

The evaluation of Y-STR loci for use in Forensics

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

UNIVERSITY of the
WESTERN CAPE

Lieze Suzette Ehrenreich

March 2005

*Supervisors: Dr. Neil Leat (PhD)
Prof. Sean Davison (PhD)*

Keywords

Forensics

Short Tandem Repeat (STR)

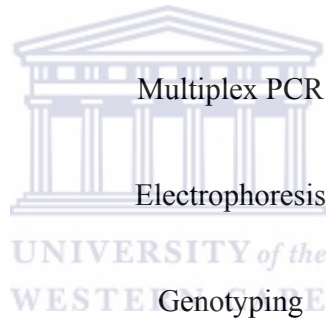
Loci

Minimal Haplotype (MH)

Database

Bio-Informatics

Polymerase Chain Reaction (PCR)



Stutter

Gene Diversity

Polymorphism

Abstract

The evaluation of Y-STR loci for use in Forensics

L S Ehrenreich

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

29/03/2005

The aim of this study was to investigate the forensic usefulness of various Y-STR loci among South African sub-populations. Three different sets of Y-STR loci were chosen for investigation. The seven loci that constitute the European 'minimal haplotype' (MH) were selected based on their established use in forensic studies. These would serve as a point of reference, against which the performance of other loci could be compared. An attempt was made to also choose loci that have been reported to consistently show high levels of variability among different populations. Generally, loci with a reported gene diversity (D) value ≥ 0.6 among various populations were selected. 14 more STRs were chosen for further investigation from the vast amount of Y-chromosomal sequence data and recently identified loci. Three factors were considered in the forensic suitability of all the non-MH loci: (1) the ease with which male specific primers could be generated for PCR amplification, (2) the levels of locus variability among three South African sub-populations and (3) the extent of stutter generated by PCR amplification. MH loci were amplified in a multiplex reaction described elsewhere (Leat et al. 2004). Three multiplex PCRs were developed as an efficient means of screening the non-MH loci. DYS385 and DYS389 of the MH were omitted from the analyses as only single-copy loci were considered. Samples were typed from 101 English-speaking Caucasian-, 88 Black Xhosa-speaking- and 77 Asian Indian males. Gene diversity values, the number of alleles identified and the average stutter was determined for each locus. The D values for the single copy loci of the MH ranged between 0.322 (DYS393) and 0.768 (DYS390), and mean stutter ranged between 4.92% (DYS19) and 11.31% (DYS392) of the primary allele. The D values of non-MH loci previously reported in the literature ranged between 0.53 (Y-GATA C4) and 0.86 (DYS449), and mean stutter ranged between 4.45% (DYS463) and 13.31% (DYS449) of the primary allele. The D values of the recently discovered loci ranged between 0.57 (DYS607) and 0.92 (DYS710), and average stutter ranged between 3.11% (DYS714) and 47.52% (DYS711) of the primary allele. The data presented here strongly suggests that some of the MH loci are not suitable for forensic purposes among South African sub-populations. The data presented here also indicates that several forensically useful loci that may compliment the MH, or replace those MH loci with low variability, are now available.

Declaration

I declare that ‘The evaluation of Y-STR loci for use in Forensics’ is my own work, that it has not been submitted for any degree or examination in any other university and that all the sources I have used or quoted have been indicated and acknowledged by complete references.

Full Name: Liezle Suzette Ehrenreich

Signed:

Date: 29/03/2005



Acknowledgements

I would like to thank my research supervisors for their guidance throughout this project. I would also like to thank them for affording me the opportunity to attend the 4th Y-User Group's Workshop in Berlin during November of 2004.

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

Many thanks also go to all the men who so generously donated biological material for the study.

My gratitude also goes to all the students and staff who made working in the Department of Biotechnology such a pleasant experience.

Thanks also go to family and friends who have supported me during this endeavour.



List of Abbreviations

DNA	Deoxyribonucleic Acid
SNP	Single Nucleotide Polymorphism
STR	Short Tandem Repeat
Y-STR	Y-chromosome STR
PAR	Pseudo-Autosomal Region
NR1	Non-recombining Y-chromosome
PBS	Primer Binding Site
MH	Minimal Haplotype
ISFG	International Society of Forensic Genetics
YHRD	Y-Chromosome Haplotype Reference Database
SWGAM	Scientific Working Group on DNA Analysis Methods
VOC	Vereenigde Oost-Indische Compagnie
YCC	Y-chromosome Consortium
PCR	Polymerase Chain Reaction
HCl	Hydrochloric Acid
KCl	Potassium Chloride
MgCl ₂	Magnesium Chloride
dNTPs	deoxy Nucleotide TriPhosphates
dATP	deoxy Adenosine TriPhosphate
BSA	Bovine Serum Albumin
DMSO	DimMethylSulfOxide
EDTA	Ethylene Diamine Tetra Acetic Di-Sodium Salt
TEMED	N, N, N', N' Tetramethyl-EthyleneDiamine
NaCl	Sodium Chloride
RFU	Relative Fluorescent Units

List of tables

Table 1.1

Comparison of three commercially available Y-STR typing kits

Table 1.2

Expected DNA content of biological samples (Lee and Ladd 2001)

Table 1.3

Racial distribution of South African populations

Table 2.1

Sequences, fluorescent dye labels and final concentration of primers used in ‘minimal haplotype’ (MH) multiplex reaction

Table 2.2

Allele and haplotype frequencies for MH loci among Asian Indian males ($n=85$) from KwaZulu-Natal, South Africa

Table 2.3

Allele and haplotype frequencies for MH loci among Afrikaner Caucasian males ($n=108$) from the Western Cape, South Africa

Table 2.4

Allele and haplotype frequencies for MH loci among males from the Coloured community ($n=107$) of the Western Cape, South Africa

Table 2.5

Comparison of gene diversity (D) values for MH loci among five South African sub-populations

Table 2.6

Haplotypes shared by more than one Asian Indian male ($n=85$)

Table 2.7

Haplotypes shared by more than one Afrikaner Caucasian male ($n=108$)



Table 2.8

Haplotypes shared by more than one Coloured male ($n=107$)

Table 2.9

Haplotypes shared by more than one English Caucasian male ($n=100$)

Table 2.10

Haplotypes shared by more than one Xhosa male ($n=99$)

Table 3.1

Primer sequences for recently identified loci as designed with the use of *Primer3* software and estimated fragment length size as observed with sequence data from contigs NT_011875.7, NT_011878.7, NT_011896.7, NT_011903.3

Table 3.2

Repeat structure nomenclature of loci

Table 3.3

Gene diversity (D) value for 20 male specific single copy loci among 46 English Caucasian males in the Western Cape, South Africa

Table 3.4

Pentanucleotide loci ranked according to longest stretch of repeats as observed from sequence data from contigs NT_011875.7, NT_011878.7, NT_011896.7, NT_011903.3

Table 3.5

Tetranucleotide loci ranked according to longest stretch of repeats as observed from sequence data from contigs NT_011875.7, NT_011878.7, NT_011896.7, NT_011903.3

Table 3.6

Trinucleotide loci ranked according to longest stretch of repeats as observed from sequence data from contigs NT_011875.7, NT_011878.7, NT_011896.7, NT_011903.3

Table 4.1

Amplification components of multiplex reactions UWC Y-Plex 1, UWC Y-Plex 2 and UWC Y-Plex 3

Table 4.2

PCR cycling conditions of multiplex reactions UWC Y-Plex 1, UWC Y-Plex 2 and UWC Y-Plex 3

Table 4.3

Sequences, fluorescent dye labels and final concentration of primers used in multiplex reactions UWC Y-Plex 1, UWC Y-Plex 2 and UWC Y-Plex 3

Table 5.1

Gene diversity values for 27 single-copy Y-STR loci among three South African sub-populations

Table 5.2

Haplotype diversity and the effect that the addition of single loci has on the accumulated haplotype diversity

Table 5.3

Stutter characteristics and sequence structure for loci used in this study. Stutter is expressed as the percentage of the stutter peak height relative to that of the primary peak.

List of Figures**Figure 1.1**

Incidence of violent crime in South Africa (April 2003 – March 2004)

Figure 1.2

(a) Number of reported rapes in South African provinces (April 2003 – March 2004) and
(b) Per capita incidence of rape in South African provinces (April 2003 – March 2004)

Figure 1.3

Y-chromosome structure indicating pseudoautosomal regions 1 and 2, and the non-recombining region of the Y-chromosome (Quintana-Murci and Fellous 2001)

Figure 1.4

A generalized protocol with guidelines on how to approach the optimisation of a multiplex PCR (Henegariu et al. 1997)

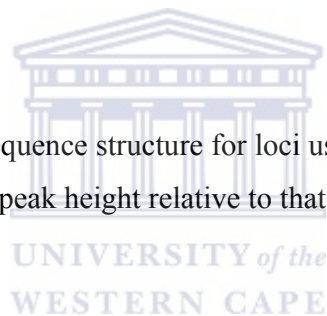


Figure 1.5

Primer design-based approach to multiplex PCR optimisation (Butler et al. 2002)

Figure 1.6

Predicted PCR product size range for loci amplified in 20-plex (Butler et al. 2002)

Figure 1.7

Predicted PCR product size range for loci amplified in 10-plex (Schoske et al. 2003)

Figure 1.8

Predicted PCR product size range for loci amplified in 20-plex (Schoske et al. 2004)

Figure 1.9

Predicted PCR product size range for loci amplified in 20-plex (Hall and Ballantyne et al. 2004)

Figure 1.10

Predicted allele size range for loci amplified in (a) Y-PLEX™ 12 kit from *Reliagene*, (b) MenType Argus Y-MH from *Biotype* and (c) PowerPlex Y from *Promega*

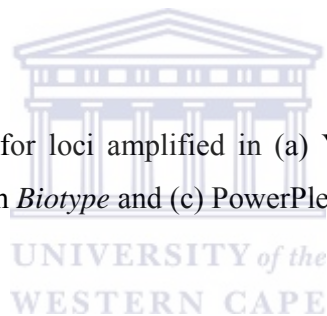


Figure 1.11

South African population distribution by home language

Figure 1.12

Neighbour joining tree showing genetic affinities between South African Bantu speakers using Y-STR haplotype data only (Lane et al. 2002)

Figure 2.1

An example of an electropherogram for the ‘minimal haplotype’ (MH) multiplex reaction

Figure 2.2

Worldwide distribution of the most common ‘MH’ among Asian Indian males from KwaZulu-Natal, South Africa (Red and orange dots on the maps indicate haplotype matches. Blue dots indicate populations for which MH data is available, but in which no matches could be found.)

Figure 2.3

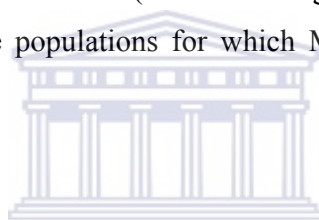
Worldwide distribution of the 2nd most common ‘MH’ among Asian Indian males from KwaZulu-Natal, South Africa (Red and orange dots on the maps indicate haplotype matches. Blue dots indicate populations for which MH data is available, but in which no matches could be found.)

Figure 2.4

Worldwide distribution of the most common ‘MH’ among Afrikaner Caucasian males from the Western Cape, South Africa (Red and orange dots on the maps indicate haplotype matches. Blue dots indicate populations for which MH data is available, but in which no matches could be found.)

Figure 2.5

Worldwide distribution of the 2nd most common ‘MH’ among Afrikaner Caucasian males from the Western Cape, South Africa (Red and orange dots on the maps indicate haplotype matches. Blue dots indicate populations for which MH data is available, but in which no matches could be found.)

**Figure 2.6**

Worldwide distribution of the most common ‘MH’ among Coloured males from the Western Cape, South Africa (Red and orange dots on the maps indicate haplotype matches. Blue dots indicate populations for which MH data is available, but in which no matches could be found.)

Figure 2.7

Worldwide distribution of the 2nd most common ‘MH’ among Coloured males from the Western Cape, South Africa (Red and orange dots on the maps indicate haplotype matches. Blue dots indicate populations for which MH data is available, but in which no matches could be found.)

Figure 2.8

Worldwide distribution of the most common ‘MH’ among English Caucasian males from the Western Cape, South Africa (Red and orange dots on the maps indicate haplotype matches. Blue dots indicate populations for which MH data is available, but in which no matches could be found.)

Figure 2.9

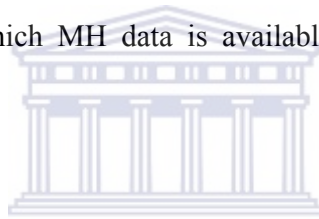
Worldwide distribution of the 2nd most common 'MH' among English Caucasian males from the Western Cape, South Africa (Red and orange dots on the maps indicate haplotype matches. Blue dots indicate populations for which MH data is available, but in which no matches could be found.)

Figure 2.10

Worldwide distribution of the most common 'MH' among Xhosa males from the Western Cape, South Africa (Red and orange dots on the maps indicate haplotype matches. Blue dots indicate populations for which MH data is available, but in which no matches could be found.)

Figure 2.11

Worldwide distribution of the 2nd most common 'MH' among Xhosa males from the Western Cape, South Africa (Red and orange dots on the maps indicate haplotype matches. Blue dots indicate populations for which MH data is available, but in which no matches could be found.)

**Figure 3.1**

20 recently identified single-copy loci ranked according to gene diversity (*D*) values among 46 English Caucasian males from the Western Cape, South Africa

Figure 4.1

Primer design-based approach to multiplex optimisation, adapted from Schoske et al. 2003

Figure 4.2

- (a) Observed allele size ranges of six loci amplified with multiplex UWC Y-Plex 1
- (b) An example of a typical electropherogram for the multiplex UWC Y-Plex 1

Figure 4.3

- (a) Observed allele size ranges of six loci amplified with multiplex UWC Y-Plex 2
- (b) An example of a typical electropherogram for the multiplex UWC Y-Plex 2

Figure 4.4

- (a) Observed allele size ranges of six loci amplified with multiplex UWC Y-Plex 3
- (b) An example of a typical electropherogram for the multiplex UWC Y-Plex 3

Figure 5.1

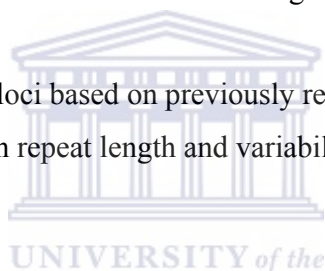
Gene diversity, number of alleles identified and frequency of the most common allele. Loci have been ranked from highest to lowest mean gene diversity



Table of Contents

Chapter 1 – Literature Review	1
1.1 Statistics for sexual assault in South Africa	1
1.2 The use of Y-STRs in the identity testing of males	3
1.3 Discovery of Y-STR loci	5
1.4 Multiplex PCR amplification of Y-STR loci	6
1.4.1 Approaches to multiplex PCR development	7
1.4.2 Validation of multiplex PCR	11
1.4.3 Non-commercial Y-STR multiplex PCR	11
1.4.4 Commercial kits for the multiplex amplification of Y-STR loci	14
1.5 Biological evidence and the use of Y-STRs in forensic casework	16
1.5.1 Biological evidence	16
1.5.2 The use of Y-STRs in forensic studies	16
1.6 Forensic casework and the YHRD	18
1.7 Populations Studies	20
1.8 Objectives of this study	22
Chapter 2 – Minimal Haplotype Analysis	22
2.1 Introduction	22
2.2 Methods and Materials	22
2.2.1 Obtaining biological samples for the study	22
2.2.2 DNA Extraction	23
2.2.3 PCR Amplification of MH loci	23
2.2.4 Fragment Analysis	24
2.2.5 Data Analysis	24
2.3 Results and Discussion	25
2.3.1 MH multiplex performance	25
2.3.2 Analysis of allele and haplotype frequencies	25
2.3.3 Comparison of allele frequencies among five South African sub-populations	30
2.3.4 Haplotype analysis and comparisons of common haplotypes with those in the YHRD	30
2.3.5 Summary	37

Chapter 3 – Finding new Y-STR loci	45
3.1 Introduction	45
3.2 Methods and Materials	46
3.2.1 Selection of STR loci from Y-chromosome sequence data	46
3.2.2 Primer design	47
3.2.3 Male specificity of primer sets	49
3.2.4 Variability of Y-STR loci and allele size ranges in a sample 46 English Caucasian males	49
3.2.5 Statistical analysis	50
3.2.6 Selection of Y-STR loci based on previously reported population data	50
3.3 Results and Discussion	50
3.3.1 Selection of STR loci from Y-chromosome sequence data	50
3.3.2 Male specificity of primer sets	52
3.3.3 Variability of Y-STR loci and allele size ranges in a sample 46 English Caucasian males	52
3.3.4 Selection of Y-STR loci based on previously reported population data	54
3.3.5 Relationship between repeat length and variability	54
3.3.6 Summary	56
Chapter 4 – Multiplex Amplification of Y-STR loci	57
4.1 Introduction	57
4.2 Methods and Materials	58
4.2.1 Loci selected for further investigation	58
4.2.2 Primer design-based approach to multiplex design	58
4.2.3 Multiplex amplification of Y-STR loci	59
4.2.4 Allelic ladders	62
4.2.5 Fragment Analysis	62
4.3 Results and Discussion	63
4.3.1 Multiplex reactions designed	63
4.3.2 Performance of multiplex UWC Y-Plex 1	65
4.3.3 Performance of multiplex UWC Y-Plex 2	66
4.3.4 Performance of multiplex UWC Y-Plex 3	67
4.3.5 Summary	68



Chapter 5 – Investigating properties of loci among three South African sub-populations

5.1	Introduction	69
5.2	Methods and Materials	70
5.3	Results and Discussion	71
5.3.1	Variability of loci among three South African sub-populations	71
5.3.2	Duplicated loci and intermediate alleles	74
5.3.3	Effect of single loci on haplotype diversity	74
5.3.4	Stutter analysis	75
5.3.5	Summary	78
	Overview and future prospects	79
	References	81
	Electronic Supplementary Resources	88
	Appendix	89



Chapter 1 – Literature Review

Sexual assault is a significant problem facing South African society. In this context, an efficient system is needed for the positive identification of criminals in incidences of sexual violence. Genetic identity testing is achieved by examining polymorphic regions of DNA. Typically sets of polymorphisms are examined together to provide a genetic profile. The polymorphic markers most commonly used are on the autosomal (1-22) chromosomes. While they have an excellent capacity to distinguish between individuals, they do also have disadvantages. In sexual assault cases it is often difficult to separate the female victim's profile from the rapist's profile. Analysis of Y-chromosome markers overcomes this by generating male specific profiles. Following will be a review on issues surrounding efficient identity testing systems. Topics covered will include (1) statistics of sexual assault in South Africa, (2) the use of Y-STR loci in identity testing of males, (3) identification of Y-STR loci, (4) multiplex PCR of Y-STR loci, (5) biological evidence and the use of Y-STRs in forensic casework, (6) forensic casework and the YHRD and (7) population aspects.

1.1 Statistics for sexual assault in South Africa

South African society is faced with high levels of violent crime. Between April of 2003 and March of 2004, a total of 652 959 acts of violence were reported to the South African Police Service (<http://www.saps.org.za>). These reports typically involved: murder, attempted murder, physical assault (I), assault with the intent to do grievous bodily harm (II), indecent assault (III), and rape. Of these reports, 52 733 were rape, making this the third most common violent crime in South Africa (Figure 1.1). The highest total reported incidence of rape occurred in Gauteng (Figure 1.2a), while the highest per capita incidence of rape occurred in the Northern Cape (Figure 1.2b). While the figures indicate a high incidence of rape, some organizations suggest that the incidence of this crime may be substantially higher.

Rape Crisis (<http://www.rapecrisis.org.za>), a non-governmental organization, suggests that police statistics underestimate sexual assault for several reasons including: (1) the fact that the definition of rape in South African law does not include rape of men and children, oral rape or rape with objects, (2) rapes within long-standing relationships are generally not viewed as rape, and are therefore less likely to be reported, and (3) because of the lack of effective policing, rapes

in particularly rural areas may be significantly under-reported. Despite this, there is agreement that the incidence of sexual violence in South Africa is unacceptably high.

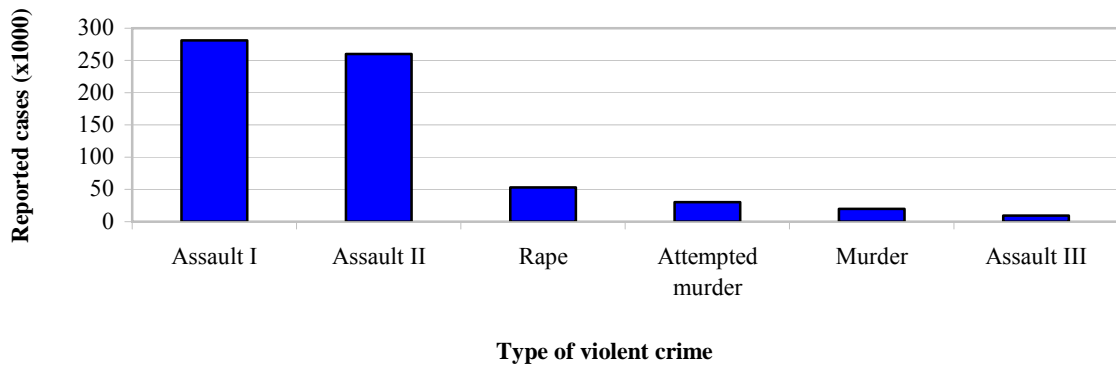


Figure 1.1. Incidence of violent crime in South Africa
<http://www.saps.org.za> (April 2003 – March 2004)

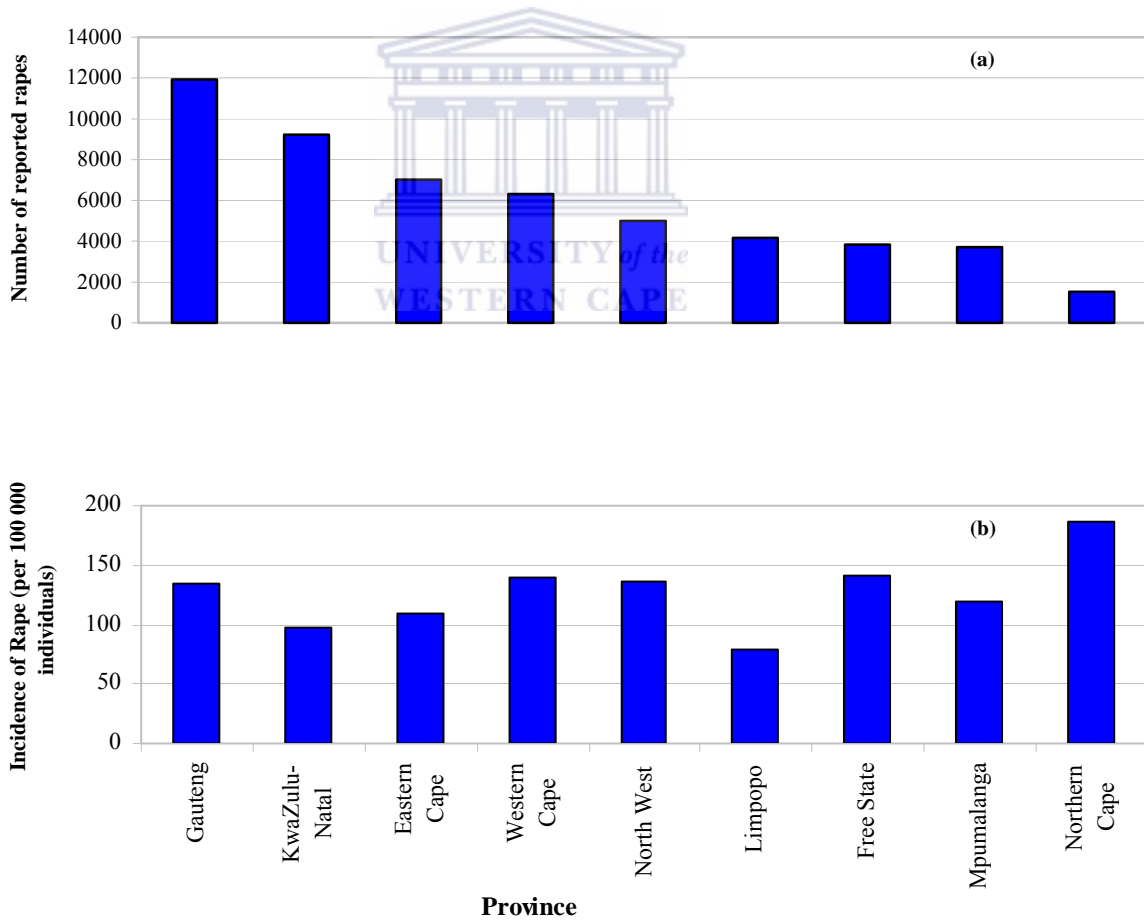


Figure 1.2. (a) Number of reported rapes in South African provinces and (b) per capita incidence of rape in South African provinces (<http://www.saps.org.za>) (April 2003 – March 2004)

1.2. The use of Y-STR loci in identity testing of males

1.2.1 STR loci as identity testing tools

Several types of genetic elements have the potential for use in human identity testing. These include single nucleotide polymorphisms (SNPs), mini-satellites and micro-satellites. SNPs are single nucleotide substitutions and are therefore bi-allelic. SNPs mutate relatively slowly and therefore tend to not be very polymorphic. Their lack of polymorphism and hence discriminatory capacity, make them less than ideal for routine human identity testing. Mini-satellites are elements that contain tandem repeats that are typically between 10 and 15 bp long. Many minisatellites are highly polymorphic due to variation in copy number of the repeat in the minisatellite (Jeffreys et al. 1985). Unfortunately, the size of the repeats and the general size of mini-satellites, hinder the ease with which these elements can be applied to human identity testing (Moxon and Wills 1999).

Micro-satellites or short tandem repeat (STR) loci are stretches of DNA containing motifs of 1-6 bp that are repeated in tandem. Approximately 3% of the human genome is occupied by STR sequences and they are found on all autosomal, as well as the sex chromosomes. The number of repeated motifs in a STR can vary between individuals, making them useful polymorphic markers in human identity testing (Moxon and Wills 1999). The use of STR loci in human identity testing has advantages over the use of other genetic elements. Because of their high mutation rate, they are generally highly polymorphic and therefore provide a good discriminatory capacity (Jobling 2001). The repeated elements and the STR loci themselves are of manageable size and can therefore be analyzed easily using techniques such as polymerase chain reaction (PCR) and electrophoresis (Roewer et al. 1992). They are therefore the most common marker of choice in human identity testing.

1.2.2 Y-STR loci as identity testing tools in males

STR loci that are found on the Y-chromosome are referred to as Y-STR loci. To understand why Y-STR loci are so ideal in the identity testing of males, one must briefly examine the Y-

chromosome (Figure 1.3). The Y-chromosome is one of the smallest chromosomes in the human genome with an average size of ~60Mb (Buhler, 1980). At the tip of both arms are pseudoautosomal regions (PARs). These PARs represent ~5% of the Y-chromosome and contain sequences that are homologous to sequences on the X chromosome. During male meiosis, these PARs are the regions where the Y-chromosome pairs and exchanges genetic material with the X chromosome (Quintana-Murci and Fellous 2001). Hence genetic elements located within these PARs, are inherited in an autosomal manner. The non-recombining region of the Y-chromosome (NRY) is made up of heterochromatin (30Mb) and euchromatin (24Mb) and is always in a haploid state. Unless a mutation occurs, the NRY is inherited intact through paternal lineages (Quintana-Murci et al. 2001).

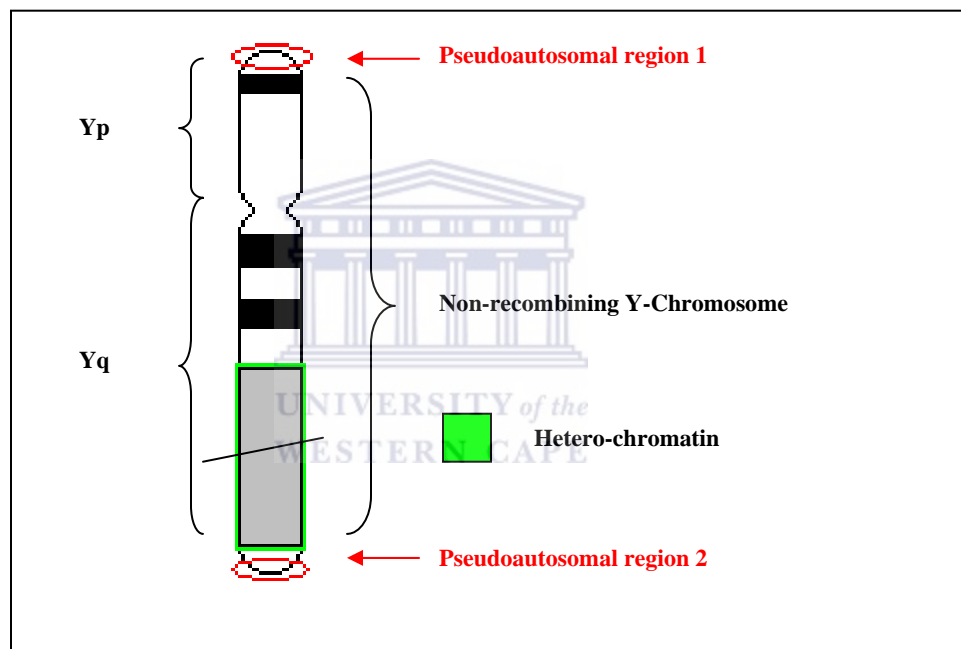


Figure 1.3. Y-chromosome structure indicating pseudoautosomal regions 1 and 2, and the non-recombining region of the Y-Chromosome (from Quintana-Murci and Fellous 2001)

The uni-parental inheritance pattern of the NRY makes it useful for several applications involving male identity testing. Some applications include: (1) paternity testing in particularly deficiency cases, (2) identification of male remains after disasters, (3) investigation of male lineages for anthropological purposes and (4) identification of male perpetrators in criminal cases, such as sexual assault (Kurihara et al. 2004; Koyama et al. 2002; Jobling and Tyler-Smith 1995; Ploski et al. 2002; Dettlaff-Kakol and Pawlowski 2002).

