

Y-STR studies of genetic genealogy, population and forensic genetics of Indian and Zulu groups in the Durban, KwaZulu-Natal area of South Africa.

by

SURINA SINGH



Submitted in fulfilment of the academic requirements for the degree of
Master of Science Biological Sciences (Forensic Genetics)
in the School of Life Sciences,
College of Agriculture, Engineering and Science,
University of KwaZulu-Natal (Westville Campus),
Durban, South Africa

June 2018

Supervisor: PROF. JENNY LAMB

As the candidate's supervisor I have approved this thesis/dissertation for submission.

Signed: 

Name: Jennifer M Lamb

Date: 02/07/2018

ABSTRACT

The combination of molecular genetics and surname analysis of short tandem repeat (STR) data has the potential to shed light on population structure and history, falling within the field of forensic deoxyribonucleic acid (DNA) analysis. Since the Y-chromosome DNA along with surnames are paternally inherited, non-related males sharing a surname should be more closely related in comparison to the general population. Currently, no surname studies based on the Indian population in South Africa exist. This study aimed to explore the genetic genealogy, population and forensic genetics of Indian (different geographic origin, religion and language) and Zulu males with different common surnames from Durban, KwaZulu-Natal. This was achieved by: (1) Collecting samples from 224 non-paternal lineage related North Indians males and generating DNA profiles, using the Yfiler® Plus kit to amplify 27 Y-chromosome STR (Y-STR) loci; (2) Comparing the genetics of the North Indian group to that of other groups with South Indian and Zulu African surnames found in the forensic lab database. Hypotheses were formulated to analyse differences in relationships at ethnic, region, religion, language and surname-based levels (Figure 1). Population and forensic genetic analyses revealed that the Yfiler® Plus gave a higher number of unique haplotypes and discrimination capacity and a lower haplotype match probability, validating its use in this study. Genetic structure was found amongst examined sub-groupings. AMOVA was significant for all levels tested, with exception to between South Indian surnames. There are no known barriers to intermarriage among people bearing these South Indian surnames. Structure and PCoA analysis showed the presence of two significant sub-populations, which were ethnic based. Population structure and diversity were not surname based, but rather at an ethnic level. This could be attributed to polyphyletic origin (many surname origin) of the analysed surnames. Surname transmission was polyphyletic for all surname groups, showing overlapping haplotypes and clades, implying multiple founders/ lineages for each specific surname investigated. The data generated in this study will contribute to the Indian DNA profiling database and could potentially serve as a baseline for further research. Further research could include sequencing autosomal STRs and hypervariable regions of mtDNA.

EXTENDED ABSTRACT

Genetic genealogy is a field of growing interest, involving genealogical testing to determine genetic relationships between individuals. The combination of molecular genetics and surname analysis of Y-STR data has the potential to shed light on population structure and history and is within the field of HID (human Identification) forensic DNA analysis. Since DNA along with surnames are passed down from our ancestors, people with the same surname should have a greater chance of sharing common ancestry when compared with the general population. There are currently no surname studies based on the Indian or Zulu populations of SA. The primary focus of this study was on genetic genealogy, i.e. co-inheritance of surnames and Y-STRs. In addition, population and forensic genetics of a sample group of mainly Indian and Zulu people from the greater Durban area of KwaZulu-Natal (KZN), SA, was investigated to search for genetic structure in comparisons of groups based on (1) Ethnicity (Indian vs Zulu), (2) Region of origin in India (North vs South), (3) Religion (Hindu vs Muslim), (4) Language (Hindi vs Muslim (Urdu) vs Tamil Indians) and, (5) Surname-based groups originating from North India (surnames Khan, Maharaj and Singh), South India (surnames Govender, Naidoo and Pillay), and Africa (surnames Buthelezi, Cele, Dlamini, Mkhize and Zulu). Further, an attempt was made to establish baseline aspects of the social history of the North Indian groups in Durban, which relate to genetic genealogy. The age profile and number of generations since the first family member arrived in the Durban area from India were investigated, along with information on whether a North Indian individual shares his surname, city, religion and language with his child and paternal and maternal forefathers.

DNA samples were collected from 224 non-parentally related North Indian males with the surnames Khan, Maharaj and Singh and the Yfiler® Plus Polymerase chain reaction (PCR) amplification kit was used to amplify 27 Y-STR loci, from which DNA profiles were generated. In order to extend the basis for comparison to other Durban area ethnic groups, Y-STR profiles, from the lab database, of South Indians (n = 90, surnames Govender, Naidoo and Pillay) and Zulus (n = 100, surnames Buthelezi, Cele, Dlamini, Mkhize and Zulu) were included in the sample set. Null Alleles were observed at 77.8 % (21 out of 27) of the different loci analysed, with most contained in loci DYS391, DYS389II and DYS448.

Population and forensic genetic analyses were used to compare four Y-STR marker sets (1) Minimal Haplotype (MHT), (2) Yfiler®, (3) Yfiler® Plus markers, and (4) Rapidly Mutating (RMu) and assess their suitability for forensic investigation in the overall sampled population. The Haplotype diversity (HD) for all marker sets was 0.999. The smaller RMu marker set (7 loci) had the highest mean genetic diversity (GD) per locus (0.82), however, the use of an increased number of marker loci (which include these 7 loci), as exemplified by the Yfiler® Plus kit (27 loci), resulted in a higher number of unique haplotypes, a higher discrimination capacity (DC) and a lower haplotype match probability (MP), validating the decision to use the Yfiler® Plus kit in this study.

One of the aims of this study is to search for the existence of genetically structured sub-groups within different sample groups, based on the following analyses: (1) Analysis of Molecular Variance (AMOVA) was used to estimate the extent of population differentiation and its significance; (2) Bayesian Analysis of Population Structure (BAPS) was carried out to estimate the number of genetic clusters in a sample group and the percentage membership of each sample member in each identified cluster; (3) Principal Co-ordinates Analysis (PCoA) was carried out to visualise genetic distance and relatedness among sample groups; and Haplotype networks were created to show mutational relationships among haplotypes within a sample set.

A relatively moderate level of genetic separation between Indian and Zulu groups was observed in all genetic structure-based analyses. In comparisons using AMOVA, 7% of the variance ($\Phi_{IPT}=0.074$, $P = 0.0001$) occurred between the two ethnic groups. This was reflected in PCoA analyses as a high degree of genetic separation between Indians and Zulus, which also formed separate groups in Bayesian analyses of population structure and to some extent in haplotype network analyses. This difference was postulated to have developed in the years following separation of the two groups by migration of the ancestors of the Indian populations out of Africa into Asia. The introduction of indentured labourers from India to the Durban region resulted in these groups again occupying the same geographic region, which led to varying degrees of interbreeding between the two groups. However, the relatively short time since the arrival of the Indians in Durban (~ 160 years), combined with cultural and legal barriers to interbreeding, has resulted in the maintenance of clearly detectable genetic structure among the two groups.

The distance separating the sites of origin of the North and South Indian samples, combined with language and cultural differences, were reflected in genetic structure among the groups of North and South Indian origin. Differences between North and South Indians were observed in all genetic structure-based analyses; AMOVA indicated that 3% of the variance occurred among North and South Indian groups ($\Phi_{IPT} = 0.029$, $P < 0.005$) whilst PCoA, Bayesian Analysis of population structure and haplotype network analysis showed a level of separation between them. Geographical proximity since the original indentured Indian labourers arrived and the consequent opportunity for the North and South Indian groups to interbreed is likely to have reduced the levels of genetic structure among them, although it is still detectible.

Genetic structure was observed between Hindu and Muslim sample members, although at a very low level. AMOVA revealed that a significant, but low 1% of the variance ($\Phi_{IPT} = 0.009$, $P = 0.0010$), occurred among sample groups.

In language-based comparisons, AMOVA revealed that a relatively low 3% of the variance ($P < 0.005$) occurred between Tamil and both Hindi and Muslim sample members, and an even lower 1% (0.001%) between Hindi and Muslim sample members. These differences were also reflected in PCoA and haplotype network analyses. As hypothesized, the existence of genetic structure based on language was supported for all comparisons (Tamil vs Hindi, Tamil vs Muslim (originally Urdu), and Hindi vs Muslim (originally Urdu), although this may have been confounded, in some cases, by region-based differences (North vs South Indians).

The Y-chromosome and surnames are paternally inherited in both North and South Indians and Zulus. As the Y-chromosome has a relatively low mutation rate per generation, it could be hypothesised that groups of people with a particular surname would be more closely related to each other than to groups with other surnames. Genetic divergence over time among surname groups based on religious and or cultural practices or region of origin are likely to be reflected in genetic divergence among surnames. AMOVA revealed that there is a broad level of genetic structure attributable to surname-based groupings, as eight percent of the variance occurred amongst all 11 surname-based groups (Khan, Maharaj, Singh, Govender; Naidoo, Pillay; Buthelezi, Cele, Dlamini, Mkhize, Zulu).

In AMOVA, 3% of the variance occurred among North Indian surname groups, with PhiPT values being low but significant. The highest PhiPT values were associated with comparisons of the Maharaj group with the Singh and Khan groups respectively. PCoA, BAPS and haplotype network analysis supported this pattern. The surname Maharaj (Brahmin caste, Hindu) appeared most distinct, possibly due to divergence based on intermarriage barriers based on caste (with Singhs) and religion (with Khans, who are Muslim).

Little genetic structure was observed amongst the South Indian Tamil surnames, Govender, Naidoo and Pillay. There are no known barriers to intermarriage among people bearing these surnames, making it unlikely that they would have diverged from one another through time, and that this would be reflected in Y-STR based genetic structure.

AMOVA revealed that 9% percent of the variance occurred among the Zulu surname groups Buthelezi, Cele, Dlamini, Mkhize and Zulu. All pairs of surname-based groups were significantly different from each other and PCoA revealed some separation of groups with the surnames Cele, Dlamini and Mkhize. Bayesian Analysis of Population Structure supported this and showed the surnames Cele and Dlamini to be particularly distinct. Consistent with this, the haplotype analysis shows one cluster consisting of only sample members with the surname Cele. Traditionally, Africans are not allowed to marry within the same in order to prevent inbreeding. Creation of such intramarriage barriers is likely to lead to high levels of diversity and low levels of genetic structure among surname-based groups. In contrast to this expectation, levels of genetic structure among Zulu surname-based groups in this study were considerably higher than those found among North Indian or South Indian surname-based groups.

One of the aims of this study was to determine the mode of inheritance of surnames, viz. whether inheritance is monophyletic or polyphyletic, whether the surname haplotypes overlap or are non-overlapping, and whether paternal transmission of surnames occurs with high or low fidelity (Jobling, 2001).

In this study, surname transmission was found to be polyphyletic for all three sets of surname groups (North Indian, South Indian and Zulu), and surname groups showed overlapping haplotypes and

clades. This implies that multiple genetically different ancestors may have founded different lineages of each specific surname investigated. Low fidelity surname transmission could also have resulted in surnames being part of multiple clades in a haplotype network, and single clades containing multiple surnames, as was observed in this study. This could happen in the case of adopted children, where the child is given the name of the adoptive father, but does not carry his Y-chromosome, or in the case of maternal transmission of surnames. The social data collected as part of this study revealed that, for the North Indian surname group, one third (Khan and Singh) to two thirds (Maharaj) of the respondents indicated that their surnames were derived from their maternal forefathers, consistent with disturbances in surname transmission. Low fidelity surname transmission which may relate to circumstances surrounding the importation of indentured labourers from India to what was then known as Natal. Anecdotal evidence suggests that surnames may have been incorrectly recorded during migration from India to SA, and that officials processing the new arrivals in Natal could not spell or pronounce some of the Indian surnames, leading them to use shortened or misspelled versions of them, or even used first names as surnames.

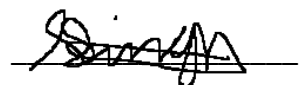
None of study samples shared haplotypes with those on the Y-chromosome Haplotype Reference Database (YHRD). Some study samples were positioned separately on the Multi-Dimensional Scaling (MDS) plot, whereas others formed groups with samples from the YHRD, indicative of common genetic origins. Overall, the Indian study samples appeared to have a South Asian origin, although the Maharaj surname appeared to be positioned as close to the European samples as to the other Asian samples, possibly indicative of a west Asian genetic origin. The positioning of the Zulu samples appeared to indicate shared origins with samples from Kenya and with Bantu Luhya samples.

Key words: Genetic genealogy, genetic structure, North and South Indian, population and forensic genetics, Y-STRs, Zulu.

PREFACE

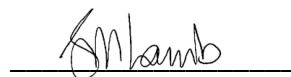
Research and lab work were conducted under the standard ethics procedures of the UKZN Biomedical Research Ethics (BE456/16, sub-study of BCA056/16). Data is being held according to the stipulations of the permit. ISFG (International Society for Forensic Genetics) and the SWGDAM (Scientific Working Group on DNA Analysis Methods) recommendations have been followed.

The work described in this thesis was carried out at the School of Life Science, University of KwaZulu-Natal (Westville campus), under the supervision of Prof. Jenny Lamb. This study represents original work by the researcher, except where the work of others is acknowledged in the text, and no part of this work has been submitted in any form to another university.



Surina Singh (MSc student)

Date: 13/11/2017



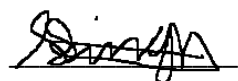
Prof. Jenny Lamb (Supervisor)

Date: 02/07/2018

PLAGIARISM DECLARATION

I, Surina Singh, declare that

- (i) The research reported in this dissertation, except where otherwise indicated or acknowledged, is my original work;
- (ii) This dissertation has not been submitted in full or in part for any degree or examination to any other university;
- (iii) This dissertation does not contain other persons' data, pictures, graphs or other information, unless specifically acknowledged as being sourced from other persons;
- (iv) This dissertation does not contain other persons' writing, unless specifically acknowledged as being sourced from other researchers. Where other written sources have been quoted, then:
 - a) their words have been re-written but the general information attributed to them has been referenced;
 - b) where their exact words have been used, their writing has been placed inside quotation marks, and referenced;
- (v) Where I have used material for which publications followed, I have indicated in detail my role in the work;
- (vi) This dissertation is primarily a collection of material, prepared by myself, published as journal articles or presented as a poster and oral presentations at conferences. In some cases, additional material has been included;
- (vii) This dissertation does not contain text, graphics or tables copied and pasted from the Internet, unless specifically acknowledged, and the source being detailed in the dissertation and in the References sections.



Signed: Surina Singh

Date: 20/06/2018

ACKNOWLEDGEMENTS

I would like to acknowledge and thank the following:

The Almighty Lord, who I believe without whom the completion of this study could have not been possible. I would like to thank God, in whatever form he may be, for giving me the strength, knowledge, mental and physical health, capability and opportunity, to undertake and complete this research project.

Prof. Jenny Lamb, my supervisor, for her overall academic assistance and guidance, and the long hours spent to ensure that the work produced is of high standards. This study would have not been conducted as robustly without her proficient knowledge in the Forensic Genetics field. Being regarded as her 'final project' as a postgraduate student, I hope I have done her proud. Truly, I could have not asked for a better supervisor and mentor than her.

Velosha Naidoo and Chelsea Mkhize for allowing the use of their profiled Tamil and Zulu surnames respectively, for comparative analyses. Seshvir Pooran for lab assistance.

Family and friends for their overall support and assistance in finding non-related male volunteers to participate in study i.e. Anil Singh, Shamin Maharaj, Abja Maharaj, Sudika Singh-Reinesch, Shahir Rajcomar, Avant Samdhan, Kaveshin Govender, Seshvir Pooran, Ruven Pillay, Nirvedh Bhole, Mr Farook Khan, Suhail Khan, Anas Hamidi, Kishen Juguth, Nevaan Singh, Serisha Puncham, Seresha Naidoo and Munchkin-Wala Singh.

NRF for financial support. Radio Al-Ansar for allowing me the opportunity to speak about my project over-air. Facebook and google surveys as a platform to find volunteers who were willing to participate, and lastly all the male volunteers who participated in this study.

ABBREVIATIONS

AMOVA	Analysis of Molecular Variance
AZFc	Azoospermia Factor c
BAPS	Bayesian Analysis of Population Genetic Structure
DC	Discrimination capacity
DNA	Deoxyribonucleic acid
FST	Fixation index
GD	Genetic diversity
HD	Haplotype diversity
HID	Human identification
ISFG	International Society for Forensic Genetics
KZN	KwaZulu-Natal
Mb	Megabytes
MDS	Multi-Dimensional Scaling
MHt	Minimal Haplotype
MJ	Median-joining
MP	Haplotype match probability
MSY	Male-specific region
mtDNA	Mitochondrial DNA
RRRY	Non-recombining region of the Y-chromosome
PAR	Pseudoautosomal region
PCoA	Principal Co-ordinates Analysis
PCR	Polymerase chain reaction
RMu	Rapidly Mutating
SA	South Africa
SNPs	Single nucleotide polymorphisms
SRY	Sex-determining region of the Y-chromosome
STRs	Short tandem repeats
SWGDAM	Scientific Working Group on DNA Analysis Methods
TMRCA	Time to Most Recent Common Ancestry
vs	Versus
YHRD	Y-chromosome Haplotype Reference Database
Y-STRs	Y-chromosome short tandem repeats

TABLE OF CONTENTS

ABSTRACT	I
EXTENDED ABSTRACT	II
PREFACE	VII
PLAGIARISM DECLARATION	VIII
ACKNOWLEDGEMENTS.....	IX
ABBREVIATIONS.....	X
TABLE OF CONTENTS	XI
1. List of Figures	XIII
2. List of Tables.....	XIV
INTRODUCTION.....	1
1. Background.....	1
2. History of South African Indians	3
2.1. Movement from India to SA.....	3
2.2. Indian Caste System	6
2.3. Marriage practices in Indians and Zulus.....	7
3. Surname studies.....	9
4. Origin and Frequency of Target surnames	11
4.1. Surnames of Indians transported to Natal.....	12
4.2. Zulu Surnames.....	14
5. Patterns of transmission of surnames and haplotypes/haplogroups.....	15
6. The Y-chromosome	17
6.1. The importance of Y-chromosome testing in forensics	18
6.2. Polymorphic Y-chromosome markers.....	21
7. YHRD (Y-chromosome Haplotype Reference Database).....	26
8. Purpose of this research	27
9. Aims, Objectives and hypotheses	28
9.1. Utility of different Y-STR marker in population and forensic genetics analyses.....	30
9.2. Genetic structure analyses based on population sub-groupings.....	30
9.3. Surname-based genetic analyses.....	32

9.4.	Population and forensic genetics based on population sub-groupings	34
9.5.	Comparison of experimental samples with samples and populations on the YHRD	34
MATERIALS AND METHODS.....		35
1.	Sampling	35
1.1.	Sample composition	36
1.2.	Comparison with other available databases	37
2.	Sampling methodology	38
2.1.	Sample collection	38
3.	DNA Profiling	39
3.1.	DNA extraction	39
3.2.	DNA Quantification	39
3.3.	Amplification of Y-STRs	40
3.4.	Capillary electrophoresis.....	41
4.	Genetic Analyses	42
4.1.	Genetic Structure	42
4.2.	Genetic diversity and Forensic parameters	44
4.3.	Comparison of experimental samples with samples and populations on the YHRD	44
5.	Social aspects: Lineage inheritance in North Indian surname groups	45
RESULTS		46
1.	Comparison of Y-STR marker sets via population and forensic genetics analyses.....	47
2.	Genetic structure analyses based on population sub-groupings	49
2.1.	Genetic Structure based on population sub-groupings	49
2.2.	Population and forensic genetics based on population sub-groupings.....	55
3.	Genetic genealogy: Surname-based genetic analyses	57
3.1.	Social aspects: Lineage inheritance in North Indian surname groups	57
3.2.	Surname-based genetic structure and surname inheritance analyses.....	62
3.3.	Surname-based population and forensic genetics analyses	69
4.	Comparisons of study samples with samples found on the Y-chromosome STR Haplotype Reference Database (YHRD)	71
DISCUSSION		75
1.	Comparison of Y-STR marker sets via population and forensic genetics analyses.....	76
2.	Genetic structure analyses based on population sub-groupings	78
2.1.	Genetic structure based on ethnicity (Indian vs Zulu)	78

2.2.	Genetic structure based on region of origin in India (North vs South India)	81
2.3.	Genetic structure based on religious groups	83
2.4.	Genetic structure based on language	84
3.	Genetic genealogy: Surname-based genetic analyses	86
3.1.	Surname-based genetic structure	86
3.2.	Surname Inheritance	89
4.	Population and forensic genetics	91
5.	Comparison of experimental samples with samples and populations on the YHRD	93
6.	Challenges and shortcomings	95
7.	Recommendations for future research	96
8.	Conclusion	98
	REFERENCES	101
	APPENDIX	111

1. List of Figures

Figure 1: Concept Map of the study setup.	2
Figure 2: Indian migration	5
Figure 3: Indian settlement in Durban	5
Figure 4: Comparison of the older Hindu and Muslim caste systems with the modern day system.	7
Figure 5: Possible relationships between Y-chromosomal haplotypes and surname transmission.	16
Figure 6: Y-chromosome structure.	18
Figure 7: SA's reported rape case trend over a nine year period	20
Figure 8: Fragment sizes and fluorescent dyes used in the amplification of 27 Y-STR loci by the Yfiler® Plus PCR amplification kit.	25
Figure 9: Sampling location map.	35
Figure 10: Electrophoresis plate setup example.	42
Figure 11: GD per locus, for 27 loci and four different marker sets (N=409)	47

Figure 12: PCoA plot visualising genetic distance and relatedness amongst different sample sub-groupings.	51
Figure 13: Bayesian analysis of genetic population structure within the entire sample group (n = 399).	53
Figure 14: Haplotype network for the entire sample group, including whites (n = 409).	54
Figure 15: Race (a), age (b) and region of birth (c) of study samples (n = 224).	58
Figure 16: Frequency distribution and age composition of the generations of Durban-area Indians in the sample since they arrived in KZN from India.	59
Figure 17: Likelihood that members of different sample groupings will share surname, city, religion and language with their children, paternal forefathers and maternal forefathers.	61
Figure 18: AMOVA: Distribution of molecular variance among and within surname-based groups of North Indians, South Indians and Zulus (n = 347).	62
Figure 19: PCoA plot for surname-based sample groupings.	64
Figure 20: Bayesian analysis of population genetic structure for three sample groups containing different surname sets (n = 347).	66
Figure 21: Haplotype networks for different surname-based groups (n = 347).	67
Figure 22: Haplotype network for shared haplotypes found in the total sample, which includes surname-based groups and random controls (n = 141).	68
Figure 23: MDS for 27 Y-STR loci for the studied and comparative populations from the YHRD.74	
Figure A 1: Research survey form	118
Figure A 2: Stand curve created using the Quantifiler Duo kit (Thermo Fisher Scientific, Waltham, Massachusetts).	120
Figure A 3: A Y-STR profile, using the Y-Filer® Plus kit, from GeneMapper® ID-X Software v1.4 (Thermo Fisher Scientific, Waltham, Massachusetts).	121
Figure A 4: Haplotype networks for individual surnames, based on Median joining method...	134

2. List of Tables

Table 1: Surname frequency world wide and in SA.....	11
Table 2: Composition of 5 different Y-STR multiplex kits.	23

Table 3: Summary of the current state of the YHRD.	27
Table 4: Sample table.....	36
Table 5: Comparative population databases (YHRD).....	37
Table 6: Null Alleles per locus (N=409).	46
Table 7: Allelic patterns for the overall study samples (N=409).....	48
Table 8: Forensic genetic parameters for the sample group (N=409) based on different marker sets.	49
Table 9: Sample sizes used in analyses of genetic structure among various subgroups of an overall sample of 399 Indian and Zulu samples.....	49
Table 10: AMOVA Results	50
Table 11: PCOA via covariance: Eigen values and percent variance explained for the principal components 1 and 2 for comparisons among different subgroups of the overall sample (n=399).	50
Table 12: Allelic patterns for the overall study sample (n = 399) and sub-grouping within this sample.....	55
Table 13: Forensic genetic parameters for the overall study sample (n = 399) and sub-grouping.	56
Table 14: Sample groupings used in surname-based genetic analyses (n = 347)	57
Table 15: Pairwise AMOVA for 27 Y-STR loci for surname-based groups.	63
Table 16: PcoA: Percent variation explained and eigenvalues for Indian, North Indian, South Indian and Zulu surname-based groups.	63
Table 17: Allelic patterns for the surname-based groups (n = 347).	70
Table 18: Forensic genetic parameters for different surname groups (n = 347).	71
Table 19: Number of haplotypes for the study samples and comparative populations (YHRD).	72
Table 20: Rst Clustering of populations.	72
Table 21: RST and significance values based on pairwise AMOVA for the study population and comparative populations from the YHRD.	73
Table A 1: The Yfiler® Plus composition.....	119
Table A 2: Standard dilution series	120
Table A 3: Allele frequencies for the overall sample group (n = 399) and sub-groupings.	122

INTRODUCTION

1. Background

Humans (*Homo sapiens*) have 46 chromosomes located inside the nucleus of cells. These comprise 22 homologous pairs, and the X and Y sex chromosomes. Every individual inherits one of each of the 22 homologous chromosome pairs and a sex chromosome from each parent. In the case of sex chromosomes, males will inherit an X chromosome from the mother and a Y from the father, whereas, females will inherit an X from each parent (Alberts *et al.*, 2002). During meiosis, recombination does not occur in most of the Y-chromosome (except for the 'pseudoautosomal' regions located at the chromosome ends). Thus, much of the Y-chromosome is passed on unchanged from father to son, unless a mutation has occurred. This means that many generations of males within a family lineage are likely to share an unchanged Y-chromosome sequence and therefore Y-STR profile (Jobling and Tyler-Smith, 1995; Jobling, 2001), and, in cultures where surnames are passed down in a patrilineal manner, will also share surnames. Rozhanskii and Klyosov (2011), using a metadatabase, found the average mutation rate for the Yfiler® marker haplotypes to be 0.00197 mutation/haplotype/generation, which was similar to the value of 0.00200 found by Klyosov (2009). The implication of this is that a Y-haplotype will mutate approximately once every 500 generations. Several aspects of human population history, including paternal lineage analysis and large-scale patterns of migration, can be explored by analysis of patterns of inheritance of STRs (short tandem repeats) associated with Y-chromosomes (Jobling and Tyler-Smith, 2003).

Driven by the popular interest in establishing family history, a number of studies based on co-inheritance of surnames and Y-haplotypes have been carried out (Sykes and Irven, 2000; Jobling, 2001; King *et al.*, 2006; McEvoy and Bradley, 2006; King and Jobling, 2009; Solé-Morata *et al.*, 2015; Martinez-Cadenas *et al.*, 2016). Genetic genealogy is a field of growing interest, involving genealogical testing to determine genetic relationships between individuals. Genealogical studies involve the use of a wide range of genetic markers to evaluate

relatedness, susceptibility to disease and individual ancestry (King and Jobling, 2009). These studies focus primarily on molecular evolution, population genetics and forensic parameters.

This study will focus on surname inheritance, population and forensic genetics of an experimental sample of North Indian males with the surnames Khan, Maharaj and Singh, and comparisons of these with similar data (from our lab) for South Indian (Govender, Naidoo and Pillay) and Zulu (Buthelezi, Cele, Dlamini, Mkhize and Zulu) surname-based groups sourced from Durban and surrounding regions of KZN SA. The samples will consist of males, owing to the focus on genetic genealogy, in particular the, co-inheritance of surnames and Y-STRs, which are found only in males. Understanding of genealogy (family history) in relation to genetic inheritance allows for a more comprehensive interpretation of genetic differences within the Indian and Zulu populations.

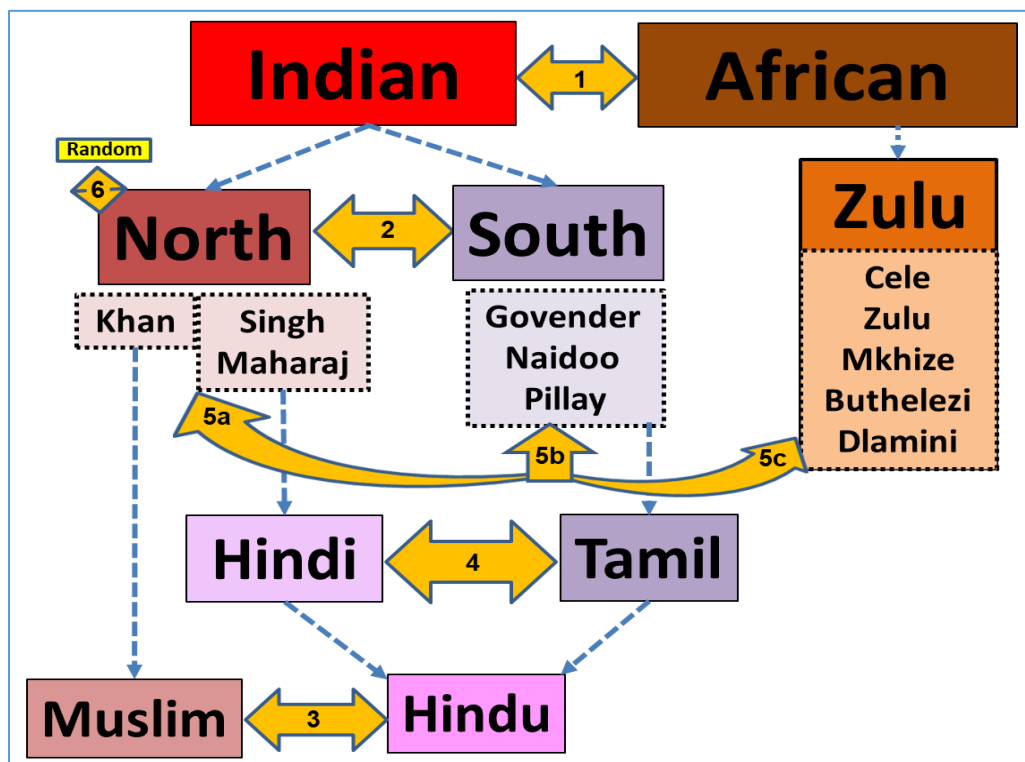


Figure 1: Concept Map of the study setup.

This figure illustrates comparisons made in this study, which will be based on 1. Ethnicity: Indian vs African; 2. Region of origin in India: North Indian vs South Indian; 3. Religion: Hindu vs Muslim; 4. Language: Hindi vs Tamil; 5. Genetic genealogy (coinheritance of surnames and Y-STRs): a. North Indian surnames (Singh, Maharaj, Khan), b. South Indian surnames (Govender, Naidoo, Pillay), c. Zulu surnames (Cele, Zulu, Mkhize, Buthelezi, Dlamini), and 6. Baseline social data for the North Indian and other random surnames used as controls.

In addition to the investigation of genetic structure among surname-based groups from India and SA, this study will also search for the existence (or not) of genetic structure among groups based on ethnicity, geographic regions of origin, religion and language (Figure 1). Also included will be a questionnaire aimed at establishing baseline aspects of the social history of the North Indian groups in Durban, which relate to genetic genealogy. The age profile and number of generations, since the first family member arrived in the Durban area from India, will be investigated. Information will also be collected on whether a North Indian individual shares his surname, city, religion and language with his child and paternal and maternal forefathers.

2. History of South African Indians

2.1. Movement from India to SA

On the 16 November 1860, the first ships transporting 600 Indians arrived in SA on the ship *Truro*, from Madras and Calcutta in India (Brain, 1985). In total 384 trips were made, bringing approximately 152 184 Indians to Port Natal (currently known as Durban) in SA, under the scheme of indenture (Figure 2), with the last ship, called *Umlazi*, arriving on 11 July 1911 (Chetty, 2010). Between 1860 and 1902, 59 662 people migrated to Durban from the South Indian region around Madras, and 35 720 arrived there from the North Indian area around Bihar, Bengal and Calcutta (Figure 2).

These Indians, who were imported by the Dutch as labourers since Africans refused to work for them, were known as 'indentured labourers'. They became urbanized in Natal and migrated towards towns in the Transvaal and the Cape Province after 1870 (Davies, 1981). After 1917, most of the Indians returned to India, although a few settled down in SA. The majority of these became owners of land along the coast of what is now known as KZN, including Durban and surrounding areas (Mukherji, 2011). They quickly established themselves as Industrial and railway workers, clerks and interpreters (SAHO, 2015).

By the 1940s the Indian community formed a major part the emerging industrial working class in KZN. The success of the settlement of these indentured labourers encouraged a new wave of migration by traders. This new group consisted of mainly Gujarati Muslims and Hindus and were known as “passenger Indians”, since they had paid to travel to SA (Mukherji, 2011). During the late 19th through to the early 20th-century, many SA Indians arrived from areas in the Indian subcontinent (Asia) that were under British Colonialism. This resulted in Indians being grouped with the broader ethno-geographic group of ‘Asians’ (Noble, 1994).

As settlements established, ethnic mixing occurred as Indians gradually started trading with other racial groups, such as whites and moved into neighbouring areas. Trading and residence competition resulted in the Indians being prohibited from settling in the Free State (Law 3 of 1885) and in restriction of the area where Indians could settle (Davies, 1981). From 1887, after extensive negotiations and legal proceedings, Indian trade and settlement was confined to certain areas, namely in the city of Durban (see Figure 3). According to the Census 2011 (Stats SA, 2011), there are 1 286 930 Indians found in SA (2.5% of the population), with more than half (756 991) living in KZN.

Currently, Durban has the highest population of Indian people outside of India (SAHO, 2015). Indian culture is followed to a high extent in Durban. However, over generations it has been segregated according to different religions and social classes (partially defined by surnames), additionally reducing strong associations and social similarity to Indians residing in India (Landy *et al.*, 2004). The majority of SA Indians are English speaking, with traditional languages mainly spoken by the older generations. Traditional Indian culture and languages are gradually dying off.

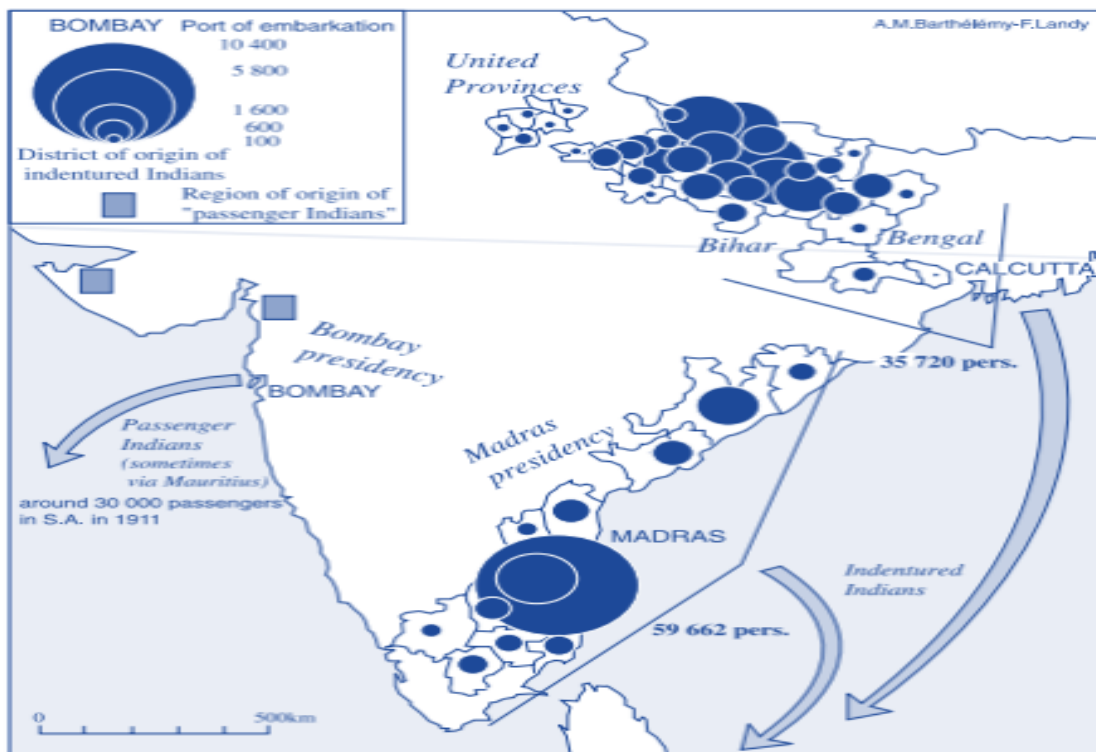


Figure 2: Indian migration.

The map shows Indian migration to KZN between 1860- 1902. Adapted from Landy *et al.* (2004).

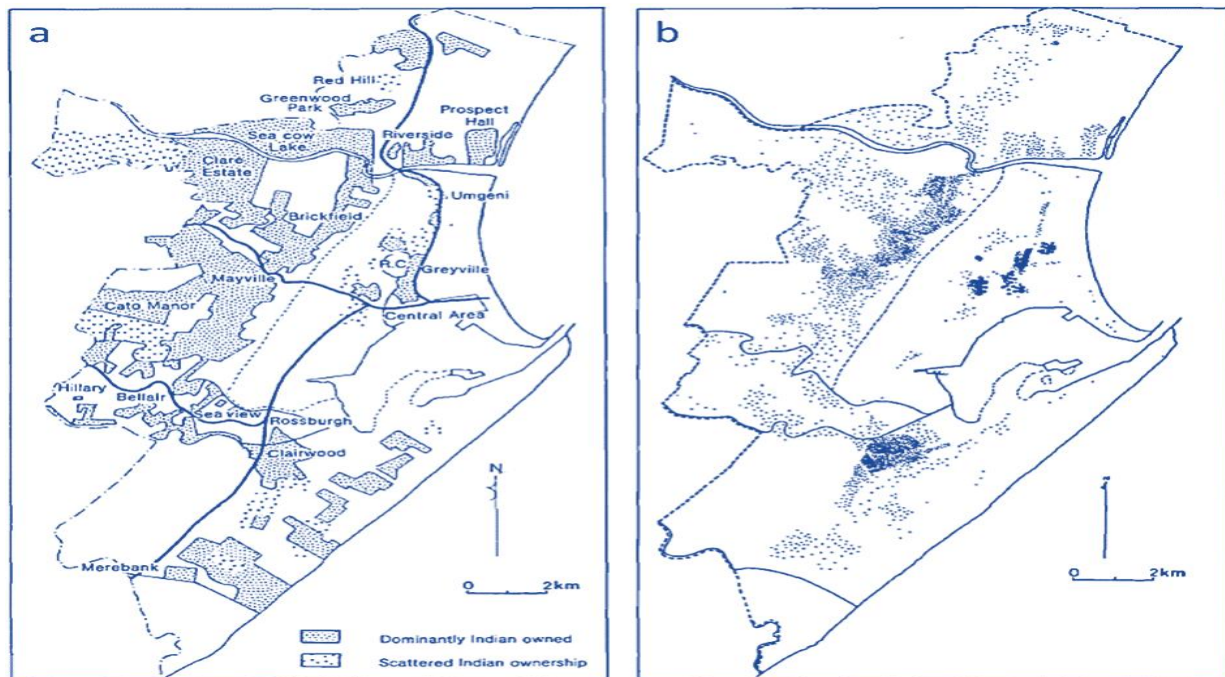


Figure 3: Indian settlement in Durban.

The map of Durban shows: a. Indian owned land in 1953, b. Indian distribution in 1951. Adapted from Davies (1981).

2.2. Indian Caste System

The word 'caste' is derived from the Portuguese word 'castas' which in translation means pure, idealising the important value in Indian culture i.e. *ritual purity* (Singh, 2017). To maintain this 'purity', Indian individuals tend to marry within the same caste. In the Indian population, there are ~ 37 000 different castes and ~ 500 different tribes (Papiha, 1996; Metspalu, 2001).

The Hindu caste system is believed to have originated from Lord Brahma (Hindu God of creation) and was created by Manu, the first of Lord Brahmas human sons. 'The Laws of Manu' were refined around 200BCE (Doniger, 1991). The Hindu caste system is divided into five groupings (Figure 4) with four main categories in the following order of descending hierarchy: Brahmins (priestly class), Kshatriyas (warrior class), Vaishyas (merchant and peasant classes) and the Shudras (labour class) (Alamy, 2017; Jayaram, 2017). The fifth grouping called 'The Dalit' (lowest of the Shudras) is detached from the other categories since this is an outcaste grouping. They were commonly referred to as the 'untouchables' or impure ones (Jayaram, 2017). Of the surnames included in this study, Maharajs are Brahmins (priestly class), whereas, Singhs are Kshatriyas (warrior class).

The Islamic religion does not embrace the Hindu caste system (Sirajudin, 2011). This led many lower caste individuals, and non-believers, to leave Hinduism and practice Islamic belief (Dirks, 2011). However, it is believed that all Indians, regardless of what religion they practice, tend to carry some vestiges of the caste system in them (Thekaekara, 2016). For example, Muslims, who follow a 'caste system', rank it according to ancestry, rather than occupation like the Hindu system.

The established Muslim system is followed mainly by South Asian Muslims Ranked in order of descending hierarchy (Figure 4) are: Ashrafs (foreigners who arrived from various regions and assimilated in Delhi, India, including; Syeds, Sheikhs, Mughals and Pathans; Ajlafs (all other converts); and Arzals (Muslim Dalits and caste renouncers) (Upadhyay, 2016).

The unfairness of the caste system has been challenged by Hindu reformist movements and has gradually changed over the years (Dirks, 2011). The modern day system is very different from the Hindu and Muslim caste systems (Figure 4) and is mainly based on wealth, education and job rankings (Alamy, 2017). However, the Hindu and Muslim caste systems are very much alive in India (Thekaekara, 2016).

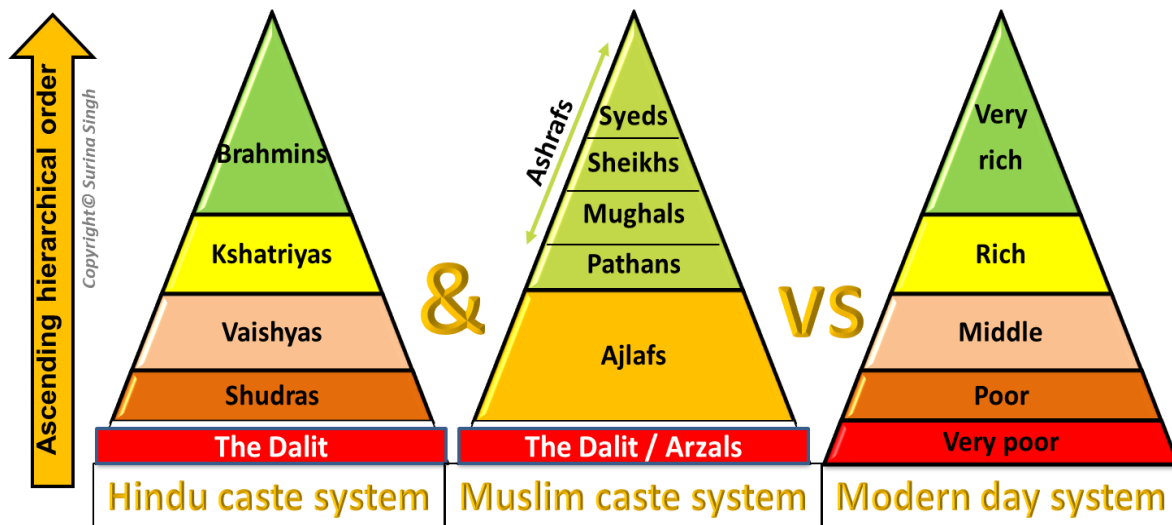


Figure 4: Comparison of the older Hindu and Muslim caste systems with the modern day system.

Indians that follow the caste system are relatively unlikely to marry out of their caste, resulting in less genetic variation between caste-based groups, and therefore genetic structuring.

Therefore, understanding the extent of adherence to the caste system can help in understanding why genetic similarities or differences are found within a particular grouping, whether it the grouping surname, regional, language or religion-based.

2.3. Marriage practices in Indians and Zulus

In SA, during Apartheid, mixed-race marriages were prohibited by the *Prohibition of Mixed Marriages Act (Act No 55 of 1949)* and the *Immorality Act of 1950*. In 1985, this act was repealed, legalising inter-racial marriages and relationships (SAHO, 2016). Generally, both Indians and Africans tend to marry within their own ethnic groupings. However, endogamy (marriage within a caste) is practiced by many Indians and exogamy (marriage out of a caste and with other castes or surnames) by Zulus.

2.3.1. Endogamy between Indians

Endogamy is the practice of marrying within a specific caste, social group or ethnic group (Bayly, 2001), and is widely practiced by Indians. Although marriage between Indians of different castes, religions, social and/or ethnic groups, is not permitted/ is looked down upon by the community, it does occur. Thus, an Indian is not encouraged to marry: (1) a non-Indian, creating an ethnic barrier between Indians and other racial/ethnic groups; (2) out of their religion, e.g. it is frowned on for a Hindu to marry a Muslim, creating a religion-based barrier; (3) out of their traditional language, e.g. a Hindi speaking person is not encouraged to marry a Tamil-speaking person, creating a language-based barrier and (4) out of their social group, e.g. a very rich person should not marry a very poor person (in accordance with the modern-day caste system (Figure 4)). An individual is also expected to marry within their 'jati' (birth caste) (Bayly, 2001).

Adherence to such social practices would tend to create genetic structuring among the different groups.

2.3.2. Exogamy in Zulus

Exogamy is widely practiced by the Zulu culture. It is against tradition for a Zulu to marry someone belonging to the same surname or caste as them, even if it is the same as that of their maternal forefather. The purpose of this practice is to prevent incest and inbreeding within the Zulu population. During Zulu weddings, the different clan names of the two families are announced to re-ensure that exogamy is practiced (Hamilton, 1997).

2.3.3. Changes/exceptions in intermarriage patterns

Hypergamy is marrying someone of a higher sub-caste/social class. Controversially, hypogamy is marrying someone of a lower sub-caste/social class (Gill, 2012). Hypergamy and hypogamy is culturally permitted in some situations provided the people involved are of the same religion and their castes are very different in terms of ranking (Ghurye, 1969). Exogamous marriages are accepted in Islam and Christianity on condition that the person from the different religion converts to Islam or Christianity.

