study samples were selected to cover all possible alleles other than those included in the first phase.

2.2 PCR amplification

Segments of the FGA locus were amplified from genomic DNA using the following oligonucleotide primers; primer 1: 5'-ATT ATC CAA AAG TCA AAT GCC CCA TAG G-3'; and primer 2: 5'-ATC GAA AAT ATG GTT ATT GAA GTA GCT G-3'. Amplification was performed in a 25 μl mixture containing 10 ng of genomic DNA, 200 μM of each dNTP (Promega), 25 pmol of each primer (MWG-Biotech AG), 1 unit of Amplitaq (Roche) in Thermophilic DNA polymerase buffer (10 mM Tris-HCl, 50 mM KCl, 0.1% Triton X100) and 1.5 mM $MgCl_2$. Samples were amplified in a GeneAmp PCR 2400 thermocycler (Applied Biosystems) using the following parameters: denaturation at 94°C for 5 min followed by 35 cycles of denaturation at 94°C for 30s, annealing at 60°C for 30s and extension at 72°C for 30s. A final extension followed at 72°C for 5 min. The PCR product (5-10 μl) was mixed with bromophenol-blue sucrose loading solution and loaded on a 2.5% agarose gel containing 0.05 μg/ml ethidium bromide. Electrophoresis followed in 1 x TBE buffer for 2 hours at 80 volts. DNA bands were visualized on a UV transilluminator, excised from the gel with a sterile surgical blade, cut into small pieces and purified using GenELute minus EtBr spin columns (Sigma). Five µl aliquots of the purified PCR product were amplified in 2nd to 4th rounds of PCR (parameters as for 1st PCR) until a pure single band was obtained.

2.3 Sequencing reactions

Single band PCR products were purified using the High Pure PCR product purification Kit (Roche) to remove primers, dNTP's and Taq polymerase. The PCR products were sequenced using the ABI Prism Big Dye Terminator Ready reaction version 3 kit (Applied Biosystems). The forward sequencing reaction was performed using primer 1 under the following conditions: 25 cycles of denaturing at 96°C for 10s, annealing at 60°C for 5s and extension at 72°C for 30s. Reverse strand sequencing was performed using primer 2 and the same parameters except for annealing at 55°C. The sequence products were precipitated using ethanol-acetate. After precipitation 25 µl of template suppression reagent (Applied Biosystems) was added to each product and denatured at 95°C for 3 min. Samples were flash cooled on ice and run on a ABI Prism 310 Genetic Analyser using POP-6 polymer. Sequences were analysed using the ABI Prism Sequencing Analysis vs.3 Software.

3 Results

In this study a total of 76 alleles (complete tetranucleotides and inter alleles), representing 27 different alleles were sequenced both by forward and reverse priming (table 1).

In the first phase, 37 samples were included: 32 Negroid, 3 Mixed Ancestry and 2 Caucasian. A total of 38 DNA fragments representing 9 different alleles were selected for investigation. Of these samples 25 were off-ladder alleles (alleles greater than the largest allelic ladder i.e. 30). According to the designation by routine STR analysis, within ladder alleles that were sequenced in this group were 16.1(n=9), 24.2(n=1) and 29.2(n=3). The off-ladder alleles that were sequenced were 31.2(n=2), 32.2(n=1), 41(n=6), 42(n=1), 44(n=14) and 45(1) respectively. After sequence characterization 8 different alleles were observed. These were 16.1(n=9), 24.2(n=1), 29.2(n=3), 31.2(n=1), 40.2(n=5), 41.2(n=1), 43.2(n=17) and 44.2(n=1). Thirteen within ladder alleles (16.1, 24.2 and 29.2) were all correctly assigned by routine STR analysis while all (25/25) of the off ladder alleles were wrongly assigned. From the 25 wrongly assigned off ladder alleles 80%(20/25) had only a 2bp difference from the sequence confirmed allele length. The remaining 20%(5/25) had a difference of four or more base pairs.

In the second phase a total of 26 specimens (22 Negroid, 2 Mixed Ancestry, 1 Caucasian and 1 SAN) representing 38 alleles were included. In this phase of the study 24 different alleles were sequenced. Only two specimens contained off-ladder alleles i.e. 31.2 and 40.2. These alleles were designated as 32 and 40 respectively by routine STR analysis. This meant a difference of 2bp from the routine STR allele designation. Of the remaining 36 within ladder alleles 11.11%(4/36) were wrongly assigned. Two of these differ only by 1 bp, one by two bps and the other one by more than four base pairs.