1.3. Discovery of Y-STR loci

By the mid 1990's only a few Y-STR loci with the potential for use in forensic studies, were known. These included the dinucleotide Y-STR loci YCA I, YCA II, and YCA III, the trinucleotide Y-STR loci DYS388 and DYS392, and the tetranucleotide Y-STR loci DYS19, DYS288, DYS385, DYS389, DYS390, DYS391, DXYS156Y, and DYS393 (Kayser et al. 1997). While some of these are single-copy loci, YCA I, YCA II, YCA III and DYS385 are duplicated on the Y-chromosome. Duplicated loci generate two fragments when subjected to PCR using one set of primers. The forward primer-binding site of DYS389 is also duplicated, yielding two products (DYS389 I and DYS389 II) that are approximately 100bp apart.

A collaborative study was undertaken to assess the suitability of these loci for forensic studies (Kayser et al. 1997). A total of 3825 males from 48 different population groups from Europe, America, Asia, Africa and Oceania, were typed. Gene diversity values ranged from low (DYS288, DYS388, DXYS156Y and YCA I), moderate (DYS391, DYS392 and DYS393), high (DYS19, DYS390, DYS389I/II and YCA II) to very high (DYS385 and YCA III). As a result, it was suggested that seven Y-STR loci (DYS19, DYS389I/II, DYS390, DYS391, DYS392, and DYS393) be used routinely in forensic applications. The inclusion of DYS385 with these loci constitutes what is commonly referred to as the 'minimal haplotype' (<http://www.yhrd.org>). Inclusion of YCA II to this core set was referred to as the 'extended haplotype'.

Prior to the release of large amounts of Y-chromosome sequence, attempts to identify novel Y-STR loci depended on a series of cloning and hybridization steps. In the most recent and almost certainly the last example of this approach, a cosmid library was constructed from flow-sorted human Y-chromosomes (White et al. 1999). Probes containing the repeated element [GATA]₁₀ or [TATC]₁₀ were used to select cosmids and subclones containing repeated GATA elements. This approach led to the identification of seven loci with the potential for use in human identity testing (Y-GATA-A4, Y-GATA-A8, Y-GATA-A10, Y-GATA-C4, Y-GATA-H4, and Y-GATA-A7.1 and Y-GATA-A7.2). At the time this work practically doubled the number of known tetranucleotide Y-STR loci.

The release of substantial amounts of Y-chromosome sequence data allowed for a more straightforward approach to the identification of novel Y-STR-loci. Ayub et al. (2000) surveyed 1.22Mb of Y-chromosome sequence, identifying 25 STR sequences. A subset of six loci was selected for further analysis in a sample of 278 Pakistani males. While gene diversity values for several of the loci were low (DYS436 – 0.064; DYS435 – 0.070; DYS434 – 0.222), three loci appeared to be relatively polymorphic (DYS437 – 0.664; DYS438 – 0.684; DYS439 – 0.728). These loci have been incorporated into commercial Y-STR typing systems and are widely accepted by the forensic community. The Scientific Working Group on DNA Analysis Methods (SWGDM) has recently recommended that DYS438 and DYS439 replace the YCA II locus in the Forensic Y User Group's 'extended haplotype'. This is largely due to technical difficulties often encountered when typing dinucleotide STR loci such as YCA II.

Recently a substantial number of STR loci have been identified from Y-chromosomal sequence data. Iida et al. (2001, 2002) characterized five loci (DYS441, DYS442, DYS443, DYS444, and DYS445). A more thorough survey was conducted by Redd et al. (2002a), resulting in the identification of 14 novel Y-STR loci (DYS446, DYS447, DYS448, DYS449, DYS450, DYS452, DYS453, DYS454, DYS455, DYS456, DYS458, DYS459, DYS463 and DYS464). By far the most comprehensive analysis of this kind has been conducted by Kayser et al. (2004).

The 23Mb of sequence surveyed by Kayser et al. (2004) represents almost all the euchromatic sequence of the Y-chromosome. Loci were selected with repeated elements ≥ 3 and ≤ 6 bp in size. Loci with dinucleotide repeated elements were avoided since they have a propensity to generate PCR 'stutter' products or 'shadow' bands. These 'stutter' products are often generated by the PCR amplification of STR loci and are generally one repeat unit shorter than the primary product. This approach resulted in the identification of 475 potential Y-STR loci of which 45 had previously been identified. PCR primers were successfully designed for 281 loci *in silico*. Of those, 166 primer-sets generated male-specific amplicons and 139 loci were demonstrated to be polymorphic in a group of eight individuals representing different binary-marker haplogroups. Using this new sequence-based approach, the pool of Y-STR loci available for evolutionary, paternity testing and forensic casework has expanded considerably.

1.4. Multiplex PCR amplification of Y-STR loci

Analysis of Y-STR loci is achieved by PCR amplification and fragment analysis to determine the differences in sizes. As a direct result of its uni-parental pattern of inheritance, the product rule used for estimating polymorphism in autosomal chromosomes cannot be applied to Y-chromosomes (Butler et al. 2002). More Y-STR loci would therefore be needed to obtain a haplotype with the same power of discrimination as an autosome of the same size and with the same density of loci (Bosch et al. 2002). It would therefore be advantageous to amplify as many polymorphic Y-STR loci so as to increase the power of discrimination of the resultant haplotype. Multiplex PCR is an effective means of achieving this objective.

Multiplex PCR is a modified form of PCR, in which two or more target DNA sequences are simultaneously amplified in the same reaction. Multiplex PCR has successfully been applied to the analysis of deletions, mutations and polymorphisms (Henegariu et al. 1997). The technique has many advantages over uniplex PCR: (1) multiplex PCR requires significantly less time than uniplex PCRs, (2) smaller amounts of PCR reagents are needed to amplify the same target sequences, and (3) consumption of the collective amount of template DNA is significantly reduced. The technique also presents challenges. Often the DNA target sequences do not amplify equally, and the presence of several primer-sets increase the chances of generating non-specific products. This not only makes the PCR less efficient, but also makes analysis of the results more difficult. A substantial amount of optimization is required for a multiplex PCR to work in such a way that all target sequences (and only the target sequences) are consistently and equally amplified. Various approaches have been developed to overcome the challenges presented by multiplex PCR and following will be a brief discussion of a few of these approaches.

1.4.1 Approaches to multiplex PCR development

One of the earlier approaches to multiplex PCR optimization investigated factors that could affect the performance of PCR (Henegariu et al. 1997). The group selected 22 primer pairs on chromosome 12, and 24 primer pairs on the Y-chromosome. A comprehensive analysis of PCR parameters was undertaken in order to determine whether these parameters could be used to improve multiplex PCR performance. They investigated cycling conditions in terms of cycle numbers, annealing temperature and annealing duration as well as extension temperatures and extension duration. The ideal concentration of PCR reagents such as Taq polymerase, PCR buffer, dNTPs and MgCl₂ as well as the ideal dNTP/MgCl₂ ratio were investigated. The effect of individual primer concentrations on the yield of individual PCR products and the effect of

additives such as BSA, glycerol and DMSO were also studied. Based on the results of the study, a generalized protocol (Figure 1.4) was designed, with guidelines on how to approach the optimization of a multiplex PCR.

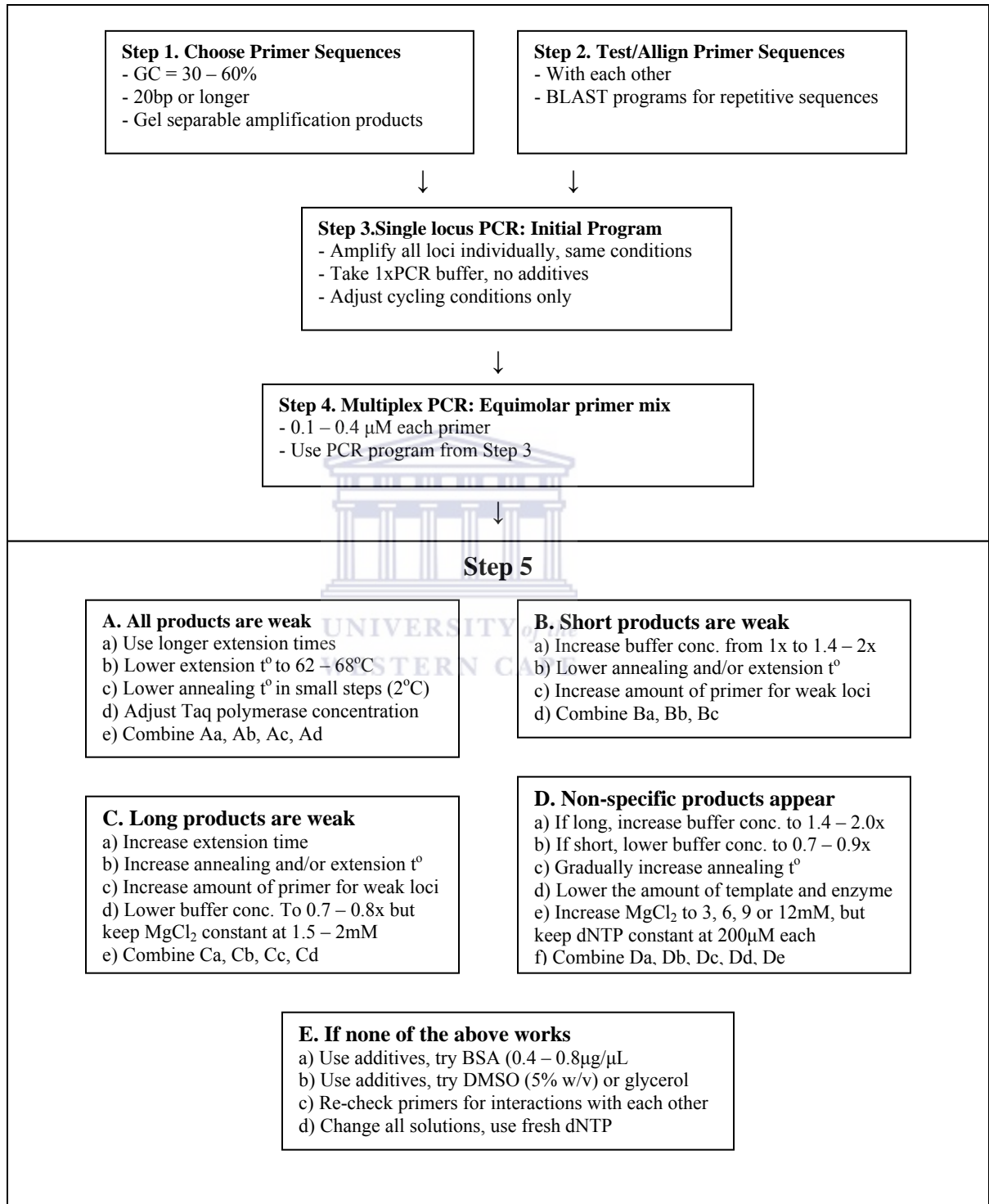


Figure 1.4. A generalized protocol with guidelines on how to approach the optimization of a multiplex PCR (Henegariu et al. 1997)

Another, more rigorous, optimization protocol was established by a group attempting to construct universal multiplex PCR systems for comparative genotyping (Wallin et al. 2002). The three areas of investigation were primer selection, PCR amplification and fluorescent allele detection. A primer selection process was based on extensive testing of numerous primers in uniplex and multiplex PCRs, as well as testing for amplification specificity and primer binding site mutations. Vast amounts of candidate primers were screened in uniplex PCRs for signal strength, and only those with the highest signal strength were chosen. Primer concentrations were varied to establish a performance window in which signal strength reached a plateau. These primers were combined in a multiplex PCR at equimolar concentrations and the balance between the PCR products for each locus was investigated. By adjusting primer concentrations, signal strength was maximized so that all amplified products labelled with the same fluorescent dye would have similar peak heights (different fluorescent dyes have different emission wavelengths and some dyes will therefore appear brighter than others). Primers that could successfully be combined in this way were then tested for amplification specificity. A series of parameters were changed that would reduce the stringency of the multiplex PCR including reducing annealing temperatures, increasing DNA template, using single stranded Chelex extracted DNA and increased MgCl₂ concentrations were used to lower the stringency. Primers that produced non-specific artifacts at a lower stringency were rejected from being used in the multiplex. The primers were then investigated for primer binding site (pbs) mutations by conducting database searches. Where pbs mutations were found, primers were either re-designed to avoid mutations or degenerate primers added to the multiplex to compensate for the mutation. Optimal PCR cycling parameters and the most appropriate composition of PCR reagents was established empirically.

More recently a multiplex design strategy focusing on primer sequence analysis was used to construct a multiplex capable of amplifying up to 20 Y-STR loci (Butler et al. 2002). The reasoning behind the approach is as follows: Since multiplex PCR uses one set of cycling conditions, primers in a multiplex PCR should possess similar annealing temperatures. These primers should also show no significant interactions within themselves, between one another or with undesirable sequences in the DNA template used. This would ensure the efficiency of the multiplex PCR by limiting the occurrence of both primer-dimers and non-specific amplified artifacts. The approach was presented in a flow diagram (Figure 1.5) divided in two parts; the first part described an *in silico* primer design and testing and the second part described experimental amplification optimization. In this way a successfully optimized multiplex PCR was created that could amplify all the desired loci and achieve similar yields between the

amplicons of all loci. The resulting profiles were free from non-specific artifacts and included amplicons that were easy to distinguish from other loci in the multiplex.

	Comments
<p style="text-align: center;">B) Multiplex PCR primer mixture optimization and testing</p> <div style="text-align: center;"> <div style="border: 1px solid black; padding: 5px; width: fit-content; margin: 0 auto;">Check primer quality</div> <div style="margin: 5px 0;">↓</div> <div style="border: 1px solid black; padding: 5px; width: fit-content; margin: 0 auto;">Test primers in singleplex PCRs</div> <div style="margin: 5px 0;">↓</div> <div style="border: 1px solid black; padding: 5px; width: fit-content; margin: 0 auto;">Combine primers of individual loci together in multiplex</div> <div style="margin: 5px 0;">↓</div> <div style="border: 1px solid black; padding: 5px; width: fit-content; margin: 0 auto;">Empirically balance primer conc. based on product yields</div> </div>	<p>Use TOF-MS. Re-order primers that are of poor quality, as they will influence PCR efficiency</p> <p>Individual primer sets should amplify respective target sequences under identical PCR conditions</p> <p>Well-balanced amplicons should be achieved by adjusting primer concentrations</p>

Figure 1.5. Primer design-based approach to multiplex optimization (Butler et al. 2002)

1.4.2 Validation of multiplex PCR

Multiplex typing systems must be optimized to the point where they meet certain performance standards. There are several governing bodies that ensure that high typing and analysis standards are maintained. Among these are the International Society for Forensic Genetics (ISFG), the Scientific Working Group on DNA Analysis Methods (SWGDM), and the European DNA Profiling Group (EDNAP). These organizations have proposed guidelines for the use and validation of multiplex PCR typing systems. Some common validation exercises include: (1) establishing that the typing system is sensitive and performs consistently using freshly prepared and stored DNA, (2) that identical results are obtained irrespective of the type of tissue from which DNA was extracted, (3) that the systems yield consistent results in several laboratories, and (4) that the system performs well when used to analyze samples similar to those encountered in forensic casework. In this regard efficient typing should be obtained for DNA extracted from body fluids mixed with commonly encountered substances (e.g. dyes, soil, leather, denim). The influence of environmental factors such as temperature, humidity and UV should also be established and the capacity of the system to analyze mixtures of male and female DNA should be examined.

1.4.3 Non-commercial Y-STR multiplex PCR

With the availability of the human genome sequence, an understanding of the parameters affecting multiplex PCR (Henegariu et al. 1997) and strategies to effectively optimize multiplex PCR (Wallin et al. 2002; Butler et al. 2002), the development of Y-STR multiplex typing systems has become less challenging. Several Y-STR multiplex PCRs able to amplify up to 21 loci in one reaction have been established.

The first of these was a Y-STR multiplex that amplified 20 loci (Figure 1.6) with the use of 17 primer pairs and a 5-dye detection system (Butler et al. 2002). This multiplex included all the loci of the MH, YCA II of the extended haplotype, as well as some loci recently identified by White et al. (1999), Ayub et al. 2000 and Redd et al. (2002a).

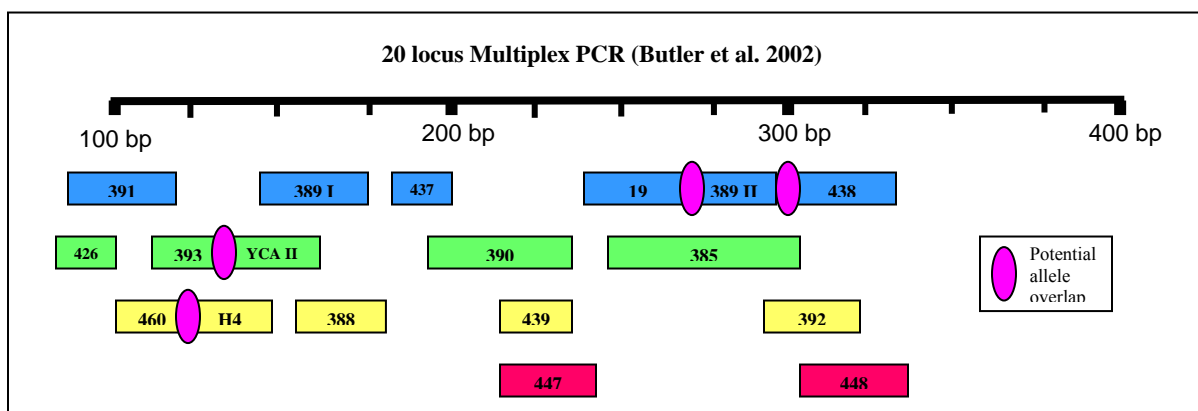


Figure 1.6. Predicted PCR product size range for loci amplified in 20-plex (Butler et al. 2002)

The next reported multiplex PCR (Figure 1.7) by the same group of researchers co-amplified 10 Y-STR loci in one multiplex PCR (Schoske et al. 2003). Only three of the MH loci (DYS19, DYS391 and DYS392) are included in this multiplex along with some other loci identified by White et al. (1999) and Ayub et al. (2000).

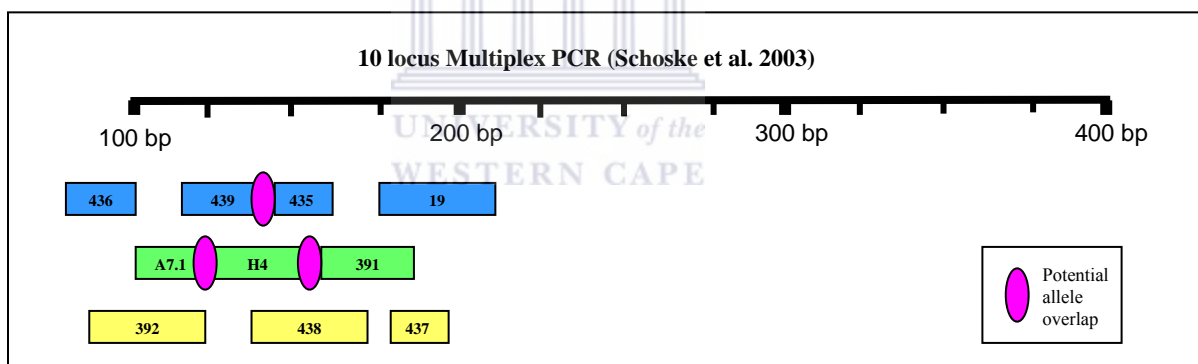


Figure 1.7. Predicted PCR product size range for loci amplified in 10-plex (Schoske et al. 2003)

The same group described yet another multiplex (Figure 1.8), one which co-amplifies 11 loci (Schoske et al. 2004). The multiplex on its own combines DYS385 of the MH with six Y-STR loci recently identified by Redd et al. (2002a), one of which is a multi-copy locus (DYS464). This multiplex was designed to be used on its own, but can also be analyzed with most of the loci of the 10-plex (shaded in Figure 1.8) previously described by this group (Schoske et al. 2003).

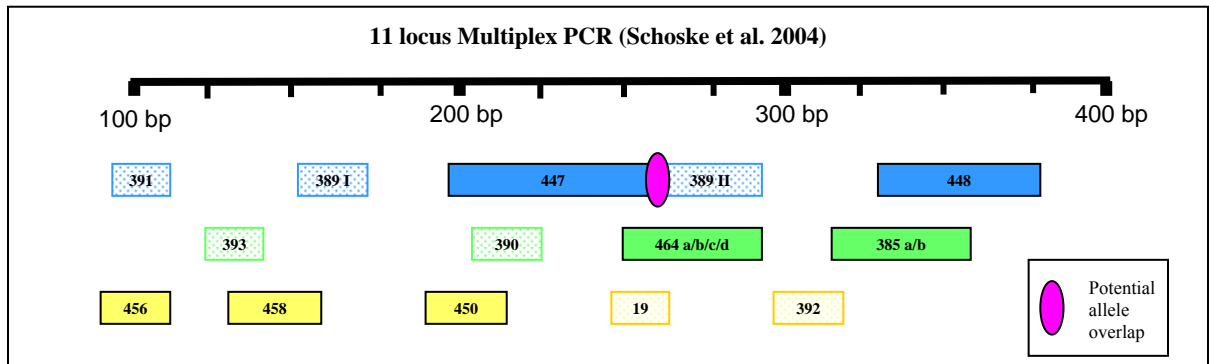


Figure 1.8. Predicted PCR product size range for loci amplified in 20-plex (Schoske et al. 2004)

The largest multiplex described to date, is a 21-locus ‘megaplex’ (Hanson and Ballantyne 2004). It does not contain any of the loci of the MH, and is made up entirely of recently identified loci (Figure 1.9). Two of these are multi-copy loci (DYS527 and DYS464). Some preliminary forensic validation in terms of specificity, male:male mixture studies, male:female mixture studies, population studies and non-probative forensic casework has been undertaken for this multiplex. It had been found that in the case of multi-copy loci DYS527 and DYS464, allele designation becomes very challenging because of overlapping alleles.

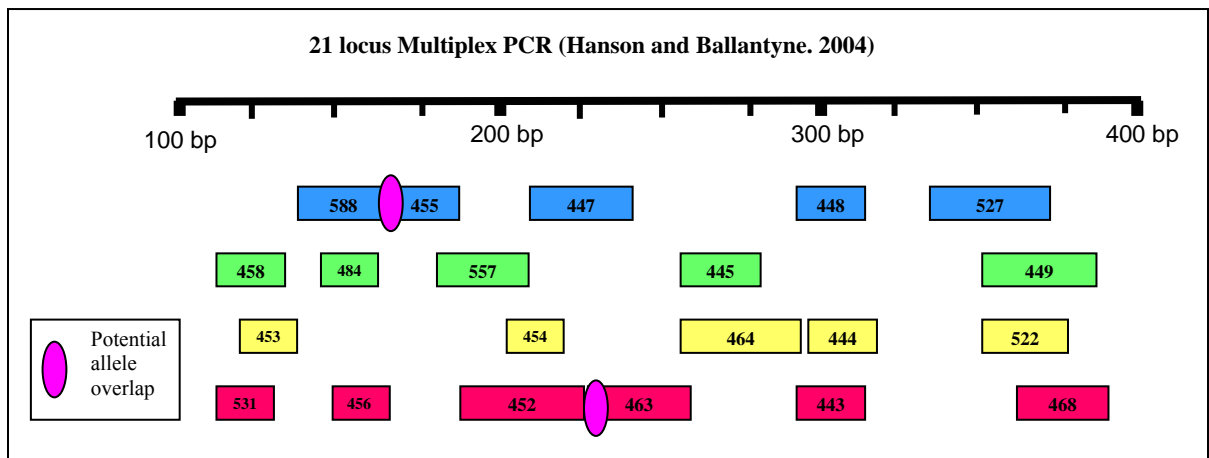


Figure 1.9. Predicted allele size range for loci amplified in 21-plex (Hanson and Ballantyne 2004)

1.4.4 Commercial Kits for the multiplex amplification of Y-STR loci

Multiplex PCR amplification of Y-STR using commercial kits may have advantages. An important advantage is that they undergo extensive testing and quality control, ensuring that they can be used with confidence. Commercial typing kits are provided with allelic ladders that facilitate accurate typing. Kits are also highly sensitive, male-specific and robust, needing very little DNA to obtain a complete profile, even in the presence of excessive amounts of female DNA. Three Y-STR typing kits, **Y-PLEX™12** (*Reliagene*), **Mentype Argus® Y-MH** (*Biotype*) and **PowerPlex Y** (*Promega*) are currently commercially available. The expected sizes of PCR products generated with each of these kits are presented in Figure 1.10 a, b, and c respectively.

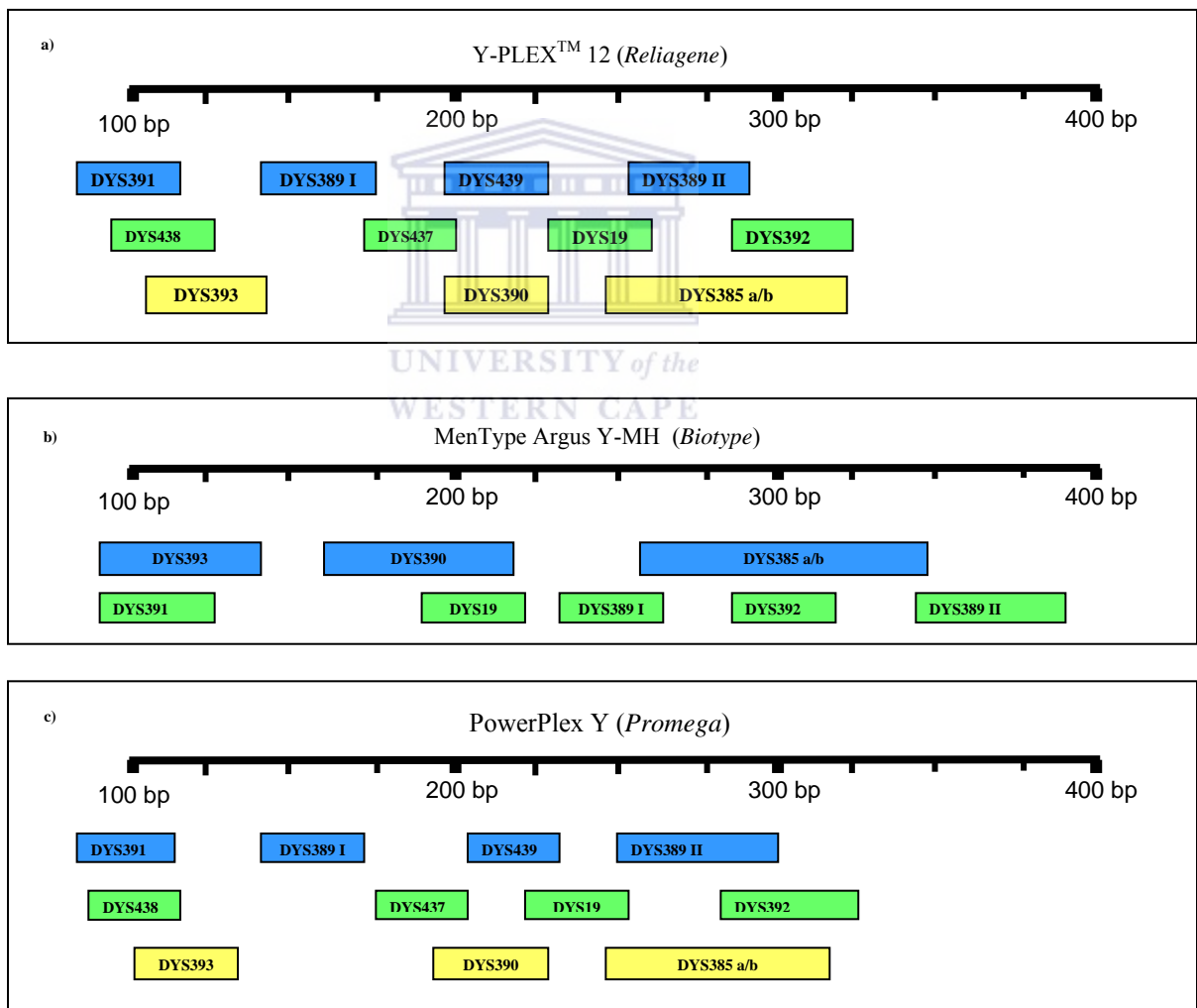


Figure 1.10. Predicted allele size range for loci amplified in (a) Y-PLEX™ 12 kit from *Reliagene*, (b) MenType Argus Y-MH from *Biotype* and (c) PowerPlex Y from *Promega*

A comparison between the three commercially available Y-STR typing kits is presented in Table 1.1. All three kits amplify the nine loci of the MH as recommended by the ISFG. **Y-PLEX™12** and **PowerPlex Y** also amplifies two additional loci (DYS438 and DYS439) that make up the extended haplotype as recommended by the SWGDAM. An added advantage of the **Y-PLEX™12** is the inclusion of the sex-determination locus Amelogenin, which also serves as an internal control for PCR and PCR inhibitors (Shewale et al. 2004). **Y-PLEX™12** has been forensically validated (Shewale et al. 2004) and the forensic validation for **PowerPlex Y** is currently underway.