Exogamous and inter-caste marriages have increased over time in different cultures due to individuals marrying within a similar educational background or job field. There are also more interactions between the youth of today, due to social media and social events, for example than there were in past generations, where females in particular were not educated and family caste or social ranking played an important role in their value in society (Leonard and Weller, 1980; Rao, 2003). According to Jacobson *et al.* (2004), the rate of exogamous marriages in SA, has increased from 1996 (303:1) to 2011 (95:1). This could be attributed to changes in societal attitudes, including greater acceptance of different cultures and ethnic groups (Jacobson *et al.*, 2004).

3. Surname studies

Many detailed surname genetic diversity (GD) studies focused on surnames of the British Isles (Sykes and Irven, 2000; King *et al.*, 2006; King and Jobling, 2009). The existence of genetic structuring among people with the same or different surnames can be confirmed or disproved; research can be directed at the history of formation of surnames in a particular geographical area or timeframe, making it easier to narrow down a search and locate a person of interest; and surname-based migration patterns can be established over generations (World Families Network, 2015).

The first genetic surname related study was published in 1875 by George Darwin, the son of Charles, who used surnames to estimate the frequency of marriages between first-cousins. He calculated the expected proportion of such marriages to be 4.5% for the upper classes and 2.25% for the general rural population (Darwin, 1875).

The isonymy method (based on the frequency of marriage between individuals of the same surname) became popular and widely used due to the affordability and ease of collecting large datasets of marriages, births and deaths among past and current populations (Jobling, 2001). This is a good method for estimating inbreeding coefficients within a population (Crow and Mange, 1965; Sykes and Irven, 2000).

The degree of co-ancestry, within surnames, is highly dependent on surname frequency (Martinez-Cadenas *et al.*, 2016). These authors found that, in the case of Y- chromosome single nucleotide polymorphisms (Y-SNPs) and Y- chromosome short tandem repeats (Y-STRs) the GD correlated positively with surname frequency (Spearman's $r = 0.896$; $p < 0.0001$ for STR haplotypes and $r = 0.749$; $p < 0.0001$ for haplogroups). However, the correlation found between surname frequencies and Y-STR inheritance differed between British and Irish surnames. British surnames, with a frequency higher than 5000 bearers at the national level, showed very little or no Y-chromosome co-ancestry, and Y-chromosome haplotype sharing increased as surname frequency decreased (King *et al.*, 2006; King and Jobling, 2009). This relationship was also observed by Martinez-Cadenas *et al.* (2016), who found that frequent Spanish surnames had a lower match probability, whereas, rare and very rare surnames have a higher match probability (Spearman's Rank Correlation: $r = 0.906$, $p < 0.0001$). In contrast, frequent Irish surnames were found to have very high Y-chromosome co-ancestry levels (McEvoy and Bradley, 2006).

Surname frequencies generally show clusters of population isolation if clans don't mix with one another and if they remain within their geographic location. This results in a lack of GD, which sometimes results in the development of a unique language or way of talking (Martinez-Cadenas *et al.*, 2016). Lane *et al.* (2002), who studied SA's Bantu-speaking groups (Pedi, Southern Sotho, Tsonga/Shangaan, Tswana, Venda Xhosa and Zulu) based on 9 autosomal and 4 Y-STR loci, found clustering within language-based groups, with the exception of the Tsonga speaking group, and high genetic distances between the different language groups. Lane *et al.* (2002) also found that genetic distances between SA's African populations correlated positively with geographical distances. In contrast, no correlation was found between language and geographical distance.

4. Origin and Frequency of Target surnames

Since the Medieval period, it has been common practice for English children to take the surname of their father (Sykes and Irven, 2000). The majority of Indian and Zulu surnames originated from a family's local language. Generally, the meaning of the surname was derived from a family's geographical origin or profession and/or social status, and in some cases the surname also indicated the caste. In these groups, surnames were also passed down from father to child.

Children taking their fathers surname has become a common practice since the English Mediaeval period. The majority of Indian and Zulu surnames originated from a family's local language. Generally, the meaning of the surname was derived from a family's geographical origin or profession and/or social status, and in some cases the surname also indicated the caste. Table 1 shows the surname frequency of the North Indian, South Indian and Zulu surname sub-groups used in this study.

Table 1: Surname frequency world wide and in SA

Ranking = Ranking position as most common surname. ~ n = Approximate number of people sharing the surname.

Adapted from Forebears (2016) and Name Stats SA (2016).

1. North Indian surnames				2. South Indian Surnames			
Surname	Location	Ranking	~ n	Surname	Location	Ranking	~ n
a. Singh	World	6 th	36 970 960	a. Govender	World	2 487 th	220 815
	SA	15 th	182 936		SA	2 nd	364 936
b. Maharaj	World	5 559 th	98 287	b. Naidoo	World	1 719 th	315 434
	SA	46 th	84 708		SA	1 st	498 108
c. Khan	World	12 th	24 514 296	c. Pillay	World	2 409 th	227 413
	SA	38 th	96 720		SA	4 th	324 376
3. Zulu surnames							
Surname	Location	Ranking	~ n	Surname	Location	Ranking	~ n
a. Cele	World	12 678 th	41 824	d. Buthelezi	World	9 716 th	55 471
b. Zulu	SA	328 th	18 252		SA	205 th	27 560
	World	1639 th	330 394	e. Dlamini	World	1 609 th	336 164
c. Mkhize	SA	177 th	30 212		SA	64 th	63 024
	World	6 552 th	83 094				
	SA	150 th	34 632				

4.1. Surnames of Indians transported to Natal

When transported to SA, the Indian settlers were asked for their surname or, in the case of those who did not understand, their 'father's name'. Confused by what this meant, many listed their father's first name as their surname. This resulted in the founding of new surnames, which creates difficulties when attempting to trace back ancestry. The surnames collected as the experimental sample in this study, Khan, Maharaj and Singh, are ranked in the top 50 commonest surnames in SA (Name Stats SA, 2016). These are regarded as 'highest caste' surnames.

4.1.1. North Indian Surnames

The Singh surname ranks as the 2nd most common surname in India (1:35 people), followed by SA (1:500 people) and the United States (1: 2 533 people) (Forebears, 2016). It is still a common surname for many North Indian Hindus. It is the most common North Indian Hindu surname in SA (Name Stats SA, 2016). Singh, originating from India, is derived from the word *simba*, meaning 'lion' in Sanskrit (Feuerstein, 2002). Singh is commonly used as a surname as well as a title and middle name. This surname is associated with power and authority, and was adopted by people of multiple castes (Chaudhary, 1995). Spelling variations include 'Sinh', and 'Sing'. In the 16th century, this surname became popular amongst Rajputs (a clan derived from the Sanskrit term 'raja-putra', meaning 'son of a king') and in 1699 it was adopted by Sikh followers of Guru Gobind Singh (Chander, 2003). Singh was used as a title by several groups in the 18th century. This surname is found throughout the Indian sub-continent, amongst several communities and religious groups (Singh, 1996) and it is only out of India where it is known to be a true surname (Brook, 2017).

The Maharaj surname has the highest frequency in SA (1: 962 people) and is the second highest frequency in India (1: 76,631 people) (Forebears, 2016). Maharaj originated from the word *Maharaja*, meaning 'great king' in Sanskrit. Spelling variations include Maraj, Maharajh and Maharaja. It has the frequency in Trinidad and Tobago, where it is the 8th most common surname (Forebears, 2016). However, there is still confusion surrounding the origin of this

surname. Most people identify a Maharaj as someone from Brahmin decent (Figure 4), however, this is not always the case. Maharaj was also used as a title to represent someone who was a master of a particular skill, and as such could be a guru (priest) or even a cook (Orie, 2013).

The Khan surname is most commonly found in Pakistan (1:15 people) (Forebears, 2016). In India 1 in 281 people bear this surname, and in SA, 1 in 734 people (Forebears, 2016). Khan originated, around 4000 BCE, as a title given to a Prince or Lord amongst Mongolian and Turkish tribesmen (Yule and Burnell, 1996; Brook, 2017). This surname was owned by Genghis Khan, an emperor also known as the 'Great Khan', who established a historical empire ranging from Turkey to China (Brook, 2017). According to Zerjal *et al.* (2003), approximately 16 million Asian men are descendants of Genghis Khan. Khan is a common surname in China, belonging to 1:734 people (Forebears, 2016). Khan is most commonly used as a surname amongst the Muslim population of Pakistan. It has a frequency of 1:15 among the Pakhtoon people. where it ranks as the most common surname. In Bangladesh, it is borne by 1:31 people and in India, 1:281 people (Forebears, 2016). The absorption of this name into the Muslim population had nothing to do with Islamic religion, but was a title given, during the British era in India, to 'loyal Muslims' (Phukan, 2011).

4.1.2. South Indian Surnames

The Govender surname is most commonly found in SA (1:254 people) and is not as frequent in India (1: 43 035 172 people) (Forebears, 2016). The name Govender originated in Lanarkshire, Scotland. They were owners of the land 'Govan'. The name was derived from two Saxon words meaning 'good wine'(Lewis, 1851). Spelling variations of this family name include 'Govan'.

The Naidoo surname is the 1st most frequency surname in SA (Table 1) (1: 180 people) and is greater than that in India (1: 20 459 344 people) (Forebears, 2016). Naidoo is a Tamil surname from the Naitea clan (also known as Navāyats) (Forebears, 2016). It is used as a title by the Telugu community (Kumari, 1998). Spelling variations include 'Naidu'.

The Pillay surname is found at greatest frequency in SA (1:273 people); it has the fourth highest frequency in India (1: 343 334 people) (Forebears, 2016). Pillay is a Tamil surname, meaning 'child'. It was derived from the word 'Pilav', which refers to an oriental dish of stewed meat and rice (Forebears, 2016). Spelling variations include 'Pillai', which is used more as a title, meaning 'prince'.

4.2. Zulu Surnames

Zulu people are of Bantu decent, and arrived in SA in the 9th century (Contralesa, 2016). They are the largest ethnic group in SA, where IsiZulu is the most common home language, spoken by 11.58 million people in SA (22.7%) (Stats SA, 2011). Y-STR data from Zulu males, generated by Siyethaba Mkhize and part of the lab database, were included in these analyses for comparative purposes. The Zulu males included in the study had the surnames Buthelezi, Cele, Dlamini, Mkhize and Zulu.

The Cele surname is most commonly found in SA (1:1410 people) (Forebears, 2016). Cele is derived from the old English word 'saelig' which means 'one who is happy and blessed' (<https://www.houseofnames.com/cele-family-crest>).

The Zulu surname is most commonly found in Zambia (1:71 people) and second most commonly in SA (1:870 people) (Forebears, 2016). Zulu means 'heaven' in the Zulu language, deriving from the Nguno Tribe (Forebears, 2016). The Zulu clan was formed in Northern KZN in 1709 by Zulu kaMalandela (Contralesa, 2016).

The Mkhize surname is most commonly found in SA (1:70 people) (Forebears, 2016). The Buthelezi surname is most commonly found in SA (1:1009 people) (Forebears, 2016). The Dlamini surname is most commonly found in Swaziland (1:7 people) and second most commonly in SA (1:382 people) (Forebears, 2016).

5. Patterns of transmission of surnames and haplotypes/haplogroups

According to Jobling (2001), surnames may be monophyletic i.e. derived from a founder (Figure 5a) or polyphyletic, derived from multiple founders (Figure 5b). Transmission of surnames occurs with low fidelity when there is a disturbance between the co-transmission of the surname and the Y-chromosome (Figure 5c). If surnames have genetically similar founders, then overlapping haplotypes would occur (Figure 5d).

Monophyletic, with high fidelity (Figure 5a) is the 'ideal' method of surname transmission. Here, each surname has a unique founder whose haplotypes are highly diverged from those of the other founders. The haplotypes, therefore, do not overlap and there will be no haplotypes shared between surnames. If each surname has a unique founder, but the founders are not highly diverged from each other, then overlapping haplotypes will occur, and people with different surnames can share Y-STR haplotypes (Figure 5d).

Transmission of surnames may be high fidelity in situations, where a child always derives its surname from its biological father. However, there are several situations which are likely to cause surnames to be transmitted with less fidelity i.e. 'low fidelity transmission'. These include situations where the child does not take the surname of the true father; the child might be conceived via a partner who is outside of the marriage (or current relationship) in which the mother is involved and takes the name of her husband/partner; the mother might choose to retain her surname and pass it on to the child; or the child might be adopted and bear the name of the adoptive parents (Figure 5c) (Jobling and King, 2004).

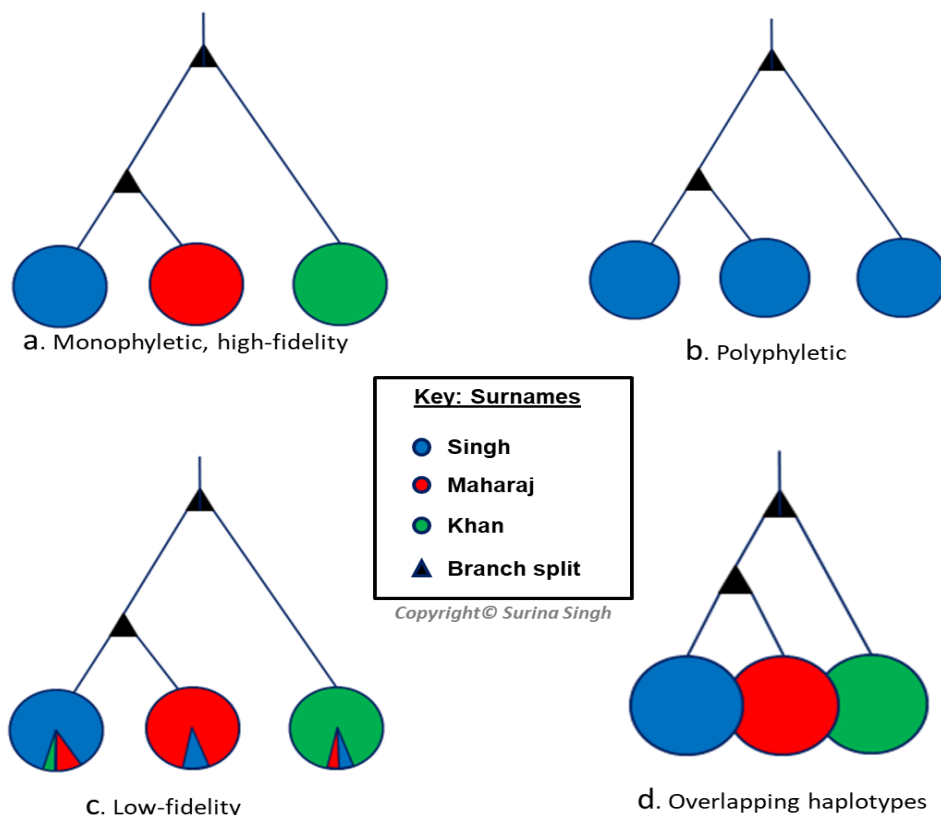


Figure 5: Possible relationships between Y-chromosomal haplotypes and surname transmission.

An illustration of the phylogenetic patterns expected in the case of different modes of surname transmission for three different surnames, based on Jobling (2001).

In many cultures, both surnames and the Y-chromosome are passed down from father to son. Thus, people with the same surname should have a greater chance of being related to each other than to members of the general population (King and Jobling, 2009). This makes it possible to trace male lineages based on surnames.

In the 'ideal' situation, each surname might form an independent Y-chromosome based haplotype cluster (this might be regarded as the equivalent of a null hypothesis relating to surname and Y-chromosome transmission). These clusters would be distinct from each other and not contain haplotypes found in other surnames i.e. monophyletic transmission with non-overlapping haplotypes (Jobling, 2001), allowing surnames to be predicted from Y-STR haplotypes, which would be of use in forensic genetics.

Several aspects of human population history, in particular large-scale patterns of migration, can be explored by analysis of the inheritance of Y-STRs (Jobling and Tyler-Smith, 2003).

Most of the Indian paternal population (> 50%) belongs to the R1a1, O2a, and H mitochondrial haplogroups (Sahoo and Kashyap, 2006). Haplogroups A, B and E are most commonly found in the African ethnic groups of SA (Motladiile, 2004). In SA, the haplotype diversity between four different ethnic groups (Zulu, Coloured, Afrikaner and Indian) was found to be 0.9981 (Tsiana, 2015). Solé-Morata *et al.* (2015), found that HD of Catalan surnames in Catalonia was positively correlated with surname frequency, and the inheritance of these surnames was monophyletic with high-fidelity. Monophyletic transmission was also observed by Sykes and Irven (2000) who found that almost half of their samples with the surname 'Sykes' in United Kingdom, shared the same Y-chromosome haplotype, even though it was predicted that this surname originated in many different regions (i.e. was predicted to be polyphyletic). This pattern was not found in the control group (random male individuals), which contained samples from the same geographical region. The average rate of non-paternity was estimated to be 1.3% per generation during the past 700 years.

6. The Y-chromosome

Discovered in 1905, the Y-chromosome (Figure 6) is a single copy sex chromosome found only in males. The Y-chromosome, about 60 megabytes (Mb) in size, is one of the smallest chromosomes found in the human genome (Skaletsky *et al.*, 2003) and comprises approximately 2% of the total Deoxyribonucleic acid (DNA) (Seyedebrahimi *et al.*, 2017). The parental ancestor of all modern Y-chromosomes is assumed to come from Africa (Out of Africa concept) (Seielstad *et al.*, 1999). The first STR polymorphism, discovered in 1992 by Roewer *et al.* (1992), lead to the foundation of Y-chromosome evolution studies (Page *et al.*, 2010; Bachtrog, 2013; Kayser, 2017).

The Y-chromosome is essentially haploid in state and is made up of the pseudoautosomal regions (PAR) at the tips and a central non-recombining region of the Y-chromosome (NRRY) (Figure 6). During meiosis, recombination is limited to the PAR i.e. PAR 1 (less than 1mb) and 2 (approximately 2.5mb), found at the tip of the short arm (Yq) and long (Yp) arm respectively (Graves *et al.*, 1998). The NRRY (Figure 6) is more commonly referred to as the male-specific region of the Y-chromosome (MSY) and comprises 95% of its length. The MSY represents a useful source of polymorphisms for the forensic analysis of male DNA (Jobling and King, 2004).

The paternally inherited MSY remains intact throughout generations, unless a mutation occurs (Quintana-Murci *et al.*, 2001; Kwak *et al.*, 2005; Butler, 2011). Since both DNA and surnames are passed down from our ancestors in many cultures and traditions, people with the same surname should have a greater chance of being related to each other than to members of the general population (King and Jobling, 2009). Since most surnames are inherited surname studies can trace male lineages.

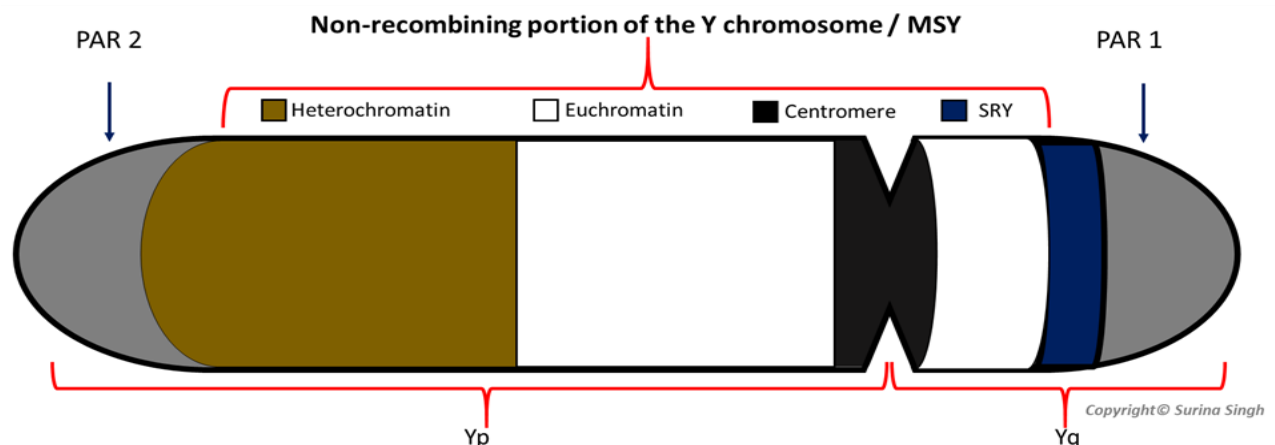


Figure 6: Y-chromosome structure.

An illustration of the Y-chromosome, indicating: the Yp (long arm), and Yq = (short arm); PAR (pseudoautosomal region) 1 and 2; the non-recombining region/ MSY (male specific region of the Y-chromosome) which contains: the heterochromatic region (non-functional genes), euchromatic region (functional genes), centromere, and the SRY (sex-determining region of the Y-chromosome).

6.1. The importance of Y-chromosome testing in forensics

Descendants from the same male lineage often share a Y-STR haplotype; this is due to the lack of recombination during meiosis, resulting in no variation or very little variation (in the case of a

mutation) from generation to generation. According to Zhivotovsky *et al.* (2004), the average Y-chromosome mutation rate is 6.9×10^{-4} per 25 year period (Butler, 2005, 2014). Thus, male family lineages will share a Y-STR haplotype for many generations until a mutation occurs.

Y-chromosome DNA profiling has important applications, especially in forensics. It can be used to eliminate and/or contribute to identifying a male perpetrator by comparison with other suspect profiles (unless the suspects come from the same paternal lineage). In paternity testing it can provide evidence on the likelihood of a male being the father of a child, as they would be expected to share the same Y-STR profile, unless a mutation had occurred; again, if other male members of the same lineage were also potential fathers, this method would not be helpful (Khan *et al.*, 2017). It is used in other paternal kinship analyses such as historical cases and for familiar searching such as the Vaatstra case. It is also used in male related missing person and victim identification cases (Kayser, 2017).

Sexual offences, specifically rape, are a serious problem especially in SA, therefore an effective system to catch perpetrators is required. SA has one the highest number of rape reports, therefore, a large growing database plays a critical role in establishing a more stable approach in how strong Y profiles can play in forensic investigations (Brenner, 2010; Andersen *et al.*, 2013; Andersen and Balding, 2017; Cereda, 2017). According to the South African Police Service annual crime report (<https://www.saps.gov.za/services/crimestats.php>) the prevalence of rape has steadily decreased since 2012/2013 (Figure 7), which could have decreased due to the development of the DNA act in 2013.

The Criminal Law (Forensic Procedures) Amendment Act No. 37 of 2013, also known as the "DNA Act", allowed for the development of a forensic DNA database in SA. This DNA act entitles the SA police services to collect DNA generate DNA profiles and match profiles of the reference samples to crime scene samples. However, getting access to a reference sample is not allowed guaranteed. This places limitations on the use of DNA as an investigative tool. SA law also prevents the use of molecular typing phenotyping i.e. DNA cannot be used in any analyses

pertaining to health/ medical testing, nor can it be used to find other any other physical information other than the gender of a person (Slabbert and Heathfield, 2018).

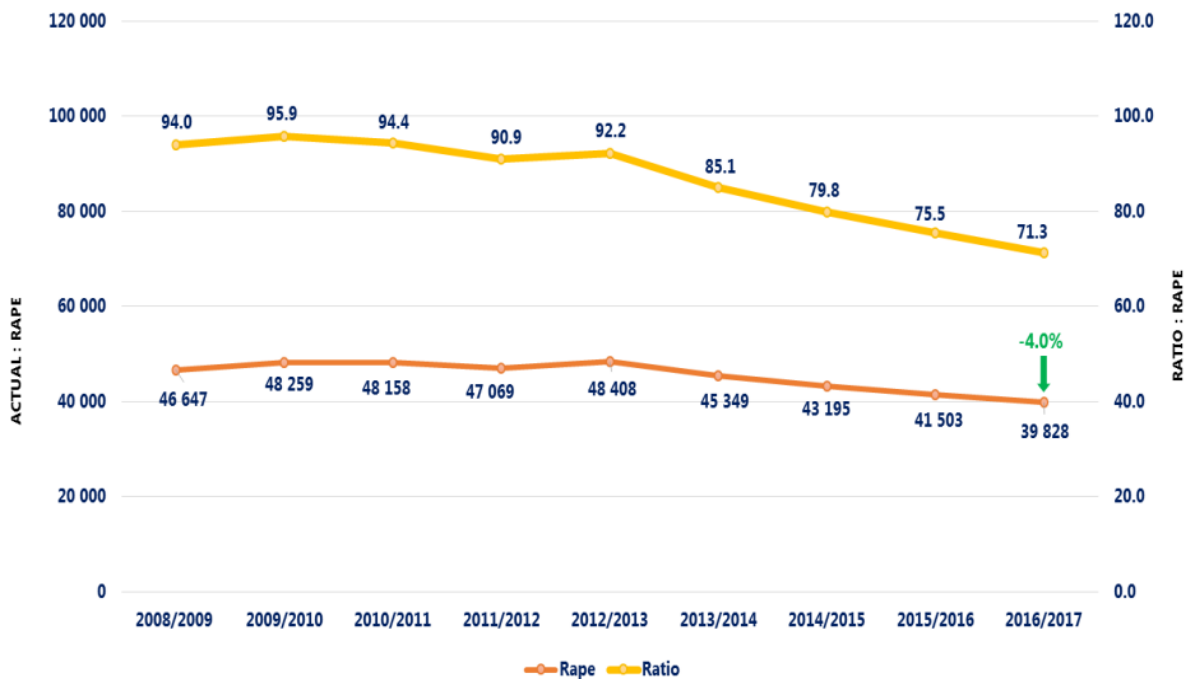


Figure 7: SA's reported rape case trend over a nine year period.

Available from: <https://www.saps.gov.za/services/crimestats.php>.

Y-STR loci are useful in investigation of rape cases as: 1) It is possible to generate the male perpetrator's Y-STR profile from biological evidence left behind at the crime even if only mixed samples are available; 2) Male Y-STR profiles are unambiguous, even in the presence of excess amounts of female DNA; and 3) The number of male contributors can be identified, as the non-recombining region is haploid (Shewale *et al.*, 2004; Khan *et al.*, 2017). In SA, D'Amato *et al.* (2011) assessed the use of 10 non-commercial Y-STR markers (DYS385a/b, DYS447, DYS449, DYS481, DYS504, DYS518, DYS612, DYS626, DYS644, and DYS710) in forensic analyses of sexual assault for the African, White and Asian/ Indian population and found these markers give higher levels of diversity and discrimination capacity (DC) when compared with common commercial markers.

Evaluation of different non-commercial Y-STR marker sets in SA revealed, as might be expected, that the greater the number of markers, and the higher their diversity, the more useful the

marker set is for forensic applications (Leat *et al.*, 2004; Ehrenreich, 2005; Leat *et al.*, 2007; D'Amato *et al.*, 2011; Tsiana, 2015).

6.2. Polymorphic Y-chromosome markers

The two major categories of Y- chromosome markers are Y-SNPs and Y-STRs (also known as microsatellites). Convenient polymorphic Y-SNPs and Y-STR markers have been identified on the non-recombining portion of the Y-chromosome (Jobling and Tyler-Smith, 1995).

SNPs are the most abundant type of polymorphism found in the human genome, and occur at approximately 1 in every 1000 bases (Sachidanandam *et al.*, 2001). Discovered in 1994 by Hammer (1994), Y-s are slow-mutating binary polymorphisms, with mutations occurring at a rate of approximately 10^{-8} per base per generation (Jobling and King, 2004). They are classified into haplogroups i.e. groups related by descent from a common ancestral SNP haplotype.

In contrast, microsatellites (STRs) are fast mutating multi-allelic markers. Large numbers of haplotypes are usually found within haplogroups and can be used to estimate the Time to the Most Recent Common Ancestor (TMRCA) (Jobling, 2001). STRs comprise ~ 3 % of the human genome, and are found in both the autosomes and sex-chromosomes, including the Y-chromosome (Ehrenreich, 2005). The number of tandem repeats varies among individuals, making these polymorphic makers useful in HID (Moxon and Wills, 1999).

The Y-STR locus DYS385 and the forward primer-binding site of DYS389 are duplicated on the Y-chromosome. Whilst undergoing polymerase chain reaction (PCR) using a single primer set, two fragments are created for DYS385, i.e. DYS385a and DYS385b, if different alleles are present on the two duplicates of this locus. DYS389 yields two products i.e. DYS389I and DYS389II, which differ in length by approximately 100bp (Ehrenreich, 2005). As Y-STRs are highly polymorphic, they are useful in identifying and comparing closely related populations (Tsiana, 2015).

The use of Y-SNPs in determining haplogroups can often be time consuming, therefore, haplogroup prediction from Y-STRs using an allele frequency approach (Athey, 2005) is more commonly used (Kwak *et al.*, 2005; Chang *et al.*, 2007; Parkin *et al.*, 2007; Frank *et al.*, 2008; Balamurugan *et al.*, 2010; Yadav *et al.*, 2011; Park *et al.*, 2012; Shrivastava *et al.*, 2017). Most of the observed variation in Y-STRs, within and among populations, is believed to be almost selectively neutral. Due to these features, Y-STRs are good markers for genetic mapping, intra-species phylogenetics and forensic analysis such as HID (Rowold and Herrera, 2003).

6.2.1. Y-STR markers

The number of Y-STR markers, used in forensic and population genetics analysis applications, has substantially increased over the past few years (Westen *et al.*, 2015). Initially, it was recommended that a minimum of 9 Y-STR markers should be used for forensic applications (De Knijff *et al.*, 1997). However, the product rule used to estimate polymorphism in autosomal chromosomes does not apply to Y-chromosomes, therefore, more Y-STR loci (than autosomal STR loci), are required to obtain the same discrimination power (Bosch *et al.*, 2002).

Multiplex Y-STR kits for amplifying many target loci are more advantageous than uniplex systems, as they are quicker to use and require less reagents and template DNA. Disadvantages include unequal amplification of target sequence DNA and an increased chance of the formation of non-specific products when several primer sets are used simultaneously (Ehrenreich, 2005). Although the increase in number of marker loci per kit may generate suspect exclusions, such kits often contain many loci which are not particularly informative (Leat *et al.*, 2007). According to Ge *et al.* (2010), combinations of a few highly mutating markers gave relatively larger Fixation index (F_{st}) values than kits with a larger number of less polymorphic markers. Investigations of primer selection, PCR amplification and fluorescent allele detection are required for the establishment of an optimized protocol for a universally accepted multiplex system for comparative genotyping (Wallin *et al.*, 2002). These systems must meet the performance standards of several authorities such as the ISFG (International Society for Forensic Genetics) and the SWGDAM (Scientific Working Group on DNA Analysis Methods).

There are many commercially available multiplex kits, and new kits are being developed all the time (Budowle *et al.*, 1997; Kayser, 2017). Kits for Y-STR testing include: MHt and RMu (non-commercial); PowerPlex® Y and Y23 System (Promega); and Yfiler® and Yfiler® Plus (Thermo Fisher Scientific, Waltham, Massachusetts) (Table 2).

Table 2: Composition of 5 different Y-STR multiplex kits.

Tick represents presence of locus in a multiplex kit. MHt = Minimal Haplotype, PPY = PowerPlex® Y System (Promega), Yfiler® = AmpFISTR® Yfiler® (Thermo Fisher Scientific, Waltham, Massachusetts), PPY = PowerPlex® Y23 System (Promega), Yfiler® Plus = Yfiler® Plus (Thermo Fisher Scientific, Waltham, Massachusetts), and RMu = Rapidly Mutating makers.

Marker/Loci	Multiplex Kit					
	Non-commercial		Promega		Thermo Fisher	
	MHt	RMu	PPY	PPY23	Yfiler®	Yfiler® Plus
DYS19	✓		✓	✓	✓	✓
DYS385 a	✓		✓	✓	✓	✓
DYS385 b	✓		✓	✓	✓	✓
DYS389I	✓		✓	✓	✓	✓
DYS389II	✓		✓	✓	✓	✓
DYS390	✓		✓	✓	✓	✓
DYS391	✓		✓	✓	✓	✓
DYS392	✓		✓	✓	✓	✓
DYS393	✓		✓	✓	✓	✓
DYS437			✓	✓	✓	✓
DYS438			✓	✓	✓	✓
DYS439			✓	✓	✓	✓
DYS448				✓	✓	✓
DYS456				✓	✓	✓
DYS458				✓	✓	✓
DYS481				✓	✓	✓
DYS460				✓		✓
DYS449		✓				✓
DYS533				✓		✓
DYS549						✓
DYS570		✓		✓		✓
DYS576		✓		✓		✓
DYS518		✓				✓
DYS526 a		✓				
DYS526 b		✓				
DYS547		✓				

FDYS635			✓	✓	✓
DYS643			✓		
Y-GATA-H4			✓		
DYS627	✓				✓
DYS612	✓				
DYS626	✓				
DYF387S1 a	✓				✓
DYF387S1 b	✓				✓
DYF399S1	✓				
DYF403S1 a	✓				
DYF403S1 b	✓				
DYF404S1	✓				
Total Loci	9	16	12	23	17
					27

The MHT consists of 9 Y-STR loci (Table 2), and was the first multiplex marker set recommended for forensic applications (De Knijff *et al.*, 1997). Ehrenreich (2005) found that the MHT loci, on their own, were not suitable for forensic investigations amongst South African sub-groups on account of their low variability and should be complemented with other kits.

The RMu marker set is a non- commercial kit consisting of 16 Y-STR loci (Table 2), entirely different from the MHT. All RMu loci have mutation rates of above 1×10^{-2} , which allows for discrimination between closely related as well as non-related males (Ballantyne *et al.*, 2012).

The PowerPlex® Y (PPY) System (Promega) consists of 12 Y-STR loci (Table 2), which include both the MHT and the Scientific Working Group on DNA Analysis Methods (SWGDAM) recommended panel of Y-STR loci (Krenke *et al.*, 2003).

The AmpFISTR® Yfiler® kit (Thermo Fisher Scientific, Waltham, Massachusetts) is a five dye multiplex system which includes the PowerPlex® Y loci and five additional loci (Table 2) for a total of 17 Y-STR loci.

The PowerPlex® Y (PPY) System (Promega) is a five dye multiplex system which includes the AmpFISTR® Yfiler® kit (Thermo Fisher Scientific, Waltham, Massachusetts) and six additional loci (Table 2), for a total of 23 Y-STR loci (Thompson and Storts, 2012).

The Yfiler® Plus PCR Amplification Kit (Thermo Fisher Scientific, Waltham, Massachusetts) is a 6-dye multiplex system which allows amplification of 27 Y-STR loci from male samples (Figure 8). This marker set includes all 17 loci found in the Yfiler® Kit (DYS19, DYS385a/b, DYS389II, DYS389I, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS635 and YGATAH4) and 10 further loci (DYS460, DYS481, DYS533, DYS576, DYF387S1a/b, DYS449, DYS518, DYS570 and DYS627). Seven of these additional loci are RMu makers (DYS576, DYF387S1a/b, DYS518, DYS570 and DYS627). This kit was used for the experimental work in this study.

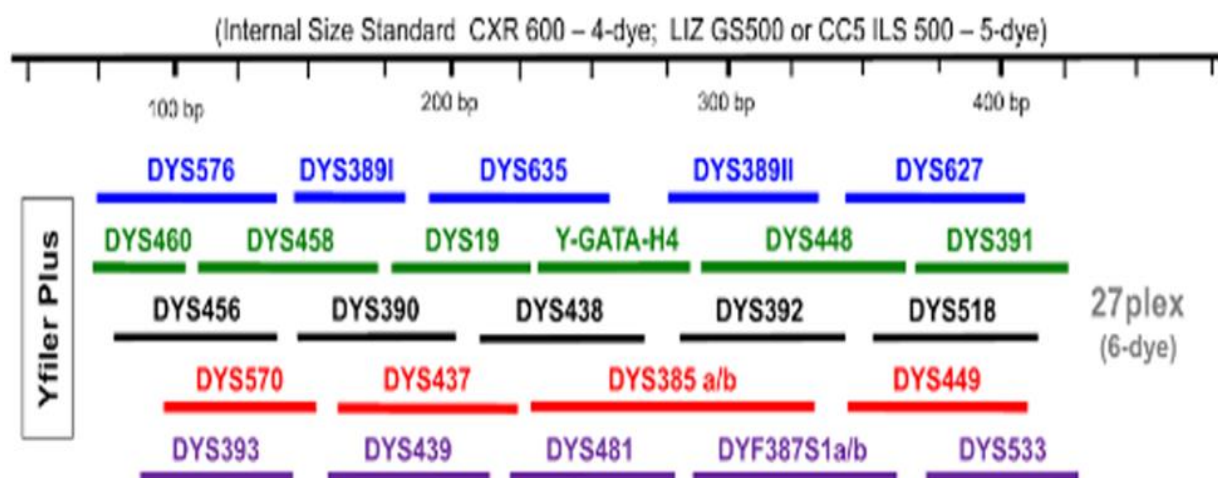


Figure 8: Fragment sizes and fluorescent dyes used in the amplification of 27 Y-STR loci by the Yfiler® Plus PCR amplification kit.

This system contains 27 Y-STR loci (DYS389I, DYS635, DYS389II, DYS627, DYS460, DYS458, DYS19, YGATAH4, DYS448, DYS391, DYS456, DYS390, DYS438, DYS392, DYS518, DYS570, DYS437, DYS385 a/b, DYS449, DYS393, DYS439, DYS481, DYF387S1 a/b, DYS533). Adapted from Butler (2014).

D’Amato *et al.* (2011), identified low GD levels and discriminatory capacity (DC) in SA populations for many commercial markers sets and found that DC increases as the number of markers in the kit increase. The addition of more Y-STR markers have been found to result in higher numbers of unique haplotypes and in turn a higher power of discrimination between individuals, making the data more reliable for forensic investigations (Ge *et al.*, 2010; D’Amato

et al., 2011; Ottaviani *et al.*, 2015). Ge *et al.* (2010) found that the unique alleles found within the studied United States (US) population increased from 56.7 %, using 10 Y-plex markers, to 92.9 % using 16 Yfiler® markers. Ottaviani *et al.* (2015) found that DC increased from 78.7% when using MHT loci to 100% when using the Yfiler® Plus kit.

The Yfiler® Plus kit presents more informative haplotypes (due to higher number of useful loci), a greater robustness, sensitivity, sensibility and higher DC than Yfiler® kit, specifically cases with high proportions of female DNA (Cainé, 2016; Phan, 2017).

Currently the Yfiler® Plus kit includes the greatest number of loci (27), which means it should have the highest discriminating power as compared to other available kits. Therefore, this kit was chosen for this study.

7. YHRD (Y-chromosome Haplotype Reference Database)

In 1994, research was directed at creating a Y-chromosome database for forensic investigation. Kayser *et al.* (1997) reported seven loci (DYS19, DYS389I/II, DYS390, DYS391, DYS392, and DYS393) suitable for forensic studies, i.e. the European 'Minimal Haplotype' (MHT). An additional locus, the duplicated YCA II locus, was incorporated to form the 'extended haplotype' (Kayser *et al.*, 1997).

The estimation of Y-STR haplotype frequencies need to be much larger than autosomal DNA referencing. Therefore a large Y-STR haplotype frequency database is required for a more reliable frequency estimates (Kayser, 2017). Building a Y-chromosome DNA data base using the MHT resulted in the establishment of the YHRD (Willuweit and Roewer, 2015; YHRD, 2018). This global Y-database currently has over 160,000 haplotypes derived from 1015 populations and 33 metapopulations (YHRD, accessed 28 July 2017). It contains Y-STR profiles from a wide range of multiplex Y-STR kits i.e. MHT, PowerPlex Y, Yfiler® (Thermo Fisher Scientific, Waltham, Massachusetts), Yfiler® Plus (Thermo Fisher Scientific, Waltham, Massachusetts), and Maximal.

Table 3: Summary of the current state of the YHRD.

Adapted from YHRD (2018). n = Number.

Dataset	Haplotypes (n)	Population samples (n)	National databases (n)	Meta-populations (n)
Minimal	197,102	1128	133	33
PowerPlex Y	158,812	879	124	32
Yfiler	145,816	796	117	32
PowerPlex Y23	39,414	236	63	28
Yfiler® Plus	22,832	108	36	28
Maximal	2,849	24	10	16

South African samples on the YHRD based on the minimal marker set and from the Cape Town region include; 108 European Afrikaners (accession number: YA003259), 100 European-English speakers (accession number: YA003257), 114 Mixed ancestry samples (Caucasian and Xhosa) (accession number: YA003260) (Leat *et al.*, 2004), and 213 Xhosa speaking individuals (Leat *et al.*, 2007). Samples from the Johannesburg region include 393 individuals from the Eastern Bantu population (contributor: Tony Lane, accession number: YA003281-1). There are also accessions on the YHRD based on the PowerPlex, Yfiler and PowerPlex Y23 markers sets for 114 Cape Town Xhosa speaking individuals (accession number: YA003258) (Leat *et al.*, 2007).

Currently, the YHRD contains no South African Yfiler®Plus data; this study has the potential to contribute South African ethnic haplotypes based on this marker set.

8. Purpose of this research

The purpose of this research is to study aspects of the genetic (inheritance Y-STRs) and social history of the Indian and Zulu population groups from the greater Durban-area i.e. assessing Y-STR data and estimating surname, ethnicity or region of origin. Y-STR studies have carried out in SA, based on evaluating the suitability of different Y-STR marker sets (Leat *et al.*, 2004; Ehrenreich, 2005; Leat *et al.*, 2007; D'Amato *et al.*, 2011; Tsiana, 2015), but no studies have been based on the Yfiler® Plus kit, nor were they surname-based.

The combination of molecular genetics and surname analysis of Y-STR data has the potential to shed light on population structure and history (King and Jobling, 2009), and is within the field of HID forensic DNA analysis. Each surname group should ideally form an independent Y-chromosome haplotype cluster. These clusters would be distinct from each other and not contain haplotypes found in other surnames, allowing surnames to be predicted from Y-STR haplotypes, which would be of use in forensic genetics. If a relationship exists between surnames and Y-STRs, the consensus sequence for each surname group could potentially be determined and can be used to predict if persons surname belongs to a surname, adding more detail to what is known about the investigated person. In forensic investigation, this type of data could serve as a “Biological witness” in criminal cases and has importance in kinship analyses and familiar searching, specifically mission person identification cases (Kayser, 2017).

There is need for such studies, as there are currently no surname-based studies on the Indian or Zulu populations of SA. The completion of this study could potentially provide a baseline for further research, as there are many different frequently occurring surnames found among different ethnic groups.

9. Aims, Objectives and hypotheses

The overall aim of this study is to explore the genetic genealogy, population and forensic genetics of (1) samples comprising male Indians with a variety of surnames, geographic regions of origins, religions, languages, and (2) male Zulus with different common surnames, all currently residing in the greater Durban area of KZN, SA.

The target Indian surnames are associated with different regions of origin in the Indian subcontinent (North vs South Indians), religions (Hindu and Islamic Muslim), and consequently languages (Hindi, Tamil and Urdu). It is possible to formulate hypotheses about potential population subdivision based on differences in geographic regions of origin, surname, and perceived barriers to marriage between people of different religions (e.g. Hindus and Islamic

Muslims). The Zulu samples comprise people from the Zulu (African) ethnic group bearing one of 5 common surnames, lending themselves to an analysis of the co-inheritance of surnames and Y-STRs.

Based on the concept map (Figure 1), this study will examine genetic relationships among and within different groupings and sample sets comprising:

- (1) Groups based on ethnicity, namely Indian vs Zulu.
- (2) Groups based on region of origin in India, namely North Indians vs South Indians.
- (3) Groups based on religion, namely Hindu vs Muslim.
- (4) Groups based on language, namely Hindi, Muslim (Urdu) and Tamil.
- (5) Surname-based groups, namely (a) North Indian (Khan, Maharaj and Singh), (b) South Indian (Govender, Naidoo and Pillay), and (c) Zulu (Buthelezi, Cele, Dlamini, Mkhize and Zulu).
- (6) Also included will be a questionnaire aimed at establishing baseline aspects of the social history of the North Indian groups in Durban, which relate to genetic genealogy. The age profile and number of generations since the first family member arrived in the Durban area from India will be investigated. Information will also be collected on whether a North Indian individual shares his surname, city, religion and language with his child and paternal and maternal forefathers.

The methodology will include the following basic steps: (1) Obtaining ethical approval for the study. (2) Sampling: Most of the samples used in this work will be sampled *de novo*, but certain analyses and comparisons will be based on already-profiled samples from the lab database, as indicated in the Materials and Methods section. (3) Y-STR profiling: In the case of material sampled for this study, Y-STR profiles will be created using the standard procedure, namely; (3a) DNA extraction using a standard commercial kit, (3b) Quantification of the amount of Human DNA in the sample by Q-PCR using the Applied Biosystems Quantifiler Duo PCR amplification kit, (3c) Amplification of 27 Y-STR loci using the Applied Biosystems Yfiler® Plus PCR Amplification Kit, (3d) Separation of the amplified fragments by capillary electrophoresis on the Applied Biosystems 3500 Genetic analyser, and (3e) Creation of DNA profiles using Applied Biosystems GeneMapper® ID-X Software v1.4 (Thermo Fisher Scientific, Waltham,

Massachusetts). (4) Comparative population genetic, forensic genetic and genetic genealogical analyses.

9.1. Utility of different Y-STR marker in population and forensic genetics analyses

This study aims to examine the population and forensic genetics of the following marker sets, which include differing Y-STR loci: (1) MHT, (2) Yfiler®, (3) Yfiler® Plus, and (4) RMu marker sets. The purpose is to assess their suitability for forensic investigation in the overall sampled population from the greater Durban metropolitan area of KZN. This will be achieved by calculating: (1) GD per locus and overall; (2) Allele frequencies; (3) Haplotype frequencies; (4) Allelic patterns such as: number of effective and private alleles, GD, and Percentage of polymorphic loci; and (5) Forensic parameters such as: HD, MP, DC and percentage of polymorphic loci.

H₁: The Yfiler® Plus Kit, chosen for use in this study, is an appropriate choice for genetic genealogy, population and forensic genetics studies of Indian and Zulu males from the greater Durban area of KZN. Genetic/haplotype diversity and discrimination capacity will be the highest for the Yfiler® Plus marker set and haplotype match probability lowest, as it includes a higher number of loci than any of the other marker sets to which it was compared (MHT, Yfiler, and RMu), and includes all of the markers in these marker sets.