In both studies a total of 27 different FGA alleles were sequenced. Three alleles 26(263bp), 28(271bp) and 43.2(333bp) had 2 sequence variants each. Three new previously undescribed alleles were also found in this study i.e. alleles 29.2, 40.2 and 41.2. The former was reported but the sequence structure was not described. In addition, one of the 43.2 alleles and the only 44.2 allele observed in this study were new sequence variants.

Based on the fragment size, two distinct groups of FGA alleles were observed; those in the allele size range from 16.1 to 31.2 and those in the size range 40.2 to 44.2. No alleles were found in the alleles size range between 31.2 and 40.2. According to sequence similarities the FGA alleles can be divided into three groups. The shorter alleles (16.1 to 26) have a [CTCC (TTCC)$_2$] motif in common at the 3' end. Exceptions were the 16.1 allele that contained a T insertion interrupting the long CTTT repeat sequence. The intermediate alleles between 26 and 31.2 contain an additional (CTTT)$_{3\text{-}5}$ repeat motif before the CTCC (TTCC)$_2$ motif at the 3' end and the 29.2, 30.2 and 31.2 alleles contain the [CTCC (TTCC)$_4$] repeat motif at the 3' end. The larger alleles (40.2 - 44.2) contain a (CTGT)$_{3\text{-}6}$ repeat motif interrupting the long CTTT repeat sequence. An additional insertion of (CTTC)$_{3\text{-}4}$ (CTTT)$_3$ repeat motif is found before the CTCC (TTCC)$_4$ motif at the 3' end. In addition all the sequenced alleles had a TTTC TTCC TTTC TTTTTT motif at the 3' end that was not included in the allele designation. No mutations were observed both at the flanking and primer binding regions of this short tandem repeat locus. The sequence formulas of all the alleles are given in Table.1.

