

# **Next-generation kinship, ancestry and phenotypic deduction for forensic and genealogical analysis**

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

Margherita Colucci

Department of Genetics and Genome Biology  
College of Life Sciences, University of Leicester  
September 2021



UNIVERSITY OF  
**LEICESTER**

## **Abstract**

### **Next-generation kinship, ancestry and phenotypic deduction for forensic and genealogical analysis**

*Margherita Colucci*

When a suspect cannot be identified by searching an investigative database of DNA profiles, attempts can be made to gather intelligence from DNA. These include kinship testing (finding relatives of the sample donor), and predicting externally visible characteristics (EVCs) and population of origin (biogeographical ancestry; BGA). This project explores the potential of genome-wide SNP chips and targeted massively parallel sequencing (MPS) in these areas.

SNP chip data arguably forms the “gold standard” for kinship analysis. In a set of eight German pedigrees, the performance of SNP-chip analysis was compared to that of the MPS ForenSeq DNA Signature Prep. Kit, which can analyse up to 230 markers, including autosomal SNPs, and autosomal, X-chromosomal and Y-chromosomal STRs. Different methods (PLINK, GENESIS, PRIMUS - SNP chip data; forrel R package - MPS data) were evaluated. Incorporating information *post facto* on X-, Y- and mtDNA SNPs added value in some scenarios. The three methods for dense autosomal SNP data performed comparably in kinship estimation and pedigree reconstruction. Kinship coefficients were estimated from the MPS data. The sequence data revealed additional variation in some complex STR arrays and SNP flanking regions, and these variants together with the set of 230 targeted markers offered higher resolution in identity-by-descent estimation. To mimic forensic scenarios, real and simulated STR data were used in an implementation of kinship estimation via likelihood for multiple searches (“blind search”), including founder inbreeding and the addition of X-STRs.

Finally, the ForenSeq kit was used to infer pigmentation phenotypes and estimate ancestry in a sample of African-Portuguese admixed individuals from Cape Verde, for whom genome-wide ancestry estimates and direct measurements of skin (melanin index) and eye colour (T-index) were available. This highlighted difficulties in both BGA and EVC estimation in admixed populations, and suggested that model-based approaches to ancestry are more useful than principal components analysis.

This research project was supervised by Prof. Mark A. Jobling, and Prof. Nuala A. Sheehan.

Work described in Chapter 3 of this thesis has been published as:

Kjelgaard Brustad, H., Colucci, M., Jobling, M.A., Sheehan, N.A., and Egeland, T. (2021). **Strategies for pairwise searches in forensic kinship analysis.** *Forensic Sci Int Genet*, 54: 102562.  
doi: <https://doi.org/10.1016/j.fsigen.2021.102562>

A copy of the paper is included in the electronic appendix.

## Acknowledgements

Heartfelt thanks and deep gratitude to my supervisors, Prof Mark A. Jobling and Prof Nuala A. Sheehan. It was a great experience to work with them on this project, discuss about this fascinating subject and be inspired by their kind, thoughtful and wise way of guiding me and the group - their example is the true picture of scientists I admire and I take as model for my emerging career. I can't express in words how grateful I am for their constant support during this adventure that was the PhD.

The best adventures are experiences that can be shared: the G2 lab at University of Leicester has been my family in these past years; a heartfelt thanks to Dr Jon Wetton, Dr Celia May, Dr Jodie Lampert, Dr Tünde Huszar, Dr Yahya Khubrani, Dr Jordan Beasley, Orie Shaw, Ettore Fedele, Emily Patterson, and others - they made great the life out of the lab too (sorry for the self-organised, endless hiking trips!). Also, thanks to the MIBTP colleagues and friends, with whom I have shared many classes and had great scientific (and non-scientific) discussions.

This project was also possible thanks to the technical knowledge of many people. I am deeply grateful to Dr Sandra Beleza for the useful discussions and for sharing data so fundamental to this study. Thanks are due to Reshma Vaghela, Dr Nic Sylvius, and Rita Neumann for the technical support and suggestions (and nice chat!) in these years. Special thanks to Dr Chiara Batini, for her kind guidance and advice, and the Genetic Epidemiology group, for interesting discussions and meetings. I wish to thank the people from our commercial partner, Verogen, and to Dr Burkhard Rolf and Dr Chris Phillips for sharing data with us. Finally, thanks to Prof Denise Syndercombe Court and Prof Louise Wain for their time and interest in this thesis and their helpful comments.

Exceptional experiences that I will always cherish were those at NMBU, with Prof Thore Egeland and Dr Hilde Kjelgaard Brustad, where I learned so much and found true friends - thank you!, and at DNA Worldwide - I am very grateful for this long-standing and amazing collaboration. I can't thank enough Dr Martin Blythe and Dr Gerard Serra-Vidal for their time, suggestions and insights (and good laugh too).

And most importantly, all this was possible thanks to the love and unconditional support from my family, always "close" to me although in a different country. Their loving belief in me is my strength. Thanks to my brother Andrea, for being funny and brave when I need it the most. Special thanks to Cinna, always there for me.

# Table of Contents

<b>CHAPTER 1: INTRODUCTION.....</b>	<b>1</b>
1.1 PATTERNS OF INHERITANCE OF THE HUMAN GENOME.....	1
1.2 AUTOSOMAL RECOMBINATION, LINKAGE AND LINKAGE DISEQUILIBRIUM.....	4
1.3 THE SEX CHROMOSOMES AND MITOCHONDRIAL DNA.....	6
1.3.1 <i>X chromosome</i> .....	6
1.3.2 <i>Y chromosome</i> .....	6
1.3.3 <i>Mitochondrial DNA</i> .....	7
1.4 VARIATION IN THE HUMAN GENOME AND ITS USE IN FORENSIC GENETICS.....	8
1.4.1 <i>Short Tandem Repeats (STRs)</i> .....	8
1.4.2 <i>Single Nucleotide Polymorphisms (SNPs)</i> .....	11
1.4.3 <i>Methods for typing forensic STRs and SNPs</i> .....	15
1.5 GENETIC RELATEDNESS .....	20
1.5.1 <i>Identity by State (IBS) and Identity by Descent (IBD)</i> .....	22
1.5.2 <i>Approaches to kinship estimation</i> .....	28
1.5.3 <i>Relatedness analysis in forensic science</i> .....	35
1.6 BIOGEOGRAPHICAL ANCESTRY AND PHENOTYPIC PREDICTION IN FORENSIC SCIENCE .....	39
1.6.1 <i>Human population history and biogeographical ancestry (BGA)</i> .....	40
1.6.2 <i>Human population history and externally visible characteristics (EVCs)</i> .....	42
1.7 AIMS AND OBJECTIVES OF THIS THESIS .....	45
<b>CHAPTER 2: KINSHIP ESTIMATION IN A HOMOGENEOUS POPULATION USING GENOME-WIDE SNP DATA.....</b>	<b>47</b>
2.1.1 <i>Summary of the features of PLINK.....</i>	48
2.1.2 <i>Summary of the features of GENESIS .....</i>	49
2.1.3 <i>Summary of the features of PRIMUS.....</i>	51
2.1.4 <i>Tools for sex-linked and uniparentally-inherited markers .....</i>	53
2.1.5 <i>Aims of this Chapter .....</i>	55
2.2 METHODS.....	55
2.2.1 <i>Sample description .....</i>	55
2.2.2 <i>Reference dataset.....</i>	56
2.3 RESULTS .....	61
2.3.1 <i>Description of pedigrees .....</i>	61
2.3.2 <i>Quality Control (QC) analysis .....</i>	65
2.3.3 <i>Population structure analysis .....</i>	69
2.3.4 <i>Kinship estimation .....</i>	75
2.4 DISCUSSION.....	103
<i>Box 1. Overview of the proposed guidance for determining kinship relatedness using genome-wide SNP data.....</i>	106
<b>CHAPTER 3: KINSHIP ESTIMATION IN A EUROPEAN POPULATION USING MASSIVELY-PARALLEL SEQUENCING DATA ON STRS AND SNPs.....</b>	<b>108</b>
3.1.1 <i>Data analysis on the Universal Analysis Software (UAS) and statistics.....</i>	110
3.1.2 <i>Summary of the features of STRait Razor v3 .....</i>	110
3.1.3 <i>Summary of the features of Familias.....</i>	111
3.1.4 <i>Summary of the features of the forrel package.....</i>	111
3.1.5 <i>Aims of this Chapter .....</i>	111
3.2 METHODS.....	111
3.2.1 <i>DNA samples and sequencing .....</i>	111
3.2.2 <i>Reference dataset.....</i>	113
3.2.3 <i>Simulated data .....</i>	113
3.2.4 <i>Data analysis .....</i>	114
3.3 RESULTS .....	120
3.3.1 <i>Sequence data quality .....</i>	120
3.3.2 <i>Autosomal STR alleles and their sequence diversity .....</i>	121

3.3.3 <i>Y STR alleles and their sequence diversity</i> .....	125
3.3.4 <i>X-STR allele description</i> .....	130
3.3.5 <i>SNP description</i> .....	132
3.3.6 <i>Custom variant calling workflow results</i> .....	135
3.3.7 <i>Reference data description (HGDP)</i> .....	138
3.3.8 <i>Evaluating the power of the markers in kinship determination</i> .....	139
3.4 DISCUSSION.....	152
<b>CHAPTER 4: STRATEGIES FOR PAIRWISE SEARCHES IN FORENSIC KINSHIP ANALYSIS.....</b>	<b>154</b>
4.1.2 <i>Overview of commonly used methods</i> .....	158
4.1.3 <i>Aims of this Chapter</i> .....	160
4.2 MATERIALS AND METHODS .....	161
4.2.1 <i>Evaluating the performance of a blind search</i> .....	161
4.2.2 <i>The problem of multiple testing</i> .....	163
4.2.3 <i>The likelihood ratio for X-chromosomal markers</i> .....	164
4.2.4 <i>Sample description</i> .....	164
4.2.5 <i>Implementation</i> .....	165
4.3 RESULTS .....	166
4.3.1 <i>Case 1: Performance of the blind search</i> .....	166
4.3.2 <i>Case 2: Blind search in DVI cases</i> .....	168
4.3.3 <i>Case 3: A blind search using real data</i> .....	170
4.3.4 <i>Case 4: Simulations incorporating founder inbreeding</i> .....	175
4.3.5 <i>Case 5: Blind search with X-chromosomal markers</i> .....	178
4.4 DISCUSSION .....	181
<b>CHAPTER 5: ESTIMATING BIOGEOGRAPHICAL ANCESTRY AND PHENOTYPES IN A PORTUGUESE-WEST AFRICAN ADMIXED POPULATION VIA MASSIVELY PARALLEL SEQUENCING .....</b>	<b>184</b>
5.1.1 <i>Ancestry inference: overview and methods</i> .....	184
5.1.2 <i>Phenotype prediction</i> .....	188
5.1.3 <i>ForenSeq kit: approach to BGA and pigmentation estimation</i> .....	190
5.1.4 <i>Aims of this Chapter</i> .....	194
5.2 MATERIALS AND METHODS.....	194
5.2.1 <i>Sample description</i> .....	194
5.2.2 <i>Sample processing and sequencing</i> .....	197
5.2.3 <i>Data analysis</i> .....	197
5.2.4 <i>Biogeographical ancestry inference</i> .....	198
5.2.5 <i>Phenotype analysis</i> .....	199
5.3 RESULTS .....	200
5.3.1 <i>Sequencing: depth of coverage, missingness, concordance with SNP chip data</i> .....	200
<i>Universal analysis software output</i> .....	200
<i>MPS workflow output</i> .....	202
5.3.2 <i>Ancestry prediction of 1000 Genomes Project samples: considerations on ForenSeq SNPs and UAS output</i> .....	206
5.3.3 <i>Ancestry prediction of an admixed population: considerations on ForenSeq kit and UAS output</i> .....	209
<i>Comparing the estimated ancestry components</i> .....	217
5.3.5 <i>HlrisPlex-S web-tool phenotypic prediction</i> .....	224
5.4 DISCUSSION .....	230
<b>CHAPTER 6: DISCUSSION .....</b>	<b>235</b>
6.1 SUMMARY OF RESULTS .....	235
6.1.1 <i>Limitations, caveats, and future work</i> .....	237
6.2 FUTURE DIRECTIONS: GENETIC GENEALOGY .....	239
6.3 FUTURE DIRECTIONS: MPS APPROACHES .....	241
<b>BIBLIOGRAPHY .....</b>	<b>243</b>
<b>ELECTRONIC APPENDICES .....</b>	<b>I</b>

APPENDIX 3A TABLE OF AUTOSOMAL STR ISOALLELES RESOLVED BY SEQUENCING.....	III
APPENDIX 3B TABLE OF AUTOSOMAL, X-CHROMOSOME AND Y-CHROMOSOME STR SEQUENCES REPORTED AS "NOVEL" BY STRAIT RAZOR V3 WITH NO CORRESPONDENCE IN STRSEQ.....	X
APPENDIX 4A. BACKGROUND INFORMATION: IMPORTANCE SAMPLING, FAMILY WISE ERROR RATE.....	XIV
APPENDIX 4C. KJELGAARD BRUSTAD, H., COLUCCI, M., JOBLING, M.A., SHEEHAN, N.A., AND EGELAND, T. (2021). "STRATEGIES FOR PAIRWISE SEARCHES IN FORENSIC KINSHIP ANALYSIS." FORENSIC SCI INT GENET 54: 102562.....	XVI

## The list of tables

### Chapter 1

- Table 1.1 Categories of SNPs useful in forensic contexts.
- Table 1.2 Repeat structure variation within 13 STR markers commonly used in forensic analysis.
- Table 1.3 IBD k coefficients for various outbred relationships.

### Chapter 2

- Table 2.1 Parameters used through the QC steps of the dataset (genome-wide SNP data) in PLINK.
- Table 2.2 Expected, pedigree-based coefficient of relationship (r) for some specific relationships.
- Table 2.3 Total number of kin relationship among the 8 analysed families, 66 samples.
- Table 2.4 Information about available kin relationships in the dataset.
- Table 2.5 Subgroups of the family panel used to check for outliers and family structure in the PCA.
- Table 2.6 This table shows the mean of the Relationship coefficient estimated by PLINK and GENESIS.
- Table 2.7 PRIMUS networks reconstruction summary.
- Table 2.8 Y haplogroups in the family set.
- Table 2.9 Table listing the haplogroups in the family.
- Table 2.10 Segment sharing on the X chromosome for sample pairs in family 5, based on 1612 SNPs.

### Chapter 3

- Table 3.1 ForenSeq DNA Signature Prep Kit (Verogen, San Diego, CA) markers.
- Table 3.2 MPS variant calling workflow.
- Table 3.3 Summary of the three runs performed on the MiSeq FGx (ForenSeq) of 70 samples.
- Table 3.4 Autosomal STR sequences reported as “Novel” by STRait Razor v3.
- Table 3.6 Y STR sequences reported as “Novel” by STRait Razor v3.
- Table 3.7 Isoalleles resolved by sequencing.

Table 3.8	X-STR sequences reported as “Novel” by STRait Razor v3 with no correspondence in STRSeq.
Table 3.9	Variants called in the flanking region of ForenSeq SNPs through the MPS workflow.

## **Chapter 4**

Table 4.1	Summary of statistics for a blind search.
Table 4.2	Example summary of 100 simulations of a (hypothetical) blind search with 20 individuals.
Table 4.3	Examples of Likelihood Ratios from of blind search including both reference and victim profiles.
Table 4.4	Alternative hypotheses analysed in the blind search.
Table 4.5	Summary of the blind search, for a LR threshold of 1000,000.
Table 4.6	Results for seven samples from a blind search on a dataset of 65 individuals, genotyped for 27 autosomal STRs.
Table 4.7	Posterior probabilities from the LRs (Table 4.7) of the blind search for seven samples on a dataset of 65 individuals, genotyped for 27 autosomal STRs.
Table 4.8	Posterior probabilities with priors from the LRs (Table 4.7) of the blind search for seven samples on a dataset of 65 individuals, genotyped for 27 autosomal STRs.
Table 4.9	Posterior probabilities averaged over 100 simulations for the comparisons between the four daughters in Figure 4.10.
Table 4.10 a	Posterior probabilities for the comparisons between the three grandmother - grandchildren pairs in Figure 4.11 considering seven relationships.
Table 4.10 b	Posterior probabilities for the comparisons between the three grandmother - grandchildren pairs in Figure 4.11 considering five relationships.

## **Chapter 5**

Table 5.1	Overview of ancestry-informative marker (AIM) sets.
Table 5.2	List of 24 variants used for phenotypic prediction in the ForenSeq system.
Table 5.3	Summary of information on the 30 Cape Verdean samples.
Table 5.4	Number of SNPs and STRs available in the set of 30 Cape Verdean samples.

Table 5.5	Description of seven ForenSeq Plex B under-performing markers, ordered according to the number of samples affected.
Table 5.6	Linkage among the clusters of phenotypic SNPs.
Table 5.7	Summary information on markers called in the ForenSeq SNP regions through the MPS workflow.
Table 5.8	Samples with no UAS phenotypic prediction.
Table 5.9	Information on eye colour prediction for the outlier sample (CVSA34).
Table 5.10	Information on eye colour prediction for the sample with intermediate eye colour (CVF323).
Table 5.11	Number of samples that have pigment predicted for eye and skin colour based on UAS and HIrisPlex-S. na: not available
Table 5.12	HIrisPlex-S results on two non-admixed populations from 1KGP based on 23 ForenSeq phenotypic SNPs.

## The list of figures

### Chapter 1

- Figure 1.1 The human genome.
- Figure 1.2 Inheritance of (a) the male-specific region of the Y (MSY), (b) the mitochondrial DNA, (c) the X-chromosome and of (d) the autosomes.
- Figure 1.3 Chromosomal crossover.
- Figure 1.4 Single nucleotide polymorphisms (SNPs) and a short tandem repeat (STR) shown on two haplotypes.
- Figure 1.5 Stepwise mutation model for STRs.
- Figure 1.6 Autosomal STRs within the GlobalFiler kit.
- Figure 1.7 Massively parallel sequencing (MPS) workflow.
- Figure 1.8 Passage of IBD segments through generations.
- Figure 1.9 The IBD triangle for outbred relationships.
- Figure 1.10 Jacquard's 9 condensed identity states.
- Figure 1.11 Pedigree representing the Habsburg family tree (1440-1740), with Philip and Joanna of Castile as founders.
- Figure 1.12 Family tree of the Ptolemaic dynasty in Egypt (305-30 BC) and IBD estimation.
- Figure 1.13 Pedigree graph.
- Figure 1.14 Likelihood Ratio use in paternity testing (paternity index, PI).
- Figure 1.15 Classic examples of population structure studies based on SNP chip data and two different analytical approaches.
- Figure 1.16 Global distribution of skin colour.
- Figure 1.17 Distribution of hair and eye colour in Europe.

### Chapter 2

- Figure 2.1 Workflow adopted by GENESIS.
- Figure 2.2 QC workflow for both the Family samples and reference dataset (1000 Genomes Project Phase 3).
- Figure 2.3 Workflow for both the Family samples and reference dataset (1000 Genomes Project Phase 3) after QC stages.
- Figure 2.4 a Diagrams of family trees for six out of the eight families (pre-QC).

- Figure 2.4 b Diagrams of family trees for two out of the eight families, Family 4 and 7 (pre-QC).
- Figure 2.5 Sex mismatches in the dataset.
- Figure 2.6 Sample cumulative non-missingness distribution plot (call rate).
- Figure 2.7 Plot comparing the proportion of missing genotypes and heterozygosity rate in the Family dataset.
- Figure 2.8 Graph of (logarithmic) portion of missing SNPs.
- Figure 2.9 SNP cumulative non-missingness distribution plot.
- Figure 2.10 PLINK PCA plot of the first three dimensions, showing the Family samples (Family) grouping with the European subset (EUR).
- Figure 2.11 PCA plot of the first three eigenvectors based on PLINK, including European panel and Family samples.
- Figure 2.12 PLINK PCA plot of the first three eigenvectors for the European panel and the subgroup 5 of unrelated individuals (Family) from the Family dataset.
- Figure 2.13 GENESIS PC-AiR plot of the first two PCs of the German-European merged dataset.
- Figure 2.14 Scatterplot of the estimated coefficient of relationship versus estimated K0 for the Family samples.
- Figure 2.15 Comparison of relationship coefficient distribution for a) parent-offspring (PO) and b) full siblings (FS) relationship based on GENESIS (pink) and PLINK (blue).
- Figure 2.16 Comparison of relationship coefficient distribution for a) half-siblings (HS) and b) avuncular relationship based on GENESIS (pink) and PLINK (blue).
- Figure 2.17 Comparison of relationship coefficient distribution for a) grandparental and b) great-grandparental relationship based on GENESIS (pink) and PLINK (blue).
- Figure 2.18 Comparison of relationship coefficient distribution for a) great-avuncular and b) first cousin relationship based on GENESIS (pink) and PLINK (blue).
- Figure 2.19 Comparison of relationship coefficient distribution for a) first cousin once and b) twice removed relationship based on GENESIS (pink) and PLINK (blue).

- Figure 2.20 Comparison of relationship coefficient distribution for a) second cousin and b) second cousin once removed relationship based on GENESIS (pink) and PLINK (blue).
- Figure 2.21 Comparison of relationship coefficient distribution for unrelated individuals based on GENESIS (pink) and PLINK (blue).
- Figure 2.22 PRIMUS output for family 4 (relatedness cutoff at 2nd degree).
- Figure 2.23 PRIMUS output for family 7 (relatedness cutoff at 2nd degree).
- Figure 2.24 a Inheritance of Y haplogroups in the eight families analysed.
- Figure 2.24 b Inheritance of Y haplogroups in the eight families analysed.
- Figure 2.25 a Mitochondrial DNA haplogroups of the families analysed.
- Figure 2.25 b Mitochondrial DNA haplogroups of the eight families analysed.
- Figure 2.26 Patterns of segment sharing within family 5 and inferred crossovers.

### **Chapter 3**

- Figure 3.1 Massively parallel sequencing (MPS) workflow.
- Figure 3.2 Workflow of the sequencing analysis.
- Figure 3.3 Read depth of 27 autosomal STRs (ForenSeq markers) across 66 samples.
- Figure 3.4 Histogram showing the number of length alleles identified (Total CE alleles) and sequence alleles in the repeat region and flanking region.
- Figure 3.5 Read depth of 24 Y- STRs (ForenSeq markers) across 28 male samples.
- Figure 3.6 Histogram showing the number of length alleles identified (Total CE alleles) and sequence alleles in the repeat region and flanking region.
- Figure 3.7 Read depth of 7 X-STRs across 65 samples (28 males, 37 females).
- Figure 3.8 Histogram showing the number of length alleles identified (Total CE alleles) and sequence alleles in the repeat region and flanking region.
- Figure 3.9 Read depths of 94 ForenSeq identity SNPs.
- Figure 3.10 Read depths of 56 ForenSeq ancestry SNPs.
- Figure 3.11 Read depths of 22 ForenSeq phenotypic SNPs.
- Figure 3.12 Simulation of 1000 parent-offspring (PO) pairs.
- Figure 3.13 Simulation of 1000 full siblings (FS) pairs.
- Figure 3.14 Simulation of 1000 half siblings (HS) pairs.
- Figure 3.15 Simulation of 1000 first cousin (FC) pairs.
- Figure 3.16 Simulation of 1000 second cousin (SC) pairs.

- Figure 3.17 Scatterplot of the 0 IBD ( $k_0$ ) proportion and 2 IBD ( $k_2$ ) for the parent-offspring (PO) relationship.
- Figure 3.18 Scatterplot of the 0 IBD ( $k_0$ ) proportion and 2 IBD ( $k_2$ ) for the full sibling (FS) relationship.
- Figure 3.19 Scatterplot of the 0 IBD ( $k_0$ ) proportion and 2 IBD ( $k_2$ ) for the grandparental (GC) relationship.
- Figure 3.20 Scatterplot of the 0 IBD ( $k_0$ ) proportion and 2 IBD ( $k_2$ ) for the first cousin (FC) relationship.
- Figure 3.21 Scatterplot of the 0 IBD ( $k_0$ ) proportion and 2 IBD ( $k_2$ ) for the second cousin (SC) relationship.
- Figure 3.22 Comparison of relationship coefficient distribution based on STRs (sequence) for (a) parent-offspring (PO) and (b) full siblings (FS) relationship based on real-world family data (pink) and 1000 simulated related pairs (blue).
- Figure 3.23 Comparison of relationship coefficient distribution based on STRs (sequence) for half-siblings (HS), grandparental and avuncular relationship based on real-world family data (pink) and 1000 simulated related pairs (blue).
- Figure 3.24 Comparison of relationship coefficient distribution based on STRs (sequence) for (a) first and (b) second cousin relationship based on real-world family data (pink) and 1000 simulated related pairs (blue).
- Figure 3.25 Scatterplot of the 0 IBD ( $k_0$ ) proportion and 2 IBD ( $k_2$ ) for five relationships based on the 9867 kinship SNPs (ForenSeq Kintelligence).

## **Chapter 4**

- Figure 4.1 Different database searches are presented here.
- Figure 4.2 Figure showing the concept of inbred founders and coefficients of inbreeding.
- Figure 4.3 ROC curve for a blind search among 40 individuals, with 10 of the pairwise comparisons in the search being true siblings.
- Figure 4.4 ROC curve from 100 simulations of a blind search with 4 sibling pairs and a total of 20 individuals.
- Figure 4.5 Pairwise comparisons performed in a blind search as in a disaster victim identification (DVI) scenario.

- Figure 4.6 This pedigree (a) represents an inbred family where individual 1 is the offspring of close relatives (siblings) corresponding to an inbreeding coefficient ( $f$ ) of 0.25.
- Figure 4.7 Boxplots showing log10 of LRs values for the 1000 half-sib pairings (4 and 5 as in Figure 4.6).
- Figure 4.8 Boxplots showing Posterior probability values for 1000 half-sib pairings (4 and 5 as in Figure 4.6).
- Figure 4.9 Boxplots showing Posterior probability values (flat prior) for the investigated individuals (4 and 5 as in Figure 4.5).
- Figure 4.10 Pedigree connecting the individuals of the analysis in Section 5.5.
- Figure 4.11 Pedigree of family 7.