Table 1.1. Comparison of three commercially available Y-STR typing kits

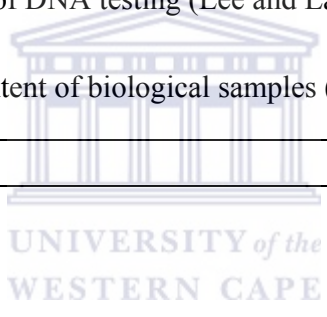
Name of the Kit	Y-Plex™ 12	Mentype Argus MH	PowerPlex Y
Number of loci amplified	12	9	12
Loci	DYS19	DYS19	DYS19
	DYS389 I	DYS389 I	DYS389 I
	DYS389 II	DYS389 II	DYS389 II
	DYS390	DYS390	DYS390
	DYS391	DYS391	DYS391
	DYS392	DYS392	DYS392
	DYS393	DYS393	DYS393
	DYS385 a/b	DYS385 a/b	DYS385 a/b
	Amelogenin		DYS437
	DYS438		DYS438
	DYS439		DYS439
Number of dyes	3	2	3
Amplification platforms	GeneAmp™ 9600 GeneAmp™ 9700 PTC 200 Peltier		
Electrophoresis platforms	ABI 310 ABI 3100 ABI 377 FMBIO III		ABI 310 ABI 3100 ABI 377
Forensically validated	Yes	Not Yet	Underway

1.5. Biological evidence, and the use of Y-STRs in forensic casework

1.5.1 Biological evidence

Biological evidence encountered in forensic casework may include bone, blood and bloodstains, semen and semen stains, tissues, organs, teeth, hair, fingernail clippings, saliva, urine and other biological fluids (Lee and Ladd 2001). Successful identification of individuals from such biological samples, depend on the quantity, degradative state and purity of the DNA in these samples. Table 1.2 gives an indication of the quantities of DNA one can expect to isolate from certain tissue types (Lee and Ladd 2001). These quantities as well as the quality of the DNA can be significantly altered by environmental exposure to light, temperature and humidity changes (Bender et al. 2004). Purity of DNA in samples can also be hampered by the presence of environmental contaminants such as soil, clothing dyes and grease. These contaminants can in turn compromise the efficacy of DNA testing (Lee and Ladd 2001).

Table 1.2. Expected DNA content of biological samples (Lee and Ladd 2001)



Type of sample	Amount of DNA
Liquid blood	20 000 - 40 000 ng/mL
Blood stain	250 - 500 ng/cm ²
Liquid semen	150 000 - 300 000 ng/mL
Post-coital vaginal swab	10 - 3000 ng/swab
Plucked hair (with root)	1 - 750 ng/root
Shed hair (with root)	1 - 10 ng/root
Liquid saliva	1 000 - 10 000 ng/mL
Oral swab	100 - 1500 ng/swab
Urine	1 - 20 ng/mL
Bone	3 - 10 ng/mg
Tissue	50 - 500ng/mg

1.5.2 Use of Y-STRs in forensic studies

While autosomal STRs have an excellent capacity to distinguish between individuals, they do also have disadvantages. It is almost always the case in investigations of sexual assault that the biological material from a male assailant is mixed with biological material from a female victim.

In such cases it is difficult to separate the female victim's profile from the rapist's profile when autosomal STRs are used. Differential extraction techniques may separate sperm and epithelial fractions, but these techniques require fresh samples. If the post-coital interval is extended, the chance of successful differential extraction decreases and becomes practically impossible after 48 hours (Hall and Ballantyne 2003b). This is not because the male component (sperm) is not present. Studies have shown that sperm can survive in the vagina for up to three days post-coitus and in the cervix up to seven days post-coitus (Willott and Allard 1982). Sperm loss in the vagina due to vaginal lavage and drainage, menstruation and the normal intra-cervico-vaginal degradative processes, may leave these surviving spermatozoa in a fragile state (Hall and Ballantyne 2003b). Sexual assault, vaginal inflammation, excessive douching and the administering of spermicides, may also add to the loss of sperm (Sibille et al. 2002). Hence, in the case of extended post-coital intervals after sexual assault, a procedure such as differential extraction is likely to fail. When such DNA evidence samples are then subjected to autosomal STR typing, the profile of the victim may mask the profile of the assailant. Analysis of Y-chromosome markers overcomes all these challenges by generating male specific profiles from very little DNA.

The characterization of sperm degradation was examined by Hall and Ballantyne (2003b). A 19 Y-STR multiplex system was used in this study. A female who had abstained from sexual intercourse for at least three days prior to the start of this investigation was recruited. Post-coital cervico-vaginal or lower to mid-vaginal swabs were taken from the subject at various time intervals (0h, 12h, 24h, 48h, 72h, 85h, 4-6 days). DNA from the swabs at the various time intervals was extracted using a differential extraction procedure as well as a non-differential extraction procedure. PCR amplification was done on DNA from the sperm fraction, the non-sperm fraction of the samples as well as DNA from the mixed sample using 3-450ng of input DNA for both multiplex reactions.

The results showed that when DNA from the mixed samples were extracted in a non-differential manner and then subjected to Y-STR typing, full male profiles could be discerned up to 48h post-coitus. In contrast to this, when DNA was extracted from the sperm fraction in a differential manner, full male profiles could only be discerned up to 12h post-coitus. It was also noted that when the post-coital interval reaches 12h, the autosomal typing of mixed sample DNA masks the profile of the male component, whereas full male profiles can be discerned from the sperm fraction up to 12h post-coitus. The results also showed that male components can be found in the

non-sperm fraction of the differentially extracted DNA at 12h post coitus, suggesting that premature lysis of sperm cells can occur at this early stage.

Shewale et al. (2004) summarized that Y-STR loci are therefore superior to autosomal STR loci in sexual assault applications because: (1) Differential extraction of sperm and non-sperm fractions in mixed samples is not necessary, (2) Analysis of samples from azoospermic or oligospermic males as well as samples from vasectomized or orchidectomized males is feasible, (3) Unambiguous male profiles can be obtained in the presence of overwhelming amounts of female DNA, (4) Because of the haploid nature of the NRY, the number of male contributors can be identified in incidents where multiple assailants are involved, (5) Y-STR use ensures speedy exclusion of suspects and (6) The 'single allele per locus' concept simplifies analysis of results.

1.6 Forensic casework and the YHRD

In 1994 an attempt was made to construct a reliable database for the practical use of the Y-chromosome loci in forensic studies. The aims of this enterprise were to: (1) identify polymorphisms capable of discriminating between the majority of unrelated lineages in a given population, (2) to establish a database representative of geographical and ethnical structure of the populations of interest, and (3) to eventually create a database of an appropriate size that would allow accurate frequency estimation for rare haplotypes (Roewer et al. 2001). This database is currently referred to as the Y-chromosome Haplotype Reference Database (YHRD).

A core set of Y-STR loci (DYS19, DYS389 I/II, DYS390, DYS391, DYS392, DYS393, and DYS385) were identified as being suitable for forensic casework. This core set was referred to as the European 'minimal haplotype' (MH) because of the authors' view that they represent the minimum requirements for sufficiently informative haplotyping in forensic casework. Addition of the duplicated YCA II locus constituted the 'extended haplotype'. Typing of the MH loci in different populations and submissions of these data to the YHRD has resulted in the establishment of the most comprehensive Y-STR haplotype database available (<http://www.yhrd.org>).

Many other European laboratories have since undertaken the use of the MH loci for forensic casework and paternity testing. To ensure accurate haplotyping and standardized use of

nomenclature by all contributors, a quality assurance test was created (Roewer et al. 2001). The test involves haplotyping five high molecular weight DNA samples for the nine loci of the MH (or the 11 loci of the 'extended haplotype'). The role of the YHRD database now is to support the presentation of Y-STR evidence in court cases, by serving as a central repository for Y-STR haplotypes which allows population frequency estimates to be obtained in a fast, non-commercial and comprehensively documented way (Roewer et al. 2001). The database can be searched for complete 'extended haplotypes', complete minimal haplotypes, and partial haplotypes – even down to one single locus. The size of the database is also now big enough to start estimating the frequency of rare haplotypes. Other useful information such as contributing laboratories, PCR primers and amplification protocols, population analysis calculations, references and links to other relevant sites are also available (Roewer et al. 2001).

Following is an example of an actual case to illustrate how valuable the YHRD is in forensic casework. In 2002 a woman was found in her Berlin apartment with a smashed skull and covered in blood. She was fortunate to survive and eventually told the police that a man who had been sub-letting the apartment next door attacked her. Evidence collected from the crime scene and the next-door apartment was analyzed with the use of autosomal STR loci, and the profiles of three individuals identified. One profile was that of the tenant who was eliminated as a suspect because he was not present at the time of the incident. Various males of European and African descent had previously used the apartment, and hence the two other autosomal profiles did not add much information at the time. Y-STR analysis was performed on all the evidence collected. The aim of this investigation was to estimate population affiliation of the two unidentified males by comparing their respective Y-STR haplotypes to the haplotypes contained in this database (Roewer 2004).

As the nationality of the tenant was already known, his Y-STR profile was used as a confirmatory test to see whether the method would work. The haplotype of the tenant accurately pointed to a population affiliation in southern Europe. The minimal haplotype of the two other suspects however did not occur as frequently in European populations (only 0.05%). Nevertheless, the haplotype strongly suggested that African descent was unlikely. The Y-STR profiles also strongly suggested that the two suspects shared a paternal lineage. With the knowledge that the tenant was Italian and that the other two suspects were likely to be European, further investigative work led to the location of one of the suspects in Italy. It turned out that the nephew of the tenant was the attacker, and was wanted in Italy for another violent crime (Roewer 2004).

1.7. Population Studies

The presentation of forensic data in court requires background information on local population structures. The results of the South African census conducted during 2001 showed that there were almost 45 million people living in South Africa. Table 1.3 shows the racial distribution of these people at the time that the census was conducted (www.census.gov.za).

Table 1.3. Racial distribution of South African populations

Race	%	Number of Individuals
Black African	79.01905	35 416 166
White	9.57978	4 293 640
Coloured	8.91237	3 994 505
Asian/Indian	2.48878	1 115 467
Total	100	44 819 778

The majority of the inhabitants of SA are referred to as ‘Black Africans’, but this definition alone does not really give information about their ancestry. Instead a more realistic approach to defining the ancestry of the ‘Black Africans’ of South Africa would be to look at their language groups. Figure 1.11 shows the distribution of the South African population based on their home language. The minority populations in South Africa speak Afrikaans and English as home languages. The other nine (of the 11 official languages) are Bantu languages (Lane et al. 2002). These are the languages spoken by ‘Black Africans’ as home language.

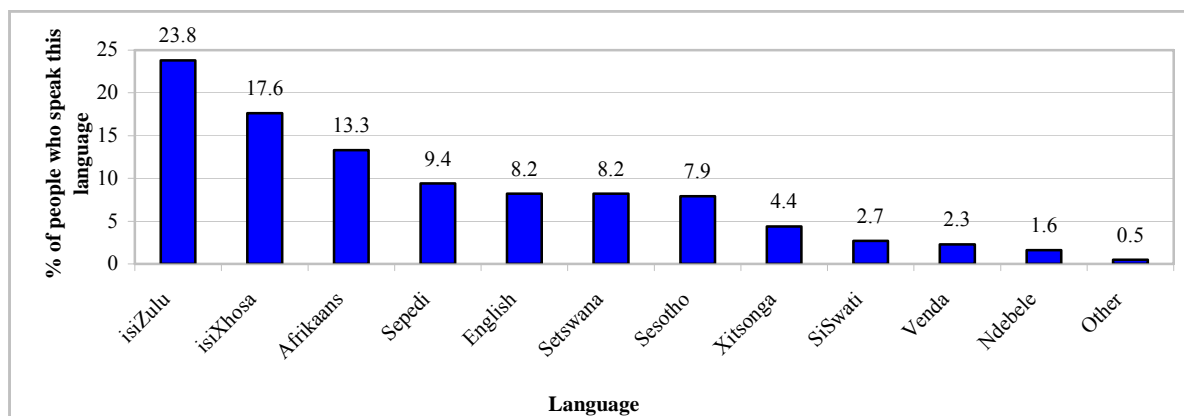


Figure 1.11. South African population distribution by language

A recent study was conducted to determine the genetic affiliations of these different Bantu speaking populations (Lane et al. 2002). Males were recruited who spoke one of the nine Bantu languages spoken in South Africa as a home language. Autosomal STR typing as well as Y-STR typing was performed. Figure 1.12 shows the genetic affinities between these Bantu speakers based on only Y-STR data.

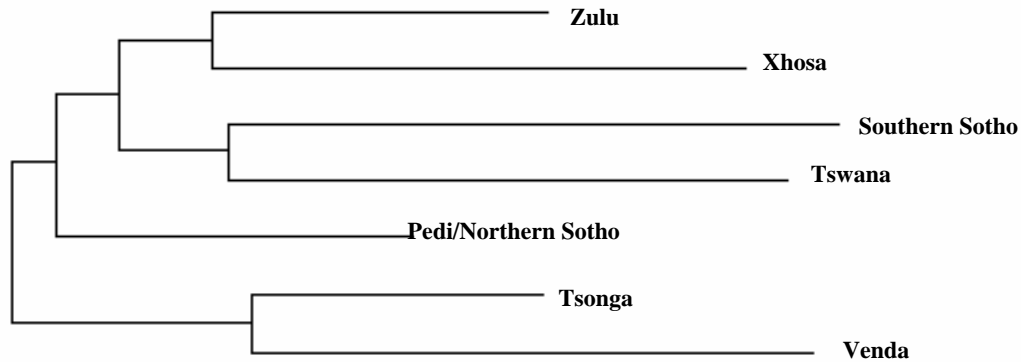


Figure 1.12. Neighbour joining tree showing genetic affinities between South African Bantu speakers using Y-STR haplotype data only (Lane et al. 2002)

1.8 Objectives of this study

The objective of the present study was to select Y-STR loci that could compliment those already in use in forensic studies, and possibly replace those with limited polymorphism among South African populations. This was prompted by an initial study revealing extremely low levels of polymorphism for two loci of the minimal haplotype (DYS391 and DYS392) in Xhosa populations in Cape Town, South Africa (Leat et al. 2004). Of the 99 individuals typed in the study only 47 unique haplotypes were observed and 13 individuals shared the most common haplotype. DYS391, DYS392 and DYS437 were also found to be virtually mono-morphic in a population surveyed in Maputo, Mozambique (Álves et al. 2003). Three criteria were used to select markers for assessment. Firstly, the single-copy markers of the minimal haplotype were selected based on their established use in forensic studies. Secondly, markers would be selected on the basis of high gene diversity values reported for several population studies. Thirdly, markers would be chosen from a survey of Y-chromosome sequence data with selections made primarily on the basis of the number of repeated elements present. Multiplex reactions would

then be developed as a means to analyze the properties and their suitability for use in forensic studies among populations likely to have distinct population history.

Chapter 2 – Investigation of ‘Minimal Haplotype’ loci among South African sub-populations

2.1 Introduction

The most widely used Y-STR loci in forensic studies are those that constitute the European ‘minimal haplotype’ (MH). An initial investigation of the MH loci among South African males was conducted by Leat et al. (2004). Two South African sub-populations, the English-speaking Caucasian- and Xhosa-speaking Black populations were included in this study. It was noted that some of the MH loci showed very low levels of variability among the Xhosa-speaking population. In a sample of 99 individuals, 93% shared allele 10 for DYS391 and 96% shared allele 11 for DYS392. This was not unlike results obtained for studies involving other sub-Saharan populations. In a sample of 112 individuals from Maputo, Mozambique, Alves et al. (2003) found that 82% shared allele 10 for DYS391 and 99% shared allele 11 for DYS392. In a sample of 31 males from Central Africa, Kayser et al. (2001) also reported that 97% shared allele 11 for DYS392. Since these loci are almost mono-morphic for the Xhosa-speaking population and possibly several other sub-Saharan populations, DYS391 and DYS392 might be of limited use in forensic studies. It was also noted that the discriminatory capacity of the MH in these two sub-populations was not ideal. Of the 100 English-speaking Caucasian individuals typed, 8% shared the most common haplotype and of the 99 Xhosa-speaking Black individuals typed, 13% shared the most common haplotype. Through collaboration with the South African Blood Transfusion services, access was gained to biological material from three more sub-populations, two from the Western Cape and a third from KwaZulu-Natal. These groups were studied to accumulate more information on MH loci among other South African sub-populations.

2.2 Methods and Materials

2.2.1 Obtaining biological samples for the study

Three other South African sub-populations were investigated in this study. These were Asian Indian males from KwaZulu-Natal, Afrikaner Caucasian males and males from the Coloured community of the Western Cape. Ethical clearance for this exercise was obtained through the Senate Research committee of the University of the Western Cape. Blood samples were collected by the nursing staff of the relevant blood transfusion services and stored in 5ml purple-top tubes (*BD Vacutainer Systems*) containing 100µl of 15% EDTA. These tubes were kept at room temperature during collection periods and then at 4°C until DNA extraction took place. Samples were collected from Asian Indian males, Afrikaner Caucasian males and males from the Coloured community. Each donor's blood containing tube was given a unique code descriptive of the sub-population to which the sample belonged. All subsequent procedures utilized the unique codes.

2.2.2 DNA Extraction

DNA extraction from whole blood was performed according to a previously described technique (Lahiri and Nurnberger 1991). During the DNA extraction procedure, appropriate protective wear (white lab-coat, two pairs of latex gloves and a translucent facial screen) was worn. To prevent possible contamination of isolated DNA by amplicons, DNA was extracted in a part of the laboratory designated to pre-PCR procedures. Working stock dilutions of 2ng/µl were made from these DNA samples for further use in experiments. Comprehensive protocols are presented in the Appendix.

2.2.3 PCR Amplification of MH loci

All nine MH loci were amplified in a single multiplex reaction. Amplifications were performed in a final volume of 10µl containing 4ng genomic DNA, 16mM Tris-HCl (pH8.3), 80mM KCl, 1.5mM MgCl₂, 200µM dNTPs, and 1U of AmpliTaq Gold (*Applied Biosystems*). Primers were synthesized by *MWG-Biotech* using previously reported sequences (Kayser et al. 1997). The primer sequences, corresponding fluorescent dye labels and the final concentration at which each primer was used, are indicated in Table 2.1.

PCR amplification was performed in a GeneAmp 2400 thermocycler (*Applied Biosystems*) as follows: 1 cycle at 95°C for 10 minutes, 30 cycles of 94°C for 1 minute, 55.5°C for 2 minutes, 68°C for 3 minutes, followed by 1 final cycle at 68°C for 75 minutes. Post-PCR modification by the non-template addition of dATP to products was enhanced by an additional step. This involved

adding 5µl of an ‘A-tailing mixture’ containing 10mM Tris-HCl (pH8.3), 50mM KCl, 1.5mM MgCl₂, 600µM dATP, and 0.5µl Taq polymerase to the amplified PCR product. These 15µl mixes were then incubated on the same thermal cycler at 68°C for 2 hours. PCR products were stored in the dark at 4°C.

Table 2.1 Sequences, fluorescent dye labels and final concentration of primers used in ‘minimal haplotype’ (MH) multiplex reaction

Primer	Primer Sequence	[Primer]
DYS19F	6-FAM-5'-cta ctg agt ttc tgt tat agt 3'	0.12µM
DYS19R	5'-atg gcc atg tag tga gga ca 3'	0.12µM
DYS389F	6-FAM-5'-cca act ctc atc tgt att atc tat 3'	0.23µM
DYS389R	5'-tct tat ctc cac cca cca ga 3'	0.23µM
DYS390F	HEX-5'-tat att tta cac att ttt ggg cc 3'	0.08µM
DYS390R	5'-tga cag taa aat gaa cac att gc 3'	0.08µM
DYS391F	HEX-5'-cta ttc att caa tca tac acc ca 3'	0.05µM
DYS391R	5'-ctg gga ata aaa tct ccc tgg ttg caa g 3'	0.05µM
DYS392F	TET-5'-tca tta atc tag ctt tta aaa aca a 3'	0.125µM
DYS392R	5'-aga ccc agt tga tgc aat gt 3'	0.125µM
DYS393F	TET-5'-gtg gtc ttc tac ttg tgt caa tac 3'	0.05µM
DYS393R	5'-aac tca agt cca aaa aat gag g 3'	0.05µM
DYS385F	HEX-5'-gtg aca gag cta gac acc atg c 3'	0.14µM
DYS385R	5'-cca att aca tag tcc tcc ttt c 3'	0.14µM

2.2.4 Fragment Analysis

Amplified fragments were analyzed using an ABI 310 Genetic Analyzer (*Applied Biosystems*). Amplified PCR fragments were prepared for analysis by adding 25µl of de-ionized formamide (*Applied Biosystems*) and 0.5µl of TAMRA 500 size standard (*Applied Biosystems*) to 3µl of PCR product. These mixes were then heat-denatured on a thermocycler (*Hybaid Omn-E*) at 95°C for 4 minutes and then immediately snap-cooled on ice for a minimum of 3 minutes. The fragments were analyzed using virtual filter-set C and run module ‘GS STR POP 4’. The time of injection was 15 seconds and the run time per sample was 26 minutes. In addition to the presence of a size standard included with every sample, allelic ladders were included with approximately every 46 samples run. The allelic ladders contained the most commonly found alleles for each MH locus.

2.2.5 Data Analysis

A quality assurance exercise as administered by the International Forensic Y-User Group (ISFG) (<http://ystr.charite.de>), was successfully completed. This ensured that genotyping corresponded with international nomenclature. Typing was conducted according to the published nomenclature and the ISFG guidelines for Y-STR analysis (Gill et al. 2001). Gene diversity (D) was calculated as $1 - \sum P_i^2$, where P_i is the allele or haplotype frequency (Nei 1987). Gene diversity values ≤ 0.5 were considered low and gene diversity values ≥ 0.7 were considered high.

2.3 Results and Discussion

2.3.1 MH multiplex performance

A total of 312 blood samples were collected and DNA was successfully extracted from 309 samples (85 Asian Indian, 109 Afrikaner Caucasian and 115 from the Coloured community). The DNA extraction technique proved to be efficient, with 90% of the stock DNA having concentrations of more than 100ng/ μ l. Amplified fragments, run on an ABI 310 instrument, typically generated electropherograms that could easily be interpreted. An example of such an electropherogram for the MH multiplex reaction is presented in Figure 2.1.

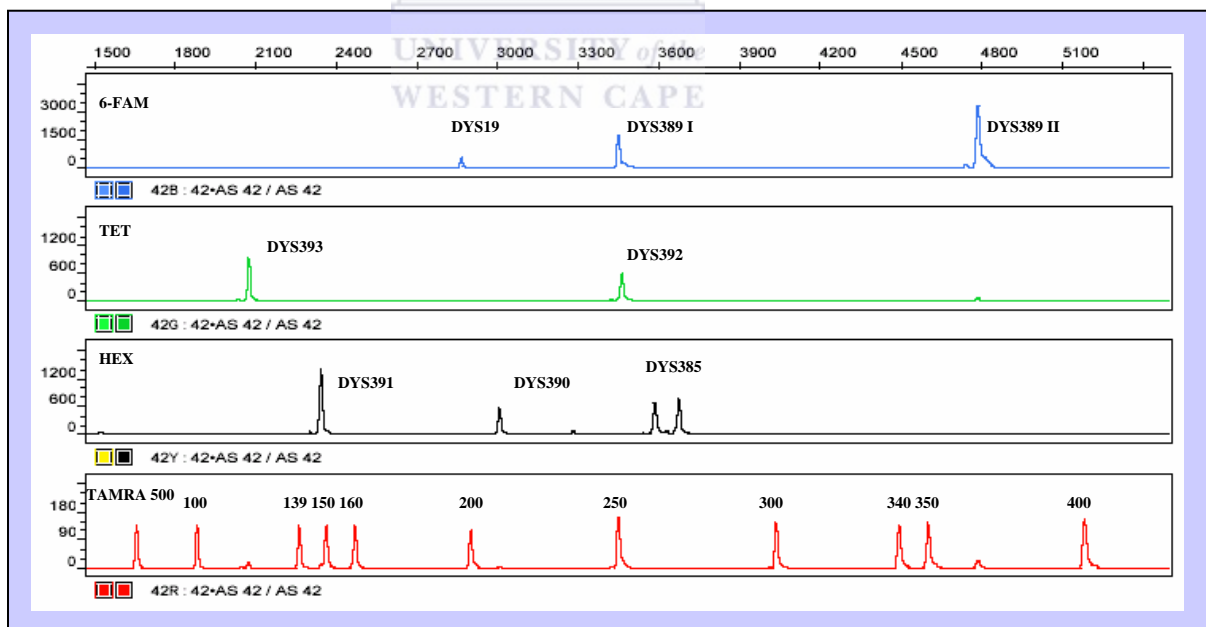


Figure 2.1. An example of a typical electropherogram of the MH multiplex reaction

2.3.2 Analysis of allele and haplotype frequencies

Gene diversity (D) values were calculated for all seven MH loci among the three sub-populations investigated. Table 2.2 presents allele and haplotype frequencies for MH loci among the Asian Indian males from KwaZulu-Natal. Table 2.3 presents allele and haplotype frequencies for MH loci among the Afrikaner Caucasian males from the Western Cape. Table 2.4 presents allele and haplotype frequencies for MH loci among males from the Coloured community of the Western Cape.

DYS385 is a duplicated locus and as such, generates complex profiles using a single primer set. While this has the advantage of providing a high discriminatory capacity for this primer set, it also means that clear locus-allele relationships cannot be established. Duplicated alleles for this locus were treated as a haplotype. The highest D value among all three sub-populations investigated was consistently obtained with DYS385. While DYS389 is also a duplicated locus, clear locus-allele relationships can be established. Alleles of this locus have been assigned to two separate loci.

2.3.2.1 Allele and haplotype frequencies for Asian Indian males

The D value for DYS385 among the Asian Indian males was 0.935 (Table 2.2). This was the highest gene diversity observed for any locus amongst all three sub-populations investigated. The highest D value for a single-copy locus was 0.734 (DYS390). The lowest D value was 0.279 (DYS391). This is to be expected as over 83% of the samples shared allele 10 for this locus. This was not unlike results obtained for a study involving an Indian population from Orissa, India. Sahoo et al. (2003) reported that 80% of 150 males from this region shared allele 10 for DYS391. Whereas all other MH loci have D values above 0.60, DYS392 has a slightly lower gene diversity value of 0.479 in this sub-population.

Table 2.2. Allele and haplotype frequencies for MH loci among Asian Indian males ($n=85$)

Allele	DYS19	DYS389-I	DYS389-II ^a	DYS393	DYS392	DYS391	DYS390	Haplotype	DYS385
9						0.012		7/15	0.012
10					0.094	0.835		7/16	0.012
11				0.082	0.706	0.153		9/16	0.012
12		0.212		0.318	0.024			11/11	0.012
13	0.059	0.506	0.012	0.400	0.047			11/14	0.153
14	0.271	0.282	0.047	0.200	0.106			11/15	0.035
15	0.447		0.553		0.012			12/14	0.024
16	0.176		0.188		0.012			13/13	0.012
17	0.047		0.200					13/16	0.012
18								13/17	0.071
19								13/18	0.047
20							0.035	13/19	0.047
21							0.412	13/20	0.035
22							0.212	14/14	0.024
23							0.118	14/15	0.024
24							0.188	14/16	0.012
25							0.024	14/17	0.082
26							0.012	14/18	0.035
								14/19	0.012
								14/20	0.024
								15/15	0.012
								15/16	0.035
								15/17	0.106
								15/18	0.012
								16/16	0.059
								16/17	0.047
								16/18	0.012
								17/18	0.012
								17/19	0.012
Gene Diversity	0.690	0.620	0.616	0.692	0.479	0.279	0.734	Gene Diversity	0.935

^a For DYS389-II, the number of repeats was obtained by subtracting the length of the corresponding DYS389-I allele.

2.3.2.2 Allele and haplotype frequencies for Afrikaner Caucasian males

The D value for DYS385 among the Afrikaner Caucasian males was 0.828 (Table 2.3). The highest D value for a single-copy locus was 0.706 (DYS19 and DYS390), although the allele distribution is more even for DYS390 than for DYS19. The lowest D value was 0.322 (DYS393),

which is to be expected as over 81% of the samples shared allele 13 for this locus. Even though the gene diversity for DYS391 is reasonable ($D = 0.509$) compared to the other MH loci, it appears practically bi-allelic, as 99% of the alleles for this locus are either allele 10 or allele 11.

Table 2.3. Allele and haplotype frequencies for MH loci among Afrikaner Caucasian males ($n=108$)

Allele	DYS19	DYS389-I	DYS389-II ^a	DYS393	DYS392	DYS391	DYS390	Haplotype	DYS385
9						0.009		9/16	0.019
10						0.481		10/14	0.019
11					0.380	0.509		11/12	0.009
12		0.194		0.037	0.102			11/13.2	0.009
13	0.111	0.676		0.815	0.463			11/13	0.019
14	0.611	0.130		0.102	0.056			11/14	0.361
15	0.194		0.028	0.046				11/15	0.130
16	0.046		0.574					11/17	0.009
17	0.037		0.259					12/13	0.009
18			0.120					12/14	0.009
19			0.019					12/15	0.009
20								12/16	0.009
21								13/14	0.083
22							0.102	13/15	0.009
23							0.324	13/16	0.009
24							0.380	13/17	0.009
25							0.185	13/18	0.009
26							0.009	14/14	0.009
								14/15	0.111
								14/16	0.009
								15/15	0.037
								15/16	0.009
								15/17	0.019
								16/16	0.009
								16/18	0.037
								16/19	0.019
								18/18	0.009
Gene Diversity	0.706	0.489	0.588	0.322	0.628	0.509	0.706	Gene Diversity	0.828

^a For DYS389-II, the number of repeats was obtained by subtracting the length of the corresponding DYS389-I allele.

2.3.2.3 Allele and haplotype frequencies for males of the Coloured community

The D value for DYS385 among the males from the Coloured community was 0.896 (Table 2.4). The highest D value for a single-copy-locus was 0.768 (DYS390). The lowest D value was 0.510 (DYS391). Gene diversity values for all MH loci investigated among this sub-population were greater than 0.5.