Table 1 Description of sequenced FGA alleles

Alle	***N***	**Repeat motif**										**bp**
16.1	11	(TTTC)$_2$	TTTT	TTCT	(CTTT)$_5$	T	(CTTT)$_3$			CTCC	(TTCC)$_2$	224
18	1	(TTTC)$_2$	TTTT	TTCT	(CTTT)$_{10}$					CTCC	(TTCC)$_2$	231
18.2	1	(TTTC)$_2$	TTTT	TT	(CTTT)$_{11}$					CTCC	(TTCC)$_2$	233
19	4	(TTTC)$_2$	TTTT	TTCT	(CTTT)$_{11}$					CTCC	(TTCC)$_2$	235
19.2	1	(TTTC)$_2$	TTTT	TT	(CTTT)$_{12}$					CTCC	(TTCC)$_2$	237
20	2	(TTTC)$_2$	TTTT	TTCT	(CTTT)$_{12}$					CTCC	(TTCC)$_2$	239
21	2	(TTTC)$_2$	TTTT	TTCT	(CTTT)$_{13}$					CTCC	(TTCC)$_2$	243
21.2	1	(TTTC)$_2$	TTTT	TT	(CTTT)$_{14}$					CTCC	(TTCC)$_2$	245
22	4	(TTTC)$_2$	TTTT	TTCT	(CTTT)$_{14}$					CTCC	(TTCC)$_2$	247
22.2	2	(TTTC)$_2$	TTTT	TT	(CTTT)$_{15}$					CTCC	(TTCC)$_2$	249
23	2	(TTTC)$_2$	TTTT	TTCT	(CTTT)$_{15}$					CTCC	(TTCC)$_2$	251
23.2	1	(TTTC)$_2$	TTTT	TT	(CTTT)$_{16}$					CTCC	(TTCC)$_2$	253
24	1	(TTTC)$_2$	TTTT	TTCT	(CTTT)$_{16}$					CTCC	(TTCC)$_2$	255
24.2	1	(TTTC)$_2$	TTTT	TT	(CTTT)$_{17}$					CTCC	(TTCC)$_2$	257
25	2	(TTTC)$_2$	TTTT	TTCT	(CTTT)$_{17}$					CTCC	(TTCC)$_2$	259
26	3	(TTTC)$_2$	TTTT	TTCT	(CTTT)$_{18}$					CTCC	(TTCC)$_2$	263
26'	1	(TTTC)$_2$	TTTT	TTCT	(CTTT)$_{12}$	CCTT	(CTTT)$_5$			CTCC	(TTCC)$_2$	263
27	1	(TTTC)$_2$	TTTT	TTCT	(CTTT)$_{13}$	CCTT	(CTTT)$_5$			CTCC	(TTCC)$_2$	267
28	1	(TTTC)$_2$	TTTT	TTCT	(CTTT)$_{20}$					CTCC	(TTCC)$_2$	271
28'	1	(TTTC)$_2$	TTTT	TTCT	(CTTT)$_{14}$	CCTT	(CTTT)$_5$			CTCC	(TTCC)$_2$	271
29	1	(TTTC)$_2$	TTTT	TTCT	(CTTT)$_{15}$	CCTT	(CTTT)$_5$			CTCC	(TTCC)$_2$	275
29.2	3	(TTTC)$_4$	TTTT	TT	(CTTT)$_{13}$	(CTTC)$_3$	(CTTT)$_3$			CTCC	(TTCC)$_4$	277
30	1	(TTTC)$_3$	TTTT	TTCT	(CTTT)$_{16}$	CCTT	(CTTT)$_5$			CTCC	(TTCC)$_2$	279
30.2	1	(TTTC)$_4$	TTTT	TT	(CTTT)$_{14}$	(CTCC)$_3$	(CTTT)$_3$			CTCC	(TTCC)$_4$	281
31.2	2	(TTTC)$_4$	TTTT	TT	(CTTT)$_{15}$	(CTCC)$_3$	(CTTT)$_3$			CTCC	(TTCC)$_4$	285
40.2	6	(TTTC)$_4$	TTTT	TT	(CTTT)$_9$	(CTGT)$_3$	(CTTT)$_{12}$	(CTTC)$_3$	(CTTT)$_3$	CTCC	(TTCC)$_4$	321
41.2	1	(TTTC)$_4$	TTTT	TT	(CTTT)$_9$	(CTGT)$_3$	(CTTT)$_{13}$	(CTTC)$_3$	(CTTT)$_3$	CTCC	(TTCC)$_4$	325
43.2	16	(TTTC)$_4$	TTTT	TT	(CTTT)$_8$	(CTGT)$_5$	(CTTT)$_{13}$	(CTTC)$_4$	(CTTT)$_3$	CTCC	(TTCC)$_4$	333
43.2'	1	(TTTC)$_4$	TTTT	TT	(CTTT)$_{12}$	(CTGT)$_6$	(CTTT)$_9$	(CTTC)$_3$	(CTTT)$_3$	CTCC	(TTCC)$_4$	333
44.2	1	(TTTC)$_4$	TTTT	TT	(CTTT)$_9$	(CTGT)$_5$	(CTTT)$_{13}$	(CTTC)$_4$	(CTTT)$_3$	CTCC	(TTCC)$_4$	337

n = number of samples sequenced

Discussion

Sequencing of STR loci yields an abundance of information about the specific locus being sequenced and about tetranucleotide repeats in general [5], also to study off-ladder alleles and to establish consistent nomenclature for new alleles [3].

Sizing

Using the ABI Prism genetic analyzer and POP-4 polymer, Buscemi *et al.* (1998) reported that PCR fragment sizes obtained by comparison with a commercial internal size standard provides very good accuracy and precision [16]. Lazaruk *et al* (1998) also reported a similar observation. The precision in sizing in their study had a standard deviation of < 0.15 bp, where 99.7% of alleles that were of equal size fall within ± 0.45 bp of the allelic ladder allele size [17]. In agreement with these reports, our study found accurate sizing and allelic assignment within ladder ranges. As confirmed by sequence analysis only 4 samples out of 49 were miss assigned. The differences in size were 1 or 2 bp. However all alleles (27/27) that were larger than the longest allelic ladder allele (>30 bp) were miss assigned. The differences in size were 2 or more bp, emphasizing the need for allelic ladders to include all possible size ranges of alleles at specific loci.