## **Chapter 5**

- Figure 5.1 UAS Phenotype and Biogeographic Ancestry (BGA) estimation, showing probabilities of various pigmentation phenotypes and ancestry prediction.
- Figure 5.2 HGDP selection browser maps for two phenotypic and ancestry SNPs.
- Figure 5.3 Read depth of 27 autosomal ForenSeq STRs.
- Figure 5.4 Analysis of ancestry in African, European, East Asian and Admixed metapopulations from the 1000 Genomes Project using the ForenSeq ancestry SNPs ( $n= 56$ ).
- Figure 5.5 PCA plots of African, European, East Asian and Admixed metapopulations from 1000 Genomes Projects using different sets of ForenSeq SNPs.
- Figure 5.6 Analysis of ancestry in Cape Verdean samples and reference set based on African, European, East Asian and Admixed metapopulations from the 1000 Genomes Project using the ForenSeq ancestry SNPs ( $n= 56$ ).
- Figure 5.7 Analysis of ancestry in Cape Verdean samples and reference set based on African, European, East Asian and Admixed metapopulations from the 1000 Genomes Project using the ForenSeq identity SNPs ( $n= 93$ ).
- Figure 5.8 Analysis of ancestry in Cape Verdean samples and reference set based on African, European, East Asian and Admixed metapopulations from the 1000 Genomes Project using the ForenSeq phenotypic SNPs ( $n= 24$ ).
- Figure 5.9 Analysis of ancestry in Cape Verdean samples and reference set based on African, European, East Asian and Admixed metapopulations from the

1000 Genomes Project using the ForenSeq ancestry and identity SNPs (n = 149).

- Figure 5.10 Analysis of ancestry in Cape Verdean samples and reference set based on African, European, East Asian and Admixed metapopulations from the 1000 Genomes Project using the ForenSeq ancestry, identity and phenotypic SNPs (n = 171).
- Figure 5.11 Correlation between the African ancestry proportion based on the SNP chip and ADMIXTURE using different combinations of markers.
- Figure 5.12 Bar plot comparing individual African ancestry proportions as estimated from different sets of markers.
- Figure 5.13 Individual African ancestral proportions based on the average from ADMIXTURE output using random subsets of 24 ancestry SNPs (out of 56).
- Figure 5.14 The qualitative classification of the phenotype assigned by the UAS model for 25 Cape Verdean individuals compared to the real phenotype (the measure of eye pigmentation, T-Index).
- Figure 5.15 The qualitative classification of the phenotype assigned by the HirisPlex model for 30 Cape Verdean individuals compared to the real phenotype.
- Figure 5.16 The qualitative classification of the phenotype assigned by the HIrisPlex model was compared to the real phenotype using an older version of the model.
- Figure 5.17 Genotypes of the 30 Cape Verdean samples at the 6 loci for eye colour prediction in the HIrisPlex-S with global map of frequency of the respective pigmentation.
- Figure 5.18 Ancestry proportions of Cape Verdean samples obtained using frappe (supervised analysis, K=2, with HapMap CEU and YRI as reference populations).
- Figure 5.19 Workflow for determining eye pigmentation based on the 6 IrisPlex SNPs.

## List of abbreviations

A:	Adenine
AIM:	Ancestry Informative Markers
aSNP:	ancestry-informative single nucleotide polymorphism
aipSNP:	ancestry, identity and phenotypic SNP
autSTR:	autosomal STR
BAM:	Binary Alignment Map
bp:	base pairs
BWA:	Burrows-Wheeler Alignment Tool (Li and Durbin 2009)
C:	Cytosine
CE:	Capillary Electrophoresis
CEPH:	Centre d'Etude du Polymorphisme Humain
CODIS:	Combined DNA Index System
dbSNP:	SNP database
DNA:	Deoxyribonucleic acid
DoC:	Depth of Coverage
EMPOP:	EDNAP mitochondrial DNA population database
EVC:	Externally-visible characteristics
FASTQ:	File format for biological sequence and their quality score
FDP:	Forensic DNA phenotyping
FGx:	Forensic mode of MiSeq Illumina sequencing platform (Verogen)
FST:	Fixation Index
FTA:	Flinders Technology Associates paper
FTDNA:	FamilyTreeDNA
G:	Guanine
GATK:	Genome Analysis Toolkit (Broad Institute, McKenna et al., 2010)
GWAS:	Genome-wide association studies
H:	Heterozygosity
$H_0$ :	Null hypothesis
$H_1$ :	Alternative hypothesis
Hg:	Haplogroup
HGDP:	Human Genome Diversity Project

HGDP-CEPH: Human Genome Diversity Project - Centre d'Etude du Polymorphisme Humain

HWE:	Hardy-Weinberg Equilibrium
IBD:	Identity by Descent
IBS:	Identity by State
iiSNP:	Identity-informative single nucleotide polymorphism
Indel:	Insertion/Deletion Polymorphism
ISFG:	International Society for Forensic Genetics
ISOGG:	International Society of Genetic Genealogy
KYA:	Thousand Years Ago
LD:	Linkage Disequilibrium
MCMC:	Markov Chain Monte Carlo
MDS:	Multi-Dimensional Scaling
MPS:	Massively Parallel Sequencing
MSY:	Male-specific region of the Y chromosome
mtDNA:	Mitochondrial DNA
NCBI:	National Center for Biotechnology Information
ng:	Nanogram
NGS:	Next-generation Sequencing
NRY:	Non-recombining portion of the Y chromosome
PAR:	Pseudoautosomal Region
PCA:	Principal Component Analysis
PCR:	Polymerase Chain Reaction
pSNP:	phenotypic-informative single nucleotide polymorphism
RFLP:	Restriction Fragment Length Polymorphism
RFU:	Relative Fluorescence Unit
SD:	Standard Deviation
SNP:	Single Nucleotide Polymorphism
SPT:	Skin phototyping
SRY:	Sex-determining Region, Y
STR:	Short-Tandem Repeat
SWG DAM:	Scientific Working Group on DNA Analysis Methods
T:	Thymine
TMRCA:	Time to most recent common ancestor

YHRD:	Y Chromosome Haplotype Reference Database
UAS:	Universal Analysis Software
VCF:	Variant Call Format, file format to store sequence variation

# Chapter 1: Introduction

This thesis explores the use of targeted Massively Parallel Sequencing (MPS) and genome-wide approaches to estimate human kinship, biogeographical ancestry and externally visible characteristics (EVCs) for forensic and genealogy applications. The main focus is on short tandem repeats (STRs) and single nucleotide polymorphisms (SNPs) on the autosomes and X-chromosome, for outbred indigenous as well as admixed populations, and for simulated inbred individuals.

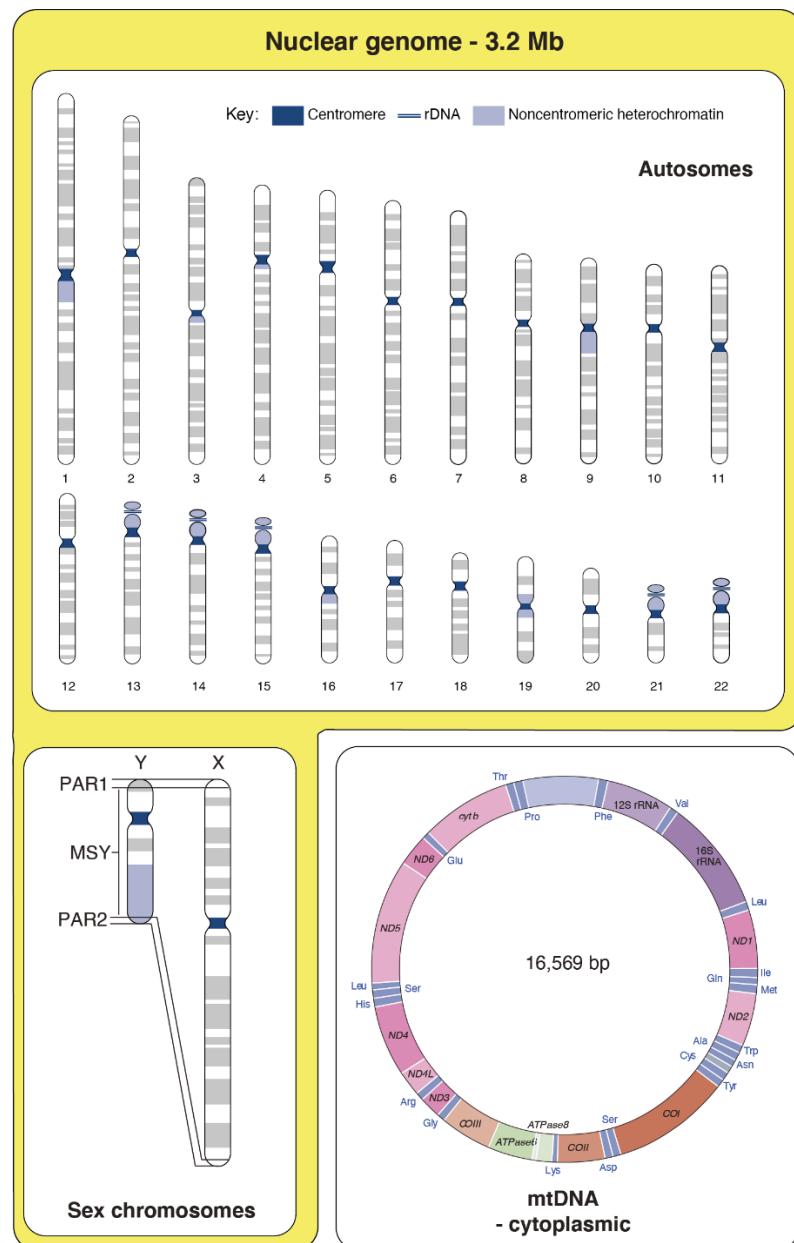
This Chapter introduces human genetic variation, how it is measured, and the general properties of markers that are used to study it, in particular short tandem repeats (STRs) and single nucleotide polymorphisms (SNPs). The forensic application of these markers in individual identification is explained, with a description of the technologies used to type them in practice.

When suspects cannot be identified from DNA database searches of crime-scene profiles, more creative approaches are taken to gather intelligence from DNA samples. These include kinship testing (finding relatives of the donor of the sample), prediction of externally visible characteristics (EVCs), and prediction of the population of origin of the donor (biogeographical ancestry; BGA). Relatedness at a population and individual level, and the genetic basis for EVCs and BGA are introduced, including how these can be useful forensic applications. Finally, the aims and objectives of this project are presented.

## 1.1 Patterns of inheritance of the human genome

Humans are diploid, and most cells contain two copies of the genome each comprising about 3.4 billion base pairs of DNA. The diploid genome is organized into 46 pairs of chromosomes: 22 pairs of autosomes and one pair of sex chromosomes (Figure 1.1). The autosomes are numbered from 1 to 22 and are shared between sexes, while the sex chromosomes differ, with females normally having two copies of the X-chromosome and males normally one X-chromosome and one Y-chromosome. DNA and genetic

information are also contained in the cell's mitochondria (more information in Section 1.3.3). A position on a chromosome is called a locus, and for an autosome carries a pair of alleles, the specific sequence information that defines the genotype at that locus. Each allele derives from one of the parents of the individual considered: if the pair of alleles is identical, the genotype is homozygous, otherwise it is heterozygous.

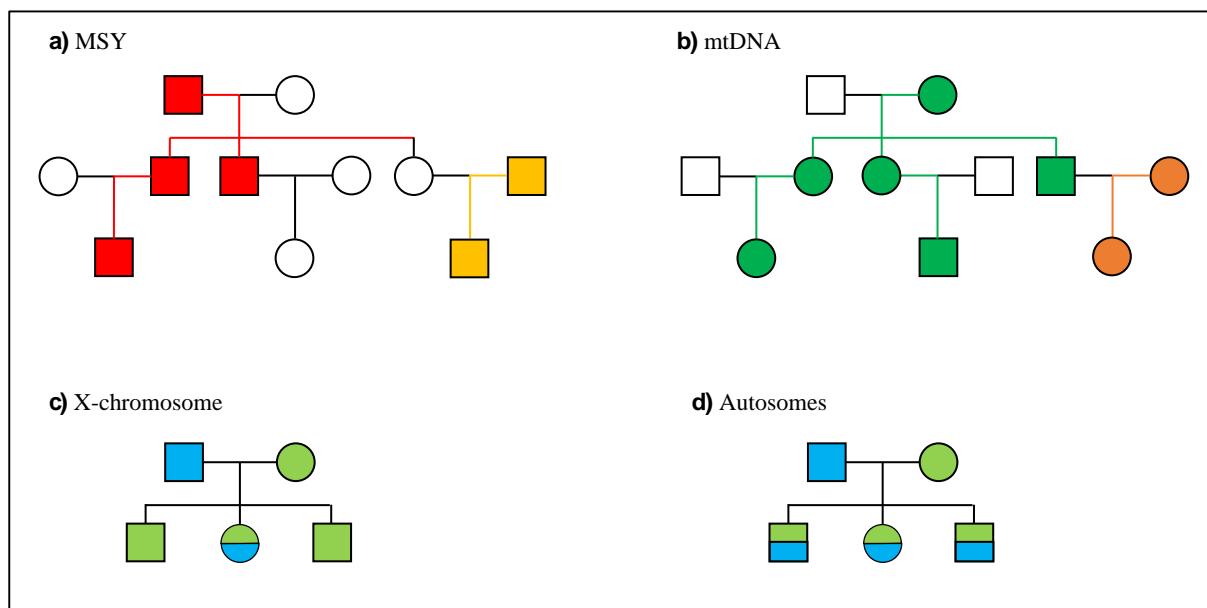


**Figure 1.1 The human genome.**

G-banded chromosomes forming the human autosomal karyogram, the sex chromosomes, and the structure of mtDNA. Sex chromosomes and mtDNA are not to scale; the X-chromosome is about the same length as chromosome 8. Figure adapted from Jobling et al. 2014.

The passage of genetic information from one generation to the next involves the production of haploid cells (gametes) through the process of meiosis. Males produce two types of gametes (X- and Y-bearing), while all female gametes carry an X-chromosome. Beyond this simple distinction, all gametes are genetically different due to the independent assortment of chromosomes, and to recombination, discussed further in the following section.

Autosomes are passed to offspring with equal contribution from both parents. In a male, the Y-chromosome is inherited from the father (inheritance through the paternal line) and X-chromosome from the mother; in a female one copy of the X-chromosome is inherited from each parent. Mitochondrial DNA (mtDNA) is inherited through the maternal line. Section 1.3. will focus on these atypical segments of the genome. The described inheritance paths are depicted in Figure 1.2.



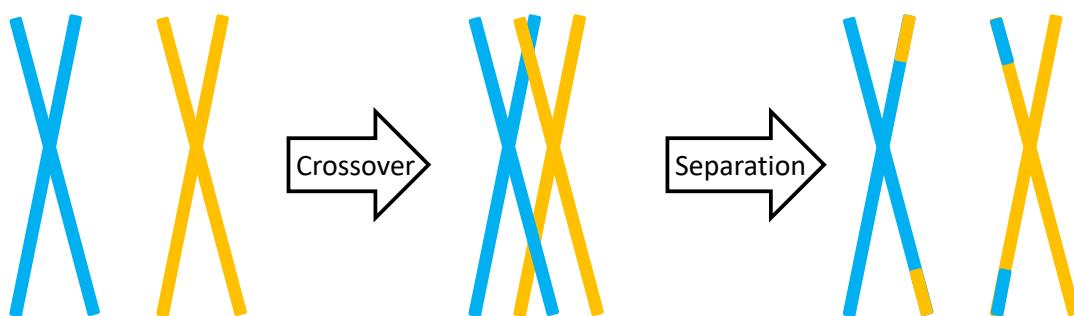
**Figure 1.2 Inheritance of (a) the male-specific region of the Y (MSY), (b) the mitochondrial DNA, (c) the X-chromosome and of (d) the autosomes.**

The colours indicate the inheritance of genetic material, squares represent males and circle represent females.

## 1.2 Autosomal recombination, linkage and Linkage Disequilibrium

Genetic recombination reshuffles DNA sequence between chromosomal homologues via crossing over (Figure 1.3) and is also essential for ensuring the proper segregation of chromosomes (Carroll 2001). On average, most chromosomes undergo one crossover event per chromosomal arm, and the further apart two markers are, the more likely they are to be separated by crossover. Genetic distance between variants can be estimated from the number of recombination events occurring between them (genetic mapping) and can be expressed in units of centiMorgans (cM), which relates to physical distance units as an average of ~1.1 cM/Mb (Kong et al. 2002).

The frequency of recombination events is not uniform by sex and across the genome (Popa et al. 2012). In fact, it is more frequent in females, at the ends of a chromosome (towards the telomeres) than towards the centromeres, and more frequent in the smallest chromosomes, likely due to variation in the initiation of double-strand breaks (DSBs) during meiotic recombination (International Human Genome Sequencing Consortium 2001). Some ‘recombination hotspots’ (generally segments of 1–2 kb) show more frequent recombination compared to adjacent larger regions (Kauppi et al. 2004). This has an impact on the distribution of sequence variants along chromosomes, on meiotic processes, and on the formation of new sequence combinations, and consequent phenotypic diversity.



**Figure 1.3 Chromosomal crossover.**

Replicated chromosomes are paired during meiosis (germ cell division in sexually-reproducing organisms) and segments are exchanged, resulting in recombinant chromosomes.

Recombination from one generation to the next needs to be taken into consideration in making deductions about kinship from genetic data. Linkage can be observed in a pedigree: a specific combination of variants along a chromosome (a haplotype) is passed to the next generation as a unit without being broken up by recombination. The property Linkage Disequilibrium (LD) can be observed at the population level as the non-random association between alleles at different loci.

This property describes alleles that are found together on the same chromosome more often than expected if the alleles were randomly segregating. This non-independence has been studied through the population-based genome-wide analysis of SNP haplotypes (International HapMap Consortium 2007; International HapMap Consortium and Donnelly 2005), and analysis of recombination events in pedigrees (Dib et al. 1996) and in sperm DNA (Jeffreys et al. 2001).

The distribution of LD in the human genome is not homogeneous, as blocks of haplotypes with high LD are alternated with regions of low LD, due to factors like crossover (including hotspots), gene conversion (the non-reciprocal transfer of variants between alleles) and natural selection. This non-homogeneity makes it possible that more shared variants are present in high LD regions than in low LD regions of the same size (Coleman et al. 2016). It is also influenced by population of origin: African populations have generally lower LD than non-Africans (International HapMap Consortium and Donnelly 2005) and phenomena such as genetic drift, admixture and inbreeding can influence the levels of LD in specific populations (Shifman et al. 2003).

**Measures of LD.** The extent of LD can be expressed by a number of measures, which describe the degree of association between a pair of markers. The measure D considers the difference between the observed frequency of two loci forming a haplotype and the expected frequency if the alleles were randomly segregating (Slatkin 2008):

$$D_{AB} = p_{AB} - p_A p_B$$

where  $p_{AB}$  is the frequency of the gametes carrying the pair of alleles A and B at two loci and  $p_A p_B$  is the product of the individual allele frequencies. When it is significantly different from zero, D indicates LD. A limit in the use of D is its dependence on allele frequencies, making comparisons of values not meaningful. An alternative is the absolute value of D, divided by its maximum possible value given the allele frequencies at the two

loci (Lewontin's  $D'$ , Lewontin 1964), which provides a measure normalised for allele frequencies. The maximum value of  $D'$  is 1, indicating complete LD; however, it is not clear how to interpret values lower than 1. Values of  $D'$  are affected by small sample sizes (which lead to overestimation) and by allele frequencies (low minor allele frequencies can lead to erroneous conclusion of LD).

Another measure,  $r^2$ , is the square of the correlation coefficient between two loci (considering biallelic loci,  $D^2$  is divided by the product of the four allele frequencies at the two loci). When this is equal to 1, it indicates perfect disequilibrium, with alleles that have not been separated by recombination. An advantage of the use of  $r^2$  is that it is more stable in cases of small sample size compared to  $D'$ . In this thesis,  $r^2$  is used for measuring LD.

## 1.3 The sex chromosomes and mitochondrial DNA

The mammalian sex chromosomes evolved from a pair of homologous autosomes (Graves 2006). Divergence initiated ~300 MYA, when one of them (the proto-Y chromosome) acquired a male sex-determining function. This was followed by repression of recombination, and subsequent degeneration of the Y chromosome, involving sequence and gene loss.

### 1.3.1 X chromosome

The X chromosome is about 150 Mb in size and contains around 1000 genes. It has a number of unusual properties. Along most of its length, it only crosses over in female meiosis, and therefore displays relatively high LD. It undergoes X-inactivation, such that one of the two X chromosomes in a normal female is somatically silenced, leading to an approximate equivalence of gene dose between males (46, XY) and females (46, XX). It also has a low effective population size ( $\frac{3}{4}$  that of autosomes), leading to relatively low genetic diversity in populations.

### 1.3.2 Y chromosome

The Y chromosome is haploid and regarded as the smallest chromosome in the genome (60 Mb). Genetically, it is divided into three different regions: Pseudoautosomal Region 1 (PAR1; 2.6 Mb, Brown 1988) and Pseudoautosomal Region 2 (PAR2; 330 kb, Kvaløy

et al. 1994), which undergo crossing over with the X chromosome in male meiosis, and the intervening male-specific region of the Y chromosome (MSY).

The MSY is gene-poor (the 23 Mb of MSY euchromatin contains 78 genes, coding for only 27 independent proteins) (Skaletsky et al. 2003), and enriched in repeated elements (SINEs, segmental duplications and endogenous retroviral sequences). It also contains large “ampliconic” segments comprising megabase-scale direct and inverted repeat units, which underlie dynamic intrachromosomal recombination processes, largely gene conversion (Trombetta and Cruciani 2017).

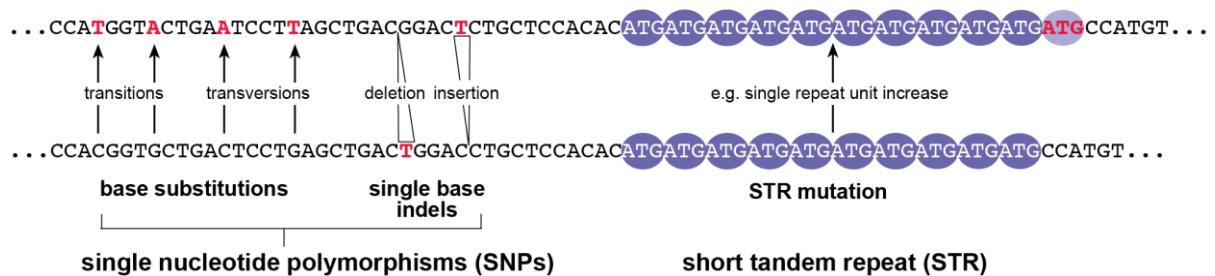
MSY is a useful marker in population genetics for describing male demographic histories and in forensic identification, due to its male-specific inheritance (Jobling and Tyler-Smith 2003, 2017a). Because of its haploidy, MSY has an effective population size one quarter that of autosomes, and this leads to a high degree of genetic drift (stochastic changes in haplotype frequencies), and relatively high geographical differentiation.

### **1.3.3 Mitochondrial DNA**

The mitochondrial DNA (mtDNA) is a circular double-stranded DNA molecule of about 16.5 kb, contained within the cytoplasmic organelles called mitochondria. Its full sequence is known (Anderson et al. 1981; Andrews et al. 1999). Average mutation rate in mtDNA is almost ten times higher than the nuclear genome, due to several factors such as the higher replication rate with longer phase in single-stranded form compared to nuclear DNA, the absence of histones that may reduce the vulnerability to mutation, and the high concentration of mutagenic oxygen free radicals linked to its function in energy generation (Jobling et al. 2014). Thanks to its maternal inheritance and high mutation rate, mtDNA is widely used in population genetics as it highly discriminates and allows investigation of female specific histories. Like the MSY, it is haploid and so has a low effective population size, which increases the geographical differentiation of haplotypes.

## 1.4 Variation in the human genome and its use in forensic genetics

Variation in the human genome exists over a very wide range of scales, from single nucleotide changes to structural variants encompassing many millions of base pairs. In forensic analysis the focus over the past three decades has been on two classes of variants in particular, SNPs and STRs.



**Figure 1.4** Single nucleotide polymorphisms (SNPs) and a short tandem repeat (STR) shown on two haplotypes.

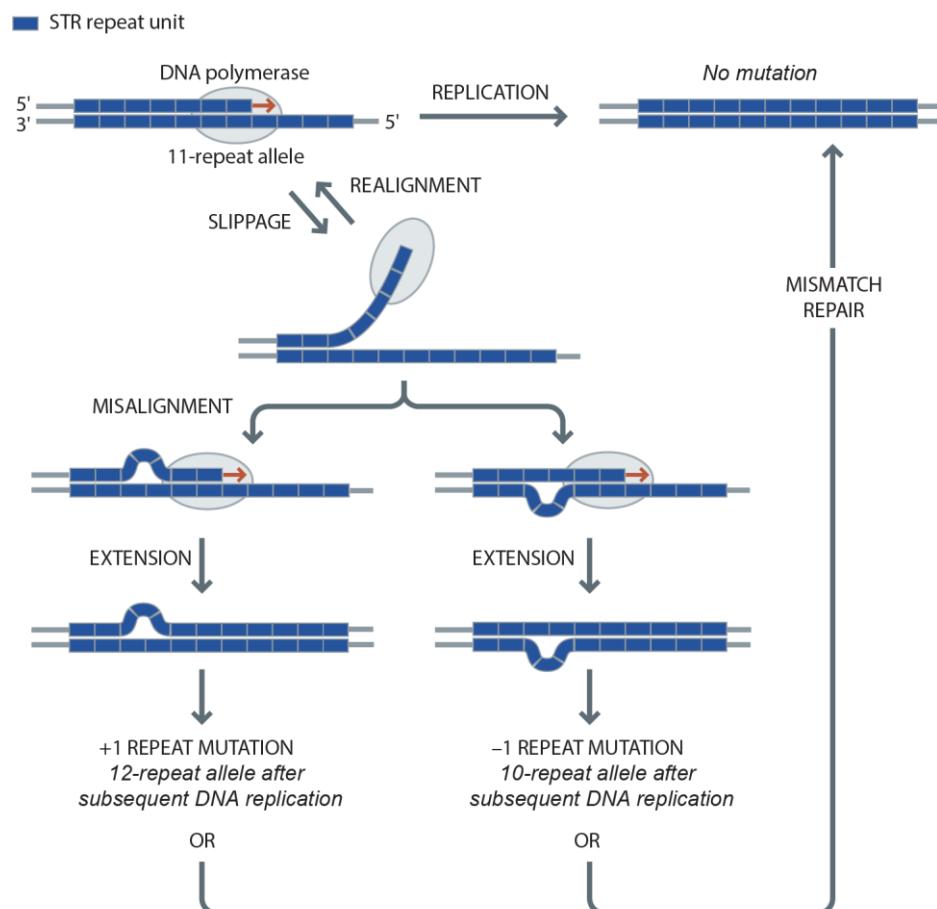
Transition, transversion, and indel SNPs are indicated. The STR is a trinucleotide (ATG) repeat, and a stepwise mutation is indicated between alleles. Adapted from Jobling et al. 2014.

#### 1.4.1 Short Tandem Repeats (STRs)

Short Tandem Repeats are sequences of DNA comprising a variable number of adjacent repeating units (typically up to 30) of 2–6 nucleotides (Figure 1.4) (Fungtammasan et al. 2015; Willems et al. 2014). These repeats comprise about 3% of the human genome (Fungtammasan et al. 2015). Many are non-polymorphic; however, a highly polymorphic subset of ~10-20 multiallelic STRs, mostly tetranucleotide repeats, has become established in forensics. Most STRs are regarded as selectively neutral, but some are in genes or regulatory regions, and are linked to disease and complex traits (Fungtammasan et al. 2015; Willems et al. 2014).

STRs show high mutation rates, likely due to the frequency of slippage events in repetitive sequences during DNA replication (Schlötterer and Tautz 1992). The rate (as measured in pedigrees) for polymorphic STRs ranges from  $10^{-4}$  to  $10^{-2}$  per STR per generation (Fan and Chu 2007). The simplest and commonly assumed mutation model is the stepwise

mutation model (Figure 1.5): in most events a single complete repeat is added or lost, with rarer events involving two or more repeats. Large-scale studies (Sun et al. 2012) show that there is a slight bias to repeat gains over losses, and that mutation rate depends on repeat array length, such that longer arrays have higher rates. There is a tendency to reach an equilibrium distribution of allele lengths, where shorter arrays may grow stepwise and longer alleles will tend to contract via deletions.