9.2. Genetic structure analyses based on population sub-groupings

This study aims to search for the existence of genetically structured sub-groups within the entire sample. (1) Pairwise Analysis of Molecular Variance (AMOVA) will be run to search for population differentiation. (2) The number of genetic clusters (population groups) in the overall population will be estimated using Bayesian Analysis of Population Genetic Structure (BAPS) V6.0.1 (Corander *et al.*, 2008). This programme will also be used to estimate the percentage membership of each sample member in each identified cluster. (3) Principal Co-ordinates Analysis (PCoA) will be carried out to visualise genetic distance and relatedness among sample

groups, using GenAlEx v6.5 (Peakall and Smouse, 2012). Haplotype networks will be constructed to examine mutational relationships between haplotypes.

H₂: Genetic structure will be found within the overall population, as barriers based on ethnicity, geographic region of origin, language and religion many have led to a level of genetic isolation, thus, allowing the Y-chromosomes of males in these groups to diversify.

9.2.1. Genetic structure based on ethnicity (Indian vs Zulu)

This study aims to examine the genetic structure amongst the Indian and Zulu components of the sample.

H_{2a}: The sample will be structured into groups based on ethnicity, as the Indian and Zulu population groups have very different geographic origins and are likely to have evolved separately prior to the introduction of Indians to the Durban region. Although interbreeding many have occurred since the Indian group arrived in the Durban area, causing the introduction of Zulu Y-chromosomes into the Indian groups and vice versa, Apartheid and various cultural practices would have served as a deterrent.

9.2.2. Genetic structure based on region of origin in India

This study aims to examine whether there is genetic structure amongst sample members whose surnames indicate origins in North vs South India.

H_{2b}: The distance separating the sites of origin of the North and South Indian samples, which originated in Calcutta and Madras, combined with language and cultural differences, and the practice of endogamy will have led to genetic diversification of North and South Indians, which will be reflected in Y-STR genetic structure among

9.2.3. Genetic structure based on religion

This study aims to examine the genetic structure amongst Hindu and Muslim Indians.

H_{2c}: In the period since the separation of the Muslim religion from the more ancient Hindu religion, Y-chromosome genetic diversification and therefore structuring will have occurred between Hindus and Muslims, maintained by perceived religious barriers to intermarriage between Muslims and Hindus.

9.2.4. Genetic structure based on language

This study aims to examine the genetic structure amongst Hindi, Muslim (Urdu) and Tamil speaking Indians.

H_{2d}: Hindi, Tamil and Urdu speakers (Muslims) will have evolved as separate groups, due to language barriers; this diversification will be reflected as Y-chromosome based genetic structure.

9.3. Surname-based genetic analyses

This study aims to determine surname-based genetic structure and mode of inheritance, and population and forensic genetics among surname groups within 3 different surname-based categories. These include: (a) the samples collected for this project, which originate from North India and (b) profiles from the lab database included for comparative purposes: One set of such data based on South Indian surnames was generated by Velosha Naidoo. The other set, based on Zulu surnames, and was generated by Siyethaba Mkhize.

This study also aims to examine social data related to inheritance in the North Indian surnames, 'Singh', 'Maharaj' and 'Khan', to provide baseline information relative to randomly chosen people with different surnames drawn from the most common ethnic groups found in SA. Frequency distributions will be run on the population sub-groups in order to track surname

inheritance, and changes in surname, religion and geographical distribution of Durban area Indians over generations.

9.3.1. Surname-based genetic structure and surname inheritance

9.3.1.1. *Surname genetic structure*

This study aims to examine whether there is genetic structure among the different surnames within each of the three groups (North Indian, South Indian, Zulu). This will include AMOVA and PCoA using GenAlEx v6.5 (Peakall and Smouse, 2012), and determination of the number of subgroups, by running Bayesian Estimation of Population Structure analyses using BAPS V6.0.1 (Corander *et al.*, 2008).

H₃: The Y-chromosome and surnames are paternally inherited in both North and South Indians and Zulus. As the Y-chromosome has a relatively low mutation rate per generation, it could be hypothesised that groups of people with a particular surname would be more closely related to each other than to groups with other surnames. Genetic divergence over time among surname groups based on religious and or cultural practices or region of origin are likely to be reflected in genetic divergence among surnames.

9.3.1.2. *Surname inheritance*

This study aims to analyse Y-STR DNA profiles to determine the manner in which surnames are inherited and the relationship between surname and Y-STR haplotype transmission based on the concepts described by Jobling (2001) (Figure 5). Of interest is whether the surnames are monophyletic or polyphyletic, whether the haplotypes associated with each surname are discrete or overlapping, and whether the surnames are transmitted with high or low fidelity.

H₄: As both surnames and the Y-chromosome are purported to be patrilineally inherited, surname inheritance will be monophyletic and high fidelity (Jobling, 2001), i.e. each surname group would

form an independent Y-chromosome haplotype cluster and these clusters would be distinct from each other and not contain haplotypes found in other surnames. This would allow surnames to be predicted from Y-STR haplotypes, which would be of use in forensic genetics. Further, surnames are transmitted from father to son with high fidelity.

9.4. Population and forensic genetics based on population sub-groupings

This study aims to examine population and forensic genetics for the overall sample group and sample sub-groups within the entire sample, where there is evidence for genetically structured groups. The sub-grouping will be based on ethnicity, region of origin in India, religion, language and surname. This will be achieved by calculating: (1): Allele frequencies; (2) Allelic patterns such as: number of effective and private alleles, GD, and percentage of polymorphic loci; and (3) Forensic parameters such as: HD, MP, DC.

9.5. Comparison of experimental samples with samples and populations on the YHRD

This study aims to determine whether any of the experimental haplotypes are also present on the YHRD (2018). Another aim is to compare the surname groups and random samples to profiles of equivalent ethnic groupings currently available on from YHRD (2018) and to assess Y-chromosome based GD and structure across these sample sets. The Multidimensional Scaling (MDS) tool found within the YHRD (2018) will be used to analyse data in terms of GD and structure across sample sets.

MATERIALS AND METHODS

1. Sampling

Samples were collected primarily from individuals residing in Durban and surrounding areas in Kwa-Zulu Natal province, SA (Figure 9); this region is known to have the largest Indian population outside of India (Mukherji, 2011).

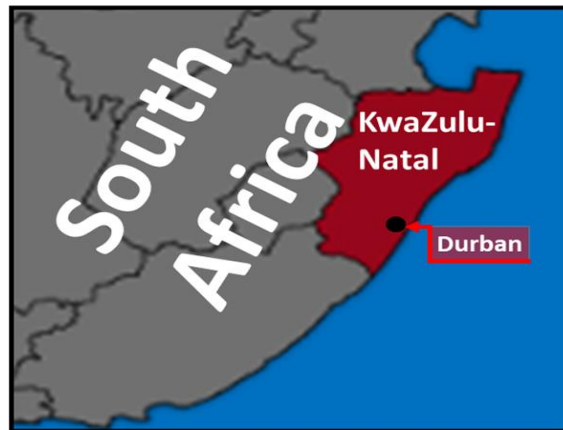


Figure 9: Sampling location map.

Samples were collected from the greater Durban metropolitan area in the province of KZN.

The minimum number of samples for each group was set at 50, in accordance to the SWGDAM guidelines for reliable data for comparative purposes the maximum number depended on the number of (2) individuals willing to participate, and (2) samples successfully profiled. DNA samples were collected from 224 non-paternal lineage related males (males who were not from the same father-line), 161 of which were of Indian descent with majority from the Durban region of KZN who shared one of three surnames, namely 'Khan' (n = 52), 'Maharaj' (n = 50) and 'Singh' (n = 59) (Table 4). Spelling variations were also accounted for, e.g. Maharaj and Singh included individuals with the spelling variation 'Maharajh' and 'Sing' respectively. These surnames are considered as 'high class' surnames in the Indian society and are among the 50 most common South African surnames (Name Stats SA, 2016). Individuals with 'high class' surnames generally marry within their class/caste, making interbreeding likely. Random samples, not chosen based on surname, and representing the population of the Durban region of KZN, were taken from 63 non-paternal lineage related males; these included people of Indian

origin (Hindus and Muslims) as well as males from three different SA ethnic groupings (African, Coloured/mixed race and White). Only one representative of each surname was included in the control group, to provide as wide a spread of surnames as possible in this group.

1.1. Sample composition

The samples collected for this study were supplemented with Y-STR profiles from the lab database to enable certain comparisons to be made (Table 4). Y-STR profiles bearing one of three South Indian (Tamil) surnames, namely Govender (30), Naidoo (30) and Pillay (30), were contributed by Velosha Naidoo, whereas Siyethaba Mkhize contributed Y-STR profiles of groups bearing the Zulu surnames Buthelezi (20), Cele (n=20), Dlamini (20), Mkhize (20) and Zulu (20) (For publication, these two students will be recognised as co-authors).

Table 4: Sample table.

Samples were collected from the greater Durban area of KZN. n = the total number of samples collected. Y-STR (N) = the number of Y-STR profiles generated and used in analyses.

Population			Category	Code	Total (n)	Y-STR (n)
A. Samples collected for this study	Indian	North Indian surnames	Khan	K	52	51
			Maharaj	M	50	48
			Singh	S	59	58
		Subtotal			161	157
	Other	Random Surnames	Hindi	RH	27	27
			Muslim	RM	10	9
			Tamil	RT	11	11
			African	RA	5	5
			White	RW	10	10
			Subtotal			63
Total				224	219	
B. Samples obtained from the lab database	Indian	South Indian (Tamil) surnames	Govender	G	30	30
			Naidoo	N	30	30
			Pillay	P	30	30
		Subtotal			90	90
	African	Zulu surnames	Buthelezi	B	20	20
			Cele	C	20	20
			Mkhize	Mk	20	20
			Dlamini	D	20	20
			Zulu	Z	20	20
			Subtotal			100
Overall Total (n)					414	409

1.2. Comparison with other available databases

The YHRD was used for: (1) Validating dataset of profiled samples, (2) Searching for matches between Y-STR haplotypes of study samples and samples from other world populations were searched for on the YHRD, (3) Running AMOVA and MDS analyses for study samples and samples from other world populations were searched for on the YHRD. For the AMOVA and MDS analyses, samples collected for use in this analysis (Table 4) were compared to only populations of the same/similar ethnic groups (to Indian, African and White) which were currently available Yfiler® Plus haplotypes in the YHRD (2018). For Indians, all the India located, and Asian grouping populations were chosen. For Africans, the Kenya (Bantu) population located and African grouping populations were chosen. For whites, the European grouping populations were chosen. This gave members of 14 other similar ethnic grouping populations (Table 5), originating from Australia, India, United States and Kenya.

Table 5: Comparative population databases (YHRD).

Samples from the Yfiler® Plus database in (YHRD, 2018) included in the analyses for purposes of comparison, originating from Australia, India, United States and Kenya. Hap (N) = number of haplotypes.

Population group	Code	Hap (N)	Accession code
Australia [Asian]	Au[As]	196	YA004231
Australia [European]	Au[Eu]	197	YA003698
India	I	19	(see below)
Assam, India [Kachari]	AI[Ka]	8	YA004030
Kerala - India [Keralite]	KI[Ke]	3	YA003946
Andhra Pradesh - India [Thoti]	API[Th]	8	YA004029
United States [Asian American]	US[As]	240	YA004084
United States [African American]	US [A]	479 (42,237)	YA003313, YA004083
United States [European American]	US[Eu]	465 (230,235)	YA003314, YA004085
Minnesota, United States [Asian American]	Mi, US[As]	96	YA004088
Minnesota, United States [African American]	Mi, US [A]	77	YA004087
Minnesota, United States [European American]	Mi, US[Eu]	67	YA004089
Kenya	Ken	128 (62,12,54)	YA004206, YA004205, YA004207
Kenya [Bantu_Luhya, Other]	Ken[BLo]	62	YA004206

2. Sampling methodology

2.1. Sample collection

For the experimental DNA profiling carried out in this study, samples were collected from 224 non-related male individuals (see Table 4). These were primarily of North Indian origin, and included males of three surname groups, Khan (52), Maharaj (50), and Singh (59); also included were 63 randomly sampled individuals with different surnames and of different racial/ethnic groups, as a control group. Sampling was done with accordance with the UKZN Biomedical Research Ethics code (BE456/16, sub-study of BCA056/16). Participants were invited to take part in the study via email, Facebook and word of mouth. The participation of the study subjects was voluntary and anonymous. The sampling methodology caused no harm to the participants. Incentives, i.e. a highlighter pen, were given to participants as a token of appreciation. Prior to participation the project was explained, and an informed consent form was given to the prospective participants.

2.1.1. Collection of DNA samples

DNA samples were collected via a buccal swab. It was ensured that participant's mouths were clean prior to sample collection i.e. mouth was rinsed if food had been recently consumed. Participants were required gently scrape the inside of the cheek with a sterile unused FLOQSwab™ (Thermo Fisher Scientific, Waltham, Massachusetts). This was done 5 times for each cheek (i.e. 10 times in total). The swabs were then allowed to dry prior to DNA extraction.

2.1.2. Answering of Questionnaires

Each participant answered a questionnaire (Figure A 1: Research survey form) related to the region, language, inheritance of surname and geographical distribution of the participants and their forefathers and included questions on basic demographic information. The purpose of the questionnaire was to generate complementary information to allow a better understanding of the genetic results.

3. DNA Profiling

DNA profiles were generated using the following steps: DNA extraction, assessment of the concentration of DNA in the extracted sample, STR-amplification and separation of the STR products by capillary electrophoresis.

3.1. DNA extraction

DNA extractions were carried out using the ZR Genomic DNA™-Tissue MiniPrep Kit (Zymo Research, USA), Qiagen DNeasy kit (Qiagen, Hilden, Germany) and PrepFiler® Forensic DNA Extraction Kits (Thermo Fisher Scientific, Waltham, Massachusetts). Prior to extraction, all three kits were tested and were found to work and provide DNA of high purity (A260/A280 ratio of >1.6 and useable concentration. DNA was extracted according to manufacturer's instructions, and stored at - 20°C in the long term, whereas working samples were stored at 4°C.

3.2. DNA Quantification

DNA concentration and purity determined using the Nanodrop spectrophotometer, ND-1000, v3.8 (Thermo Fisher Scientific, Waltham, Massachusetts); the A260/A280 ratio was assessed in order to obtain an estimate of the purity of the extracted DNA (samples with ratios of greater than 1.6 were deemed useable). The Quantifiler® Duo DNA Quantification Kit (Thermo Fisher Scientific, Waltham, Massachusetts) was used to obtain a real-time PCR estimate of the concentration of human DNA in the sample according to manufacturer's instructions), in conjunction with the ABI7500 real time PCR machine (Thermo Fisher Scientific, Waltham, Massachusetts).

A standard dilution series (Table A 2) allowed for a standard curve to be created (Figure A 2), from which the concentration of the unknown experimental samples would be determined. Knowledge of the concentration of human DNA in the sample was necessary in order to allow an appropriate amount to be used in the Y-STR amplification step. Use of too much DNA in the amplification step can cause problems which are observed after fragment separation by capillary electrophoresis; these include off scale peaks, inconsistent peak heights and areas,

'pull up' peaks and split peaks, due to incomplete adenylation. Low DNA concentrations can result in partial or no profiles.

3.3. Amplification of Y-STRs

Prior to Y-STR amplification, the concentration of the extracted DNA was diluted to ~ 1ng/μl per sample. The Yfiler® Plus PCR Amplification Kit (Thermo Fisher Scientific, Waltham, Massachusetts) was tested according to manufacturer's instructions, performing one third, half and full reactions. It was found that all three reaction amounts gave full Y-STR profiles. The purpose of this was to determine whether it would be possible to generate full profiles with lower amounts of reagents, to save laboratory work costs. It was decided that one third reactions would be used.

Y-STR PCR amplifications were performed using the Yfiler® Plus PCR Amplification Kit (Thermo Fisher Scientific, Waltham, Massachusetts), which amplified 27 loci (Figure 8). Reactions were carried out, according to the manufacturer's instruction, as described in the Yfiler® Plus PCR Amplification Kit instruction manual (Thermo Fisher Scientific, Waltham, Massachusetts), with the exception that one-third reactions were used. Internal validation was done using one third, half and full reaction volumes, with all resulting in full profiles being generated with no allele drop outs, validating the use of one third reactions being used due to limitations in the financial resources available for the project. Thus, instead of 25 μl reactions, 8.3 μl reactions were formulated, using the same reagent composition. The 8.3 μl reaction consisted of: 0.3 μl sample DNA (~ 1ng/μl), 3.3 μl Master mix, 1.7 μl Primer Mix and 2 μl distilled water. Samples were loaded into a MicroAmp® Optical 96-well Reaction Plate (Applied Biosystems) and amplified using the Applied Biosystems Verity PCR machine (Thermo Fisher Scientific, Waltham, Massachusetts). Thermal cycling was carried out in 9600 Emulation mode, as per manufacturer's instructions: Initial incubation = 95°C for 1 minute; followed by 30 cycles of denaturation at 94°C for 4 seconds and annealing at 61.5°C for 1 minute; final extension = 60°C for 22 minutes; and final hold = 4°C until run was stopped. The amplified products were then stored at 4°C until further use.

3.4. Capillary electrophoresis

Capillary electrophoresis was performed, according to the manufacturer's instruction, as described in the Yfiler® Plus PCR Amplification Kit instruction manual (Thermo Fisher Scientific, Waltham, Massachusetts), in order to separate the amplified Y-STR allelic fragments. Prior to capillary electrophoresis, samples were prepared by adding 9.6 µl Hi-Di™ formamide and 0.4 µl GeneScan™ 600 LIZ® Size standard v2.0 (Thermo Fisher Scientific, Waltham, Massachusetts) to each 1 µl amplified sample product. This was then heated for 3 minutes at 95°C, using the Applied Biosystems Verity standard PCR machine, in order to ensure that the DNA was denatured, and then placed on ice until use. The 96 well plate consisted of: one positive control, one negative control, 4 allelic ladders, prepared sample DNA (Figure 10). Allelic ladders were added 1 per 3 injections (1 ladder every third column), to accurately genotype samples. The products were separated using the Applied Biosystems 3500 Capillary Electrophoresis machine. Parameters were set up as per manual (Thermo Fisher Scientific, Waltham, Massachusetts): Sizing range = Partial ranging from 60 bp to 460 bp; and Size calling method = Local Southern Method.

The outputs were analysed using GeneMapper® ID-X Software v1.4 (Thermo Fisher Scientific, Waltham, Massachusetts) to generate a Y-STR profile chromatogram. The analytical thresholds were set according to the Yfiler® Plus PCR Amplification Kit instruction manual (Thermo Fisher Scientific, Waltham, Massachusetts). The genotype for each sample was then determined after comparing output to the Yfiler® Plus allelic ladder and control samples to define off ladders (See Table A 1: The Yfiler® Plus composition). Measures, such as lab sterile techniques, were put in place to prevent contamination. If contamination was determined, samples were re-run to get an accurate profile.

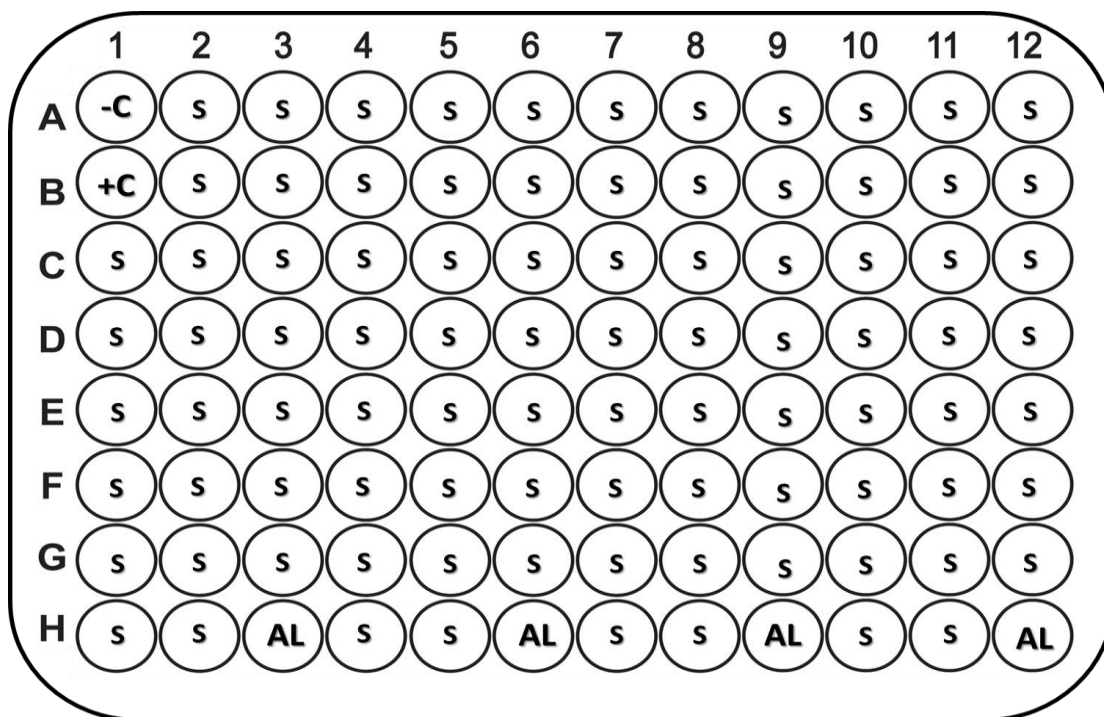


Figure 10: Electrophoresis plate setup example.

An illustration of a typical 96 well plate set up where: -C = Negative control, +C = Positive control, S = Sample and AL = Allelic ladder.

4. Genetic Analyses

Prior to analysing, the data obtained from 414 male SA individual DNA profiles were evaluated using YHRD (2018). The evaluation process allowed for the change in the uploaded file coding/structure to suit the database's specific analyses and eliminated profiles lacking allelic information. Five samples were removed due to the lack of allelic information: 1 Singh (S41), 2 Maharaj (M22, M49) 1 Khan (K39), and 1 Random (RM03) sample. These samples were removed in further analyses.

4.1. Genetic Structure

AMOVA and PCoA were carried out in GenAlEx v6.5 (Peakall and Smouse, 2012). AMOVA was used to estimate genetic differentiation between populations based on the PhiPT statistic, which shuffles whole population among regions. Pairwise PhiPT AMOVA analyses were calculated to examine where there were significant differences between the different sub-groups. Significance was set as $p = 0.05$ and calculations were set at 9999 per mutations. PhiPT

was calculated using the formula: $\Phi_{iPT} = \frac{AP}{(WP+AP)} = \frac{AP}{TOT}$, where AP = Estimated variation Among Populations, and WP = Estimated variation Within Populations. Haploid Nm values were calculated using the formula: $Nm = \frac{(1/\Phi_{iPT}) - 1}{2}$. PCoA was used as an exploratory tool to visualise similarities and dissimilarities among sample groups based on 27 Y-STR loci.

Bayesian Estimation of Population Structure analysis, which accounts for linked loci present in Y-STR genetic structure analyses, was used to estimate the number of genetic clusters (population groups) in the overall population, using BAPS V6.0.1 (Corander *et al.*, 2008). The most likely K (genetic cluster membership) was determined for mixture and admixture analyses, using the following parameters: (1) Mixture analyses were run for (a) Linked loci and (b) Clustering of individuals, where the input upper bond number of populations was set at the standard value = 20. (2) If admixture was present between different K populations, the admixture analysis was run with the following standard parameters: pop = 5, iteration number = 50, reference individual from each pop no. = 50, reference individual from each pop iteration number = 10. Significant clustering was determined using $p = 0.05$. A phylogenetic tree was generated using the Neighbour-Joining method, if $K > 2$, to display the genetic clustering structure distance.

SplitsTree V4.14 (Huson and Bryant, 2005) was used to construct a haplotype network for the overall sampled population and the 3 different surname groupings (North, South and Zulu surnames), using the Neighbor Net method. This method combines aspects of the neighbor joining (NJ) and SplitsTree, using a genetic distances to compute Neighbor Net splits into a network rather than a tree (Bryant and Moulton, 2002). Network v5 (Forster *et al.*, 2017) was used to construct a Median-joining (MJ) haplotype network each surname group. Prior to construction, a star conduction was calculated to reduce data complexity by identifying clusters and shrinking nodes. The MJ haplotype network was calculated with default recommended settings (parameters set as 5, loci weight automatically set at 10 for all 27 loci, epsilon = 0). Maximum parsimony (MP) trees was then calculated to clean up the contracted network by removing links not used to generate the shortest tree in the network. As per manual instructions (<http://www.fluxus-engineering.com/Network5000userguide.pdf>): The network

tree for shared haplotypes was drawn based on frequencies > 1, for a less complex tree; The Median Joining square option was selected for more MP links; and Distance was calculated by the connection cost method.

4.2. Genetic diversity and Forensic parameters

GD for each locus was calculated according to Nei and Tajima (1981) using the formula: $GD =$

$\frac{N(1 - \sum_i x_i^2)}{N-1}$, Where x_i^2 is the (relative) allele frequency of each haplotype in the sample and N is the sample size. Allelic Patterns, Allele frequencies and GD were computed using GenAlEx v6.5 (Peakall and Smouse, 2012). Haplotype frequencies were calculated using the counting method (Rapone *et al.*, 2016).

Haplotype diversity (HD), also known as gene diversity, is a measure of haplotype uniqueness within a specific population. This was calculated according to Nei and Tajima (1981), using the

formula: $HD = \frac{N(1 - \sum_i p_i^2)}{N-1}$, where p_i^2 is the (relative) frequency of each haplotype in the sample and N is the sample size. Haplotype match probability (MP) was calculated using the formula $MP = \sum_i p_i^2$ and discrimination capacity (DC) was calculated by: $DC = \frac{H_{Ob}}{N}$, where H_{Ob} is the number of haplotypes observed (Rapone *et al.*, 2016).

4.3. Comparison of experimental samples with samples and populations on the YHRD

The YHRD (2018) was used run comparative analyses based on closely related populations found on the database (Table 5). Study sample Y-STR profiles were searched for by inputting the samples in the YHRD to see if there were any haplotype matches amongst YHRD Yfiler® Plus samples. At the time of searching (11 February 2018) the Yfiler® Plus database contained 22 832 haplotypes, 108 sample populations, 36 national databases and 28 metapopulations. The studied population was compared to 14 other similar ethnic groups (Table 5) (selection of these groupings described in 1.2 Comparison with other available databases). AMOVA (Excoffier *et al.*, 1992) was carried in YHRD (2018) out to measure the variance in the number of repeat units

found across 27 Y-STRs, within and between populations, taking in to account the molecular relationship of alleles. Significant P values ($p = 0.05$) were also calculated at 10,000 permutations. Multi-Dimensional Scaling (MDS) analysis, based on Kruskal's non-metric MDS algorithm (Kruskal, 1964), was carried out to visualize the level of similarity between the different populations (Table 4 and Table 5). The analysis was based on pairwise Rst genetic distances. The Rst threshold was set as 0.05 for clustering, using 3 as a minimal cluster size.

5. Social aspects: Lineage inheritance in North Indian surname groups

Survey data (Figure A 1: Research survey form) was collected for the North Indian and random samples (224 samples) of this study (see Table 4). The data was analysed using IBM SPSS Statistics V21 (SPSS, 2012) to obtain relevant frequencies for comparison, track surname inheritance, and changes in surname, religion and geographical distribution of North Indian and random samples over generations. The results were further compared to genetic structure (AMOVA, PCoA, structure and Network) analyses based on the North Indian surname sub-groups.

RESULTS

The initial study population, including experimental samples collected for the purposes of this study (North Indian surname-based groups and random samples) and profiles included from the lab database (South Indian and Zulu surname-based groups) consisted of a total 414 samples. Five Indian samples were removed after the validation process using YHRD (2018), leaving a sample of 409 Y-STR DNA profiles, based on the Yfiler® Plus kit (27 loci). An example of a Y-STR profile output is shown in Figure A 3.

Null Alleles were observed at 77.8 % (21 out of 27) of the different loci analysed, with most contained in loci DYS391, DYS389II and DYS448, where they occurred in 4.89 %, 4.89 % and 4.65 % of the samples respectively (Table 6).

Table 6: Null Alleles per locus (N=409).

Distribution of samples containing null alleles at each locus.

Loci	Total	%	Samples null allele for a specific locus was found
DYS391	20	4.89	S11, S17, S38, M10, M18, M21, M31, M34, M36, M42, K01, K17, K23, K33, K44, K52, RM10, RW05, RW08 and RW10.
DYS389II	20	4.89	S25, S38, M08, M18, M30, M31, M34, M42, K01, K15, K17, K23, K38, K40, RH15, RM09, RM10, RW04, RW05 and RW10.
DYS448	19	4.65	S14, S21, S38, M10, M18, M30, M31, M35, M36, M42, M47, K01, K15, K20, K23, K38, K44, RW04 and RW10.
DYS533	16	3.91	S21, S36, S38, M13, M18, M36, M48, K01, K15, K17, K20, K23, K33, K40, RW04 and RW10.
DYS449	14	3.42	S04, S38, M18, M20, M31, M36, M42, K01, K15, K23, RH15, RM10 and RW05.
DYS627	13	3.18	S25, S38, M08, M18, M30, M36, K01, K17, K23, RH15, RM10, RW05 and RW10.
DYS392	12	2.93	S01, S38, M18, M20, M34, M35, K01, K15, K17, K23, RM10 and RW04.
DYS518	11	2.69	S01, S23, S38, M18, M20, M21, M36, K23, RH15, RM10 and RW10.
YGATAH4	10	2.44	S57, M01, M35, M36, M42, M47, K17, K38, K44 and RW04.
DYS385	8	1.96	M18, M31, K01, K15, K17, K23, RM10 and RW04.
DYS481	7	1.71	S38, M18, M42, K15, K33, RW04 and RW10.
DYS19	5	1.22	S57, M35, M36, M47 and K33.
DYS438	5	1.22	M18, M20, M36, K15 and K17.
DYS439	4	0.98	K01, K15, K23 and RW04.
DYS390	3	0.73	K23, RW04 and RW10.
DYS393	2	0.49	K01 and K15.
DYS458	2	0.49	K01 and K15.
DYS635	2	0.49	K23 and RW04.
DYS389I	1	0.24	M42.
DYS437	1	0.24	M42.
DYS460	1	0.24	K01.

1. Comparison of Y-STR marker sets via population and forensic genetics analyses

GD for the Yfiler® Plus kit, was reported across the total sample set for different loci and groups of markers loci ranging between 0.363-0.884 (Figure 11) with a mean \pm standard deviation of 0.712 ± 0.143 . The Yfiler® Plus kit contained 27 loci, which were also present in other kits with fewer loci, namely the Yfiler® (23 loci), MHT (9 loci) and RMu (7 loci) marker sets with overall GD's of 0.673 ± 0.147 , 0.689 ± 0.184 , and 0.821 ± 0.034 respectively. The overall diversity was the highest for locus DYS449 (found in the Yfiler® Plus and RMu marker sets) and lowest for locus DYS91 (found in the Yfiler® Plus, Yfiler® and MHT marker sets). RMu markers, had the highest overall GD (loci > 0.75, mean GD = 0.821) (Figure 11).

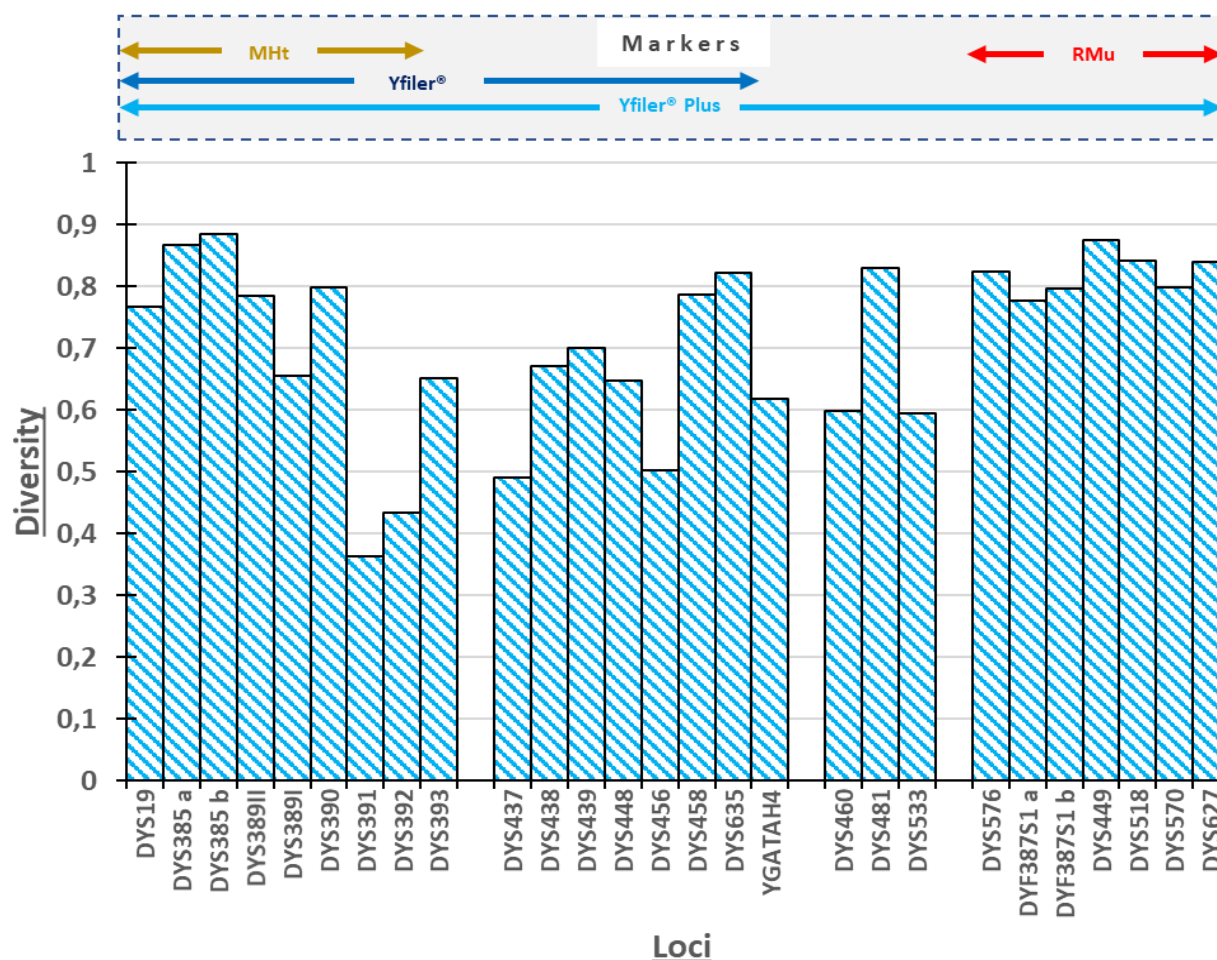


Figure 11: GD per locus, for 27 loci and four different marker sets (N=409).

GD is reported for all 27 Y-STR loci amplified by the Yfiler® Plus PCR amplification kit. Also indicated are the markers contained in the Minimal Haplotype (MHT), Yfiler®, Yfiler® Plus, and Rapidly Mutating (RMu) markers sets.

Allelic Patterns (Table 7) were reported for the overall sample set based on different Y-STR marker sets. Per locus, the RMu marker set had the highest mean number of alleles with frequency > 5 % (5.86), mean number of effective alleles (Ne) (5.78) and GD (0.82), whereas the MHT marker set showed the highest mean number of alleles and private alleles (12.11). The Yfiler® kit, which contained the second highest number of loci, yielded the lowest values for all of the above variables. The Percentage of polymorphic loci was 100% for all marker sets. Allelic Patterns (Table 7) were reported for the overall sample set based on different Y-STR marker sets. Per locus, the RMu marker set had the highest mean number of alleles with frequency > 5 % (Nf = 5.86), mean number of effective alleles (Ne = 5.78) and GD (0.82), whereas the MHT marker set showed the highest mean number of alleles and Private alleles (PA = 12.11). The Yfiler® kit, which contained the second highest number of loci, yielded the lowest values for all of the above variables. The percentage of polymorphic loci was 100% for all marker sets.

Table 7: Allelic patterns for the overall study samples (N=409).

Loci (N) = Number of loci. Na = Mean no. alleles per locus. Nf = Frequency of alleles > 5 %. Ne = Mean no. effective alleles per locus. Private Alleles =. Number of alleles found only in a single sample population; GD= genetic diversity. % P = Percentage of polymorphic loci. MHT = Minimal Haplotype markers. RMu = Rapidly Mutating markers.

Marker	Loci (N)	Na	Nf	Ne	PA	GD	% P
Yfiler® Plus	27	11.04	4.67	4.32	11.04	0.71	100%
Yfiler®	17	10.65	4.24	3.83	10.65	0.67	100%
MHT	9	12.11	4.67	4.34	12.11	0.69	100%
RMu	7	12.00	5.86	5.78	12.00	0.82	100%

Forensic parameters (Table 8) were reported for the overall study sample. Haplotypes were observed between one and six times. HD for all markers was 0.999. The Yfiler® Plus kit had the highest number of observed haplotypes (345) and DC (0.844), whereas it had the lowest MP. In contrast, the RMu marker set had the lowest number of observed haplotypes (345) and highest MP (0.0038).

Table 8: Forensic genetic parameters for the sample group (N=409) based on different marker sets.

HD = Unbiased haplotype diversity; MP = Haplotype match probability and DC = discrimination capacity (formulas in method section). Mht = Minimal haplotype markers. RMu = Rapidly Mutating markers.

Marker sets	Number of observed haplotypes							HD	MP	DC
	Total	Once	Twice	3 X	4 X	5 X	6 X			
Yfiler® Plus	345	281	64	0	0	0	0	0.999	0.0032	0.844
Yfiler®	334	261	72	0	1	0	0	0.999	0.0034	0.817
MHt	313	231	74	4	2	0	0	0.999	0.0036	0.765
RMu	315	234	73	6	0	1	1	0.999	0.0038	0.77

2. Genetic structure analyses based on population sub-groupings

The analyses of genetic structure among sample groups and sub-grouping were based on 399 samples (white ethnic group samples removed) overall, with lower samples for some comparisons, depending on availability (Table 9).

Table 9: Sample sizes used in analyses of genetic structure among various subgroups of an overall sample of 399 Indian and Zulu samples.

Analysis based on	Comparisons between	Total (N)	Population sub-groupings
1. Ethnicity	Indians vs Zulus	399	294 Indian and 105 Zulu individuals
2. Region of origin	North vs South Indians	294	193 North Indian and 101 South Indian individuals
3. Religion	Hindu vs Muslim Indians	294	234 Hindu and 60 Muslim individuals
4. Language	Hindi vs Tamil Indians	234	133 Hindi and 101 Tamil-speaking individuals
5. Religion and Language	a. Hindi vs Muslim Indians	193	133 Hindi, 60 Muslim (Urdu) and 101 Tamil-speaking individuals
	b. Muslim vs Tamil Indian	161	

2.1. Genetic Structure based on population sub-groupings

AMOVA was carried out for all comparisons listed in Table 9. Molecular variance among groups ranged from 1% to 7%, with the highest among group variance (7%) being ethnically based and occurring between the Indian and Zulu sub-grouping, and the lowest (1%) being religion-based (Hindus vs Muslims) and Religion/Language-based (Hindis vs Muslims). The PhiPT values (an estimate of differentiation among groups) were all significant (range = 0.0001 to 0.001) and

ranged from 0.074 to 0.0009 (Table 9). PhiPT is an analogue of Fst, where the higher the value the greater the genetic differentiation; PhiPT values ranging from 0.15–0.25; 0.05–0.15 and values below 0.05 reflect great, moderate and little genetic differentiation respectively (Yaacov *et al.*, 2012). Thus, although PhiPT was significant for all comparisons (based on ethnicity, region of origin, religion, language, and a combination of religion and language (Table 9), the greatest amount of structure, equating to a moderate level of structure, was found among Indian and Zulu ethnic groups. Corresponding this, the Nm value was the lowest (6.261) for the ethnicity comparison, and greatest for the religion comparison (Hindu vs Muslim) (52.709).

Table 10: AMOVA Results

Distribution of molecular variance among and within sample groups, PhiPT, Nm (effective number of migrants), Df (degrees of freedom) and P (significance) value. AMOVA was based on 9999 permutations. Bold* = significant p-values (i.e <0.05). AMOVA based on 9999 permutations using GenAlEx v6.5 (Peakall and Smouse, 2012).

Comparison based on	Sample groups	Molecular Variance		PhiPT AMOVA values			
		Among groups	Within groups	PhiPT	Nm	Df	p-value
1. Ethnicity	1. Indian vs Zulu	7%	93%	0.074	6.261	398	0.0001
2. Region of origin	2. North vs South Indian	3%	97%	0.029	16.462	293	0.0001
3. Religion	3. Hindu vs Muslim	1%	99%	0.009	52.709	293	0.0010
4. Language	4. Hindi vs Tamil	3%	97%	0.034	14.011	233	0.0001
5. Religion and Language	5a. Hindi vs Muslim	1%	99%	0.011	44.874	192	0.0010
	5b. Muslim vs Tamil	3%	97%	0.027	18.160	160	0.0001

PCoA analyses revealed that 12.97 to 15.24% of the variance was explained by the first two components (Table 11), which had the highest eigenvalues.

Table 11: PCOA via covariance: Eigen values and percent variance explained for the principal components 1 and 2 for comparisons among different subgroups of the overall sample (n=399).

Population group	Eigen value		Variance explained		
	PCoA 1	PCoA 2	PCoA 1	PCoA 2	Total
1. Indian vs Zulu	13.98	11.45	7.13%	5.84%	12.97%
2. North vs South Indian	12.99	10.46	7.60%	6.12%	13.72%
3. Hindu vs Muslim	12.99	10.46	7.60%	6.12%	13.72%
4. Hindi vs Tamil	12.00	9.93	7.94%	6.57%	14.51%
5a. Hindi vs Muslim	12.44	8.55	9.03%	6.21%	15.24%
5b. Muslim vs Tamil	11.16	7.34	8.82%	5.80%	14.62%

PCoA is used here to visualise genetic distance and relatedness among sample groups. The PCoA plots showed varying degrees of overlap and separation of the sample groups being compared (Figure 12). The greatest degree of separation and correspondingly least overlap was observed for the Indian vs Zulu populations; in this case the separation was much greater than the overlap, and there were distinct clusters of Indian only, and Zulu only data points (Figure 12.1). In the comparison of North vs South Indians (Figure 12.2), there was a higher degree of overlap, although a part of the North Indian groups did cluster separately from all South Indians. The language-based comparison plot for the Hindi vs Tamil groups (Figure 12.4) was similar to that of the North vs South Indians, with a separate clustering of Hindis in addition to the regions of overlap which contained most Tamils and some Hindis. There was a high degree of overlap of the Hindu and Muslim religious groupings (Figure 12.3), and of the Hindi and Muslim language-based groups (Figure 12.5a) and Tamil and Muslim language-based groups (Figure 12.5b).

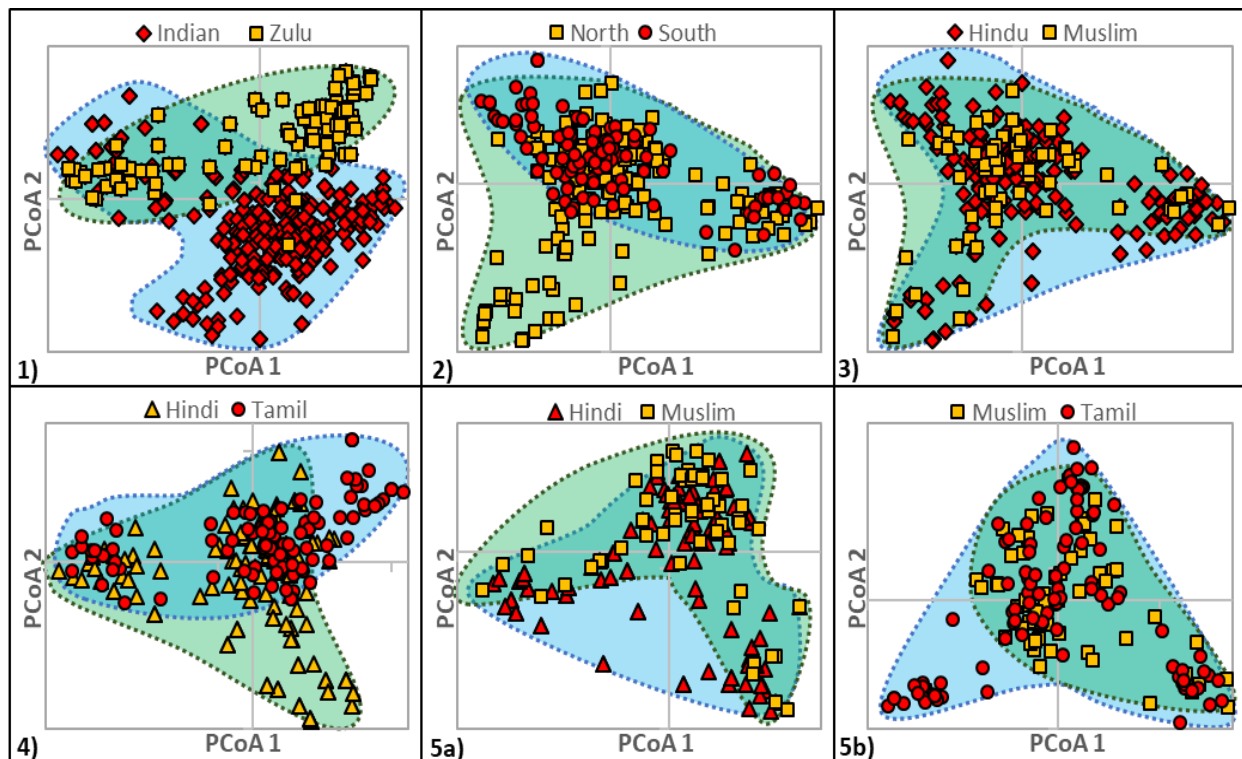


Figure 12: PCoA plot visualising genetic distance and relatedness amongst different sample sub-groupings.

Comparison plots were based on 1 - Ethnicity, 2 - Region of origin, 3 - Religion, 4 - Language, 5a and 5b - A combination of language and religion. Blue shading = area where red group only is found. Green area = area where yellow group only is found. Greenish blue area = zone where both yellow and red groups are found.

Genetic structure analyses were carried out in BAPS V6.0.1 (Corander *et al.*, 2008). The whole sample set ($n = 399$, Table 9) was analysed by this method, which attempts to identify the number of genetically similar groups (K) within the sample, and to assign individual samples to these groups.