Sequence variants

Dauber *et al* (2000) reported 17 different alleles at the FGA locus and Barber *et al* reported 22 alleles ranging in size from 168 to 249 bp [13]. Lazaruk *et al* 2001 reported 36 alleles and 4 sequence variants at this locus. Griffiths *et al* 1998 also reported 28 alleles at this locus that could be used to improve the allelic ladder. They reported that the sequence of allele 27 they used was different from the one that was reported by Barber *et al* [23]. Additionally, a STR fact sheet documented 42 alleles and 1 sequence variant. In our study 27 alleles including 3 sequence variants and 3 previously sequence-undescribed alleles were found. Combining the later three reports, there were 46 sequence described alleles at the FGA locus. Among these 16 were complete tertanucleotide repeats, the remaining 30 were interalleles that vary by 2 bp from the common sequence including the exceptional alleles 16.1 and 22.3 that differ in one bp. The observed increased polymorphism obviously increases the discriminatory capability of this locus. High sequence variation was observed both at the complete tetranucleotide and interalleles. Alleles 27, 43.2 and 44.2 have three sequence variants each, while alleles 18,24, 28, 32, 42.2, 46.2, 47.2 and 50.2 have two sequence variants each according to different studies on different population groups. Two to six allelic sequence variants were also reported for STR loci D3S1358, D5S230, D8S639, D8S1132, D11S488, D12S391, D14S229, D21S11, vWF2 and vWA [3,8,12,14,16,19,20,21,22]. The presence of these increased allelic sequence variations in a population will possibly pose a problem both in forensic and paternity result interpretation. The problem might be magnified when there is an exclusion at only one locus. Also if sequence variants of common alleles (such as alleles between 19 and 25) are found at higher frequencies in a population. Alleles between 19 to 25 of the FGA locus are found at the frequency of > 0.1000 in the majority of the world population groups.

Ethnic variation

It has been suggested that for any profile, differences in allele proportion between ethnic groups could be analyzed to form an ethnic inference system. Therefore, information regarding the probable ethnicity of an unknown offender may direct investigation and the setting of priorities [24]. Dauber *et al* (2001) reported that the larger alleles in their study were exclusively found in the Afro-Caribbean population [13]. Barber *et al* (1995) also showed that some alleles were exclusive to some ethnic groups [4]. In our study, larger alleles (> 30) were found in the Negroid, Mixed ancestry and Caucasian specimens, whereas all (n = 11) of the 16.1 alleles were observed in the Negroid population samples indicating a possible ethnic restriction of this allele in this local population. However, this allele was reported and sequenced from other population groups with a similar sequence structure. In conclusion: 1) Knowledge of the sequence structure of alleles in specific STR loci enables accurate allelic designation. In addition it also helps to identify new alleles that could be incorporated in allelic ladders that would ease allelic assignment. 2) Sequence studies also disclose variants. Since sequence variants are also described in some of the commonly used STR loci, there is a need to conduct population-based studies. The outcome of such studies would aid in accurate analysis and interpretation of paternity and forensic data. It could also aid in setting up of certain criteria as to when sequencing would be used to supplement difficult routine DNA marker profile in paternity and forensics. 3) Identification and use of ethnically restricted alleles and sequence variants will help to focus forensic investigation.

References

[1] P.M. Schneider, Basic issues in forensic DNA typing, Forensic Sci. Int, 88 (1997) 17-22.

[2] E. Momhinweg, C. Luckenbach, R. Fimmers, H. Ritter, D3S1358: Sequence analysis and gene frequency in a German population, Forensic Sci. Int. 95 (1998) 173-178.

[3] K. Lazaruk, J. Wallin, C. Holt, T. Nguyen, P Sean, Walsh, Sequence variation in humans and other primates at six short tandem repeat loci used in forensic identity testing, Forensic Sci. Int. 119 (2001) 1-10.

[4] M. D. Barber, B. J. McKeown, B. H. Parkin, Structural variation in the alleles of a short tandem repeat system at the human alpha fibrinogen locus, Int. J. Legal Med. 108 (1996) 180-185.

[5] http://www.promega.com/pnotes/58/5189c.html New approaches to DNA fingerprint analysis

[6] M. Klintschar, B. Glock, E. M. Dauber, W. R. Mayr, Genetic variation and sequence studies of a highly variable short tandem repeat at the D17S976 locus, Int. J. Legal Med. 112 (1998) 50-54.

[7] F. Neuhuber, M. Klintschar, M. Radacher, A collaborative geneic study on the STR system FGA in two Austrian population samples, Forensic Sci. Int. 91 (1998) 1-6.

[8] T. Yoshimoto, K. Tamaki, S. Katsumata. X. L. Huang, R. Uchihi, M. Tanaka, H. Uchida, T. Yamamoto, S. Chen, J. A. L. Armour, Y. Katsumata, Sequence analysis of alleles at a microsatellite locus D14S299 (wg1c5) and population genetic comparison, Int. J. Legal Med. 113 (1999) 15-18.