**Figure 1.5 Stepwise mutation model for STRs.**

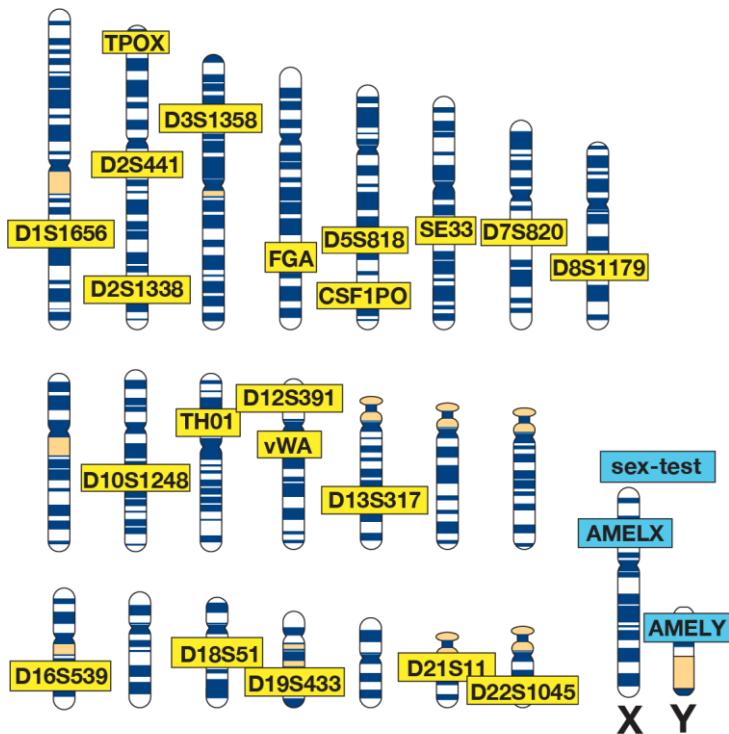
Source: Jobling et al. 2014.

The back-and-forth mutation processes at STRs mean that there are multiple mutational pathways to any given allele; therefore, the same STR alleles in two different genomes are not due to the same common ancestor (identity by state), this will be described in greater detail in Section 1.5.1.2.

#### **1.4.1.1 Applying STR analysis in forensic genetics**

Autosomal STRs are widely used in the forensic field as they are highly polymorphic and informative (Churchill et al. 2016), and can be amplified by PCR on short (~200-450 bp) amplicons, allowing the analysis of degraded DNA. The most common detection method is measurement of fragment lengths following size separation by capillary electrophoresis (CE) (see section 1.8.2.1). Combining several STRs in multiplexes provides genotypes that are individually identifying: discriminatory power is assessed by calculating the random match probability (RMP), which is the probability that the DNA profile from a random sample from the population has the same profile as the queried sample. In order to evaluate the rarity of the DNA profile, allele frequency information is needed for an appropriate reference population, and the product rule is used (i.e. the locus frequency estimates are multiplied). The product rule assumes Hardy-Weinberg and linkage equilibrium, and yields RMPs as low as  $10^{-26}$  for Globalfiler (Life Technologies 2012), a commonly used STR kit. STR genotypes are easy to store and query in databases, which led to the development of large investigative databases such as the UK's National DNA Database (Werrett 1997).

Specific STRs have been selected to form a core set routinely tested, for example the 13 markers selected in the 1990s to be part of the Combined DNA Index System (CODIS) in the USA (Edge et al. 2017). The number of autosomal forensic STRs commonly used in commercial kits has increased over time, reducing RMP values; a current example is GlobalFiler [Applied Biosystems], containing STRs that are shown in Figure 1.6. In order to prevent biased and unreasonable searches, these markers are not linked (or very weakly linked) to individual information such as phenotype or biogeographical ancestry.



**Figure 1.6 Autosomal STRs within the GlobalFiler kit.**

Loci are labelled on the human G-banded karyogram in their approximate locations. The kit also analyses the amelogenin sex-test loci (shown in blue), and two additional Y-chromosomal markers (not shown).

Due to their high degree of polymorphism among human populations, autosomal STR genotypes offer high discriminative power in kinship analysis too. Paternity testing generally adopts the same markers used for individual identification (Butler 2006), while for other close relationships (i.e. sibship) additional STR markers are used (Mo et al. 2018). However, for more distant relationships, the currently available STRs are not enough (Amorim and Pereira 2005a; Mo et al. 2018; Schneider 2012a).

#### 1.4.2 Single Nucleotide Polymorphisms (SNPs)

About 85% of human genetic variation is due to SNPs (Budowle et al. 2004) which are an alteration of a single nucleotide. Most commonly analysed SNPs are observed to be biallelic, and the minor allele frequency (MAF) is an important property, describing the frequency at which the less common allele is found within a given population.

Both base substitutions and single base insertions and deletions (indels) are considered SNPs, although these types of variants arise through different mechanisms that have different rates (Jobling et al. 2014). Base substitutions arise through the mis-incorporation of nucleotides during DNA replication, occurring with a frequency of  $10^{-9}$ - $10^{-11}$  per nucleotide per replication event, or mutagenesis due to chemical modification or physical damage. Point mutations can also occur during repair synthesis that follows DNA damage. Chemical damage includes endogenous chemical processes (e.g. deamination of cytosines into uracils, or oxidation, methylation and depurination) and the action of chemical mutagens (e.g. base analogs, base-modifying agents, intercalating agents, and cross-linking agents). There are two types of base exchange: transitions (pyrimidine to pyrimidine or purine to purine changes) and transversions (pyrimidine to purine, or vice versa, see Figure 1.4).

The current (2021) release of the dbSNP database (build 135) of the National Center for Biotechnology Information (NCBI) for short genetic variations contains 52,345,594 validated SNPs, each of which is prefixed with the letters rs, for RefSNP. There are very many more submitted SNPs yet to be validated (ssSNPs; total 178,140,838). These SNPs were discovered through a wide variety of sequence-based approaches, and currently found efficiently via whole-genome sequencing. Each sequenced genome contains around 4-5 million SNPs compared to the reference sequence (The 1000 Genomes Project Consortium 2015). From these sequences the average nucleotide diversity ( $\pi$ , representing the likelihood that a given nucleotide position differs across two randomly chosen sequences) can be estimated: this value is around  $7.6 \times 10^{-4}$  (one SNP expected per 1250 bp) in European populations, and  $9.9 \times 10^{-4}$  (one SNP per 1006 bp) in African populations.

The measurement of nuclear base substitution rates has been facilitated by the sequencing of whole genomes in human pedigrees via high-throughput DNA sequencing technologies, offering a direct estimate of  $1.2 \times 10^{-8}$  per base per generation (about 10 times more frequent than indels, depending on the locus) (Kong et al. 2012). This low rate means that the presence in two genomes of the same base at a SNP can be assumed to be due to inheritance from the same common ancestor; this is identity by descent (instead of identity by state), and will be described in greater detail in Section 1.5.1.2.

While most SNPs can be regarded as selectively neutral, some play a role in Mendelian diseases, when a single variant in a gene can cause a phenotype, as well as in common and complex traits, when an effect is obtained by the interaction of multiple SNPs and the environment (e.g. Alzheimer’s disease, cancer, type 2 diabetes, coronary artery disease) (Mu and Zhang 2013). Particular SNPs are also involved in “normal” phenotypes not linked to diseases (e.g. height, body mass index, externally visible characteristics). Some of these, used in forensics for the prediction of externally visible characteristics, will be discussed further in Chapter 5.

Many different methods exist to type SNPs; this thesis uses sequencing-based methods, and also data from genome-wide SNP microarrays (“SNP chips”) that analyse many thousands of SNPs simultaneously. These will be described in Section 1.4.3.4.

#### **1.4.2.1 Applying SNP analysis in forensic genetics**

SNPs can provide useful information on identity, ancestry, phenotype, and pharmacogenetics, suitable for investigative and information gathering purposes (Churchill et al. 2016). Forensically useful SNP categories as proposed by Budowle and Van Daal (2018) are summarised in Table 1.1. Since they are single-base variants, it is possible to type SNPs in short amplicons, as small as 50 bp, which is advantageous in cases where the DNA is degraded (Budowle and Van Daal 2018).

Because SNPs are generally biallelic, a larger number of SNPs (>50) must be typed for obtaining sufficiently low RMPs compared to multiallelic STRs (Gill 2001; Mo et al. 2018).

**Table 1.1 Categories of SNPs useful in forensic contexts.**

Source: Budowle and Van Daal 2008.

Class	Description	Application
identity-testing	high heterozygosity and low population heterogeneity, low linkage disequilibrium, and PCR design compatibility	individualization
lineage informative	tightly linked SNPs that function as haplotype markers	identification of missing individuals, kinship analyses
ancestry informative	requiring low heterozygosity and high population differentiation	establishing high probability of biogeographical ancestry, inferring phenotypic characteristics for investigative lead value
phenotype informative	variants that affect function of specific genes involved in externally visible phenotypes	establishing high probability of phenotypic characteristics for investigative lead value (skin color, hair color, eye color)
pharmacogenetic	variants that affect function of specific genes involved in drug metabolism	cause of death determination

Identity-testing SNPs may have broad applications, such as analysis of highly degraded samples from human remains (i.e. bones, teeth, and hair), and several researchers have proposed SNP panels: Dixon and co-workers in 2005 (21-SNP panel) (Dixon et al. 2005), Sanchez and co-workers in 2006 (52-SNP panel) (Sanchez et al. 2006), and Kidd et al. in 2006 (>40 SNPs) (Budowle and Van Daal 2018; Kidd et al. 2006).

Many authors have suggested a combined use of STRs and SNPs (Amorim and Pereira 2005b; Børsting and Morling 2011; Kling et al. 2012b; Mo et al. 2018; Schneider 2012b; Silvia et al. 2017), and the possibility of using Massively Parallel Sequencing (MPS) technology would overcome several obstacles here, including the possibility of processing both SNPs and STRs in parallel (Silvia et al. 2017). This will be discussed in the following sections.

Although SNP analysis is sometimes applied in real cases of kinship testing, systematic assessments of their performance are rare compared to studies focused on STRs (Cho et al. 2017), and databases able to provide pedigree information are rare or non-existent. Examples of SNP systems are SNPforID multiplex assay (52 markers) (Sanchez et al. 2006) and the IISNP (92 markers) (Kidd et al. 2006; Pakstis et al. 2010), and a 273-SNP system (designed to analyse the Chinese Han population)(Zhang et al. 2017).

### **1.4.3 Methods for typing forensic STRs and SNPs**

The possibility of simultaneously typing several genetic markers in one PCR analysis and in an automatic way (multiplex assays) efficiently exploits DNA evidence that may be limited in availability (small amounts, degraded), reduces contamination opportunities, and reduces sample utilisation to enable re-testing (Budowle and Van Daal 2018).

#### **1.4.3.1 STR analysis via capillary electrophoresis (CE)**

The method traditionally used in forensic identification of individuals is based on Capillary Electrophoresis (CE), used for separating and detecting STRs following PCR amplification on the basis of fragment length. CE systems involve three main steps: injection, separation and detection. Denatured DNA fragments move through a capillary filled with polymer (i.e. polydimethyl acrylamide) due to the applied voltage. The separation of fragments is determined by size (i.e. smaller negatively charged fragments travel faster than the larger ones), and their differentiation and detection is possible thanks to the capture of the light signal due to the fluorescent dyes attached to one PCR primer from each locus-specific pair.

Standard kits normally contain 15-23 autosomal STRs. For sexual assault cases, Y-STRs may be included to solve DNA mixtures, and X-STRs may help in some kinship cases (Li et al. 2019). Applied Biosystems has offered various capillary array systems, from the ABI 310 single-capillary system based on four fluorescent dye colours to the ABI 3500 multi-capillary system based on 6-dye detection (Butler 2012).

These CE systems are highly sensitive (~1-2.5 ng of DNA sample, Butler 2005) and can be high throughput, and are used widely. One disadvantage of this technology is the technical constraint due to the limited number of loci (and of dyes) that can be used (e.g. commonly, target amplicons range between 80 and 500 bp, as in GlobalFiler™ Express Life Technologies 2012; PowerPlex® ESI 17 Pro System, Promega 2017; Investigator® 24plex QS, QIAgen 2016), which may have reached its maximum with the Spectrum CE system (24 autosomal STRs and 11 Y-STRs). This limit in the numbers of dyes and, therefore, of loci that can be simultaneously detected can be overcome by massively parallel sequencing technology (MPS, introduced in the next section). Moreover, CE considers only the allele length of an STR. The internal sequence information, including

indels, SNPs in a repeat unit and variation in the flanking region may be identified via MPS.

#### **1.4.3.2 Analysis of STRs and SNPs via Massively Parallel Sequencing (MPS)**

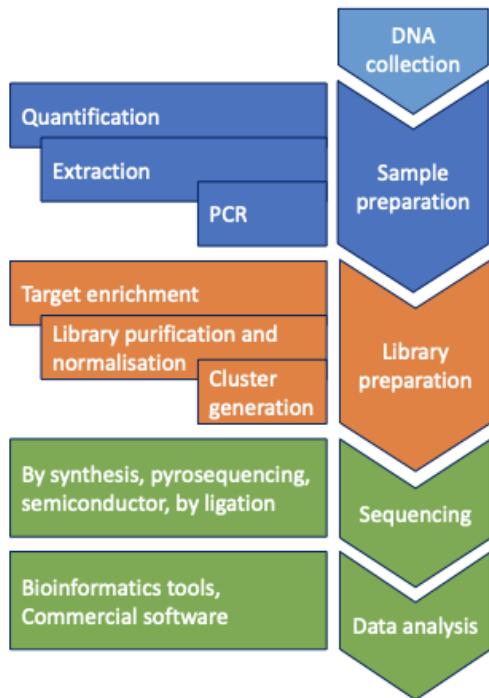
MPS technology offers the possibility of obtaining large amounts of genome sequence data quickly and cheaply. Medical genetics, including clinical applications (e.g. genetic diagnoses, personalized medicine) and population genetics have greatly benefited from MPS, generating whole genome or whole exome sequences for less than \$1000 (<https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost>). In forensics, it has generally been applied to the targeted sequencing of PCR amplicons, where its characteristics offer an increased discrimination power, making it advantageous for some complex cases, including kinship determination and mixture deconvolution. Also, the low amount of DNA required makes it possible to use MPS in cases with limited or degraded DNA material (Bruijns and Tiggelaar 2018). Sequencing STRs may reveal additional polymorphisms; these can involve flanking DNA, or variants within the repeat array. Such repeat pattern variation is locus specific and depends on the internal repeat structures (Table 1.2). The inclusion of specific SNPs may also offer information on ancestry, paternity or phenotype, and mitochondrial DNA can also be sequenced (Bruijns and Tiggelaar 2018).

**Table 1.2 Repeat structure variation within 13 STR markers commonly used in forensic analysis.**

This table shows the general characteristics of the repeat structure of the markers, referring to the CODIS database. Source: Butler 2015, adapted from Urquhart et al. 1994. More complex structures tend to reveal additional repeat pattern variants on sequencing.

Category	Description	Repeat structure (example)	CODIS locus
Simple repeats	units are of identical length and sequence	GATA-GATA-GATA	TPOX, CSF1PO, D5S818, D13S317, D16S539
Simple repeats with non-consensus alleles	non-consensus alleles between repeats of identical length and sequence	GATA- <b>GAT</b> -GATA	TH01, D18S51, D7S820
Compound repeats	two or more adjacent simple repeats	GATA-GATA- <b>GAC</b> <b>A</b>	vWA, FGA, D3S1358, D8S1179
Complex repeats	units of several repeat block of variable length	GATA- <b>GAC</b> <b>A</b> - <b>CA</b> - <b>CATA</b>	D21S11

MPS methods, as the name suggests, allow many sequencing reactions to be carried out in parallel. There are different types of sequencers, of which the most suitable for forensic applications are Verogen's MiSeq FGx, ThermoFisher's Ion Torrent PGM, and Ion S5. Most MPS technologies share the same workflow overall: genomic DNA preparation, "library" preparation, amplification, sequencing and data analysis (Figure 1.7).



**Figure 1.7 Massively parallel sequencing (MPS) workflow.**

This chart shows the main steps in a generic forensic MPS application.

In targeted approaches, the generated DNA fragments are used to create a library: adapters are ligated to the fragments through one or two PCR reactions. This step adds identifying sequences for clonal amplification of the library as well as unique barcodes that label individual samples within a set that can be simultaneously sequenced. Given that per-nucleotide error rates are higher than that of Sanger sequencing, read-depth (also known as coverage – the mean number of times a nucleotide is covered by a sequence read) is an important factor. As coverage increases, so does confidence in the determined sequence.

A number of different technologies have been developed, the most dominant of which is currently sequencing by synthesis (Illumina), which is used to analyse amplicons within Verogen's ForenSeq kit, and is utilized in this project. In the Illumina method, the pool of library molecules is hybridized via adapter oligonucleotides to a lawn of complementary primers on the surface of a flow-cell, and then each molecule is amplified by process called bridge amplification to form a cluster of thousands of identical molecules. It is in these clusters that sequencing occurs via the incorporation of reversible fluorescent-labelled dye terminators and laser-induced detection. The maximum number

of bases sequenced in each cluster is the read length, and this varies with each technology, and in Illumina sequencing, with the particular platform used for Illumina MiSeq (and Verogen) technology it is ~150 bp.

The possibility of capturing sequence variation is a great advantage that may have an impact in many applications in forensics by increasing the number of detectable alleles (Gettings et al. 2015). For STRs, there is no longer a need to space the lengths of amplicons to allow them to be separated, and no limitation of the CE-based number of fluorescent channels, since each amplicon is detected via its sequence, regardless of length. There is also clearer understanding of stutter ratios: because the stutter ratio for an allele is linked to the specific sequence of that allele, the sequence data may be used to predict stutter patterns (Ballard et al. 2020). Amplicons can be as short as possible, allowing the analysis of degraded samples (Gettings et al. 2015), and the combination of SNPs and STRs. In principle, hundreds of markers can be co-amplified in a single reaction, reducing the amount of template DNA needed for analysis.

The forensic field has been slow in adapting to MPS technology in casework (Just et al. 2017), due to the fact that existing systems are so well established, and that new ones require investment, training and validation before acceptance. MPS kits for forensic applications have been developed by three companies and include the CODIS loci (Almalki et al. 2017). The PowerSeq Auto System (Promega, Madison WI) includes 22 autosomal STRs, one Y-STR and Amelogenin and runs on the Illumina MiSeq platform (Zeng et al. 2015), the ForenSeq DNA Signature Prep Kit (Verogen, San Diego, CA) includes two primer mixes, and it will be further described in Chapter 3. The Precision ID (Thermo Fisher Scientific, Waltham, MA) runs on the Ion PGM and S5 Systems and includes three panels (Ancestry and Identity SNP panels, and GlobalFiler targeting 33 STR markers (Almalki et al. 2017).

Population-based studies are necessary to fully understand and exploit the new information captured via MPS, establishing allele frequencies (Gettings et al. 2018a; Novroski et al. 2016). There is as yet no consensus for thresholds, coverage, and nomenclature standards (Ballard et al. 2020; Gettings et al. 2018a; Phillips et al. 2018a).

#### **1.4.3.3 SNP arrays**

Chip-based microarrays for the simultaneous typing of hundreds of thousands of pre-selected SNPs were developed to support medical genetics projects, and have powered a large number of genome-wide association studies detecting associations between many complex diseases and specific regions of the genome.

Most chip designs work by annealing of the test DNA to an oligonucleotide array, followed by single-base primer extension and fluorescent detection. For diploid (autosomal) DNA, homozygous SNPs produce single-colour fluorescence, while heterozygous SNPs should produce approximately equal ratios of two colours, allowing the genotype at each SNP to be called. Because of the need to scan the whole genome effectively and cheaply, SNPs were chosen on a “tag SNP” design, in which haplotype blocks identified from the International HapMap Project could be tagged with just a few SNPs (International HapMap Consortium 2007), since all SNPs within a block are correlated via linkage disequilibrium. As well as tag SNPs on the autosomes and the X-chromosome, many SNP chips contain mtDNA and MSY SNPs.

SNP chips were designed to be used with plentiful amounts of DNA of good quality and would therefore be expected to perform poorly in forensically relevant samples. However, some samples have yielded useful data which have been used to query large reference SNP chip databases and thereby find putative relatives. This is described further below.

## **1.5 Genetic relatedness**

This section will describe how DNA information can be used to investigate kinship, and the theory behind kin relatedness in humans. The study of relatedness is relevant in many fields: it is not only of importance for many aspects of life (as in marriage and inheritance laws), but also in agriculture (additive and dominance components of variance estimation), animal breeding programmes, human genetics in mapping disease genes and linkage studies, heritability estimation, evolutionary genetics, in ecology for conservation programmes, in genealogy inference and in forensic science (Conomos et al. 2016; Powell et al. 2010; Shem-Tov and Halperin 2014; Wang 2002; Weir et al. 2006). Pedigree information, which would allow us to easily estimate relationships, is often absent

(especially in fieldwork for natural populations), and the pairwise relatedness between individuals can be estimated only by using molecular markers (Lynch and Ritland 1999; Wang 2002).

A broad definition of relatedness is linked to the average genetic distance, which reflects the most recent time in the past that DNA was present in a single ancestor (Lawson and Falush 2012; McVean 2009). Within the framework of a known pedigree, degrees of relatedness can be defined as the number of meioses between two individuals within the pedigree, and this normally involves a short time to a most recent shared ancestor. Individuals from a population who are not within the same pedigree may also have shared ancestors, but with greater time-depth. Therefore, it is possible to distinguish between “relatedness”, which refers to genetic sharing and can be described by parameters, and “relationship”, which is specific to a pedigree connection between individuals and is described by genealogical relationships (i.e. parent-offspring, siblings etc.).

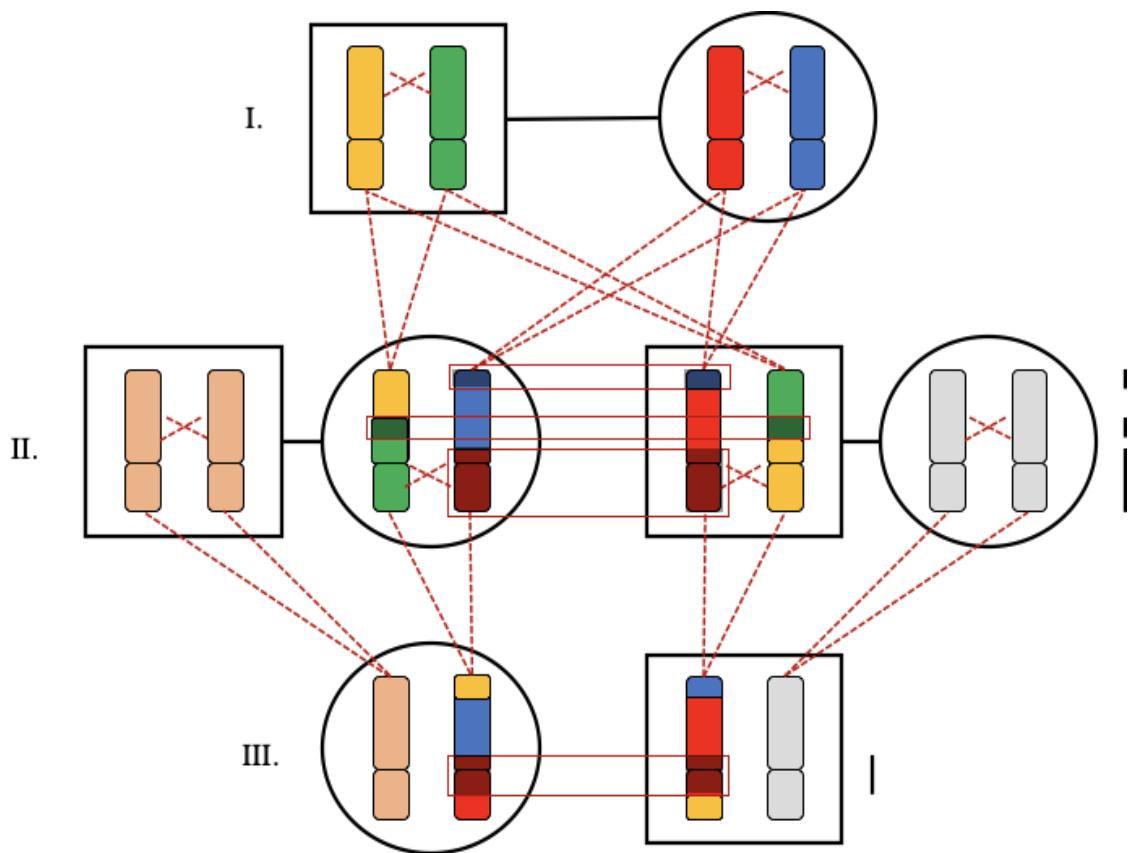
A pedigree is considered as a set of related individuals together with a full specification of all the relationships among them (Thompson 1985, 2000). Individuals in a pedigree can be divided into founders (persons whose parents are not in the pedigree) and non-founders (descendants of the founders) (Sanchez et al. 2008); (Cheung 2013). In a pedigree, according to Mendel’s First Law, during the process of transmission of a DNA copy (meiosis), each parent passes one of two alleles to an offspring, with each allele having the same chance to be transmitted (Cheung 2013). Pedigrees can have graphical representation: squares represent male individuals, circles represent female individuals, horizontal lines indicate marriages and offspring are connected to their parents via vertical lines. Generations are denoted by Roman numerals and include individuals in the same level in the pedigree.

In a finite population, any two individuals show a certain degree of relatedness, as they share a common ancestor at some point in the past. This means that there is a “background relatedness” in the population in addition to more recent family relatedness. However, this is generally low for human populations, even if effects can be seen in linkage studies, such as a reduction of heterozygosity compared to expectations, often depending on population of origin and its customs, e.g. with regard to mate-choice (Bittles and Black 2010). The background relatedness can be much higher in non-human populations and,

for example in conservation biology, may affect the fitness of offspring if the relatedness between potential mates is ignored and, thus, underestimated.

### 1.5.1 Identity by State (IBS) and Identity by Descent (IBD)

IBS is defined as the proportion of shared alleles at the same loci between two independent genomes. IBD suggests that the shared alleles have been inherited from a common ancestor (Figure 1.8). The average proportion of genome shared IBD between two individuals separated by  $m$  meioses is  $2^{-(m-1)}$ , while the length of IBD segments has a mean of  $100 m^{-1}$  (approximately exponential distribution) (Browning 2012).



**Figure 1.8 Passage of IBD segments through generations.**

Starting from parents at generation I, the two homologous autosomal chromosomes undergo meiosis and crossover before a recombinant copy is inherited by the offspring (II). Three of the parental segments (in the same colour for both individuals and highlighted in the figure) are entirely inherited without recombination: these segments are shared IBD between the two siblings. The offspring of the two siblings (generation III) inherit an unrelated chromosome and a recombinant chromosome which shows segments traceable to their grandparents (generation I). There is just one segment (in red) shared IBD between these first cousins (generation III); they may share approximately 12.5% of their genomes IBD.