The mixture analyses for clustering of individuals (Figure 13.1b) revealed 3 genetically different populations ($K = 3$), which were based on region of origin i.e. North Indians, South Indians and Zulu's. The mixture analyses for linked loci (Figure 13.1a) showed clear separation of the Zulu population, in comparison to the North and South Indians, which appeared to be more homogeneous. North Indians showed the presence of dark blue, red and yellow bands, which were not present in the South Indians. The neighbour joining tree (Figure 13.1c) showed that the Zulu's formed a distinct cluster, which was sister to a cluster containing the North and South Indians. Thus, North Indians and South Indians are more genetically similar to each other than they are to Zulu's, with the South Indians showing more mutational differences from the common ancestor of the Indian group.

Admixture was found in the sampled population (Figure 13.2). The admixture analyses for linked loci (Figure 13.2a) showed clear separation of the Zulu population from the North and South Indians. However, the admixture analyses for clustering of individuals (Figure 13.2b) revealed two genetically different populations ($K = 2$), namely Indians and Zulus. Admixture was observed in the subgroup with the surname Singh, which contained genetic elements from the Zulu population (K_2).

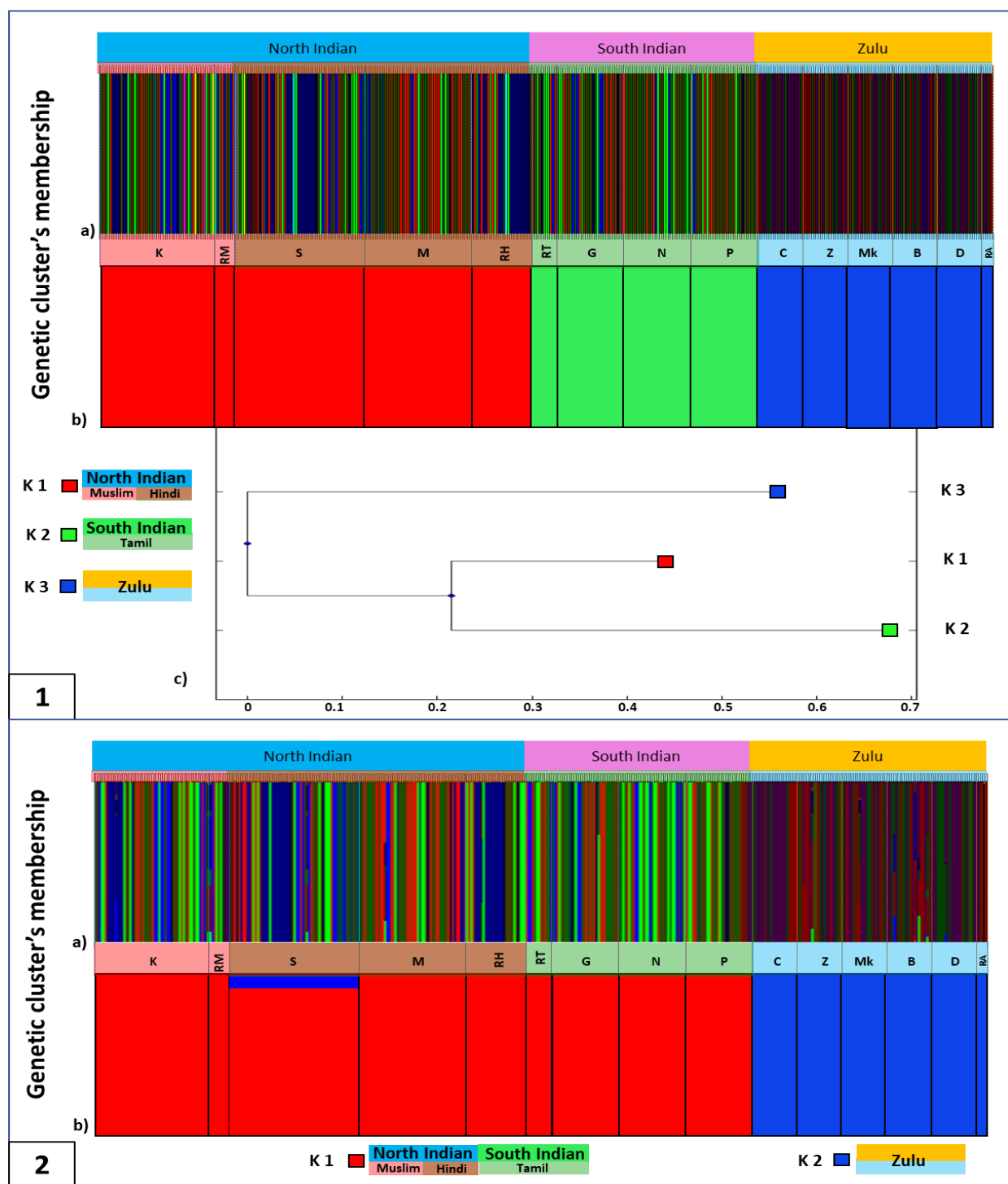


Figure 13: Bayesian analysis of genetic population structure within the entire sample group (n = 399).

Genetic clustering based on: 1. Mixture model for (a) Linked loci (where each vertical line represents an individual), (b) Clustering of individuals where K (number of genetic groups) = 3, and (c) neighbour joining (NJ) phylogram for clustering of individuals. 2. Admixture model for (a) Linked loci, and (b) Clustering of individuals and where K = 2. North Indians comprise Muslims with the surname Khan (K), and Hindus with the surnames Singh (S) and Maharaj (M). South Indians comprise Tamils, with the surnames Govender (G), Naidoo (N), and Pillay (P). The African Zulu group has the surnames Buthlezi (B), Cele (C), Dlamini (D), Mkhize (Mk) and Zulu (Z). R = Random samples with no pre-specified surname, ethnicity, religion or region of origin; H = Hindi; M = Muslim; T = Tamil; A = African.

Haplotype network analyses of the overall sample were based on a total 409 samples and 345 haplotypes (Figure 14). The Zulu group form a major cluster (SW direction, green) and a minor cluster (NE direction, green) in combination with a group of random Hindis and Singhs (pink). The Indian group forms two major clusters oriented towards the NW and E direction of the network plot. The white samples tend to cluster to the northwest of the plot, along with Indians of the surname Maharaj and Khan.

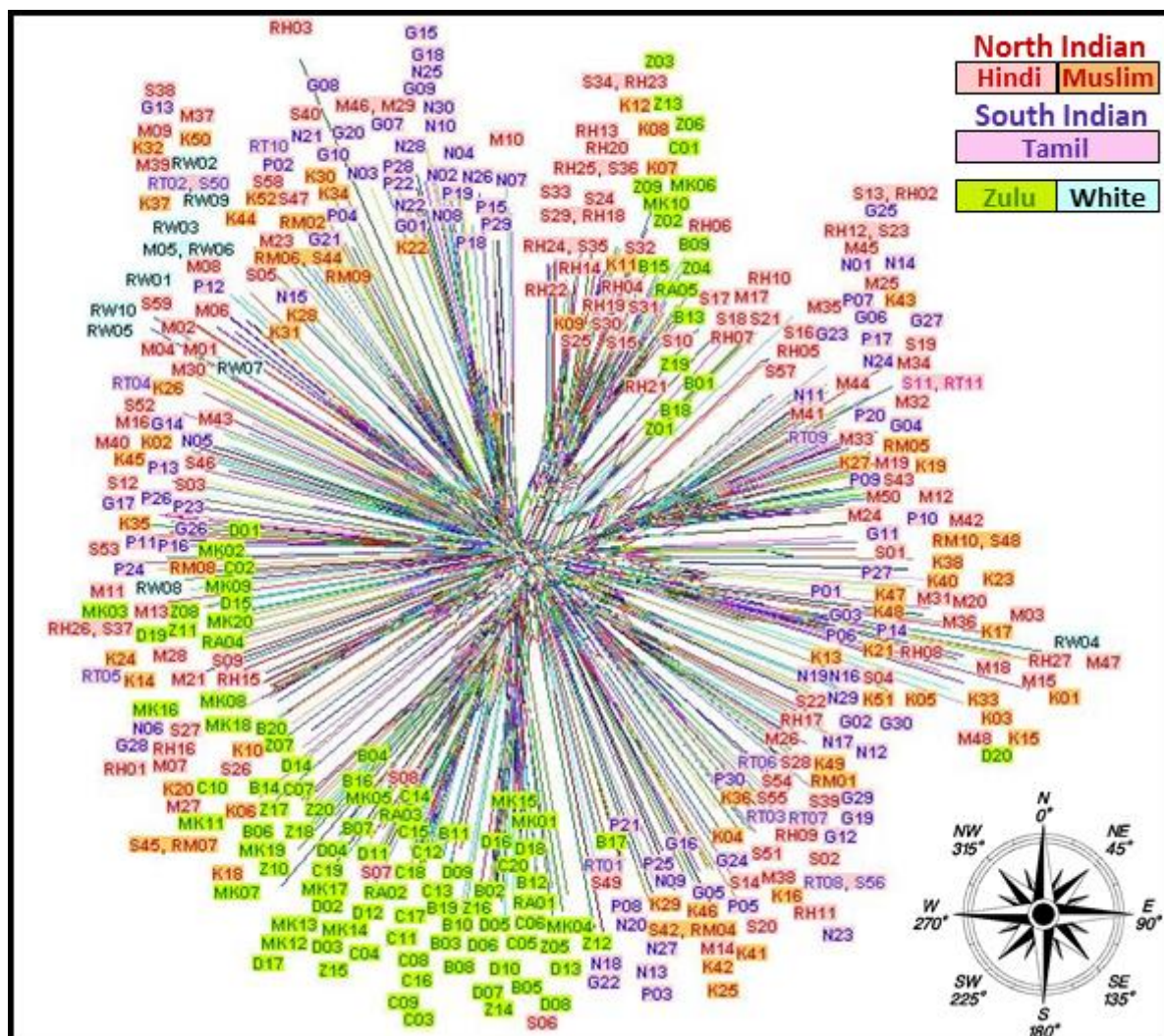


Figure 14: Haplotype network for the entire sample group, including whites (n = 409).

The plot was based on the Neighbor Net method and generated in SplitsTree V4.14 (Huson and Bryant, 2005). Each line represents a haplotype. North Indians comprise Hindis with the surnames Singh (S) and Maharaj (M), and Muslims with the surname Khan (K). South Indians comprise of Tamils with the surnames Govender (G), Naidoo (N), and Pillay (P). Zulus have the surnames Buthelezi (B), Cele (C), Dlamini (D), Mkhize (Mk) and Zulu (Z). R = Random, H = Hindi, M = Muslim, T = Tamil, A = African and W = White.

2.2. Population and forensic genetics based on population sub-groupings

Allele frequencies were reported for all 27 loci across the sample sub-groups (Table A 3.1) and ranged from 0.001-0.952. The highest allele frequency for the Zulu sub-group was found in DYS437 (allele 14, 0.952). Allele 10 of locus DYS391 displayed the highest allele frequency for the overall sample group (0.776), Indian (0.730), South Indian (0.802), Hindu (0.739), Muslim (0.692), and Tamil (0.802) groups. Allele 11 of DYS392 displayed the highest frequency in the case of the North Indian (0.757) and Hindi (0.797) groups.

Allelic Patterns (Table 12) were reported for the overall sample group, as well as for different sub-groups. The number of alleles (Na), frequency of alleles $\geq 5\%$ (Nf) and number of effective alleles (Ne), private alleles (PA) and GD, ranged between 6.593 - 11.222, 3.444 - 4.741, 3.062-4.329, 0.704 - 11.222, 0.591 - 0.719, respectively. Percentage of polymorphic loci was 100% for all groupings. The overall sample group had the highest Na, Nf, Ne, and PA, whereas the Indian population had the highest GD (0.712). The Zulu population had the lowest Na, Nf, Ne, H, and % P, whereas the South Indian population had the lowest PA (0.926). When comparing the different sub-groupings, the GD was higher for (1) Indians as compared to Zulus, (2) South as compared to North Indians, (3) Hindus as compared to Muslims, (4) Tamil as compared to Hindi, (5) Muslim as compared to Hindi.

Table 12: Allelic patterns for the overall study sample (n = 399) and sub-grouping within this sample.

The mean values of Na, Nf, Ne, PA, H and % P for each sub-group are given in the table. N = Number of individuals, Na = No. Alleles, Nf = Frequency of alleles $\geq 5\%$, Ne = No. Effective Alleles, PA = Private Alleles i.e. the number of alleles found only in a single sample population, GD = average genetic diversity. % P = Percentage of polymorphic loci. Total = Overall sample group. Superscript numbers represent sample groups based on: 1 = Ethnic group; 2 = Region; 3 = Religion of origin in India; 4 = Language. 5 = Religion and Language.

	Total	Indian ¹	Zulu ¹	North ²	South ²	Hindu ³	Muslim ^{3,5}	Hindi ^{4,5a}	Tamil ^{4,5b}
N	399	294	105	193	101	234	60	133	101
Na	11.222	10.148	6.593	9.222	7.148	9.444	6.778	8.333	7.148
Nf	4.667	4.444	3.444	4.333	4.407	4.556	4.741	4.370	4.407
Ne	4.329	4.263	3.062	4.081	4.121	4.227	4.038	3.941	4.121
PA	11.222	4.630	1.074	3.000	0.926	3.370	0.704	2.296	1.111
GD	0.712	0.719	0.591	0.706	0.710	0.715	0.713	0.692	0.710
% P	100%	100%	100%	100%	100%	100%	100%	100%	100%

Forensic genetic parameters (Table 13) were reported for the whole sample group, as well as for sub-groupings within this. Most commonly haplotypes occurred once only, although some occurred a maximum of twice within a sample grouping. The Zulu (African) (n = 105) and Muslim (n = 60) sub-groupings contained only unique haplotypes. For the following groups, some haplotypes occurred twice (numbers in brackets = number of haplotypes that occurred twice); Indian (n = 51), North Indian (n = 39), South Indian (n = 2), Hindu (n = 37), Hindi (n = 25), and Tamil (n = 2). HD ranged between 0.997 and 1.000. The MP was the highest for the Muslim group (0.0167) and lowest for the overall sampled population (0.0032). DC ranged between 0.798 and 1.00.

Table 13: Forensic genetic parameters for the overall study sample (n = 399) and sub-grouping.

HD = Unbiased haplotype diversity, MP = Haplotype match probability, and DC = discrimination capacity (formulas in method section). Superscript numbers represent sample groups based on: 1 = Ethnicity; 2 = Region of origin in India; 3 = Religion; 4 = Language; 5 = Religion and Language.

Groups	Observed haplotypes (n)			HD	MP	DC
	Total	Once	Twice			
Overall	345	281	64	0.999	0.0032	0.844
Indian¹	243	192	51	0.999	0.0046	0.827
Zulu¹	105	105	-	1.000	0.0095	1.000
North²	193	115	39	0.998	0.0073	0.798
South²	99	97	2	1.000	0.0103	0.980
Hindu³	197	160	37	0.999	0.0056	0.842
Muslim^{3,4,5}	60	60	-	1.000	0.0167	1.000
Hindi^{4,5}	108	83	25	0.997	0.0103	0.812
Tamil^{4,5}	99	97	2	1.000	0.0103	0.980

3. Genetic genealogy: Surname-based genetic analyses

The analyses below were based on 347 samples, divided into three surname-based groups (Table 14): North Indian surnames (Khan, Maharaj, and Singh); South Indian surnames (Govender, Naidoo, and Pillay); and Zulu surnames (Buthelezi, Cele, Dlamini, Mkhize, and Zulu).

Table 14: Sample groupings used in surname-based genetic analyses (n = 347)

Groups	Total (n)	Language	Population sub-groupings		
			Surname	Code	(N)
North Indian	157	Hindi	Khan	K	51
			Maharaj	M	48
		Muslim	Singh	S	58
South Indian	90	Tamil	Govender	G	30
			Naidoo	N	30
			Pillay	P	30
Zulu (African)	100	Zulu	Buthelezi	B	20
			Cele	C	20
			Dlamini	D	20
			Mkhize	Mk	20
			Zulu	Z	20

3.1. Social aspects: Lineage inheritance in North Indian surname groups

Social data was based on individuals who answered the questionnaire (Figure A 1: Research survey form) related to the religion, language, inheritance of surname and geographical distribution of the participants and their forefathers, and including questions requesting basic demographic information. This consisted of 224 males i.e. 161 with the surnames Khan, Maharaj and Singh, as well as 63 random individuals.

Most of the sampled population were Indians (93.3 %), whereas 4.5 % percent were Caucasian (White) and 2.2 % percent were African. 52.1 percent of the sample members were aged between 18 and 27 years, and the rest between 18 and 87 years old (Figure 15). Of the sample members, 92.8 % were from KZN province and 77.6% were from the Durban Metropolitan Area (Figure 15).

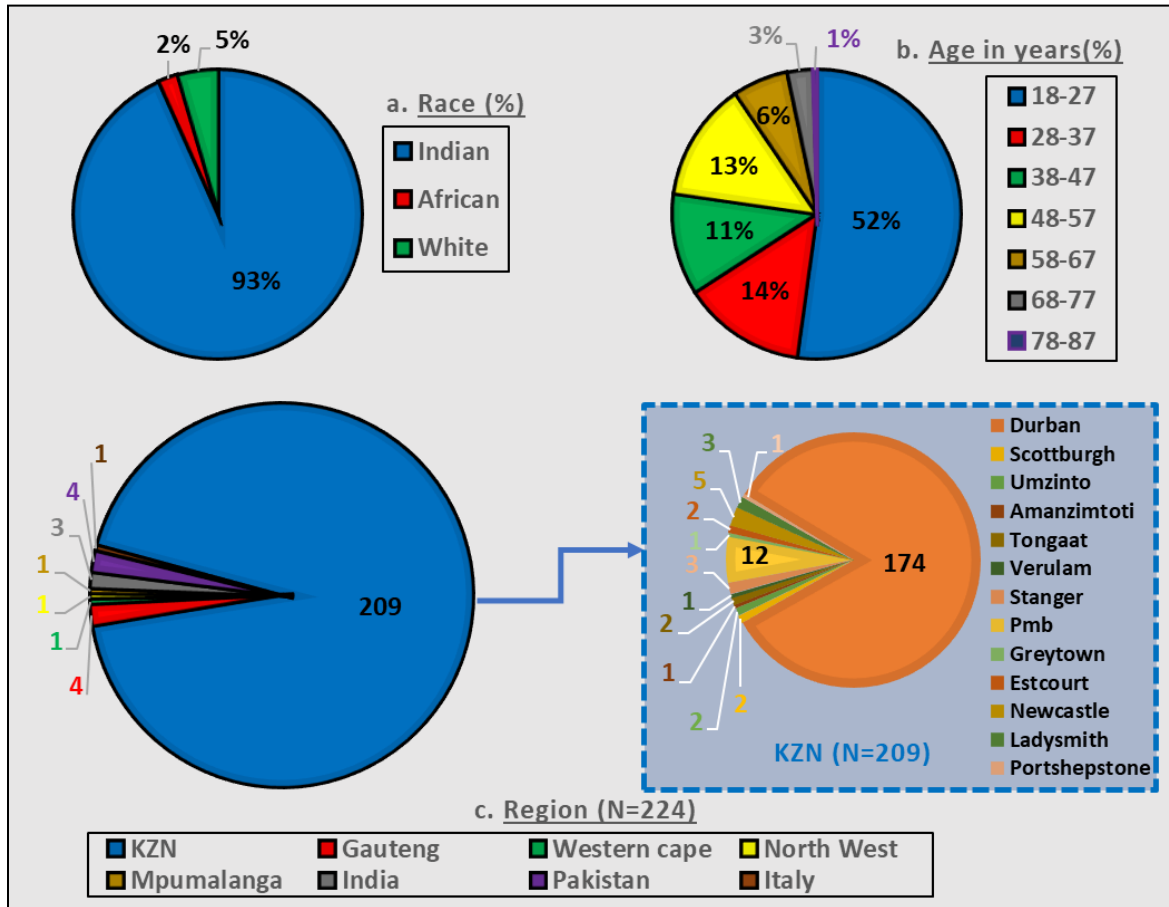


Figure 15: Race (a), age (b) and region of birth (c) of study samples (n = 224).

Survey data as indicated by responses to a questionnaire (Figure A 1: Research survey form).

Most of the individuals sampled were from the 4th (40.2 %) and 5th generation (33.2 %) descended from ancestors who immigrated to KZN from India, i.e. their great-great grandfathers or great-great-great grandfathers came from India (Figure 16). Most Indian individuals in the 18- 27-year age group (23.9 %) belonged to the 5th generation in SA (Figure 16).

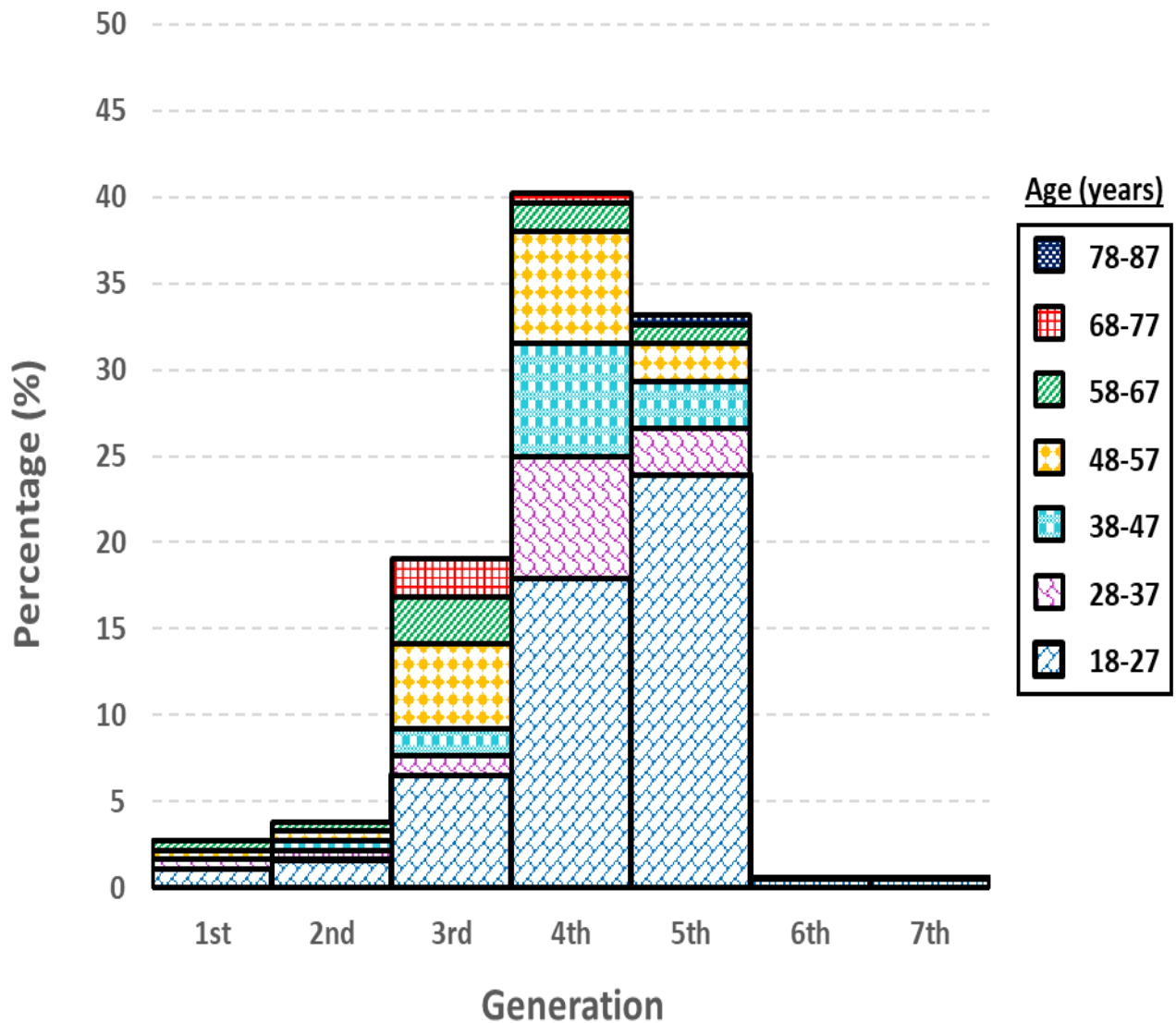


Figure 16: Frequency distribution and age composition of the generations of Durban-area Indians in the sample since they arrived in KZN from India.

Generations are labelled as: 1st = Current generation came to KZN from India, 2nd = Father immigrated from India, 3rd = Grandfather, 4th = Great grandfather, 5th = Great-great grandfather, 6th = Great-great-great grandfather and 7th = Great-great-great-great grandfather. The bar representing each generation is subdivided according to the current age profile of the members of that generation.

Survey responses (based on Figure A 1: Research survey form), showed that only 15.4 % Indians (i.e. 32 out of 208 Indians) had relatives in India, with whom they were in contact. These relatives were from a wide range of areas with most Muslims being from Allahabad (6.3 %) and Hindis from Bihar (9.4%) and Surat (15.6 %).

Generally, all groups are most likely to share their surnames with their children and paternal forefathers, and unlikely to share their surnames with their mothers, as might be expected in cultures/societies where surnames are paternally transmitted (Figure 17). Sample members are more likely than not to live in the same city as their children, paternal and maternal forefathers, with the exception of random Africans, who do not generally live in the same city as their children. Sample members are reasonably likely to share their religion and language with their children, maternal and paternal forefathers (Figure 17).

Overall Random Hindi (RH) and Random Tamil (RT) sample individuals were slightly less likely to share their surnames with their children (Figure 17.1a). All Random Muslim (RM), Random African (RA) and Random White (RW) males indicated that they passed on their surnames to their child (Figure 17.1a). All Singh (S) and (RM) males inherited their surname from their paternal forefathers (Figure 17.1b). Sample members with the surname Maharaj (M) were most likely to have inherited their surnames from their maternal forefathers (63 %), followed by RW (55.6 %), whereas no RM and RT inherited their surnames from their maternal forefathers (Figure 17.1.c). Most Indians (> 50%) were born and brought up in the same city as their child (Figure 17.2), however, the reverse was observed for RA (0 %) and RW (33.3 %). More than 50% of the sample was born and brought up in the same city as their paternal (Figure 17.2b) and maternal forefathers (Figure 17.2c), with the exception of RH, of whom only 37.55 % were born and brought up in the same city as their maternal forefathers (Figure 17.2c). All sample groups were more likely than not to share religion (Figure 17.3) and traditional language (Figure 17.4) with their children and paternal and maternal forefathers (greater than 50%).

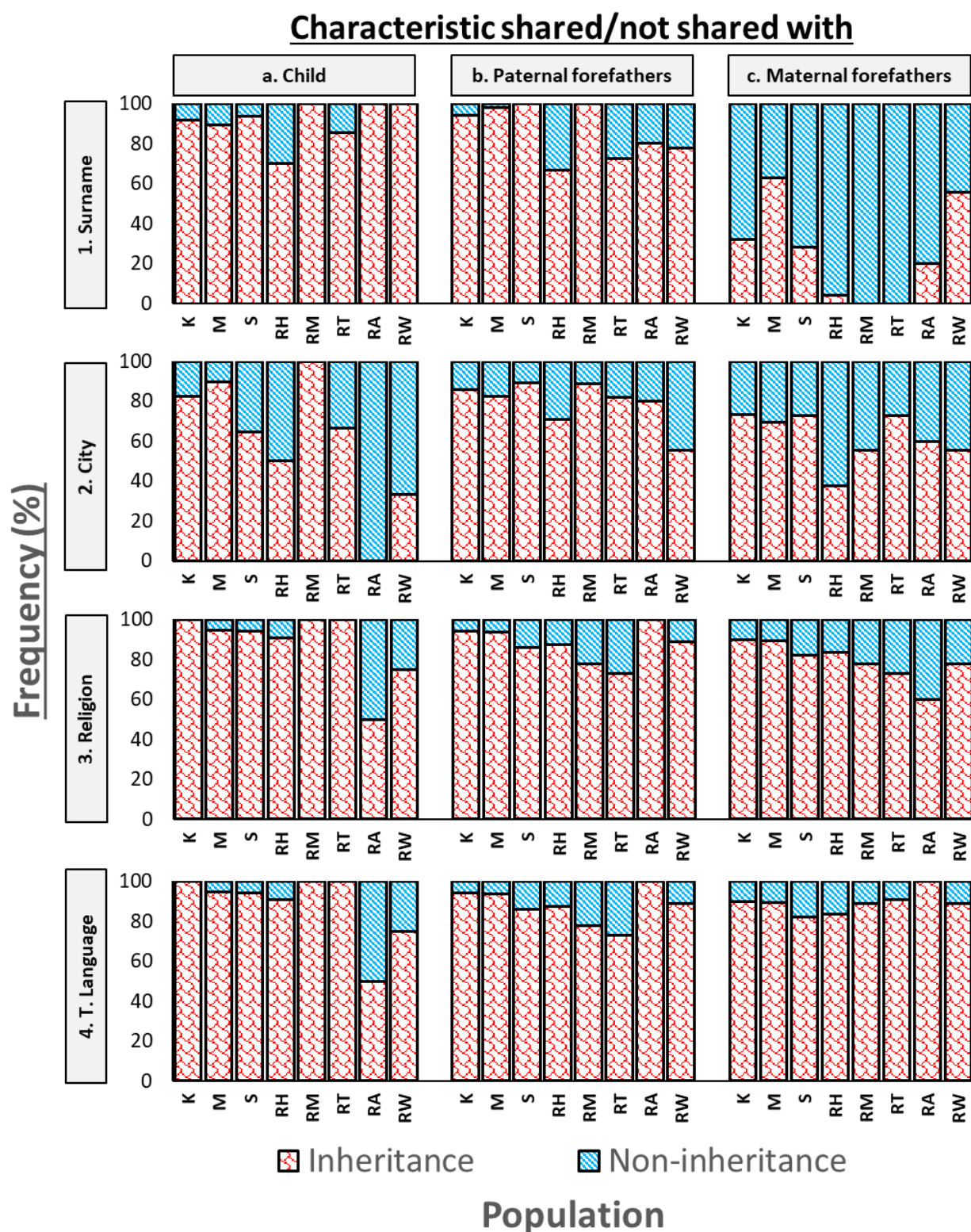


Figure 17: Likelihood that members of different sample groupings will share surname, city, religion and language with their children, paternal forefathers and maternal forefathers.

K = Khan, M = Maharaj, S = Singh; Random, H = Hindi, M = Muslim, T = Tamil, A = African, and W = White.

3.2. Surname-based genetic structure and surname inheritance analyses

In an analysis of a sample set based on all samples subdivided into 11 surname-based groups (Khan, Maharaj, Singh, Govender; Naidoo, Pillay; Buthelezi, Cele, Dlamini, Mkhize, Zulu), 92% of the molecular variance occurred within surname-based groups, whereas 8 % occurred between these groups. AMOVA analyses were then carried out for the North Indian, South Indian and Zulu groups in order to search for genetic structure among the surnames within each group (Table 14). Molecular variance among the surnames within these groups ranged from 1% to 9%, with the highest among group variance (9%) occurring between Zulu surnames, and the lowest (1%) occurring between South Indian surnames (Figure 18).

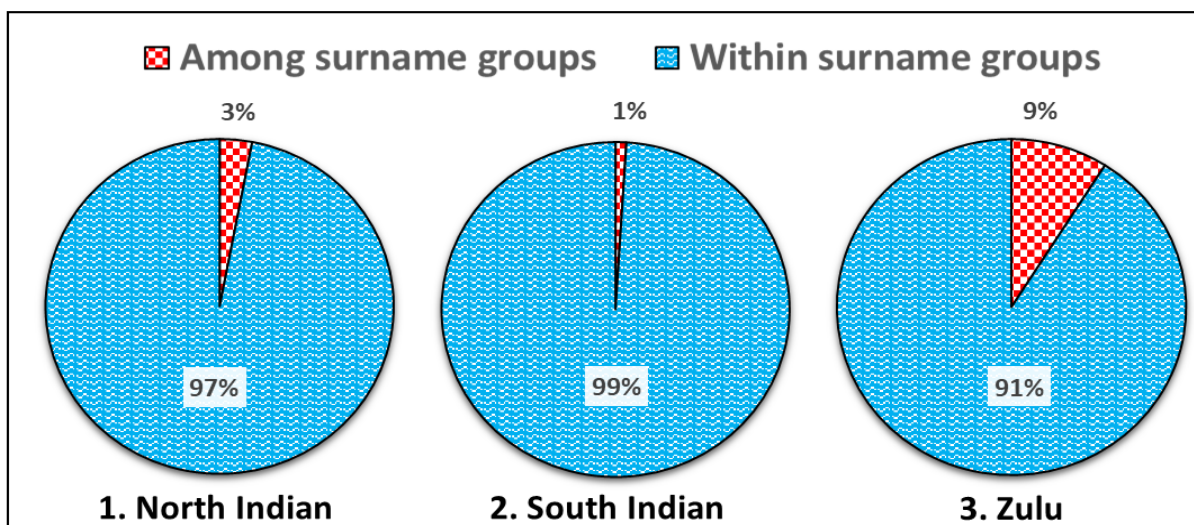


Figure 18: AMOVA: Distribution of molecular variance among and within surname-based groups of North Indians, South Indians and Zulus (n = 347).

A pairwise AMOVA (Table 15) was carried out to show which surnames were significantly genetically different from each other. (1) Significant genetic differences were observed among all pairwise combinations of the North Indian surnames Khan, Maharaj and Singh. (2) There were no significant pairwise differences among the South Indian surnames Govender, Naidoo and Pillay. (3) All pairwise combinations of the Zulu surnames Buthelezi, Cele, Dlamini, Mkhize and Zulu were significantly different from each other.

Table 15: Pairwise AMOVA for 27 Y-STR loci for surname-based groups.

Bold* = significant p -values (i.e <0.05); Nm = effective number of migrants; p -value = significance value. Pairwise AMOVA was based on 9999 permutations using GenAlEx v6.5 (Peakall and Smouse, 2012).

Surname group	Surname comparison			Nm	PhiPT	p -value
1 North Indian	Khan	vs	Maharaj	18.4889	0.02633	0.000
	Khan	vs	Singh	35.706	0.01381	0.005
	Maharaj	vs	Singh	12.4955	0.03847	0.000
2. South Indian	Govender	vs	Naidoo	69.8957	0.0071	0.173
	Govender	vs	Pillay	59.3406	0.00836	0.125
	Naidoo	vs	Pillay	74.9411	0.00663	0.175
3. Zulu	Buthlezi	vs	Cele	4.34366	0.10323	0.001
	Buthlezi	vs	Dlamini	6.93244	0.06727	0.003
	Buthlezi	vs	Mkhize	10.9544	0.04365	0.014
	Buthlezi	vs	Zulu	12.3096	0.03903	0.043
	Cele	vs	Dlamini	4.00134	0.11108	0.000
	Cele	vs	Mkhize	4.36141	0.10285	0.002
	Cele	vs	Zulu	2.3955	0.17268	0.000
	Dlamini	vs	Mkhize	4.67731	0.09658	0.001
	Dlamini	vs	Zulu	4.18226	0.10679	0.001
	Mkhize	vs	Zulu	14.2387	0.03392	0.039

The first two principal components, PCoA 1 and 2 (Table 16), which had the highest eigenvalues, explained the greatest percent of the variance among groups with different surnames. The total molecular variance explained for the surname-based groups ranged between 13.73 and 28.90% (Table 16).

Table 16: PcoA: Percent variation explained and eigenvalues for Indian, North Indian, South Indian and Zulu surname-based groups.

Surname group	Eigen value		Variance explained		
	PCoA 1	PCoA 2	PCoA 1	PCoA 2	Total
1. Indian Surnames	12.13	8.40	7.69%	5.33%	17.71%
a. North Indian	9.71	7.54	7.73%	6.01%	13.73%
b. South Indian	11.12	7.08	12.21%	7.78%	19.99%
2. Zulu Surnames	15.82	8.01	19.19%	9.71%	28.90%

The PCoA plot showed varying degrees of overlap between the different surname-based groups compared (Figure 19). The least separation was observed for South Indian surnames (Govender, Naidoo and Pillay), which showed an almost complete overlap in their distributions (Figure 19.1b). Within the other three groups (Indian, North Indian and South Indian surnames), there was a degree of separation between surname-based groups, although in all cases a greater degree of overlap was observed. In Figure 19.1 and Figure 19.1a, a separate group of Singhs and Khans was observed, distinct from a common group comprising sample members with surnames Khan, Maharaj and Singh. Figure 19.1 also contained a separate group comprising primarily sample members with the surnames Govender and Naidoo. Figure 19.2 (Zulu surnames) showed a large degree of overlap of all five surnames, but three smaller separate areas where sample members with the surnames Cele, Dlamini and Mkhize were located.

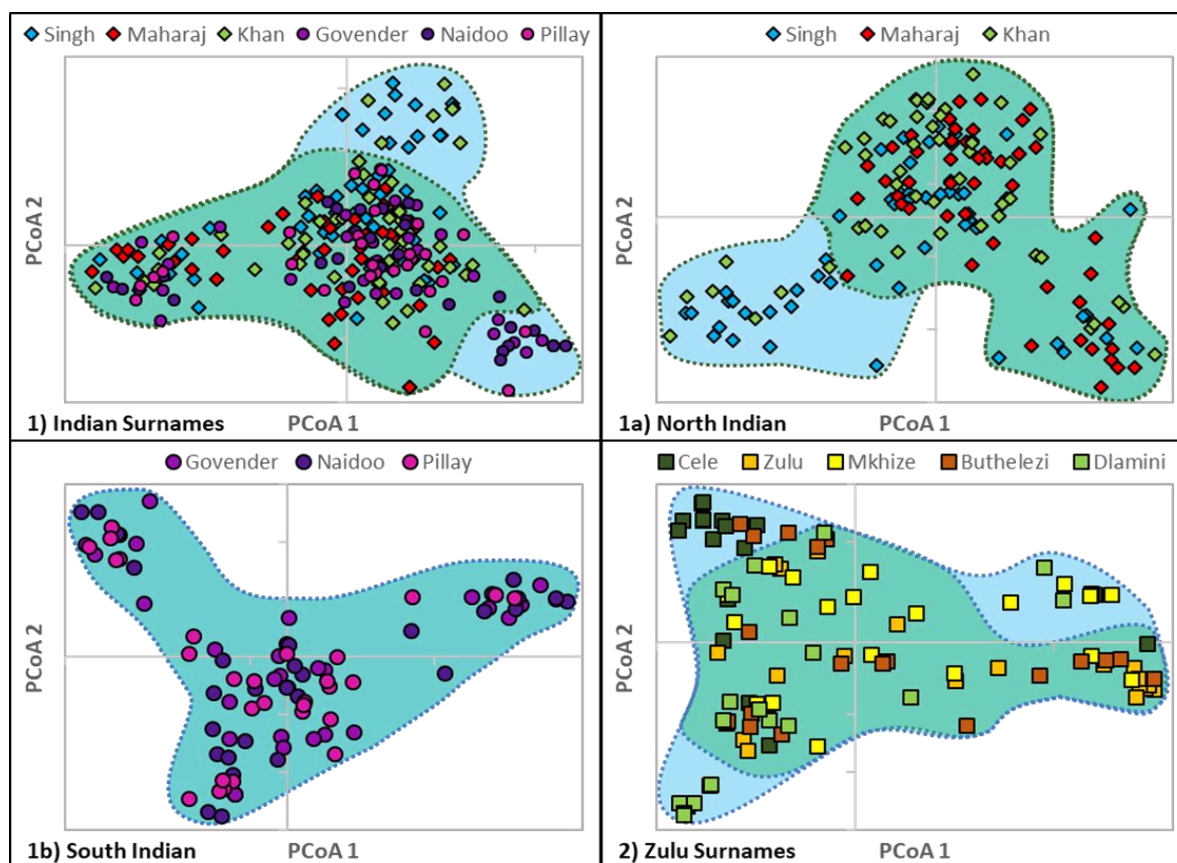


Figure 19: PCoA plot for surname-based sample groupings.

The Indian surname group (19.1) comprised the surnames Govender, Naidoo and Pillay, Maharaj, Singh and Khan. The North Indian group (19.1a) comprised the surnames Maharaj, Singh and Khan whereas the South Indian group (19.1b) comprised the surnames Naidoo, Govender and Pillay. The Zulu surname group (19.1c) comprised the surnames Buthelezi, Cele, Dlamini, Mkhize and Zulu. Overlap between all groupings is represented by the greenish-blue area. A degree of surname-based separation is represented by blue area.

Bayesian Estimation of Population Structure among surname-based groups was carried out using BAPS V6.0.1 (Corander *et al.*, 2008). Region-based analyses of genetic structure showed North Indians, South Indians and Zulus to be genetically distinct groups (Figure 13.1b). Mixture analyses for linked loci for were carried out separately for these three groups in an attempt to detect whether or not genetic structure existed among surname-based groupings within them.

The mixture analysis for North Indians (Figure 20.1) showed the presence of a large cluster of royal blue bands for Singhs and Khans ($n = 15$ and $n = 7$ respectively), but only one royal blue band for Maharajs. The Maharajs show a presence of a large cluster of blackish-blue bands ($n = 16$), and a large cluster of khaki bands ($n = 14$), which were not observed for Singhs and Khans. Mixture analysis for South Indians (Figure 20.2) revealed little genetic separation of people with the surnames Naidoo, Govender and Singh. Mixture analysis for the Zulus (Figure 20.3) showed some relatively distinct patterns among the surname-based groups. The surname Cele was distinguished by membership of the red genetic group ($n = 10$), Zulu by the yellow genetic group ($n = 8$), Mkhize by membership of groups represented by two shades of green ($n = 10$) and Dlamini by purple and blue ($n = 15$). The surname Buthelezi appeared to have a more varied genetic group membership.

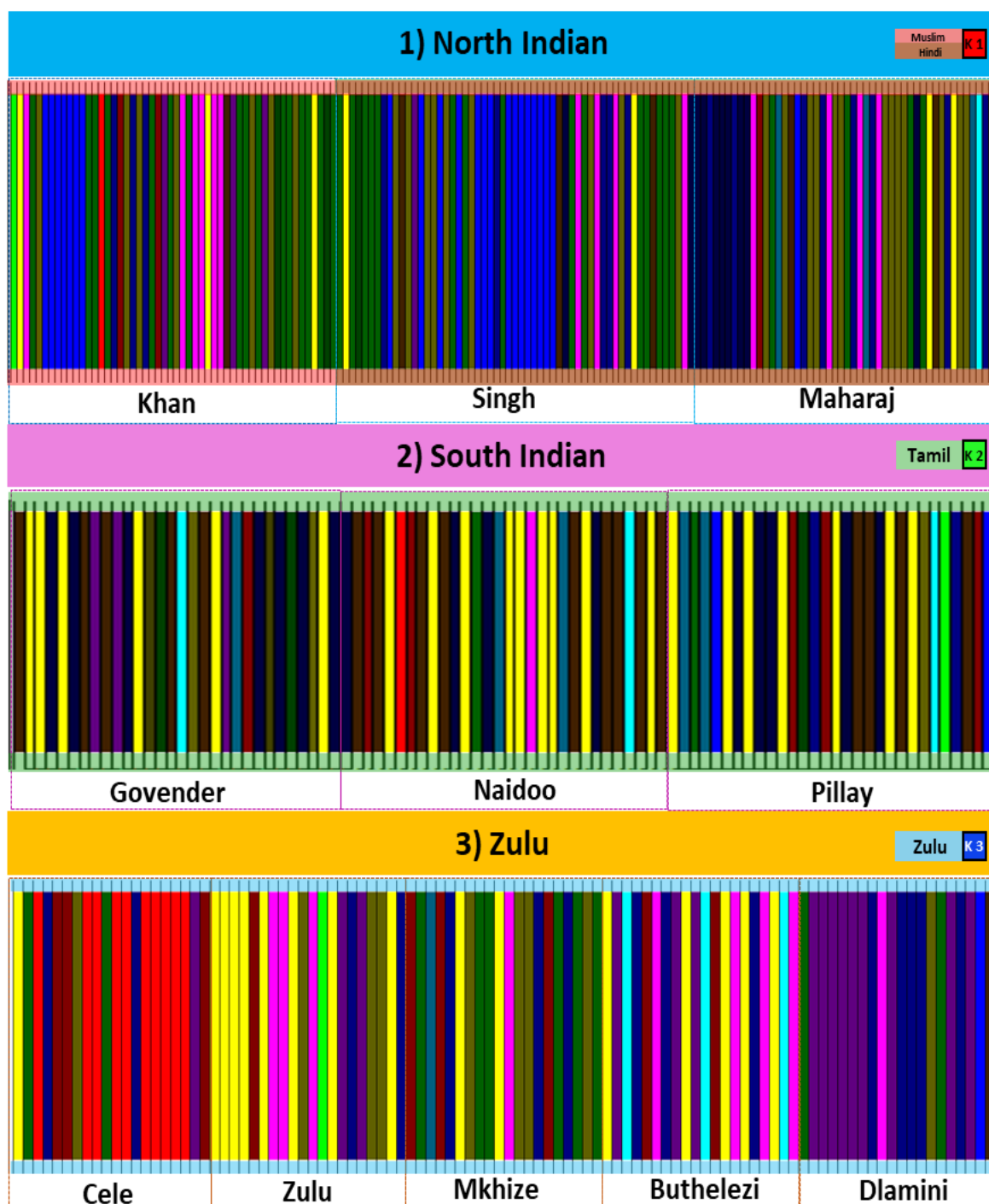


Figure 20: Bayesian analysis of population genetic structure for three sample groups containing different surname sets (n = 347).

Genetic clustering based on the Mixture model for Linked loci for three groups comprising different surname sets, i.e. 1. North Indian, 2. South Indian, and 3. Zulu surnames. North Indians comprise the surnames Khan (K), Maharaj (M), and Singh (S). South Indians comprise the surnames Govender (G), Naidoo (N), and Pillay (P). Zulus include the surnames Buthelezi (B), Cele (C), Dlamini (D), Mkhize (Mk) and Zulu (Z). Each colour represents the analysed sub-populations found within the population. Each line represents an individual.