Table 2.4. Allele and haplotype frequencies for MH loci among males of Coloured community ($n=107$)

Allele	DYS19	DYS389-I	DYS389-II ^a	DYS393	DYS392	DYS391	DYS390	Haplotype	DYS385
9				0.009	0.009	0.037		10/15	0.009
10		0.009		0.009	0.019	0.607		11/13	0.028
11		0.009		0.084	0.402	0.346		11/14	0.290
12		0.243		0.607	0.112	0.009		11/15	0.047
13	0.019	0.542		0.187	0.421			11/16	0.009
14	0.477	0.187		0.103	0.037			12/12	0.009
15	0.336	0.009	0.028					12/14	0.019
16	0.140		0.551					12/15	0.019
17	0.028		0.262					12/18	0.009
18			0.140					12/20	0.009
19			0.019					12/23	0.009
20							0.019	13/14	0.019
21							0.178	13/15	0.019
22							0.093	13/16	0.019
23							0.308	13/17	0.009
24							0.299	13/18	0.009
25							0.075	14/14	0.009
26							0.028	14/15	0.028
								14/16	0.028
								14/17	0.019
								14/18	0.019
								14/19	0.009
								15/15	0.028
								15/16	0.047
								15/17	0.028
								15/19	0.009
								15/20	0.047
								15/21	0.009
								16/16	0.019
								16/17	0.056
								16/18	0.019
								16/19	0.019
								16/20	0.019
								17/17	0.009
								17/18	0.037
								17/20	0.009

Gene Diversity	0.639	0.612	0.607	0.578	0.647	0.510	0.768	Gene Diversity	0.896
----------------	-------	-------	-------	-------	-------	-------	-------	----------------	-------

^a For DYS389-II, the number of repeats was obtained by subtracting the length of the corresponding DYS389-I allele.

2.3.3 Comparison of allele frequencies among five South African sub-populations

Table 2.5 compares the observed gene diversity (D) values of the MH loci among five South African sub-populations, three from this study and two from a previous study (Leat et al. 2004). In four out of five sub-populations studied, the highest gene diversity for a single-copy locus was obtained with DYS390. The highest gene diversity for a single-copy locus in the Xhosa-speaking males was obtained with DYS19. For both Asian Indian and Xhosa populations, lowest gene diversity values were found with DYS391 and DYS392, suggesting they may be of limited use for these two sub-populations. Lowest gene diversity values in the English and Afrikaner Caucasian populations were obtained with DYS393. This suggests that in these two sub-populations, DYS393 may be of limited use. The other MH loci have reasonable gene diversity values, making them useful when combined in a haplotype. In general, the lowest individual gene diversity values for MH loci seem to occur among the English-speaking Caucasian males and the highest individual gene diversity values for MH loci seem to occur among males from the Coloured community.

Table 2.5. Comparison of D values of MH loci among five different South African sub-populations

Population Group	n	DYS19	DYS389-I	DYS389-II	DYS393	DYS392	DYS391	DYS390	DYS385
Afrikaner Caucasian	108	0.706	0.489	0.588	0.322	0.628	0.509	0.706	0.828
Asian	85	0.690	0.620	0.616	0.692	0.479	0.279	0.734	0.935
Coloured Community	107	0.639	0.612	0.607	0.578	0.647	0.510	0.768	0.896
English Caucasian ^a	100	0.460	0.530	0.437	0.325	0.574	0.544	0.662	0.820
Xhosa ^a	99	0.700	0.680	0.683	0.550	0.080	0.130	0.610	0.918

^a Populations investigated by Leat et al. 2004

2.3.4 Haplotype analysis and comparisons of common haplotypes with those recorded in the YHRD

The Y-chromosome Haplotype Reference Database (YHRD) (<http://www.yhrd.org>) has become an invaluable tool for comparing MH data from different regions in the world. Version 15 of this

database is currently available and it contains 28 650 haplotypes. To ensure that genotyping nomenclature in the database is standardized, a quality assurance exercise must first be successfully completed. The database accepts population data, the definition of a population being a representative group of more than 50 individuals living in the same area.

Some historical background to the populations investigated in this study will be provided. For each of the three sub-populations investigated, the most two common observed haplotypes have been compared with those in the YHRD. This was done in an attempt to trace the origin of the paternal lineages of these sub-populations. Some maps (Figure 2.2 to Figure 2.11) have been provided to indicate regional haplotype matches (YHRD). Red and orange dots on the maps indicate haplotype matches. Blue dots indicate populations for which MH data is available, but in which no matches could be found. For easy comparisons, these maps can be found at the end of this chapter (pages 39 to 44).

2.3.4.1 Asian Indians

a) Population History of Asian Indians in South Africa

The Asian Indian community of present-day KwaZulu-Natal can largely trace their ancestry to the labourers who came to South Africa during the late 1800's. In 1843 Natalia became a British colony. The Boers who had lived there moved to the Transvaal, leaving British colonialists as a significant presence. In order to survive and make full use of this seemingly fertile territory, they experimented with the cultivation of tea, wattle and sugarcane plantations. Sugarcane in particular, and to a lesser extent tea, soon proved to be good sources of income. Manual labour was needed for these ventures to succeed. Sir George Grey, Governor of the Cape Colony at the time, was approached about this labour shortage and after negotiations with Indian authorities, it was agreed that labourers would be brought from India (Calpin 1949).

The first indentured Indian labourers arrived in Natal during November of 1860. They were mainly Hindu. In addition, some Muslim and Christian males also arrived from South India and Calcutta. They were under contract for three years, which was eventually extended to five years. They were then given three choices: (1) they could re-indenture themselves for a further period, (2) they could go back to India at the expense of the government, or (3) they could be awarded a piece of 'Crown' land equivalent to the value of their return trip to India. Many chose to stay in

South Africa. By 1866 the import of indentured labourers was stopped, partly because of complaints to the Indian government of ill treatment of its people by the British. As more labourers became free men, the labour shortage once again became a problem and seeking labourers from India again became a necessity (Bagwandeem 1989).

In 1866 import of indentured labourers resumed again, but with a few amendments to their contracts. For every 100 men who came to South Africa, 40 women had to accompany them. The term of service was now five years, but Indians were not allowed to return to their homeland before spending a minimum of 10 years in South Africa. This ensured that reliable labour would be available for other industries such as the coalmines, the railways, brickyards and various other industries. By this time the Indian community in Natal consisted of three types of inhabitants: (1) formerly indentured labourers that had become free men, (2) labourers who were still under contract and (3) 'passenger' Indians (mainly political exiles who returned to India after their term of exile were expired). These individuals came to South Africa at their own expense and enjoyed the same rights as the British in the province. The influx of Asian Indians into Natal caused the size of the community to increase from the 693 that arrived in November of 1860 to 35 763 by 1891 (Calpin 1949; Bagwandeem 1989). While Asian Indians remain a minority, they still represent a well-established part of the population of South Africa. Most of the Asian Indians in KwaZulu-Natal today are descendants of former labourers.

b) Haplotypes from Asian Indians and comparisons with those in the YHRD

For the Asian Indian males of KwaZulu Natal, the MH diversity was 0.984537. Unique haplotypes were found for 70 of the 85 males investigated (82.35%). In light of this favourable discriminatory capacity and the fact that most individual loci in the MH have reasonable gene diversity values, these results suggest that the MH could be useful for identity testing among this particular population.

When searching the YHRD for geographical matches to haplotypes, it does not appear as though complete MH population data from India have been submitted. From literature searches, it also appears as though population data on Asian Indians in India have focused mainly on the use of autosomal STRs (Kashyap et al. 2004) or Y-SNPs (Ramana et al. 2001; Redd et al. 2002b). When searching the YHRD for population matches to haplotypes, some data is available for

metapopulations from Indo-Iranian and Indian descent with which the results here will be compared.

Five Asian Indian haplotypes occurred at least twice (Table 2.6 – page 39). The most common haplotype occurred four times (4.71%). 73 worldwide matches to this haplotype were found in the YHRD, seven of these matching haplotypes from Indian and Malay populations in Asia (Malaysian – 5, Singapore – 1, Kabardinian – 1). However this haplotype matched 64 others from European communities (Figure 2.2 – page 39), of which 40 were haplotype matches to Indo-Iranian metapopulations. Among the 498 Indo-Iranian haplotypes in the YHRD, this haplotype is present at a frequency of 8.03% and is the third most common haplotype among Asians in Malaysia (~1.72%). The second most common haplotype in this sub-group matched six Asian haplotypes in Asia (Malaysia – 5 and Singapore – 1) (Figure 2.3 – page 39). Analysis of this data can be improved if complete MH data for reasonably sized populations from India were available, particularly in the YHRD.

2.3.4.2 Afrikaner Caucasians

a) Population History of Afrikaner Caucasians in South Africa

The arrival of the Dutch in the Cape during the 17th century represented the beginning of the establishment of an Afrikaner community in South Africa. During this time, the ‘Vereenigde Oost-Indische Compagnie’ (VOC) was the world’s biggest trading company. It owned 1755 ships that sailed for trade from the Netherlands to Jakarta between 1602 and 1699. The distance travelled required the building of a refreshment station at the ‘Cape of Good Hope’. For this task the VOC appointed Jan van Riebeeck who arrived at the Cape in 1652. With him were his wife, son and ninety other individuals (Giliomee 2003). His task was to build a fort large enough to house eighty men, who in turn would plant fruit trees, and ensure that arable and pasture land were made available. This fort was to serve as a place where sailors could recover from the long periods at sea and where ships could be restocked with fresh produce (Böeseken 1989).

The diversity of the European population at the Cape changed with the arrival of other European settlers. In 1688 some 180 French Huguenots fleeing religious prosecution also arrived at the Cape. The Dutch ‘burghers’ at the Cape soon became nervous about this outside influence. Simon van der Stel, Cape governor at the time accordingly settled most of these immigrants in the ‘French

Hoek' (Franschoek) and the Drakenstein (Paarl). Instructions were given that the French immigrants be interspersed among Dutch burghers so that they would "learn their language and morals". With no country to go back to, they had to either adapt to this new country or disappear. This strategy was successful for the Dutch – by 1750 few under the age of 40 could speak French (Giliomee 2003). Germans also made their appearance at the Cape during this time. Most of the Germans were single males who spoke diverse dialects. After the arrival of the French at the Cape, most appear to have married French women. Their children however spoke a language referred to as Afrikaans-Dutch (Giliomee 2003).

The term Afrikaner was first used in 1707 to describe an Afrikaans-speaking person of European descent. With a German father, Dutch mother and a mixed race half sister (from his father's relations with a Khoi-Khoi woman), 'Afrikaner' was the name that Hendrik Biebouw (who was born at the Cape) gave himself. The term 'Afrikaner' had previously only been used to refer to offspring of the indigenous inhabitants and slaves of the country (Giliomee 2003). Soon the identity of a new group of people was realised. The common denominator within this group was not so much their origin, but their language – Afrikaans. This language was derived from a mixture between mostly Dutch, with the language of the Khoi-Khoi people, the original inhabitants of the Cape (van Bruwer 1964). Most people who refer to themselves as Afrikaners in South Africa today, have a diverse historical background, but are bound as a unity by their language.

b) Haplotypes from Afrikaner Caucasians and comparisons with those in the YHRD

The YHRD had been useful in comparing haplotypes from the Afrikaner males investigated in this study with those from the rest of particularly Western Europe. The overall haplotype diversity for the MH in the Afrikaner Caucasian sub-population was 0.983539. However, the discriminatory capacity for the MH among this sub-population seems to be limited. Only 59 of the 108 Afrikaner males investigated (54.63%), had unique haplotypes. It could therefore be assumed that the use of the MH for identity testing in this particular sub-population would be of limited use.

Twenty Afrikaner Caucasian haplotypes occurred at least twice (Table 2.7 – page 40). The most common haplotype was found in five individuals (4.63%). A total of 246 matches to this haplotype have been found in regions all over Western Europe (Figure 2.4– page 41).

Interestingly, of the 246 European matches found to this haplotype, 117 (47.56%) are from two populations (German – 105 and Dutch – 12). Among the populations where this haplotype is found, it appears to be twice as frequent in the Dutch group (3.74%) than in the German group (2.03%). A much bigger Dutch sample size is needed to confirm this. The second most common haplotype was found in four individuals (3.70%). A total of 63 matches to this haplotype have been found evenly distributed in regions all over Western Europe (Figure 2.5 – page 41).

The most commonly found haplotypes for the Afrikaner Caucasian sub-population are also commonly found among Dutch and German individuals. The MH data presented here therefore appears to be consistent with the historical background of this sub-population. The results also suggest that the Afrikaner Caucasian sub-population in South Africa is a genetically diverse group.

2.3.4.3 Coloured community

a) Population History of the Coloured community in South Africa

The arrival of the Dutch in the Cape during the 17th century at the same time gave rise to the Coloured community of South Africa. When Jan van Riebeeck arrived at the Cape, he was urged by the VOC to at all times live in peace with the indigenous communities (Giliomee 2003). Soon after the Dutch arrived, they started to farm with cattle. The Khoi-Khoi people of the Cape were nomadic herdsman who also farmed predominantly with cattle. The first contact of the European immigrants with the Khoi-Khoi people of the Cape was as a direct result of cattle trade (van Bruwer 1964).

To provide labour in the Cape Colony, the Dutch officials also decided to import slaves. The slaves were bought from Madagascar, Mozambique, the East African coast as well as Malaysia and India between 1680 and 1731, and the slaves were generally sold to the Dutch ‘burghers’ (Giliomee 2003). During the first 30 years after the Cape was founded, the European immigrants in the region were single males. For the 36 years before the French arrived in South Africa few other women were present in the Cape other than the Khoi-Khoi women. At first, relationships with the Khoi-Khoi women were encouraged to boost the numbers of the ‘European community’ at the Cape. Contact between European immigrants, the original Khoi-Khoi people of the Cape

and slaves brought in from Malaysia and Africa ultimately lead to the establishment of the diverse community currently identified as the Coloured population (van Bruwer 1964).

Among the Coloured community in South Africa, distinct sub-groups have developed. These were the Cape Coloured, Griquas, Malay, Oorlams and Koranas. These groups are culturally and geographically distinct from one another. The Oorlams no longer exist. The Koranas have moved up the west coast of South Africa and now reside in Namibia where they have been incorporated into the local Nama people (van Bruwer 1964). The Malay have remained a distinct entity within the Coloured community, partly because of their religion – Malay people are generally Muslim, whereas other Coloured people are generally Christian – although these divisions are less than static. The Griquas live mostly in the northern Cape. The Cape Coloured people are the most complex of all these groups as their origins are so varied. In a blood group study done by Botha et al. (1972), it was found that the Cape Coloured community possessed a blood group pattern containing 36% Western European, 34% South African Black and 30% Asian genes.

b) Haplotypes from Coloured community and comparisons with those in the YHRD

The overall haplotype diversity for the MH loci among males of the Coloured community was 0.985414, the highest among the three sub-populations investigated. Unique haplotypes were found for 83 of the 107 (77.57%) males investigated. Eight haplotypes occurred at least twice (Table 2.8 – page 42). When comparing the non-unique haplotypes with those in the YHRD, all matched other haplotypes in the database.

The two most common haplotypes from the Coloured community were found in five individuals each (4.67%). The first of these two haplotypes (Table 2.8 – page 42) is the same as the most common haplotype in the Afrikaner group (Table 2.7 – page 40), and 246 European matches to this haplotype were found in the YHRD. This suggests that some of the male lineage of this admixed population may be the same as the Afrikaner community of South Africa, which would be consistent with the historical origin of this group.

The other most common haplotype matched 606 other haplotypes in the YHRD of which 478 were European. Of these 478 European matches found to this haplotype, 55.44% are from three populations (German – 135, Spanish – 118 and Dutch – 12). Among the populations in which this haplotype is observed, it is much more frequent in the Spanish group (8.53%) than in the Dutch

group (5.12%) and the German group (2.94%). This again suggests that some of the male lineage of this admixed population may be the same as the Afrikaner community of South Africa. It also introduces another addition to the male lineage of this group – the Spanish. From the history of South Africa it has been known that Portuguese have also sailed to the Cape even before the Dutch (de Kock 1989). When one considers how the maps of countries have changed over the years, it may be reasonable to expect population admixture between Spain and Portugal.

The fifth most common haplotype in the Coloured population (Table 2.8 – page 42) is also the most common haplotype among the Xhosa-speaking Black population in the Western Cape (Table 2.10 – page 44) (Leat et al. 2004). This suggests contributions from both European and indigenous African paternal lineages to the Coloured community.

2.3.5 Summary

When the most common haplotypes for all the sub-populations groups studied recently are considered, a few things become apparent.

- 1.** The two most common haplotypes in the Asian Indian community are found in South East Asia. However, the most commonly found haplotype in Asian Indians of South Africa is also commonly found in Europe among Indo-Iranian metapopulations. To establish whether this haplotype is common to India or South Asia, and to determine the frequency of these haplotypes in this region, it would be helpful if complete MH data were available for Asian Indians in India.
- 2.** The two most common haplotypes in the Xhosa community are only found in Southern Africa. This would suggest that this haplotype might be unique to the male lineages of this region. It could also suggest that the Xhosa community in Southern Africa might be a group with a distinct population history.
- 3.** When comparing the most common haplotypes of the other three sub-populations investigated, their paternal lineage seems intertwined. The most common haplotype in the Afrikaner males is also one of the two most common haplotypes in the Coloured community. The other most common haplotype of the Coloured community is also one of the third most common haplotypes in the Afrikaner community, which is in turn the most common haplotype shared by the English Caucasian males in the Western Cape. Matches to these haplotypes in the YHRD have been found in German, Dutch and also Spanish populations. This points not only to the

genetically diverse male lineage of the Coloured community, but also to the genetically diverse male lineage of the Afrikaner and English Caucasian communities of the Western Cape.

When analyzing the D values of individual MH loci (Table 2.5), it can be observed that some of these loci show low variability among the South African sub-populations investigated. For the Xhosa and Asian Indian communities, DYS391 and DYS392 may be of limited use in forensic studies. For both the English and Afrikaner Caucasian communities, DYS393 may be of limited use in forensic studies. Despite the advantage of increased discriminatory capacity, duplicated loci generate complex profiles when used in the analysis of mixtures. They are therefore not ideal for forensic studies. For these reasons it was decided to investigate other loci for properties that would be useful in forensic studies.



Table 2.6. Haplotypes shared by more than one Asian Indian male ($n=85$)

Haplotype	n	Frequency
DYS19, <i>DYS389-I</i> , <i>DYS389-II</i> ^a , <i>DYS393</i> , <i>DYS392</i> , <i>DYS391</i> , <i>DYS390</i> , <i>DYS385</i> ^b		
15-14-16-12-11-10-22-(15/17)	4	0.047
14-12-16-11-14-10-22-(14/17)	3	0.035
15-12-17-13-11-11-25-(13/17)	2	0.024
17-13-18-13-11-11-25-(12/14)	2	0.024
14-13-16-14-11-10-23-(14/17)	2	0.024
15-12-16-12-11-10-22-(15/17)	2	0.024

^a For *DYS389-II*, the number of repeats was obtained by subtracting the length of the corresponding *DYS389-I* allele.

^b Haplotypes for the duplicated locus *DYS385* are presented in parenthesis



Figure 2.2. Worldwide distribution of the most common ‘minimal haplotype’ among Asian Indian males from KwaZulu-Natal (Red and orange dots indicate haplotype matches, blue dots indicate populations for which MH data is available, but in which no matches could be found.)

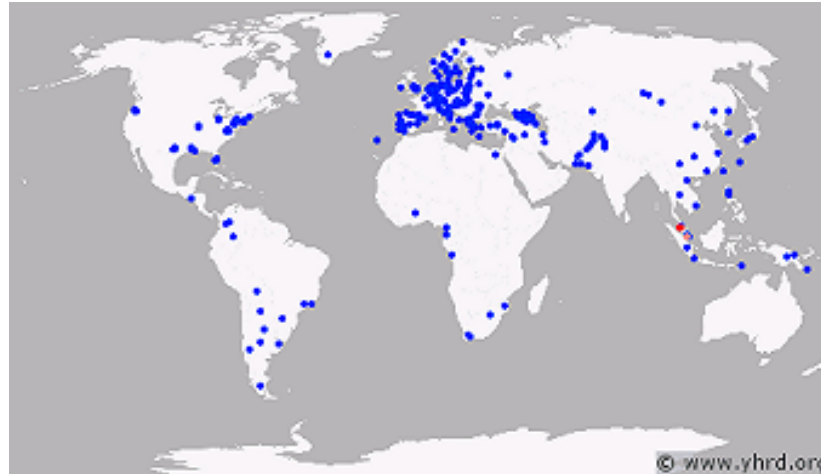


Figure 2.3. Worldwide distribution of the 2nd most common ‘minimal haplotype’ among Asian Indian males from KwaZulu-Natal (Red and orange dots indicate haplotype matches, blue dots indicate populations for which MH data is available, but in which no matches could be found.)

Table 2.7. Haplotypes shared by more than one Afrikaner Caucasian male ($n=108$)

Haplotype	n	Frequency
DYS19, DYS389-I, DYS389-II ^a , DYS393, DYS392, DYS391, DYS390, DYS385 ^b		
14-13-16-13-13-11-23-(11/14)	5	0.046
14-13-16-13-13-10-24-(11/15)	4	0.037
13-12-16-13-11-11-24-(14/15)	3	0.028
14-13-16-13-13-11-24-(11/15)	3	0.028
14-13-16-13-13-11-24-(11/14)	3	0.028
15-13-18-13-11-11-25-(11/14)	3	0.028
13-13-16-13-13-10-23-(11/14)	2	0.019
14-12-16-13-11-10-22-(13/14)	2	0.019
14-12-16-13-11-10-22-(14/15)	2	0.019
14-12-17-13-11-10-22-(13/14)	2	0.019
14-13-16-13-13-11-23-(11/15)	2	0.019
14-13-16-13-14-11-25-(11/13)	2	0.019
14-13-17-13-11-10-25-(16/18)	2	0.019
14-13-17-13-13-11-23-(11/14)	2	0.019
14-13-18-14-12-10-23-(15/15)	2	0.019
14-14-16-13-13-11-24-(9/16)	2	0.019
15-13-17-13-11-11-25-(11/14)	2	0.019
15-13-16-13-13-11-23-(11/14)	2	0.019
15-14-18-14-12-10-23-(14/15)	2	0.019
16-13-19-13-11-10-25-(11/14)	2	0.019

^a For DYS389-II, the number of repeats was obtained by subtracting the length of the corresponding DYS389-I allele.

^b Haplotypes for the duplicated locus DYS385 are presented in parenthesis

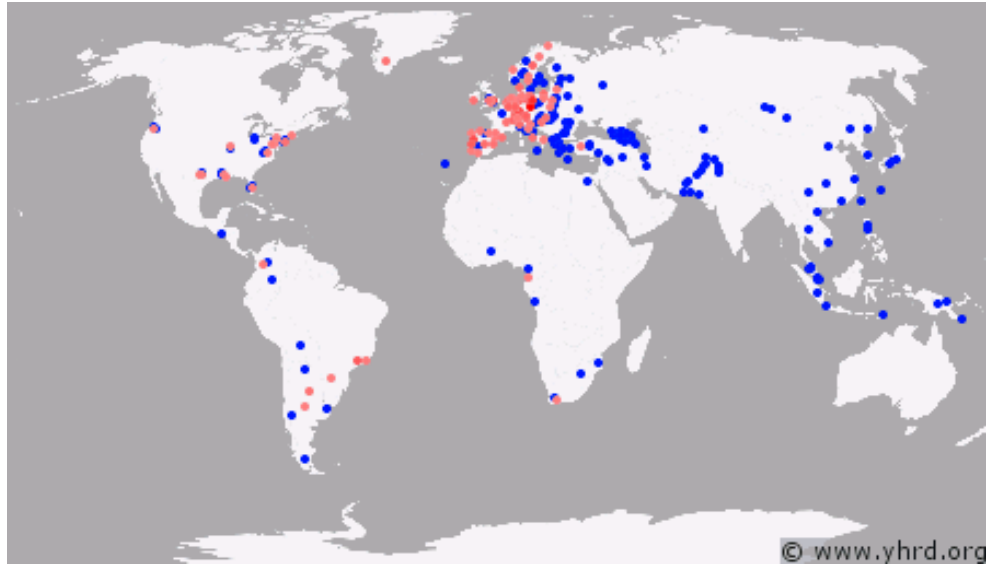


Figure 2.4. Worldwide distribution of the most common ‘minimal haplotype’ among Afrikaner Caucasian males from the Western Cape (Red and orange dots indicate haplotype matches, blue dots indicate populations for which MH data is available, but in which no matches could be found.)

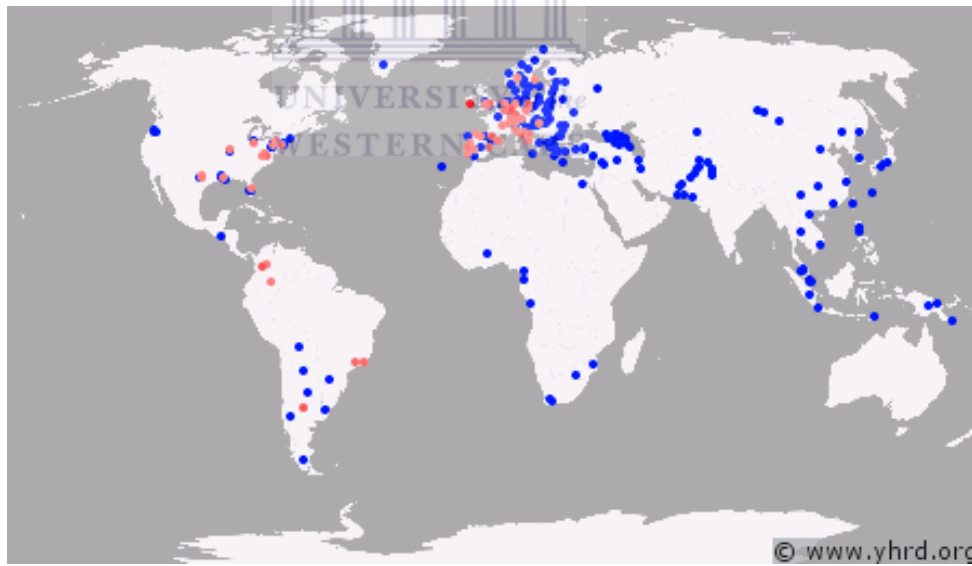


Figure 2.5. Worldwide distribution of the 2nd most common ‘minimal haplotype’ among Afrikaner Caucasian males from the Western Cape (Red and orange dots indicate haplotype matches, blue dots indicate populations for which MH data is available, but in which no matches could be found.)

Table 2.8. Haplotypes shared by more than one Coloured community male ($n=107$)

Haplotype	n	Frequency
DYS19, <u>DYS389-I</u> , <u>DYS389-II</u> ^a , <u>DYS393</u> , <u>DYS392</u> , <u>DYS391</u> , <u>DYS390</u> , <u>DYS385</u> ^b		
14-13-16-13-13-11-23-(11/14)	5	0.047
14-13-16-13-13-11-24-(11/14)	5	0.047
14-13-16-13-13-11-23-(11/15)	3	0.028
14-13-16-13-13-10-23-(11/14)	3	0.028
14-12-16-13-11-10-26-(15/20)	2	0.019
14-13-16-13-13-11-24-(11/13)	2	0.019
14-13-17-13-13-11-24-(11/14)	2	0.019
15-14-18-13-11-10-21-(15/16)	2	0.019

^a For DYS389-II, the number of repeats was obtained by subtracting the length of the corresponding DYS389-I allele.

^b Haplotypes for the duplicated locus DYS385 are presented in parenthesis

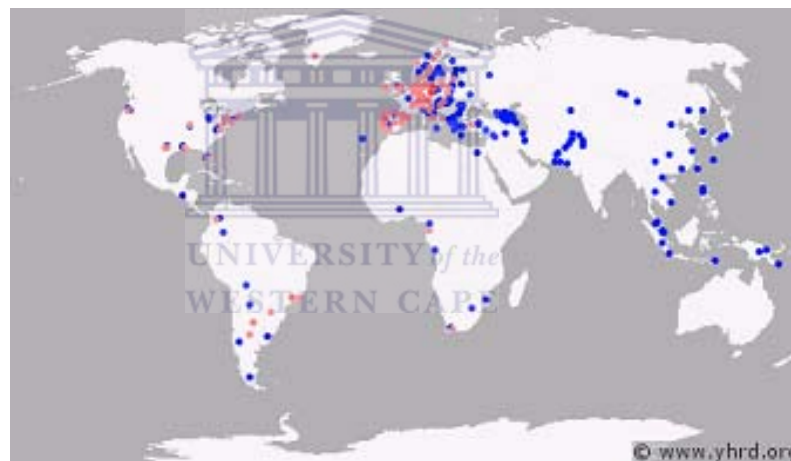


Figure 2.6. Worldwide distribution of the most common ‘minimal haplotype’ among Coloured males from the Western Cape (Red and orange dots indicate haplotype matches, blue dots indicate populations for which MH data is available, but in which no matches could be found.)