[9] A. Urquhart, C. P. Kimpton, T. J. Downes, P. Gill, Variation in short tandem repeat sequences - a survey of twelve microsatellite loci for use as forensic identification markers, Int. J. Legal Med. 107 (1994) 13-20.

[10] P. Gill, B. Brinkmann, E. d'Aloja, J. Andersen, W. Bar, A. Carracedo, B. Dupuy, B. Eriksen, M. Jangblad, V. Johnsson, A. D. Kloosterman, P. Lincon, N. Morling, S. Rand, M. Sabatier, R. Scheithauer, P. Schnider, M. C. Vide, Consideration from the European DNA profiling group (EDNAP) concerning STR nomenclature, Forensic Sci. Int. 87 (1997) 185-192.

[11] S. Watson, A. Allsop, L. Foreman, Z. Kelsey, P. Gill, Sequenced allelic ladders and population genetics of a new STR multiplex system, Forensic Sci. Int. 115 (2001) 207-217.

[12] L. Henke, R. Fimmers, J. Reinhold, M. Dulmer, S. Cleef, J. Arnold, J. Henke, Sequence analysis and population data on the 'new' short tandem repeat locus D5S2360, Forensic Sci. Int. 116 (2001) 55-58.

[13] E. M. Dauber, B. Glock, D. W. M. Schwartz, W. R. Mayer, Further sequence and length variation at the STR loci HumFES/FPS, HumVWA, HumFGA and D12S391, Int. J. Legal Med. 113 (2000) 76-80.

[14] C. Sidel, S. Muller, O. Jager, E. Seifried Sequence analysis and population data of short tandem repeat polymorphism at loci D8S639 and D11S488, Int. J. Legal Med. 112 (1999) 355-359.

[15] K. A. Mills, D. A. Even, J. C. Murray, Tetranucleotide repeat polymorphismat the human alpha fibrinogen locus (FGA), Human Mole Genet 1 (1992) 779

[16] L. Buscemi, A. Tagliabracci, C. Sassaroli, F. Bianchi, S. Canestrari, D. Rodriguez, Polymerase chain reaction typing of D21S11 short tandem repeat polymorphism by capillary electrophoresis. Allele frequencies and sequencing data in a population sample from central Italy, Forensic Sci. Int. 92 (1998) 251-258.

[17] K. Lazaruk, P.S. Walsh, F. Oaks, D. Gilbert, B.B. Rosenblum, S. Menchen, D. Scheibler, H. M. Wenz, C. Holt. J. Wallin, Genotyping of forensic short tandem repeat (STR) systems based on sizing precision in a capillary electrophoresis instrument, Electrophoresis 19 (1998) 86-93

[18] http://www.cstl.nist.gov/biotech/strbase/str_fga.htm

[19] S. Hering, E. Muller, New alleles and mutational events in D12S391 and D8S1132 sequence data from an eastern German population, Forensic Sci. Int.124 (2001) 187-191.

[20] A. P. Haddad, R. L. Sparrow The short tandem repeat locus VWF2 in Intron 40 of the von Willebrand factor gene consists of two polymorphic sub-loci Forensic Science International 119 (2001) 299-304.

[21] F. X. Xiao, A. Gilissen, J-J. Cassiman, R. Decorte, Quadruplex fluorescent STR typing system (HUMVWA, HUMTHO1 and HPRT) with sequence-defined allelic ladders Identification of a new allele at D21S11 Forensic Science International 94 (1998) 39-46.

[22] H-G. Zhou, K. Sato, Y. Nishimaki, L. Gang, H. Hasekura. The HumD21S11system short tandem repeat DNA polymorphisms in Japanese and Chinese Forensic Science International 86 (1997) 109-118.

[23] R.A.L. Griffiths, M.D. Barber, P.E. Johnson, S.M. Gillbard, M.D. Haywood, C.D. Smith, J Arnold, T.Burke, A.J. Urquhart, P. Gill, New reference allelic ladders to improve allelic designation in a multiplex STR system, Int. J. Legal Med, 111 (1998) 267 –272.

[24] A.L. Lowe, A. Urquhart. L. A. Foreman, I.W. Evett, Inferring ethnic origin by means of an STR profile, Forensic Sci. Int. 119 (2001) 17 –22.