Haplotype network analyses were based on a total 374 taxa belonging to surname-based groupings (Figure 21). In the North Indian group there were no major clades which comprised exclusively one surname, although there were two clades which were enriched for samples with the surname Maharaj (one approximately due north and the other due east, Figure 21a). There was also a cluster towards the northeast which included mostly samples with the surname Singh. The remainder of the clades contained a mixed distribution of the surnames Khan, Maharaj and Singh. The South Indian group (Figure 21b) showed very little structuring of clades, which generally contained mixtures of samples with the surnames Naidoo, Govender and Pillay. The Zulu group contained a clear 'Cele only' cluster slightly west of South (Figure 21c), a Dlamini rich clade to the northwest and a 'Zulu' rich clade to the southeast.

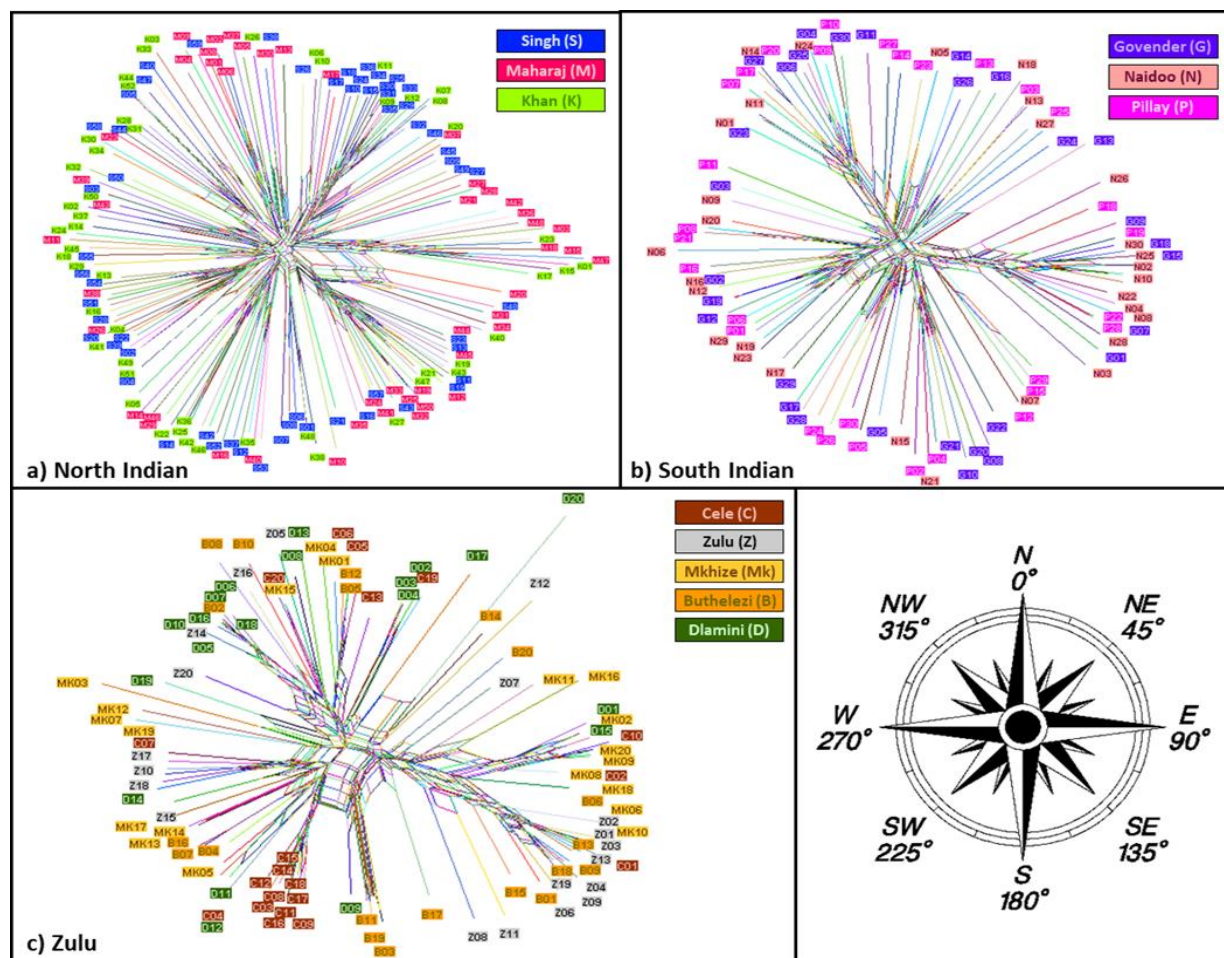


Figure 21: Haplotype networks for different surname-based groups (n = 347).

Haplotype networks based on the Neighbor Net method were constructed for 3 different surname-based groups i.e. a) North Indian (n = 157), b) South Indian (n = 100), and c) Zulu (n = 100) using SplitsTree V4.14 (Huson and Bryant, 2005). Each line represents a haplotype. North Indians comprise the surnames Khan (K), Maharaj (M), and

Singh (S). South Indians comprise the surnames Govender (G), Naidoo (N), and Pillay (P). Zulus include the surnames Buthelezi (B), Cele (C), Dlamini (D), Mkhize (Mk) and Zulu (Z).

The haplotype network for shared haplotypes (Figure 22), was based on only haplotypes with a frequency > 1 i.e. 141 taxa and 61 haplotypes. Overall, the haplotype network comprised majority of samples with the surname Singh, which shared haplotypes mainly with random Hindu and Tamil samples. Singhs also shared haplotypes with Khans, random Muslims and samples with the surname Zulu. The majority of samples named Maharaj shared haplotypes with random Whites. Zulus with the surname Cele tended to share haplotypes with each other, or with those with the surname Zulu. South Indians with the surname Govender tended to share haplotypes with each other, and with sample members named Singh and Pillay.

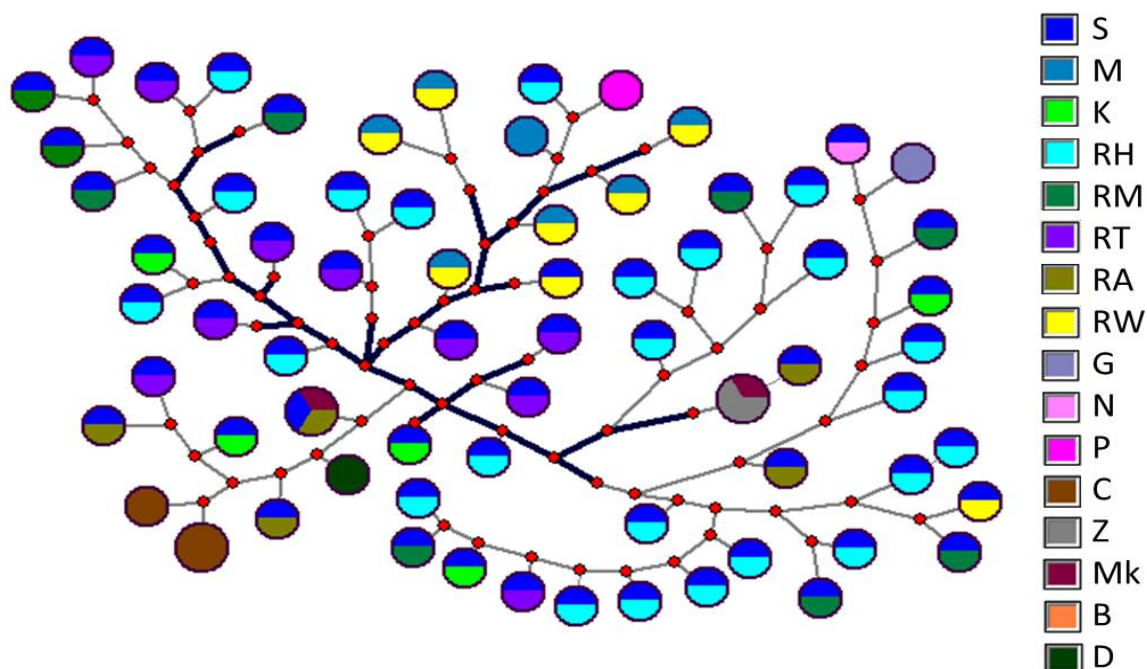


Figure 22: Haplotype network for shared haplotypes found in the total sample, which includes surname-based groups and random controls (n = 141).

The MJ Network was calculated based on star conduction and maximum parsimony to derive the simplest parsimony tree using Network v5 (Forster *et al.*, 2017). The size of each circle is proportional to the number of individuals sharing a particular haplotype. Only haplotypes with a frequency > 1 are represented above. K = Khan, M = Maharaj, S = Singh; G = Govender; N = Naidoo, P = Pillay; B = Buthelezi, C = Cele, D = Dlamini, Mk = Mkhize, Z = Zulu; Random, H = Hindi, M = Muslim, T = Tamil, A = African, and W = White.

Haplotype networks were also constructed for each individual surname. All surname groups contained multiple haplotypes, and multiple clades, consistent with multiple founders Figure A 4). Haplotypes were mostly individual-specific, consistent with a high HD, although some haplotypes were represented by more than one sample. Singh (Figure A 4a), Khan (Figure A 4c), Naidoo (Figure A 4e), Pillay (Figure A 4f), Zulu (Figure A 4h), Mkhize (Figure A 4i), and Buthelezi (Figure A 4j) showed no shared haplotypes between any individuals.

Maharaj (Figure A 4b) had one shared haplotype (M29 and M46). Govender (Figure A 4d) had one shared haplotype (G02 and G19). Cele (Figure A 4g) had two shared haplotypes (C09 and C17; and C08, C11, C12, C14, C16). Dlamini (Figure A 4k) had one shared haplotype (D05 and D18).

3.3. Surname-based population and forensic genetics analyses

Allele frequencies were reported for all 27 loci across the surname-based sub-groups (Table A 3.2**Error! Reference source not found.**) and ranged from 0.009-1.0. The overall highest allele frequency (1.0) was found in DYS437, allele 14, for samples with the surname Buthelezi from the Zulu sub-group. The highest allele frequency was DYS391 allele 10, for Khan (0.75), Govender (0.867), Naidoo (0.767), Pillay (0.833), Cele (1.0) and Dlamini (1.0). The highest allele frequency was DYS456 allele 15, for Dlamini (1.0). The highest allele frequency was DYS392 allele 11, for Cele (1.0), and Mkhize (0.950). The highest allele frequency (1.0) was found in DYS437 allele 14, for Zulu, and Buthelezi.

Allelic Patterns (Table 17) and were reported for the different surname-based sub-groups. The number of alleles (Na), frequency of alleles $\geq 5\%$ (NF) and number of effective alleles (Ne), private alleles (PA) and percentage of polymorphic loci (% P), ranged from 2.56 - 6.59, 2.56 - 4.48, 2.04 - 4.02, 0 - 0.7, 0.44 - 0.71, and 85 -100 %, respectively. The highest number of private alleles was found for the surname Khan (0.70), followed by Maharaj (0.52). No private alleles were found for Singh. Overall, the GD was the highest for Khan (0.71) and lowest for Cele

(0.44). Within each surname group GD was highest in (1) Khan (0.71) for North Indians, (2) G (0.70) for South Indians, and (3) Mk (0.57) for Zulu surnames.

Table 17: Allelic patterns for the surname-based groups (n = 347).

N = Number of individuals, Na = No. Alleles, Nf = Frequency of alleles $\geq 5\%$. Ne = No. Effective Alleles, PA = Private Alleles i.e. Number of alleles found only in a single sample population among the broader collection of 16 sub-groups, GD= genetic diversity. % P = Percentage of polymorphic loci. K = Khan, M = Maharaj, S = Singh, G = Govender; N = Naidoo, P = Pillay; B = Buthelezi, C = Cele, D = Dlamini, Mk = Mkhize, and Z = Zulu.

	1. North Indian			2. South Indian			3. Zulu				
	K	M	S	G	N	P	B	C	D	Mk	Z
N	51	48	58	30	30	30	20	20	20	20	20
Na	6.59	6.56	6.44	5.56	5.52	5.59	3.59	3.52	4.19	4.48	3.89
Nf	4.37	4.37	4.44	4.41	4.33	4.41	3.59	3.52	4.19	4.48	3.89
Ne	4.02	3.73	3.77	3.93	3.72	3.79	2.70	2.04	2.44	2.81	2.58
PA	0.70	0.52	0	0.22	0.07	0.22	0.04	0.07	0.22	0.26	0.19
GD	0.71	0.70	0.68	0.70	0.68	0.68	0.55	0.44	0.52	0.57	0.55
% P	100%	100%	100%	100%	100%	100%	96%	93%	93%	100%	96%

Forensic parameters (Table 18) were reported for the surname-based groups. HD for all surname groups ranged from 0.998-1.000. The surname groups consisted of all unique haplotypes, with the exception of the surnames Maharaj, Govender and Pillay, in which one haplotype occurred twice. Haplotype match probabilities ranged from 0.0500 to 0.0172. The Zulu surname groupings had the highest haplotype match probabilities (MP) (0.0500), and Singh (S) had the lowest MP (0.0172). For all surnames, the DC was 1.000, with the exception of the surname Maharaj (0.979). Based on a low MP (with a relatively high HD and DC), the likelihood of a matching haplotype in descending order was: Singh, Khan, Maharaj, Naidoo and Pillay, and The Zulu surnames (Cele, Zulu, Mkhize, Buthelezi, and Dlamini).

Table 18: Forensic genetic parameters for different surname groups (n = 347).

HD = Unbiased haplotype diversity by Population, MP = Haplotype match probability, and DC = discrimination capacity (formulas in method section). K = Khan, M = Maharaj, S = Singh; G = Govender; N = Naidoo, P = Pillay; B = Buthelezi, C = Cele, D = Dlamini, Mk = Mkhize, and Z = Zulu.

Surname		Observed haplotype (n)			HD	MP	DC
		Total	Once	Twice			
North Indian	K	51	51	0	1.000	0.0196	1.000
	M	47	46	1	0.999	0.0217	0.979
	S	58	58	0	1.000	0.0172	1.000
South Indian	G	30	28	1	0.998	0.0356	1.000
	N	30	30	0	1.000	0.0333	1.000
	P	30	28	1	0.998	0.0356	1.000
Zulu	B	20	20	0	1.000	0.0500	1.000
	C	20	20	0	1.000	0.0500	1.000
	D	20	20	0	1.000	0.0500	1.000
	Mk	20	20	0	1.000	0.0500	1.000
	Z	20	20	0	1.000	0.0500	1.000

4. Comparisons of study samples with samples found on the Y-chromosome STR Haplotype Reference Database (YHRD)

Matches between Y-STR haplotypes of study samples and samples from other world populations were searched for on the YHRD. No haplotype matches were found from 18,921 haplotypes in the database (YHRD, 2018). Validation using YHRD indicated that all alleles were valid as they fell within the allele range of each of the 27 loci investigated. At markers DYS635 and DYS390 the alleles 15 (found at M36) and 16 (found at M47) respectively, were not present on the YHRD. The experimental sample (consisting of 346 haplotypes) was compared to all related populations (Table 5) found in the YHRD for the Yfiler® Plus loci (consisting of 2045 other haplotypes in total) (Table 19). The purpose of comparing the experimental samples with this range of samples was to investigate the likely origin of the experimental samples.

Table 19: Number of haplotypes for the study samples and comparative populations (YHRD).

Values were obtained from YHRD (2018) using the Yfiler® plus search mode. K = Khan, M = Maharaj, S = Singh; G = Govender; N = Naidoo, P = Pillay; B = Buthelezi, C = Cele, D = Dlamini, Mk = Mkhize, Z = Zulu; R = Random, H = Hindi, M = Muslim, T = Tamil, A = African, and W = White. Australia [Asian] = Au[As], and Australia [European] = Au[Eu]. India = I, Assam - India [Kachari] = AI[Ka], Kerala - India [Keralite] = KI[Ke], and Andhra Pradesh - India [Thoti] = API[Th]. US[As] = United States [Asian American], US [A] = United States [African American], US[Eu] = United States [European American], Mi,US[As] = Minnesota, United States [Asian American], Mi,US [A] = Minnesota, United States [African American] and Mi,US[Eu] = Minnesota, United States [European American]. Ken = Kenya., and BLo = Bantu_Luhya, Other.

No. of Haplotypes	Combined Population																Total
	K	M	S	RH	^R M	RT	RA	RW	G	N	P	B	C	D	B	Z	346
	40	30	52	25	7	11	5	6	30	30	30	18	17	13	18	17	
	Au [As]	Au [Eu]	I	AI [Ka]	KI [Ke]	API [Th]	US [As]	US [A]	US [Eu]	Mi,U S[As]	Mi,U S [A]	Mi,U S[Eu]	Ken	Ken [BLo]	2045		
	19 6	19 7	19	8	3	8	24 0	47 9	46 5	96	77	67	12 8	62			

Cluster analyses were based on the experimental samples and those downloaded from the YHRD, using the online AMOVA clustering tool available on YHRD (2018), which uses the discrete Laplace method. This resulted in four clusters (Table 20). The clusters were formed based on their ethnicity i.e. clusters 1 and 3 consisted of only Indians/Asians, whereas clusters 2 and 4 consisted of only Africans. Significant difference ($p < 0.005$) was observed amongst all the non-clustered Indian and African sub-groups (Table 20).

Table 20: Rst Clustering of populations.

Clusters identified from YHRD (2018). K = Khan, S = Singh, R = Random, H = Hindi, M = Muslim, T = Tamil, and A = African. G = Govender, B = Buthelezi, C = Cele, D = Dlamini and Mk = Mkhize. . Australia [Asian] = Au[As]. Assam - India [Kachari] = AI[Ka], Kerala - India [Keralite] = KI[Ke], and Andhra Pradesh - India [Thoti] = API[Th]. Mi,US[As] = Minnesota, United States [Asian American], Ken = Kenya., and BLo = Bantu_Luhya, Other.

Cluster	N	Sub-group	Ethnic group
Cl1	8	K, S, RH, RM, RT, G, KI[Ke], API[Th]	All Indian
Cl2	4	B, C, RA, Mk	All African
Cl3	3	AI[Ka], Au[As], Mi,US[As]	All Indian/Asian
Cl4	3	D, Ken[BLo], Ken	All African (Kenya)

Rst values generated by AMOVA (Table 21), using the online AMOVA tool available on YHRD (2018), were mostly significant. The greatest difference was observed between Zulus and

Mi,US[Eu] ($R_{st} = 0.495$, $p < 0.005$). No Significant difference ($p > 0.05$) was found between: (1) the entire white study population i.e. Random White (RW) and the European populations (Au [Eu]; Mi,US [Eu] and US [Eu]); (2) Maharaj (M) and RW; (3) US Africans and Minnesota US Africans; (4) Zulus and Cluster 2, which consisted of all the study Zulus with the exception of Dlamini (Table 21).

Table 21: RST and significance values based on pairwise AMOVA for the study population and comparative populations from the YHRD.

The values were obtained from YHRD (2018) using the Yfiler® plus AMOVA and MDS mode. K = Khan, M = Maharaj, S = Singh, G = Govender; N = Naidoo, P = Pillay; B = Buthelezi, C = Cele, D = Dlamini, Mk = Mkhize, Z = Zulu; Random, H = Hindi, M = Muslim, T = Tamil, A = African, and W = White. US[As] = United States [Asian American], US [A] = United States [African American], US[Eu] = United States [European American], Mi,US[As] = Minnesota, United States [Asian American], Mi,US [A] = Minnesota, United States [African American] and Mi,US[Eu] = Minnesota, United States [European American]. Cl1 = K, S, RH, RM, RT, G, KI[Ke], API[Th]; Cl2 = B, C, RA, Mk; Cl3 = AI[Ka], Au[As], Mi,US[As]; and Cl4 = D, Ken[BLo], Ken (see Table 20 for clustering groups). R_{st} below diagonal and P (probability) above diagonal. **Bold** = significant p -values (i.e. < 0.05), grey = > 0.05 .

	M	RW	N	P	Z	Au [Eu]	Mi,US [A]	US [A]	US [As]	Mi,US [Eu]	US [Eu]	Cl 1	Cl 2	Cl 3	Cl 4
M	-	0.117	0.010	0.022	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
RW	0.106	-	0.001	0.001	0.000	0.137	0.000	0.000	0.000	0.098	0.107	0.001	0.000	0.000	0.000
N	0.084	0.276	-	0.807	0.002	0.000	0.001	0.000	0.000	0.000	0.000	0.002	0.000	0.000	0.000
P	0.079	0.288	-0.037	-	0.001	0.000	0.004	0.004	0.000	0.001	0.001	0.021	0.000	0.003	0.000
Z	0.231	0.642	0.223	0.246	-	0.000	0.003	0.006	0.000	0.000	0.000	0.032	0.179	0.000	0.005
Au [Eu]	0.098	0.041	0.222	0.175	0.469	-	0.000	0.000	0.000	0.956	0.231	0.000	0.000	0.000	0.000
Mi,US[A]	0.118	0.337	0.151	0.129	0.178	0.237	-	0.090	0.000	0.000	0.000	0.000	0.001	0.000	0.000
US [A]	0.175	0.320	0.159	0.125	0.162	0.229	0.007	-	0.000	0.000	0.000	0.000	0.002	0.000	0.000
US [As]	0.117	0.219	0.151	0.126	0.334	0.158	0.167	0.171	-	0.000	0.000	0.000	0.000	0.000	0.000
Mi,US[Eu]	0.077	0.063	0.194	0.150	0.458	-0.006	0.213	0.209	0.147	-	0.722	0.000	0.000	0.000	0.000
US [Eu]	0.131	0.052	0.232	0.176	0.495	0.001	0.266	0.250	0.179	-0.003	-	0.000	0.000	0.000	0.000
Cl 1	0.071	0.278	0.095	0.083	0.114	0.203	0.041	0.073	0.122	0.173	0.241	-	0.000	0.000	0.000
Cl 2	0.252	0.408	0.210	0.189	0.064	0.452	0.126	0.091	0.298	0.411	0.485	0.145	-	0.000	0.001
Cl 3	0.121	0.234	0.105	0.089	0.238	0.180	0.138	0.145	0.022	0.164	0.203	0.100	0.253	-	0.000
Cl 4	0.290	0.502	0.302	0.304	0.132	0.409	0.081	0.073	0.293	0.394	0.436	0.137	0.099	0.234	-

MDS dimension stress values ranged from 0-1; the lower the value the more reliable the spread of the MDS plot dimensions. The low MDS dimension stress value obtained (0.029) (Figure 23), indicates that the groups obtained from R_{st} clustering were suitable for dimension demonstration. MDS (Figure 23) revealed that overall each sample subgroup was associated with sub-populations within their ethnic groups. Indian samples tended to be located in the mid region of the plot (M, N, P, US [As], Cl 1 and Cl 3), African samples tended to be in the positive dimensions (upper right) and whites in the negative dimension region (lower left). The sub-

group Maharaj was located close to the White ethnic group. Pillay and Naidoo (Tamil surnames) group closest to each other (lower right of the Indian/Asian grouping). Asians group closer together (upper left). Cl 1, consisting of Indian subpopulations, is located at a similar distance to both Maharaj (M) and the Pillay (P) /Naidoo (N) cluster, however M and P&N are a little closer to each other than to Cl 1(Figure 23).

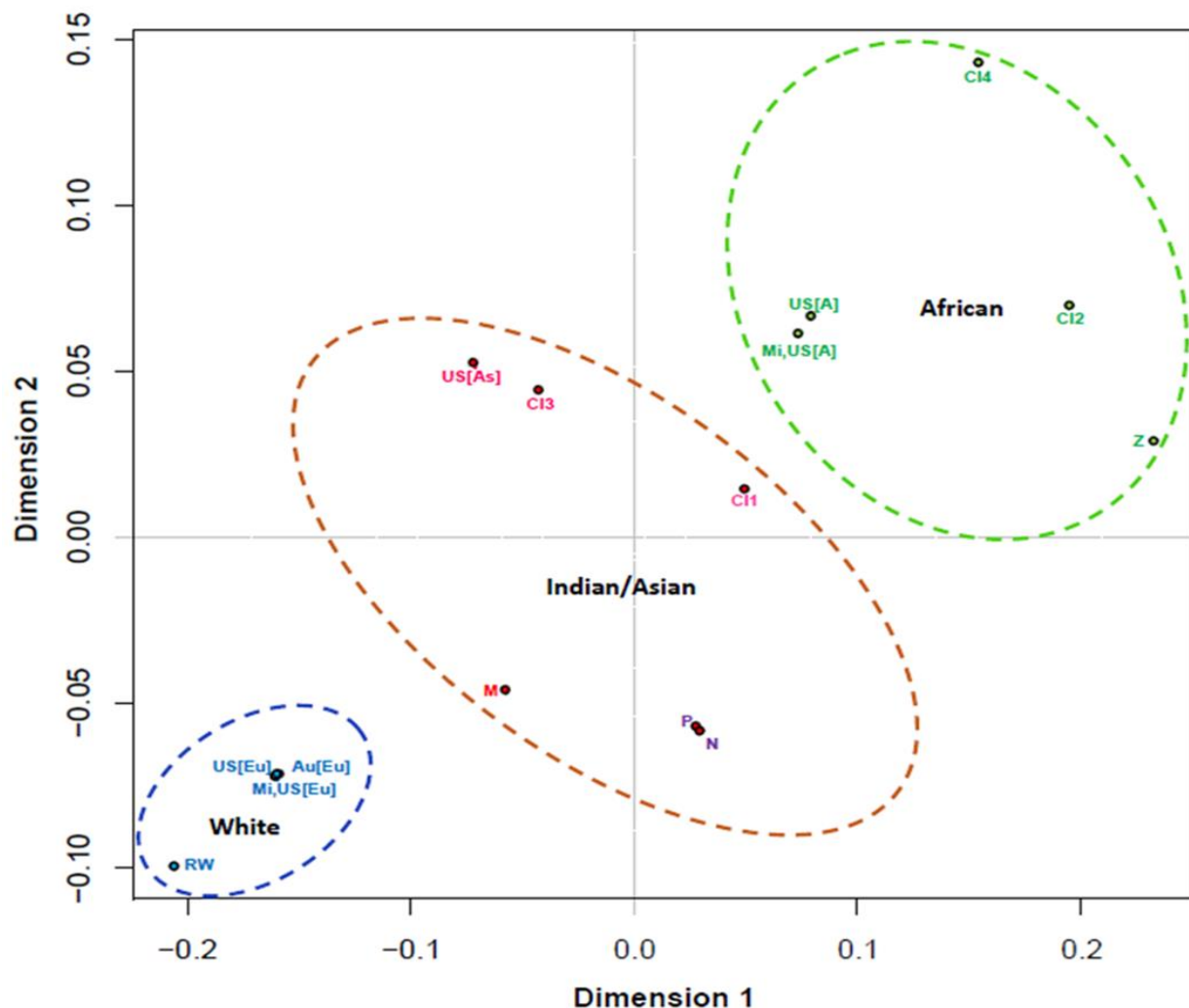


Figure 23: MDS for 27 Y-STR loci for the studied and comparative populations from the YHRD.

The values were obtained from YHRD (2018) using the Yfiler® plus AMOVA and MDS mode. Dimension stress = 0.02932. K = Khan, M = Maharaj, S = Singh, G = Govender; N = Naidoo, P = Pillay; B = Buthelezi, C = Cele, D = Dlamini, Mk = Mkhize, Z = Zulu; Random, H = Hindi, M = Muslim, T = Tamil, A = African, and W = White. US[As] = United States [Asian American], US [A] = United States [African American], US[Eu] = United States [European American], Mi,US[As] = Minnesota, United States [Asian American], Mi,US [A] = Minnesota, United States [African American] and Mi,US[Eu] = Minnesota, United States [European American]. Cl1 = K, S, RH, RM, RT, G, KI[Ke], API[Th]; Cl2 = B, C, RA, Mk; Cl3 = Al[Ka], Au[As], Mi,US[As]; and Cl4 = D, Ken[BLo], Ken (see Table 20 for clustering groups).

DISCUSSION

The growing field of genetic genealogy provides useful information for population and forensic genetics. The overall aim of this study was to explore the genetic genealogy, population and forensic genetics of (1) samples comprising male Indians with a variety of surnames, geographic regions of origins and religions and (2) male Zulus with different common surnames, all currently residing in the greater Durban area of KZN, SA.

The target Indian surnames are associated with different regions of origin in the Indian subcontinent (North vs South Indians), religions (Hindu and Islamic Muslim), and consequently languages (Hindi, Tamil and Urdu). It is possible to formulate hypotheses about potential population subdivision based on differences in geographic regions of origin, surname, and perceived barriers to marriage between people of different religions (e.g. Hindu and Islamic Muslims). The African samples comprise people from the Zulu ethnic group bearing one of 5 common surnames, lending themselves to an analysis of the co-inheritance of surnames and Y-STRs.

Null Alleles were observed in 77.8 % of the examined loci (21 out of 27 loci), primarily found in the loci DYS391, DYS389II and DYS448 (Table 6). Null alleles, specifically at locus DY448 (Table 6), have been found in many studies based on Y-STRs (Purps *et al.*, 2014; Olofsson *et al.*, 2015; Ottaviani *et al.*, 2015; Rapone *et al.*, 2016; Zgonjanin *et al.*, 2017). Ottaviani *et al.* (2015), found 2 null alleles at DYS448 from 250 samples of unrelated males from North Italy, when using the Yfiler® kit and PowerPlex Y23 system. Purps *et al.* (2014), in a global analysis of RMu Y-STRs, observed a total of 133 null alleles at 17 loci (frequency = 0.03%); a large deletion was also found in the AMEL region of the Y-chromosome of Asian samples. This type of deletion was also observed in other studies (Chang *et al.*, 2007; Parkin *et al.*, 2007; Takayama *et al.*, 2009). The presence of null alleles or deletions could be due to the rearrangement of the Azoospermia Factor c (AZFc) region (a highly polymorphic region where massive recurrent deletions occur in men with very low sperm count) (Turrina *et al.*, 2015). According to Budowle *et al.* (2008), the

presence of null alleles does not affect the interpretation of Y-STR results, as current methods for Y-STR profile interpretation can accommodate their presence.

1. Comparison of Y-STR marker sets via population and forensic genetics analyses

Basic population and forensic genetic parameters were examined using different marker sets (MHT, Yfiler®, RMu and Yfiler® Plus) in order to assess their relative utility and suitability for use with the present dataset.

H₁: The Yfiler® Plus Kit, chosen for use in this study, is an appropriate choice for genetic genealogy, population and forensic genetics studies of Indian and Zulu males from the greater Durban area of KZN. Genetic/haplotype diversity and discrimination capacity will be the highest for the Yfiler® Plus marker set and haplotype match probability lowest, as it includes a higher number of loci than any of the other marker sets to which it was compared (MHT, Yfiler, and RMu), and includes all of the markers in these marker sets

The Yfiler® Plus kit consists of markers with a relatively high GD (Figure 11) with a mean±standard deviation of 0.712±0.143, which makes these markers useful in determination of genetic differences. Based on the samples in this study (Figure 11), marker DYS385b had the highest GD (0.884), a result consistent with many studies making similar comparisons, which found the loci DYS385a/b to have the highest GD owing to their high levels of polymorphism (Ehrenreich, 2005; Parkin *et al.*, 2007; Frank *et al.*, 2008; Balamurugan *et al.*, 2010; Ge *et al.*, 2010; Yadav *et al.*, 2011; Rapone *et al.*, 2016; Al-Azem *et al.*, 2017; Shrivastava *et al.*, 2017). DY449 was found to have the second highest GD (0.875), again consistent with other studies which found the GD of DY449 to be relatively high (> 0.8) (Hanson and Ballantyne, 2007; Rodig *et al.*, 2008; D'Amato *et al.*, 2011).

The locus DYS391, showed the lowest GD (0.369) (Figure 11). Leat *et al.* (2004), who examined the Xhosa population in Cape Town, SA (GD=0.320), Ehrenreich (2005), who examined the SA

population with Asian Indians as a sub-grouping ($GD = 0.279$), and Purps *et al.* (2014), who examined global populations, also found this locus to show the lowest GD. The occurrence of the highest GD at DYS385a/b and lowest at DY391 was also observed in other studies on Indians (Ehrenreich, 2005; Frank *et al.*, 2008; Balamurugan *et al.*, 2010; Yadav *et al.*, 2011; Shrivastava *et al.*, 2017).

The 7 RMu markers yielded the highest number of alleles with a frequency of greater than 5%, the highest number of expected alleles, the highest number of private alleles, as found by Ottaviani *et al.* (2015), and the highest diversity, as expected (Table 7). However, the addition of more Y-STR markers as exemplified by the Yfiler® plus kit, based on 27 markers (Table 8), resulted in a higher number of unique haplotypes, a higher DC and a lower MP. A high DC combined with a low MP makes data more reliable for forensic investigations (Ge *et al.*, 2010; D'Amato *et al.*, 2011; Ottaviani *et al.*, 2015). Thus, the Yfiler® Plus kit, based on the addition of 7 RMu markers to the Yfiler® set of markers (Rapone *et al.*, 2016; Zgonjanin *et al.*, 2017) is likely to be most useful for studies of genetic genealogy, population and forensic genetics. Loci found in the RMu marker set (Figure 11) gave a relatively high GD (> 0.75 , $\text{mean} \pm \text{standard deviation}$ $GD = 0.821 \pm 0.034$). RMu loci have been found to increase DC between non-related individuals, between males (Ballantyne *et al.*, 2010; Ottaviani *et al.*, 2015), and between non-related individuals from the same patrilineage, which could potentially allow for discrimination between relatives of the same patrilineage (Ballantyne *et al.*, 2012).

As the number of marker loci increase, so do the number of unique haplotypes (Table 8). This in turn increases the DC. This pattern was also observed in other studies: Based on a global analysis using the PowerPlex Y23 system, Purps *et al.* (2014), observed high levels of unique haplotypes; Ge *et al.* (2010), who studied the US population, found that unique markers increased in number when using the 16 Yfiler® markers compared with 10 Y-plex markers (56.7 %); Ottaviani *et al.* (2015), who studied unrelated males from North Italy, found that the DC increased when using the Yfiler® Plus kit compared with the MHT.

As in this study, (Ottaviani *et al.*, 2015) found MP to be lower when based on the Yfiler® Plus kit than the MHT marker set. García *et al.* (2016), who studied 6 surnames from the Basque country, showed the suitability of the Yfiler® Plus kit to improve the male lineage DC (0.999996) when compared with the MHT marker set, as was found in this study. Khubrani *et al.* (2018), who studied paternal inheritance in males from different regions of Saudi Arabia, found that the Yfiler® Plus kit (based on 27 loci) provided a higher DC (95.3%) than the Yfiler® kit (17 loci).

2. Genetic structure analyses based on population sub-groupings

It is to be noted that all conclusions drawn for each sub-grouping were not based on the overall population within that sub-group, but rather to the target surnames in this study that fall within that specific sub-group.

2.1. Genetic structure based on ethnicity (Indian vs Zulu)

H_{2a}: The sample will be structured into groups based on ethnicity, as the Indian and Zulu population groups have very different geographic origins and are likely to have evolved separately prior to the introduction of Indians to the Durban region. Although interbreeding many have occurred since the Indian group arrived in the Durban area, causing the introduction of Zulu Y-chromosomes into the Indian groups and vice versa, Apartheid and various cultural practices would have served as a deterrent.

A relatively moderate level of genetic separation between Indian and Zulu groups was observed in all genetic structure-based analyses. In comparisons using AMOVA, 7% of the variance ($\Phi_{IPT} = 0.074$, $p = 0.0001$) (Table 10.1) occurred between the two ethnic groups. This was reflected in PCoA analyses (Figure 12.1) as a high degree of genetic separation between Indians and Zulus, which also formed separate groups in Bayesian analyses of population structure (Figure 13). Haplotype network analyses (Figure 14), showed a strong Zulu-only cluster, similar to the findings of Lane *et al.* (2002), who studied the South African Bantu speaking groups based on 9 autosomal and 4 Y-STR loci, and found clustering within the Zulu group; other haplotype

clusters contained both Indian and Zulu sample members, consistent with the analyses mentioned above. Ge *et al.* (2010), in a study of the US African and Asian populations, also found a strong genetic difference between Indians and Africans.

This is consistent with the 'out of Africa' theory (Cavalli-Sforza *et al.*, 1994), in which Indians are proposed to have migrated out of Africa to Eurasia (Sahoo *et al.*, 2006) around 60 000 years ago (Kivisild *et al.*, 2003).

From that time until the Indians returned as indentured labourers to SA, the two groups would have evolved separately, accumulating genetic differences. This would have led to the degree of genetic structuring observed in this study, which occurred despite the likelihood of a level of interbreeding among these groups since the Indians travelled from India to the Durban area. Such interbreeding is reflected in the Bayesian analysis of population structure (Figure 13.2), where a level of admixture is observed between the Zulu and Indian groups, as the Indian group with the surname Singh shows some membership of the Zulu group. Network analysis (Figure 14) revealed a primarily Zulu cluster, some primarily Indian clusters, and some clusters which contained both Indian and Zulu membership (Figure 14). Such mixed Indian/Zulu clades are consistent with a level of admixture between the two groups, as was found in the Bayesian analysis of population structure.

Consistent with these results, other genetic studies have found that Indian caste populations originating in Europe or Asia (Mountain *et al.*, 1995; Bamshad *et al.*, 1996; Quintana-Murci *et al.*, 1999; Bamshad *et al.*, 2001) showed slight admixture with African populations (Bamshad *et al.*, 1996).

MtDNA sequence based analyses suggested that global Indian populations shared a common late Pleistocene maternal ancestry in India (Misra, 2001; Kivisild *et al.*, 2003), exhibiting the macro haplogroups M and N, whereas Y-chromosome markers suggest paternal ancestry from Central or West Eurasia (Sahoo *et al.*, 2006). These authors, using Y-SNPs, found paternal inheritance was of South Asian origin for the Indian caste communities, closer to the north and

west regions of India. Prior settlement of South Asia was most likely over the Southern route from Africa as the haplogroup M is found in this population and is absent in East and Southwest Asia (Quintana-Murci *et al.*, 1999; Kivisild *et al.*, 2003; Sahoo *et al.*, 2006). Thus, it appears that Indian populations have diverse ancestry, stemming from India generally, its north and west regions, and central and or west Eurasia, and that different populations followed different colonization routes.

In this study the Indian vs Zulu comparison showed a moderately high degree of genetic structuring ($\Phi_{IPT} = 0.074$, $p = 0.0001$) (Table 10.1) , with Indians having a higher GD than Zulus (Table 12). This is possibly due to the diverse origins of the Indian populations, from west Eurasia to India. The Zulu population, however, is a relatively tightly knit Nguni grouping, and is therefore likely to show lower GD.

Genetic separation between Indians and the Zulu populations is likely to stem primarily from their diversification due to geographical separation caused by the out of Africa migrations (around ~ 60 000 years ago) of groups which ended up as modern-day Indians, a subset of which then arrived in Durban from India ~ 160 years ago. Present day Zulus and Durban area Indians came to inhabit the same geographical space in the years following the first introduction of indentured Indian labourers to the Durban area; as a consequence of this there is likely to have been some interbreeding, which would have reduced the genetic separation which had developed between the two groups. However, there are factors which may have prevented a greater degree of homogenization of the two groups. These include (1) Cultural tendencies which lead groups to marry/interbreed with other members of the same group, (2) Endogamy among Indians (marriage within a specific caste) (Bayly, 2001) and, (3) Banning of inter-racial marriages by Apartheid laws. A survey by Posel (2001) found that most Africans found other ethnic groups to be 'untrustworthy' (56 %), uncomfortable to be around (46.8 %) and 'hard to imagine ever being friends' (52.7 %). This makes Africans less likely to interbreed with Indians, maintaining the genetic separation of the two groups.

Despite Apartheid laws inter-racial interbreeding did occur, resulting in mixed race offspring, Thus, explaining genetic admixture between different race groups (Slabbert and Heathfield, 2018), such as was found between Indian and Zulu groups in this study. Motladiile (2004), who studied Y-STRs of the SA coloured population using a meta-database, found that high diversity levels in coloured (mixed race) populations were a result of genetic admixture from African, Asian and European ancestries. Some alleles/loci found in the SA Coloureds (Y-STR loci DYS391 (allele 10) and DYS392 (allele 11)) were from African and Asian ancestral contributions, respectively (Motladiile, 2004).

The post-Apartheid era, in which intermarriage barriers no longer exist, may have resulted in further interbreeding between the Indian and Zulu groups, reducing the level of genetic structuring among them. In order to better understand the dynamics of GD and structuring, studies based on autosomal STRs and sequencing of the hypervariable regions of the mtDNA should be carried out.

2.2. Genetic structure based on region of origin in India (North vs South India)

This study aimed to examine whether there was genetic structure amongst sample members whose surnames indicate origins in North vs South India.

H_{2b} : The distance separating the sites of origin of the North and South Indian samples, which originated in Calcutta and Madras, combined with language and cultural differences, and the practice of endogamy will have led to genetic diversification of North and South Indians, which will be reflected in Y-STR genetic structure among these groups.

Genetic differences between North and South Indians were observed in all genetic structure-based analyses. AMOVA indicated that 3% of the variance occurred among North and South Indian groups, and PhiPT (0.029) indicated a low but significant ($p < 0.005$) level of genetic structuring (Table 10.2). Consistent with this, the PCoA plot (Figure 12.2) showed a considerable degree of overlap between the North and South Indian samples, although there was a relatively

large group of North Indians which appeared genetically separate from the overlapping samples; these are likely to be responsible for the 3% of the variance that occurs among these groups in the AMOVA (Table 10.2). Bayesian Analysis of population structure (Figure 13.1) also found the North and South Indians to be separate genetic groups. The distinction between these groups was discernible in the haplotype network analysis (Figure 14), as there was a North Indian rich cluster, although mostly clusters consisted of a mixture of North and South Indian samples. In summary, there appears to be a discernible, but relatively weak level of genetic structure among the North and South Indian groups. These groups, which come from geographically separated areas, are likely to have diverged not only due to lack of physical opportunity to meet and interbreed, but also due to barriers created by differences in language, culture and religion in the North and South of India.

Most of the Indians currently in SA, particularly KZN, are from the 4th and 5th generation of Indians (Figure 16), depending on age, which originally came down from India. On arrival in the Durban area of SA, the groups of Indians from the North and South lived in the same area, and therefore had the opportunity to interbreed. The close proximity of these two groups was further maintained during the Apartheid years, as the Group Areas act confined Indians to certain areas where they were allowed to live. Interbreeding between the two groups would have lowered the level of genetic structuring among them, although relatively weak structuring is still detectable, as discussed above. Sahoo *et al.* (2006), using Y-SNPs, found paternal inheritance was closer in the north and west regions of India. Ghosh *et al.* (2011), who studied the Indian populations from India using the Y-filer kit, also found that haplotype clustering was high within region-based populations.

Shrivastava *et al.* (2017), who used Y-STRs to study the central Indian (Madhya Pradesh) population, found a high amount of genetic variation among different tribes and castes, as did Ghosh *et al.* (2011), who studied different region and language-based Indian populations in India, and Yadav *et al.* (2011), who studied the North India population. Khubrani *et al.* (2018), who studied paternal lineages of males from different regions in Saudi Arabia, also found that region of origin is related to GD and structure. People from the different regions were highly

genetically differentiated, with low diversity in the north and centre and high diversity in the west and east. Saudi Arabian and Indian populations commonly engage in endogamous marriages, fostering high levels of inbreeding (Scott *et al.*, 2016), consistent with regional population differentiation. In contrast, Rapone *et al.* (2016), using the Yfiler® Plus kit, found low, non-significant genetic distances ($RST < 0.011$) among northern central and southern Italians. The difference could be based on lower levels of ethnically based barriers to intermarriage in the Italian population, as these are not practiced as strongly as in the past (Coffey, 2004). Sahoo *et al.* (2006), using Y-SNPs, found paternal inheritance for the Indian caste communities was better explained at region-based level than at religion, language and caste-based levels.

2.3. Genetic structure based on religious groups

This study aimed to examine the genetic structure amongst Hindu and Muslim Indians.

H_{2c}: In the period since the separation of the Muslim religion from the more ancient Hindu religion, Y-chromosome genetic diversification and therefore structuring will have occurred between Hindus and Muslims, maintained by perceived religious barriers to intermarriage between Muslims and Hindus.

The scriptures of the world's 5 popular religions (Hinduism, Buddhism, Islam, Christianity and Judaism) all have a similar belief system which limits sexual behaviour to marrying within same culture, and does not permit adultery, which results in higher paternity certainty (Strassmann *et al.*, 2012). In Hinduism, cuckoldry/adultery is not allowed according to the 'The Laws of Manu' (Doniger, 1991). In Islam, confusion of paternity is prevented by the Quran, which states that women are only allowed to re-marry 3 menstrual periods after divorce (Strassmann, 1992). Based on the above religious practices, one would therefore expect to observe a level of genetic separation (structuring) among Hindu and Muslim members of the sample group. A similar principle is followed by Indians who converted to Christianity, as the Bible prohibits adultery and the reproduction of extramarital children (Coogan *et al.*, 2010).

Genetic structure was observed between Hindu and Muslim sample members, although at a very low level. AMOVA revealed that a significant but low 1% of the variance ($\Phi_{IPT} = 0.009$, $p = 0.0010$) (Table 10.3) occurred among sample groups. PCoA (Figure 12.3) showed a very high degree of overlap between these two religious' groups, although there was a small group of Hindus (comprising south Indian Tamils), which were found outside the area of overlap. In contrast, neither Bayesian Analysis of population structure (Figure 13) nor haplotype network analysis (Figure 14) revealed a clear distinction between Hindu and Muslim samples.

Hinduism is the world's oldest living religion, originating around the 15th – 5th century BCE (Leser, 2018), and at some point, Muslims are likely have converted from Hinduism to Islam (Terreros *et al.*, 2007). According to Islamic religion, which originated in the 7th century (Leser, 2018), Muslims are allowed to marry within the same family line, causing high inbreeding levels (El-Hazmi *et al.*, 1995), and the development of genotypes/haplotypes which are somewhat Muslim-specific, even though they might have had their origins in common Hindu genetics. This is likely to have led to a level of divergence of Muslim Indians from the Hindu grouping.

2.4. Genetic structure based on language

This study aimed to examine the genetic structure amongst Hindi, Muslim (Urdu) and Tamil speaking Indians.

H_{2d}: Hindi, Tamil and Urdu speakers (Muslims) will have evolved as separate groups due to language barriers; this diversification will be reflected as Y-chromosome based genetic structure.