Inbreeding is said to occur when individuals are more related than expected if they were coming from a randomly-mating population (i.e. if parents are related) (Thompson 2000). The degree of inbreeding is measured by the inbreeding coefficient ( $f$ ), which is the probability of two alleles, from two different individuals, being IBD. When the two alleles are IBD, an individual is defined autozygous at a locus (Thompson 2000). Due to the higher chance of passing the same deleterious recessive allele to the next generation, this can cause inbreeding depression. Incest represents a case of extreme inbreeding involving close relatives; this topic is covered in the following section.

### 1.5.1.1 Coefficients of relatedness

The kinship (or pedigree) coefficient ( $\phi$ ) is the probability that an allele randomly selected in one individual is IBD with a random allele at the same genomic position in another individual. It is possible to distinguish between the expected (pedigree) kinship coefficient, which is the probability that a pair of alleles at a locus in two individuals are IBD, and the realised kinship, which is the actual proportion of genome shared IBD (Dou et al. 2017). The realised kinship differs from expected pedigree kinship due to stochastic behaviour of recombination and segregation; therefore, estimates of realised kinship may be more realistic than those of pedigree kinship, and dense SNP genotypes may give the necessary accuracy (Wang et al. 2017). The statistics estimated from the data are usually compared to the theoretical values (Waples et al. 2019).

Outbred pairwise relationships can be described by IBD coefficients (Table 1.3) introduced by Cotterman (Cotterman 1941). IBD coefficients  $k_0$ ,  $k_1$  and  $k_2$  represent the probability that two individuals share 0, 1 or 2 alleles identical by descent (these probabilities will be also defined as IBD0, IBD1 and IBD2 in this thesis). These coefficients are usually estimated from patterns of observed IBS, as IBD status cannot be directly observed: alleles are shared IBD not only because of recent common ancestry, but also because of more ancient common ancestry, or recurrent mutation and gene conversion (Waples et al. 2019). For example, if two individuals have genotypes AA and BB, it is possible to say that they show no alleles IBS or IBD at this locus (IBS0), while with AA and AB genotypes, they share one allele in common (IBS1), which may or may not be inherited from a recent common ancestor (IBD1) (Stevens et al. 2011). When two individuals share both alleles from a common ancestor, they are IBD2. When deriving

IBD from IBS, mutation rates and recombination events should be taken into account (Tiret and Hospital 2017).

The kinship coefficient ( $\phi$ ) can be easily derived from IBD coefficients, or probabilities: if  $k0_{i,j}$ ,  $k1_{i,j}$ ,  $k2_{i,j}$  define the proportions of genome shared IBD by two individuals i and j, then the kinship coefficient is

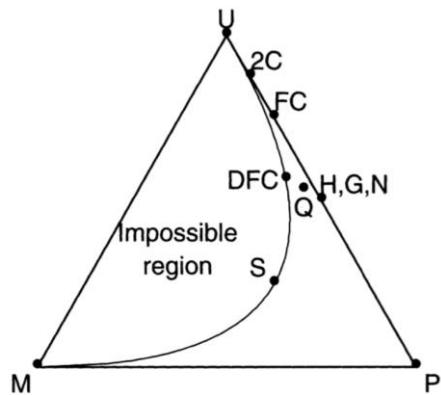
$$\phi_{i,j} = \frac{k1_{i,j}}{4} + \frac{k2_{i,j}}{2} .$$

**Table 1.3 IBD k coefficients for various outbred relationships.**

Adapted from Thompson 2000.

Pedigree relationship	<b>k0, P(IBD=0)</b>	<b>k1, P(IBD=1)</b>	<b>k2, P(IBD=2)</b>
Parent-offspring	0	1	0
Full sibling	0.25	0.5	0.25
Half sibling, avuncular, grandparental	0.5	0.5	0
First cousin, great-grandparents, great avuncular, half avuncular	0.75	0.25	0
Double first cousin	0.5625	0.375	0.0625
Quadruple half first cousin	0.5312	0.4375	0.0312
Distantly related	varies	varies	0
Unrelated	1	0	0

The relationships expressed by these coefficients can be visually represented: Thompson (2000) proposed a triangle with vertices corresponding to unrelated pairs of individuals (where  $k0=1$ ), parent-offspring pairs (where  $k1=1$ ) and identity or monozygotic twins ( $k2=1$ ) (Figure 1.9). In this Thesis, a scatterplot of IBD0 against either IBD2 or against the coefficient of relationship is used, representing the IBD triangle (see, for example, Figure 1.12 b). It is important to highlight that each pedigree relationship is determined by a point in the triangle, however the same  $k$  probabilities may define different pedigree relationships. For example, grandparent-grandchild, half-sibs, and aunt-niece all have  $k = (1/2, 1/2, 0)$ .



**Figure 1.9 The IBD triangle for outbred relationships.**

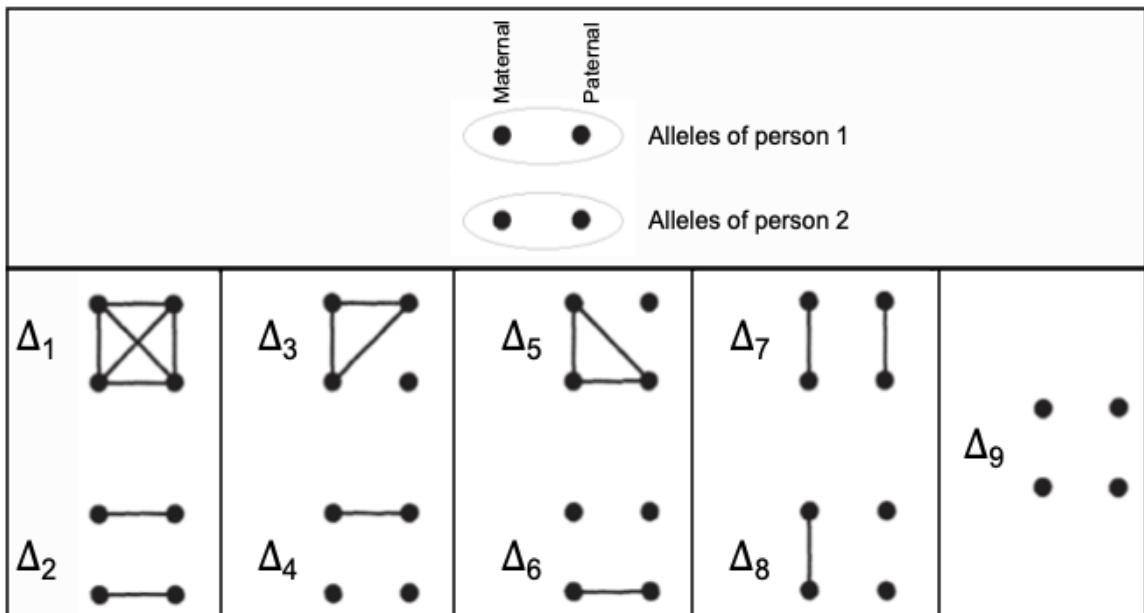
IBD values are plotted against each other causing relationships with the same IBD values to cluster together in specific points of the plot. These specific points are labelled: M = monozygotic twins, P = parent-offspring, S = full siblings, H,G,N = half-siblings/grandparental/aunt-niece, DFC = double first cousin, Q = quadruple-half-first-cousins, FC = first cousin, 2C = second cousin, U = unrelated. Source: Thompson 2000.

There are no standard definitions of IBD probabilities, and these depend heavily on calculation methods and assumptions. Powell and co-workers (Powell et al. 2010) highlight how the definition of “unrelated” linked to IBD0 sharing is a contradiction as, at a certain point in history, all individuals share close population inheritance. This may lead to the loss of important information and inaccurate estimates of genome-wide inbreeding and of genetic variance. In this way, the IBD concept appears to contrast with the idea that all alleles can be traced back to a common ancestor at some point in the past (coalescence theory). Applying IBD to the analysis of recent common ancestors, and coalescence theory to that of distant ones may be a solution. However, this may not be feasible when using dense markers, as the two types of ancestors tend to converge (Powell et al. 2010).

Typically, a reference population (also called the “base” population) is used to define that two alleles are IBD. In the case of a known pedigree, the pedigree founders would be the relevant baseline population. However, this is often not applied, and SNP data are commonly used for estimation with no reference to a known pedigree.

When inbreeding is present, additional higher-order coefficients of relatedness other than IBD0, 1 and 2 better explain the level of relatedness (Lynch and Ritland 1999). In order to accommodate this, Jacquard’s nine coefficients ( $\Delta$ ) may be used (Jacquard 1972).

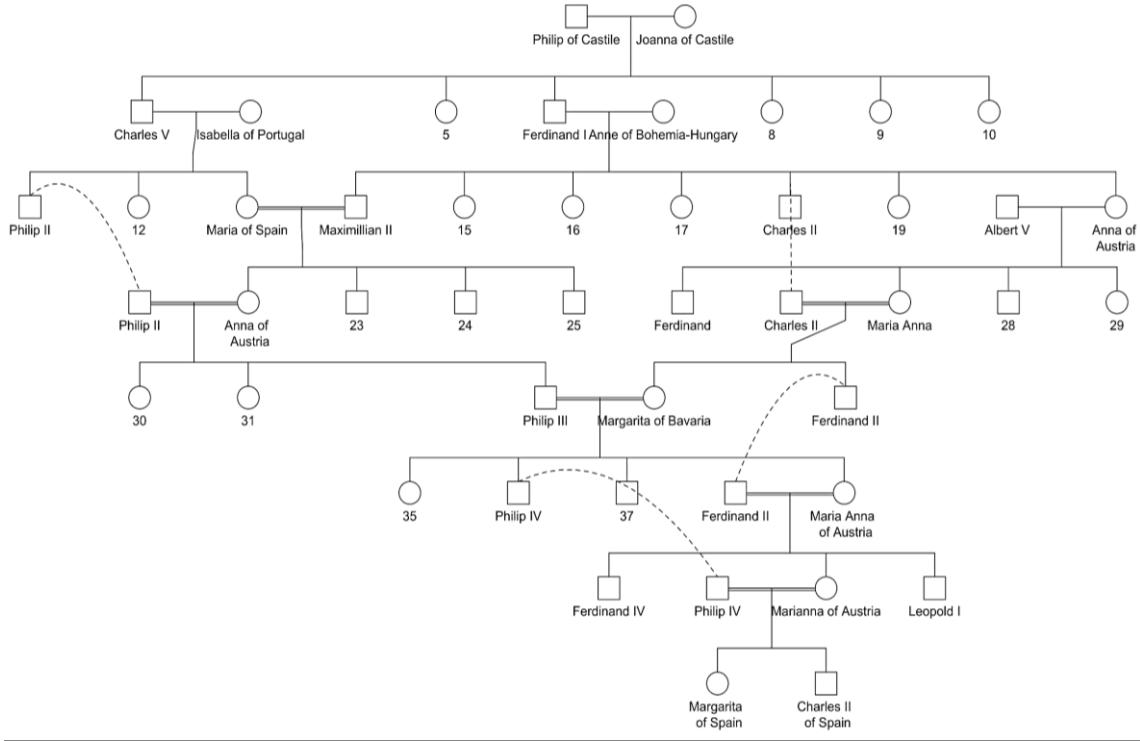
These parameters represent the probability that two individuals show IBD at one of 1 to 9 states (Figure 1.10). The first six states ( $\Delta_1$  through  $\Delta_6$ ) are generally zero in a large, random-mating outbred population, as they are associated with inbreeding.  $\Delta_7$  represents the probability that two individuals share both of their alleles IBD (i.e. full siblings).



**Figure 1.10 Jacquard's 9 condensed identity states.**

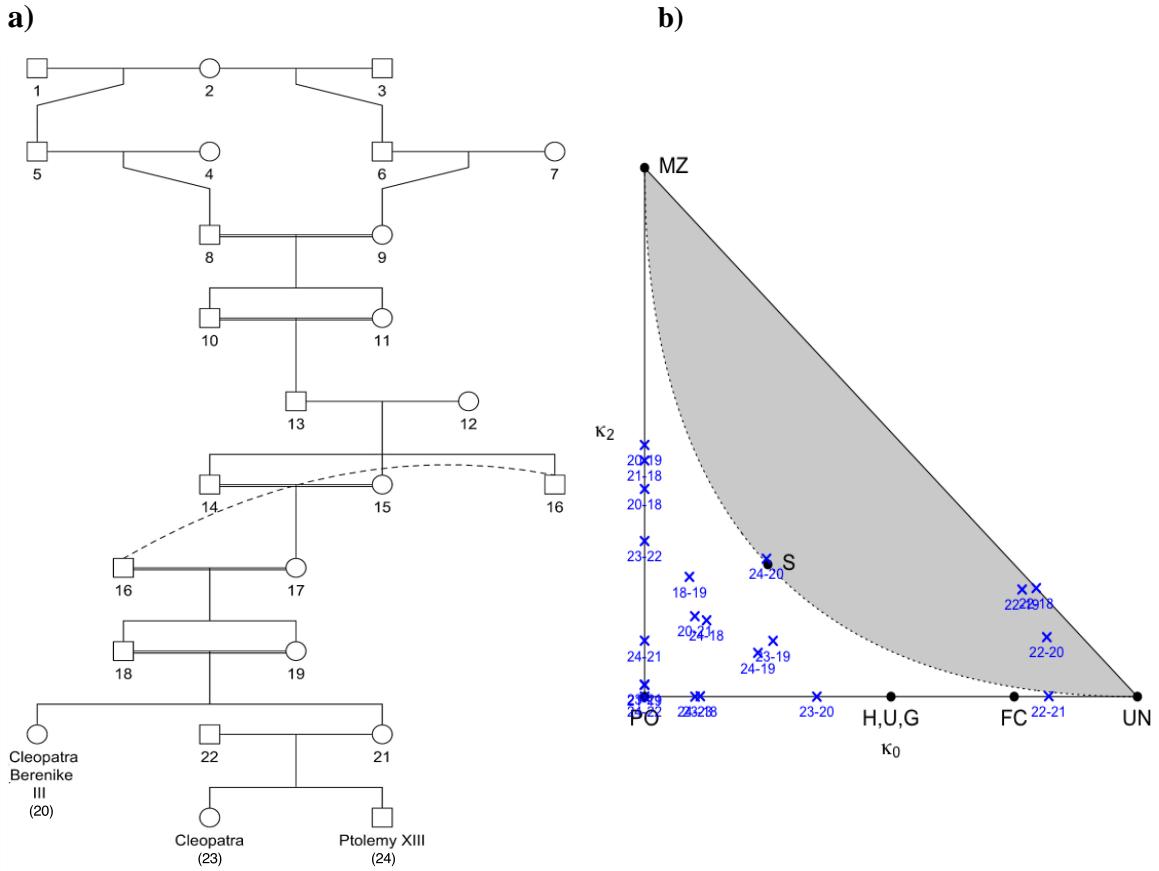
Each dot represents an allele, either maternally or paternally inherited. The lines connect alleles that are in IBD. The first six states are non-zero only when inbreeding occurs. For outbred individuals, only the last three states are possible and correspond to IBD2 ( $\Delta_7$ , where both alleles for both individuals are IBD, expressed by the segment connecting the dots), IBD1 ( $\Delta_8$ , where only one allele is IBD) and IBD0 ( $\Delta_9$ , no alleles are IBD, and therefore no dots are connected).

An extreme case of inbreeding is incest. Mating within the family creates complex relationship pedigree graphs with “loops”. Here, two historical examples are presented: the Habsburg pedigree (Figure 1.11) and the Ptolemaic dynasty (Figure 1.12). The pedigrees present several loops, and inbreeding is highlighted by double lines representing marriage between related individuals. This has an impact on the kinship coefficients (Figure 1.12 shows calculated IBD on simulated data for an example of loop pedigree).



**Figure 1.11 Pedigree representing the Habsburg family tree (1440-1740), with Philip and Joanna of Castile as founders.**

The dotted lines highlight that the same individual has several positions in the pedigree (e.g. Philip II is both uncle and spouse of Anna of Austria), creating closed “loops”. The double lines represent a within family (inbreeding) marriage. The calculated inbreeding coefficient is around 0.025 for Philip I and 0.254 for Charles II (a value expected for offspring of parent-child or sibling pairs).



**Figure 1.12 Family tree of the Ptolemaic dynasty in Egypt (305-30 BC) and IBD estimation.**  
 (a) In the family tree, the dotted lines highlight that the same individual has several positions in the pedigree creating closed “loops”. The double lines represent a within family (inbreeding) marriage. (b) Scatterplot of the 0 and 2 IBD proportions for different related pairs. The individuals are defined by numbers: Cleopatra Berenike III = 20, Cleopatra = 23, Ptolemy XIII = 24; the data were simulated using forrel v 1.3.0 R package (ped suite, <https://github.com/magnusdv/forrel>) on 16 STRs (D3S1358, D8S1179, D18S51, D21S11, FGA, TH01, vWA, D2S441, D10S1248, D22S1045, D1S1656, D12S391, D2S1338, D16S539, D19S433, SE33) with European allele frequencies (1000 Genomes Project, <http://spsmart.cesga.es/popstr.php>). It is possible to notice that Cleopatra Berenike III and Ptolemy XIII (avuncular) cluster towards full sibling (S) instead of the avuncular/half-sibling/grandparental (H,U,G).

## 1.5.2 Approaches to kinship estimation

There are three main approaches that use molecular marker information to estimate relatedness: methods that consider IBD, pedigree reconstruction methods, and likelihood approaches, which calculate and compare the likelihoods of relationship hypotheses given the marker data.

### 1.5.2.1 IBD-based approaches

The approaches to identify recent common ancestry through IBD sharing may be divided into three main classes (Browning 2012): methods that consider the length of sharing, model-based methods for estimating IBD probabilities, and methods that consider haplotype frequencies (considering whether the estimated frequency of shared haplotypes is below a threshold; this class of methods will not be discussed in this thesis).

**Length-based methods.** Two haplotypes can be considered IBD across the whole segment when they are identical at most alleles in a large segment. These methods can either consider genotype data or haplotype data.

Genotype-based methods, generally, search for long chromosomal segments where at least one allele is in common among the pair of individuals (Miyazawa et al. 2007), stopping where individuals show different homozygous alleles at a locus. The length threshold must be chosen in order to obtain a balance between false positives and false negatives, as a high threshold leads to low power and a low threshold to a high false positive rate (Browning 2012). For example, in a family-based study on a Mendelian disorder, a threshold of 3.0 cM was reported to keep both false negatives and positives at a low rate (Browning 2012; Miyazawa et al. 2007), while, for inferring haplotype phase with low false positive rate, a threshold of 10 cM was proposed (Kong et al. 2008).

Haplotype-based methods can directly use haplotype information to identify shared haplotypes, considering their chromosomal length. Examples of tools that do this are Germline (Gusev et al. 2009), which considers the length of shared haplotypes.

There are some limits to both haplotype and genotype-based methods. The former are more suitable for detecting short IBD segments (2-10 cM in length), but long IBD segments may be fragmented due to errors in determining haplotypes from genotypes (phasing) (Zhou et al. 2020). Accurate phasing is not necessary for genotype-based methods, which can detect long segments ( $\geq 15$  cM), suitable for first- and second-degree relative identification.

It is important to take into consideration that segment length varies, affected by recombination at each meiosis (with a rate of approximately one event per Morgan per meiosis). Therefore, long or rare haplotypes shared between two individuals are more likely to be IBD, indicating recent relatedness, compared to short IBD segments, which may be an indicator of more ancient relatedness (Zhou et al. 2020). IBD segments shared between populations may also signal some specific aspect of migration history.

As relationships become more distant, they share fewer autosomal segments IBD (Huff et al. 2011; Staples et al. 2016). Staples and coworkers (2016) (Staples et al. 2016) discuss a limit at ninth degree (i.e. fourth cousin), highlighting that tenth-eleventh degree need different approaches and existing pairwise comparison tools are not enough. For example, individuals separated by 12 meioses have a small chance of sharing autosomal segments of expected length 8 Mb (Browning and Browning 2012); individuals separated by 20 meioses have a very low chance (probability of ~0.001) of sharing any segments, but if they do share segments, these have an expected length of 5 Mb (Brown et al. 2012). Dense SNP marker data may provide more precision.

Methods based on autosomal segment length have not been used in this project, but were applied in analysing X-chromosome SNP data in Chapter 2.

**Model-based IBD approaches.** These methods estimate the unobserved IBD status at each marker (Browning 2012; Zhou et al. 2020). IBD states generally considered are:

- Binary, recording presence or absence of IBD;
- Ternary, counting the pairs of haplotypes shared IBD (0, 1 or 2);
- Multiple values, such as 9 coefficients for unphased data (Jacquard 1972) or 15 states for phased data (Thompson 2008).

Changes in IBD status are usually modelled using a Markov Model (IBD sharing is dependent only on the alleles at a locus and on the state of the previous locus), which is performed on the genetic distance scale (cM) accounting in this way for recombination, or a hidden Markov Model (HMM) where the hidden state corresponds to the IBD status, considering the observed genotypes to be independent, and assuming Hardy-Weinberg equilibrium. Probabilities of genotype errors can be incorporated, and LD models may be introduced. Maximum Likelihood estimates (MLEs) can be also used. MLEs are heavily influenced by the sample size and are computationally intensive (Milligan 2003;

Thompson 1975), but more precise than method of moments estimators (MMEs), which are more computationally efficient and can take ancestral allele fractions into account (Astle and Balding 2009; Milligan 2003). For example, HMM is applied in PLINK (Purcell et al. 2007) where IBD status is deduced from IBS sharing. LD models can be introduced (as in RELATE, Albrechtsen et al. 2009; IBDLD, Han and Abney 2011; BEAGLE extension, Browning et al. 2018, Browning and Browning 2007). Most IBD estimation methods assume that all markers are unlinked and in linkage equilibrium. There are some approaches that account for the possibility of linkage. Merlin's approach is based on sparse inheritance trees (Abecasis et al. 2002b). Kling and coworkers (2015) proposed a method that uses two Markov chain approaches, one for inheritance patterns and the other for founder allele patterns that accounts for LD, implemented in FamLinkX for X-chromosomal markers.

Another aspect to consider in kin relatedness estimation in genome-wide SNP datasets is the presence of substructure due to different population groups and possible admixture. The correct identification of population structure is essential for many research applications that use genome-wide SNP datasets such as association studies (Thornton 2015), population genetics, and medical genomics (including personalised medicine) (Conomos et al. 2015a; De la Cruz and Raska 2014).

Generally, in order to identify underlying population structure and correct for stratification in genome-wide association studies (GWAS), the most widely used approach is principal components analysis (PCA) (Conomos et al. 2015a; Patterson et al. 2006; Price et al. 2006; Thornton 2015). PCA performs a dimension reduction to principal components (PCs) of genotype values reflecting the variability of the data. Top PCs reflect population structure, but, when samples contain related individuals (cryptic relatedness or family data), these vectors are confounded by family structure: due to the correlated genotypes among relatives, artifactual PCs may be induced (Conomos et al. 2015a; De Andrade et al. 2015; Thornton 2014). When individuals are known to be related (e.g. pedigrees), often eigenvectors are calculated considering only the founders in the dataset. However, cryptic relatedness as well as mis-specified and incomplete genealogical information cannot be handled, and founders' genotypes are fundamental (but not always available). In addition, a subset of founders may not be entirely representative of the dataset's ancestries (Chen et al. 2013; Conomos et al. 2015a).

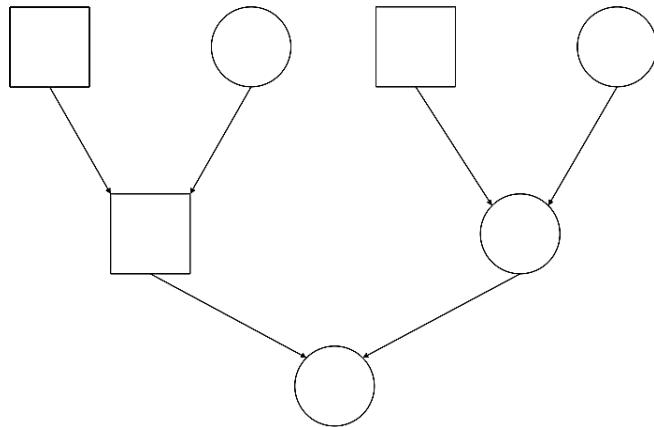
Therefore, the ability to identify a set of unrelated individuals or most distantly related individuals via IBD pairwise estimates in a family dataset is of great relevance.

More details on tools used in this thesis are given in Chapter 2 and 3.

### 1.5.2.2 Pedigree reconstruction methods

Pairwise estimates for constructing pedigree relationships have some limits. Multiple pedigrees may be compatible with the genetic data, and pairwise methods may not be able to identify the correct structure nor possible inconsistencies with a reported pedigree (for example, when identifying a 4<sup>th</sup> degree relationship, it is difficult to distinguish between a great-avuncular and first cousin). Also, manually reconstructing multigenerational pedigrees from pairwise data is error-prone and time-consuming, and inconsistent samples are commonly excluded, possibly losing important data (Staples 2014).

Another approach is to use pedigree reconstruction tools, using the information of family structure. Pedigree reconstruction has many applications including heritability estimation, family-based association, pedigree-aware SNP imputation and phasing, ecology, genealogy inference, and victim identification (Shem-Tov and Halperin 2014). Generally, the most common pedigree reconstruction scenario uses the genotypes of several generations as a starting point for estimating the pedigree structure that best fits the observed data (Shem-Tov and Halperin 2014). Reconstruction can utilise the pedigree graph (Figure 1.13), where individuals are represented as nodes, and connections between parents and children are edges. As in standard pedigree representations, females and males are shown respectively as circles and squares at nodes (Kirkpatrick et al. 2011).



**Figure 1.13 Pedigree graph.**

Nodes are represented by squares and circles (individuals).

Initial approaches to pedigree reconstruction were based on a structured machine learning approach that maximizes the probability of observing the data (Thompson 1985), and on HMM-based methods (Stankovich et al. 2005; Thatte and Steel 2008). Common issues in pedigree reconstruction are parental errors (i.e. a wrong pedigree structure), phenotype (or Mendelian) errors (i.e. the phenotype is incompatible with the known genotype) and non-Mendelian errors (Sanchez et al. 2008).

#### 1.5.2.3 Evaluating relationship hypotheses via the Likelihood Ratio (LR)

Likelihood ratios compare two contrasting hypotheses and express how much more likely one hypothesis is in explaining the genetic data than the other. Likelihoods are computed, using underlying allele frequencies, and then compared (Gjertson et al. 2007b). Hypotheses are generally defined as  $H_1$  (or the prosecution hypothesis, e.g. “the tested man is the father of the child”) and  $H_2$  (or the null, or defense hypothesis, e.g. “an unrelated untested man is the father of the child”):

$$LR = \frac{Pr(data | H_1)}{Pr(data | H_2)}.$$

The phrasing is crucial, as it may affect the LR (Egeland et al. 2000).

Mutation rates of forensic STRs are low, between 0.01-0.22% per transmission for the 15 CODIS loci ([www.cstl.nist.gov/biotech/strbase/mutation.htm](http://www.cstl.nist.gov/biotech/strbase/mutation.htm)), but mutation may in principle affect the LR results. The effect of mutation is most important for parent-

offspring relationships, where it can lead to a mismatch between profiles resulting in a likelihood of zero. One simple practical approach is to exclude the discordant STR and to re-calculate the likelihood (Laurent et al. 2022); alternatively, calculations can be done using tools that include STR mutation models (Morimoto et al. 2020; Kling et al. 2014).