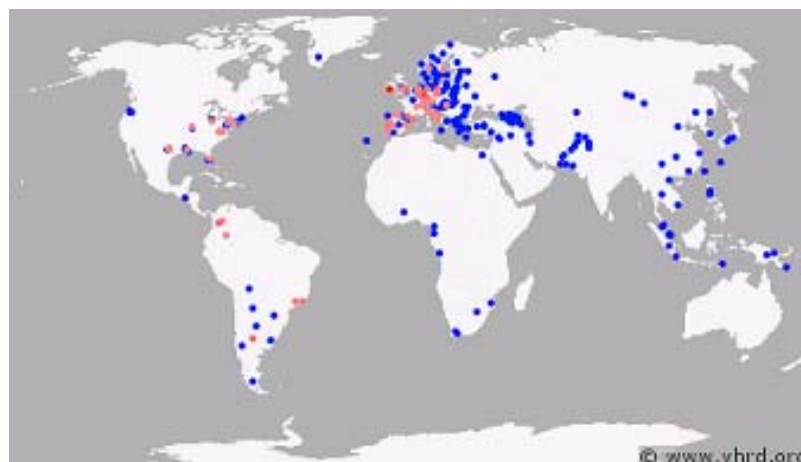


Figure 2.7. Worldwide distribution of the 2nd most common ‘minimal haplotype’ among Coloured males from the Western Cape (Red and orange dots indicate haplotype matches, blue dots indicate populations for which MH data is available, but in which no matches could be found.)

Table 2.9. Haplotypes shared by more than one English male ($n=100$) (Leat et al. 2004)

Haplotype	n	Frequency
DYS19, <i>DYS389-I</i> , <i>DYS389-II</i> ^a , <i>DYS393</i> , <i>DYS392</i> , <i>DYS391</i> , <i>DYS390</i> , <i>DYS385</i> ^b		
14-13-16-13-13-11-24-(11/14)	8	0.08
14-13-16-13-13-10-24-(11/14)	6	0.06
14-12-16-14-11-10-23-(14/15)	2	0.02
14-12-16-13-11-10-23-(13/14)	2	0.02
14-13-16-13-13-11-24-(12/14)	2	0.02
14-13-16-13-13-11-23-(11/15)	2	0.02
14-13-16-12-13-11-24-(11/14)	2	0.02
14-12-16-13-11-10-22-(13/14)	2	0.02
14-13-16-13-13-10-24-(11/11)	2	0.02
14-14-16-13-13-11-24-(11/14)	2	0.02
14-12-16-13-11-10-23-(14/15)	2	0.02
15-13-17-13-13-11-24-(11/14)	2	0.02

^a For *DYS389-II*, the number of repeats was obtained by subtracting the length of the corresponding *DYS389-I* allele.

^b Haplotypes for the duplicated locus *DYS385* are presented in parenthesis

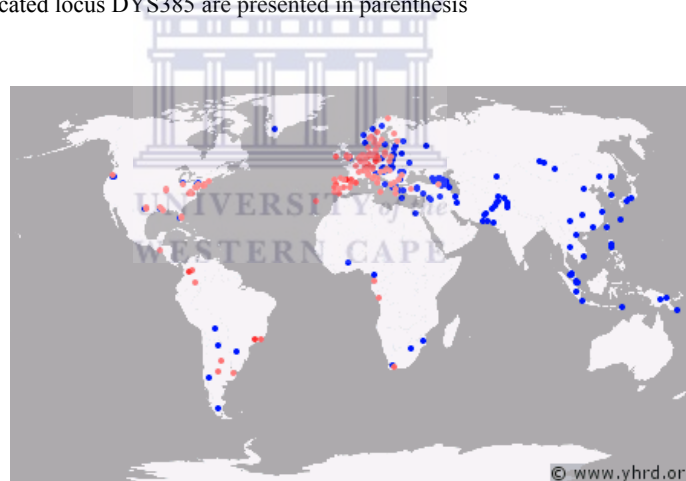


Figure 2.8. Worldwide distribution of the most common ‘minimal haplotype’ among English Caucasian males from the Western Cape (Leat et al. 2004) (Red and orange dots indicate haplotype matches, blue dots indicate populations for which MH data is available, but in which no matches could be found.)

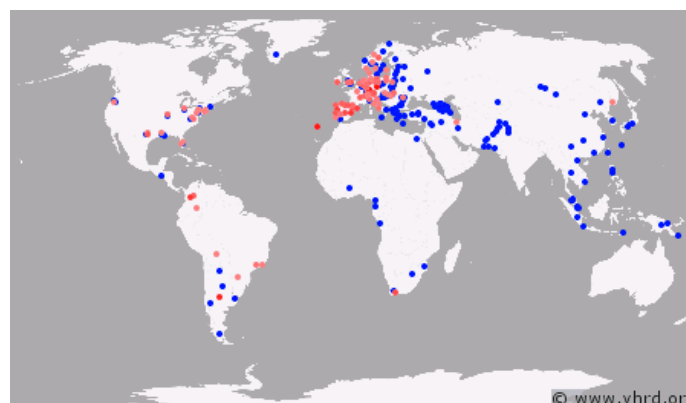


Figure 2.9. Worldwide distribution of the 2nd most common ‘minimal haplotype’ among English Caucasian males from the Western Cape (Leat et al. 2004) (Red and orange dots indicate haplotype matches, blue dots indicate populations for which MH data is available, but in which no matches could be found.)

Table 2.10. Haplotypes shared by more than one Xhosa male ($n=99$) (Leat et al. 2004)

Haplotype	n	Frequency
DYS19, <i>DYS389-I</i> , <i>DYS389-II</i> ^a , <i>DYS393</i> , <i>DYS392</i> , <i>DYS391</i> , <i>DYS390</i> , <i>DYS385</i> ^b		
14-12-16-13-11-10-26-(15/20)	13	0.131
15-14-16-13-11-10-21-(15/17)	7	0.071
16-13-18-13-11-10-21-(16/17)	4	0.04
16-14-17-15-11-10-21-(17/21)	3	0.03
16-14-17-15-11-10-21-(16/20)	3	0.03
14-12-16-13-11-10-25-(15/20)	3	0.03
16-14-17-15-11-10-21-(17/20)	3	0.03
15-14-18-13-11-10-24-(11/11)	3	0.03
15-14-18-13-11-10-21-(15/16)	3	0.03
17-13-17-14-11-10-21-(16/18)	2	0.02
15-13-16-14-11-10-21-(16/16)	2	0.02
16-13-18-13-11-10-21-(15/18)	2	0.02
16-13-17-15-11-10-22-(17/17)	2	0.02
15-13-16-12-11-10-22-(15/18)	2	0.02

^a For *DYS389-II*, the number of repeats was obtained by subtracting the length of the corresponding *DYS389-I* allele.

^b Haplotypes for the duplicated locus *DYS385* are presented in parenthesis

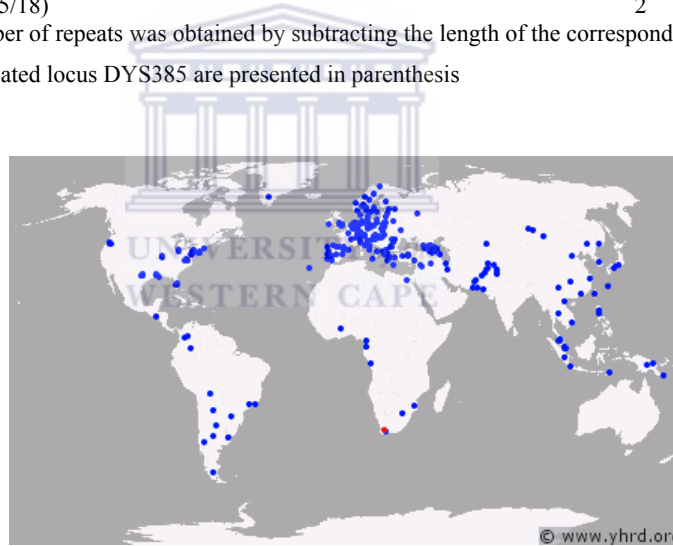


Figure 2.3.10. Worldwide distribution of the most common ‘minimal haplotype’ among Black Xhosa males from the Western Cape (Leat et al. 2004) (Red and orange dots indicate haplotype matches, blue dots indicate populations for which MH data is available, but in which no matches could be found.)

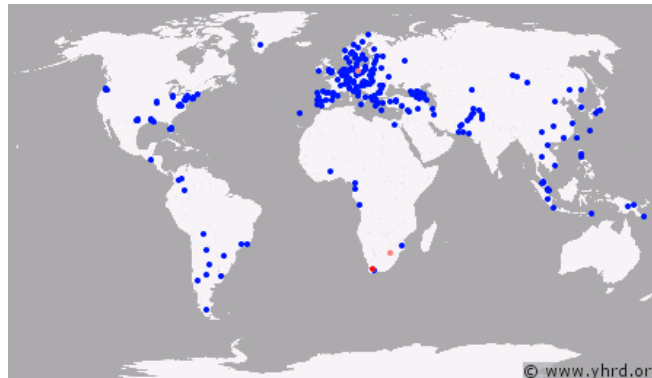


Figure 2.11. Worldwide distribution of 2nd the most common ‘minimal haplotype’ among Black Xhosa males from the Western Cape (Leat et al. 2004) (Red and orange dots indicate haplotype matches, blue dots indicate populations for which MH data is available, but in which no matches could be found.)

Chapter 3 – Selecting STR loci with which to investigate South African populations

3.1 Introduction

The loci of the ‘minimal haplotype’ (MH) have been well characterized in terms of allele nomenclature, gene diversity, population data and mutation rates. An extensive database, the Y-chromosome Haplotype Reference Database or YHRD (<http://www.yhrd.org>), has also been established with MH data from diverse populations. More recently, the Scientific Working Group on DNA Analysis Methods (SWGAM) has recommended the inclusion of DYS438 and DYS439 to the core MH. Reports have shown that the addition of these two loci has increased the discriminatory capacity of the MH (Grignani et al. 2000; Gusmão et al. 2002). These two additional loci have been included in both the Yfiler™ kit (*Applied Biosystems*) as well as the PowerPlex Y kit (*Promega*). In addition, the Powerplex Y kit includes DYS437, while the Yfiler™ kit includes DYS437, DYS448, DYS456, DYS458, DYS635 (Y-GATA C4) and Y-GATA H4.

The objective of the present study was to select loci that could compliment those already in use and possibly replace those with limited variability among South African populations. This was prompted by an initial study revealing extremely low levels of polymorphism for two MH loci (DYS391 and DYS392) in Xhosa populations in Cape Town, South Africa (Leat et al. 2004)(see Chapter 2). Of the 99 individuals typed in the study, only 47 unique haplotypes were observed and 13 individuals shared the most common haplotype. DYS391, DYS392 and DYS437 were also found to be virtually mono-morphic in a population surveyed in Maputo, Mozambique (Álves et al. 2003)(see Chapter 2).

A subset of recently identified loci could be selected to compliment existing Y-STR typing systems. Recent surveys of Y-chromosome sequence data have resulted in the identification of at least 191 polymorphic Y-STR loci (Ayub et al. 2000; Iida et al. 2001; Iida et al. 2002; Redd et al. 2002a; Kayser et al. 2004). Several factors should be considered in the selection of Y-STRs.

These include variability, the ease with which male specific PCR primers can be designed, the extent to which stutter artifacts are generated during PCR, and copy number. Multi-copy loci generate complex profiles using a single primer set. While this has the advantage of providing a high discriminatory capacity for a given primer set, clear locus-allele relationships cannot be established and the analyses of mixtures are therefore complicated. The present study aimed to assess the properties of polymorphic, single-copy loci for which male specific primers could be designed and which generated limited stutter artifacts during PCR.

At the time of initiating the study, approximately 49 loci had been well characterized (Redd et al. 2002a). Primer sequences for a substantial number of novel Y-STR loci had also been submitted to the Genome Database (GDB) (<http://www.gdb.org>) and are the subject of a recent publication (Kayser et al. 2004). In considering which loci should be selected for assessment, the most obvious approach was to start with the loci of the minimal haplotype that could serve as a point of reference (see Chapter 2). Additional loci could then be selected on the basis of gene diversity values reported for several populations. This approach alone would have failed to consider the large number of STR loci recently identified on the Y-chromosome. Therefore an attempt was also made to select STR loci from Y-chromosome sequence data.

A comprehensive survey of most Y-STRs was considered impractical given the available resources, so an attempt was made to increase the probability of selecting polymorphic loci. STR loci have a good chance of being variable provided that a consecutive homologous stretch with greater than eight repeated units are present (Moxon and Wills 1999). Therefore one of the most obvious characteristics on which a selection could be made was the number of repeated elements at each locus. It was hypothesized that selecting loci with long stretches of repeated elements would at least increase the chances of selecting more variable STRs. Once a subset of loci had been selected they could be compared with entries in the GDB to establish if they were novel. Using this approach a set of widely used loci could be assessed, while also testing a subset recently identified from Y-chromosome sequence data.

3.2 Methods and Materials

3.2.1 Selection of STR loci from Y-chromosome sequence data

Tandem Repeat Finder software (v2.02) was used to identify short tandemly repeated elements from approximately 23Mb of euchromatic Y-chromosome sequence (Benson 1999; <http://c3.biomath.mssm.edu/trf.html>). Four contigs (NT_011875.7, NT_011878.7, NT_011896.7, NT_011903.3) were screened. The following *Tandem Repeat Finder* parameters were used (Alignment match: 2, Alignment mismatch: 7, Alignment indel: 7, Minimum alignment score to report: 30, Maximum period size: 5). Simple programs were written in C++ to process the data further. Sequences with 80% to 90% homology between subunits were included if 20 or more repeats were present. Sequences with greater than 90% homology between subunits were included if seven or more repeats were present. For each STR locus, 400bp of sequence was recovered on either side of the repeated elements. The data were sorted into three files for trinucleotide-, tetranucleotide- and pentanucleotide loci respectively.

Two ranking systems were implemented. The first system used two parameters generated by *Tandem Repeats Finder* (percent matches between repeated elements and the number of repeated elements present). The second ranked sequences according to number of perfectly repeated elements in a maximum of two blocks. In both systems, loci were ranked from those with the highest number of repeat units to those with the lowest. An attempt was made to select single-copy loci with high rankings on either system.

An *in silico* indication of copy number was obtained by comparing the sequences within each repeat class with one another using the batch processing facility of the stand-alone version of BLAST (<http://www.ncbi.nlm.nih.gov>). Multi-copy loci were assumed to be those that generated equivalent high BLAST scores against themselves and one or more other loci. Single-copy loci were assumed to be those that generated a high BLAST score only against themselves. This approach only served to increase the probability of selecting single-copy loci, since it depended on correct contig assembly. It should also be noted that PCR products from multi-copy loci could be distinguished by careful primer design focusing on minor sequence differences.

3.2.2 Primer Design

In order to design primers that would flank the repeated elements of the recently identified loci, the *Primer3* software was used (Rozen and Skaletsky 2000; http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi). Primers were chosen to be between 25bp and 30bp in length, and have a melting temperature between 60°C and 64°C. The difference in melting temperatures

between the two primers was specified to be no more than 4°C. Also, self-complementarity was specified to be less than 8bp and 3' complementarity between primers was specified to be less than 3bp. Expected size range of the PCR fragments were generally between 100bp and 400bp. It was important to try and establish whether the chosen primer sequences would be unique to the Y-chromosome, and whether they would amplify more than one fragment on the Y-chromosome. BLAST searches were used to compare the primer sequences with available human genome sequence. Only those primers that appeared to be unique to a single locus on the Y-chromosome were selected for further use. The selected primer sequences as produced by *Primer3*, and expected sizes as observed with contigs NT_011875.7, NT_011878.7, NT_011896.7, NT_011903.3 are presented in Table 3.1.

Table 3.1. Primer sequences of recently identified loci using *Primer3* software

Locus	Forward Primer	Reverse Primer	Estimated Fragment Size ^a
DYS481	5' aaa agg aat gtg gct aac gct gtt c 3'	5' aat tca ttg cag att ctt ggt cca c 3'	205bp
DYS485	5' gag atg aga aat cac tca ttg gac ac 3'	5' gca gtg agc aagh act gtg c 3'	242bp
DYS612	5' aag ttt cac aca ggt tca gag gtt g 3'	5' gct gcc ttt tgt tca ttc ttg tg 3'	239bp
DYS711	5' ctg etc aga gag gtg gtt att aca 3'	5' tgc tgt cat tgt atc tet tca etc c 3'	387bp
DYS614	5' gaa tat ggc taa ctc gat tca gag 3'	5' cca cca aaa ggt ttt cag act ca 3'	339bp
DYS570	5' tga cta ggt aga aat cct ggc tgt g 3'	5' tca gca tag tca aga aac cag aca ac 3'	179bp
DYS712	5' caa gaa cag cct ggg taa cag tg 3'	5' tta tat ggt aca gcc cat gaa cac tt 3'	170bp
DYS514	5' ttt etc taa cac aca cct gtc cat tt 3'	5' cac caa ttg cct atg tga aga gta tg 3'	250bp
DYS715	5' atg gtt gga aga aag cat tga tga 3'	5' cta ggt aat tag cta cct agt tag 3'	196bp
DYS557	5' gcc cag cat gtg ttt tga cta ttt 3'	5' ggt gtt tga ccc agt gat atg ttc t 3'	241bp
DYS710	5' gag gtc aag gct gca aga atc tat ga 3'	5' cat act etc tcc etc cct etc ttt ttt c 3'	225bp
DYS607	5' agc ata cag cgt aat cac agc tca c 3'	5' etc aaa acc cat aat tca ggt cca g 3'	234bp
DYS518	5' ggc aac aca agt gaa act get tet 3'	5' get ctt acc atg ggt gat ttc ttt c 3'	276bp
DYS713	5' gtg caa gcc aag ggc ttt ata agt 3'	5' cct ggg tga cag act cca tet taa a 3'	320bp
DYS626	5' gac aga gtg caa gac ccc ata g 3'	5' tgt cga cat taa gaa gaa ttt tgg 3'	271bp
DYS521	5' aag atg atc atg aac tcc tgg gat caa 3'	5' atg gac act agc cat gac atg ttg ct 3'	331bp
DYS644	5' aga tgc tga ctt cgg ggt agt tc 3'	5' gag gca gaa agt caa gga aat caa 3'	228bp
DYS714	5' aaa ttt acc cac ctt ctg cat cg 3'	5' ttt tct acc tat gat gcc ctt tg 3'	250bp
DYS717	5' ggc cga gag aat gga att gat 3'	5' ccc gaa ctt cag cac tat gaa atg 3'	250bp
DYS716	5' taa atc aga att cct ttc caa tcc a 3'	5' tct ggg ttt cag agt ggg ata att t 3'	224bp

^a From sequence on contigs NT_011875.7, NT_011878.7, NT_011896.7, and NT_011903.3

The e-PCR facility of GDB was initially used to determine if a sequence selected for analysis had been assigned a GDB designation. Since recent updates have not been conducted on the e-PCR

facility, an additional test for novelty was used. A BLAST database was compiled from a file containing sequences for the following primer sets: (1) primers designed during the present study for the loci selected from Y-chromosome sequence data and (2) primer sequences provided by GDB for all Y chromosome amplimers submitted to the database before the 25th February 2005. The stand-alone version of BLAST was used to compare the loci selected from Y-chromosome sequence data against the primer database.

3.2.3 Male specificity of primer sets

Primers were tested to gain an indication of copy-number and male-specificity of the loci they amplified. For this purpose, genomic DNA from two males and one female was PCR-amplified. Primers were synthesized by *MWG-Biotech*, *Inqaba Biotech* or *Applied Biosystems*. Primer sequences have been presented in Table 3.1. Amplifications were performed in a final volume of 25µl containing 8ng genomic DNA, 0.5µM of the forward primer, 0.5µM of the reverse primer, 10mM Tris-HCl (pH8.3), 50mM KCl, 1.5mM MgCl₂, 0.2mM dNTPs, and 0.25µl of Taq polymerase. PCR reactions were performed in a GeneAmp 2400 thermocycler (*Applied Biosystems*) as follows: 1 cycle at 94°C for 2 minutes, 30 cycles of 94°C for 30 seconds, 55°C, 58°C, or 60°C for 30 seconds, 72°C for 30 seconds, followed by 1 final cycle at 72°C for 10 minutes. Fragment analysis was achieved by means of standard agarose gel electrophoresis and ethidium bromide staining.

3.2.4 Variability of Y-STR loci and allele size ranges in a sample of 46 English Caucasian males

An initial indication of variation and relative allele size ranges was obtained for 20 loci selected from Y-chromosome sequence. This was achieved by analyzing DNA from 46 English Caucasian males. It was expected that some of the loci investigated in this manner would have low gene diversity. Unlabelled primers and PAGE fragment analysis were therefore used in this part of the study to avoid the cost of labelling primers that would not be used for further investigation. PCR amplification with unlabeled PCR primers in uniplex reactions was performed as for the male specificity test (Section 3.2.3). Fragment analysis was achieved by means of PAGE analysis and silver staining as adapted from Sambrook et al. (1989). Full details of reagents and procedures are provided in Appendix. In order to determine the relative sizes of the amplicons from each locus the largest and smallest alleles in each case were amplified and run on a single PAGE gel. This

data would later be useful to plan the multiplex PCR primer sets so as to reduce the probability generating amplicons from different loci with overlapping size ranges.

3.2.5 Statistical Analysis

Gene Diversity (D) was calculated as $(1 - \sum P_i^2)$, where P_i is the allele frequency (Nei 1987). P_i for each of the alleles was determined by simple allele counting.

3.2.6 Selection of Y-STR loci based on previously reported population data

In order to select loci from previously published population data, gene diversity values were tabulated for 30 previously studied Y-STR loci not included in the minimal haplotype (Hou et al. 2001; Mohyuddin et al. 2001; Alvarez et al. 2002; Bosch et al. 2002; Gusmão et al. 2002; Iida et al. 2002; Redd et al. 2002a; Alves et al. 2003; Gusmão et al. 2003; Lee et al. 2003; Zhu et al. 2003). Loci with gene diversity values consistently above 0.6 were selected.

3.3 Results and Discussion

3.3.1 Selection of STR loci from Y-chromosome sequence data

Y-chromosome sequence data was screened for STR sequences. This yielded 232 trinucleotide-, 437 tetranucleotide-, and 118 pentanucleotide STR sequences. An attempt was made to prioritize single-copy loci with high numbers of repeated elements. 20 apparently single-copy Y-STR loci with high rankings on either of the ranking systems were selected for further analysis (indicated in blue in Table 3.2). The selection included 17 loci with complex repeat structures and three with simple repeat structures (Table 3.2). Among these were five trinucleotide-, 11 tetranucleotide- and four pentanucleotide loci. One locus, DYS710, which contained a short tract of dinucleotide repeats flanked by tetranucleotide repeats was included (Table 3.2.). It was selected as part of the present study while the duplicated dinucleotide locus YCA II was still recommended as part of the extended haplotype. It was hoped that the short dinucleotide repeat tract would lead to increased polymorphism with a fairly modest stutter.

Two approaches were used to establish if the loci identified from Y-chromosome sequence data were novel. The first used the GDB e-PCR facility, while the second used a BLAST search against primer sequences for all Y-chromosome amplimers submitted to GDB before the 25th February 2005. In every case primers designed during the present study matched the appropriate locus during the BLAST searches. In addition, eleven loci matched primers for previously reported loci. Eight loci did not match previously described primer sets and appear to be novel (Bold and blue in Table 3.2). The BLAST and GDB e-PCR searches yielded identical results.

Table 3.2. Locus Repeat Structures

Locus	Repeated elements
DYS711 ^a	(ctt)₄(cttt)₁(ctf)₃₀(ctc)₃(ctf)₁(ctc)₁(ctf)₃(ctcctt)₄(ctccta)₁(ctf)₂₅(ctc)₂(ctf)₃
DYS612 ^b	(cct) ₅ (ctt) ₁ (tct) ₄ (cct) ₁ (tct) ₂₅₋₂₉
DYS710 ^a	(aaag)₁₇(ag)₁₃(aaag)₁₁
DYS481 ^b	(ctt) ₂₂
DYS614 ^b	(ctt) ₂ (cct) ₃ (ctt) ₃ (cttcttt) ₂ (ctt) ₂ (ctg) ₁ (ctt) ₂ (ctgctt) ₂ (ctgct) ₁ (ctt) ₁₈
DYS449 ^c	(tttc)_nn₅₀(tttc)_n
DYS518 ^b	(aaag) ₃ (gaag) ₁ (aaag) ₁₆ (ggag) ₁ (aaag) ₄
DYS392 ^b	(aat) ₁₁₋₁₃ (aaat) ₂
DYS458 ^c	(gaaa)_n
DYS712 ^a	(agat)₁₇(agac)₅
DYS570 ^b	(ttc) ₁₄₋₂₁
DYS713 ^a	(ctf)₁₈tc(tctf)₂tctgtctttttcttctct(fctf)₁₂(ttct)₅tcc(ttct)₂(ctttf)₂tfta(ttat)₄
Y-GATA-A10	(tttc)₂(tct)₁(ccat)₂(atct)₁₂
DYS607 ^a	(gaag) ₁₅ (gaaagaag) ₂ gatg(gaag) ₂
DYS557 ^b	(ttc) ₄ (ttct) ₁ (tttc) ₄ (ttc) ₁ (tttc) ₁₄₋₂₃
DYS390 ^b	(agat) ₃ (gata) ₄ (gaca) ₁ (gata) ₈₋₁₂ (gaca) ₈ (gata) ₂
DYS635 ^b	(tcta)₄(tgta)₂(tcta)₂(tgta)₂(tcta)₁₁
DYS393 ^b	(agat) ₁₁₋₁₃
DYS391 ^b	(tctg) ₃ (tcta) ₉₋₁₁
DYS626 ^b	(gaaa) ₁₆₋₂₇ (ggaa) ₅₋₆
DYS439 ^b	(gata)₁₀₋₁₃
DYS446 ^c	(tctct)_n
DYS19 ^b	(tacc) ₉₋₁₃ (tacc) ₁ (tacc) ₃
DYS452 ^c	(tatac)₂(tgtac)₂(tatac)_n(catac)₁(tatac)₁(catac)₁(tatac)₃(catac)₂ (tatac)₃(catac)₁(tatac)₃
DYS463 ^c	(aaagg)_n(aagg)_n(aagga)₂
DYS644 ^b	(tttta) _{11-...} ttta (tttta) _{11-...}
DYS714 ^a	(tttct)₁₉(cttct)₂(tttct)₂(cttct)₂(tttct)₂(t)₁₁
DYS715 ^a	(ggat)₂aggta(gata)₁₆aagatgatagatgtgat(ggat)₉
DYS485 ^b	(tta) ₁₂₋₁₆
DYS716 ^a	(tcac)₅(tccat)₁₁cctattctattgaactccatt(ccact)₂
DYS521 ^b	(ctt) ₄ (cttt) ₁ (ctt) ₇₋₁₂ (ctct) ₂ (cttt) ₂
DYS717 ^a	(tgtat)₂(tattg)₂(factg)₆(tattg)₁₀
DYS514	(gaaa) ₁₅ (ggaa) ₃ (gaaa) ₂

^a Repeat elements observed for sequence on contigs NT_011875.7, NT011878.7, NT_011896.7, NT_011903.3.

^b Repeated elements as reported in the online-only tab-delimited data set associated with Kayser et al. (2004)

^c Repeated elements as reported by Redd et al. (2002a)

The objective of the present study was not specifically to identify novel STRs but rather to assess the properties of loci that have been widely studied as well as those recently identified from Y-chromosome sequence data. Despite this, it appears that the approach used resulted in the identification of eight novel markers. There are three reasons why these markers may not have been identified in previous studies: (1) the markers include repeated elements rejected in previous studies. In the present study loci with dinucleotide repeat elements were typically rejected. However, DYS710 was not eliminated because its short dinucleotide repeat tract is flanked by tetranucleotide repeat elements (Table 3.2) The presence of the dinucleotide repeat tract would have resulted in the rejection of this locus by Redd et al. (2002a) and Kayser et al. (2004). (2) Differences in primer selection parameters. Slight differences in primer selection parameters may have allowed for the selection of a functional primer set in the present study but not in previous studies. Primer design was particularly difficult for DYS710, DYS715 and DYS716 due to the presence of repetitive sequences. (3) Differences in the sequence information used for marker identification. Three of the four contigs used in the present study were also used by Kayser et al. (2004) (NT_011878, NT_011896, NT_011903). The studies differed in that Kayser et al. (2004) used NT_0113 while the present study used NT_011875. The loci DYS710, DYS712, DYS714, DYS715 and DYS717 all lie on NT_011875.

3.3.2 Male specificity of primer sets

Unlabeled PCR primers were designed for the 20 loci selected from Y-chromosome sequence data. The male-specificity of PCR primers was assessed by amplifying DNA from two male and one female donor in uniplex PCR reactions. All primer sets generated apparently single-copy male-specific amplicons from genomic DNA.

3.3.3 Variability of Y-STR loci and allele size ranges in a sample of 46 English Caucasian males

An initial indication of variability and relative allele size ranges was obtained for the 20 loci selected from Y-chromosome sequence data. This was achieved by analyzing DNA from 46 Caucasians using unlabeled PCR primers in uniplex reactions followed by PAGE analysis and silver staining. Figure 3.1 ranks these 20 loci according to *D* value within this population sample,

and Table 3.3 gives the actual D values. Gene diversity values ranged from 0.235 (DYS717) to 0.90 (DYS710) while the number of alleles identified per locus ranged from 3 to 15 (Table 3.3).