It should be noted that although South African Muslims are descended from a population of Muslims in India who would have been Urdu speakers, probably a minority of those who currently reside in the greater Durban area currently speak Urdu.

Haplotype network analysis (Figure 14) revealed clusters which contained primarily Hindi and primarily Tamil speaking sample members. Consistent with this, PCoA (Figure 12.3) revealed an area of overlap between the two groups, but also areas containing Hindi-only and Tamil-only samples. AMOVA revealed that a relatively low 3% but significant part of the variance ($p < 0.005$) occurred between Tamil and both Hindi (Table 10.4) and Muslim (Table 10.5b) sample members, and an even lower (1%) but still significant (0.001%) (Table 10.5a) part of the variance occurred between Hindi and Muslim sample members.

Overall it would appear that there is a level of genetic structure between all three language-based groups. There is more genetic separation between the South Indian Tamils and the North Indian Hindus and Muslims than there is between the two North Indian groups (Hindi/Muslim). Previously, it was concluded that there was genetic structuring between sample members originating in North vs South India. Thus, it appears that the region-based comparison (North vs South Indian) has an impact on the language-based comparison. The lowest, but still significant, degree of language-based genetic structuring occurs between the North Indian Hindi and Muslims, for which there is no region-based confounding factor. This contrasts with the outcomes of other studies which have shown that lineage clustering was not based on language (Passarino *et al.*, 1996; Bamshad *et al.*, 2001; Kivisild *et al.*, 2003). According to Martinez-Cadenas *et al.*, 2016, region-based clusters (such as observed in the study between North Indian language-based groups) occur if clans within a region don't mix, which sometimes results in the development of a unique language (Martinez-Cadenas *et al.*, 2016). The Hindi and Urdu languages are very similar as most words have the same meaning in both languages, which is not the case with the Tamil language, which is completely different to Hindi and Urdu. The lower levels of genetic structure among the North Indian Hindi and Muslim (originally Urdu) groups may also reflect the relative similarity of the Hindi and Urdu languages and therefore the ease of communication and interaction between these groups, who were also geographically proximate. In contrast, the greater structure amongst the Tamil group and the Hindi and Muslim groups may reflect the greater distinctiveness of the Tamil language from the Hindi and Urdu languages, whose speakers were separated by both geographic and language (communication) barriers.

3. Genetic genealogy: Surname-based genetic analyses

In this study the genetics of surname-based groups of North Indians (Khan, Maharaj and Singh), South Indians (Govender, Naidoo and Pillay) and Zulus (Buthelezi, Cele, Dlamini, Mkhize and Zulu) was studied.

Social data, related to inheritance of the North Indian surnames 'Khan' , 'Maharaj' and 'Singh' , was also analysed to provide baseline information relative to randomly chosen people with different surnames drawn from the most common ethnic groups found in SA.

3.1. Surname-based genetic structure

This was investigated for three groups of surnames from different regions/ethnicities, namely (1) North Indian surnames, (2) South Indian surnames and (3) Zulu surnames.

H₃: The Y-chromosome and surnames are paternally inherited in both North and South Indians and Zulus. As the Y-chromosome has a relatively low mutation rate per generation, it could be hypothesised that groups of people with a particular surname would be more closely related to each other than to groups with other surnames. Genetic divergence over time among surname groups based on religious and or cultural practices or region of origin are likely to be reflected in genetic divergence among surnames.

AMOVA revealed that eight percent of the variance occurred amongst all 11 surname-based groups (Khan, Maharaj, Singh, Govender; Naidoo, Pillay; Buthelezi, Cele, Dlamini, Mkhize, Zulu) which were significantly different from each other. This shows that there is a broad level of genetic structure attributable to surname-based groupings. This is investigated in greater detail in the sections below, which focus on three of the sub-groupings, namely North Indian, South Indian and Zulu surname-based groups.

9.5.1. Genetic structure among North Indian surname-based groups.

AMOVA revealed that three percent of the variance occurred North Indian among surname groups (Figure 18.1). Although Φ_{IPT} values were all less than 0.05, indicative of relatively low levels of genetic structure, pairwise comparisons among all three surnames were significant. The highest Φ_{IPT} values and greatest significance was associated with comparisons of the Maharaj group with the Singh ($p = 0.000$) and Khan ($p = 0.000$) groups respectively (Table 15.1). PCoA revealed a region of overlap of all three surnames, and a separate grouping of Singhs and Khans only (Figure 19.1a). Bayesian Analysis of Population Structure (Figure 20) and the haplotype network (Figure 21a, Figure 22) supported this pattern. Thus, it appears that the Maharaj group is clearly distinct from some of the Singhs and Khans, whereas the Singhs and Khans are less distinct from each other. The reason for this may be that the inter-caste marriage barrier is strictest for the Maharajs, as they are from a Brahmin (high) caste, causing people with this surname to have evolved separately and diverged somewhat from other groups, and thus resulting in inbreeding. This divergence may then have been maintained by high fidelity co-inheritance of surnames and Y STRs, without confounding factors such as maternal transmission of surnames due to adoption or choice. Bamshad *et al.* (1996) also found the Brahmin caste to be more distinct than other South Indian castes in India. The distinction between the Maharaj and Khan groups may also reflect their differing religious affiliations (Muslim vs Hindu respectively), although it has been shown previously in this study that genetic structuring between Muslims and Hindus is very weak, although significant.

9.5.2. Genetic structure among South Indian surname-based groups.

AMOVA revealed that only one percent of the variance occurred among South Indian surname groups (Figure 18.2). Consistent with this, none of the pairwise surname group comparisons (Pillay vs Govender, Pillay vs Naidoo, and Govender vs Naidoo) were significantly different (Table 15.2); the PCoA showed a complete spatial overlap of all three surname-based groups (Figure 19.1b); and Bayesian Analysis of Population Structure (Figure 20) and the haplotype

network (Figure 21b) showed no clearly discernible difference in group membership among surname-based groups.

Consistent with these results, Ramana *et al.* (2001), observed no significant structure in the distribution of Y-SNPs in the South Indian population of India, consisting of different castes and tribes. Watkins *et al.* (2008), in a comprehensive study of genetic variation in South Indians, found no significant difference between various Tamil South Indian castes, which only differ by 0.96 % in STR variance. Further, in an AMOVA, Balamurugan *et al.* (2010) found that 99% of the variance occurred within and only 1% among 5 Tamil groups from India. The virtual absence of genetic structure observed amongst the surnames Govender, Naidoo and Pillay in this study supports the suggestion that there is a common Y gene pool for males of South Indian origin (Balamurugan *et al.*, 2010).

9.5.3. Genetic structure among Zulu surname-based groups .

AMOVA revealed that 9% percent of the variance occurred among the Zulu surname groups Buthelezi, Cele, Dlamini, Mkhize and Zulu (Figure 18.3). All pairs of surname-based groups were significantly different from each other with significance levels ranging from $p < 0.005$ to 0.043 (Table 15.3). PCoA revealed a region of overlap of all 5 surname groups, and some separation of groups with the surnames Cele, Dlamini and Mkhize (Figure 19.1c). Bayesian Analysis of Population Structure (Figure 20) supported this and showed the surnames Cele and Dlamini to be particularly distinct. Haplotype analysis (Figure 21c, Figure 22) shows one cluster consisting of only sample members with the surname Cele, although this did not include all sample members with the surname Cele; there was a further cluster consisting of sample members with the surnames Cele and Zulu.

Traditionally, Africans are not allowed to marry within the same patrilineage as this is regarded as incest; doing so results in punishment by the elders of their community (Strassmann *et al.*, 2012). Creation of such intramarriage barriers, is likely to lead to high levels of diversity and low levels of genetic structure among surname-based groups. In contrast to this expectation, levels

of genetic structure among Zulu surname-based groups in this study were considerably higher than those found among North Indian or South Indian surname-based groups.

3.2. Surname Inheritance

One of the aims of this study was to determine the mode of inheritance of surnames, viz. whether inheritance is monophyletic or polyphyletic, whether the surname haplotypes overlap or are non-overlapping, and whether paternal transmission of surnames occurs with high or low fidelity (Jobling, 2001) (Figure 5).

H₄: As both surnames and the Y-chromosome are purported to be patrilineally inherited, surname inheritance will be monophyletic and high fidelity (Jobling, 2001), i.e. each surname group would form an independent Y-chromosome haplotype cluster and these clusters would be distinct from each other and not contain haplotypes found in other surnames. This would allow surnames to be predicted from Y-STR haplotypes, which would be of use in forensic genetics. Further, surnames are transmitted from father to son with high fidelity.

The research hypothesis, and also the simplest explanation, is that co-inheritance of surnames and Y-STRs, in the absence of mutation, would lead to surnames being monophyletic and transmitted with high fidelity, if no disturbances in surname transmission from father to son has occurred (Jobling, 2001) (Figure 5a). This relationship was observed by Sykes and Irven (2000), who found that almost half of their samples with the surname 'Sykes' shared the same Y-chromosome haplotype (were monophyletic), despite the surname originating in many different regions (i.e. should ideally be polyphyletic).

However, this relationship between Y-STRs and surname inheritance was not observed in this study, where surname transmission was polyphyletic for all three sets of surname groups (North Indian, South Indian and Zulu), and surname groups showed overlapping haplotypes and clades (Figure 21, Figure 22, Figure A 4). This implies that multiple genetically different ancestors are likely to have founded different lineages of each specific surname investigated

(Kayser, 2017). High levels of polyphyletic transmission were also observed by King *et al.* (2006), in their study of a British surname, and reported on by King and Jobling (2009), in a review of genetic genealogy. However, low fidelity surname transmission could also have resulted in surnames being part of multiple clades in a haplotype network, and single clades containing multiple surnames, as was observed in this study. This could happen in the case of adopted children (Sykes and Irven, 2000; Solé-Morata *et al.*, 2015), where the child is given the name of the adoptive father, but does not carry his Y-chromosome. Another explanation is that surnames might be maternally transmitted, as might occur in the case of single mothers who give the child their surname although the child will bear the Y-chromosome of his father. The social data collected as part of this study revealed that, for the North Indian surname group, one third (Khan and Singh) to two thirds (Maharaj) of the respondents indicated that their surnames were derived from their maternal forefathers, consistent with disturbances in surname transmission. The average rate of non-paternal surname transmission was estimated to be 1.3% per generation during the past 700 years by Solé-Morata *et al.* (2015), who studied Catalan surnames from Spain using Y-SNPS.

There are a number of other explanations for low fidelity surname transmission which may relate to circumstances surrounding the importation of indentured labourers from India to what was then known as Natal. Anecdotal evidence suggests that surnames may have been incorrectly recorded during migration from India to SA, and that officials processing the new arrivals in Natal could not spell or pronounce some of the Indian surnames, leading them to use shortened or misspelled versions of them, or even used first names as surnames. Solé-Morata *et al.* (2015), who studied Catalan surnames from Spain using Y-SNPS, found that the introgressive hybridization rate into a surname caused by the above factors was estimated to be 1.5 – 2.6% per generation.

Although people sharing a similar surname may be more closely related than random unrelated individuals, the correlation may not be strong enough to hold much weight in use of surnames to determine genetic relatedness (King and Jobling, 2009; Andersen and Balding, 2017). It

appears that the relationship between surname and Y-STR inheritance is not strongly observed in this study (Figure 21).

4. Population and forensic genetics

Allele frequencies per loci ranged between ranging from 0.001-0.952 for the overall sample and sub-groups (Table 11). The overall sample HD (Table 13) was relatively high (0.999); it was of a similar level to that found by Tsiana (2015), who found the HD between South Zulu, Coloured, Afrikaner and Indian groups to be 0.998 in analyses based on the University of the Western Cape 10 Y-STR locus set. In this study, the DC of the overall study sample comprising Indian and Zulu sample members was moderately high (0.844) (Table 13). In contrast, D'Amato *et al.* (2011), found a slightly higher DC (0.9145), in a study based on the Yfiler® kit, between South African Xhosa, European and Indian groups.

Comparison of forensic genetic parameters between Indians and Zulus: In this study, the HD and DC of the Zulu grouping were higher those of the Indian group (Table 13). This contrasted with the results of other studies; for example D'Amato *et al.* (2011) found that SA Indians had a higher DC than Xhosa Africans, and Purps *et al.* (2014) found that Asians had a higher DC than Africans. One would expect a more diverse population, as the Indians are postulated above to be, to have a higher HD and therefore higher DC. That this was not the case in this study could be a matter of chance in the selection of sample subjects. It could be due to the closer relatedness amongst the North and South Indians investigated, making DC lower, as compared to the Zulu population, where GD is promoted in Zulus by the practice of exogamy.

Comparison of forensic genetic parameters between groups originating from different regions in India (North vs South): DC was higher for South Indians (0.980) than North Indians (0.798) (Table 13). Bindu *et al.* (2007), also found a high DC for the South Indian population of Andhra Pradesh in India (0.999) based on autosomal DNA.

Comparison of forensic genetic parameters between groups which differ in language and or religion: DC was higher for Muslim Indians (1.0) as compared to (1) Hindu Indians (0.842) (Table 13), and (2) Tamil (0.980) and Hindi (0.812) Indians (Table 13). Muslim groups show high Y chromosomal diversity (Rosser *et al.*, 2000) and mtDNA admixture (Terrerros *et al.*, 2007) with Hindu groupings. This may be due to the conversion from Hinduism (oldest religion in the world) to Islam.

Comparison of forensic genetic parameters between surname groups: DC for surname-based groups were high for all surname groups (Table 18) where DC was 1 for all groups, with exception to Maharaj (DC = 0.979), which could be a result of high genetic similarity found within this surname group.

Y-STR haplotype frequencies data is required at a much larger scale than autosomal Y-STR to result in higher reliability and strength in forensic based analyses (Kayser, 2017). These frequencies can be used to identify more information about an unknown individual, such as placing the person within a specific social background i.e. ethnic, region / religion grouping, which gives rise to information that was not previously known. This is permitted in countries such as Netherlands (legislation amended in 2003), US no legislation limiting the use of DNA) and Texas (legislation in place) (Slabbert and Heathfield, 2018). However, SA laws forbid the use of molecular typing phenotyping for purposes other than finding out the gender of an individual to be used in criminal investigation, but does not have a legislation against its use for research purposes.

One of the main ethical concerns of molecular phenotyping is the invasion of privacy, especially if it is used to reveal disease related traits that an individual is not comfortable disclosing. Biomedical ethical procedures ensures that all genetic variation research have both ethical and social implications in its application, particularly in non-medical research (Cho and Sankar, 2004). Although there are many ethical concerns, it has many advantages in forensic investigation and serves as a “Biological witness “for criminal cases. It is also important in

kinship analyses and familiar searching, specifically mission person identification cases (Kayser, 2017).

The restricted use of DNA for phenotypic purposes in SA has been criticised as imposing limitations, with regard to potential of using DNA as efficient tool for forensic investigation in future development (Butler, 2011). Therefore, it recommended by this study and previous literature (Butler, 2011; Slabbert and Heathfield, 2018) that SA laws should be amended to allow additional phenotypic characterises, such as ethnicity, under standard controlled procedures for the use of forensic investigations (Cho and Sankar, 2004).

5. Comparison of experimental samples with samples and populations on the YHRD

None of the haplotypes found in this study were found to match haplotypes on the Y-chromosome STR Haplotype Reference Database (YHRD, 2018), indicating a level of uniqueness of the study samples relative to those on this database.

Based on AMOVA, some of the study samples clustered with groups found on the YHRD indicating a degree of genetic similarity with them. Samples named Singh, Khan and Govender, along with random Hindu, Muslim and Tamil samples formed groups with Keralite and Thoti samples from India, indicative of common genetic origins. Similarly, samples named Dlamini formed a group with Kenyan and Bantu Luhya samples. Other study samples, including those with the surnames Maharaj, Naidoo, Pillay and Zulu, and a cluster containing Cele, Mkhize and Buthelezi surnames, did not cluster with YHRD samples, indicating a degree of genetic distinctness which was supported by RST values which were largely indicative of moderate to high levels of genetic structure (Table 21).

The distinctness of these sample members was supported by their separate positioning in the MDS Plot (Figure 23), in similar regions to YHRD samples and sample groups of similar

ethnicities or from similar geographic regions. For example, samples with the surnames Naidoo, Pillay and Maharaj, were located centrally within the plot along with other samples and groups of Asian origin, consistent with their common geographic origin and probably evolutionary history. Balamurugan *et al.* (2010), in a similar study, also found that Indian samples clustered with Asian populations from the YHRD. This may be due to the fact that many SA Indians are descended from South Asia (Noble, 1994); for example the surnames Naidoo and Pillay are of Sri Lankan descent.

Samples with the surname Zulu were positioned in the top right of the MDS plot, along with other samples and groups of African origin, as was a group comprising sample members with the surnames Cele, Mkhize and Buthelezi. Random White samples from this study were positioned at the bottom left, as were other samples of European origin. In contrast, García *et al.* (2016), who studied the genetics of 6 surnames from the Basque country using the Yfiler® Plus kit, found these to be significantly different to similar (Spanish) populations found in the YHRD.

It was interesting to note that samples with the surname Maharaj, which showed a level of distinctness from North Indian sample members named Khan and Singh in the surname analyses, were positioned separately and closest to the bottom left region of the plot where the European samples were located. This could be consistent with origins within Asia, but on the western side, closer to Europe, or with a level of interbreeding between this group and people of European origin. Bamshad *et al.* (2001), using mtDNA and Y-chromosome markers, found that paternally inherited Y-chromosome variation in upper caste Indians is more similar to Europeans than to Asians, consistent with the position of Maharaj samples relatively close to the European groups in the MDS plot. Further, Zerjal *et al.* (2007) found genetic isolation and drift within the upper castes in Jaunpur, but not the lower castes, which suggests the influence of founder effects and social factors for upper caste surnames (Zhao *et al.*, 2009), such as Maharaj.

6. Challenges and shortcomings

Precautions were taken to prevent contamination however partial profiles were formed for some samples i.e. only some genotypes could be obtained. This may be due to: Volunteers not swabbing hard enough, DNA degradation/PCR inhibitors present, null alleles. This decreases the significance of a match due to the presence of less markers. To minus this, profiles were scored according to the number of alleles and concentration present in each profile for each test to identify which samples needed to be amplified or recollected for an extraction. Some samples were recollected and extracted, however getting a second swab from certain participants were quite difficult due to non-response or not being able to meet for a second swab taking. Samples that could not be reobtained/give a good profile had to be removed from the projects genetic analyses.

Small extra peaks (more than one allele per loci) were observed in some profiles, which were not necessary caused by contamination since there was no signs in the negative control. Allele duplications or triplications have been reported in many other studies (Sanchez *et al.*, 2004; Butler *et al.*, 2005; Leat *et al.*, 2007; Kirsch *et al.*, 2008; Balamurugan *et al.*, 2010; Ge *et al.*, 2010; Purps *et al.*, 2014). The relevant frequency of these duplications are estimated as ~1 % (Butler *et al.*, 2005). However, this an underestimate of its real existence since most studies generally do not mention these multi-peaks (Balamurugan *et al.*, 2010). Theses extra peaks could be due to: Artefacts caused by technology used, formation of stutter bands/shoulder peaks, or even Incomplete Adenylation of products.

DC of this study was moderately high (0.84), but may not be as high as expected, due to some male individuals sharing a haplotype, even though this was avoided where possible and appeared by random sampling (Khubrani *et al.*, 2018). Comparative data found on YHRD are limited for Yfiler® Plus and are not as wide spread across populations as compared to the Yfiler database. This shows the need for more Yfiler® Plus population databases.

7. Recommendations for future research

This study of the genetic genealogy, population and forensic genetics of North Indian, South Indian and Zulu groups residing in the greater Durban area of KZN was based on 27 Y-STR loci amplified by the Yfiler[®] Plus Kit (Thermo Fisher Scientific, Waltham, Massachusetts). This kit was chosen for use in this study as it included the highest number of loci of commercially available kits at the time the study was begun. The reason this study was based on Y-STRs was because one of the primary aims was to study genetic genealogy, based on co-inheritance of surnames and the Y-chromosome (Jobling and Tyler-Smith, 1995; Jobling, 2001). However, in order to obtain a more complete picture of the genetics of the study groups, it will be necessary to complement this study with studies based on autosomal STRs and sequencing of the hypervariable regions of the control region of the mitochondrial DNA (mtDNA).

Y-STR markers are not as informative as autosomal STR markers in many respects (Butler, 2005). Unlike autosomal STRs, Y-STR loci are linked and not independent of one another (Watkins *et al.*, 2008); which does not make not possible to calculate the random match probability over all loci (Hameed *et al.*, 2014), and therefore to calculate a statistic which allows identification of a suspect to individual level. Males from the same paternal lineage share the same Y-STR haplogroup (when no mutations have taken place), therefore Y-STR analyses are unlikely to be able to distinguish between males of the same paternal lineage (Rozhanskii and Klyosov, 2011; Andersen and Balding, 2017). Most commonly used autosomal STR kits amplify STRs which are on different chromosomes, and are therefore unlinked, enabling match probabilities to be calculated and suspects to be identified to individual level (Butler, 2014; Hameed *et al.*, 2014).

Sequencing of the mitochondrial hypervariable (HV) regions 1, 2 and 3 is likely to add a further dimension to a study of the genetics of the people of the greater Durban, KZN area, as mtDNA is maternally inherited, in contrast to Y-STRs, which are paternally inherited and autosomal STRs, which are biparentally inherited. Although HV1, HV2 and HV3 are less variable than STRs,

sequencing of these regions allows determination of the haplogroup on the human mitochondrial phylogenetic tree to which the sample belongs (Senafi *et al.*, 2014). This information is useful in tracking the ancestry of samples.

It is likely that population history based on human Y-Chromosomes is different from that based on mtDNA because cultural barriers determining mating structures might differ between males and females, and because there are likely to have been behavioural differences between males and females in migrations, wars and colonisations (Jobling and Tyler-Smith, 1995). Maternal lineages may share the same haplotype (with exception to the occurrence of mutations). However, traces of paternal inheritance of mtDNA have also been found (Kraytsberg *et al.*, 2004). Therefore, mtDNA should also be analysed for a more comprehensive study.

When used in conjunction with autosomal DNA and or mtDNA, Y-STR profiling could potentially be used to predict the ethnic and geographic regions of origin of an individual (Lao *et al.*, 2006), which is useful in forensic investigations (Kayser, 2017). Therefore, further research including autosomal and mtDNA profiling needs to be conducted in order to provide a more rounded and comprehensive picture of the genetics of the major populations of the greater Durban areas of KZN, SA. Surname-based studies should include a wide variety of non-related individuals sharing a common clan name (Greeff *et al.*, 2010). The scope of this study could also be extended beyond North and South Indians, and Zulus to include other population/ethnic groups found in the Durban area, such as people of mixed race (Coloureds) and people of European and Caucasian descent (Whites).

8. Conclusion

This study contributes to the Indian DNA profiling database and could potentially serve as a baseline for further research opportunities as there are many different common surnames among different ethnic groups. SA has one the highest number of rape reports, therefore, a large growing database plays a critical role in establishing a more stable approach in how strong Y profiles can play in forensic investigations (Brenner, 2010; Andersen *et al.*, 2013; Andersen and Balding, 2017; Cereda, 2017).

Although the smaller RMu marker set (7 loci) gave the highest mean GD, use of an increased number of marker loci, as exemplified by the Yfiler® Plus kit (27 loci), resulted in a higher number of unique haplotypes, a higher DC and a lower MP, making the data most reliable for forensic investigation (Ge *et al.*, 2010; D'Amato *et al.*, 2011; Ottaviani *et al.*, 2015). HD and DC are the highest for the Yfiler® marker set and MP lowest relative to other commonly used marker sets (MHt, Yfiler, and RMu). This also validates the decision to use the Yfiler® Plus kit in this study. The Yfiler® Plus kit presents more informative haplotypes (due to higher number of useful loci), a greater robustness, sensitivity, sensibility and higher DC than Yfiler® kit, specifically cases with high proportions of female DNA (Cainé, 2016; Phan, 2017).

Conclusions based on the study's sub-groups were based on the target surnames chosen for this study, therefore they cannot be considered as an overall population within each sub-grouping. As hypothesised, the Zulu and Indian sample groups were shown in multiple analyses to exhibit considerable genetic structuring. This difference, within these surname oriented sub-groupings, postulated to have developed in the years following separation of the two groups by migration of the ancestors of the Indian populations out of Africa into Asia. The introduction of indentured labourers from India to the Durban region resulted in these groups again occupying the same geographic region, which led to varying degrees of interbreeding between the two groups. However, the relatively short time (~160 years) since the arrival of the Indians in Durban, combined with various cultural and legal barriers to interbreeding, has resulted in the

maintenance of clearly detectable Y-chromosome genetic structure among the two groups. The hypothesis, that there is still detectable genetic structure among the Durban area Indians whose surnames indicate origins in North vs South India, was supported. The hypothesis that Y-chromosome based genetic structure occurs among Muslims and Hindus, within these surname oriented sub-groupings, from the greater Durban, KZN area was weakly accepted. As hypothesized, the existence of genetic structure based on language, within surname oriented sub-groupings, was supported for all comparisons (Tamil vs Hindi, Tamil vs Muslim (originally Urdu speaking), and Hindi vs Muslim (originally Urdu), although this may have been confounded, in some cases, by region-based differences (North vs South Indians).

Relatively low levels of genetic structure were found among the surnames Khan, Maharaj and Singh, somewhat supporting the hypothesis above. The surname Maharaj (Brahmin caste, Hindu) appeared most distinct, possibly due to divergence owing to intermarriage barriers based on caste (with Singhs) and religion (with Khans, who are Muslim). Little genetic structure was observed amongst the South Indian Tamil surnames, Govender, Naidoo and Pillay. There are no known barriers to intermarriage among people bearing these surnames, making it unlikely that they would have diverged from one another through time, and that this would be reflected in Y-STR based genetic structure. There was a relatively high degree of genetic structure among groups with the surnames Buthelezi, Cele, Dlamini, Mkhize and Zulu, despite cultural rules forbidding marriage within patrilineages. The research hypothesis that each surname group would form an independent Y-chromosome haplotype cluster and be monophyletic with high fidelity transmission, was rejected.

None of the study samples shared haplotypes with those on the YHRD. Some study samples were positioned separately on the MDS plot, whereas others formed groups with samples from the YHRD, indicative of common genetic origins. The samples which were separate were still grouped in proximity to other samples of similar geographic/ethnic origin. Overall, the Indian study samples appeared to have a South Asian origin, although the Maharaj surname appeared to be positioned as close to the European samples as to the other Asian samples, possibly

indicative of a west Asian genetic origin. The positioning of the Zulu samples appeared to indicate shared origins with samples from Kenya and with Bantu Luhya samples.

Overall African/Zulu and or Indian based sub-groupings, have the highest DC (with the highest HD and lowest MP), with the African's having the lowest DC (with the lowest HD and highest MP). African sub-groupings, within this study and from YHRD, are the most distant from other sub groupings, supporting the "out of African" theory. Surname inheritance was not monophyletic with high fidelity was assumed. However, the combination of Y-STRs and surnames are useful in forensic practices. Genetic structure and diversity analyses revealed that patterns were not surname based, but better explained at an ethnic and region level. Analyses of autosomal and mtDNA would help to explain the population histories in more depth.

REFERENCES

- Al-Azem, M., El Andari, A., and Mansour, I. (2017). Estimation of Allele and Haplotype Frequencies for 23 YSTR Markers in the Lebanese Population. *Forensic Research & Criminology International Journal*, 5(2). doi:10.15406/frcij.2017.05.00150
- Alamy. (2017). What is India's caste system? *BBC News*. India. 20 July 2017. 29 January 2018. Retrieved from <http://www.bbc.com/news/world-asia-india-35650616>.
- Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). *Molecular Biology of the Cell* (4th Edition ed.). Garland Science, New York.
- Andersen, M. M., and Balding, D. J. (2017). How Convincing Is A Matching Y-Chromosome Profile? *bioRxiv*, 131920.
- Andersen, M. M., Caliebe, A., Jochens, A., Willuweit, S., and Krawczak, M. (2013). Estimating trace-suspect match probabilities for singleton Y-STR haplotypes using coalescent theory. *Forensic Science International: Genetics*, 7(2), 264-271.
- Athey, T. W. (2005). Haplogroup Prediction from Y-STR Values Using an Allele-Frequency Approach. *Journal of Genetic Genealogy*, 1, 1-7.
- Bachtrog, D. (2013). Y-chromosome evolution: emerging insights into processes of Y-chromosome degeneration. *Nature Reviews Genetics*, 14(2), 113-124.
- Balamurugan, K., Suhasini, G., Vijaya, M., Kanthimathi, S., Mullins, N., Tracey, M., and Duncan, G. (2010). Y chromosome STR allelic and haplotype diversity in five ethnic Tamil populations from Tamil Nadu, India. *Legal Medicine*, 12(5), 265-269.
- Ballantyne, K. N., Goedbloed, M., Fang, R., Schaap, O., Lao, O., Wollstein, A., Choi, Y., van Duijn, K., Vermeulen, M., and Brauer, S. (2010). Mutability of Y-chromosomal microsatellites: rates, characteristics, molecular bases, and forensic implications. *The American Journal of Human Genetics*, 87(3), 341-353.
- Ballantyne, K. N., Keerl, V., Wollstein, A., Choi, Y., Zuniga, S. B., Ralf, A., Vermeulen, M., de Knijff, P., and Kayser, M. (2012). A new future of forensic Y-chromosome analysis: rapidly mutating Y-STRs for differentiating male relatives and paternal lineages. *Forensic Science International: Genetics*, 6(2), 208-218.
- Bamshad, M., Fraley, A. E., Crawford, M. H., Cann, R. L., Busi, B. R., Naidu, J. M., and Jorde, L. B. (1996). mtDNA variation in caste populations of Andhra Pradesh, India. *Human biology*, 1-28.
- Bamshad, M., Kivisild, T., Watkins, W. S., Dixon, M. E., Ricker, C. E., Rao, B. B., Naidu, J. M., Prasad, B. R., Reddy, P. G., and Rasanayagam, A. (2001). Genetic evidence on the origins of Indian caste populations. *Genome research*, 11(6), 994-1004.
- Bayly, S. (2001). *Caste, society and politics in India from the eighteenth century to the modern age* (Vol. 3): Cambridge University Press.
- Bindu, G. H., Trivedi, R., and Kashyap, V. (2007). Allele frequency distribution based on 17 STR markers in three major Dravidian linguistic populations of Andhra Pradesh, India. *Forensic science international*, 170(1), 76-85.
- Bosch, E., Lee, A. C., Calafell, F., Arroyo, E., Henneman, P., de Knijff, P., and Jobling, M. A. (2002). High resolution Y chromosome typing: 19 STRs amplified in three multiplex reactions. *Forensic science international*, 125(1), 42-51.
- Brain, J. (1985). *125 Years-The Arrival of Natal's Indians in Pictures*. Natalia: Natal Society Foundation.

- Brenner, C. H. (2010). Fundamental problem of forensic mathematics—the evidential value of a rare haplotype. *Forensic Science International: Genetics*, 4(5), 281-291.
- Brook, C. (2017). *The internet surname database*. Name Origin Research website. Retrieved 22 March 2017, Database site: <http://www.surnamedb.com/>
- Bryant, D., and Moulton, V. (2002). *NeighborNet: An agglomerative method for the construction of planar phylogenetic networks*. Paper presented at the International Workshop on Algorithms in Bioinformatics.
- Budowle, B., Aranda, X. G., Lagace, R. E., Hennessy, L. K., Planz, J. V., Rodriguez, M., and Eisenberg, A. J. (2008). Null allele sequence structure at the DYS448 locus and implications for profile interpretation. *International journal of legal medicine*, 122(5), 421-427.
- Budowle, B., Nhari, L. T., Moretti, T. R., Kanoyangwa, S. B., Masuka, E., Defenbaugh, D. A., and Smerick, J. B. (1997). Zimbabwe black population data on the six short tandem repeat loci—CSF1PO, TPOX, THO1, D3S1358, VWA and FGA. *Forensic science international*, 90(3), 215-221.
- Butler, J. M. (2005). *Forensic DNA typing: biology, technology, and genetics of STR markers*: Academic Press.
- Butler, J. M. (2011). *Advanced topics in forensic DNA typing: methodology*: Academic Press.
- Butler, J. M. (2014). *Advanced topics in forensic DNA typing: interpretation*: Academic Press.
- Butler, J. M., Decker, A. E., Kline, M. C., and Vallone, P. M. (2005). Chromosomal duplications along the Y-chromosome and their potential impact on Y-STR interpretation. *Journal of Forensic Science*, 50(4), JFS2004481-2004487.
- Cainé, L. S. R. M. (2016). Y-STR markers, haplotype discrimination and sensibility in sexual assault cases. The impact of different technologies.
- Cavalli-Sforza, L. L., Menozzi, P., and Piazza, A. (1994). *The history and geography of human genes*: Princeton university press.
- Cereda, G. (2017). Impact of model choice on LR assessment in case of rare haplotype match (frequentist approach). *Scandinavian Journal of Statistics*, 44(1), 230-248.
- Chander, P. (2003). *India: Past and Present*: APH Publishing.
- Chang, Y. M., Perumal, R., Keat, P. Y., Yong, R. Y., Kuehn, D. L., and Burgoyne, L. (2007). A distinct Y-STR haplotype for Amelogenin negative males characterized by a large Y p 11.2 (DYS458-MSY1-AMEL-Y) deletion. *Forensic science international*, 166(2), 115-120.
- Chaudhary, P. (1995). Using surnames to conceal identity. *The Times of India*. 21 February 2009.
- Chetty, K. (2010). *My Roots: memoirs and little tit-bits*. UKZN. Head - Documentation Centre. Retrieved from [http://scnc.ukzn.ac.za/doc/B/Roots/Family Trees/Chetty K/Chetty K My roots and O ther titbits.pdf](http://scnc.ukzn.ac.za/doc/B/Roots/Family%20Trees/Chetty%20K/Chetty%20K%20My%20roots%20and%20O%20ther%20titbits.pdf)
- Cho, M. K., and Sankar, P. (2004). Forensic genetics and ethical, legal and social implications beyond the clinic. *Nature genetics*, 36.
- Coffey, S. (2004). *Intra- and Inter- Marriage Between Ethnic Groups in Beverly in 1895, 1900, and 1905*. Retrieved 31 May 2018, Database site: <http://primaryresearch.org/intra-and-inter-marriage-between-ethnic-groups-in-beverly-in-1895-1900-and-1905/>

- Contralesa. (2016). *One of the Most diverse Nations in the world: South Africa, The place we call home*. Call IT Services. Retrieved 22 February 2018, Database site: <http://contralesa.org/south-african-culture/>
- Coogan, M. D., Brettler, M. Z., Newsom, C. A., and Perkins, P. (2010). *The New Oxford Annotated Bible: New Revised Standard Version: with the Apocrypha: an Ecumenical Study Bible*: Oxford University Press, USA.
- Corander, J., Marttinen, P., Sirén, J., and Tang, J. (2008). Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. *BMC bioinformatics*, 9(1), 539.
- Crow, J. F., and Mange, A. P. (1965). Measurement of inbreeding from the frequency of marriages between persons of the same surname. *Eugenics Quarterly*, 12(4), 199-203.
- D'Amato, M. E., Bajic, V. B., and Davison, S. (2011). Design and validation of a highly discriminatory 10-locus Y-chromosome STR multiplex system. *Forensic Science International: Genetics*, 5(2), 122-125.
- Darwin, G. H. (1875). Marriages between first cousins in England and their effects. *Journal of the Statistical Society of London*, 38(2), 153-184.
- Davies, R. J. (1981). The spatial formation of the South African city. *GeoJournal*, 2, 59-72.
- De Knijff, P., Kayser, M., Caglia, A., Corach, D., Fretwell, N., Gehrig, C., Graziosi, G., Heidorn, F., Herrmann, S., and Herzog, B. (1997). Chromosome Y microsatellites: population genetic and evolutionary aspects. *International journal of legal medicine*, 110(3), 134-140.
- Dirks, N. B. (2011). *Castes of mind: Colonialism and the making of modern India*: Princeton University Press.
- Doniger, W. (1991). *The laws of Manu*: Penguin UK.
- Ehrenreich, L. S. (2005). *The evaluation of Y-STR loci for use in forensics*. University of the Western Cape.
- El-Hazmi, M., Al-Swailem, A., Warsy, A., Al-Swailem, A., Sulaimani, R., and Al-Meshari, A. (1995). Consanguinity among the Saudi Arabian population. *Journal of medical genetics*, 32(8), 623-626.
- Excoffier, L., Smouse, P. E., and Quattro, J. M. (1992). Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics*, 131(2), 479-491.
- Feuerstein, G. (2002). *The yoga tradition: Its history, literature, philosophy and practice*: United Nations Publications.
- Forebears. (2016). *Surname Meaning and Statistics*. Retrieved 8 July 2016, Database site: <http://forebears.io/surnames/>
- Forster, P., Bandelt, H., and Röhl, A. (2017). Network 5.0. 0.1: Phylogenetic Network Software (Version 5.0. 0.1), Fluxus Technology Ltd 1999-2017. Sterner (MP) algorithm developed by Tobias Polzin and Vahdati Daneshmand. Retrieved from www.fluxus-engineering.com.
- Frank, W. E., Ralph, H. C., and Tahir, M. A. (2008). Y chromosome STR haplotypes and allele frequencies in a southern Indian male population. *Journal of forensic sciences*, 53(1), 248-251.
- García, O., Yurrebaso, I., Mancisidor, I., López, S., Alonso, S., and Gusmão, L. (2016). Data for 27 Y-chromosome STR loci in the Basque Country autochthonous population. *Forensic Science International: Genetics*, 20, e10-e12.

- Ge, J., Budowle, B., Planz, J. V., Eisenberg, A. J., Ballantyne, J., and Chakraborty, R. (2010). US forensic Y-chromosome short tandem repeats database. *Legal Medicine*, 12(6), 289-295.
- Ghosh, T., Kalpana, D., Mukerjee, S., Mukherjee, M., Sharma, A. K., Nath, S., Rathod, V. R., Thakar, M. K., and Jha, G. N. (2011). Genetic diversity of 17 Y-short tandem repeats in Indian population. *Forensic Science International: Genetics*, 5(4), 363-367.
- Ghurye, G. S. (1969). *Caste and race in India*: Popular Prakashan.
- Gill, R. (2012). *The structure of Indian society: Then and now*: JSTOR.
- Graves, J. A., Wakefield, M. J., and Toder, R. (1998). The origin and evolution of the pseudoautosomal regions of human sex chromosomes. *Human molecular genetics*, 7(13), 1991-1996.
- Greeff, F. A., Greeff, A. S., Harris, Y., Rinken, L., and Welgemoed, D. (2010). Clan, tribe and household: Y-DNA & one name studies. *Journal of Genetic Genealogy*, 6(1).
- Hameed, I. H., Jebor, M. A., Ommer, A. J., Yoke, C., Zaidian, H., Al-Saadi, A. H., and Abdulazeez, M. A. (2014). Genetic variation and DNA markers in forensic analysis. *African Journal of Biotechnology*, 13(31).
- Hamilton, C. (1997). Restructuring within the Zulu royal house: clan splitting and the consolidation of royal power and resources under Shaka. *African Studies*, 56(2), 85-113.
- Hammer, M. F. (1994). A recent insertion of an alu element on the Y chromosome is a useful marker for human population studies. *Molecular Biology and Evolution*, 11(5), 749-761.
- Hanson, E. K., and Ballantyne, J. (2007). An ultra-high discrimination Y chromosome short tandem repeat multiplex DNA typing system. *PloS one*, 2(8), e688.
- Huson, D. H., and Bryant, D. (2005). Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution*, 23(2), 254-267.
- Jacobson, C. K., Amoateng, A. Y., and Heaton, T. B. (2004). Inter-racial marriages in South Africa. *Journal of comparative family studies*, 443-458.
- Jayaram, V. (2017). *Hinduism and Caste System*. Unique Resource on Hindu and Bauddha Dharmas. Retrieved 30 January 2018., Database site: http://www.hinduwebsite.com/hinduism/h_caste.asp
- Jobling, M. A. (2001). In the name of the father: surnames and genetics. *TRENDS in Genetics*, 17(6), 353-357.
- Jobling, M. A., and King, T. E. (2004). *The distribution of Y-chromosomal haplotypes: forensic implications*. Paper presented at the International Congress Series.
- Jobling, M. A., and Tyler-Smith, C. (1995). Fathers and sons: the Y chromosome and human evolution. *TRENDS in Genetics*, 11(11), 449-456.
- Jobling, M. A., and Tyler-Smith, C. (2003). The human Y chromosome: an evolutionary marker comes of age. *Nature Reviews Genetics*, 4(8), 598-612.
- Kayser, M. (2017). Forensic use of Y-chromosome DNA: a general overview. *Human genetics*, 1-15.
- Kayser, M., Caglià, A., Corach, D., Fretwell, N., Gehrig, C., Graziosi, G., Heidorn, F., Herrmann, S., Herzog, B., and Hidding, M. (1997). Evaluation of Y-chromosomal STRs: a multicenter study. *International journal of legal medicine*, 110(3), 125.
- Khan, K., Siddiqi, M. H., Abbas, M., Almas, M., and Idrees, M. (2017). Forensic applications of Y chromosomal properties. *Legal Medicine*, 26, 86-91.

- Khubrani, Y. M., Wetton, J. H., and Jobling, M. A. (2018). Extensive geographical and social structure in the paternal lineages of Saudi Arabia revealed by analysis of 27 Y-STRs. *Forensic Science International: Genetics*, 33, 98-105.
- King, T. E., Ballereau, S. J., Schürer, K. E., and Jobling, M. A. (2006). Genetic signatures of coancestry within surnames. *Current Biology*, 16(4), 384-388.
- King, T. E., and Jobling, M. A. (2009). What's in a name? Y chromosomes, surnames and the genetic genealogy revolution. *TRENDS in Genetics*, 25(8), 351-360.
- Kirsch, S., Münch, C., Jiang, Z., Cheng, Z., Chen, L., Batz, C., Eichler, E. E., and Schempp, W. (2008). Evolutionary dynamics of segmental duplications from human Y-chromosomal euchromatin/heterochromatin transition regions. *Genome research*, 18(7), 1030-1042.
- Kivisild, T., Rootsi, S., Metspalu, M., Mastana, S., Kaldma, K., Parik, J., Metspalu, E., Adojaan, M., Tolk, H.-V., and Stepanov, V. (2003). The genetic heritage of the earliest settlers persists both in Indian tribal and caste populations. *The American Journal of Human Genetics*, 72(2), 313-332.
- Klyosov, A. A. (2009). DNA Genealogy, mutation rates, and some historical evidences written in Y-chromosome. I. Basic principles and the method. *Journal of Genetic Genealogy*, 5, 186-216.
- Kraytsberg, Y., Schwartz, M., Brown, T. A., Ebralidse, K., Kunz, W. S., Clayton, D. A., Vissing, J., and Khrapko, K. (2004). Recombination of human mitochondrial DNA. *Science*, 304(5673), 981-981.
- Krenke, B. E., Fulmer, P. M., Miller, K. D., and Sprecher, C. J. (2003). The PowerPlex® Y System. *Profiles DNA*, 6(2), 6-9.
- Kruskal, J. B. (1964). Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29(2), 115-129.
- Kumari, A. V. (1998). *Social Change Among Balijas: Majority Community of Andhra Pradesh*: MD Publications Pvt. Ltd.
- Kwak, K. D., Jin, H. J., Shin, D. J., Kim, J. M., Roewer, L., Krawczak, M., Tyler-Smith, C., and Kim, W. (2005). Y-chromosomal STR haplotypes and their applications to forensic and population studies in east Asia. *International journal of legal medicine*, 119(4), 195-201.
- Landy, F., Maharaj, B., and Mainet-Valleix, H. (2004). Are people of Indian origin (PIO) "Indian"? A case study of South Africa. *Geoforum*, 35(2), 203-215.
- Lane, A., Soodyall, H., Arndt, S., Ratshikhopha, M., Jonker, E., Freeman, C., Young, L., Morar, B., and Toffie, L. (2002). Genetic substructure in South African Bantu-speakers: Evidence from autosomal DNA and Y-chromosome studies. *American journal of physical anthropology*, 119(2), 175-185.
- Lao, O., van Duijn, K., Kersbergen, P., de Knijff, P., and Kayser, M. (2006). Proportioning whole-genome single-nucleotide-polymorphism diversity for the identification of geographic population structure and genetic ancestry. *The American Journal of Human Genetics*, 78(4), 680-690.
- Leat, N., Benjeddou, M., and Davison, S. (2004). Nine-locus Y-chromosome STR profiling of Caucasian and Xhosa populations from Cape Town, South Africa. *Forensic science international*, 144(1), 73-75.
- Leat, N., Ehrenreich, L., Benjeddou, M., Cloete, K., and Davison, S. (2007). Properties of novel and widely studied Y-STR loci in three South African populations. *Forensic science international*, 168(2-3), 154-161.