More details on LR characteristics are in Chapter 4.

#### **1.5.2.4 Linkage Disequilibrium in kinship analysis**

Linkage Disequilibrium (LD) introduces additional background sharing (Henden et al. 2016). Markers that are close together may have the same or similar genealogies due to the dependence between alleles induced by LD, unlike markers that are distant and have, consequently, different ancestral genealogies due to recombination.

LD affects estimation of relatedness and of population stratification, introducing bias in the statistical calculations and leading to possibly incorrect interpretation if this factor is not taken into account (Boyles et al. 2005; Kling et al. 2015c). This is due to the fact that allele frequencies for alleles at closely linked loci are not truly independent. Not accounting for LD when using tools that assume linkage equilibrium may lead to the false detection of IBD, particularly where pedigrees have few genotyped individuals (Browning and Browning 2011b). It also biases chromosome-specific and genome-wide tests of homozygosity (Coleman et al. 2016).

Genome-wide SNP panels typically contain hundreds of thousands of markers, and therefore considering LD is of particular importance. The general approach when handling SNP data is to exclude regions of extended LD (e.g. the HLA region) during the quality check phase: pruning LD among SNP markers is suggested for both obtaining an unbiased detection of cryptic relatedness and population stratification, and easing the computational burden for later analysis (Anderson et al. 2010; Weale 2010). Generally, pruning can be performed excluding SNPs that are linked (e.g.  $r^2 > 0.2$ ) in a specified window (e.g. 50 kb) (Anderson et al. 2010), but there is no general consensus on how to choose the best marker set with minimal LD and enough power to detect distant relatives (Ko and Nielsen 2017). Overall, no guidelines exist regarding approaches able to balance between marker pruning and retaining sufficient information (Ko and Nielsen 2017).

Some methods claim to account and overcome the effects of LD, applying a Markov process (as in Albrechtsen et al. 2008). Other tools implement the Lander-Green algorithm (Lander and Green 1987), which bases likelihood calculations on pedigree structures, genetic marker data and models linkage (marker dependency) with Markov chains, but is unable to model mutations (Kurbasic and Hössjer 2008). The Merlin tool (Abecasis et al. 2002b) examines clusters of markers in LD, but, again, cannot account for mutation or recombination among markers within a cluster (Abecasis and Wigginton 2005).

### **1.5.3 Relatedness analysis in forensic science**

The application of genetic information to relatedness analysis is linked to the development of DNA fingerprinting (Jeffreys et al. 1985a; Jeffreys et al. 1985b) and the discovery that minisatellite bands were inherited in a Mendelian fashion: the first use of this new discovery was in immigration case in which it was necessary to verify the family relationship of a boy attempting to re-enter the UK to his alleged mother, a UK citizen. Paternity testing is much more commonly performed (Section 1.5.3.1).

Kinship analysis may help in solving identification cases when conventional direct comparisons of reference and queried material cannot be carried out: these include disaster victim and missing individual identification cases. Profiles from human remains can be compared to those from putative relatives to support an identification. However, because the number of autosomal STRs usually analysed is small, these markers lack power to identify relationships beyond parent-child and full sibling. For more distant relationships, most lineage-based genetic markers (mtDNA and the MSY) can be useful in specific cases where there is respectively matrilineal and patrilineal transmission, but depending on the haplotypes involved, the evidential weight can be low.

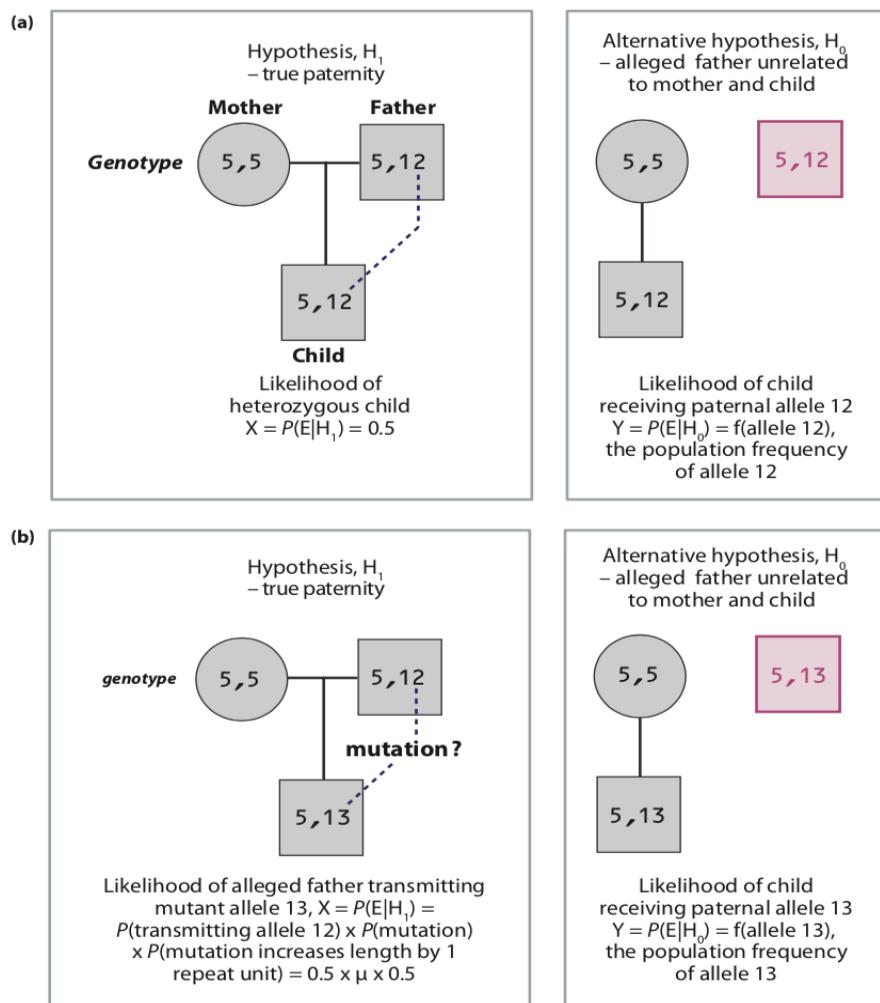
#### **1.5.3.1 Paternity testing**

The simplest, and most frequently carried out forensic kinship analysis is the paternity test, in which the weight of evidence is expressed as likelihood ratio, defined as paternity index (PI; Figure 1.14 a):

$$PI = \frac{P(E|H_1)}{P(E|H_0)},$$

where the hypothesis that the tested man is the father ( $H_1$ ) is tested against the hypothesis that a random man is the father ( $H_0$ ) given the types observed (E).

Standard paternity testing is done using a set of autosomal STRs, and here there is a finite chance that an STR mutation has taken place between a parent and child: gain or loss of a repeat, if not taken into account, can lead to incorrect paternal or maternal exclusion. In fact, mutations have a significant impact on the LR calculations for parent-offspring, as mismatches between the profiles corroborate the hypothesis of non-relationship. A parent and a child have to share at least one allele at each locus, when mutation is not present: a mismatch that is ignored in the calculation will lead to a total likelihood of all loci of zero. In practice, mutation rates for autosomal STRs are well characterized, and mutation can be accounted for in the likelihood ratio. The principle is illustrated in Figure 1.14 b. Other issues to take into consideration in the biostatistical calculations, like possible silent alleles, are not considered here.



**Figure 1.14 Likelihood Ratio use in paternity testing (paternity index, PI).**

These trio pedigrees show genotypes at a single autosomal locus (numbers are the repeat units at the analysed STR). (a) The LR is calculated as  $X/Y = P(E|H_1)/P(E|H_0) = 1/[2f(\text{allele 12})]$ ; with lower allele frequencies (rarer alleles) there is stronger evidence in favour of paternity. (b) The LR considers single-step mutation and is calculated as  $X/Y = P(E|H_1)/P(E|H_0) = 1/[4f(\text{allele 13})/\mu]$ . Source: (Jobling et al. 2014).

Y-chromosomal STRs are sometimes used in more complex cases, defined as deficiency paternity testing, where the potential father cannot be traced or is dead and relatives on the male line (i.e. grandfather, paternal uncle) are tested. Similarly, it is possible to consider the maternal line and to analyse mtDNA.

### **1.5.3.2 Familial search**

Familial search is a well-established investigative tool for identifying close kin relationships, such as parental and sibship, to aid criminal investigations when no direct matches between a suspect's trace and database profiles are obtained (Debus-Sherrill and Field 2017, 2019; Slooten and Meester 2014).

Familial DNA search includes searching an unknown forensic profile against the profiles in a criminal database: this leads to a variable number of "partial matches" sufficiently similar to the queried profile to represent close relatives of the suspect (Gershaw et al. 2011; Slooten and Meester 2014). The matches are ranked using likelihood ratios and can be used as investigative lead, and putative relatives may be interviewed. The number of "hits" depends on the frequencies of alleles making up the queried profile, and can be large (Granja and Machado 2019; Greely et al. 2006). In the UK, searches are often geographically restricted when the case warrants this, in order to reduce the number of partial matches.

Familial search has been utilised in the UK since 2002 (Stokes 2008) in serious criminal cases with single-source DNA, when searches through the National DNA Database fail to identify a match. In the United States, it was first applied in California (2008), then in Colorado (2009), Virginia (2011), and Texas (2012) (Moreau 2012), and it is regulated in other countries such as France and New Zealand (Slooten and Meester 2014). Some well-known cases are the Grim Sleeper and the killing of lorry driver Michael Little in 2003 (examples given in Egeland et al. 2015, Miller 2010). A significant difference between missing person identification and familial search is that, in the first case, the relatives are willingly present in the reference database, while in the second situation the search is made among profiles obtained without consent within the framework of the criminal justice system.

### **1.5.3.3 The rise of investigative genetic genealogy**

As noted above, standard STR-based familial search can only detect very close relatives. For more distant relatives, a larger number of markers is needed. The availability of SNP chips and the ability to obtain SNP-chip profiles from some forensic samples (section 1.4.3.3) has led to the development of investigative genetic genealogy. This takes advantage of large accessible databases of SNP-chip profiles generated through direct-to-

consumer testing of members of the public. One such database, GEDmatch, contains 1.4 million profiles (in 2020) generated by a range of different companies, and uploaded by customers, as well as tools to allow comparisons with a newly input profile, and the generation of a list of “matches” that may represent relatives. These may be distant – it is claimed up to 4<sup>th</sup>-5<sup>th</sup> cousin (Kling et al. 2021).

The most famous case in which this was applied was that of the so-called Golden State Killer (Phillips 2018; Wickenheiser 2019). Joseph DeAngelo, the serial murderer and rapist who was active in California during the 1970s-80s, was identified in April 2018 and convicted thanks to the use of DNA evidence preserved from the crime scenes and the profile subsequently uploaded into GEDmatch. On this database, it was possible to find “matches” (i.e. possible distant relatives) and narrow the research down through genealogical “triangulation” before confirming the identity with classical forensic techniques. Triangulation consists in finding matches (e.g. a probable fourth cousins in this case) and reconstructing the family tree to find the link (i.e. the common ancestor) between the matches and the target individual. However, the successful closure of this forty-year-old cold case poses many questions about genetic privacy, ownership and use of DNA data and evidence as well as database searching by law enforcement. The ethics and issues that concern the applications of genetic genealogy and DTC testing methods in forensics are covered in Chapter 6.

## **1.6 Biogeographical ancestry and phenotypic prediction in forensic science**

In a no-suspect case where a DNA sample is available, the identification of relatives is not the only investigative avenue that can be followed. Genetically based prediction of the ancestry and phenotypic features of the donor can also be attempted. In both cases, standard forensic STRs are of little or no use. While STR genotypes show some geographical differentiation at the continental level, this is noisy (Phillips et al. 2011), and the choice of STRs has deliberately avoided any potential phenotypic associations.

In practice, approaches to both biogeographical ancestry (BGA) and externally visible characteristics (EVCs) have focused on specifically chosen SNPs that form sensitive multiplexes for forensically relevant samples, but genome-wide SNP chips can also offer useful information in both areas. In Chapter 5 of this thesis BGA and EVCs are considered in an admixed population, and further background is given in the introduction to that chapter. Here some general considerations are introduced.

### **1.6.1 Human population history and biogeographical ancestry (BGA)**

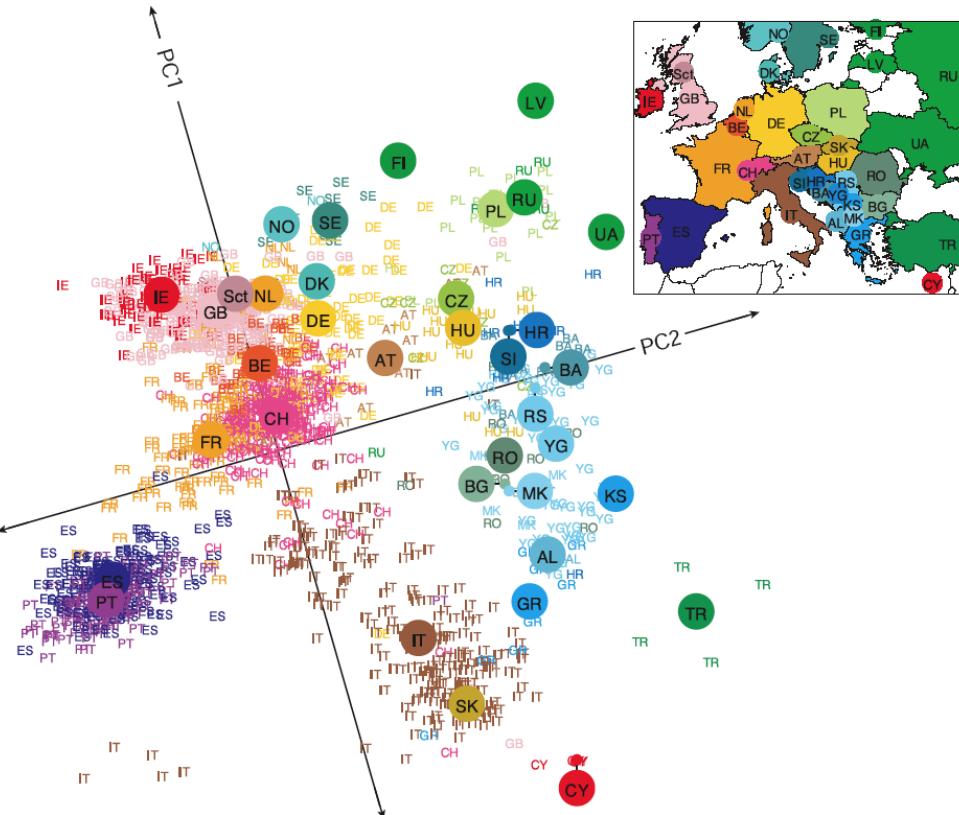
Anatomically modern humans originated in Africa around 200 KYA (thousand years ago), and by ~45 KYA a subset had migrated out of Africa into the Eurasia and Australia (Belle et al. 2006; Jobling et al. 2014). Later migrations colonised the Americas from around 20 KYA (Dulik et al. 2012; Reich et al. 2012), and the islands of the Pacific more recently. As a consequence of this history, African populations today are the most genetically diverse, and there is a reduction of diversity in indigenous populations with increasing distance from Africa (Ramachandran et al. 2005). During the process of migration and since that time, allele frequencies have become differentiated between populations largely through genetic drift. Selection has also played a role at some loci, discussed further in the next section.

Classical results from blood group and protein analysis (Lewontin 1972), and later DNA analysis (Barbujani et al. 1997), showed that most (85%) genetic variation is within, rather than between populations, but nonetheless the signal of the remaining 15% inter-population variation can provide information about ancestry. The power to distinguish between populations depends on how different the populations are (largely related to geographical distance), and the number and type of genetic markers used. Ancestry estimation from genetic data has several applications in different fields of genetics such as association studies, population genetics, personalized and medical genomics and, more recently, in forensic investigation (Conomos et al. 2015a).

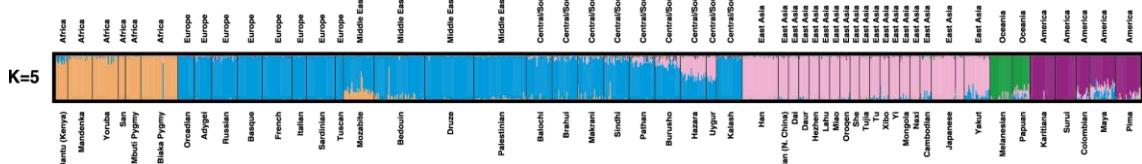
In forensic genetics, targeted SNP panels have been created for BGA estimation, focused on SNPs that show particularly high allele frequency difference between populations. In population genomics more generally, many studies have used genome-wide SNP chip data to study ancestry, together with analytical methods such as principal components analysis (PCA), which displays individuals from different populations in a way that

correlates with geography (e.g. Novembre et al. 2008; Figure 1.15 a), and model-based approaches such as STRUCTURE and ADMIXTURE, which can assign populations to genetic clusters that also have geographical sense (e.g. Rosenberg et al. 2002; Figure 1.15 b).

a)



b)



**Figure 1.15 Classic examples of population structure studies based on SNP chip data and two different analytical approaches.**

(a) Principal component analysis of 1,387 Europeans, from Novembre et al. 2008. Individuals are the small two-letter symbols with letters representing country codes, and the large circles are centroids for countries. Note the resemblance of the scatterplot to the inset map. (b) Estimated ancestral proportions for ~1000 HGDP individuals when considering five clusters (K=5) from Rosenberg et al. 2002; individuals are vertical lines, colours corresponds to the K components. Organisation into populations and regions was done *post facto*.

Throughout human history, populations that have previously been separated and undergone genetic drift have mixed – this process of admixture was particularly marked during the last 500 years, driven by large-scale colonization and practices such as slave trading. In these settings, genetic diversity of admixed populations is high compared to parental populations, and can be deconvoluted using statistical methods and genome-wide SNP data. This subject is returned to in Chapter 5, where an African-European admixed population from Cape Verde is studied.

### **1.6.2 Human population history and externally visible characteristics (EVCs)**

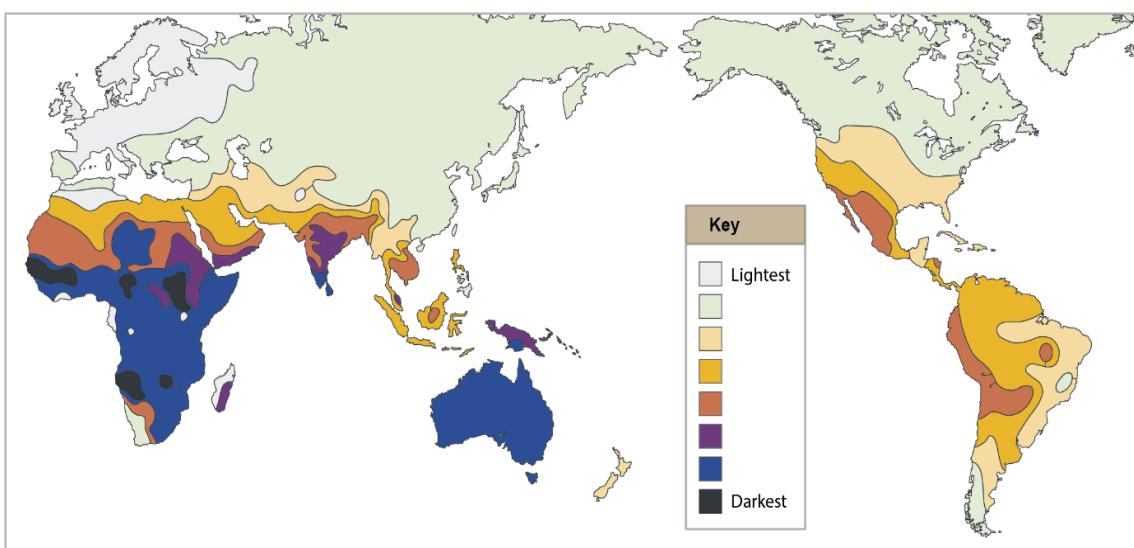
Human externally visible characteristics (EVCs) that have been commented upon by anthropologists and others for centuries include stature, shape of face and facial features, hair type, and colour of skin, eyes and hair (pigmentation). All of these could form part of a witness statement that describes a suspect or victim, and since all have some basis in genetics, the field of forensic science has been interested in the extent to which they can be predicted from DNA samples.

None of the listed traits are determined by single genes, and some are highly complex. Although some work has been done on facial features and hair morphology, most attention has focused on pigmentation, a complex trait influenced by both genetics and other factors such as environment and age. In skin, hair, and the eye's iris it is determined by melanin produced in subcellular compartments, the melanosomes: differences in pigmentation are due to their size, number, melanin composition and distribution. Natural selection, based on a balance between protection against the damaging effects of UV radiation and its importance in vitamin D synthesis, has shaped much of the variation in global skin pigmentation. Sexual selection may also have played a role.

Pigmentation in the eye is one of the traits with highest colour variability in Europe and it can be predicted with high degree of precision. Variants in *HERC2* and *OCA2* are the most relevant for eye colour (according to GWAS studies on Europeans). Skin colour is one of the most complex pigmentation phenotypes. The most important genes involved are *SLC24A5*, *SLC45A2* and *KITLG*, which explain a large portion of the skin

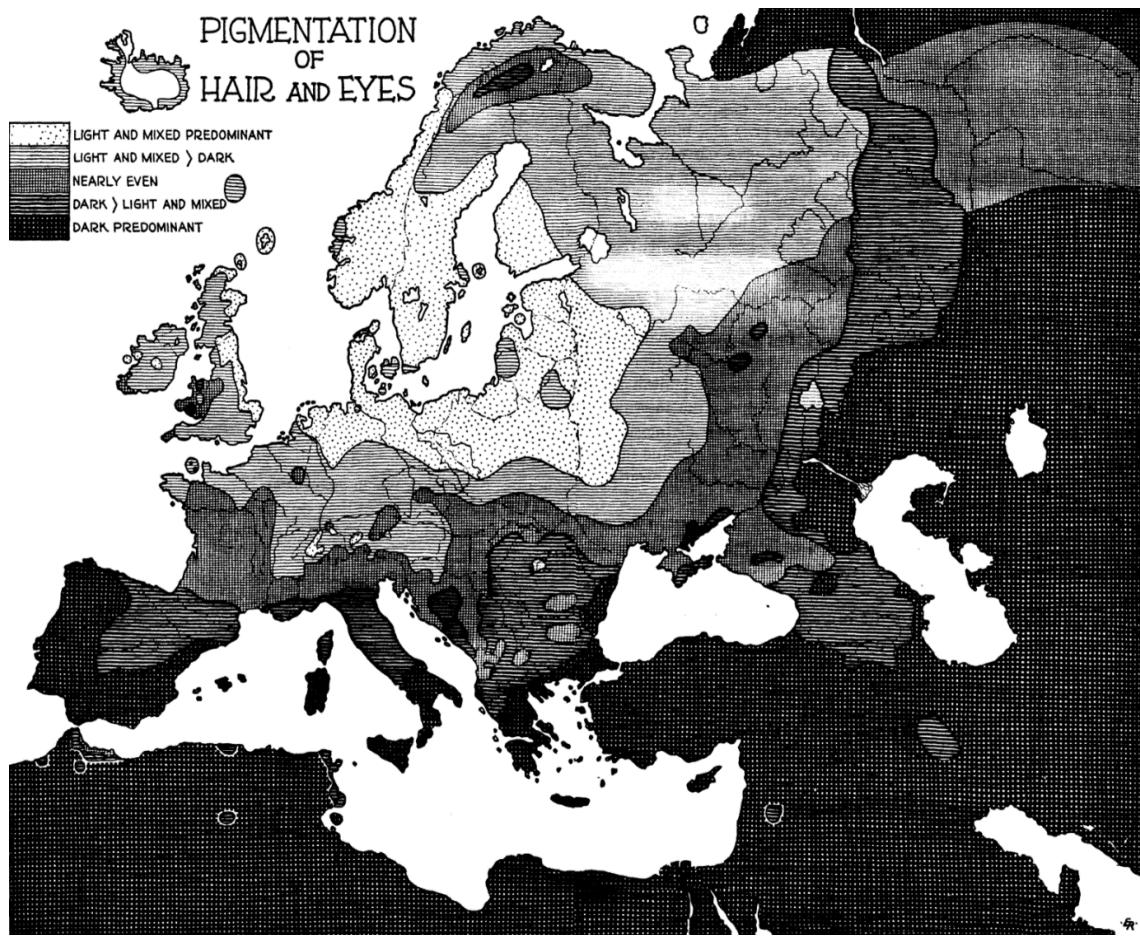
pigmentation differences observed between European and West African populations (Edwards et al. 2010). Forensic plexes targeting these traits have been developed and will be discussed in Chapter 5.

EVCs, including pigmentation, are very geographically differentiated (Figure 1.16 and 1.17): for example, as a trait, skin pigmentation shows 60% differentiation between populations (Relethford 1992), as opposed to the 15% seen for neutral variation. Indeed, the EVCs of forensic interest form the basis of traditional racial classifications. Therefore, BGA and EVCs are closely entangled, and indeed (as discussed in Chapter 5), markers used for BGA overlap to some degree with those used for EVC prediction.



**Figure 1.16 Global distribution of skin colour.**

Based on Jurmain et al. 2018 and Biasutti 1941. Source: Jobling et al. 2014.



**Figure 1.17 Distribution of hair and eye colour in Europe.**

In Europe there is a high degree of variation for hair and eye pigmentation compared to skin colour, which shows more variation globally (Figure 1.16) than in Europe. Source: Coon 1939.

## 1.7 Aims and objectives of this thesis

This thesis explores alternative methods of individual identification using both massively parallel sequencing and SNP chip data for kinship, biogeographical ancestry and phenotypic estimation for forensic and genealogical applications. Both homogenous and admixed datasets are considered, also supported by data simulations. Autosomal and sex-linked markers are analysed.

**Aim 1:** To assess available methods for kinship estimation and pedigree reconstruction based on SNP chip data of real-world family samples containing a wide range of different pedigree relationships. This is described in Chapter 2.

**Objectives:**

- To process SNP chip data, including quality control checks, considering LD and potential substructure in the dataset, which has an impact on subsequent analyses;
- To reconstruct relationships from autosomal data estimating IBD and using pedigree reconstruction approaches that are applied in non-forensic settings;
- To consider X-, Y- and mtDNA-SNP data in the framework of relationship estimation.

**Aim 2:** To assess the possible use of a forensic MPS kit for kinship estimation, proposing an efficient analysis pipeline. This is described in Chapter 3.

**Objectives:**

- To sequence real-world family samples using the ForenSeq Signature Prep DNA kit, Primer mix B (Verogen);
- To combine the available STR and SNP markers, considering the additional variation given by the sequence data;
- To compare the obtained results against known pedigree information of samples and estimates obtained from simulated data.

**Aim 3:** To implement multiple database searches for kinship estimation considering searches performance and using both real and simulated data. This is described in Chapter 4.

**Objectives:**

- To implement a Bayesian approach and parametric representation of the likelihood calculation;
- To propose a method to evaluate the performance of the searches;
- To consider pedigrees with possible founder inbreeding;
- To include both autosomal and X-chromosome data (STRs);
- To make these approaches available through an R package.