Table 3.3. D value for 20 male-specific single-copy loci among 46 English Caucasian males (loci chosen for further investigation are in bold)

Locus	Number of alleles observed	Gene diversity
DYS710	15	0.9007571
DYS711	12	0.8941407
DYS626	10	0.8487719
DYS712	11	0.8482995
DYS713	10	0.801513
DYS481	7	0.7977321
DYS518	7	0.7873551
DYS570	8	0.7807189
DYS714	8	0.7674865
DYS557	7	0.7476375
DYS614	8	0.739131
DYS612	6	0.7372405
DYS644	9	0.6984885
DYS607	7	0.6994334
DYS715	5	0.6389417
DYS514	7	0.6285452
DYS485	6	0.5630275
DYS716	3	0.5472593
DYS521	5	0.2778833
DYS717	3	0.23535



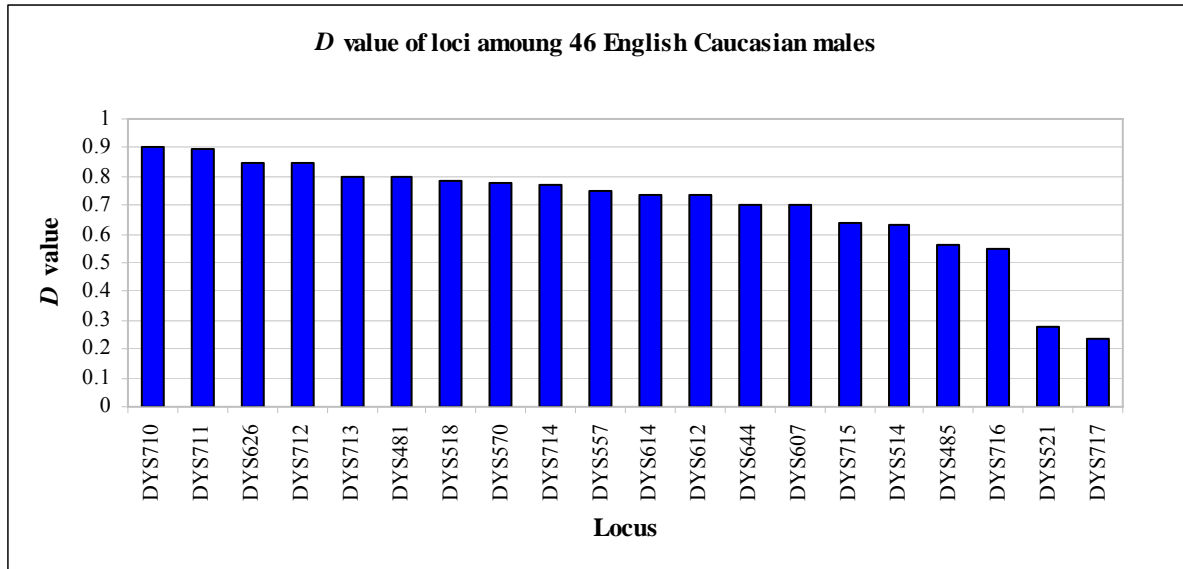


Figure 3.1. 20 male specific single-copy loci ranked according to *D* values among 46 English Caucasian males

Fourteen loci with gene diversity values above 0.65 (bold in Table 3.3) were selected for further analysis. In order to determine the relative sizes of the amplicons from each locus the largest and smallest alleles in each case were amplified and run on a single PAGE gel. This data would be used to plan the multiplex PCR primer sets so as to reduce the probability of generating amplicons from different loci with overlapping size ranges.

3.3.4 Selection of widely studied Y-STR loci

A set of relatively widely studied Y-STR loci were selected on the basis of high gene diversity values reported in a number of populations (Hou et al. 2001; Mohyuddin et al. 2001; Alvarez et al. 2002; Bosch et al. 2002; Gusmão et al. 2002; Iida et al. 2002; Redd et al. 2002a; Alves et al. 2003; Gusmão et al. 2003; Lee et al. 2003; Zhu et al. 2003). The selected loci included DYS439, DYS446, DYS447, DYS448, DYS449, DYS452, DYS458, DYS463, DYS635 (Y-GATA-C4) and Y-GATA-A10.

3.3.5 Relationship between repeat length and variability

STR loci have a good chance of being variable, provided that a consecutive homologous stretch with greater than eight repeated units are present (Moxon and Wills 1999). It has also been observed that STR loci with shorter units are more variable than longer units, making dinucleotide repeat loci more variable than for example pentanucleotide loci. There also seems to

be a direct relationship between variability and the average number of consecutively repeated units in a STR locus (Kayser et al. 2004). STR loci with higher average number of consecutive repeats would generally be more variable than those with a lower average number of consecutive repeats.

In the study done by Kayser et al. (2004), the factors influencing variability of STR loci were also investigated by means of statistical analyses. Repeat variance was used as a means for predicting variability of STR loci. For simple STR loci, it had been found that the average number of consecutive repeats in a locus was the overwhelming factor influencing its variability. Simple STR loci with higher average number of consecutive repeats would generally be more variable than those with a lower average number of consecutive repeats. For complex STR loci, the correlations were not as straight forward. In general, if two STR loci with the same consecutive number of repeated units exist, one being a simple repeat locus and the other being part of a complex locus, the complex locus is predicted to show more variability than the simple locus. Also, if two STR loci exist with the same total amount of repeated units, one being a simple locus and the other being a complex locus, the simple locus is predicted to generally be more variable than the complex locus. This suggests that even in complex loci, the average number of consecutive repeats in a locus greatly influences its variability.

Table 3.4 ranks the pentanucleotide loci according to the number of the longest consecutive stretch of repeated units. Table 3.5 ranks the simple and complex tetranucleotide loci according to the number of the longest consecutive stretch of repeated units. Table 3.6 ranks the trinucleotide loci according to the number of the longest consecutive stretch of repeated units. The numbers of repeated elements in these loci are as observed from sequence data on contigs NT_011875.7, NT_011878.7, NT_011896.7, and NT_011903.3. Multiple complex loci have been omitted. The study by Kayser et al. (2004) used the average number of repeats from various individuals to calculate repeat variance. In the absence of sequencing data, this study has used only the allele appearing on the contigs investigated. Even so, it seems as though there is generally a direct correlation between repeat length and variability among the tetra- and pentanucleotides. The only exception to this observation is DYS612.

Table 3.4. Pentanucleotide loci ranked according to longest stretch of repeats

Rank	Locus	Longest stretch of repeats	Gene Diversity
------	-------	----------------------------	----------------

1	DYS714	(CTTT)18	0.7674865
2	DYS644	(TTTA)16	0.6984885
3	DYS716	(TCCAT)12	0.5472593
4	DYS717	(TATTG)10	0.2353500

Table 3.5. Tetranucleotide loci ranked according to longest stretch of repeats

Rank	Locus	Longest stretch of repeats	Gene Diversity
1	DYS712	(AGAT)17	0.8482995
2	DYS570	(TTTC)17	0.7807189
3	DYS557	(TTTC)16	0.7476375
4	DYS514	(GAAA)15	0.6285452
5	DYS521	(CTTT)9	0.2778833

Table 3.6. Trinucleotide loci ranked according to longest stretch of repeats

Rank	Locus	Longest stretch of repeats	Gene Diversity
1	DYS711	(TCT)31	0.8941407
2	DYS612	(TCT)25	0.7372405
3	DYS481	(CTT)22	0.7977321
4	DYS614	(CTT)18	0.739131
5	DYS485	(TTA)16	0.5630275

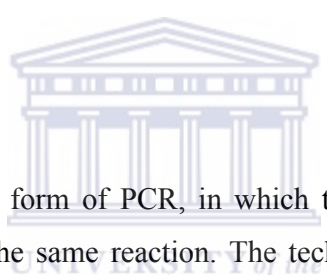
3.3.6 Summary

In order to obtain a more comprehensive assessment of the properties of the Y-STR loci being studied, a more substantial population study is needed. From the two approaches followed here, 24 loci were considered for further investigation in such a population study. From the available Y-chromosome sequence data and some preliminary polymorphism testing, 14 loci were selected for further investigation. These were *DYS710*, *DYS711*, *DYS626*, *DYS712*, *DYS713*, *DYS481*, *DYS518*, *DYS570*, *DYS714*, *DYS557*, *DYS614*, *DYS612*, *DYS607* and *DYS644*. They were chosen because they typically showed a gene diversity value of 0.65 and higher among the 46 English-speaking males tested. From the literature searches at the time, 10 loci were selected. These were *DYS439*, *DYS446*, *DYS447*, *DYS448*, *DYS449*, *DYS452*, *DYS458*, *DYS463*, *DYS635* (Y-GATA C4) and Y-GATA A10. They were chosen because they typically had a reported gene diversity value of ~0.60 and higher in a range of population studies. Multiplex

reactions would be used as a means to facilitate the analysis of the loci chosen for further investigation.

Chapter 4 – Multiplex Amplification of Y-STR Loci

4.1 Introduction



Multiplex PCR is a modified form of PCR, in which two or more target DNA sequences are simultaneously amplified in the same reaction. The technique has many advantages. Multiplex PCR facilitates the simultaneous typing of a range of STR loci in less time, using fewer reagents than would be the case with uniplex PCRs. However, the technique also presents challenges. Often the DNA target sequences do not all amplify equally and in some cases, some of them do not amplify at all. Multiplex PCR may also generate non-specific products that makes the reaction less efficient and also complicates analysis of the results. A substantial amount of optimization is required to develop a multiplex PCR typing system in such a way that all target sequences (and only the target sequences) are consistently and equally amplified.

In addition to the ‘minimal haplotype’ (MH) loci (see Chapter 2), a total of 24 Y-STR loci were chosen for further investigation (see Chapter 3). 14 loci (DYS710, DYS711, DYS626, DYS712, DYS713, DYS481, DYS518, DYS570, DYS714, DYS557, DYS614, DYS612, DYS607 and DYS644) were selected from the available sequence data and some preliminary polymorphism testing. 10 loci (DYS439, DYS446, DYS447, DYS448, DYS449, DYS452, DYS458, DYS463, DYS635 (Y-GATA C4) and Y-GATA A10) were selected from the literature searches at the time. In order to obtain a more comprehensive assessment of the properties of the Y-STR loci

being studied, a more substantial investigation of these loci was needed. For this reason, properties such as polymorphism among populations likely to have distinct population histories, and stutter generated by PCR was required. The study would assess samples from English-speaking Caucasian-, Black Xhosa-speaking- males and Asian Indian males.

Multiplex PCR was considered an efficient way of analyzing these loci. Several reports were consulted to find a robust approach for this technique (Henegariu et al. 1997; Butler et al. 2002; Wallin et al. 2002; Schoske et al. 2003). Three basic multiplex reactions were constructed according to protocols which took into account the expected size ranges of the fragments, the fluorescent dye labels of primers, similarity in amplification conditions as well as post-PCR modifications of amplified fragments (Brownstein et al. 1996; Magnuson et al. 1996; Butler et al. 2002; Schoske et al. 2003).

4.2 Methods and Materials

4.2.1 Loci selected for further investigation

In addition to the MH loci, a total of 24 Y-STR loci were chosen for further investigation. 14 recently identified loci (DYS710, DYS711, DYS626, DYS712, DYS713, DYS481, DYS518, DYS570, DYS714, DYS557, DYS614, DYS612, DYS607 and DYS644) were selected for further investigation from the available sequence data and some preliminary polymorphism testing. These loci were chosen because they typically showed a gene diversity value of ~0.65 and higher among the 46 English-speaking Caucasian males investigated. 10 loci (DYS439, DYS446, DYS447, DYS448, DYS449, DYS452, DYS458, DYS463, DYS635 (Y-GATA C4) and Y-GATA A10) were selected from previous reports. They were chosen because they typically had reported *D* values of ~0.60 and higher in a range of population studies (Hou et al. 2001; Mohyuddin et al. 2001; Alvarez et al. 2002; Bosch et al. 2002; Gusmão et al. 2002; Iida et al. 2002; Redd et al. 2002a; Alves et al. 2003; Gusmão et al. 2003; Lee et al. 2003; Zhu et al. 2003).

4.2.2 Primer design-based approach to multiplex design

In order to design multiplex reactions for the amplification of loci, a primer design-based approach was followed. Figure 4.1 is a flow diagram that presents this approach (adapted from Schoske et al. 2003). Three basic factors (allele size range, fluorescent dye labels and

amplification conditions) were considered for multiplex design. In order to accommodate this design strategy, the 14 recently identified loci were re-analyzed to assess the likelihood of designing alternative primers. Amplicon size reduction through alternative primer design was possible for six loci. Of all the alleles found for each of these loci, the biggest and smallest alleles were run on a polyacrylamide gel as described in Chapter 3. In the six cases where two primer sets were available, amplification of alleles was done with both sets of primers. This gave an indication as to how loci could be combined into the multiplex reaction. For the 10 loci chosen on the basis of reported gene diversity values, allele size ranges were obtained from either the original report or from the Genome Database (GDB) (<http://www.gdb.org>). Primer sequences for DYS446 and DYS449 were as reported by Redd et al. (2002a). Primer sequences for GATA A10 and DYS635 (GATA C4) were as described in GDB. *Primer3* software was used to re-design primer sets for the other loci using the same criteria as described in Chapter 3.

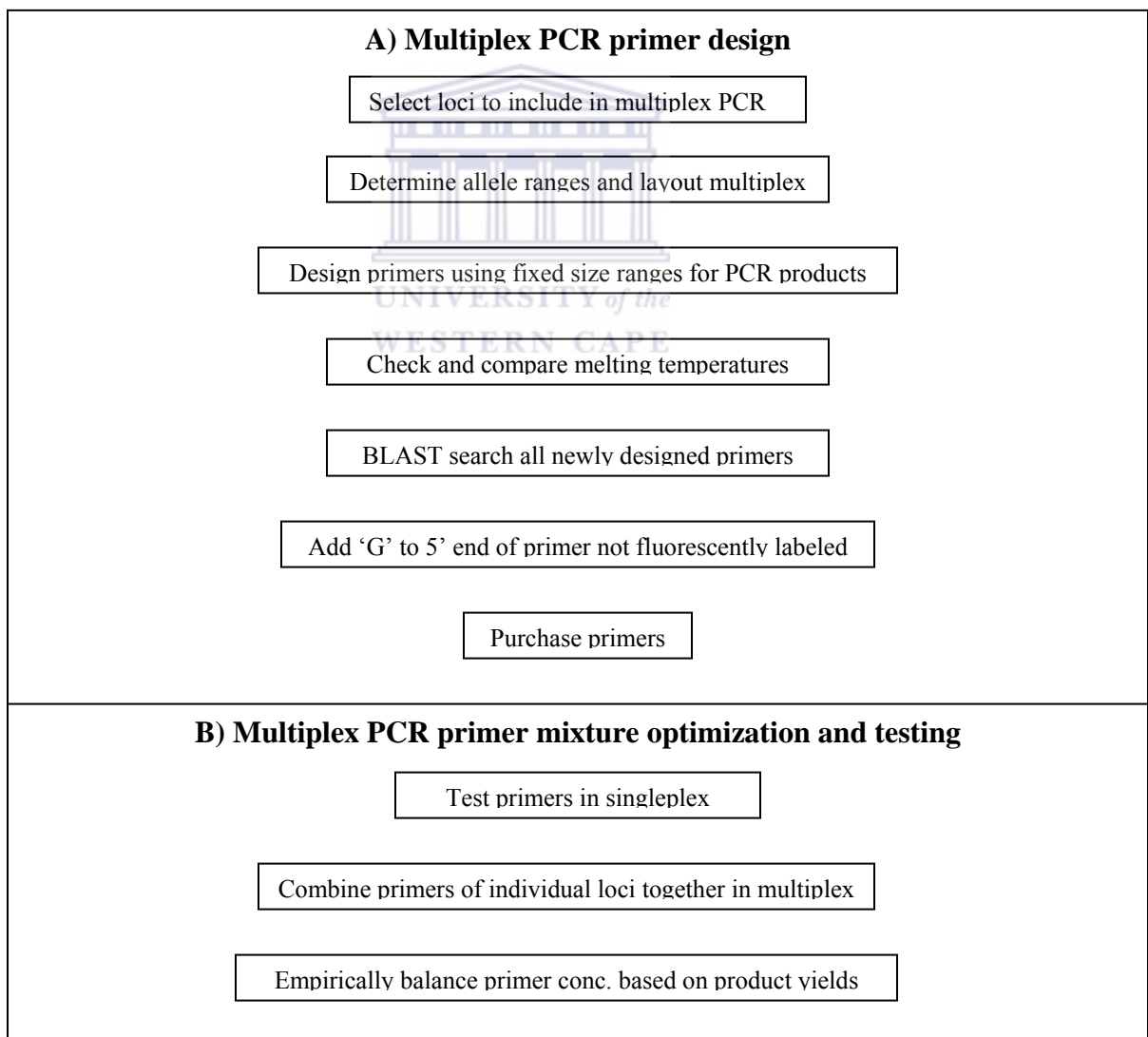


Figure 4.1. Primer design-based approach to multiplex optimization

(adapted from Schoske et al. 2003)

All primers were designed with similar annealing temperatures. Post-PCR modification to avoid the occurrence of split peaks was also considered. One approach to address this was to add a 'G' nucleotide on to the 5' end of unlabeled primers in a pair (Schoske et al. 2003). All primers were tested for male specificity and copy number as previously described in Chapter 3. Primers for DYS448 were omitted from further analysis as these primers failed to generate a PCR product from genomic DNA, even after rigorous optimization steps.

For the multiplex reactions described here, 3-dye systems (6-FAM, VIC and NED) were used, with an additional dye (ROX) for the internal lane size standard. Using this approach, three basic multiplex reactions (UWC Y-Plex 1, 2 and 3) were constructed. UWC Y-Plex 1 amplified six loci, UWC Y-Plex 2 amplified 11 loci and UWC Y-Plex 3 amplified eight loci. Primers for DYS447 were omitted from further analysis as these primers dramatically impeded the amplification of other loci in UWC Y-Plex 3.

4.2.3 Multiplex amplification of Y-STR loci

Table 4.1 presents concentration of reagents used for amplification of UWC Y-Plex 1, 2 and 3. Table 4.2 presents the cycling conditions used to amplify UWC Y-Plex 1, 2 and 3. PCR primers were synthesized by *Inqaba Biotech*, *MWG Biotech* or *Applied Biosystems*. Table 4.3 presents the primer sequences, final concentration in the multiplex PCR, and size range of alleles as determined by the internal lane standard ROX 500. Post-PCR modification (the non-template addition of dATP to PCR products) was also performed as described in Chapter 2. Modified PCR products were stored at 4°C in the dark until fragment analysis took place.

Table 4.1. Amplification components of multiplex reactions UWC Y-Plex 1, 2 and 3

PCR Cocktail	UWC Y-Plex 1	UWC Y-Plex 2	UWC Y-Plex 3
Genomic DNA (Total)	4ng	4ng	4ng
Tris-HCl, pH 8.3 (Final Conc.)	10mM	10mM	10mM
KCl (Final Conc.)	50mM	50mM	50mM
MgCl ₂ (Final Conc.)	1.5mM	1.5mM	1.5mM
dNTPs (Final Conc.)	200µM	200µM	200µM
Taq Polymerase (Total)	0.5U	0.5U	0.5U

Final volume 10µl 10µl 10µl

Table 4.2. Cycling conditions of multiplex reactions UWC Y-Plex 1, 2 and 3

Amplification Conditions	Number of Cycles	UWC Y-Plex 1	UWC Y-Plex 2	UWC Y-Plex 3
Thermal Cycler		GeneAmp 2700 (<i>Applied Biosystems</i>)		
Denaturation (11 minutes)	1	95°C	95°C	95°C
Denaturation (1 minute)	34	94°C	94°C	94°C
Annealing (1.5 minutes)		58°C	60°C	62°C
Extention (1.5 minutes)		72°C	72°C	72°C
Final Extention (75 minutes)	1	68°C	68°C	68°C

Table 4.3 Sequences, fluorescent dye labels and final concentration of primers used in multiplex reactions, UWC Y-Plex 1, 2 and 3

Locus	Primer Sequence, Dye Label and Orientation	Multiplex Sets			Allele size range (bp)
		Set 1	Set 2	Set 3	
DYS19	F: 6-FAM-5'-cta ctg agt ttc tgt tat agt 3' R: 5'-atg gcc atg tag tga gga ca 3'	0.12µM			
DYS713	F2: 5' gtg caa gcc aag ggc ttt ata agt 3' R2: 5' FAM cct ggg tga cag act cca tct taa a 3'	0.5µM			293-344
DYS712	F2: 5' g caa gaa cag cct ggg taa cag tg 3' R2: 5' VIC tta tat ggt aca gcc cat gaa cac tt 3'	0.07µM			148-210
DYS714	F2: 5' gta tta ggc cat ctt gcc agc 3' R2: 5' NED ttt tct acc tat gat gcc ctt tg 3'	0.15µM			152-205
DYS626	F2: 5' gca aga ccc cat agc aaa ag 3' R2: 5' NED aag aag aat ttt ggg aca tgt tt 3'	0.3µM			233-281
DYS570	F2: 5' g tga cta ggt aga aat cct ggc tgt g 3' R2: 5' FAM tca gca tag tca aga aac cag aca ac 3'		0.075µM		162-197
DYS710	F2: 5' FAM gag gtc aag gct gca aga atc tat ga 3' R2: 5' g cat act ctc tcc ctc cct ctc ttt ttt c 3'		0.5µM		197-255
DYS518	F: 5' ggc aac aca agt gaa act gct tct 3' F: 5' FAM gct ctt acc atg ggt gat ttc ttt c 3'		0.5µM		258-311
DYS711	F2: 5' cag agc cca gca cct agg tta agt 3' R2: 5' FAM tgc tgt cat tgt atc tct tca ctc c 3'		0.5µM		321-375
DYS481	F2: 5' VIC aaa agg aat gtg gct aac gct gtt c 3' R2: 5' gct cac cag aag gtt gca aga ctc a 3'		0.2µM		120-154
DYS612	F2: 5' VIC aag ttt cac aca ggt tca gag gtt g 3' R2: 5' gac act tgc cat ggg tat cta gag c 3'		0.3µM		179-222

DYS557	F: 5' gcc cag cat gtg ttt tga cta ttt 3' R: 5' VIC ggt gtt tga ccc agt gat atg ttc t 3'	0.15µM 0.15µM	219-259
DYS614	F1: 5' gaa tat ggc taa ctc gat tca gag 3' R1: 5' VIC cca cca aaa ggt ttt cag act ca 3'	0.5µM 0.5µM	315-342
DYS607	F2: 5' NED agc ata cag cgt aat cac agc tca c 3' R2: 5' gct caa aac cca taa ttc agg tcc ag 3'	0.1µM 0.1µM	215-244
DYS446	F2: 5' g tat ttt cag tct tgt cct gtc 3' R2: 5' NED aaa tgt atg gcc aac ata gca aaa cca 3'	0.5µM 0.5µM	284-328
GATA-A10	F: 5' FAM cct gcc atc tct att tat ctt gca 3' R: 5' g ata aat gga gat agt ggg tgg att 3'	0.096µM 0.096µM	151-184
DYS452	F: 5' FAM cat tgg tgg tgt tct gat gag gat aat 3' R: 5' gag ttt tac atg tag caa ata ggt t 3'	0.24µM 0.24µM	226-267

Table 4.3 Continued ...

Locus Name	Primer Sequence, Dye Label and Orientation	Multiplex Sets			Allele size range (bp)
		Set 1	Set 2	Set 3	
DYS644	F2: 5' ggg tag ttc cag gcc cta att cat 3' R2: 5' VIC gtt gtg tca ctg acc tcc aac ct 3'			0.24µM 0.24µM	152-225
DYS439	F: 5' VIC tgt cct gaa tgg tac ttc cta ggt t 3' R: 5' g atg cct ggc ttg gaa ttc ttt tac 3'			0.24µM 0.24µM	243-263
(DYS635) GATA-C4	F: 5' g cac tgt att tca gct tga gtg atg g 3' R: 5' VIC ctc ttg gct tct cac ttt gca tag aat 3'			0.24µM 0.24µM	269-310
DYS458	F: 5' gag caa cag gaa tga aac tcc aat 3' R: 5' NED cat gag cca cca cgc cca c 3'			0.048µM 0.048µM	118-152
DYS463	F2: 5' NED tga tgt aga cta aga gcc aca gag c 3' R2: 5' gag gtt gtg tga ctt gac tga ctc ct 3'			0.24µM 0.24µM	164-205
DYS449	F2: 5' NED tgg agt ctc tca agc ctg ttc ta 3' R2: 5' g cct gga agt gga gtt tgc tgt 3'	0.5µM 0.5µM	0.5µM 0.5µM	0.5µM 0.5µM	340-384

4.2.4 Allelic ladders

Allelic ladders were constructed to facilitate consistent allele typing. For the loci identified from Y-chromosome sequence data, amplicons of different sizes were identified for 46 individuals by PAGE analysis. Corresponding genomic DNA samples representing the most common alleles were mixed and amplified with the relevant primer set. This resulted in a single completed PCR reaction containing a range of reference amplicons. For the loci chosen on the basis of published

gene diversity values, an initial group of 36 individuals was typed and the same approach followed. Sequencing of the alleles of each ladder was not undertaken. This was due to the fact that the study aimed to examine key properties of polymorphism, stutter and copy number. Allelic ladder sequencing will be undertaken for loci judged suitable for future use. An artificial allele nomenclature was used with the smallest allele, identified in initial studies, designated as allele 1.

4.2.5 Fragment Analysis

Amplified fragments were analyzed using an ABI 377 Genetic Analyzer (*Applied Biosystems*). Samples were prepared for electrophoresis by mixing 1µl of the PCR product with 1.5µl of loading dye mix. The loading dye mix consisted of de-ionized formamide, ROX 500 size standard and dextran blue (all *Applied Biosystems*) in a ratio of 5:1:1. For the allelic ladders, the same mixing procedure was followed. These mixes were then heat-denatured at 95°C for 2 minutes and snap-cooled immediately until the products were loaded.

To perform gel electrophoresis, manufacturer's instructions were followed. Plates were assembled, using 0.2mm spacers. A gel solution was prepared containing 1X TBE (pH 8.3), 36% w/v urea and a final concentration of 5% Long Ranger polyacrylamide (*Applied Biosystems*). After the urea was dissolved, 25ml of the gel solution was passed through a 0.22-micron filter and the gel solution de-gassed. Aliquots of 125µl of 10% APS (Ammonium Persulphate) and 17.5µl of TEMED (N, N, N', N' – Tetramethyl – Ethylenediamine) were added to the gel mixture. The gels were poured and allowed to polymerize for a minimum of two hours.

Gels were run according to manufacturer's instructions. Data was collected with the *ABI 377 collection* software (*Applied Biosystems*) and analyzed using *GeneScan 3.1* (*Applied Biosystems*) software. Sized fragments were then converted to allele numbers, with the use of the *Genotyper 2.5* (*Applied Biosystems*) software. An artificial allele nomenclature was used with the smallest allele in the initial studies designated as allele 1. Allele numbers were designated according to the known allele numbers of the fragments in the allelic ladders that were included in every run.

4.3 Results and Discussion

4.3.1 Multiplex reactions designed

Three multiplex reactions (UWC Y-Plex 1, UWC Y-Plex 2 and UWC Y-Plex 3) were designed. Allele size ranges for loci of all three multiplex reactions have been presented in Table 4.2.3.3.

UWC Y-Plex 1 amplified six Y-STRs. These included two 6-FAM labelled primer sets (DYS19 and DYS713), one VIC labelled primer set (DYS712) and three NED labelled primer sets (DYS714, DYS626 and DYS449). Figure 4.2 (a) presents the observed allele size ranges for UWC Y-Plex 1 loci. Figure 4.2 (b) depicts an example of a typical electropherogram for the multiplex, UWC Y-Plex 1.

UWC Y-Plex 2 amplified 11 Y-STRs. These included four 6-FAM labelled primer sets (DYS570, DYS710, DYS518, and DYS711), four VIC labelled primer sets (DYS481, DYS612, DYS557 and DYS614) and three NED labelled primer sets (DYS607, DYS446 and DYS449). Figure 4.3 (a) presents the observed allele size ranges for UWC Y-Plex 2 loci. Figure 4.3 (b) depicts an example of a typical electropherogram for the multiplex, UWC Y-Plex 2.

UWC Y-Plex 3 amplified eight Y-STRs. These included two 6-FAM labelled primer sets (Y-GATA A10 and DYS452), three VIC labelled primer sets (DYS644, DYS439 and DYS635) and three NED labelled primer sets (DYS458, DYS463 and DYS449). Figure 4.4 (a) presents the observed allele size ranges for UWC Y-Plex 3 loci. Figure 4.4 (b) depicts an example of a typical electropherogram for the multiplex, UWC Y-Plex 3.

There have been two incidences where undesirable primer-primer interactions appear to have occurred. The first occurred in UWC Y-Plex 1 where the undesirable interactions of primers caused the formation of PCR artifacts. The second occurred in UWC Y-Plex 3 where the entire multiplex reaction failed if one primer set (DYS447) was not removed from the reaction. Such undesirable primer-primer interactions can be avoided by investigating the impact that different primers in a multiplex reaction will have on the formation of amplified PCR products. This approach can now be automated with the use of computer programs such as 'AutoDimer', which increases the probability of designing efficient multiplex reactions (Vallone and Butler 2004).

4.3.2 Performance of multiplex UWC Y-Plex 1

Figure 4.2 (b) depicts a typical electropherogram obtained with UWC Y-Plex 1. All six loci amplified well and peak heights were reasonably balanced. In addition to six peaks representing six loci, two other 6-FAM-labelled peaks were also routinely observed. These PCR artifacts (indicated by arrows in the diagram) are most likely caused by undesirable interactions between primers in the multiplex reaction. Because the sizes of both these artifacts did not overlap with the size ranges of alleles to be typed, they were generally disregarded during analysis. The amplification of DYS714 was slightly less efficient than the other two NED-labeled loci. Further optimization should address this imbalance.