- Leonard, K., and Weller, S. (1980). Declining subcaste endogamy In India: the Hyderabad Kayasths, 1900-75. *American Ethnologist*, 7(3), 504-517.
- Leser, S. (2018). *The 8 Oldest Religions in the World*. Retrieved 26 July 2018, Database site: <https://theculturetrip.com/asia/articles/the-8-oldest-religions-in-the-world/>
- Lewis, S. (1851). *A Topographical Dictionary of Scotland: Comprising the Several Counties, Islands, Cities, Burgh and Market Towns, Parishes, and Principal Villages, with Historical and Statistical Descriptions: Embellished with Engravings of the Seals and Arms of the Different Burghs and Universities* (Vol. 2): S. Lewis and Company.
- Martinez-Cadenas, C., Blanco-Verea, A., Hernando, B., Busby, G. B., Brion, M., Carracedo, A., Salas, A., and Capelli, C. (2016). The relationship between surname frequency and Y chromosome variation in Spain. *European Journal of Human Genetics*, 24(1), 120-128.
- McEvoy, B., and Bradley, D. G. (2006). Y-chromosomes and the extent of patrilineal ancestry in Irish surnames. *Human genetics*, 119(1-2), 212-219.
- Metspalu, M. (2001). *Common maternal legacy of Indian caste and tribal populations*. Department of Evolutionary Biology. Tartu University, Faculty of Biology and Geography, Institute of Molecular and Cell Biology.
- Misra, V. (2001). Prehistoric human colonization of India. *Journal of Biosciences*, 26(4), 491-531.
- Motladiile, T. W. (2004). *Y-chromosome variation in the South African 'coloured' population*.
- Mountain, J. L., Hebert, J. M., Bhattacharyya, S., Underhill, P. A., Ottolenghi, C., Gadgil, M., and Cavalli-Sforza, L. L. (1995). Demographic history of India and mtDNA-sequence diversity. *American journal of human genetics*, 56(4), 979.
- Moxon, E. R., and Wills, C. (1999). DNA microsatellites: agents of evolution? *Scientific American*, 280(1), 94-99.
- Mukherji, A. (2011). Durban largest 'Indian' city outside India. *The Times of India*. 23 June 2011. 9 July 2016. Retrieved from <http://timesofindia.indiatimes.com/city/mumbai/Durban-largest-Indian-city-outside-India/articleshow/9328227.cms?referral=PM>
- Name Stats SA. (2016). *The most common surnames in South Africa*. Name statistics South Africa. Retrieved 29 April 2016, Database site: <http://www.name-statistics.org>
- Nei, M., and Tajima, F. (1981). DNA polymorphism detectable by restriction endonucleases. *Genetics*, 97(1), 145-163.
- Noble, K. B. (1994). Fearing Domination by Blacks, Indians of South Africa Switch Loyalties. *New York Times*, 22. 22 April 1994. 7 November 2017.
- Olofsson, J. K., Mogensen, H. S., Buchard, A., Børsting, C., and Morling, N. (2015). Forensic and population genetic analyses of Danes, Greenlanders and Somalis typed with the Yfiler® Plus PCR amplification kit. *Forensic Science International: Genetics*, 16, 232-236.
- Orie, L. (2013). Maharaj: Brahmin by any other name. *Trinidad and Tobago's – Newsday*. 22 March 2017. Retrieved from <http://www.newsday.co.tt/commentary/0,174357.html>
- Ottaviani, E., Vernarecci, S., Asili, P., Agostino, A., and Montagna, P. (2015). Preliminary assessment of the prototype Yfiler® Plus kit in a population study of Northern Italian males. *International journal of legal medicine*, 129(4), 729-730.
- Page, D. C., Hughes, J. F., Bellott, D. W., Mueller, J. L., Gill, M. E., Larracuente, A., Graves, T., Muzny, D., Warren, W. C., and Gibbs, R. A. (2010). Reconstructing sex chromosome evolution. *Genome biology*, 11(S1), I21.
- Papiha, S. (1996). Genetic Variation in India. *Human biology*, 68(5), 607.

- Park, M. J., Lee, H. Y., Yang, W. I., and Shin, K.-J. (2012). Understanding the Y chromosome variation in Korea—relevance of combined haplogroup and haplotype analyses. *International journal of legal medicine*, 126(4), 589-599.
- Parkin, E. J., Kraayenbrink, T., Opgenort, J. R. M., van Driem, G. L., Tuladhar, N. M., de Knijff, P., and Jobling, M. A. (2007). Diversity of 26-locus Y-STR haplotypes in a Nepalese population sample: isolation and drift in the Himalayas. *Forensic science international*, 166(2), 176-181.
- Passarino, G., Semino, O., Modiano, G., Bernini, L., and Benerecetti, A. S. (1996). mtDNA provides the first known marker distinguishing proto-Indians from the other Caucasoids; it probably predates the diversification between Indians and Orientals. *Annals of human biology*, 23(2), 121-126.
- Peakall, R., and Smouse, P. E. (2012). GenAIEx 6.5: Genetic analysis in Excel. Population genetic software for teaching and research—an update (Version 6.5). *Bioinformatics*, 28, 2537-2539.
- Phan, A. (2017). *Validation of YFiler Plus Amplification Kit for the San Diego Police Department*.
- Phukan, S. (2011). *The story of the Khan. Writings/ Articles of Dr Satyakam Phukan*. Retrieved from <https://drsatyakamphukan.wordpress.com/the-story-of-the-khan/>
- Posel, D. (2001). What's in a name? Racial categorisations under apartheid and their afterlife. *TRANSFORMATION-DURBAN-*, 50-74.
- Purps, J., Siegert, S., Willuweit, S., Nagy, M., Alves, C., Salazar, R., Angustia, S. M., Santos, L. H., Anslinger, K., and Bayer, B. (2014). A global analysis of Y-chromosomal haplotype diversity for 23 STR loci. *Forensic Science International: Genetics*, 12, 12-23.
- Quintana-Murci, L., Semino, O., Bandelt, H.-J., Passarino, G., McElreavey, K., and Santachiara-Benerecetti, A. S. (1999). Genetic evidence of an early exit of Homo sapiens sapiens from Africa through eastern Africa. *Nature genetics*, 23(4), 437.
- Quintana-Murci, L. s., Krausz, C., and McElreavey, K. (2001). The human Y chromosome: function, evolution and disease. *Forensic science international*, 118(2), 169-181.
- Ramana, G. V., Su, B., Jin, L., Singh, L., Wang, N., Underhill, P., and Chakraborty, R. (2001). Y-chromosome SNP haplotypes suggest evidence of gene flow among caste, tribe, and the migrant Siddi populations of Andhra Pradesh, South India. *European Journal of Human Genetics*, 9(9), 695.
- Rao, S. (2003). The dollar brides—Indian girls marrying NRIs often escape to a hassle-free life. *The Telegraph*, 18.
- Rapone, C., D'Atanasio, E., Agostino, A., Mariano, M., Papaluca, M. T., Cruciani, F., and Berti, A. (2016). Forensic genetic value of a 27 Y-STR loci multiplex (Yfiler® Plus kit) in an Italian population sample. *Forensic Science International: Genetics*, 21, e1-e5.
- Rodig, H., Roewer, L., Gross, A., Richter, T., de Knijff, P., Kayser, M., and Brabetz, W. (2008). Evaluation of haplotype discrimination capacity of 35 Y-chromosomal short tandem repeat loci. *Forensic science international*, 174(2-3), 182-188.
- Roewer, L., Amemann, J., Spurr, N., Grzeschik, K.-H., and Epplen, J. (1992). Simple repeat sequences on the human Y chromosome are equally polymorphic as their autosomal counterparts. *Human genetics*, 89(4), 389-394.
- Rosser, Z. H., Zerjal, T., Hurles, M. E., Adojaan, M., Alavantic, D., Amorim, A., Amos, W., Armenteros, M., Arroyo, E., and Barbujani, G. (2000). Y-chromosomal diversity in Europe

- is clinal and influenced primarily by geography, rather than by language. *The American Journal of Human Genetics*, 67(6), 1526-1543.
- Rowold, D. J., and Herrera, R. J. (2003). Inferring recent human phylogenies using forensic STR technology. *Forensic science international*, 133(3), 260-265.
- Rozhanskii, I. L., and Klyosov, A. A. (2011). Mutation Rate Constants in DNA Genealogy (Y Chromosome). *Advances in Anthropology*, 1(2), 26-34.
- Sachidanandam, R., Weissman, D., Schmidt, S. C., Kakol, J. M., Stein, L. D., Marth, G., Sherry, S., Mullikin, J. C., Mortimore, B. J., and Willey, D. L. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, 409(6822), 928-933.
- SAHO. (2015, 16 November 2015). The first Indian people, 197 men, 89 women and 54 children, arrive in Natal. Retrieved from <http://www.sahistory.org.za>
- SAHO. (2016, 6 July 2017). The Prohibition of Mixed Marriages Act commences. Retrieved from <http://www.sahistory.org.za/dated-event/prohibition-mixed-marriages-act-commences>
- Sahoo, S., and Kashyap, V. (2006). Phylogeography of mitochondrial DNA and Y-Chromosome haplogroups reveal asymmetric gene flow in populations of Eastern India. *American journal of physical anthropology*, 131(1), 84-97.
- Sahoo, S., Singh, A., Himabindu, G., Banerjee, J., Sitalaximi, T., Gaikwad, S., Trivedi, R., Endicott, P., Kivisild, T., and Metspalu, M. (2006). A prehistory of Indian Y chromosomes: evaluating demic diffusion scenarios. *Proceedings of the National Academy of Sciences of the United States of America*, 103(4), 843-848.
- Sanchez, J. J., Brión, M., Parson, W., Blanco-Verea, A. J., Børsting, C., Lareu, M., Niederstätter, H., Oberacher, H., Morling, N., and Carracedo, A. (2004). Duplications of the Y-chromosome specific loci P25 and 92R7 and forensic implications. *Forensic science international*, 140(2), 241-250.
- Scott, E. M., Halees, A., Itan, Y., Spencer, E. G., He, Y., Azab, M. A., Gabriel, S. B., Belkadi, A., Boisson, B., and Abel, L. (2016). Characterization of Greater Middle Eastern genetic variation for enhanced disease gene discovery. *Nature genetics*, 48(9), 1071.
- Seielstad, M., Bekele, E., Ibrahim, M., Touré, A., and Traoré, M. (1999). A view of modern human origins from Y chromosome microsatellite variation. *Genome research*, 9(6), 558-567.
- Senafi, S., Ariffin, S. H. Z., Din, R. D. R., Wahab, R. M. A., Abidin, I. Z. Z., and Ariffin, Z. Z. (2014). Haplogroup Determination Using Hypervariable Region 1 and 2 of Human Mitochondrial DNA. *Journal of Applied Sciences*, 14(2), 197-200.
- Seyedebrahimi, R., Esfandiari, E., Rashidi, B., Salehi, R., Dahghi, A. G., Dabiri, S., and Kheirollahi, M. (2017). Comparison of the Frequency of Y-short Tandem Repeats Markers between Sadat and Non-Sadat Populations in Isfahan Province of Iran. *Advanced biomedical research*, 6.
- Shewale, J. G., Nasir, H., Schneida, E., Gross, A. M., Budowle, B., and Sinha, S. K. (2004). Y-Chromosome STR system, Y-Plex™ 12, for forensic casework: development and validation. *Journal of Forensic Science*, 49(6), JFS2004024-2004013.
- Shrivastava, P., Jain, T., and Trivedi, V. B. (2017). Haplotype data for 17 Y-STR loci in the population of Madhya Pradesh, India. *Forensic Science International: Genetics*, 26, e31-e32.

- Singh, K. S. (1996). *Communities, segments, synonyms, surnames and titles* (Vol. 8): Oxford University Press.
- Singh, S. (2017). *Caste System: BR Ambedkar's Perspective*. Asstt. Prof., Punjabi University Guru Kashi College, Damdama Sahib (Bathinda).
- Sirajudin, M. (2011). *Introduction to Hinduism*. Slideshare. Retrieved 22 February 2018, Database site: <https://www.slideshare.net/hafizi88/hinduism-9934799>
- Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P. J., Cordum, H. S., Hillier, L., Brown, L. G., Repping, S., Pyntikova, T., Ali, J., and Bieri, T. (2003). The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature*, 423(6942), 825-837.
- Slabbert, N., and Heathfield, L. J. (2018). Ethical, legal and social implications of forensic molecular phenotyping in South Africa. *Developing world bioethics*.
- Solé-Morata, N., Bertranpetit, J., Comas, D., and Calafell, F. (2015). Y-chromosome diversity in Catalan surname samples: insights into surname origin and frequency. *European Journal of Human Genetics*, 23(11), 1549-1557.
- SPSS. (2012). IBM SPSS statistics (Version 21). *Boston, Mass: International Business Machines Corp*, 126.
- Stats SA. (2011). *Census 2011*. Database site: http://www.statssa.gov.za/census/census_2011/census_products/Census_2011_Census_in_brief.pdf
- Strassmann, B. I. (1992). The function of menstrual taboos among the Dogon. *Human Nature*, 3(2), 89-131.
- Strassmann, B. I., Kurapati, N. T., Hug, B. F., Burke, E. E., Gillespie, B. W., Karafet, T. M., and Hammer, M. F. (2012). Religion as a means to assure paternity. *Proceedings of the National Academy of Sciences*, 109(25), 9781-9785.
- Sykes, B., and Irven, C. (2000). Surnames and the Y chromosome. *The American Journal of Human Genetics*, 66(4), 1417-1419.
- Takayama, T., Takada, N., Suzuki, R., Nagaoka, S., Watanabe, Y., Kumagai, R., Aoki, Y., and Butler, J. M. (2009). Determination of deleted regions from Yp11. 2 of an amelogenin negative male. *Legal Medicine*, 11, S578-S580.
- Terreros, M. C., Rowold, D., Luis, J. R., Khan, F., Agrawal, S., and Herrera, R. J. (2007). North Indian Muslims: enclaves of foreign DNA or Hindu converts? *American journal of physical anthropology*, 133(3), 1004-1012.
- Thekaekara, M. M. (2016). India's caste system is alive and kicking – and maiming and killing. *The Guardian*. 15 August 2016. 1 February 2018. Retrieved from <https://www.theguardian.com/commentisfree/2016/aug/15/india-caste-system-70-anniversary-independence-day-untouchables>.
- Thompson, J., and Storts, D. (2012). The PowerPlex® Y23 System: A new Y-STR multiplex for casework and database applications. *Profiles in DNA 2012*.
- Tsiana, K. J. (2015). Y-STR profiling of four South African populations using the University of the Western Cape 10 locus set.
- Turrina, S., Caratti, S., Ferrian, M., and De Leo, D. (2015). Deletion and duplication at DYS448 and DYS626 loci: unexpected patterns within the AZFc region of the Y-chromosome. *International journal of legal medicine*, 129(3), 449-455.

- Upadhyay, A. (2016). *What is the Indian caste system and how does it work?* Pakistani American Proud of his Indian Heritage. Retrieved 23 February 2018, Database site: <https://www.quora.com/What-is-the-Indian-caste-system-and-how-does-it-work>
- Wallin, J. M., Holt, C. L., Lazaruk, K. D., Nguyen, T. H., and Walsh, P. S. (2002). Constructing universal multiplex PCR systems for comparative genotyping. *Journal of Forensic Science*, 47(1), 52-65.
- Watkins, W. S., Thara, R., Mowry, B. J., Zhang, Y., Witherspoon, D. J., Tolpinrud, W., Bamshad, M., Tirupati, S., Padmavati, R., and Smith, H. (2008). Genetic variation in South Indian castes: evidence from Y-chromosome, mitochondrial, and autosomal polymorphisms. *BMC genetics*, 9(1), 86.
- Westen, A. A., Kraaijenbrink, T., Clarisse, L., Grol, L. J., Willemse, P., Zuniga, S. B., de Medina, E. A. R., Schouten, R., van der Gaag, K. J., and Weiler, N. E. (2015). Analysis of 36 Y-STR marker units including a concordance study among 2085 Dutch males. *Forensic Science International: Genetics*, 14, 174-181.
- Willuweit, S., and Roewer, L. (2015). The new Y chromosome haplotype reference database. *Forensic Science International: Genetics*, 15, 43-48.
- World Families Network. (2015). *Understand DNA Testing*. World Families Network (WFN). Retrieved 5 May 2016, Database site: <http://www.worldfamilies.net/dnatesting>
- Yaacov, D. B., Arbel-Thau, K., Zilka, Y., Ovadia, O., Bouskila, A., and Mishmar, D. (2012). Mitochondrial DNA variation, but not nuclear DNA, sharply divides morphologically identical chameleons along an ancient geographic barrier. *PloS one*, 7(3), e31372.
- Yadav, B., Raina, A., and Dogra, T. D. (2011). Haplotype diversity of 17 Y-chromosomal STRs in Saraswat Brahmin community of North India. *Forensic Science International: Genetics*, 5(3), e63-e70.
- YHRD. (2018). *Y-chromosome Haplotype Reference Database*. Database site: <http://www.yhrd.org>
- Yule, H., and Burnell, A. C. (1996). *Hobson-Jobson: The Anglo-Indian Dictionary*: Wordsworth Editions.
- Zerjal, T., Pandya, A., Thangaraj, K., Ling, E. Y., Kearley, J., Bertoneri, S., Paracchini, S., Singh, L., and Tyler-Smith, C. (2007). Y-chromosomal insights into the genetic impact of the caste system in India. *Human genetics*, 121(1), 137-144.
- Zerjal, T., Xue, Y., Bertorelle, G., Wells, R. S., Bao, W., Zhu, S., Qamar, R., Ayub, Q., Mohyuddin, A., and Fu, S. (2003). The genetic legacy of the Mongols. *The American Journal of Human Genetics*, 72(3), 717-721.
- Zgonjanin, D., Alghafri, R., Antov, M., Stojiljković, G., Petković, S., Vuković, R., and Drašković, D. (2017). Genetic characterization of 27 Y-STR loci with the Yfiler (®) Plus kit in the population of Serbia. *Forensic science international. Genetics*.
- Zhao, Z., Khan, F., Borkar, M., Herrera, R., and Agrawal, S. (2009). Presence of three different paternal lineages among North Indians: a study of 560 Y chromosomes. *Annals of human biology*, 36(1), 46-59.
- Zhivotovsky, L. A., Underhill, P. A., Cinnioğlu, C., Kayser, M., Morar, B., Kivisild, T., Scozzari, R., Cruciani, F., Destro-Bisol, G., and Spedini, G. (2004). The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. *The American Journal of Human Genetics*, 74(1), 50-61.

APPENDIX

University of KwaZulu-Natal Informed Consent document



Dear research participant

You are being invited to consider participating in a study that involves research in Forensic Genetics. Before agreeing to participate in this research study, it is important that you read and understand the following explanations of the purpose, procedure, potential risks and benefits of the study.

Project title: Genetic genealogy – Surname study using YSTRs, autosomal STRs, mitochondrial DNA sequencing and SNPs.

Principal Investigator: Jennifer Margaret Lamb, Associate Professor in School of Life sciences, College of Agriculture, Engineering and Science- University of KwaZulu-Natal (Westville campus).

Co-investigator: Surina Singh, a masters student in School of Life sciences, College of Agriculture, Engineering and Science- University of KwaZulu-Natal (Westville campus).

Contact details for Surina and her supervisor are listed below:

Name	Designation	Contact	Email
Ms Surina Singh	Masters student	0814597274	212519970@stu.ukzn.ac.za
Prof. Jenny Lamb	Supervisor	031 260 3038	Lambj@ukzn.ac.za

Purpose of the study: You are being invited to consider participating in a study that involves research in Forensic Genetics. The aim of this study is to use paternally-inherited Y-STRs and/or Y-SNPs to trace the inheritance through generations of Indians sharing a common surname and to determine the number of lineages within each surname group. This is likely to be related to the number of separate introductions of a particular surname to South Africa from India in over the last 160 years. Autosomal DNA will be used to generate a DNA profile of the participants. Mitochondrial DNA may also be analyzed to trace maternal relationships. The combination of molecular genetics and surname analysis sheds more light on population structure and history, and falls within the field of HID (human identification) forensic DNA analysis. There is a need for such studies as there are currently no surname studies based on the Indian population in South Africa.

Study procedures: The study is expected to enroll 400 representatives from Indian males sharing common surnames. The work will involve anonymous DNA profiling of each participant. If you participate, your role will be to provide a tissue sample by gently swabbing the inside of both of your cheeks with a sterile swab. This is painless and safe.

Participants will be anonymous. Your name will not be recorded. Stored tissue samples and DNA will be destroyed as soon as they have been processed to form a DNA profile. Participation in this research is voluntary and participants may withdraw at any point.

Benefits/risks of this study: This study will yield information on the heritage and history of different surname groups within the Indian community of Durban, and will also increase our knowledge of autosomal and Y-STR profiles of this section of the population, which will contribute to the advancement of the criminal justice system in South Africa. There are no foreseeable risks involved in the study.

This study has been ethically reviewed and approved by the UKZN Biomedical research Ethics Committee (approval number: BE456/16). In the event of any problems or concerns/questions you may contact the above researchers at The School of Life Sciences, University of KwaZulu-Natal, Westville using the contact details given above, or the UKZN Biomedical Research Ethics Committee, contact details as follows:

BIOMEDICAL RESEARCH ETHICS ADMINISTRATION
Research Office, Westville Campus
Govan Mbeki Building
Private Bag X 54001
Durban
4000
KwaZulu-Natal, SOUTH AFRICA
Tel: 27 31 2604769 - Fax: 27 31 2604609
Email: BREC@ukzn.ac.za

CONSENT

I hereby confirm that I voluntarily give consent to participate in the study entitled: Genetic genealogy – Surname study using YSTRs and SNPs".

I understand the purpose and procedures of the study.

I have been given an opportunity to answer questions about the study and have had answers to my satisfaction.

I declare that my participation in this study is entirely voluntary.

If I have any further questions/concerns or queries related to the study I understand that I may contact the researcher at (provide details).

If I have any questions or concerns about my rights as a study participant, or if I am concerned about an aspect of the study or the researchers then I may contact:

BIOMEDICAL RESEARCH ETHICS ADMINISTRATION
Research Office, Westville Campus
Govan Mbeki Building
Private Bag X 54001
Durban
4000
KwaZulu-Natal, SOUTH AFRICA
Tel: 27 31 2604769 - Fax: 27 31 2604609
Email: BREC@ukzn.ac.za

1. I have read and give consent to the above mentioned *

Check all that apply.

☐ Yes

2. Full name of participant *

(Name, middle name & surname)

Eligibility

3. Is your gender male? *

Mark only one oval.

☐ Yes

☐ If no, unfortunately you cannot take part in the study *After the last question in this section, stop filling out this form.*

4. Has anyone from your paternal lineage taken part in this study? *

(i.e. someone with the same father and/ same paternal forefathers as you)

Mark only one oval.

☐ No

☐ If yes, unfortunately you cannot take part in this study *Stop filling out this form.*

Collection of DNA sample

DNA samples will be collected from you via a buccal (cheek) swab.

You will be required to rub the buccal swab 5 times against the inside of each cheek.

Please state whether/ not a DNA sample has been collected as yet

5. DNA sample collected? *

Mark only one oval.

☐ No

☐ Yes *Skip to question 9.*

Contact information

Please provide the following so that you can be contacted to arrange the collection of DNA sample.

If your application meets the required criteria you will be contacted shortly to arrange for a DNA sample collection.

6. Email address *

7. Cell Number *

8. Address (optional)

Skip to question 10.

DNA sample identity

Please state the unique code given to your DNA sample

This will be found on the buccal swab given to you.

9. Unique DNA code *

Demographic Info

Please answer the following demographic questions for statistical purposes only.
Choose one answer per question.

10. Age group (years)? *

Mark only one oval.

- ☐ 18-27
☐ 28-37
☐ 38-47
☐ 48-57
☐ 58-67
☐ 68-77
☐ 78-87
☐ 88-97

11. What is your race? *

Mark only one oval.

- ☐ Indian
☐ Other: _____

Surname inheritance & distribution related questions

Please answer the following surname inheritance and distribution related questions.
Select the applicable options

1. Surname

- * If "Other" please specify
- * Forefathers refers to your grandfather, great grandfather, great great grandfather.

12. What is your surname? *

Mark only one oval.

- ☐ Singh
☐ Maharaj
☐ Khan
☐ Other: _____

13. Is this the same as your child (if applicable)?

Mark only one oval.

- ☐ Yes
☐ No

14. Is this the same as your father and paternal forefathers? *

Mark only one oval.

- ☐ Yes
☐ No

15. Is this the same as your mother's maiden surname and maternal forefathers? *

Mark only one oval.

- ☐ Yes
☐ No

16. If no for any of the above, specify generations that differ and their surnames

2. Location

* If "Other" please specify

* Forefathers refers to your grandfather, great grandfather, great great grandfather.

17. In which city/ town were you born and brought up? *

Mark only one oval.

- ☐ Durban
☐ Pietermaritzburg
☐ Johannesburg
☐ Cape Town
☐ Other: _____

18. Is this the same as your child (if applicable)?

Mark only one oval.

- ☐ Yes
☐ No

19. Is this the same as your father and paternal forefathers? *

Mark only one oval.

- ☐ Yes
☐ No

20. Is this the same as your mother's and maternal forefathers? *

Mark only one oval.

- ☐ Yes
☐ No

21. If no for any of the above, specify generations that differ and their location

3. Religious belief

* If "Other" please specify

* Forefathers refers to your grandfather, great grandfather, great great grandfather.

22. Are you? *

Mark only one oval.

- ☐ Hindu
☐ Muslim
☐ Christian
☐ Other: _____

23. Is this the same as your child (if applicable)?

Mark only one oval.

- ☐ Yes
☐ No

24. Is this the same as your father and paternal forefathers? *

Mark only one oval.

- ☐ Yes
☐ No

25. Is this the same as your mother's and maternal forefathers? *

Mark only one oval.

- ☐ Yes
☐ No

26. If no for any of the above, specify generations that differ and their religious belief

4. Traditional language

- * If "Other" please specify
- * Forefathers refers to your grandfather, great grandfather, great great grandfather.

27. What is your traditional language? *

Please note that this does not refer to your primary spoken language
Mark only one oval.

- ☐ Hindi
- ☐ Tamil
- ☐ Urdu
- ☐ Other: _____

28. Is this the same as your child (if applicable)?

Mark only one oval.

- ☐ Yes
- ☐ No

29. Is this the same as your father and paternal forefathers? *

Mark only one oval.

- ☐ Yes
- ☐ No

30. Is this the same as your mother's and maternal forefathers? *

Mark only one oval.

- ☐ Yes
- ☐ No

31. If no for any of the above, specify generations that differ and their traditional language

Answer if applicable

If your forefathers were from India please answer the following questions
(migration from India to SA took place around 160 years ago)

* Forefathers refers to your grandfather, great grandfather, great great grandfather

32. Which generation came down from India?

Mark only one oval.

- ☐ Father
☐ Grandfather
☐ Great grandfather
☐ Great great grandfather
☐ Other: _____

33. Do you have any close relatives that are still in India?

Mark only one oval.

- ☐ Yes
☐ No

34. - If yes, state the city In India they from

Contact Details

35. Email *

36. Cell (optional)

37. I would like to see the results of this study *

Mark only one oval.

- ☐ Yes
☐ No

Please click SUBMIT below

If you have any questions feel free to contact Surina Singh - surinsingh@hotmail.co.za



Figure A 1: Research survey form

This form was created using Google surveys and consists of: a) Participation consent and b) questionnaire. Link:
https://docs.google.com/forms/d/e/1FAIpQLSfyVeXpdjHPduZN6YubjPuCXX_ndktqdSL-asqlcgRuwU1Ytw/viewform

Table A 1: The Yfiler® Plus composition

The fluorescent marker dyes and its corresponding amplified loci for the Yfiler® Plus is shown. The Yfiler® Plus Allelic Ladder was used to genotype the analysed samples. The alleles contained in the allelic ladder and the DNA profile of the Control DNA 007 are also listed. Adapted from the Yfiler® Plus manual.

Dye	Loci	Alleles in Allelic Ladder	Control
6-FAM™	DYS576	10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25	19
	DYS389I	9, 10, 11, 12, 13, 14, 15, 16, 17	13
	DYS635	15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30	24
	DYS389II	24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35	29
	DYS627	11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27	21
VIC®	DYS460	7, 8, 9, 10, 11, 12, 13, 14	11
	DYS458	11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24	17
	DYS19	9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19	15
	YGATAH4	8, 9, 10, 11, 12, 13, 14, 15	13
	DYS448	14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24	19
	DYS391	5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16	11
NED™	DYS456	10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24	15
	DYS390	17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29	24
	DYS438	6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16	12
	DYS392	4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20	13
	DYS518	32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49	37
TAZ™	DYS570	10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26	17
	DYS437	10, 11, 12, 13, 14, 15, 16, 17, 18	15
	DYS385	6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28	11,14
	DYS449	22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40	30
SID™	DYS393	7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18	13
	DYS439	6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17	12
	DYS481	17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32	22
	DYF387S1	30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44	35,37
	DYS533	7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17	13

Table A 2: Standard dilution series

As recommend in the Quantifiler Duo kit manual (Thermo Fisher Scientific, Waltham, Massachusetts). Std.= Standard. Conc. = concentration. Control sample = 50 μL of 200 ng/ μL stock.

Standard	Conc. (ng/ μL)	Amounts (a+b)		Dilution facto
		a) Std.	b) $T_{10}E_{0.1}$ /glycogen buffer	
Std. 1	50	Control sample	150 μL	4×
Std. 2	16.7	50 μL of Std. 1	100 μL	3×
Std. 3	5.56	50 μL of Std. 2	100 μL	3×
Std. 4	1.85	50 μL of Std. 3	100 μL	3×
Std. 5	0.62	50 μL of Std. 4	100 μL	3×
Std. 6	0.21	50 μL of Std. 5	100 μL	3×
Std. 7	0.068	50 μL of Std. 6	100 μL	3×
Std. 8	0.023	50 μL of Std. 7	100 μL	3×

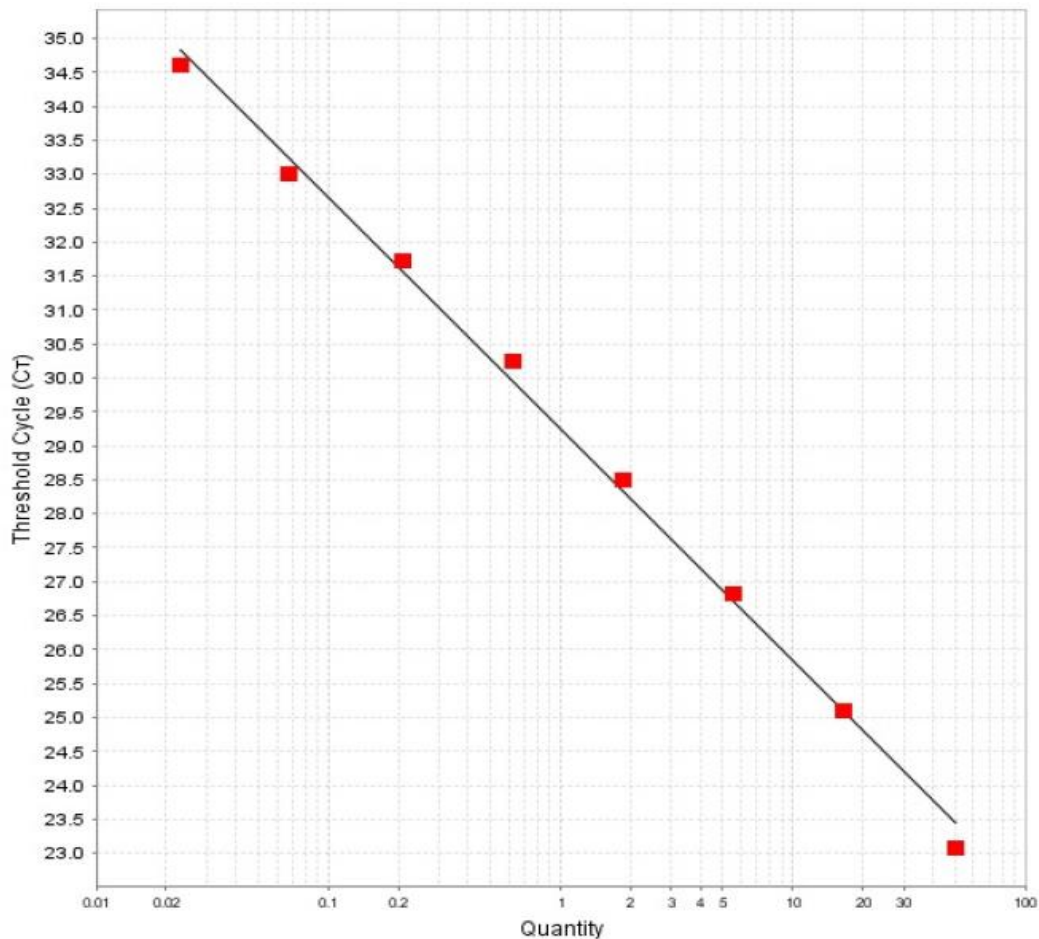


Figure A 2: Stand curve created using the Quantifiler Duo kit (Thermo Fisher Scientific, Waltham, Massachusetts).

Red square = Standard. Curve slope = -3.413. Y-intercept = 29.239. $R^2 = 0.997$.

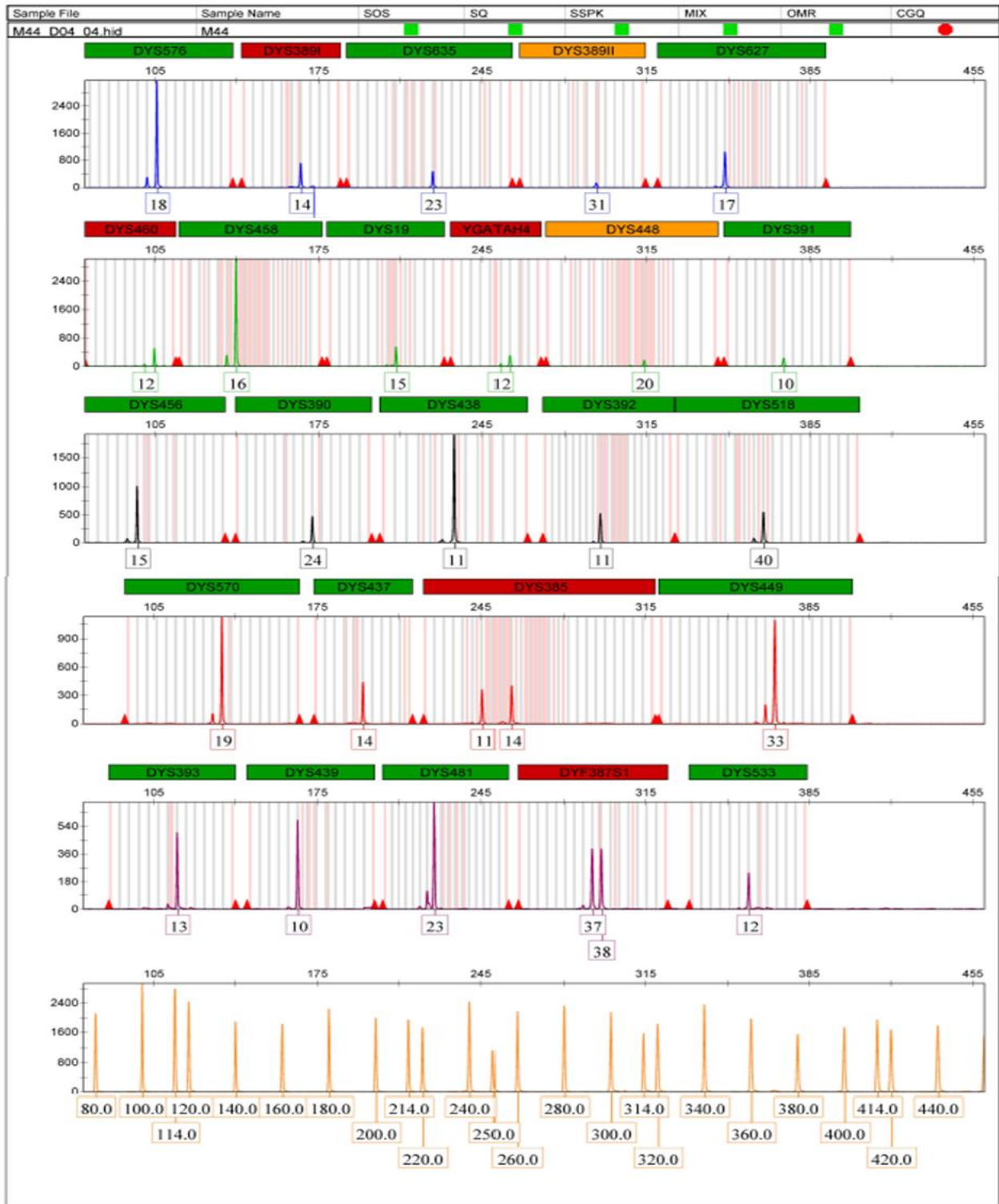


Figure A 3: A Y-STR profile, using the Y-Filer® Plus kit, from GeneMapper® ID-X Software v1.4 (Thermo Fisher Scientific, Waltham, Massachusetts).

Table A 3: Allele frequencies for the overall sample group (n = 399) and sub-groupings.

Allele frequencies are reported for all loci except DYS385a/b and DYS387S1a/b, where genotype frequencies are shown instead (calculated as the combination of the two alleles). K = Khan, M = Maharaj, S = Singh; G = Govender; N = Naidoo, P = Pillay; B = Buthelezi, C = Cele, D = Dlamini, Mk = Mkhize, and Z = Zulu. a = Ethnic group; b = Region; c = Religion of origin in India; d = Language; e = Religion and Language. North = North Indian surname, South = South Indian surnames, Zulu = Zulu surnames. Dash represents the absence of a specific allele in a sub-group.

Locus	Allele / genotyp e	Overall sample	1. Sample sub-groupings								2.Surname based sub-groupings											
			Indian a	Zulu a	North b	South b	Hindu c	Muslim (Urdu) c,d,e	Hindi d,e	Tamil d,e	K North	M North	S North	G South	N South	P South	B Zulu	C Zulu	D Zulu	Mk Zulu	Z Zulu	
DYS19	11	0.005	0.003	0.01	-	0.01	0.004	-	-	0.01	-	-	-	0.033	-	-	-	-	-	0.05	-	-
	12	0.005	0.003	0.01	0.005	-	0.004	-	0.008	-	-	0.023	-	-	-	-	-	0.05	-	-	-	
	13	0.01	0.014	-	0.022	-	0.009	0.034	0.016	-	0.041	0.045	-	-	-	-	-	-	-	-	-	
	13.2	0.005	0.003	0.01	0.005	-	0.004	-	0.008	-	-	0.023	-	-	-	-	0.05	-	-	-	-	
	13.3	0.012	-	0.048	-	-	-	-	-	-	-	-	-	-	-	-	0.05	-	-	0.05	0.15	
	14	0.367	0.387	0.276	0.419	0.327	0.384	0.397	0.43	0.327	0.429	0.409	0.397	0.333	0.433	0.267	0.3	0.15	0.05	0.4	0.45	
	14.2	0.012	0.007	0.029	0.011	-	0.004	0.017	0.008	-	0.02	0.023	-	-	-	-	-	0.1	-	0.05	-	
	14.3	0.197	0.139	0.371	-	0.396	0.175	-	-	0.396	-	-	-	0.367	0.467	0.5	0.4	0.65	0.45	0.25	0.2	
	15	0.195	0.258	0.019	0.371	0.05	0.218	0.414	0.352	0.05	0.367	0.364	0.379	-	-	0.033	-	-	-	-	-	
	15.2	0.002	-	0.01	-	-	-	-	-	-	-	-	-	-	-	-	0.05	-	-	-	-	
	15.3	0.007	-	0.029	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.1	-	0.05	
	16	0.142	0.139	0.162	0.118	0.178	0.162	0.052	0.148	0.178	0.061	0.114	0.155	0.233	0.067	0.2	0.2	0.05	0.3	0.1	0.15	
	17	0.035	0.045	0.01	0.048	0.04	0.035	0.086	0.031	0.04	0.082	-	0.069	0.033	0.033	-	-	-	-	0.05	-	
	18	0.002	-	0.01	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.05	-	
	19	0.002	-	0.01	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.05	-	
DYS385a/b	7	0.001	0.002	-	-	0.005	0.002	-	-	0.005	-	-	-	0.017	-	-	-	-	-	-	-	
	9	0.005	0.007	-	0.011	-	0.009	-	0.004	-	0.032	-	0.009	-	-	-	-	-	-	-	-	
	10	0.003	0.004	-	0.006	-	0.005	-	-	-	0.021	-	-	-	-	-	-	-	-	-	-	
	11	0.08	0.099	0.006	0.099	0.099	0.104	0.083	0.123	0.099	0.043	0.205	0.070	0.150	0.050	0.117	-	-	0.033	-	-	
	12	0.026	0.034	-	0.03	0.04	0.032	0.042	0.016	0.04	0.074	0.023	0.018	0.033	0.017	0.067	-	-	-	-	-	
	12.1	0.003	0.004	-	-	0.01	0.005	-	-	0.01	-	-	-	0.033	-	-	-	-	-	-	-	

	12.2	0.001	0.002	-	0.003	-	0.002	-	0.004	-	-	0.011	-	-	-	-	-	-	-	-	
	13	0.095	0.122	-	0.119	0.129	0.11	0.167	0.111	0.129	0.117	0.148	0.105	0.067	0.167	0.167	-	-	-	-	#VALUE!
	13.1	0.005	0.007	-	-	0.02	0.009	-	-	0.02	-	-	-	0.067	-	-	-	-	-	-	-
	13.2	0.003	0.004	-	0.006	-	0.005	-	0.004	-	0.011	-	0.009	-	-	-	-	-	-	-	-
	14	0.153	0.172	0.056	0.169	0.178	0.173	0.167	0.19	0.178	0.117	0.307	0.132	0.183	0.233	0.100	-	0.059	0.067	0.156	-
	15	0.113	0.131	0.062	0.122	0.149	0.119	0.175	0.111	0.149	0.149	0.091	0.123	0.100	0.117	0.233	-	0.029	0.100	0.094	0.118
	16	0.163	0.138	0.258	0.138	0.139	0.131	0.167	0.139	0.139	0.149	0.045	0.175	0.083	0.217	0.117	0.342	0.265	0.333	0.125	0.206
	16.1	0.004	0.002	0.011	-	0.005	0.002	-	-	0.005	-	-	-	0.017	-	-	-	0.029	0.033	-	-
	16.2	0.005	0.007	-	0.011	-	0.009	-	0.012	-	-	-	0.018	-	-	-	-	-	-	-	-
	17	0.143	0.122	0.225	0.133	0.104	0.126	0.108	0.139	0.104	0.138	0.057	0.167	0.050	0.067	0.150	0.342	0.265	0.133	0.125	0.235
	17.1	0.013	0.011	0.022	-	0.03	0.014	-	-	0.03	-	-	-	0.100	-	-	-	0.059	-	0.031	0.029
	17.2	0.003	0.004	-	0.006	-	0.005	-	0.008	-	-	-	0.009	-	-	-	-	-	-	-	-
	17.2	0.001	0.002	-	-	0.005	-	0.008	-	0.005	-	-	-	-	0.017	-	-	-	-	-	-
	18	0.067	0.048	0.135	0.052	0.04	0.052	0.033	0.052	0.04	0.032	0.068	0.061	0.050	0.050	0.017	0.079	0.088	0.200	0.125	0.176
	19	0.036	0.032	0.051	0.039	0.02	0.036	0.017	0.016	0.02	0.096	0.011	0.026	0.033	0.017	0.017	0.079	0.088	0.033	0.063	-
	20	0.058	0.032	0.146	0.039	0.02	0.036	0.017	0.048	0.02	0.021	0.034	0.053	0.017	0.033	-	0.079	0.059	0.067	0.281	0.235
	21	0.004	-	0.017	-	-	-	-	-	-	-	-	-	-	-	-	0.053	0.029	-	-	-
	22	0.005	0.007	-	0.006	0.01	0.005	0.017	0.004	0.01	-	-	0.009	-	0.017	0.017	-	-	-	-	-
	23	0.003	-	0.011	-	-	-	-	-	-											
	24	0.001	0.002	-	0.003	-	0.002	-	0.004	-	-	-	-	-	-	-	0.026	0.029	-	-	-
	28	0.003	0.004	-	0.006	-	0.005	-	0.008	-	-	-	0.009	-	-	-	-	-	-	-	-
	30	0.003	0.004	-	0.006	-	0.005	-	0.008	-	-	-	0.009	-	-	-	-	-	-	-	-
DYS389II	3	0.003	0.004	-	0.006	-	0.005	-	0.009	-	-	0.027	-	-	-	-	-	-	-	-	-
	9	0.003	0.004	-	0.006	-	0.005	-	0.009	-	-	0.027	-	-	-	-	-	-	-	-	-
	20	0.003	0.004	-	0.006	-	0.005	-	0.009	-	-	0.027	-	-	-	-	-	-	-	-	-
	25	0.003	-	0.01	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.05
	27	0.011	0.015	-	0.012	0.02	0.018	-	0.017	0.02	-	-	0.018	0.033	0.033	-	-	-	-	-	-
	28	0.257	0.257	0.267	0.274	0.228	0.257	0.255	0.282	0.228	0.25	0.081	0.327	0.233	0.3	0.167	0.3	0.15	0.118	0.263	0.45
	29	0.212	0.238	0.119	0.173	0.347	0.266	0.118	0.197	0.347	0.114	0.324	0.145	0.3	0.3	0.433	0.3	-	0.176	0.053	0.1