**Aim 4:** To explore the biogeographical ancestry and phenotypic prediction output of a forensic MPS tool with unrelated admixed samples. This is described in Chapter 5.

**Objectives:**

- To sequence unrelated samples with European-African ancestry (Cape Verdeans) using the ForenSeq Signature Prep DNA kit, Primer mix B (Verogen);
- To consider PCA and model-based approaches for estimating the ancestral origins of the samples;
- To consider the results from other platforms like the HIrisPlex-S system (Erasmus Medical Center, <https://hirisplex.erasmusmc.nl/>) for eye and skin colour.

# **Chapter 2: Kinship estimation in a homogeneous population using genome-wide SNP data**

The possibility of identifying related individuals in a dataset solely from their genetic profiles is of importance in several research areas (Conomos et al. 2016). For close (first-degree) relationships, as in paternity tests, a small number of autosomal short-tandem repeats (STRs) is sufficient, and probabilities of true paternity can be established with near certainty (Gjertson et al. 2007a). Kinship estimation becomes more challenging when more distant kin relationships are involved, as with each additional generation that separates two individuals, the expected proportion of the genome shared identical by descent (IBD) halves.

Additional power can be gained by increasing the number of autosomal DNA markers analysed, for example by using genome-wide SNP chips, which typically analyse several hundred thousand SNPs simultaneously (Kling and Tillmar 2019; Skare et al. 2009; Gill 2001) . Such chips were designed for genome-wide association studies (GWAS), and in such studies analytical tools offer a check of cryptic relatedness in the dataset, since related individuals can be either excluded if they are regarded as a “nuisance” factor in such studies, or require methods capable of leveraging the added power they can provide (e.g. in family-based studies). A variety of methods have been developed that consider kin relationships by determining relatedness, identifying pairwise relationships, or constructing whole pedigrees, and these find application in, for example, genetic genealogy and forensics.

In Section 1.5.1, a description of the most commonly used approaches was given: models for IBD probabilities, length-based methods, haplotype frequency methods, pedigree reconstruction methods, and LR. This Chapter focuses on approaches that use IBD models and pedigree reconstruction tools for autosomal SNP chip data, applying these to a real SNP-based pedigree dataset, and highlighting advantages and disadvantages and the rationale behind the tools used in the analysis, and offering guidance for kinship analysis in this setting. The tools used were PLINK (Purcell et al. 2007), GENESIS

(Conomos et al. 2015a) and PRIMUS (Staples et al. 2014) (summarised in the following sections), all of which are suitable for dense SNP chip data. PLINK is routinely used in association studies, often as a quality-control tool. GENESIS and PRIMUS offer approaches to identify the most distantly related or unrelated individuals through an IBD model approach and a pedigree reconstruction approach, respectively. This is important for several reasons: it is necessary to identify possible substructure due to different population groups and/or admixture in the dataset because it has an impact on identifying the correct reference dataset for estimating the correct allele frequencies to be used in further analysis (i.e. IBD estimation), and commonly used methods to identify underlying population structure are affected by the presence of clustering due to family structures (e.g. in a PCA, the correlated genotypes among relatives may induce artefactual PCs, Conomos et al. 2015a; De Andrade et al. 2015; Thornton 2014). Other ways to handle this issue have limitations: for example, it is possible to consider only the pedigree founders when family structures are known, but it is assumed that there is no cryptic relatedness, mis-specified or incomplete genealogical information, and that the founders are indeed representative of the dataset's ancestries (Conomos et al. 2015a; Chen et al. 2013). Finally, since many SNP chips contain X-linked, Y-linked and mitochondrial SNPs as well as autosomal SNPs, approaches to analysing and interpreting these data within a pedigree dataset are also described.

### 2.1.1 Summary of the features of PLINK

PLINK v1.90b6.3 is a software run in C++ (Purcell et al. 2007), which can rapidly analyse large genome-wide datasets. It is widely used in GWAS, providing a range of quality control measures and summary statistics, including: genotyping rates, allele and genotype frequencies, Hardy-Weinberg equilibrium tests, single-SNP Mendelian error summaries (for family data), individual heterozygosity rates, sex-checking (based on X-chromosome heterozygosity), and several tests for association studies. Relevant tests for this study are described in more depth in the following paragraphs.

**LD pruning.** Through a repeated sliding-window procedure, it is possible to obtain a subset of independent variants that are in approximate linkage equilibrium, based on correlations between genotype allele counts. PLINK considers only samples flagged as “founders”, and this option does not consider phase. The necessary parameters are an  $r^2$  value (see Chapter 1), set as a threshold to determine if two variants are in LD, the window

size (kb), defining the section under analysis, and a variant count to shift the window: when a pair of variants in the analysed window has a squared correlation greater than the threshold, one of them is noted and pruned.

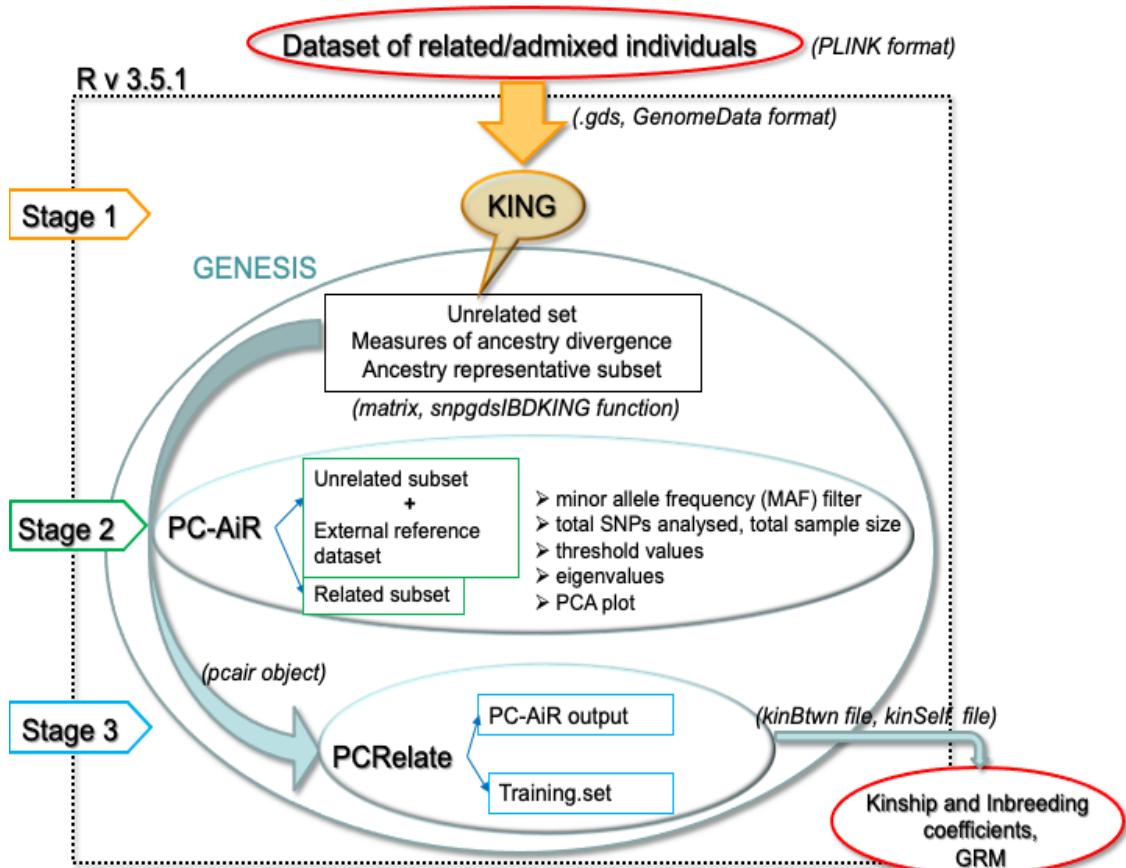
**Population stratification.** Among the approaches offered, PLINK performs Principal components analysis.

### 2.1.2 Summary of the features of GENESIS

In the presence of cluster-correlated genome-wide data (for example, family data), the underlying structure is not adequately addressed by a standard correction for population stratification through principal components adjustments (De Andrade et al. 2015). In a genome-wide association analysis workflow, the family structure is generally unknown, and often only founders are included in the PCA, or individuals are considered to be unrelated. These assumptions may be a reasonable approximation when applied to a homogeneous population, but may not be adequate for admixed or highly structured datasets (De Andrade et al. 2015).

One tool capable of handling cryptic relatedness and population structure in genome-wide SNP data is the R package GENESIS (<https://www.bioconductor.org/packages/release/bioc/html/GENESIS.html>) by Conomos and co-workers (Conomos et al. 2018). This package is based on functions from the “GWAStools” (Gogarten et al. 2012) and "gdsfmt" (Zheng et al. 2012) packages and provides an innovative PCA approach called “PC-AiR” (principal components analysis in related samples) (Conomos et al. 2015a) which detects population structure with known or cryptic relatedness, providing ancestry inference while conditioning out genetic similarity due to recent family (pedigree) relatedness. It also includes an IBD estimation method called “PCRelate” (Conomos et al. 2016), which accounts for representative PCs, and estimates recent ancestry (kinship coefficients, IBD sharing probabilities, and inbreeding coefficients).

**Estimation of recent ancestry and population stratification.** In R v3.5 (R Core Team 2014), the PC-AiR function of the GENESIS package uses as input the KING kinship estimates (Kinship based INference for GWAS) (Manichaikul et al. 2010) and a set of unrelated individuals estimated by KING from the same input dataset (it also includes the pairs most “distantly related” from each family) to calculate principal components and then kinship coefficients. The various steps performed can be summarised into three stages (Figure 2.1).



**Figure 2.1 Workflow adopted by GENESIS.**

A set of different functions (in brackets) must be run by the user (PLINK files were converted into GDS and GenomeData format, `snpgdsIBDKING` function was used to create kinship matrix from the KING algorithm, PC-AiR to create a `pcair` object and PCRelate to create `kinBtwn` and `kinSelf` files that contain all the necessary information on kinship and inbreeding coefficients and the genetic relationships matrix [GRM]).

**(i) Obtaining the initial kinship estimates:** The empirical kinship coefficient estimates are obtained from KING, a tool able to handle samples from structured populations. KING will output a negative estimate when samples belong to different discrete subpopulations: this is used by PC-AiR to ensure unrelated individuals are also population-structure representative.

**(ii) Selection of an informative set of unrelated samples, PCA, and projection onto relatives:** PC-AiR identifies a representative subset for ancestry inference from the genetic data only - no reference population panels or genealogical information are needed.

It chooses two non-overlapping subsets, one made up of diverse, mutually unrelated and ancestry-representative individuals from the entire input dataset, and the other consisting of related individuals, with one relative in the other subset. This is done using measures of pairwise relatedness and ancestry divergence. The ancestry-representative subset is used to infer the axes of variation, and then the coordinates for all other samples are predicted on the basis of genetic similarities with the individuals in the ancestry-representative subset (Conomos et al. 2015a).

(iii) ***Estimation of kinship coefficients adjusted for population structure using PCs***: PC-Relate is a model-free approach for inferring genetic relatedness, including kinship coefficients, probabilities of IBD sharing and inbreeding coefficients (Conomos et al. 2016). The algorithm considers previously calculated PCs and divides correlations among samples into two components, allele sharing due to familial relatedness and recent ancestors (based on IBD), and allele sharing reflecting population structure and distant common ancestors.

### 2.1.3 Summary of the features of PRIMUS

PRIMUS (Staples et al. 2014) is suitable for both SNP-chip and next-generation sequencing data, and reconstructs pedigrees from pairwise IBD estimates, usually supplied from the tools PLINK, KING, or REAP (a method for estimating IBD among samples from structured populations; Thornton et al. 2012). PRIMUS models multiple-generation pedigrees of different sizes without prior knowledge of family structure, considering relatedness between connected individuals up to 3rd degree and identifying the maximum unrelated individuals set (Maximum Unrelated Set Identification algorithm, IMUS). It also accommodates the presence of half-siblings and non-genotyped individuals, and incorporates additional information (age and sex) where available (Ko and Nielsen 2017; Staples et al. 2014). A disadvantage of PRIMUS is its low accuracy for pedigrees with a large proportion (30% or greater) of missing individuals (Ko and Nielsen 2017). However, partial reconstructions (creating smaller familiar nuclei using a higher relatedness threshold limited to 1st- or 2nd-degree relationships) may still give useful information (Staples et al. 2014).

**Estimation of recent ancestry.** PRIMUS's pedigree-reconstruction approach can be divided into three stages:

- i) ***IBD estimation step***: In the “prePRIMUS” stage, IBD estimates are calculated using PLINK v 1.9. Quality control measures are applied to the SNPs and samples, then PCA is performed on a subset of unrelated individuals in order to choose the most appropriate reference population dataset from within the HapMap3 panel (<https://www.sanger.ac.uk/resources/downloads/human/hapmap3.html>). Allele frequencies are then calculated, and IBD is estimated based on these frequencies.
- ii) ***Family network and relationship prediction***: Individuals are assigned to groups according to the expected mean IBD0, IBD1 and IBD2 (covered in Chapter 1, Section 1.5.1). PRIMUS supports six kin relationship classes (parent-offspring; full siblings; half-siblings/avuncular/grandparental; first cousin/great-grandparental/great-avuncular/half-avuncular; distantly related; and unrelated). Through a Kernel-Density-Estimation function, considering IBD0 and IBD1, vectors of the likelihoods corresponding to each relationship used for reconstruction and ranking of possible pedigrees are obtained.
- iii) ***Pedigree reconstruction***: Likelihood vectors are used to construct the pedigrees, incorporating pairwise relationships from the family network and sex data. This starts from close relationships (parent-offspring and full-sibling combinations), then adds second-degree relationships (half-sibling, avuncular, and grandparental) and missing individuals where necessary, and finally third-degree relationships (first-cousin, half-avuncular, great-avuncular, and great-grandparents). The degree of the identified relationships is assigned according to a coefficient of relationship,  $r$ , defined as the proportion of genome shared. The reconstruction stops when all possible relationship combinations are in the output, or when there are no other possible outputs. If the output presents no pedigree, the minimum likelihood threshold is lowered to 0.2 (the default value is 0.3, the minimum is 0.01) (Staples 2014). If the output presents a partial reconstruction (which may happen when crucial individuals are missing or pedigrees are very large), the relatedness cut-off is set to 0.375, allowing smaller, first-degree pedigrees to be output.

Finally, PRIMUS outputs possible pedigrees. If only one has been created, this is the “true pedigree”, up to consideration of 3rd degree relationships. However, there may be more than one structure that can fit the genetic data, so all the possible outputs are ranked according to a pedigree score (sum of the log of the likelihood values of each relationship

in the pedigree), and to additional information, such as age. Partial pedigrees of large families (created because the pedigree is too large, or because too many pedigrees could fit the genetic information) can be recovered by lowering the likelihood threshold to 0.01, and/or creating smaller, 1st-degree networks by using a relatedness cutoff of 0.375 (corresponding to parent-offspring and full-siblings), then connecting together the smaller pedigrees into one complete pedigree, using other tools as BEAGLE (Browning and Browning 2011a; Browning and Browning 2010) and ERSA (Huff et al. 2011).

**Cutoff.** The threshold level of relatedness to be considered in a dataset can be adjusted by the users, as mentioned above. Staples and coworkers (2014) suggest to take into consideration the sensitivity of the method used for calculating pairwise relatedness: PLINK assigns a coefficient of relatedness ( $r$ ) cutoff of 0.1 (the minimum sharing for a pair to be considered related, roughly corresponding to a 3rd-degree or first cousin relationship that is 0.09375 - in PRIMUS), as this tool is not accurate in estimating more distant relationships than first cousins; KING has a cutoff of 0.05 ( $r/2$ ).

#### **2.1.4 Tools for sex-linked and uniparentally-inherited markers**

Additional information relevant to pedigree structure may be gained by including SNPs on Y and X chromosomes and mtDNA. The analysed dataset in this project includes variants in these regions and offers the opportunity to consider the inclusion of these SNPs in the kinship estimation analysis.

Several issues need to be considered when straying from the traditional sets of autosomal variants, and in particular for Y and mtDNA SNPs. Firstly, the rationale for including particular SNPs on a chip is often unclear, and poorly chosen SNPs can affect the informativeness of the data and the performance of methods such as haplogroup estimation. Technically under-performing SNPs may also be relatively frequent given the complex repetitive structure of the Y chromosome (Jobling and Tyler-Smith 2017b) and the small size and high nucleotide diversity of mtDNA (Soares et al. 2009), so genotyping errors may be relatively common. For mtDNA, SNPs at hypermutable sites may not be ideal for identifying maternal line inheritance and could introduce mutations inconsistent within a family structure.

In this Chapter a set of tools is described, and their performance is considered in Section 2.4.

#### **2.1.4.1 XIBD**

IBD detection on the X chromosome is not commonly undertaken, and is generally based on measures of IBS sharing (Browning and Browning 2013; Gusev et al. 2009; Henden et al. 2016). As highlighted in Chapter 1 (Section 1.5), IBD partitions for autosomes and X-chromosomes are different, and the sex of the individuals must be considered for the latter (for outbred individuals, there are three IBD partitions considering autosomes, two IBD X-partitions for two males, two IBD X-partitions for a male/female pair and three IBD X-partitions for two females) (Pinto et al. 2012).

The XIBD tool ((Henden et al. 2016), <https://github.com/bahlolab/XIBD>) utilizes a hidden Markov model (HMM) to infer IBD due to a recent common ancestor (25 generations) for both unphased SNP chip and MPS data. HapMap data (11 populations) offer allele frequencies. The choice of correct reference is fundamental, as it must be representative of the samples; small reference datasets or admixed populations cannot be handled. Two approaches can model LD, however the final input is pruned and possible LD excluded. Pseudoautosomal regions are excluded.

#### **2.1.4.2 yHaplo**

The yHaplo (Poznik 2016) software package (<https://github.com/23andMe/yhaplo>) offers Y-chromosome haplogroup calling from sequence data or SNP chip data, accommodating missing data, genotype errors, and mutation recurrence. The algorithm involves two steps: firstly, based on phylogenetically informative SNPs from the International Society of Genetic Genealogy (ISOGG) database (hg19/GRCh37 reference assembly), a Y-chromosome phylogeny is drafted, then the path on the phylogeny is assigned for each individual. Samples are assumed to be male.

#### **2.1.4.3 Haplogrep**

Haplogrep v2 (Kloss-Brandstätter et al. 2010; Weissensteiner et al. 2016) predicts haplogroups (with quality scores) from sequence or SNP data, based on the mtDNA phylogenetic tree Phylotree (build 17, including 5437 haplogroups; van Oven 2015a). It has both a web interface (<https://haplogrep.i-med.ac.at/app/index.html>) and a command-line package (<https://github.com/seppinho/haplogrep-cmd>).

### **2.1.5 Aims of this Chapter**

This Chapter aims to assess available methods for accurate determination of relationships up to 7th degree (second cousin once removed) when using genome-wide SNP data.

Genome-wide SNP data from a real-world, anonymised dataset of families will be analysed to reconstruct relationships among the individuals. For the autosomal data, three different tools will be used: PLINK (Purcell et al. 2007), which is a commonly used software for GWAS; PRIMUS (Staples et al. 2014), which is specific for pedigree reconstruction; and GENESIS (Conomos et al. 2018), which is able to specifically handle relatedness and population structure.

Firstly, the main characteristics of the dataset will be determined; then, each software will be described, and the workflow created, the results obtained and discussion of the analysis will be presented.

Following from the autosomal analysis, information available from X-, Y- and mtDNA-SNPs will be extracted, analysed and assessed.

## **2.2 Methods**

Analyses were performed on the ALICE high-performance computing (HPC) cluster of the University of Leicester.

### **2.2.1 Sample description**

DNA was extracted from buccal swabs by Eurofins from 72 consenting donors of European (German) origin and typed at Eurofins using the Illumina HumanOmniExpressExome-8 v1.2 SNP chip (964,193 SNPs, of which 1506 Y SNPs, 22927 X SNPs, 218 mtDNA SNPs). Data, pedigree structures and DNA samples were received for further analysis and are described in Section 2.3.1.

## 2.2.2 Reference dataset

A reference dataset is required to provide external allele frequency data for the analyses, and to check population affinity of the tested German samples. Data were obtained from the 1000 Genomes Project Phase 3 (1000 GP, TGPC 2015; <http://www.internationalgenome.org/1000-genomes-browsers>; <ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/>), sub-setting the same markers present in the pedigrees. Phase 3 includes 2504 individuals from five different meta-populations: African (AFR); Ad Mixed American (AMR); East Asian (EAS); European (EUR); South Asian (SAS). The European sub-set (n=503) was used for most analysis steps here, comprising unrelated individuals from: Tuscan in Italy (TSI; n=107); Finnish in Finland (FIN; n=99); British in England and Scotland (GBR; n=91); Iberian population in Spain (IBS; n=107); and Utah residents with Northern and Western European ancestry (CEU; n=99). Notably, 1000 Genomes Project samples do not include a population from Germany, but the CEU, as a north European sample, might be expected to show the closest genetic affinity.

Cryptic relatedness is kin relationship not known before genotyping (Hormozdiari et al. 2014). Several studies have pointed out the presence of cryptic relatedness within the 1000 Genomes Project samples, with relationships closer than first cousins (Al-Khudhair et al. 2015; Fedorova et al. 2016; Gazal et al. 2015; Schlauch et al. 2017). These cryptic relatives were also detected here (data not shown), and were filtered out prior to analysis. Related individuals, originally included in the set, can be found on the 1000 GP website ([ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/related\\_samples\\_vcf/related\\_samples\\_panel.20140910.ALL.panel](ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/related_samples_vcf/related_samples_panel.20140910.ALL.panel), [ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/related\\_samples\\_vcf/](ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/related_samples_vcf/)).

## 2.2.3 Data analysis workflow

The data analysis workflow included five steps:

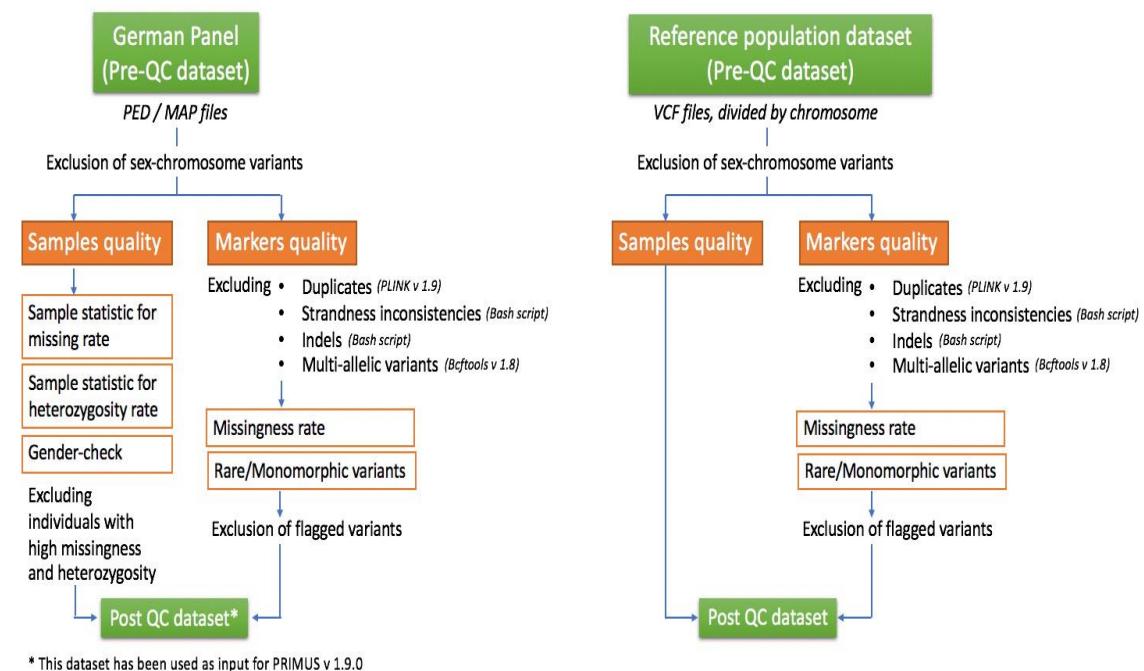
- (i) pre-processing of genome-wide SNP data for sample and marker quality control using PLINK (Purcell et al. 2007) and bash; see Table 2.1 and Figures 2.2 and 2.3;
- (ii) population analysis in order to identify possible (unexpected) substructure: Principal Component Analysis (PCA) was performed using PLINK and PC-AiR (GENESIS).

Linkage Disequilibrium (LD), which may affect subsequent analysis, is addressed using PLINK, excluding redundant variants from the population analysis;

(iii) Identity by Descent (IBD) pattern identification: IBD analysis in PLINK (based on a Hidden Markov model and Method of Moments); kinship coefficients estimation adjusted for population structure and admixture using representative PCs (from PCA) using PC-Relate (GENESIS);

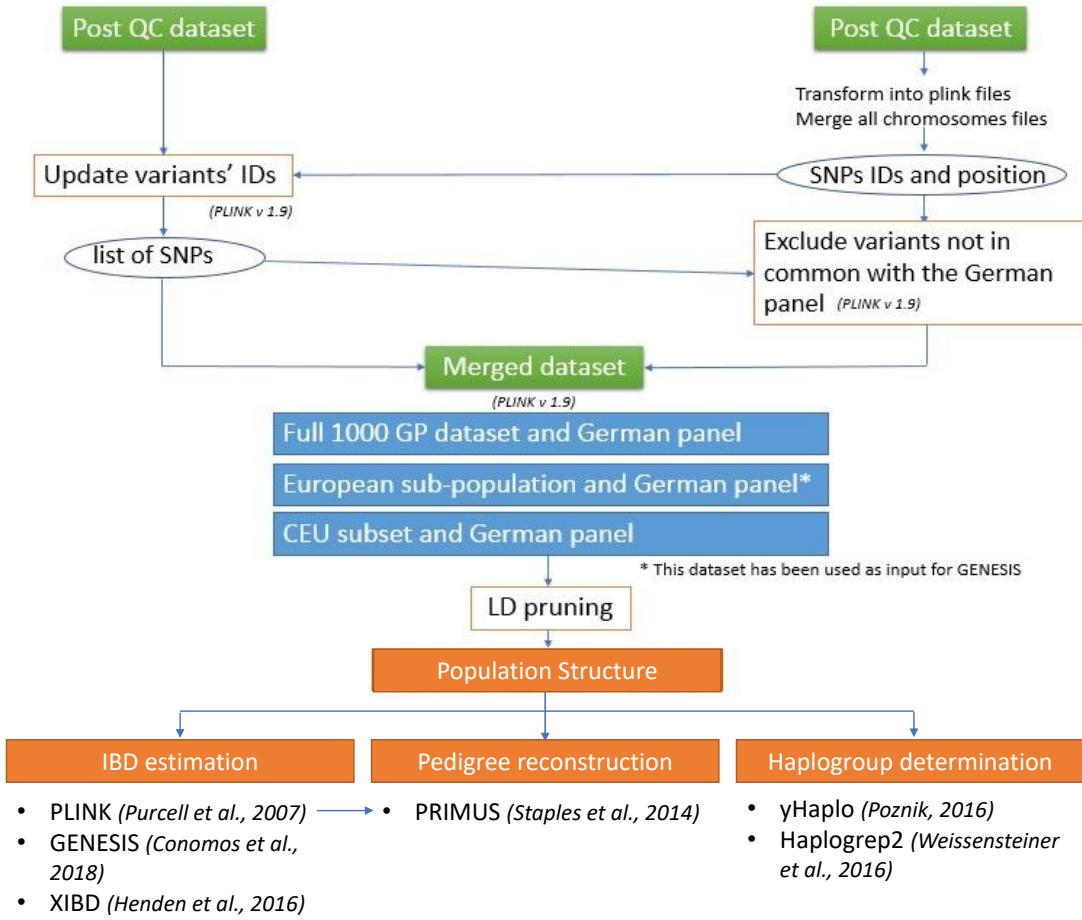
(iv) pedigree reconstruction through a family network and pedigree reconstruction method as applied by PRIMUS (Staples et al. 2014);

(v) inclusion of sex-linked and uniparentally-inherited markers *post facto*.