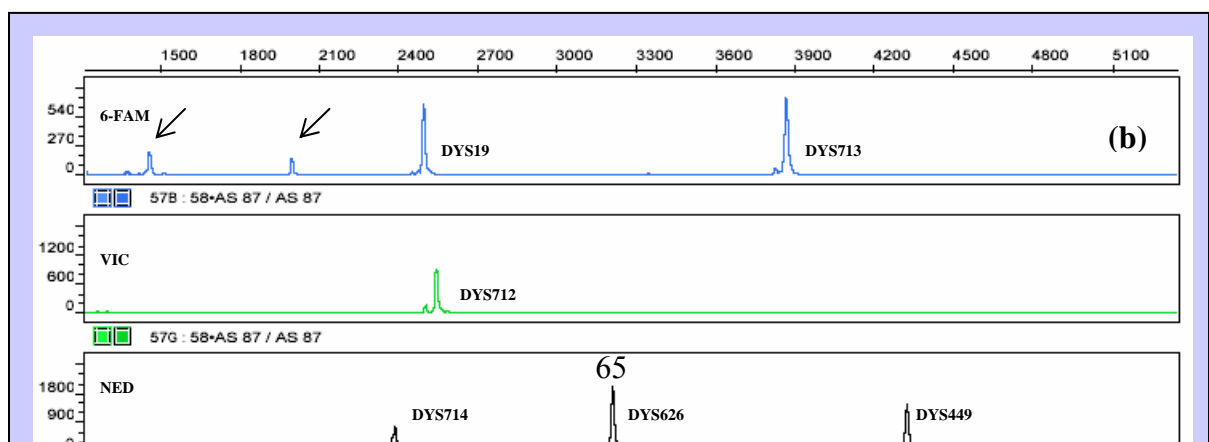
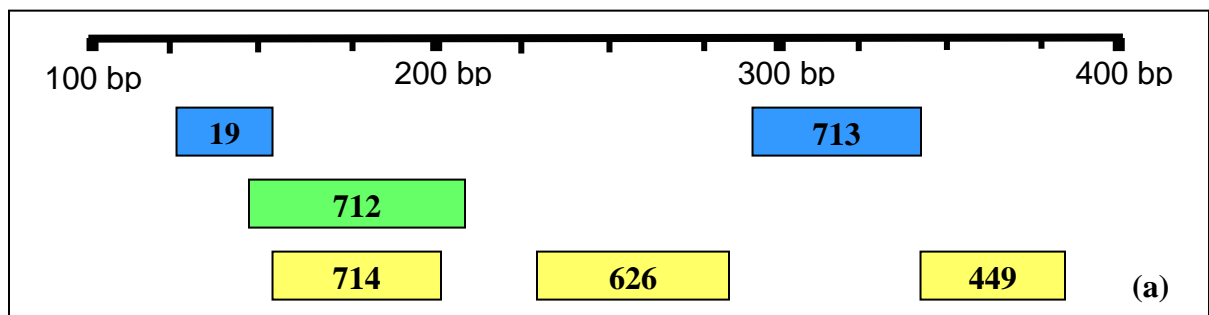


Figure 4.2. (a) Observed allele size ranges of six loci amplified with multiplex UWC Y-Plex 1 and (b) An example of a typical electropherogram for the multiplex, UWC Y-Plex 1 (Arrows indicate PCR artifacts generated by the multiplex reaction)

4.3.3 Performance of multiplex UWC Y-Plex 2

Figure 4.3 (b) depicts a typical electropherogram obtained with UWC Y-Plex 2. All 11 loci consistently amplified and no PCR artifacts were observed. In each dye lane, there was one STR that did not amplify in a balanced manner. Of the 6-FAM-labeled STRs, the peak of DYS518 was generally brighter than the rest of the STRs labeled with this dye. Of the VIC-labeled STRs, the peak for DYS481 was generally much fainter than the rest of the STRs labeled with this dye. Of the NED-labeled STRs, the peak for DYS449 was generally slightly fainter than the rest of the STRs labeled with this dye. More rigorous optimization could balance the amplification products for these loci among the other loci in this multiplex reaction. Because all 11 loci consistently amplified, this multiplex was deemed acceptable as a means to generate population data.

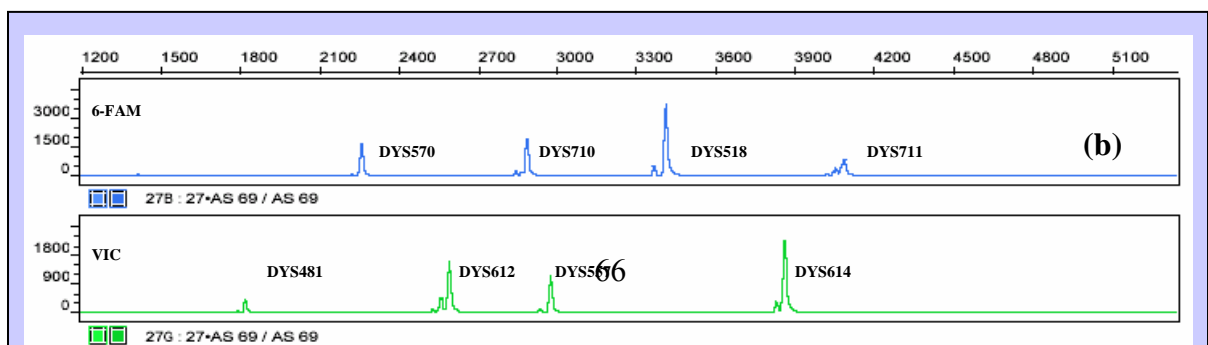
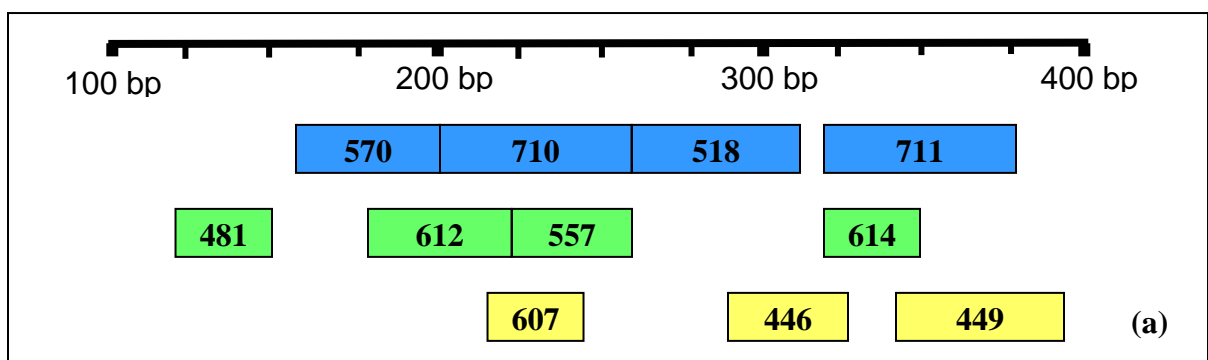


Figure 4.3. (a) Observed allele size ranges of 11 loci amplified with multiplex UWC Y-Plex 2 and (b) An example of a typical electropherogram for the multiplex, UWC Y-Plex 2

4.3.4 Performance of multiplex UWC Y-Plex 3

Figure 4.4 (b) depicts a typical electropherogram obtained with UWC Y-Plex 3. Primers designed for the amplification of DYS447 were omitted from this multiplex as they generated substantial artifacts that prevented other loci from being amplified. The other eight loci consistently amplified without generating PCR artifacts. In each dye lane, there was one STR that over-amplified relative to the rest of the loci. These were GATA A10, DYS644 and DYS463. Once again, more rigorous optimization should be able to balance the PCR products of this multiplex reaction. Since eight loci amplified consistently, this multiplex was used as a means to generate population data.

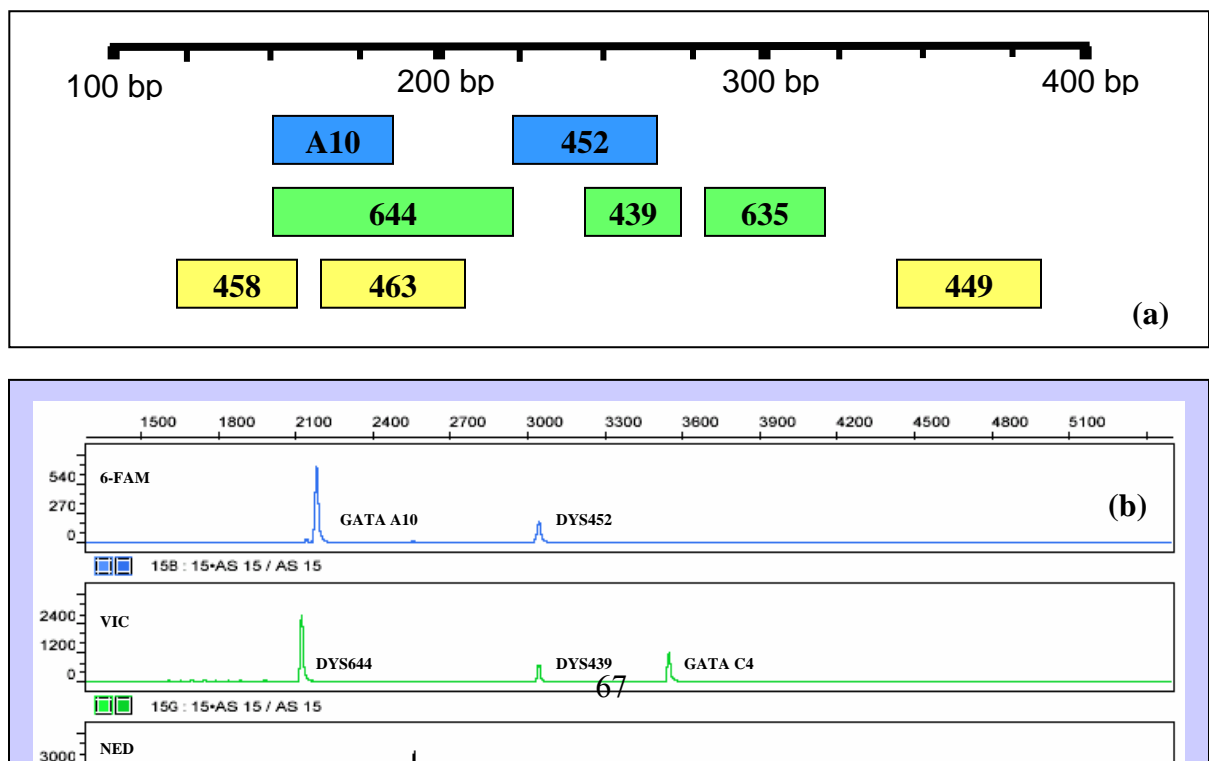


Figure 4.4. (a) Observed allele size ranges of eight loci amplified with multiplex UWC Y-Plex 3 and (b) An example of a typical electropherogram for the multiplex UWC Y-Plex 3

4.3.5 Summary

Approaches for the development of Y-STR typing systems that include up to 20 loci are well established (Butler et al. 2002; Hall and Ballantyne 2003a; Schoske et al. 2003; Schoske et al. 2004; Daniels et al. 2004; Hanson and Ballantyne 2004). While these multiplex reactions are able to amplify large numbers of loci, most of these multiplex reactions amplify at least some loci that do not show high variability. The multiplex reactions designed here amplify loci shown to be highly variable (see Chapter 3), thereby increasing the discriminatory capacity of haplotypes created with these loci.

Three multiplex reactions (UWC Y-Plex 1, UWC Y-Plex 2 and UWC Y-Plex 3) were designed. UWC Y-Plex 1 consistently amplified six loci, UWC Y-Plex 2 consistently amplified 11 loci and UWC Y-Plex 3 consistently amplified eight loci. These multiplex reactions were therefore accepted as efficient tools with which to investigate properties of loci such as polymorphism among South African sub-populations and stutter generated by PCR amplification. Allele sizes generally ranged between 120bp and 350bp, making these multiplex reactions useful for the analysis of degraded DNA. The two exceptions are the loci DYS449 and DYS711 for which allele sizes can range up to 384bp.

Chapter 5 – Investigating properties of loci among three South African sub-populations

5.1 Introduction

A number of studies have indicated that Y-STR haplotype diversity is limited in several populations. In a study by Mohyuddin et al. (2001), it has been found that haplotype analysis with as many as 16 loci could not adequately distinguish between some Pakistani males. Genotyping of 726 Pakistani males from 12 ethnic groups was conducted with seven ‘minimal haplotype’ (MH) loci (the duplicated locus DYS385 was excluded) as well as nine other loci (DYS388, DYS425, DYS426, DYS434, DYS435, DYS436, DYS437, DYS438 and DYS439). From the results of this study, it was shown that 13 of the 90 individuals (14.4%) from the Parsi ethnic group shared the same haplotype and 17 of the 107 individuals (15.9%) from the Brahui ethnic group shared another haplotype. When investigating the gene diversity values for loci among these ethnic groups, it was observed that three of the non-MH loci (DYS434, DYS435 and DYS436) are practically mono-morphic and have gene diversity values below 0.3.

More recently, a study by Hedman et al. (2004), reported that haplotype analysis with 16 loci could also not adequately distinguish between some Finnish males. When the MH loci were used to investigate 400 Finnish males, 22.5% shared the same haplotype. The number of loci was then increased to 16 with the addition of seven more loci (DYS435, DYS436, DYS437, DYS438, DYS439, DYS460 and Y-GATA H4). When 200 males for which MH data was available were

typed with the additional loci, 13% still shared the same haplotype. When one examines the gene diversity of the individual loci, it can be seen that three non-MH loci (DYS435, DYS436 and DYS438) have diversity values below 0.2. It was observed that the inclusion of these loci did not add to the diversity of the haplotypes generated, and could therefore have been omitted. Three more loci have a gene diversity value of less than 0.5 and probably contributed little to the observed haplotype diversity.

The reduced observed Y-STR haplotype diversity within these populations can be attributed to at least two factors. Firstly, these are genetically distinct groups originating from smaller founder populations (Sajantila et al. 1996; Qamar et al. 2002). Secondly, some of the STR loci chosen to investigate these populations exhibit low variability even among populations that are known to be genetically diverse (Bosch et al. 2002).

In order to maximize the discriminatory capacity of a Y-STR typing system, it seems appropriate include as many highly variable loci as possible. Factors to consider when choosing such loci include: (1) the ease with which male specific primers can be designed to amplify loci, (2) the extent to which stutter is generated during PCR and (3) individual locus variability. In recent years it has become common practice for researchers to construct multiplex reactions using large numbers of STR loci (Butler et al. 2002; Hall and Ballantyne, 2003; Schoske et al. 2004; Schoske et al. 2004; Daniels et al. 2004; Hanson and Ballantyne 2004). It is often the case that some of the loci would exhibit high levels of variability, and some would exhibit very low levels of variability. Omitting loci that exhibit low levels of variability seems appropriate, as they are unlikely to increase haplotype diversity (Hedman et al. 2004).

The present study aimed to investigate the properties of a collection of highly variable Y-STR loci. The study was initiated by examining the variability of the loci in three populations likely to have distinct population histories. The loci selected for the study included five single copy loci from the MH, eight loci selected from previous reports and 14 loci recently identified from Y-chromosome sequence data and preliminary polymorphism testing. The effect of individual loci on overall haplotype diversity was also investigated as well as the extent to which stutter was generated by PCR.

5.2 Methods and Materials

Population data was generated for three South African sub-populations likely to have distinct populations histories. These were the English-speaking Caucasian-, the Xhosa-speaking Black- males from the study by Leat et al. (2004) as well as the Asian Indians from KwaZulu-Natal. Genotyping was attempted for 101 English-speaking Caucasian-, 88 Xhosa-speaking Black- and 77 Asian Indian males. Multiplex reactions as described in Chapter 4 facilitated the analysis of 27 single copy loci. These included the single-copy loci of the MH (DYS19, DYS390, DYS391, DYS392, DYS393), eight loci selected on the basis of published gene diversity values (DYS439, DYS446, DYS449, DYS452, DYS458, DYS463, DYS635 (Y-GATA C4) and Y-GATA A10) and the 14 most variable loci selected from Y-chromosome sequence data (DYS710, DYS711, DYS626, DYS712, DYS713, DYS481, DYS518, DYS570, DYS714, DYS557, DYS614, DYS612, DYS607 and DYS644). Gene- and haplotype diversity values were calculated as described previously in Chapters 2 and Chapter 3. By ranking the 27 single copy loci from the highest mean gene diversity to lowest mean gene diversity and calculating the haplotype diversity with the continual addition of a locus, the effect of individual loci on haplotype diversity was also investigated. An indication of PCR stutter was obtained by analyzing the profiles for ten samples with peak heights above 400 RFUs.

5.3 Results and Discussion

5.3.1 Variability of loci among three South African sub-populations

In order to obtain a more comprehensive assessment of the properties of the Y-STR loci being studied, a more substantial population study was conducted. An attempt was made to select three populations likely to have distinct population histories. The study assessed samples from English-speaking Caucasian males (101 samples), Xhosa-speaking Black males (88 samples) and Asian Indian males (77 samples). Multiplex reactions facilitated the analysis of 27 loci, including the single-copy loci of the MH (DYS19, DYS390, DYS391, DYS392, DYS393), eight loci selected on the basis of published gene diversity values (DYS439, DYS446, DYS449, DYS452, DYS458, DYS463, DYS635 (Y-GATA C4) and Y-GATA A10) and the 14 most variable loci selected from Y-chromosome sequence data (DYS710, DYS711, DYS626, DYS712, DYS713, DYS481, DYS518, DYS570, DYS714, DYS557, DYS614, DYS612, DYS607 and DYS644). Gene diversity values for all 27 loci among the three sub-populations are presented in Table 5.1.

Three parameters were assessed for each locus in each population group: gene diversity, the number of alleles identified and the frequency of the most common allele (Figure 5.1). Mean gene diversity values ranged from 0.317 (DYS391) to 0.887 (DYS711) and the number of alleles ranged from 3 (DYS391, DYS392) to 21 (DYS710) (Figure 5.3.1). The novel STRs DYS710, DYS711, DYS712, DYS713 and DYS714 are among the most variable loci while the widely used single-copy Y-STRs DYS391, DYS392 and DYS393 are among the least variable loci. The greatest variation in gene diversity between populations was observed for the least variable markers particularly DYS391 and DYS392.

Table 5.1. Gene diversity values for 27 single-copy STR loci among three South African sub-populations

Locus	Caucasian	Asian	Xhosa	Mean Gene diversity
DYS711	0.900	0.890	0.870	0.887
DYS710	0.920	0.920	0.800	0.880
DYS518	0.800	0.830	0.860	0.830
DYS626	0.830	0.800	0.790	0.807
DYS714	0.790	0.850	0.760	0.800
DYS449	0.770	0.860	0.770	0.800
DYS612	0.760	0.800	0.810	0.790
DYS713	0.740	0.820	0.800	0.787
DYS712	0.860	0.830	0.640	0.777
DYS570	0.740	0.830	0.700	0.757
DYS458	0.780	0.780	0.690	0.750
DYS557	0.710	0.850	0.670	0.743
DYS644	0.710	0.760	0.760	0.743
DYS614	0.690	0.850	0.650	0.730
DYS481	0.700	0.670	0.800	0.723
DYS446	0.640	0.810	0.670	0.707
DYS607	0.710	0.810	0.570	0.697
DYS635	0.600	0.830	0.640	0.690
DYS463	0.650	0.760	0.580	0.663
DYS390	0.660	0.730	0.560	0.650
Y-GATA-A10	0.530	0.670	0.710	0.637
DYS19	0.480	0.700	0.710	0.630
DYS439	0.610	0.710	0.560	0.627
DYS452	0.580	0.660	0.590	0.610
DYS393	0.310	0.690	0.570	0.523
DYS392	0.570	0.450	0.040	0.353
DYS391	0.540	0.280	0.130	0.317

For 15 of the 27 loci, more alleles were found with the Asian Indian population than for the other two populations investigated. For 17 loci, the Asian Indian population exhibits the highest gene diversity values. This suggests that the Asian Indian population was the most diverse of the three groups investigated. This confirms the results obtained using the MH (see Chapter 2). For 19 of the 27 loci, there are fewer alleles found with the Xhosa population than for the other two populations investigated. For 16 loci, this population exhibits the lowest gene diversity values. This suggests that the Xhosa group was the least diverse of the three groups investigated.



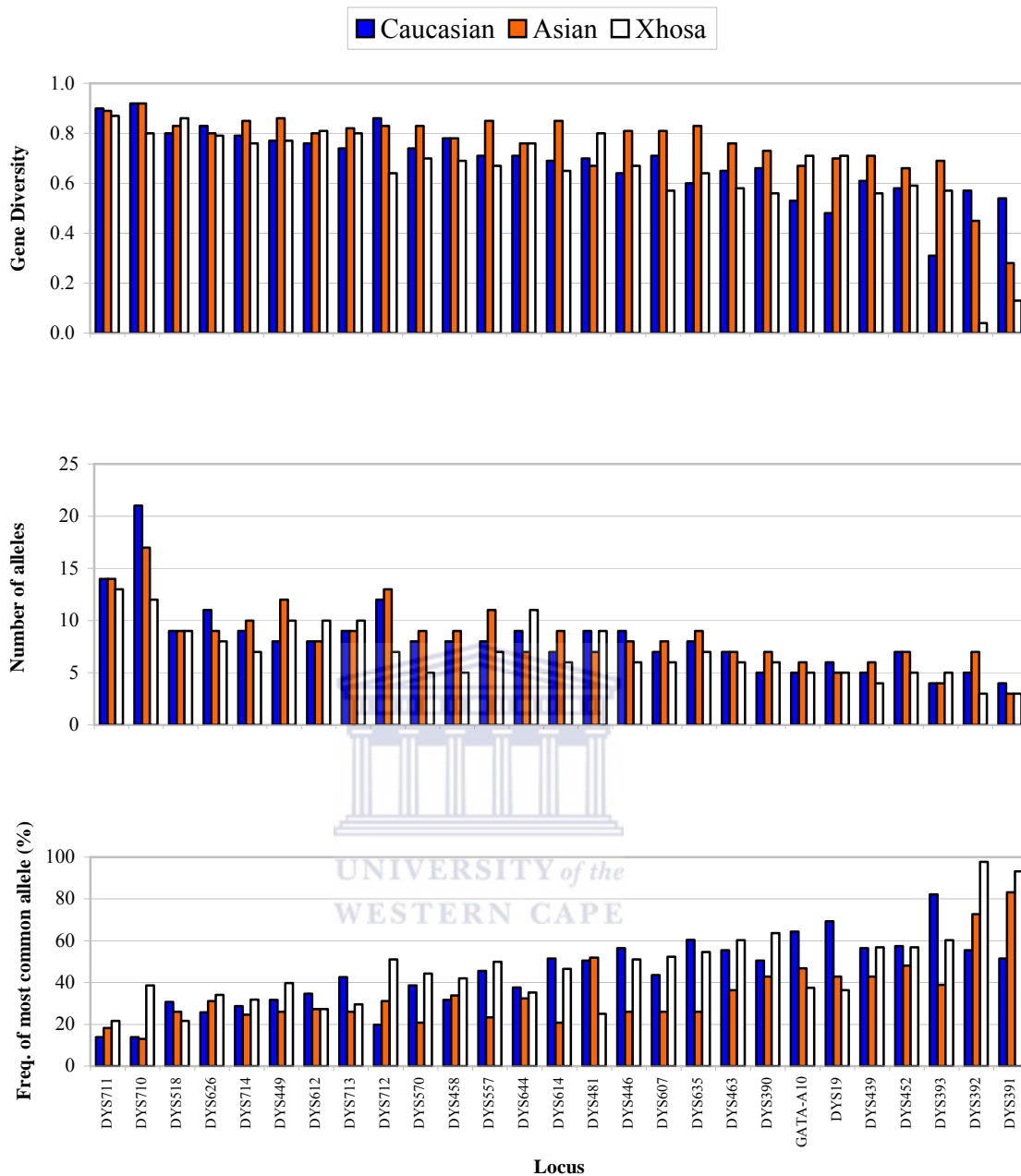


Figure 5.1 Gene diversity, number of alleles identified and frequency of the most common allele. Loci have been ranked from highest to lowest mean gene diversity.

The frequency of the most common allele and the number of alleles identified gives a simple indication of the distribution of allele frequencies (Figure 5.1). As expected the least variable markers have spiked allele frequency distributions with a high frequency for one allele and few additional alleles. This is especially the case for *DYS391*, *DYS392* and *DYS393* where the most common allele is present at a frequency above 80% for at least one sub-population in each case. Broad lower-lying distributions were observed for the more variable loci. For the seven loci with the highest mean gene

diversity values the frequency of the most common allele did not exceed 40% (Figure 5.1). There is currently very little data available with which to compare the results from the recently identified loci. Gene diversity values for four of these recently identified loci (DYS481, DYS557, DYS570 and DYS612) are available from the eight individuals investigated in the study by Kayser et al. (2004). For all four loci the gene diversity value was 0.857 among these eight individuals. The mean gene diversity value found for these same loci in the sample of 266 individuals described here ranged between 0.723 and 0.790.

5.3.2 Duplicated loci and intermediate alleles

Most loci analyzed in this study generated profiles consistent with a single copy on the Y chromosome. There were two instances where profiles for a specific locus suggested a duplication event. These involved DYS446 and DYS713 typed in different individuals from the Xhosa community. While DYS518 appeared to generate a profile consistent with a single-copy in all samples analyzed, duplication events have been reported for this locus for individuals from binary marker haplogroups A and E (Kayser et al. 2004).

Intermediate alleles (alleles that appeared to have at least one incomplete repeat) were observed for several loci. The same intermediate allele was observed five times for DYS607, three different DYS714 intermediate alleles were observed for three samples, and one intermediate allele was observed for DYS518. Sequencing of these alleles is required in order to identify the sequence elements responsible for generating amplicons of intermediate length. In all cases where these intermediate alleles or duplication events were observed, the PCR was repeated twice to make sure that the result was correct.

5.3.3 Effect of single loci on haplotype diversity

Table 5.2 ranked the 27 single-copy loci from highest mean gene diversity to lowest gene diversity, and shows the effect that individual loci have on the haplotype diversity. Only 73 of the 77 (94.8%) Asian Indian males could be positively identified, but this discriminatory capacity was reached with only the five most variable loci. The possibility exists that some paternally related males have been included in the investigation. This could be the reason for the fact that even with 27 loci, full discriminatory capacity is not reached in this sub-population. Only 83 of the 88 (94.3%) Xhosa males could be positively identified, but 18 loci were required to reach this

discriminatory capacity. This could be indicative of apparently low haplotype diversity for this community. Some loci (DYS557, DYS446, DYS607, DYS463 and DYS390) did not contribute to the haplotype diversity. All 101 English-speaking Caucasian males could be positively identified and this discriminatory capacity was reached with only the eight most variable loci. This may point to the genetic diversity of this group. From this data, it can be observed that combining even a few highly variable loci in a haplotype, discriminatory capacity can be substantially increased as compared to haplotype analysis with loci that are not particularly variable.

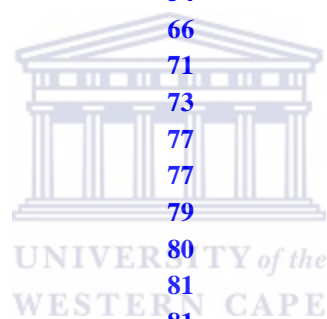
5.3.4 Stutter Analysis

The amplification of STR loci by PCR often generates ‘stutter’ products or ‘shadow’ bands. These stutter products are generally one repeat unit shorter than the primary product. It has also been reported that the degree to which a STR locus stutters depends primarily on the number of consecutively repeated homogenous units in the locus and the size of the repeated subunit (Klitschar and Wiegand 2003; Shinde et al. 2003). For example, loci with longer stretches of homogenous consecutively repeated units will tend to stutter more than those with shorter stretches of homogenous consecutively repeated units. Samples used in forensic evidence often involve mixtures of biological material. ‘Stutter’ products complicate the analysis of mixtures particularly when there is a major and minor contributor. For this reason, it is desirable to do forensic testing with STR loci that do not generate substantial stutter amplicons.

An indication of PCR stutter was obtained by analyzing the profiles for ten samples with peak heights above 400 RFUs (Table 5.3). As expected, loci with long stretches of trinucleotide repeat units generated the highest stutter peaks, while pentanucleotide loci generated the lowest stutter peaks. Optimizing the reaction conditions and reducing the cycle number may reduce PCR stutter. However, it is unlikely that PCR stutter for DYS710 and DYS711 will be reduced to a generally accepted level. Despite this, the following points should be considered before rejecting the loci. Firstly, the process of typing material from a single contributor will always be straightforward even in the presence of stutter artifacts, since both loci appear to be present as a single-copy. Secondly, when mixtures are involved, allele assignment will only be complicated if the products are separated by one or two repeat units. It is possible that the positive attributes of polymorphism and broad low-lying allele frequency distributions may outweigh the negative implications of PCR stutter for a small subset of markers in a multiplex typing system.

Table 5.2. 27 single-copy loci ranked from highest mean gene diversity to lowest gene diversity, and the effect that individual loci have on the haplotype diversity among three South African sub-populations

Locus	Asian Indian (n=77)		Xhosa (n=88)		English Caucasian (n=101)	
	# of Haplotypes	Haplotype Diversity	# of Haplotypes	Haplotype Diversity	# of Haplotypes	Haplotype Diversity
DYS518	9	0.833193	9	0.861312	9	0.800706
DYS626	37	0.958678	29	0.94344	42	0.958337
DYS714	64	0.981278	41	0.956612	72	0.981472
DYS449	71	0.984652	54	0.971849	90	0.987158
DYS713	73	0.985664	66	0.978822	98	0.989511
DYS712	73	0.985664	71	0.981921	99	0.989707
DYS570	73	0.985664	73	0.982696	100	0.989903
DYS458	73	0.985664	77	0.985279	101	0.990099
DYS557	73	0.985664	77	0.985279	101	0.990099
DYS644	73	0.985664	79	0.986054	101	0.990099
DYS614	73	0.985664	80	0.986312	101	0.990099
DYS481	73	0.985664	81	0.98657	101	0.990099
DYS446	73	0.985664	81	0.98657	101	0.990099
DYS607	73	0.985664	81	0.98657	101	0.990099
DYS635	73	0.985664	82	0.986829	101	0.990099
DYS463	73	0.985664	82	0.986829	101	0.990099
DYS390	73	0.985664	82	0.986829	101	0.990099
Y-GATA-A10	73	0.985664	83	0.987087	101	0.990099
DYS19	73	0.985664	83	0.987087	101	0.990099
DYS439	73	0.985664	83	0.987087	101	0.990099
DYS452	73	0.985664	83	0.987087	101	0.990099
DYS393	73	0.985664	83	0.987087	101	0.990099
DYS392	73	0.985664	83	0.987087	101	0.990099
DYS391	73	0.985664	83	0.987087	101	0.990099



observed for DYS714 during the study suggesting that mononucleotide tract is not particularly polymorphic.