	30	0.252	0.26	0.228	0.286	0.218	0.216	0.451	0.214	0.218	0.455	0.243	0.255	0.3	0.2	0.2	0.1	0.1	0.647	0.105	0.2
	31	0.196	0.167	0.287	0.179	0.149	0.165	0.176	0.179	0.149	0.182	0.135	0.2	0.133	0.133	0.133	0.3	0.75	0.059	0.316	-
	32	0.053	0.041	0.089	0.042	0.04	0.05	-	0.06	0.04	-	0.108	0.036	-	0.033	0.067	-	-	-	0.263	0.2
	33	0.008	0.011	-	0.018	-	0.014	-	0.026	-	-	0.027	0.018	-	-	-	-	-	-	-	-
DYS389I	11	0.017	0.024	-	0.032	0.01	0.004	0.102	-	0.01	0.12	-	-	0.033	-	-	-	-	-	-	-
	12	0.336	0.326	0.381	0.358	0.267	0.362	0.186	0.435	0.267	0.16	0.283	0.448	0.267	0.4	0.167	0.45	0.15	0.15	0.5	0.6
	13	0.437	0.426	0.448	0.379	0.515	0.422	0.441	0.351	0.515	0.46	0.435	0.362	0.467	0.367	0.633	0.45	0.75	0.75	0.1	0.2
	14	0.193	0.199	0.171	0.2	0.198	0.185	0.254	0.176	0.198	0.26	0.261	0.155	0.2	0.233	0.2	0.1	0.1	0.1	0.4	0.2
	15	0.015	0.021	-	0.026	0.01	0.022	0.017	0.031	0.01	-	0.022	0.034	0.033	-	-	-	-	-	-	-
	21	0.002	0.003	-	0.005	-	0.004	-	0.008	-	-	-	-	-	-	-	-	-	-	-	-
DYS390	11	0.003	0.003	-	0.005	-	0.004	-	0.008	-	-	-	-	-	-	-	-	-	-	-	-
	16	0.003	0.003	-	0.005	-	0.004	-	0.008	-	-	0.023	-	-	-	-	-	-	-	-	-
	17	0.003	-	0.01	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.05
	18	0.003	0.003	-	-	0.01	0.004	-	-	0.01	-	-	-	-	-	0.033	-	-	-	-	-
	19	0.01	0.01	0.01	0.011	0.01	0.013	-	0.016	0.01	-	0.023	0.017	0.033	-	-	-	-	-	-	-
	20	0.003	-	0.01	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.05	-	-
	21	0.276	0.167	0.596	0.21	0.089	0.162	0.19	0.219	0.089	0.204	0.068	0.259	-	0.133	0.167	0.65	0.85	0.75	0.474	0.25
	22	0.201	0.261	0.038	0.21	0.356	0.253	0.293	0.172	0.356	0.306	0.159	0.207	0.4	0.433	0.2	0.15	-	-	-	0.05
	23	0.133	0.178	0.01	0.161	0.208	0.166	0.224	0.133	0.208	0.224	0.136	0.155	0.233	0.167	0.2	-	-	0.05	-	-
	24	0.133	0.16	0.029	0.167	0.149	0.175	0.103	0.195	0.149	0.102	0.386	0.103	0.133	0.133	0.2	-	0.05	-	0.105	-
	24.3	0.003	0.003	-	0.005	-	-	0.017	-	-	0.02	-	-	-	-	-	-	-	-	-	-
	25	0.221	0.202	0.279	0.215	0.178	0.214	0.155	0.242	0.178	0.122	0.205	0.241	0.2	0.133	0.2	0.2	0.1	0.15	0.421	0.6
	26	0.008	0.007	0.01	0.011	-	0.004	0.017	0.008	-	0.02	-	0.017	-	-	-	-	-	-	-	-
	27	0.003	-	0.01	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.05
DYS391	5	0.003	-	0.01	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.05	-
	8	0.003	0.004	-	0.006	-	0.005	-	0.009	-	-	0.029	-	-	-	-	-	-	-	-	-
	8.3	0.003	0.004	-	-	0.01	0.005	-	-	0.01	-	-	-	-	-	0.033	-	-	-	-	-
	9	0.021	0.03	-	0.03	0.03	0.032	0.019	0.034	0.03	-	0.029	0.054	-	0.033	-	-	-	-	-	-
	10	0.776	0.73	0.922	0.686	0.802	0.739	0.692	0.684	0.802	0.75	0.657	0.696	0.867	0.767	0.833	0.75	1	1	0.9	0.95

	11	0.185	0.219	0.068	0.254	0.158	0.206	0.269	0.248	0.158	0.227	0.229	0.25	0.133	0.2	0.133	0.25	-	-	0.05	0.05
	12	0.005	0.007	-	0.012	-	0.009	-	0.017	-	-	0.057	-	-	-	-	-	-	-	-	-
	15	0.003	0.004	-	0.006	-	-	0.019	-	-	0.023	-	-	-	-	-	-	-	-	-	-
	18	0.003	0.004	-	0.006	-	0.005	-	0.009	-	-	-	-	-	-	-	-	-	-	-	-
DYS392	5	0.003	-	0.01	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.05	-	-
	6	0.003	-	0.01	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.05	-
	10	0.074	0.097	0.019	0.096	0.099	0.076	0.185	0.057	0.099	0.152	0.05	0.088	0.133	0.067	0.1	-	-	0.05	-	0.05
	11	0.742	0.716	0.867	0.757	0.644	0.728	0.667	0.797	0.644	0.696	0.625	0.842	0.633	0.533	0.667	0.900	1.000	0.600	0.950	0.850
	12	0.041	0.022	0.095	0.028	0.01	0.013	0.056	0.016	0.01	0.043	-	0.035	-	0.033	-	0.100	-	0.300	-	0.100
	13	0.054	0.047	-	0.057	0.03	0.058	-	0.081	0.03	-	0.225	0.018	0.033	0.033	0.033	-	-	-	-	-
	14	0.077	0.108	-	0.051	0.208	0.116	0.074	0.041	0.208	0.087	0.100	0.018	0.200	0.333	0.167	-	-	-	-	-
	15	0.005	0.007	-	0.006	0.01	0.009	-	0.008	0.01	-	-	-	-	-	0.033	-	-	-	-	-
	37	0.003	0.004	-	0.006	-	-	0.019	-	-	0.022	-	-	-	-	-	-	-	-	-	-
DYS393	10	0.015	0.017	0.01	0.026	-	0.009	0.052	0.015	-	0.061	0.021	0.017	-	-	-	-	-	-	0.050	-
	11	0.062	0.086	-	0.037	0.178	0.099	0.034	0.038	0.178	0.041	0.085	-	0.167	0.300	0.133	-	-	-	-	-
	12	0.116	0.162	-	0.226	0.04	0.12	0.328	0.182	0.04	0.347	0.085	0.241	-	-	-	-	-	-	-	-
	12.1	0.059	0.082	-	-	0.238	0.103	-	-	0.238	-	-	-	0.267	0.233	0.300	-	-	-	-	-
	13	0.556	0.515	0.638	0.6	0.356	0.545	0.397	0.689	0.356	0.367	0.681	0.672	0.400	0.300	0.300	0.650	0.800	0.350	0.700	0.650
	14	0.116	0.11	0.133	0.089	0.149	0.099	0.155	0.061	0.149	0.143	0.106	0.052	0.100	0.133	0.233	0.100	0.050	0.300	0.150	0.100
	15	0.074	0.024	0.219	0.016	0.04	0.026	0.017	0.015	0.04	0.020	0.021	0.017	0.067	0.033	0.033	0.250	0.150	0.350	0.100	0.250
	16	0.002	0.003	-	0.005	-	-	0.017	-	-	0.020	-	-	-	-	-	-	-	-	-	-
DYS437	10	0.002	-	0.01	-	-	-	-	-	-	-	-	-	-	-	-	-	0.050	-	-	-
	11	0.005	0.007	-	0.011	-	-	0.035	-	-	0.042	-	-	-	-	-	-	-	-	-	-
	12	0.002	0.003	-	0.005	-	-	0.018	-	-	0.021	-	-	-	-	-	-	-	-	-	-
	13	0.015	0.021	-	0.032	-	0.009	0.07	0.015	-	0.083	0.043	-	-	-	-	-	-	-	-	-
	14	0.68	0.597	0.952	0.64	0.515	0.622	0.491	0.705	0.515	0.479	0.574	0.741	0.467	0.467	0.567	1.000	0.950	0.950	0.850	1.000
	15	0.196	0.245	0.029	0.175	0.376	0.275	0.123	0.197	0.376	0.125	0.298	0.172	0.367	0.433	0.367	-	-	-	0.150	-
	16	0.092	0.121	-	0.127	0.109	0.09	0.246	0.076	0.109	0.229	0.085	0.086	0.167	0.100	0.067	-	-	-	-	-
	18	0.002	-	0.01	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.050	-	-

	19	0.002	0.003	-	0.005	-	-	0.018	-	-	0,021								
	20	0.002	0.003	-	0.005	-	0.004	-	0.008	-	-	-	-	-	-	-	-	-	-
DYS438	8	0.01	0.014	-	0.005	0.03	0.013	0.017	-	0.03	0.02	-	-	-	0.033	0.067	-	-	-
	9	0.191	0.262	0.01	0.238	0.307	0.259	0.276	0.22	0.307	0.306	0.163	0.276	0.3	0.267	0.267	-	-	0.053
	10	0.264	0.287	0.214	0.286	0.287	0.281	0.31	0.276	0.287	0.306	0.163	0.293	0.233	0.4	0.3	0.3	0.05	0.053
	11	0.47	0.395	0.718	0.411	0.366	0.395	0.397	0.417	0.366	0.367	0.465	0.397	0.467	0.3	0.333	0.7	0.85	0.789
	12	0.058	0.035	0.058	0.049	0.01	0.044	-	0.071	0.01	-	0.163	0.034	-	-	0.033	-	0.1	0.105
	13	0.005	0.004	-	0.005	-	0.004	-	0.008	-	-	0.023	-	-	-	-	-	-	-
	24	0.003	0.004	-	0.005	-	0.004	-	0.008	-	-	0.023	-	-	-	-	-	-	-
DYS439	7	0.003	-	0.01	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.053
	10	0.173	0.24	-	0.262	0.198	0.242	0.228	0.277	0.198	0.208	0.356	0.241	0.267	0.100	0.167	-	-	-
	11	0.398	0.389	0.408	0.433	0.307	0.368	0.474	0.415	0.307	0.521	0.333	0.466	0.200	0.300	0.400	0.450	0.250	0.211
	12	0.323	0.253	0.515	0.219	0.317	0.268	0.193	0.231	0.317	0.167	0.222	0.241	0.333	0.267	0.400	0.450	0.700	0.684
	13	0.073	0.073	0.068	0.059	0.099	0.078	0.053	0.062	0.099	0.042	0.089	0.034	0.100	0.200	0.033	0.100	0.050	0.053
	14	0.025	0.035	-	0.011	0.079	0.039	0.018	0.008	0.079	0.021	-	0.017	0.100	0.133	-	-	-	-
	15	0.005	0.007	-	0.011	-	0.004	0.018	0.008	-	0.021	-	-	-	-	-	-	-	-
	23	0.003	0.003	-	0.005	-	-	0.018	-	-	0.021	-	-	-	-	-	-	-	-
DYS448	11	0.005	0.007	-	0.012	-	0.005	0.019	0.008	-	0.023	-	0.017	-	-	-	-	-	-
	18	0.055	0.07	0.02	0.041	0.119	0.077	0.038	0.042	0.119	0.047	0.029	0.052	0.1	0.1	0.167	0.05	-	0.053
	18.4	0.003	0.004	-	0.006	-	-	0.019	-	-	0.023	-	-	-	-	-	-	-	-
	19	0.509	0.577	0.314	0.556	0.614	0.591	0.519	0.571	0.614	0.512	0.514	0.569	0.567	0.7	0.6	0.4	0.15	0.105
	20	0.22	0.246	0.147	0.257	0.228	0.232	0.308	0.235	0.228	0.302	0.4	0.207	0.233	0.2	0.2	0.05	0.1	0.263
	21	0.202	0.092	0.51	0.123	0.04	0.091	0.096	0.134	0.04	0.093	0.057	0.155	0.1	-	0.033	0.5	0.75	0.579
	22	0.003	0.004	-	0.006	-	0.005	-	0.008	-	-	-	-	-	-	-	-	-	-
	23	0.003	-	0.01	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.053
DYS456	10	0.002	0.003	-	0.005	-	0.004	-	0.008	-	-	-	-	-	-	-	-	-	-
	12	0.015	0.02	-	0.021	0.02	0.021	0.017	0.023	0.02	0.02	0.021	0.017	0.067	-	-	-	-	-
	13	0.029	0.034	0.019	0.047	0.01	0.026	0.067	0.038	0.01	0.078	0.021	0.052	-	-	-	-	-	0.1
	14	0.078	0.102	0.01	0.098	0.109	0.085	0.167	0.068	0.109	0.176	0.125	0.052	0.167	0.033	0.167	-	0.05	-

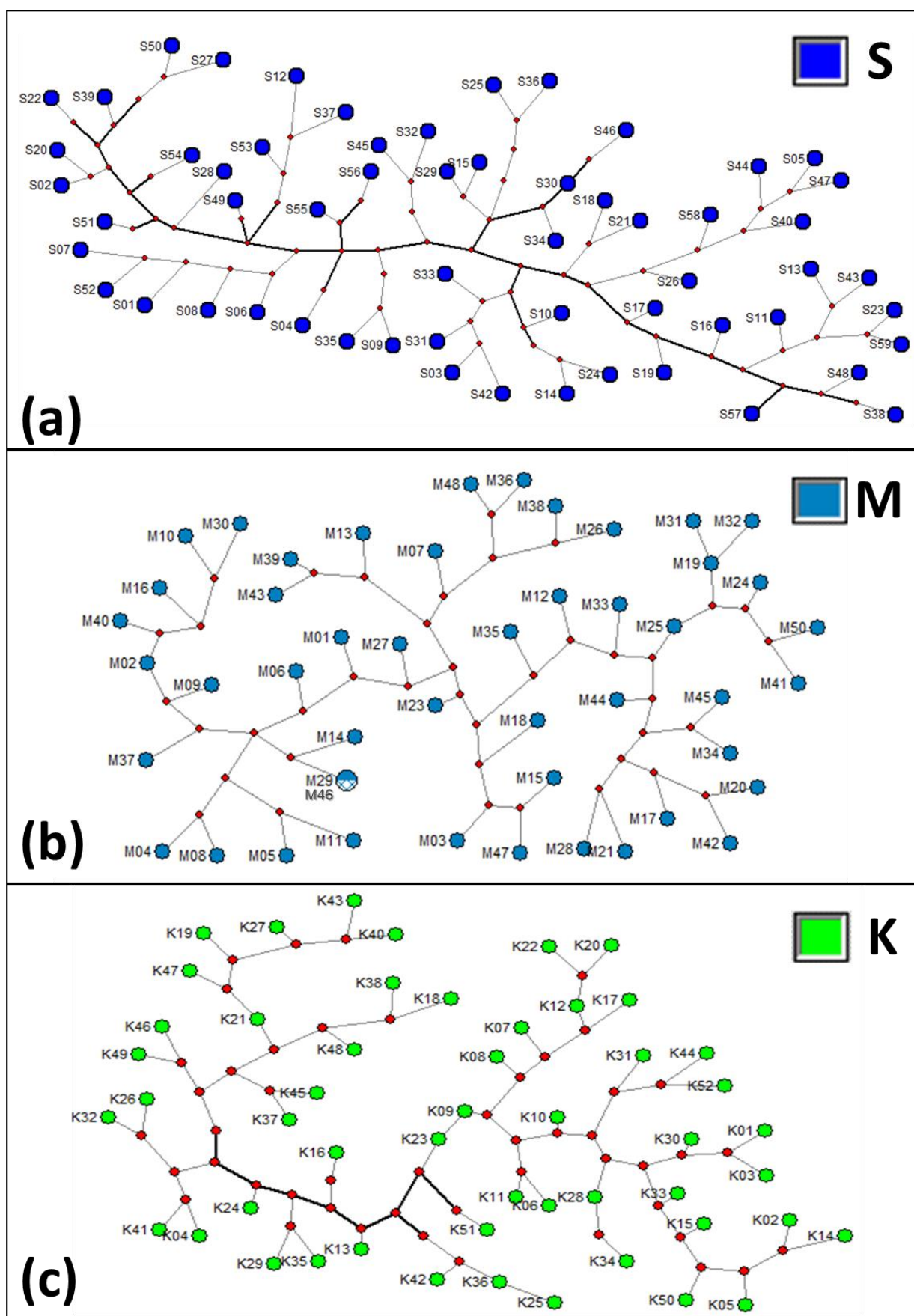
	15	0.685	0.616	0.905	0.658	0.535	0.624	0.583	0.692	0.535	0.549	0.583	0.724	0.333	0.733	0.567	0.95	0.9	1	0.8	0.85
	15.3	0.02	0.02	0.019	-	0.059	0.026	-	-	0.059	-	-	-	0.1	0.033	0.067	-	0.05	-	-	0.05
	16	0.134	0.163	0.038	0.114	0.257	0.179	0.1	0.12	0.257	0.118	0.188	0.103	0.3	0.2	0.2	0.05	-	-	0.05	0.1
	17	0.024	0.024	0.01	0.036	-	0.017	0.05	0.03	-	0.039	0.042	0.034	-	-	-	-	-	-	0.05	-
	18	0.01	0.014	-	0.016	0.01	0.017	-	0.023	0.01	-	0.021	0.017	0.033	-	-	-	-	-	-	-
	20	0.002	0.003	-	0.005	-	-	0.017	-	-	0.02	-	-	-	-	-	-	-	-	-	-
DYS458	11	0.002	0.003	-	0.005	-	0.004	-	0.008	-	-	0.021	-	-	-	-	-	-	-	-	-
	12	0.007	0.01	-	0.005	0.02	0.013	-	0.008	0.02	-	0.021	-	-	0.067	-	-	-	-	-	-
	13	0.007	0.007	-	0.005	0.01	0.009	-	0.008	0.01	-	0.021	-	0.033	-	-	-	-	-	-	-
	14	0.022	0.021	0.029	0.016	0.03	0.026	-	0.023	0.03	-	-	0.017	0.033	0.067	-	0.1	-	-	0.05	-
	14.3	0.01	-	0.039	-	-	-	-	-	-	-	-	-	-	-	-	-	0.235	-	-	-
	15	0.139	0.172	0.039	0.159	0.198	0.155	0.246	0.121	0.198	0.271	0.17	0.086	0.2	0.233	0.167	-	-	0.05	0.15	-
	16	0.306	0.29	0.343	0.296	0.277	0.318	0.175	0.348	0.277	0.146	0.511	0.293	0.267	0.1	0.433	0.25	0.412	0.55	0.2	0.3
	16.3	0.02	0.003	0.069	-	0.01	0.004	-	-	0.01	-	-	-	0.033	-	-	-	0.118	-	0.15	0.1
	17	0.179	0.193	0.147	0.169	0.238	0.193	0.193	0.159	0.238	0.208	0.191	0.172	0.233	0.3	0.133	0.15	-	0.1	0.15	0.25
	17.2	0.007	-	0.029	-	-	-	-	-	-	-	-	-	-	-	-	0.05	-	-	0.1	-
	18	0.254	0.252	0.284	0.307	0.149	0.227	0.351	0.288	0.149	0.354	0.021	0.379	0.133	0.167	0.2	0.45	0.176	0.25	0.2	0.35
	19	0.032	0.034	0.01	0.032	0.04	0.034	0.035	0.03	0.04	0.021	0.043	0.034	0.067	0.067	-	-	0.059	-	-	-
	20	0.01	0.01	0.01	0.005	0.02	0.013	-	0.008	0.02	-	-	0.017	-	-	0.033	-	-	0.05	-	-
	21	0.002	0.003	-	-	0.01	0.004	-	-	0.01	-	-	-	-	-	0.033	-	-	-	-	-
DYS635	15	0.002	0.003	-	0.005	-	0.004	-	0.008	-	-	0.022	-	-	-	-	-	-	-	-	-
	16	0.002	0.003	-	-	0.01	0.004	-	-	0.01	-	-	-	0.033	-	-	-	-	-	-	-
	17	0.005	-	0.019	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.1	-
	18	0.007	0.007	0.01	0.005	0.01	0.004	0.017	-	0.01	0.02	-	-	-	-	0.033	-	-	0.05	-	-
	19	0.007	0.01	-	0.016	-	0.013	-	0.023	-	-	0.022	0.017	-	-	-	-	-	-	-	-
	20	0.117	0.135	0.076	0.112	0.178	0.13	0.153	0.093	0.178	0.16	0.044	0.138	0.2	0.167	0.1	0.1	0.15	0.05	0.05	-
	21	0.216	0.17	0.352	0.128	0.248	0.174	0.153	0.116	0.248	0.18	0.133	0.138	0.133	0.2	0.367	0.3	0.15	0.6	0.3	0.45
	22	0.156	0.125	0.257	0.154	0.069	0.117	0.153	0.155	0.069	0.16	0.133	0.155	-	0.1	0.133	0.3	0.25	0.25	0.45	0.05
	23	0.216	0.246	0.095	0.25	0.238	0.257	0.203	0.271	0.238	0.2	0.489	0.155	0.333	0.2	0.233	0.05	0.4	-	-	-

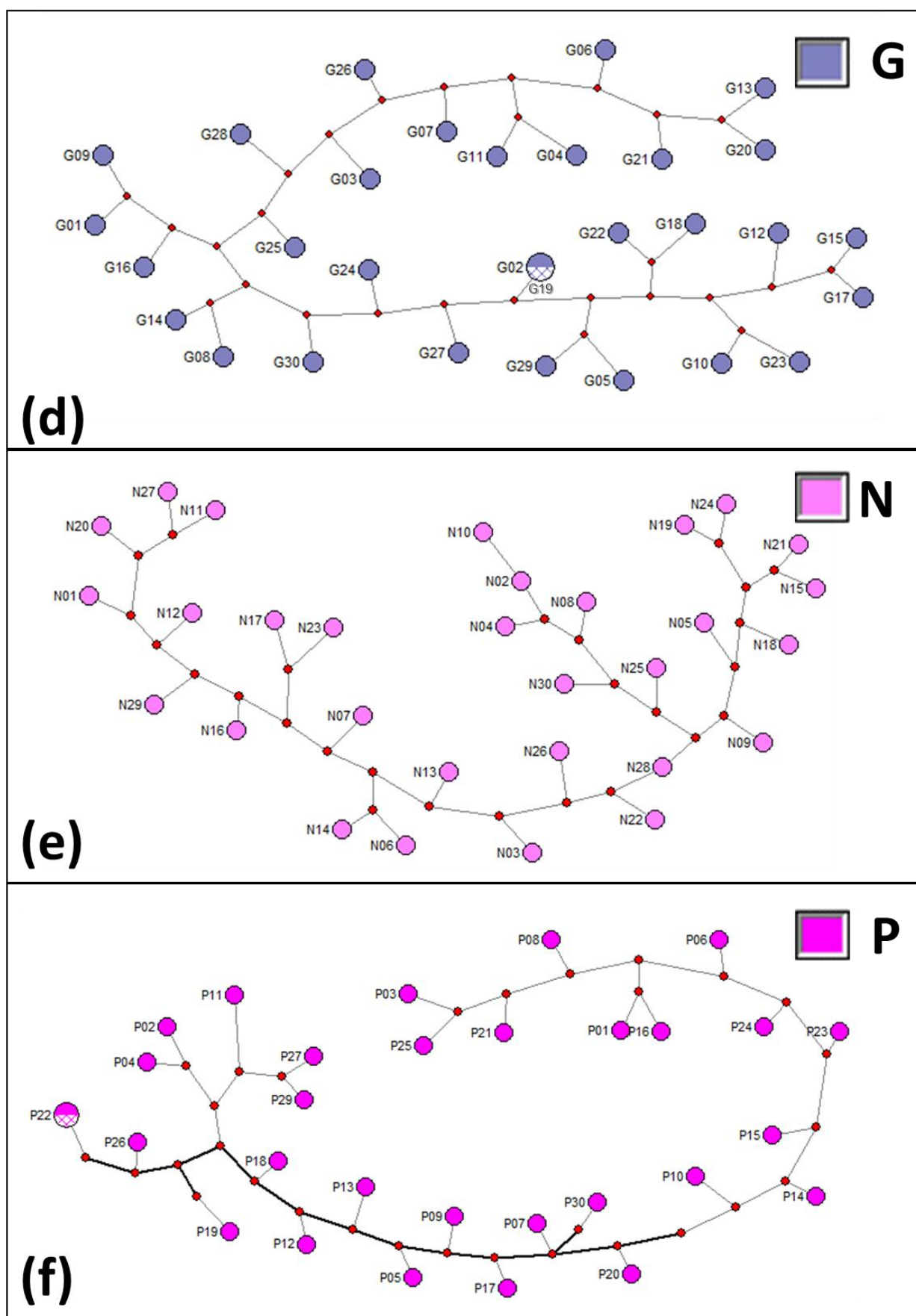
	24	0.208	0.218	0.19	0.271	0.119	0.213	0.237	0.287	0.119	0.2	0.111	0.345	0.133	0.133	0.067	0.25	0.05	0.05	0.1	0.5
	25	0.045	0.059	-	0.043	0.089	0.052	0.085	0.023	0.089	0.08	0.022	0.034	0.067	0.167	0.067	-	-	-	-	-
	26	0.012	0.017	-	0.005	0.04	0.022	-	0.008	0.04	-	-	0.017	0.1	0.033	-	-	-	-	-	-
	27	0.002	0.003	-	0.005	-	0.004	-	0.008	-	-	0.022	-	-	-	-	-	-	-	-	-
	36	0.002	0.003	-	0.005	-	0.004	-	0.008	-	-	-	-	-	-	-	-	-	-	-	-
YGATAH4	10	0.02	0.011	0.049	0.011	0.01	0.009	0.018	0.008	0.01	0.021	-	0.017	-	-	0.033	-	0.05	0.176	-	0.05
	11	0.426	0.376	0.578	0.414	0.307	0.376	0.375	0.432	0.307	0.404	0.293	0.466	0.4	0.267	0.2	0.7	0.35	0.529	0.6	0.7
	12	0.429	0.45	0.353	0.381	0.574	0.451	0.446	0.352	0.574	0.426	0.366	0.397	0.467	0.7	0.567	0.3	0.6	0.235	0.35	0.25
	13	0.115	0.152	0.01	0.177	0.109	0.15	0.161	0.184	0.109	0.149	0.293	0.121	0.133	0.033	0.2	-	-	-	0.05	-
	14	0.008	0.007	0.01	0.011	-	0.009	-	0.016	-	-	0.049	-	-	-	-	-	-	0.059	-	-
	15	0.003	0.004	-	0.006	-	0.004	-	0.008	-	-	-	-	-	-	-	-	-	-	-	-
DYS460	7	0.002	0.003	-	0.005	-	-	0.017	-	-	0.02	-	-	-	-	-	-	-	-	-	-
	9	0.02	0.021	0.019	0.026	0.01	0.017	0.034	0.023	0.01	0.02	0.043	0.017	-	0.033	-	-	-	0.05	0.05	-
	10	0.409	0.399	0.438	0.421	0.356	0.414	0.339	0.458	0.356	0.32	0.261	0.569	0.233	0.3	0.367	0.25	0.8	0.3	0.55	0.25
	11	0.475	0.464	0.505	0.432	0.525	0.453	0.508	0.397	0.525	0.56	0.565	0.31	0.5	0.633	0.6	0.7	0.2	0.6	0.4	0.65
	11.1	0.002	0.003	-	0.005	-	-	0.017	-	-	0.02	-	-	-	-	-	-	-	-	-	-
	12	0.079	0.096	0.038	0.089	0.109	0.103	0.068	0.099	0.109	0.04	0.087	0.103	0.267	0.033	0.033	0.05	-	0.05	-	0.1
	13	0.005	0.007	-	0.011	-	0.004	0.017	0.008	-	0.02	0.022	-	-	-	-	-	-	-	-	-
	15	0.007	0.007	-	0.011	-	0.009	-	0.015	-	-	0.022	-	-	-	-	-	-	-	-	-
DYS481	9	0.003	0.004	-	0.005	-	0.004	-	0.008	-	-	-	-	-	-	-	-	-	-	-	-
	17	0.003	0.004	-	0.005	-	0.004	-	0.008	-	-	0.024	-	-	-	-	-	-	-	-	-
	21	0.03	0.039	-	0.033	0.05	0.04	0.034	0.032	0.05	0.041	0.049	0.017	-	0.067	0.100	-	-	-	-	-
	22	0.116	0.141	-	0.137	0.149	0.146	0.121	0.144	0.149	0.102	0.220	0.138	0.100	0.133	0.200	-	-	-	-	-
	23	0.27	0.335	0.115	0.322	0.356	0.345	0.293	0.336	0.356	0.286	0.390	0.310	0.400	0.367	0.300	0.100	0.150	-	0.150	0.100
	24	0.225	0.268	0.125	0.295	0.218	0.257	0.31	0.288	0.218	0.367	0.146	0.293	0.300	0.233	0.100	0.150	0.050	-	0.100	0.350
	25	0.106	0.099	0.125	0.104	0.089	0.084	0.155	0.08	0.089	0.122	0.098	0.103	0.167	0.033	0.033	0.300	-	0.053	0.050	0.250
	26	0.093	0.053	0.212	0.033	0.089	0.058	0.034	0.032	0.089	0.041	0.073	0.017	0.033	0.133	0.133	0.100	0.100	0.579	0.050	0.250
	27	0.098	0.035	0.279	0.044	0.02	0.035	0.034	0.048	0.02	0.041	-	0.069	-	-	0.067	0.200	0.600	0.105	0.500	-
	28	0.053	0.021	0.144	0.022	0.02	0.022	0.017	0.024	0.02	-	-	0.052	-	-	0.067	0.150	0.100	0.263	0.150	0.050

	30	0.003	0.004	-	-	0.01	0.004	-	-	0.01	-	-	-	-	0.033	-	-	-	-	-	
DY5533	10	0.081	0.104	0.02	0.102	0.109	0.115	0.058	0.12	0.109	0.070	0.073	0.138	0.100	0.067	0.100	-	-	-	0.053	0.056
	11	0.364	0.252	0.687	0.282	0.198	0.217	0.404	0.232	0.198	0.442	0.171	0.259	0.167	0.200	0.267	0.722	0.800	0.842	0.632	0.444
	12	0.512	0.586	0.293	0.559	0.634	0.597	0.538	0.568	0.634	0.488	0.683	0.534	0.733	0.633	0.567	0.278	0.200	0.158	0.316	0.500
	13	0.042	0.054	-	0.051	0.059	0.066	-	0.072	0.059	-	0.073	0.069	-	0.100	0.067	-	-	-	-	-
	16	0.003	0.004	-	0.006	-	0.004	-	0.008	-	-	-	-	-	-	-	-	-	-	-	-
DY5576	12	0.002	-	0.01	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.05	-
	13	0.007	0.01	-	0.01	0.01	0.004	0.033	-	0.01	0.039	-	-	-	-	0.033	-	-	-	-	-
	14	0.054	0.017	0.162	0.021	0.01	0.021	-	0.03	0.01	-	-	0.052	-	-	0.033	0.15	0.15	0.15	0.35	-
	15	0.237	0.17	0.448	0.233	0.05	0.175	0.15	0.271	0.05	0.176	0.104	0.293	-	0.033	0.1	0.45	0.55	0.25	0.3	0.7
	16	0.139	0.122	0.2	0.119	0.129	0.124	0.117	0.12	0.129	0.098	0.063	0.172	0.133	0.1	0.067	0.2	0.1	0.5	0.05	0.2
	17	0.215	0.235	0.162	0.218	0.267	0.214	0.317	0.173	0.267	0.333	0.167	0.207	0.367	0.333	0.133	0.2	0.2	0.1	0.2	0.05
	18	0.196	0.248	0.019	0.223	0.297	0.239	0.283	0.195	0.297	0.255	0.292	0.172	0.333	0.267	0.3	-	-	-	0.05	0.05
	19	0.11	0.146	-	0.13	0.178	0.162	0.083	0.15	0.178	0.078	0.313	0.069	0.133	0.2	0.233	-	-	-	-	-
	20	0.027	0.034	-	0.026	0.05	0.043	-	0.038	0.05	-	0.063	0.017	0.033	0.067	0.067	-	-	-	-	-
	21	0.002	0.003	-	-	0.01	0.004	-	-	0.01	-	-	-	-	-	0.033	-	-	-	-	-
	22	0.007	0.01	-	0.016	-	0.009	0.017	0.015	-	0.02	-	0.017	-	-	-	-	-	-	-	-
	31	0.002	0.003	-	0.005	-	0.004	-	0.008	-	-	-	-	-	-	-	-	-	-	-	-
DYF387S1a/b	11	0.001	0.002	-	0.003	-	0.002	-	0.004	-	-	-	-	-	-	-	-	-	-	-	-
	12	0.001	0.002	-	0.003	-	0.002	-	0.004	-	-	-	-	-	-	-	-	-	-	-	-
	30	0.001	-	0.006	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.028	-	-
	32	0.001	-	0.006	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.028	-	-
	34	0.009	0.007	0.006	0.006	0.01	0.007	0.008	0.008	0.01	-	0.024	-	0.017	0.017	-	-	-	-	-	0.033
	35	0.073	0.085	0.006	0.074	0.104	0.078	0.108	0.078	0.104	0.054	0.122	0.063	0.133	0.150	0.067	-	-	-	0.042	-
	36	0.147	0.126	0.185	0.128	0.124	0.129	0.117	0.123	0.124	0.141	0.183	0.116	0.150	0.100	0.133	0.083	0.375	0.167	0.125	0.167
	37	0.209	0.244	0.113	0.253	0.228	0.247	0.233	0.254	0.228	0.250	0.268	0.241	0.167	0.233	0.233	0.333	0.031	0.083	0.083	0.033
	38	0.249	0.256	0.244	0.29	0.198	0.272	0.2	0.287	0.198	0.315	0.195	0.313	0.183	0.183	0.217	0.250	0.156	0.278	0.208	0.233
	39	0.201	0.175	0.31	0.165	0.193	0.168	0.2	0.156	0.193	0.185	0.134	0.188	0.133	0.117	0.283	0.194	0.344	0.306	0.375	0.367
	40	0.084	0.079	0.107	0.068	0.099	0.081	0.075	0.082	0.099	0.022	0.073	0.071	0.183	0.100	0.050	0.139	0.094	0.111	0.083	0.133

		41	0.018	0.022	0.006	0.011	0.04	0.014	0.05	0.004	0.04	0.033	-	0.009	0.033	0.083	0.017	-	-	-	0.042	-
		42	0.003	0.002	0.006	-	0.005	-	0.008	-	0.005	-	-	-	-	0.017	-	-	-	-	0.042	-
		43	0.001	-	0.006	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.033
DYS449	16	0.003	0.004	-	0.006	-	-	0.018	-	-	0.021	-	-	-	-	-	-	-	-	-	-	-
	23	0.003	0.004	-	0.006	-	-	0.018	-	-	0.021	-	-	-	-	-	-	-	-	-	-	-
	24	0.01	0.014	-	0.017	0.01	0.009	0.036	0.008	0.01	0.043	-	0.018	0.033	-	-	-	-	-	-	-	-
	25	0.013	0.018	-	0.011	0.03	0.018	0.018	0.008	0.03	0.021	-	-	-	0.033	0.067	-	-	-	-	-	-
	26	0.049	0.062	0.02	0.017	0.139	0.077	-	0.025	0.139	-	0.077	-	0.133	0.200	0.133	-	-	0.056	0.050	-	-
	27	0.062	0.025	0.167	0.017	0.04	0.023	0.036	0.008	0.04	0.043	-	0.018	0.033	0.067	0.033	0.150	0.200	0.111	0.300	0.053	-
	28	0.171	0.109	0.353	0.137	0.059	0.109	0.109	0.15	0.059	0.106	0.051	0.161	0.033	0.033	0.133	0.250	0.100	0.444	0.450	0.579	-
	29	0.114	0.12	0.069	0.114	0.129	0.127	0.091	0.125	0.129	0.064	0.128	0.143	0.167	0.067	0.100	0.050	-	0.111	-	0.158	-
	30	0.132	0.12	0.157	0.126	0.109	0.136	0.055	0.158	0.109	0.064	0.128	0.161	0.033	0.100	0.200	0.250	0.500	-	-	-	-
	31	0.109	0.116	0.078	0.126	0.099	0.1	0.182	0.1	0.099	0.191	0.179	0.071	0.100	0.133	0.100	0.050	0.100	0.167	0.050	0.053	-
	32	0.181	0.228	0.069	0.206	0.267	0.235	0.2	0.208	0.267	0.191	0.231	0.232	0.300	0.267	0.133	0.100	0.050	0.111	0.050	0.053	-
	33	0.109	0.123	0.078	0.16	0.059	0.113	0.164	0.158	0.059	0.149	0.179	0.143	0.067	0.100	-	0.150	0.050	-	0.050	0.105	-
	34	0.034	0.047	-	0.051	0.04	0.045	0.055	0.05	0.04	0.064	0.026	0.054	0.100	-	0.033	-	-	-	-	-	-
	35	0.003	0.004	-	0.006	-	-	0.018	-	-	0.021	-	-	-	-	-	-	-	-	-	-	-
	36	0.005	0.004	0.01	-	0.01	0.005	-	-	0.01	-	-	-	-	-	0.033	-	-	-	0.050	-	-
	37	0.003	0.004	-	-	0.01	0.005	-	-	0.01	-	-	-	-	-	0.033	-	-	-	-	-	-
DYS518	11	0.005	0.004	0.01	0.006	-	0.005	-	0.008	-	-	-	0.018	-	-	-	-	-	-	-	-	-
	28	0.003	-	0.01	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	0.050
	36	0.023	0.029	-	0.045	-	0.023	0.053	0.041	-	0.061	0.025	0.036	-	-	-	-	-	-	-	-	-
	37	0.077	0.104	0.01	0.062	0.178	0.113	0.07	0.058	0.178	0.082	0.075	0.054	0.300	0.067	0.167	0.050	-	-	-	-	-
	38	0.152	0.104	0.245	0.101	0.109	0.095	0.14	0.083	0.109	0.143	0.175	0.054	0.167	0.100	0.067	0.100	0.550	0.176	0.150	0.300	-
	39	0.201	0.222	0.147	0.208	0.248	0.221	0.228	0.198	0.248	0.265	0.200	0.250	0.100	0.267	0.300	0.150	0.050	0.176	0.200	0.050	-
	40	0.258	0.276	0.225	0.337	0.168	0.279	0.263	0.372	0.168	0.224	0.275	0.357	0.200	0.133	0.200	0.400	0.250	0.412	0.050	0.100	-
	41	0.101	0.1	0.108	0.073	0.149	0.095	0.123	0.05	0.149	0.122	0.025	0.071	0.100	0.300	0.100	0.050	-	0.176	0.200	0.100	-
	42	0.075	0.072	0.088	0.073	0.069	0.077	0.053	0.083	0.069	0.041	0.125	0.071	0.033	0.067	0.067	-	0.100	0.059	0.300	-	-
	43	0.077	0.075	0.088	0.084	0.059	0.081	0.053	0.099	0.059	0.041	0.075	0.089	0.067	0.067	0.067	0.200	-	-	0.050	0.200	-

	44	0.018	0.011	0.039	0.006	0.02	0.014	-	0.008	0.02	-	0.025	-	0.033	-	0.033	-	0.050	-	0.050	0.100
	45	0.008	-	0.029	0.006	-	-	0.018	-	-	-	-	-	-	-	-	0.050	-	-	-	0.100
	47	0.003	0.004	-	-	-	-	-	-	-	0.020	-	-	-	-	-	-	-	-	-	-
DYS570	10	0.002	0.003	-	0.005	-	0.004	-	0.008	-	-	0.021	-	-	-	-	-	-	-	-	-
	13	0.015	0.017	0.01	0.021	0.01	0.009	0.05	0.008	0.01	0.059	0.021	-	-	-	0.033	-	-	0.050	-	-
	14	0.005	0.003	0.01	0.005	-	-	0.017	-	-	0.020	-	-	-	-	-	-	-	0.050	-	-
	15	0.084	0.113	-	0.073	0.188	0.121	0.083	0.069	0.188	0.098	0.063	0.069	0.167	0.267	0.200	-	-	-	-	-
	16	0.079	0.086	0.058	0.068	0.119	0.099	0.033	0.084	0.119	0.039	0.146	0.052	0.167	0.133	0.067	-	-	0.200	0.053	0.050
	17	0.183	0.202	0.125	0.178	0.248	0.194	0.233	0.153	0.248	0.216	0.125	0.207	0.267	0.233	0.200	0.200	-	0.300	0.105	0.050
	18	0.331	0.315	0.365	0.393	0.168	0.315	0.317	0.427	0.168	0.275	0.375	0.431	0.167	0.100	0.167	0.400	0.300	0.050	0.526	0.550
	19	0.207	0.164	0.337	0.173	0.149	0.151	0.217	0.153	0.149	0.235	0.208	0.138	0.100	0.133	0.233	0.350	0.650	0.100	0.263	0.250
	19.3	0.035	0.021	0.077	-	0.059	0.026	-	-	0.059	-	-	-	0.067	0.100	0.033	0.050	0.050	0.250	-	0.050
	20	0.042	0.055	0.01	0.079	0.01	0.056	0.05	0.092	0.01	0.059	0.042	0.103	-	-	-	-	-	-	0.053	-
	21	0.015	0.017	0.01	-	0.05	0.022	-	-	0.05	-	-	-	0.067	0.033	0.067	-	-	-	-	0.050
	41	0.002	0.003	-	0.005	-	0.004	-	0.008	-	-	-	-	-	-	-	-	-	-	-	-
DYS627	12	0.003	0.004	-	0.006	-	0.005	-	0.008	-	-	-	-	-	-	-	-	-	-	-	-
	15	0.01	0.011	0.01	0.006	0.02	0.009	0.018	-	0.02	0.021	-	-	-	-	0.067	-	-	-	0.05	-
	16	0.008	0.007	0.01	-	0.02	0.009	-	-	0.02	-	-	-	0.033	0.033	-	-	-	0.053	-	-
	17	0.085	0.12	-	0.115	0.129	0.127	0.091	0.126	0.129	0.085	0.211	0.089	0.133	0.2	0.067	-	-	-	-	-
	18	0.244	0.309	0.087	0.391	0.168	0.3	0.345	0.412	0.168	0.362	0.211	0.429	0.267	0.067	0.133	0.3	-	0.053	-	0.1
	18.2	0.003	0.004	-	0.006	-	0.005	-	0.008	-	-	0.026	-	-	-	-	-	-	-	-	-
	19	0.184	0.113	0.375	0.115	0.109	0.095	0.182	0.084	0.109	0.17	0.105	0.107	0.033	0.2	0.1	0.1	0.75	0.632	0.2	0.2
	20	0.187	0.145	0.308	0.132	0.168	0.141	0.164	0.118	0.168	0.149	0.105	0.161	0.1	0.167	0.2	0.45	0.1	0.105	0.45	0.4
	21	0.114	0.12	0.106	0.098	0.158	0.123	0.109	0.092	0.158	0.128	0.105	0.071	0.233	0.1	0.2	0.15	-	0.053	0.05	0.3
	21.2	0.005	0.004	0.01	-	0.01	0.005	-	-	0.01	-	-	-	0.033	-	-	-	-	0.053	-	-
	22	0.096	0.102	0.038	0.086	0.129	0.114	0.055	0.101	0.129	0.043	0.184	0.089	0.133	0.067	0.167	-	0.05	-	0.15	-
	23	0.041	0.036	0.058	0.023	0.059	0.036	0.036	0.017	0.059	0.043	-	0.036	0.033	0.067	0.067	-	0.1	0.053	0.1	-
	24	0.016	0.018	-	0.017	0.02	0.023	-	0.025	0.02	-	0.026	0.018	-	0.067	-	-	-	-	-	-
	25	0.005	0.007	-	0.006	0.01	0.009	-	0.008	0.01	-	0.026	-	-	0.033	-	-	-	-	-	-





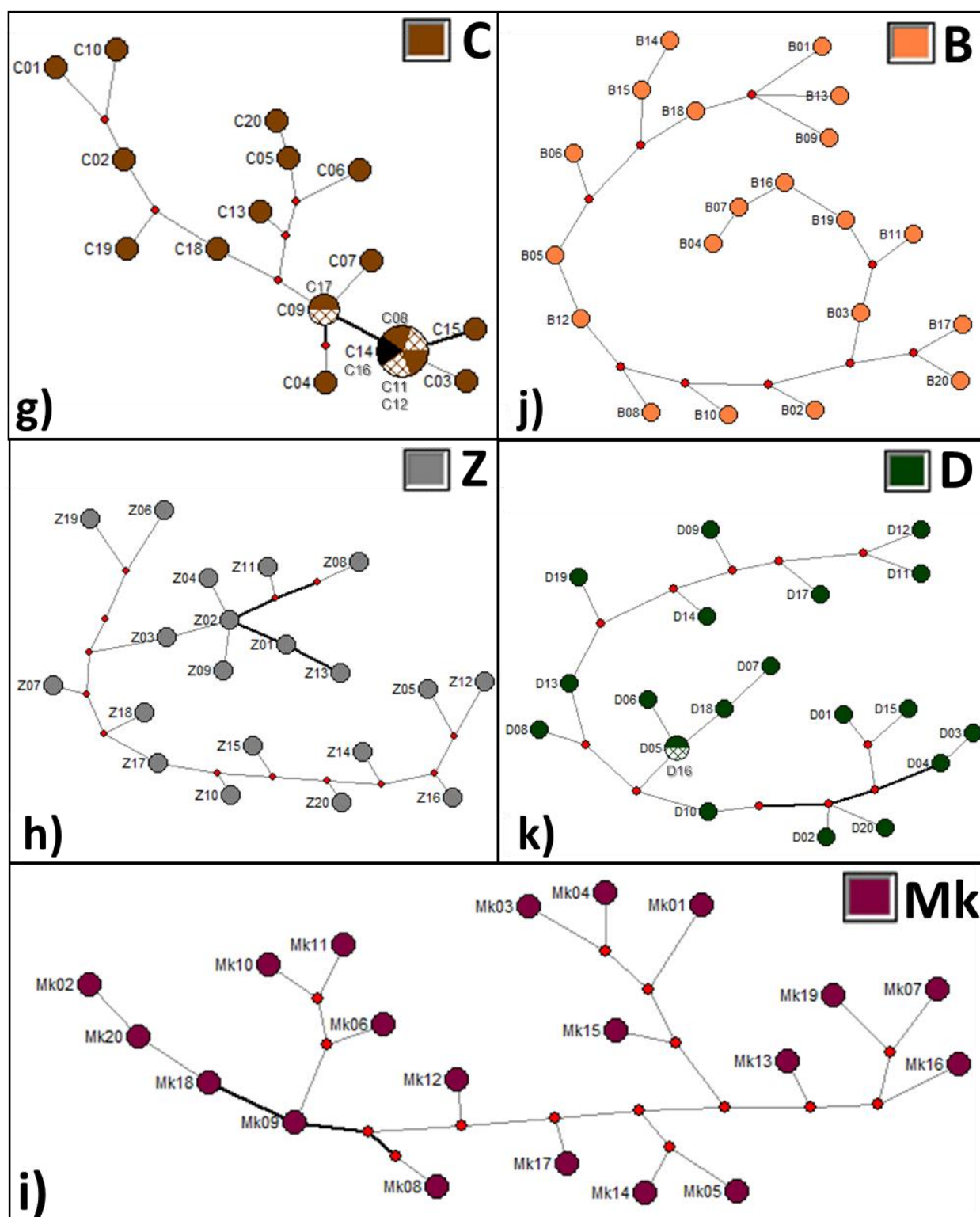


Figure A 4: Haplotype networks for individual surnames, based on Median joining method.

The MJ Network was calculated, based on star conduction and maximum parsimony (MP) to get the simplest parsimony tree, for each surname: a. Singh (S), b. Maharaj (M), c. Khan (K), d. Govender (G), e. Naidoo (N), f. Pillay (P), g. Cele (C), h. Zulu (Z), i. Mkhize (Mk), j. Buthelezi (B), and k. Dlamini (D). Each circle is proportional to the number of individuals. The thick black line outlines the trees torso (i.e. backbone).