**Figure 2.2 QC workflow for both the Family samples and reference dataset (1000 Genomes Project Phase 3).**

This workflow shows the step-by-step sample and marker filtering procedures in order to obtain a “cleaned” dataset ready for further analysis (kinship and population structure analysis in PLINK and GENESIS, and family network determination with PRIMUS).



**Figure 2.3 Workflow for both the Family samples and reference dataset (1000 Genomes Project Phase 3) after QC stages.**

This figure shows the procedure used to merge the datasets and files for kinship, and population structure analysis prior kinship determination analyses through different approaches.

### 2.2.3.1 Data analysis workflow: Pre-processing of genome-wide SNP data for sample and marker quality control

Before estimating kinship, the datasets must undergo several quality control (QC) steps. The Family samples and the reference dataset were separately processed through PLINK. Firstly, samples' sex must be checked in order to identify mismatch from the reported sex due, for example, to sample mix-up or genotyping errors. The samples identified in this way are also included in the next QC steps. Sex chromosomes are then excluded from further analysis in this workflow, but are considered later (step [v] in the summary above). Next, sample quality checks are applied: low-quality samples were identified using the metrics of missingness rate and heterozygosity, and excluded from the dataset. Samples which had sex mis-labelled were excluded as well. Marker quality was checked in order to exclude duplicated variants, possible insertions and deletions and strand inconsistencies in genotyping. In addition, all multiallelic SNPs were eliminated using

bcftools v1.8 (samtools, <http://www.htslib.org/doc/#publications>) (Li 2011), as PLINK cannot handle them. Rare variants (with Minor Allele Frequency <1%) and monomorphic alleles were excluded. Parameters applied to both datasets are listed in Table 2.1.

**Table 2.1. Parameters used through the QC steps of the dataset (genome-wide SNP data) in PLINK.**

Parameter	Description	Threshold	PLINK command
<b>Variant missingness</b>	filters out all variants with missing call rates exceeding the provided value	0.05	--geno
<b>Sample missingness</b>	filters out all samples with missing call rates exceeding the provided value	0.05	--mind
<b>Minor allele frequency</b>	filters out all variants with minor allele frequency below the provided threshold	0.01	--maf

At this point, the two datasets were merged, in order to allow PLINK to calculate valid allele frequencies (as the software extrapolates this information directly from the dataset) (Figure 2.3). To allow comparisons, only the variants present in the Family samples were retained in the reference dataset, and SNP IDs in the Family samples were updated according to the reference dataset (to account for different allele and genotype designations between platforms). Three subsets were created: a set with Family samples and full reference panel (which includes five meta-populations), a set with Family samples and European sub-population (EU), and a set containing the Family samples and the French (CEU) sub-set.

### 2.2.3.2 Data analysis workflow: Population structure analysis

**Pruning for Linkage Disequilibrium.** Subsequent analyses are influenced by the inclusion of non-independent SNPs, as patterns of LD confound estimates of IBD and cause kin relationships to be overestimated (Anderson et al. 2010; Browning and Browning 2016; Coleman et al. 2016; Henden et al. 2016; Ko and Nielsen 2017; Weale 2010). A window of 50 variants was used, with a shift of 5 variants between windows, and an  $r^2$  cut-off of 0.2.

**Principal components analysis.** PCA in PLINK was performed on the three subsets (German samples and full 1000 GP, German samples and EU, German samples and CEU), producing 20 dimensions.

Then the PC-AiR (GENESIS) method was used: the dataset was suitable for this because it had already undergone quality control filtering, had rare variants excluded, and LD pruning. This prevents high-density genotyping arrays with highly correlated SNPs clusters confounding individual PCs (Conomos et al. 2015a). In order to produce PCs able to represent the ancestries of all individuals, without confounding via relatedness, a subset of mutually unrelated and ancestry-representative individuals was used, according to measures of ancestry divergence calculated by the KING-robust kinship coefficient estimator (Manichaikul et al. 2010). Then, using the KING matrix, the PC-AiR algorithm divides the dataset into an “unrelated” (ancestry-representative) and a “related” subset. To investigate the underlying structure in each PC, a reference dataset was included (EU subset from 1000 GP).

### 2.2.3.3 Data analysis workflow: IBD estimation

IBD sharing was estimated through PLINK. PLINK estimates IBD from IBS information and hence estimates the coefficient of relationship  $r$  (also called the PI\_HAT value), where  $r$  is defined as:  $P(\text{IBD}=2) + 0.5 * P(\text{IBD}=1)$  and is interpreted as the proportion of genome shared IBD. In Table 2.2, the expected, pedigree-based  $r$  is shown for four relationship degrees.

These proportions of sharing can be plotted, showing clusters corresponding to relationships (see section 2.3.3).

**Table 2.2. Expected, pedigree-based coefficient of relationship ( $r$ ) for some specific relationships.**

<b>r</b>	<b>Percentage IBD</b>	<b>Relationship</b>
<b>1.0</b>	100%	Identical twins, duplicates
<b>0.5</b>	50%	First-degree
<b>0.25</b>	25%	Second-degree
<b>0.125</b>	12.5%	Third-degree

In addition, recent genetic relatedness was estimated through PC-Relate, which can adjust for population structure using the ancestry-representative principal components (PCs) calculated from PC-AiR. An unrelated subset was specified (based on initial measures of kinship and ancestry divergence) and used as a training set for ancestry adjustment; this included all mutually unrelated individuals from the family dataset (as calculated by GENESIS) and individuals from the reference panel (1000 GP). From the output file, it is possible to extract a table of pairwise relatedness estimates and a table of individual inbreeding coefficients.

#### **2.2.3.4 Data analysis workflow: Pedigree reconstruction**

The dataset passing QC in PLINK (Section 2.2.3.1) was used as input in PRIMUS. The workflow followed the suggested steps in the PRIMUS manual. Firstly, the default degree relatedness cutoff of 3rd-degree relationship was used, then it was lowered to 2nd degree.

#### **2.2.3.5 Uniparentally-inherited and sex-linked markers**

Mitochondrial DNA SNPs were used to classify haplogroups via the software Haplogrep2 v2.2 (Kloss-Brandstaetter et al. 2010; Weissensteiner et al. 2016) based on Phylotree 17. Y-SNPs were used to define haplogroups using the yHaplo tool (Poznik 2016). The dataset was LD pruned, allowing the use of the XIBD tool to analyse X-chromosome data (Henden et al. 2016) with the model that assumes LE.

## **2.3 Results**

The families are labelled from 1 to 8, and each individual is labelled with “F”, the number of the family and “P”, followed by the individual’s number in the family tree (e.g. the first person in family 1 would be called “F1P1”). This is reflected in the pedigree figures.

### **2.3.1 Description of pedigrees**

Information regarding family pedigree structures was known: 72 individuals forming 8 pedigrees of known structure were available (Figure 2.4, Table 2.3, Table 2.4),

comprising a variety of pairwise kin relationships up to 7<sup>th</sup> degree (considered as the number of separating meioses between each person, Steffens et al. 2006). Among these pedigrees, most founders were not genotyped (diagrams of family trees are shown in Figure 2.4); specifically, in family 3 two founders are typed, while only one is present in each of family 4, family 5, family 7 and family 8.

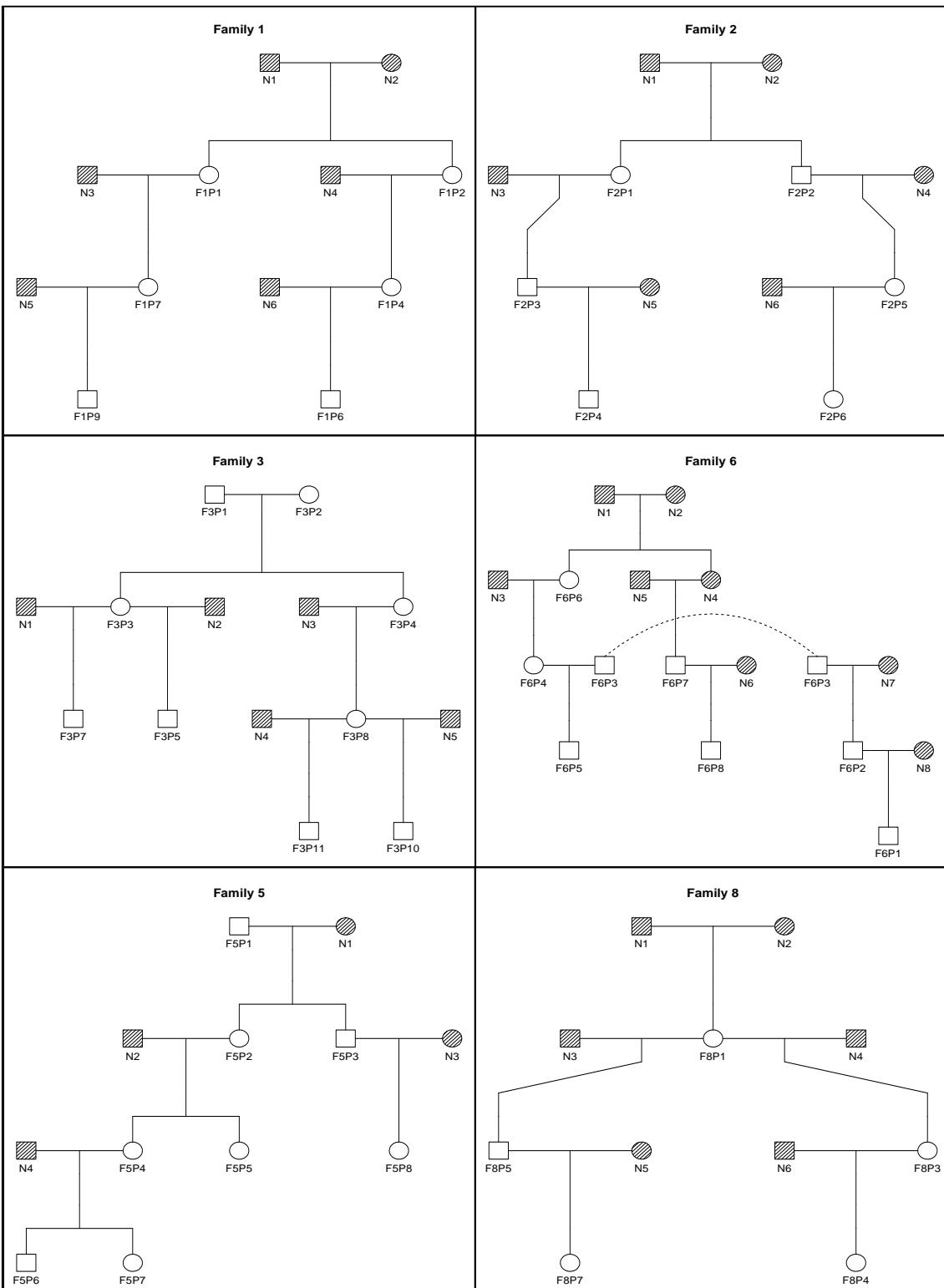
Before any QC step and consequent samples exclusion, there are 56 parent-offspring (PO), 14 full siblings (FS), 6 half-siblings (HS), 31 grandparental, 41 avuncular relationships, 6 great-grandparental, 18 great-avuncular, 30 first cousin, 38 first cousin once removed, 6 first cousin twice removed, 14 second cousin, 4 second cousin once removed and 82 unrelated pairs (UN).

The samples were genotyped by Eurofins on the Illumina HumanOmniExpressExome-8 v1.2 chip, containing 964,193 SNP markers (which as well as autosomal SNPs includes 22,927 X-, 1506 Y-, 218 mtDNA-, and 464 pseudoautosomal SNPs). Eurofins reported that two out of 72 chips failed, and this was confirmed here in the QC analysis (Table 2.3).

**Table 2.3. Total number of kin relationship among the 8 analysed families, 66 samples.**

This table is based on Table 1.3 (Chapter 1) and reports the pedigree relationship, the corresponsive IBD status (IBD 0, 1 and 2) and the number of pairs (after applying QC).

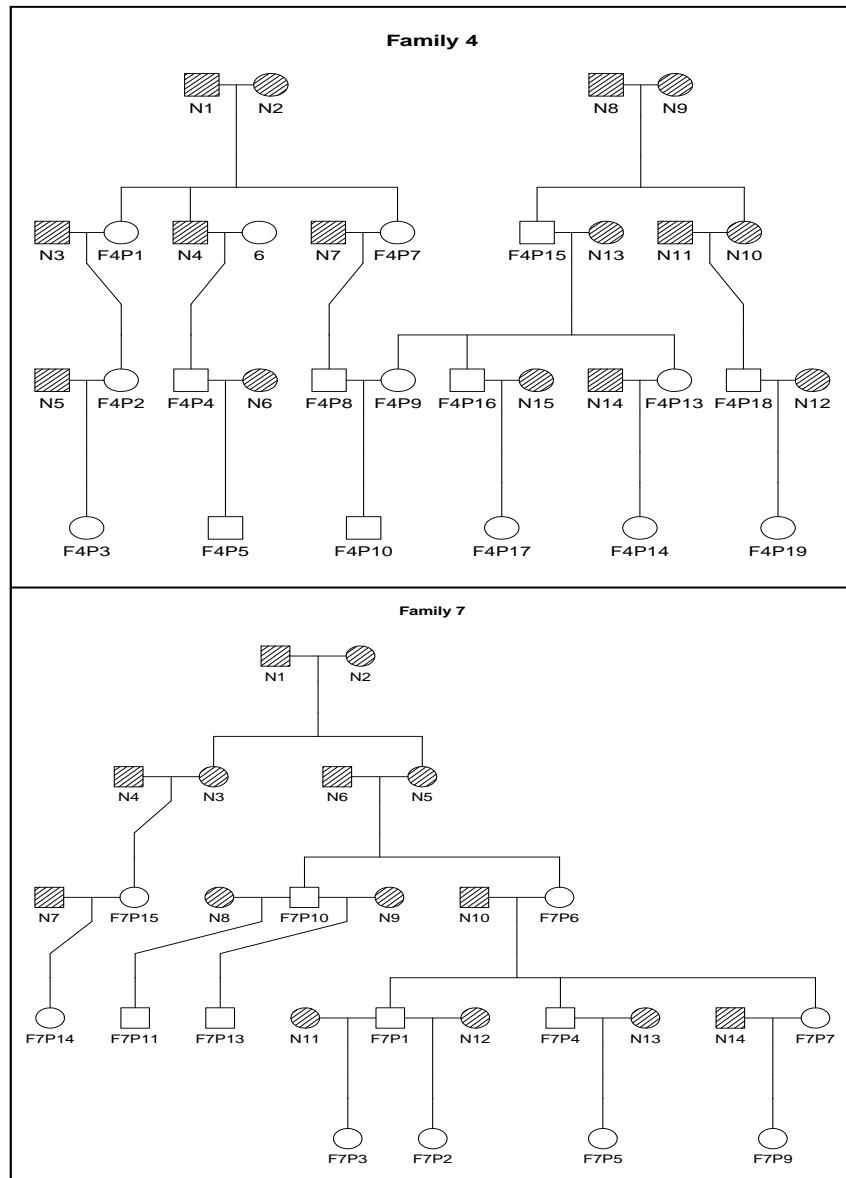
Pedigree relationships	r	IBD 0	IBD 1	IBD 2	Total
<b>PO</b>	0.5	0	1	0	46
<b>FS</b>	0.5	0.25	0.5	0.25	13
<b>HS</b>	0.25	0.5	0.5	0	5
<b>grandparental</b>	0.25	0.5	0.5	0	27
<b>avuncular</b>	0.25	0.5	0.5	0	34
<b>great-grandparental</b>	0.125	0.75	0.25	0	6
<b>great-avuncular</b>	0.125	0.75	0.25	0	18
<b>first cousin</b>	0.125	0.75	0.25	0	24
<b>first cousin once removed</b>	0.0625	0.875	0.125	0	32
<b>first cousin twice removed</b>	0.03125	0.9375	0.0625	0	4
<b>second cousin</b>	0.03125	0.9375	0.0625	0	14
<b>second cousin once removed</b>	0.015625	0.96875	0.03125	0	4
<b>UN</b>	0	1	0	0	71



**Figure 2.4 a Diagrams of family trees for six out of the eight families (pre-QC).**

The shaded individuals are samples that were not collected and typed, and they are labeled with a number and ‘N’. Typed samples have the family number and a number assigned: individual number 1 in family 1 corresponds to “F1P1” in the analyses. These pedigrees show all available samples that will also be considered in Chapter 3, however, samples that failed the chip typing (F2P3, F3P7) and the QC steps (F1P7, F3P4, F6P7) are then excluded from the analysis. The

dashed line in Family 6 highlights that the same individual (3) has offspring from individual 4 and N7.



**Figure 2.4 b Diagrams of family trees for two out of the eight families, Family 4 and 7 (pre-QC).**

The shaded individuals are samples that were not collected and typed, and they are labeled with a number and “N”. Typed samples have the family number and a number assigned: individual number 1 in family 1 corresponds to “F1P1” in the analyses. These pedigrees show all available samples, however, sample F4P8 was excluded from the analysis after QC steps.

**Table 2.4 Information about available kin relationships in the dataset.**

Overview of family members and relationships in each family, considering the pairwise comparison within the family and in the full dataset after QC (which excluded six individuals and are not here reported).

Family number	Number of typed family members	Number of females	Number of males	Number of pairwise relationships	Most distant relationships
1	5	3	2	10	second cousin
2	5	3	2	10	second cousin
3	7	3	4	21	first cousin once removed
4	16	10	6	120	second cousin
5	8	5	3	28	first cousin once removed
6	7	2	5	21	second cousin
7	13	8	5	78	second cousin once removed
8	5	4	1	10	first cousin
Total	66	38	28	2145	-

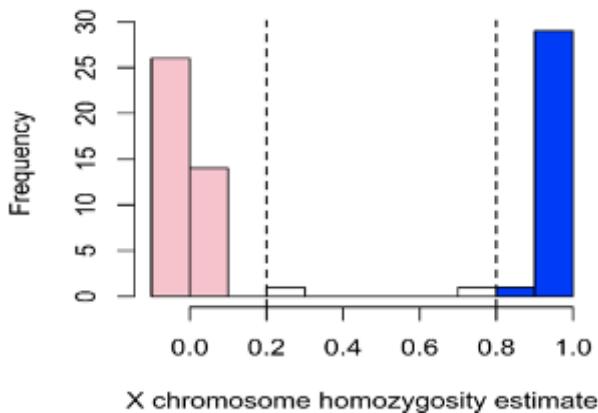
In the provided pedigree of Family 7, the female F7P7 was mislabelled as a founder, while DNA data showed that she is actually part of the pedigree, as a child of F7P6 and a sibling of F7P1 and F7P4. Individuals belonging to different families, F3P7 and F4P15, appeared distantly related based on an initial PLINK analysis, but after a stringent QC (Section 2.3.2), F3P7 failed the quality controls and was excluded from further analysis.

### 2.3.2 Quality Control (QC) analysis

Starting with a total of 964,193 variants, only autosomal SNPs were considered for further QC checks. Prior to QC, the Family dataset contained 72 samples (41 females and 29 males) and 937,880 variants, but only 66 samples (~92%, 38 females and 28 males, 6 founders) and 625,494 variants (66.7%) passed QC. Details of QC are given below.

**Sex mismatch.** PLINK's sex mismatch check calculates the mean homozygosity across X chromosome markers for each individual, assigning "female" status if the proportion of X heterozygous genotypes ( $F$ ) value is under 0.2, and "male" if the value is above 0.8. Based on these thresholds, two ambiguous samples were found: F2P3 and F6P7. Individual F2P3 was listed as female by PLINK, but, due to known genotyping failure, data were not sufficient to confirm this ( $F=0.2159$ ); individual F6P7 ( $F=0.7995$ ) was

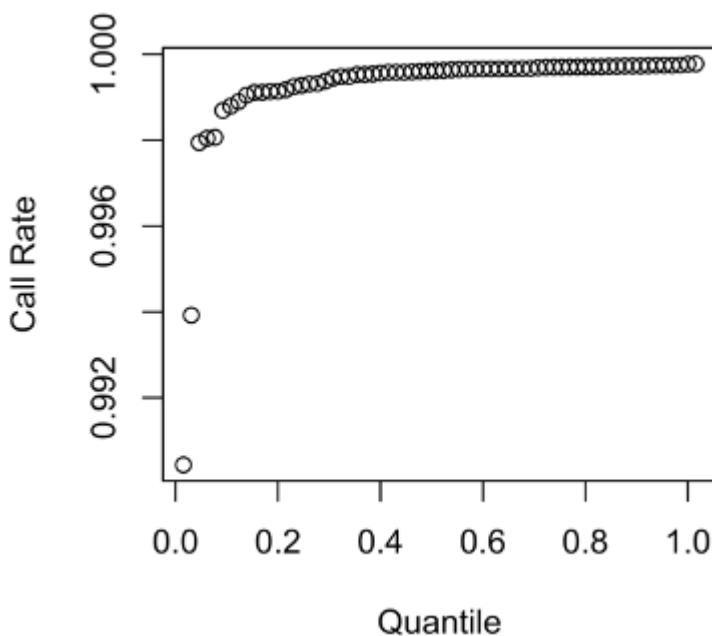
initially wrongly listed as female in the known family tree. For both samples, sex could not be inferred using the above rules, and these individuals have been excluded from further analysis.



**Figure 2.5 Sex mismatches in the dataset.**

Females are called when  $F < 0.2$ , and males called when  $F > 0.8$ . The dashed lines highlight the  $F$  threshold values, and between these lie the two ambiguous individuals.

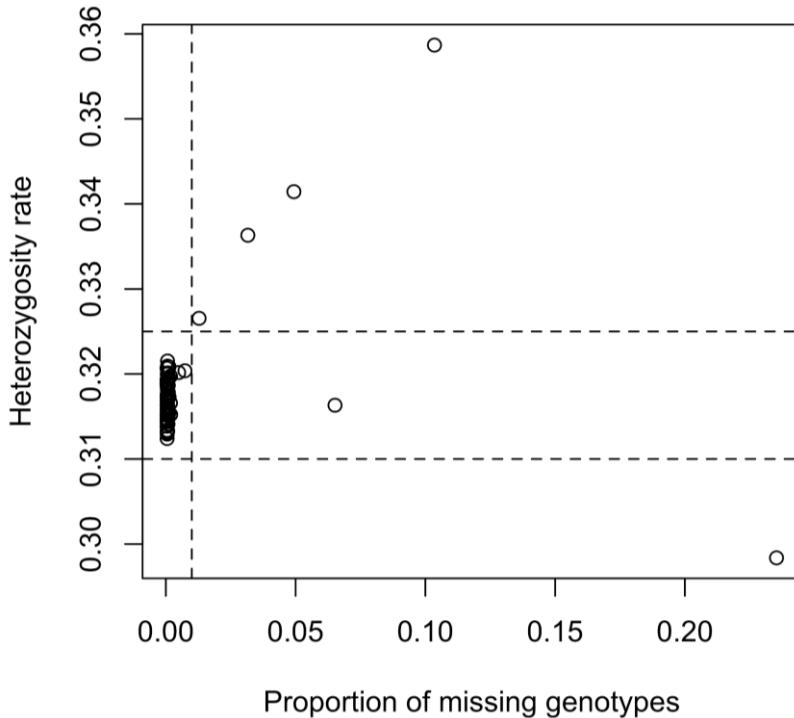
**Call rate checking (samples).** Call rate (as the proportion of non-missing genotypes) was calculated per individual through PLINK and plotted (Figure 2.6). All individuals have a call rate above 99%. Roslin and co-workers (Roslin et al. 2016) suggest a threshold of 97%.



**Figure 2.6 Sample cumulative non-missingness distribution plot (call rate).**

The commonly accepted threshold is at 97% (Roslin et al. 2016).

**Summary of sample filtering.** Sample missingness (x-axis) was plotted against heterozygosity rate per individual (y-axis), in order to identify samples failing these measures (Figure 2.7).



**Figure 2.7 Plot comparing the proportion of missing genotypes and heterozygosity rate in the Family dataset.**

The plot shows six individuals with extreme values: individuals with heterozygosity rate above 0.325 and below 0.31 and also those with missing rate above 0.01 are manually selected as outliers. The thresholds were defined by eye.

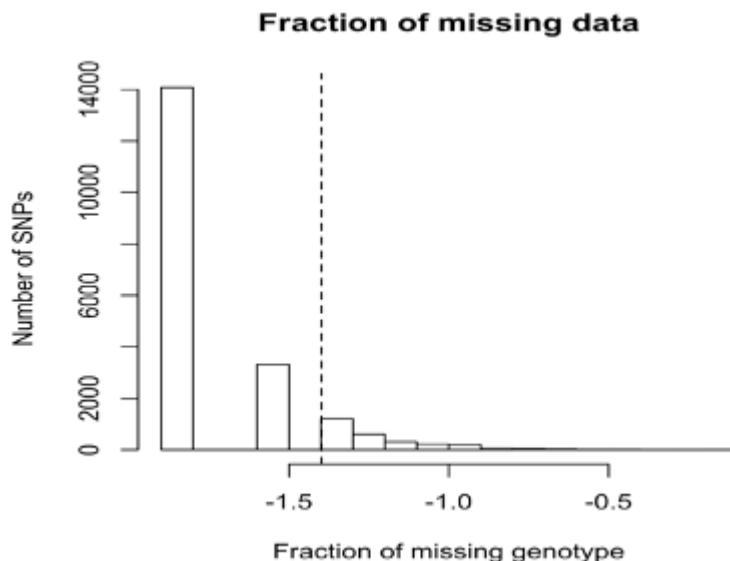
Six outliers with extreme values (showing missing rate >0.01) were identified: F1P7, F2P3, F3P4, F3P7, F4P8, F6P7. Of these individuals, a probe failure was known for F2P3 and F3P4 (who failed both the missingness rate and heterozygosity rate checks). F6P7 was flagged as a sex mismatch.

At this point, the panel contained 66 individuals (~92% of samples passing filter) and 937,879 variants (~97%). Inferred sex was consistent with provided sex for 70 of the 72 samples. There was evidence of sample failure, but no contamination.