5.3.5 Summary

A great deal of variability in autosomes can be attributed to the recombination that they undergo during meiosis. However, The NRY does not recombine during male meiosis and variability of Y-chromosomes is largely the result of mutations that have accumulated over time. As a direct result of its uni-parental pattern of inheritance, more Y-STR loci would therefore be needed to obtain a haplotype with the same discriminatory capacity as an autosome of the same size and with the same density of loci (Bosch et al. 2002). It would therefore be advantageous to amplify as many polymorphic Y-STR loci to increase the discriminatory capacity of the resultant haplotype.

It would appear that choosing the most suitable STR loci for forensic purposes might be complex. For increased power of discrimination, the main consideration should be variability. This is a characteristic achieved by having long homogenous stretches of repeated units. Unfortunately, when STR loci are chosen for variability, by taking into consideration the amount of consecutively repeated units in a locus, the likelihood of selecting loci that would generate high average percentages of stutter also increases. As variable trinucleotide loci seem to generate high average percentages of stutter, it would appear that these loci might not be suitable for forensic purposes unless this problem can be efficiently addressed.

From this data, it can be observed that combining even a few highly variable loci in a haplotype, discriminatory capacity can be substantially increased as compared to haplotype analysis with loci that are not particularly variable. It was also observed that this phenomenon is population dependant. Ultimately markers should be selected, not only on the basis of their polymorphism and physical properties but also on the extent to which they increase haplotype resolution in the populations of interest against the background of Y-STRs already in use.

Overview and future prospects

When analyzing the D values of individual MH loci, it can be observed that some of these loci show low variability among the South African sub-populations investigated. For the Xhosa and Asian Indian communities, DYS391 and DYS392 may be of limited use in forensic studies. For both the English and Afrikaner Caucasian communities, DYS393 may be of limited use in forensic studies. The other single copy MH loci have reasonable gene diversity values, making them useful when combined in a haplotype. Despite the advantage of increased discriminatory capacity, duplicated loci generate complex profiles when used in the analysis of mixtures. They are therefore not ideal for forensic studies. For these reasons it was decided to investigate other loci for properties that would be useful in forensic studies among South African sub-populations.

From the two approaches followed in this study, 22 loci were ultimately selected for such an investigation. From the available Y-chromosome sequence data and some preliminary polymorphism testing, 14 loci were selected for further investigation. These were DYS710, DYS711, DYS626, DYS712, DYS713, DYS481, DYS518, DYS570, DYS714, DYS557, DYS614, DYS612, DYS607 and DYS644. They were chosen because they typically showed a gene diversity (D) value of 0.65 and higher among the 46 English-speaking males tested. From the literature searches at the time, eight loci were selected. These were DYS439, DYS446, DYS449, DYS452, DYS458, DYS463, DYS635 (Y-GATA C4) and Y-GATA A10. They were chosen because they typically had a reported D value of ~ 0.60 and higher in a range of population studies. Multiplex reactions were considered as an efficient means to facilitate the analysis of these loci.

Three multiplex reactions (UWC Y-Plex 1, UWC Y-Plex 2 and UWC Y-Plex 3) were designed. UWC Y-Plex 1 consistently amplified six loci, (DYS19, DYS713, DYS712, DYS714, DYS626 and DYS449). UWC Y-Plex 2 consistently amplified 11 loci (DYS570, DYS710, DYS518, DYS711, DYS481, DYS612, DYS557, DYS614, DYS607, DYS446 and DYS449). UWC Y-Plex 3 consistently amplified eight loci (Y-GATA A10, DYS452, DYS644, DYS439, DYS635 (Y-GATA C4), DYS458, DYS463 and DYS449). These multiplex reactions were accepted as efficient tools with which to investigate properties of loci such as polymorphism among South African sub-populations and stutter generated by PCR amplification.

In the present study variability, stutter and copy number was investigated for 27 single-copy loci in three South African populations using dye labeled primers. An indication of the size ranges over which alleles were observed was also been presented. Samples were typed from 101 English-speaking Caucasians, 88 Xhosa individuals and 77 Asian Indians. The *D* values for the single copy loci of the MH ranged between 0.322 (DYS393) and 0.768 (DYS390), and mean stutter ranged between 4.92% (DYS19) and 11.31% (DYS392) of the primary allele. The *D* values of non-MH loci previously reported in the literature ranged between 0.53 (Y-GATA C4) and 0.86 (DYS449), and mean stutter ranged between 4.45% (DYS463) and 13.31% (DYS449) of the primary allele. The *D* values of the recently identified loci ranged between 0.57 (DYS607) and 0.92 (DYS710), and mean stutter ranged between 3.11% (DYS714) and 47.52% (DYS711) of the primary allele.

The data presented here strongly suggests that some of the MH loci are not suitable for forensic purposes among South African sub-populations. The data presented here also indicates that several forensically useful loci that may compliment the MH, or replace those MH loci with low variability, are now available. It has been observed that by combining even a few highly variable loci in a haplotype, discriminatory capacity can be substantially increased as compared to haplotype analysis with loci that are not particularly variable. It was also observed that this phenomenon is population dependant.

The loci of the ‘minimal haplotype’ (MH) are in use, partly because they have been well characterized in terms of allele nomenclature, population data and mutation rates. Future prospects for this study will involve investigations that aim to characterize the loci described here to the same extent as the MH loci. Sequencing of different sized alleles from different sub-populations should be conducted to standardize allele nomenclature. Entries of these loci into a database similar to the YHRD and standardization of the allele nomenclature by all researchers in the field may become imperative for the effective investigation of these loci. Mutation rates of these loci should also be investigated among confirmed father-son pairs of multi-generational pedigrees. Once these loci have been extensively investigated among diverse populations, constructive changes may be made to the minimal haplotype.

Three other studies have recently described an in-depth investigation of recently identified loci. The study by Kayser et al. (2004) has described the investigation of 139 polymorphic loci among eight individuals from diverse binary-marker haplogroups. A study by Lim et al. (2004) has

described the investigation of 52 loci among 75 individuals from the Y-chromosome Consortium (YCC) panel. A study by Butler et al. (2004) has described the investigation of 27 loci among three populations from the United States. The data presented here should complement the findings of comprehensive surveys such as these and prove useful in the selection of Y-STRs for refined typing systems.

Ultimately loci should be selected, not only on the basis of their polymorphism and physical properties but also on the extent to which they increase haplotype resolution in the populations of interest against the background of Y-STRs already in use.

References

Alvarez S, Soledad Mesa M, Lopez A, et al, (2002) STR data for nine Y-chromosomal loci in Guinea Equatorial (central Africa). *Forensic Sci Int* 127: 142-144

Álves C, Gusmao L, Barbosa J, et al, (2003) Evaluating the informative power of Y-STRs: a comparative study using European and new African haplotype data. *Forensic Sci Int* 134: 126-133

Ayub Q, Mohyuddin A, Qamar R, et al, (2000) Identification and characterisation of novel human Y-chromosomal microsatellites from sequence database information. *Nucl Acids Res* 28: e8

Bagwandeem D (1989) Historical Perspectives. In: Arkin AJ, Magyar KP, Pillay GJ (ed) *The Indian South Africans*. Owen Burgess Publishers, Pine Town, pp 1-22

Bender K, Farfán, Schneider P (2004) Preparation of degraded human DNA under controlled conditions. *Forensic Sci Int* 139: 135-140

Benson G (1999) Tandem Repeats Finder: a program to analyze DNA sequences. *Nucl Acids Res* 27: 573-580

Böeseken A (1989) The arrival of van Riebeeck at the Cape. In: Muller CFJ (ed) *500 Years – a history of South Africa*. Academica Publishers, Pretoria Cape Town, pp 18-34

Botha M (1972) Blood Group Gene Frequencies: an indication of the genetic constitution of population samples in Cape Town. *Am J Roentgenol Radium Ther Nucl Med* 115: Suppl: 1-27

Bosch E, Lee A, Calafell F, et al, (2002) High resolution Y chromosome typing: 19 STRs amplified in three multiplex reactions. *Forensic Sci Int* 25: 42-51

Brownstein M, Carpten J, Smith J (1996) Modulation of non-templated nucleotide addition by Taq DNA polymerase: primer modifications that facilitate genotyping. *Biotech* 20: 1004-6, 1008-1010

Buhler E (1980) A synopsis of the human Y Chromosome. *Hum Gen* 55: 145-175

Butler J, Decker A, Vallone, et al, (2004) Allele frequencies for 27 Y-STR loci with U.S. Caucasian, African American, and Hispanic samples. *Forensic Sci Int* (submitted for publication)

Butler J, Schoske R, Vallone P, et al, (2002) A novel multiplex for simultaneous amplification of 20 Y chromosome STR markers. *Forensic Sci Int* 129: 10-24

Calpin G (1949) *Indians in South Africa*. City Printing Works, Pietermaritzburg

Daniels D, Hall A, Ballantyne J (2004) SWGDAM developmental validation of a 19-locus Y-STR system for forensic casework. *J Forensic Sci* 49: 1-16

De Kock (1989) *Explorers and Circumnavigators of the Cape*. In: Muller CFJ (ed) *500 Years – a history of South Africa*. Academica Publishers, Pretoria Cape Town, pp 18-34

Dettlaff-Kakol A, Pawlowski R (2002) First Polish DNA "manhunt" - an application of Y-chromosome STRs. *Int J Legal Med* 116: 289-291

Giliomee H (2003) *The Afrikaners*. Tafelberg Publishers, Cape Town

Gill P, Brenner C, Brinkmann B, et al, (2001) DNA Commission of the International Society of Forensic Genetics: recommendations on forensic analysis using Y-chromosome STRs. *Forensic Sci Int* 124: 5-10

Grignani P, Peloso G, Fattorini P, et al, (2000) Highly informative Y-chromosomal haplotypes by the addition of three new STRs DYS437, DYS438 and DYS439. *Int J Legal Med* 114: 125-129

Gusmao L, Alves C, Beleza S, et al, (2002) Forensic evaluation and population data on the new Y-STRs DYS434, DYS437, DYS438, DYS439 and GATA A10. *Int J Legal Med* 116: 139-147

Gusmao L, Sanchez-Diz P, Alves C, et al, (2003) Results of the GEP-ISFG collaborative study on the Y chromosome STRs GATA A10, GATA C4, GATA H4, DYS437, DYS438, DYS439, DYS460 and DYS461: population data. *Forensic Sci Int* 135:150-157

Hall A, Ballantyne J (2003a) The development of an 18-locus Y-STR system for forensic casework. *Anal Bioanal Chem* 376: 1234-146

Hall A, Ballantyne J (2003b) Novel Y-STR typing strategies reveal the genetic profile of the semen donor in extended interval post-coital cervico-vaginal samples. *Forensic Sci Int* 139: 58-72

Hanson E, Ballantyne J (2004) A highly discriminating 21 locus Y-STR “megaplex” system designed to augment the minimal haplotype loci for forensic casework. *J Forensic Sci* 49: 1-12

Hedman M, Pimenoff V, Lukka M, et al, (2004) Analysis of 16 Y-STR loci in the Finnish population reveals a local reduction in the diversity of male lineages. *Forensic Sci Int* 142: 37-43

Henegariu O, Heerema N, Dlouhy S, et al, (1997) Multiplex PCR: critical parameters and step-by-step protocol. *Biotech* 23: 504-511

Hou Y, Zhang J, Li Y, et al, (2001) Allele sequences of six new Y-STR loci and haplotypes in the Chinese Han population. *Forensic Sci Int* 118: 147-152

Iida R, Tsubota E, Matsuki T (2001) Identification and characterization of two novel human polymorphic STRs on the Y-Chromosome. *Int J Legal Med* 115: 54-56

Iida R, Tsubota E, Sawazaki K, et al, (2002) Characterization and haplotype analysis of the polymorphic Y-STRs DYS443, DYS444 and DYS445 in a Japanese population. *Int J Legal Med* 116: 191-194

Jeffreys A, Wilson V, Thein S (1985) Hypervariable 'minisatellite' regions in human DNA. *Nature* 1985 314: 67-73

Jobling M (2001) Y-chromosomal SNP haplotype diversity in forensic analysis. *Forensic Sci Int* 118: 158-162

Jobling M, Tyler-Smith C (1995) Fathers and sons: the Y chromosome and human evolution. *TIGS* 11: 449-56

Kashyap V, Chattopadhyay P, Dutta R, et al, (2004) Genetic structure and affinity among eight ethnic populations of Eastern India: based on 22 polymorphic DNA loci. *Am J Hum Biol* 16: 311-327

Kayser M, Caglia A, Corach D, et al, (1997) Evaluation of Y-chromosomal STRs: a multicenter study. *Int J Legal Med* 110: 125-133, 141-149

Kayser M, Kittler R, Erler A, et al, (2004) A comprehensive survey of human Y-chromosomal microsatellites. *Am J Hum Genet* 74: 1183-1197

Kayser M, Krawczak M, Excoffier L, et al, (2001) An extensive analysis of Y-chromosomal microsatellite haplotypes in globally dispersed human populations. *Am J Hum Genet* 68: 990-1018

Klitsch M, Wiegand P (2003) Polymerase slippage in relation to the uniformity of tetrameric repeat stretches. *Forensic Sci Int* 135: 163-166

Koyama H, Iwasa M, Tsuchimoshi T, et al, (2002) Utility of Y-STR haplotype and mtDNA sequence in personal identification of human remains. *Am J Med Path* 23: 181-185

Kurihara R, Yamamoto T, Uchihi R, et al, (2004) Mutations in 14 Y-STR loci among Japanese father-son haplotypes. *Int J Legal Med* 118: 125-131

Lahiri D, Nurnberger J (1991) A rapid non-enzymatic method for the preparation of HMW DNA from blood for RFLP studies. *Nucl Acids Res* 19: 5444

Lane A, Soodyall H, Arndt S, et al, (2002) Genetic Substructure in South African Bantu Speakers: Evidence from autosomal DNA and Y-chromosome studies. *Am J Phys Anthro* 119: 175-185

Leat N, Benjeddou M, Davison S (2004) Nine-locus Y-chromosome STR profiling of Caucasian and Xhosa populations from Cape Town, South Africa. *Forensic Sci Int* 144: 73-75

Lee H, Ladd C (2001) Preservation and collection of biological evidence. *Croat Med J* 42: 225-228

Lee H, Oh J, Han G, et al, (2003) Allele frequencies and haplotypes of six new Y-specific STR loci in Koreans. *Forensic Sci Int* 136: 89-91

Lim S, Xue Y, Tyler-Smith C (2004) Evaluation of 52 novel Y-STRs for forensic and population-genetic studies. IV International Forensic Y-User Workshop

Magnuson V, Ally D, Nylund S, et al, (1996) Substrate nucleotide-determined non-templated addition of adenine by Taq DNA polymerase: implications for PCR-based genotyping and cloning. *Biotech* 21: 700-709

Mohyuddin A, Ayub Q, Qamar R, et al, (2001) Y-chromosomal STR haplotypes in Pakistani populations. *Forensic Sci Int* 118: 141-146

Moxon E, Wills C (1999) DNA microsatellites: agents of evolution? *Sci Am* 280: 72-77

Nei M (1987) *Molecular Evolutionary Genetics*. Columbia University Press, New York

Ploski R, Wozniak M, Pawlowski R, et al, (2002) Homogeneity and distinctiveness of Polish paternal lineages revealed by Y chromosome microsatellite haplotype analysis. *Hum Gen* 110: 592-600

Qamar R, Ayub Q, Mohyuddin A, et al, (2002) Y-chromosomal DNA variation in Pakistan. *Am J Hum Genet* 70: 1107-1124

Quintana-Murci L, Fellous M (2001) The Human Y Chromosome: The Biological Role of a 'Functional Wasteland'. *J Biomed Biotech* 1: 18-24

Quintana-Murci L, Krausz C, McElreavey K (2001) The human Y chromosome: function, evolution and disease. *Forensic Sci Int* 118: 169-181

Ramana GV, Su B, Jin L, et al, (2001) Y-chromosome SNP haplotypes suggest evidence of gene flow among caste, tribe, and the migrant Siddi populations of Andhra Pradesh, South India. *Eur J Hum Genet* 9: 695-700

Redd A, Agellon A, Kearney V, et al, (2002a) Forensic value of 14 novel STRs on the human Y chromosome. *Forensic Sci Int* 130: 97-111

Redd A, Roberts-Thomson J, Karafet T, et al, (2002b) Gene flow from the Indian subcontinent to Australia: evidence from the Y chromosome. *Curr Biol* 12: 673-677

Roewer (2004) Male DNA fingerprints say more. *Profiles in DNA* 7: 14-15

Roewer L, Arnemann J, Spurr N, et al, (1992) Simple repeat sequences on the human Y chromosome are equally polymorphic as their autosomal counterparts. *Hum Gen* 89: 389-394

Roewer L, Krawczak M, Willuweit S, et al, (2001) Online reference database of European Y-chromosomal short tandem repeat (STR) haplotypes. *Forensic Sci Int* 118: 106-113

Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. *Methods Mol Biol* 132: 365-86

Sajantila A, Salem AH, Savolainen P, et al, (1996) Paternal and maternal DNA lineages reveal a bottleneck in the founding of the Finnish population. *Proc Natl Acad Sci USA* 93: 12035-12039

Sahoo S, Chainy G, Kashyap V (2003) Allele frequency of eight Y-Chromosome STR loci in Oriya population of India. *J Forensic Sci* 48: 245-252

Sambrook J, Fritsch E, Maniatis T (1989) Detection and analysis of proteins expressed from cloned genes. In: Molecular cloning – a laboratory manual. Cold Spring Harbor Laboratory Press, New York, pp18.56-18.57

Schoske R, Vallone P, Kline M, et al, (2004) High-throughput Y-STR typing of U.S. populations with 27 regions of the Y chromosome using two multiplex PCR assays. Forensic Sci Int 139: 107-121

Schoske R, Vallone P, Ruitberg C, et al, (2003) Multiplex PCR design strategy used for the simultaneous amplification of 10 Y chromosome short tandem repeat (STR) loci. Anal Bioanalytical Chem 375: 333-343

Shewale J, Nasir H, Schneida E, et al, (2004) Y-chromosome STR system, Y-PLEX™ 12, for forensic casework: development and validation. J Forensic Sci 49: 1-13

Shinde D, Lai Y, Sun F, et al, (2003) Taq DNA polymerase slippage mutation rates measured by PCR and quasi-likelihood analysis: (CA/GT)_n and (A/T)_n microsatellites. Nucl Acids Res 31: 974-980

Sibille I, Duverneuil C, Lorin de la Grandmaison G, et al, (2002) Y-STR DNA amplification as biological evidence in sexually assaulted female victims with no cytological detection of spermatozoa. Forensic Sci Int 125: 212-216

Vallone P, Butler J (2004) AutoDimer: a screening tool for primer-dimer and hairpin structures. Biotech 37: 226-231

Van Bruwer (1964) Onstaansgeskiedenis. In: Theron E (ed) Die Kleurling bevolking van Suid Afrika. Citadel Press, Cape Town, pp 1-7

Wallin J, Holt C, Lazaruk K, et al, (2002) Constructing universal multiplex PCR systems for comparative genotyping. J Forensic Sci 47: 52-65

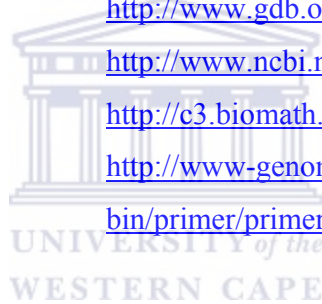
White P, Tatum O, Deaven L, et al, (1999) New, male-specific microsatellite markers from the human Y chromosome. Genom 57: 433-437

Willot G, Allard J (1982) Spermatozoa – their persistence after sexual intercourse. *Forensic Sci Int* 19: 135-154

Zhu Q, Tang J, Gao Y, et al, (2003) Distributions of allelic frequencies and haplotypes of two novel Y-chromosome STR in a Chinese population. *J Forensic Sci* 48: 457

Electronic Supplementary Resources

South African Police Service: <http://www.saps.org.za>
Rape Crisis: <http://www.rapecrisis.org.za>
Census South Africa: <http://www.census.gov.za>
YHRD: <http://www.yhrd.org>
International Forensic Y-User Group: <http://ystr.charite.de>
Genome Database (GDB): <http://www.gdb.org>
BLAST: <http://www.ncbi.nlm.nih.gov>
Tandem Repeat Finder: <http://c3.biomath.mssm.edu/trf.html>
Primer3: http://www-genome.wi.mit.edu/cgi-bin/primer/primer3_www.cgi



Appendix

1 DNA Extraction, Quantification and Working Stock Solutions

1.1 Reagents (To be autoclaved immediately after preparation)

TKM I (100ml)

1M Tris, pH 8.0 (<i>Merck Laboratory Supplies</i>)	1ml
100mM KCl (<i>Merck Laboratory Supplies</i>)	10ml
200mM MgCl ₂ (<i>Merck Laboratory Supplies</i>)	5ml
100mM EDTA (<i>Merck Laboratory Supplies</i>)	2ml
Distilled H ₂ O	82ml

TKM I + Nonidet P40 (100ml)

1M Tris, pH 8.0 (<i>Merck Laboratory Supplies</i>)	1ml
100mM KCl (<i>Merck Laboratory Supplies</i>)	10ml
200mM MgCl ₂ (<i>Merck Laboratory Supplies</i>)	5ml
100mM EDTA (<i>Merck Laboratory Supplies</i>)	2ml
Distilled H ₂ O	82ml
Nonidet P40 (<i>Sigma</i>)	2.25ml

TKM II (100ml)

1M Tris, pH 8.0 (<i>Merck Laboratory Supplies</i>)	1ml
100mM KCl (<i>Merck Laboratory Supplies</i>)	10ml
200mM MgCl ₂ (<i>Merck Laboratory Supplies</i>)	5ml
100mM EDTA (<i>Merck Laboratory Supplies</i>)	2ml
2M NaCl (<i>Merck Laboratory Supplies</i>)	20ml
Distilled H ₂ O	62ml

10% w/v SDS (100ml)

SDS (<i>Merck Laboratory Supplies</i>)	10g
Distilled H ₂ O	100ml

2M NaCl (100ml)

NaCl (<i>Merck Laboratory Supplies</i>)	11.67g
Distilled H ₂ O	100ml

1.2 Procedure

1. 0.5ml of carefully mixed whole blood was transferred to a clean, dry, APPROPRIATELY LABELLED 1.5ml microfuge tube.
2. 0.5ml of sterile [TKM I + Nonidet P-40] was added to the 0.5ml blood and the contents of the tube mixed gently by inversion.
3. The tube was then placed in a bench-top centrifuge (*Eppendorf, 5415 D*) and the mixture centrifuged at 5000 rpm for 10 minutes to pellet the nuclei.
4. The supernatant was then carefully removed, paying attention to not disturb the pellet.
5. The pellet was then washed by adding 0.5ml sterile TKM I to it.
6. The tube was centrifuged for 10 minutes at 5000 rpm.
7. The supernatant was removed again. This time, extra care was taken to remove as much of it as possible, without disturbing the pellet.
8. 70µl of sterile TKM II was added to the pellets and tube vortexed (*Stuart Scientific, vortex mixer SA3*) until the pellet was completely re-suspended in the liquid.
9. 4.37µl of sterile 10% w/v SDS was added to this mix and the tube vortexed briefly.
10. The tube was then incubated in a water-bath (*Memmert*) at 55°C for 10 minutes.
11. After incubation, 264µl of sterile 2M NaCl was added and the tube vortexed briefly.
12. The tube was then centrifuged at 13000 rpm for five minutes to pellet extra-cellular components.
13. After centrifugation, the supernatant was transferred to a clean, dry and CORRECTLY LABELLED, 1.5ml microfuge tube.
14. To this supernatant, 677µl of absolute ethanol (room temperature) was added and the contents of the tube mixed gently by inverting it a few times. At this point the DNA became visible.
15. The tube was then centrifuged at 13000 rpm for five minutes to pellet the DNA. (The DNA did not always end up at the bottom of the tube, but sometimes got stuck on the side of the tube).
16. The supernatant was then removed and the tube centrifuged again at 13000 rpm.

17. The supernatant was removed after centrifugation and 250µl of ice-cold 70% ethanol added to wash the DNA.
18. The tube was then centrifuged at 13000 rpm for five minutes and the supernatant removed.
19. Step 17 and 18 were repeated.
21. The DNA was dried at room temperature, re-suspended in 100µl of sterile distilled water and the DNA allowed to go into solution at room temperature overnight.

All the blood waste from the DNA extraction process were decanted into appropriate containers, sealed properly, clearly labelled as blood waste and discarded in an appropriate manner by *Waste Tech*, an accredited waste removal company.

1.3 Quantifying DNA and working stock dilutions

1. 5µl of each of the extracted DNA aliquots were diluted in 100µl of sterile distilled water.
2. Each of the dilutions was then placed in a quartz cuvette in a spectrophotometer (*Milton Roy, Genesis 5*) and the absorbance readings taken over a range of wavelengths from 240nm to 300nm.
3. The DNA concentration was calculated from the absorbance readings at 260nm.
4. Working stock dilutions at 2ng/µl were made from all the DNA samples.
5. The original DNA stocks as well as working stock solutions were stored at -20°C

2. Silver stain analysis

2.1 Reagents

Binding Solution (1ml)

Binding Saline (<i>Promega</i>)	2.5µl
Ethanol (<i>Merck Laboratory Supplies</i>)	50µl
Glacial Acetic Acid (<i>Merck Laboratory Supplies</i>)	950µl

10X TBE Buffer (2L)

Tris (<i>Merck Laboratory Supplies</i>)	216g
Boric Acid (<i>Merck Laboratory Supplies</i>)	110g
EDTA (<i>Merck Laboratory Supplies</i>)	14.88g
Distilled H ₂ O	up to 2L

4% Polyacrylamide Gel Mix (80ml)

Urea (<i>Merck Laboratory Supplies</i>)	36g
40% 19:1 Polyacrylamide solution (<i>Promega</i>)	8ml
10X TBE Buffer	4ml
Distilled H ₂ O	up to 80ml

2X Loading Buffer (10ml)

Bromophenol Blue (<i>Merck Laboratory Supplies</i>)	0.05g
Xylene Cyanol (<i>Merck Laboratory Supplies</i>)	0.05g
Ethanol (<i>Merck Laboratory Supplies</i>)	9.5ml
NaOH	0.1mmol

Silver stain Solution (2L)

AgNO ₃ (<i>Merck Laboratory Supplies</i>)	2g
Distilled H ₂ O	2L

Developing Solution (2L)

NaOH Pellets (<i>Merck Laboratory Supplies</i>)	30g
15% Formaldehyde (<i>Merck Laboratory Supplies</i>)	20ml
Distilled H ₂ O	up to 2L

2.2 Preparation of gel running plates

In order to perform PAGE, glass plates had to be prepared in such a way that the polyacrylamide gel would remain bound to the one plate and separate easily from the other plate. This was done so that staining could be performed easily after PAGE, with the gel bound to the one glass plate. Care was always taken that the plates were absolutely free from any dirt. The glass plate to which the gel was bound was cleaned three times with ethanol after which 1ml of the binding solution

was applied to the entire plate surface and left on the plate for five minutes. After the five minutes waiting period, the excess of binding solution was cleaned from the plate approximately 4 to 5 times using ethanol. The other glass plate was also cleaned with ethanol and then *See Thru*, a commercially available reagent used to repel rain from windscreens, was applied according to the manufacturer's instructions. The plates were then clamped together with the treated sides on the inside using 0.2mm spacers. 75ml of 4% polyacrylamide gel solution was filtered through a 0.45micron filter prior to use. 500µl of 10% APS and 50µl of TEMED was added to this gel solution, mixed well and the gel poured using a syringe. The comb was then inserted and clamped in between the glass plates. The gels were normally fully polymerized after approximately 45 minutes.

2.3 Fragment Analysis

After amplification polyacrylamide gel electrophoresis (PAGE) was performed on the PCR products. 3µl of the PCR product was thoroughly mixed with 3µl of 2X loading buffer. This mixture was then heat-denatured at 95°C for 3 minutes and immediately snap-cooled on ice until loading took place. On each gel, one sample was loaded at three positions across the gel to assist with allele identification of slightly distorted gels. The gels were run in 0.5X TBE buffer and allowed to reach a temperature of at least 40°C before samples were loaded. In order to facilitate easy loading of the gels, it was imperative that the wells were rinsed from excess urea and possible gel pieces before loading. 3-5µl of the PCR product-loading dye mixture was loaded on the 4% polyacrylamide gel and electrophoresis performed.

Following PAGE, the gels were silver-stained using a method adapted from Sambrook et al. (1989). Plates were separated and the glass plate, to which the polyacrylamide gel was chemically bound by the binding solution, was rinsed with distilled water to remove the excess running buffer. The gel was then soaked in a 0.1% w/v AgNO₃ solution for 10–15 minutes. The gel was rinsed in distilled water again to remove excess AgNO₃ solution and placed into a developing solution containing 1.5% w/v NaOH and 0.15% v/v formaldehyde. The use of NaOH facilitated the separation of gels from plates after staining. This was done until the bands started to appear and the edges of the gel started to separate from the glass plates. This last soak generally lasted about 15 minutes. The gel was rinsed in distilled water, detached from the glass plate with chromatography paper (*Whatmann 3MM*) and air-dried.