**Marker filtering.** Duplicate SNPs were found through the `--list-duplicate-vars` PLINK command. In the Family dataset, 19,288 (~2%) duplicates were excluded

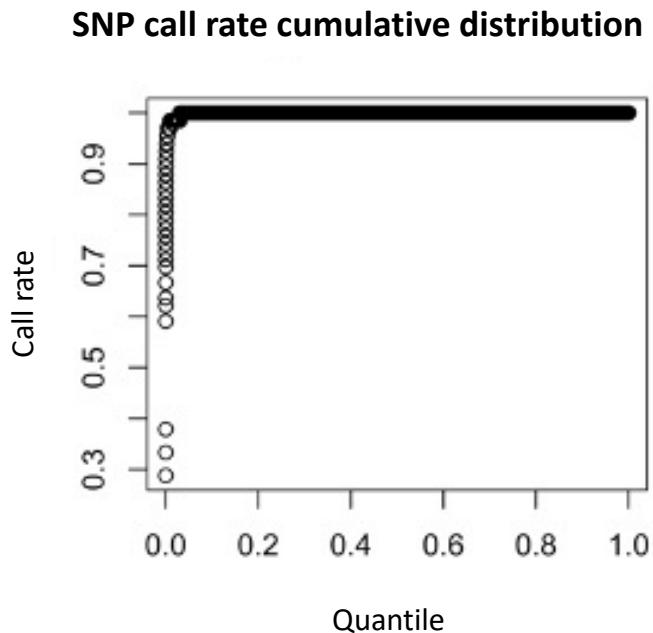
leaving 918,591 variants. Datasets were screened for strand inconsistencies with a bash script to create a list of A-T/C-G variants to be excluded. In total, 6789 variants were excluded, leaving 911,803 variants. Insertion/deletions (not handled by PLINK) were excluded with a bash script, finding 250,717 indels. After this filtering step, there were 661,086 variants remaining.

**Call rate checking (variants).** Call rate at the SNP level (proportion of non-missing genotypes per SNP) was checked and the threshold of maximum missingness set at 4% (Figure 2.8) (generally 3-7%, Anderson et al. 2010): 6018 SNPs (0.9 %) did not pass this filter.



**Figure 2.8 Graph of (logarithmic) portion of missing SNPs.**

The graph shows the proportion of missing SNPs on a logarithmic scale against the number of SNPs considered. The threshold is applied at -1.4 (4% missingness).



**Figure 2.9 SNP cumulative non-missingness distribution plot.**  
The threshold used is 97%.

**Summary of variant filtering.** Finally, variants with missing data as recognised in the previous steps and with MAF < 1% (monomorphic and rare alleles) were excluded from the dataset, leaving 625,494 variants and 66 individuals.

The Reference panel, after QC steps, contains 9,997,177 variants and 2186 individuals (475 EUR; 88 CEU).

Finally, the Family samples and reference dataset were merged, retaining variants in common to both sets (the Family dataset contains 307,319, 31.87% of original number of variants).

### 2.3.3 Population structure analysis

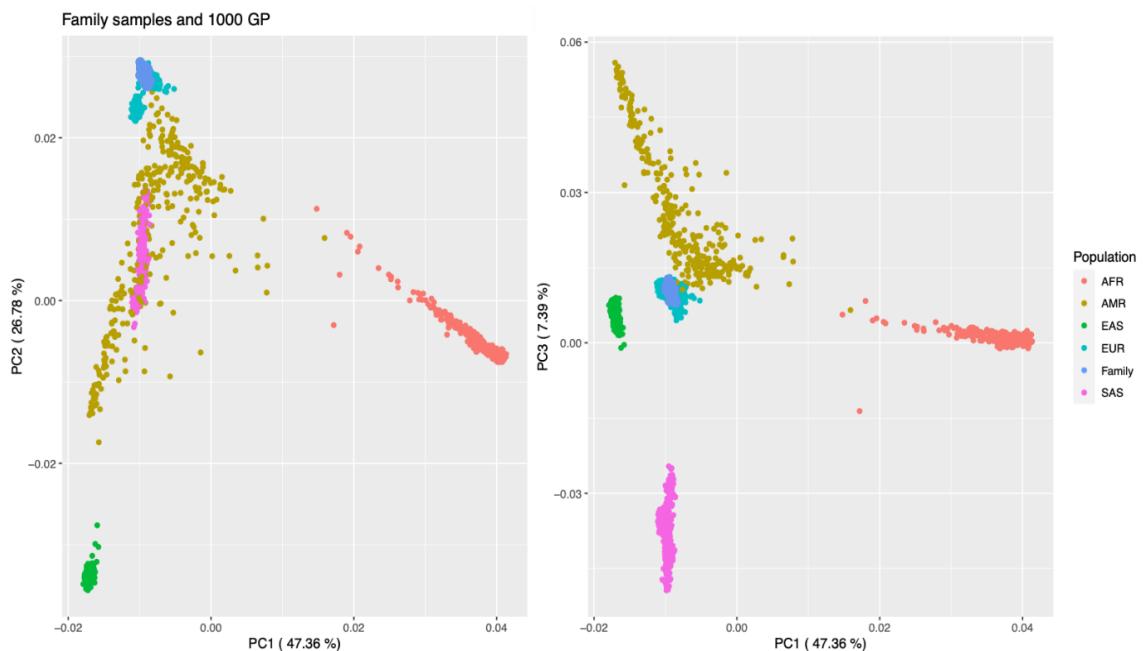
Considering ancestry information and population structure in a sample prior to kinship estimation is important, since it affects the appropriate choice of reference allele frequency data.

Autosome-wide SNP data on the Family samples, together with reference data, were first pruned for LD using PLINK. After pruning the full reference dataset merged with the

Family samples, 197,796 of 307,319 variants were removed and 2252 individuals remained. In the European subset merged with the Family dataset, 218,111 of 307,319 variants were removed and 541 individuals remained. In the CEU sub-group merged with the Family dataset, 222,713 variants of 307,319 were removed, while 154 individuals passed filters.

Population structure was assessed by visual inspection of PCA plots of the Family samples together with various combinations of reference data.

Population distributions of the reference data in the PCA plots of Figure 2.10 are consistent with previous studies (Jakobsson et al. 2008; Li et al. 2008): PC1 differentiates sub-Saharan Africans (AFR) from all other individuals; PC2 reflects a west-east spread of individuals across Eurasia from EUR to EAS; and AMR and SAS lie on the PC2 axis between EAS and EUR. As expected, the Family samples cluster with the EUR reference.

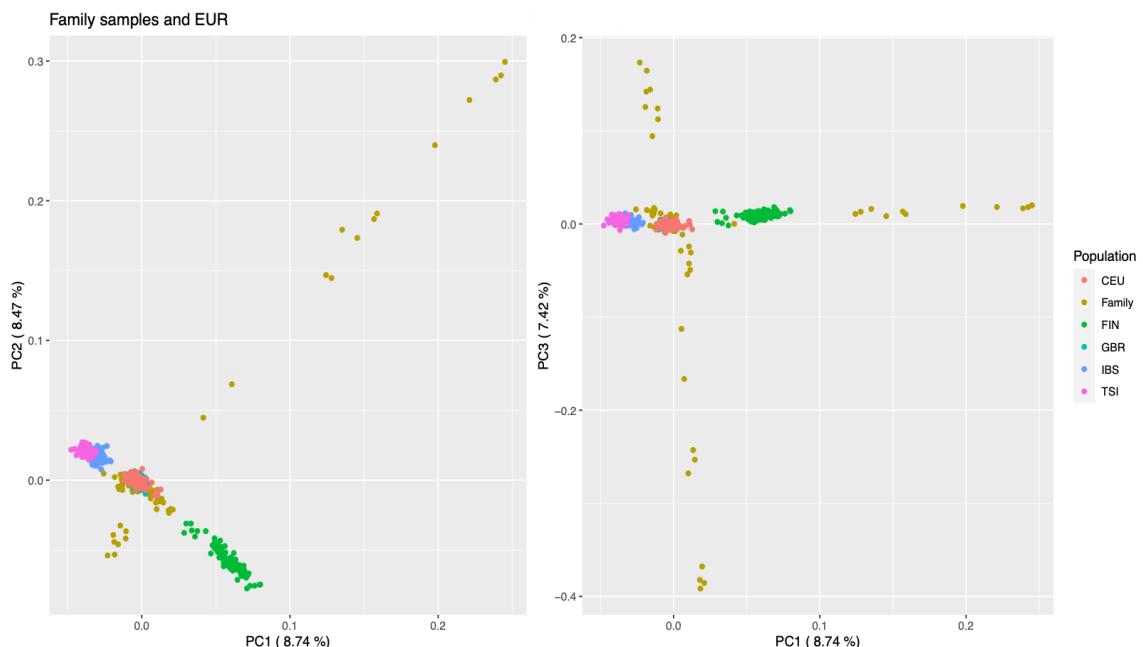


**Figure 2.10** PLINK PCA plot of the first three dimensions, showing the Family samples (Family) grouping with the European subset (EUR).

The other populations are: African (AFR), East Asian (EAS), South Asian (SAS), Ad Mixed American (AMR).

When only the EUR populations of 1000 GP samples was used as a reference, most individuals within the Family samples clustered most closely with CEU and GBR subsets. However, some individuals are widely spread across a diagonal axis, with an apparently small degree of variation in both PC1 and PC2 (Figure 2.11). This likely reflects the presence of related individuals within the Family dataset influencing the clustering, with the first two PCs reflecting the familial groups instead of population structure (Conomos et al. 2018; Thornton et al. 2014; Thornton and Bermejo 2014).

Again, it was possible to perform the same analysis but at a finer scale using only the CEU subgroup (Appendix 2a).



**Figure 2.11 PCA plot of the first three eigenvectors based on PLINK, including European panel and Family samples.**

The PCA shows the Family samples (Family) mainly grouping with the CEU and GBR cluster. Outlier clustering is due to axes reflecting family grouping instead of population substructure.

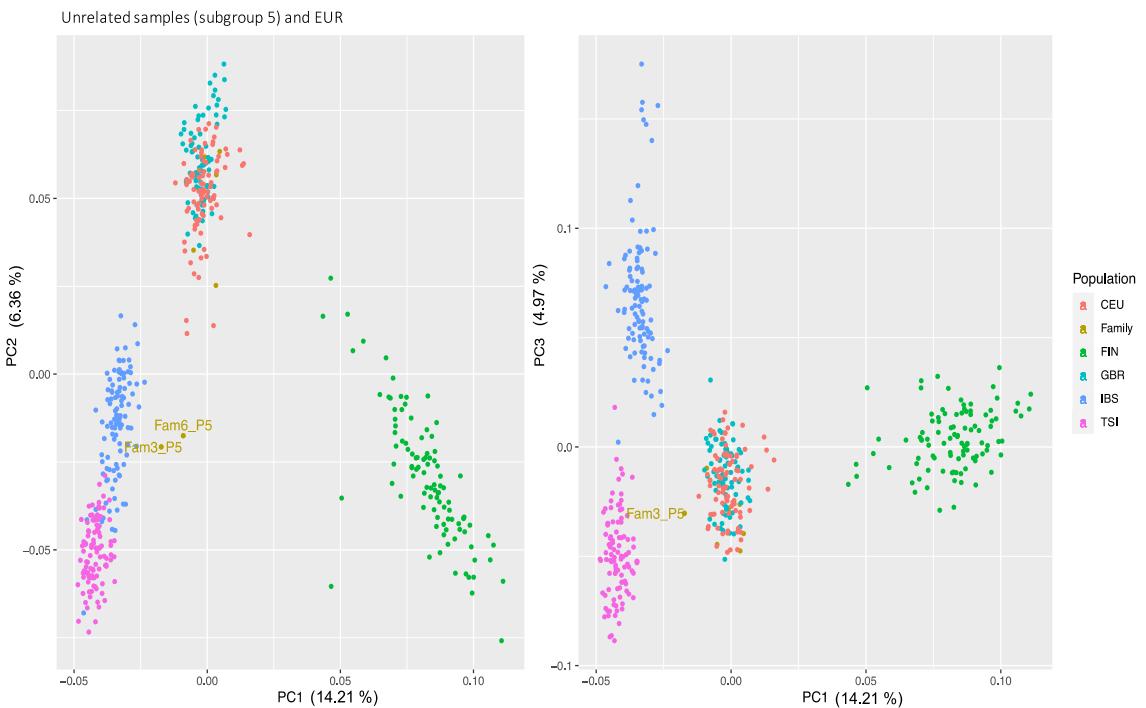
In order to avoid any effects of related individuals on the PCA, subsets of unrelated individuals within the Family dataset were identified. Initially, this subgrouping was led by the output of the Maximum Unrelated Set Identification algorithm in PRIMUS (Appendix 2b), however, the samples identified were still distantly related (second cousin and second cousin once removed) causing spurious grouping in the PCA. For this reason, only small groups of unrelated individuals from the Family dataset were then considered

(Table 2.5). Only subgroup 5 is here reported in Figure 2.12 as an example (Appendix 2c).

**Table 2.5. Subgroups of the family panel used to check for outliers and family structure in the PCA.**

These individuals were chosen to create subsets of completely unrelated individuals in order to perform PCA able to reflect population structure and not family structure.

Subgroup of family panel	Unrelated individuals
1	F1P1, F2P1, F3P1, F4P1, F5P1, F6P1, F7P1, F8P1, F4P17
2	F1P2, F2P2, F3P2, F4P2, F5P2, F6P2, F7P2, F4P16
3	F3P3, F4P3, F5P3, F6P3, F7P3, F8P3, F4P14
4	F1P4, F2P4, F4P4, F5P4, F6P4, F7P4, F8P4
5	F2P5, F3P5, F4P5, F5P5, F6P5, F7P5, F8P5
6	F1P6, F2P6, F4P6, F5P6, F6P6, F7P6, F4P15
7	F4P7, F5P7, F7P7, F8P7, F4P13
8	F3P8, F4P18, F5P8, F6P8, F7P15

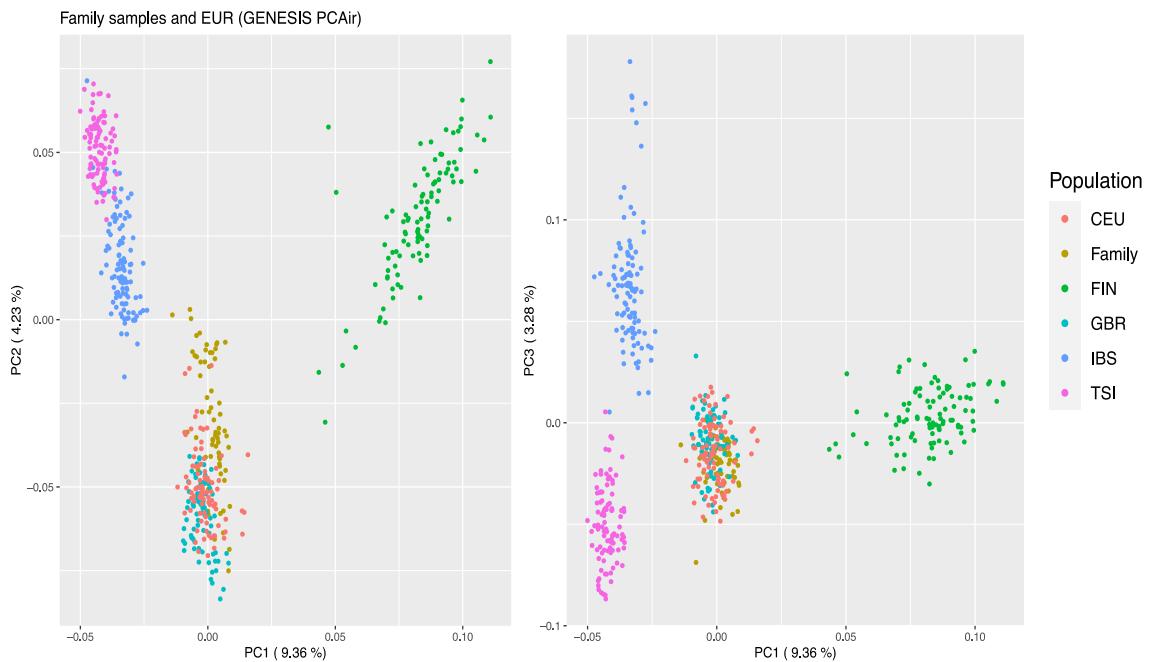


**Figure 2.12 PLINK PCA plot of the first three eigenvectors for the European panel and the subgroup 5 of unrelated individuals (Family) from the Family dataset.**

This shows that the unrelated samples from subgroup 5 mainly group with the CEU and GBR cluster. Two possible outliers are highlighted (F3P5 and F6P5).

From the results, some outliers are identified: F3P5 and F6P5 (visible in Figure 2.12, showing the PCA focusing on subgroup 5), F6P4, F5P4 and F5P7 (identified in the other subgroups, but not shown here).

The indication from the above PLINK-based analysis that there are some outlying individuals in the Family dataset means that it useful to undertake an analysis with GENESIS, which is better suited to accommodating population structure and admixture. Analysis with PC-AiR within GENESIS showed that, as expected, most Family samples cluster with the CEU. However, some show more affinity with the IBS population: six individuals appeared to be outliers from the main family cluster; these included individual 5 from family 3. This confirms that F3P5 is an outlier in ancestry terms, showing some genetic affiliation to the IBS population.



**Figure 2.13 GENESIS PC-AiR plot of the first two PCs of the German-European merged dataset.**

German samples cluster mainly with the GBR and CEU group, however, some individuals appear closer to the IBS group.

The known pedigree structures were checked to find a possible explanation for the samples that appeared to show unexpected population affinity in the PCA analysis. In Family 3, individual 5 is offspring of 3 (typed and confirmed of North-European origin) and 2, who is untyped, and may be reasonably of Iberian origin (relatives as close as full siblings may show different ancestry distributions if their parents come from different populations, Thornton et al. 2012). In Family 5, individual 3 is offspring of typed founder 1, with individual 2 as sibling, who doesn't appear to be an outlier. In Family 6, individual 3 is a founder (appearing to be an outlier) having offspring 5 (outlier) with typed 4 and offspring 2 with untyped parent.

For  $n$  homogeneous population, reliable relatedness estimation would be expected for first, second and many third-degree kin relationships, with overlap between third (i.e. great-grandparental, great-avuncular, first cousin) and fourth degree (i.e. first cousin once removed) relationships and between fourth degree and unrelated (Thornton et al. 2012). These overlaps may increase if population homogeneity is assumed and possible substructure is not considered. A possible outlier was identified when the reference subgroup IBS was included; this shows the importance of including all relevant reference

populations when analysing patterns of variation within a country (Salmela et al. 2008), and the need to choose methods more suitable to accommodate the presence of subgroups (like GENESIS).

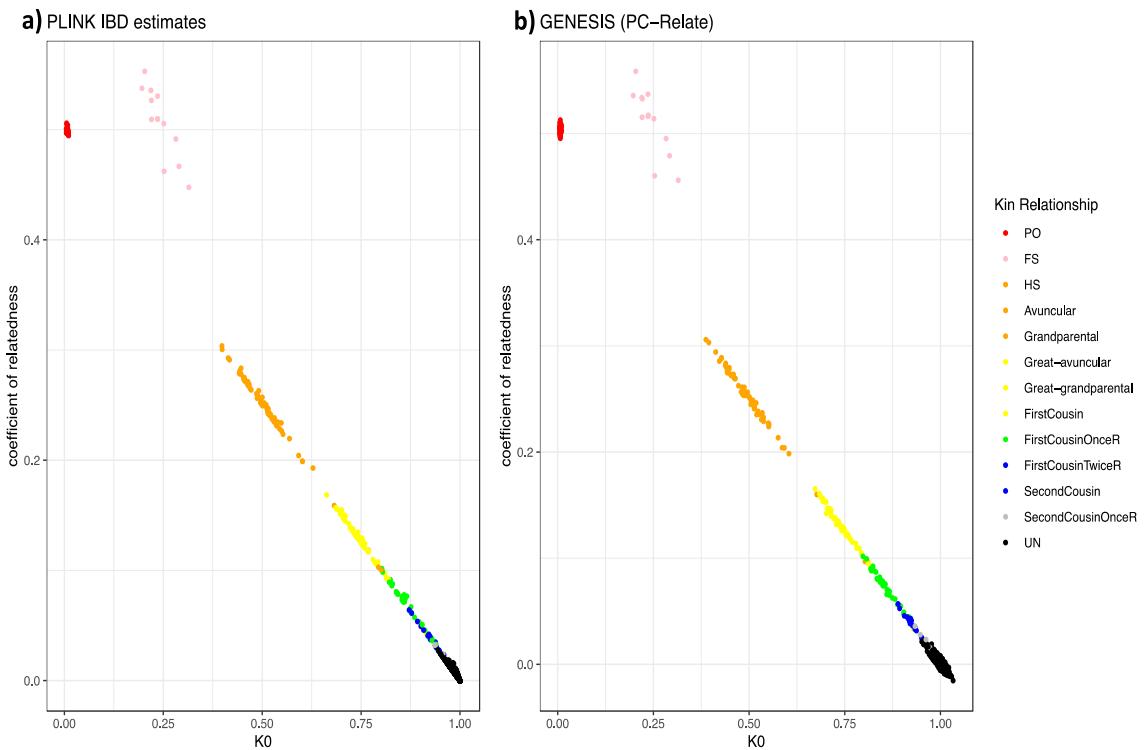
In practice, given the slight heterogeneity in clustering with European reference sub-populations, it was decided to use the whole European panel (EUR) as a reference.

### 2.3.4 Kinship estimation

After identifying the most appropriate population affiliation, it is possible to use the relevant allele frequency information in the estimation of IBD coefficients and to reconstruct pedigree structure.

#### 2.3.4.1 IBD estimation

Estimated IBD coefficients were obtained through PLINK. Estimates of the probability of sharing 0 alleles IBD and estimated proportion of genome shared IBD,  $r$ , were plotted for the dataset containing the Family samples and EUR group as reference dataset, using both PLINK and PC-Relate within GENESIS (Figure 2.14) and for the dataset containing the Family samples and CEU group as reference dataset (data not shown). The GENESIS kinship estimation output (Figure 2.14 b) appears very similar to the PLINK output (a). From Figure 2.14, it is possible to see some avuncular pairs (in orange) and one first cousin pair (in yellow) clustering with more distant relationships (lower estimated relationship coefficient than expected): these pairs are F8P3 and F8P7, F8P4 and F8P5, and F6P1 and F6P5 (half-avuncular due to an half-sibship in the family tree); and 8P4 and 8P7 for first (half)cousin. Stochastic variation may explain another avuncular pair who appears more distantly related: F1P1 and F1P4 in Figure 2.14 b (GENESIS analysis).

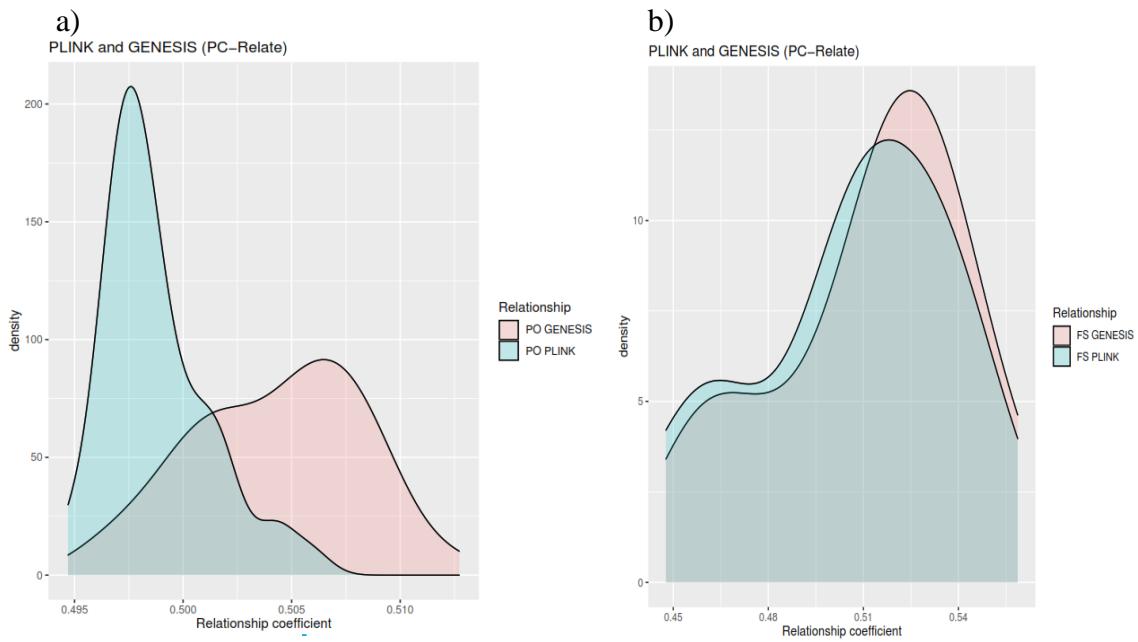


**Figure 2.14 Scatterplot of the estimated coefficient of relationship versus estimated K0 for the Family samples.**

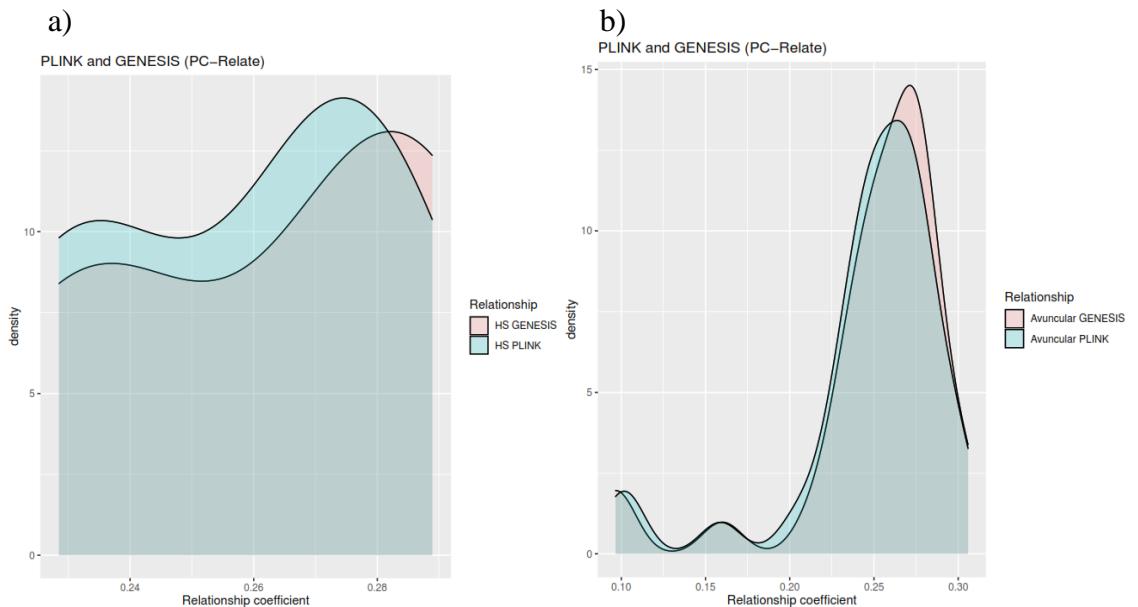
The plots represent the different types of relationships in different colours according to the information from the given pedigree. PO: parent-offspring; FS: full-sibling; HS: half-sibling; UN: unrelated. The European sub-population dataset was used to obtain allele frequencies and was included in the analysis, but not included in the plots (only family pairs are here reported).

(a) As calculated by PLINK; (b) As calculated by GENESIS.

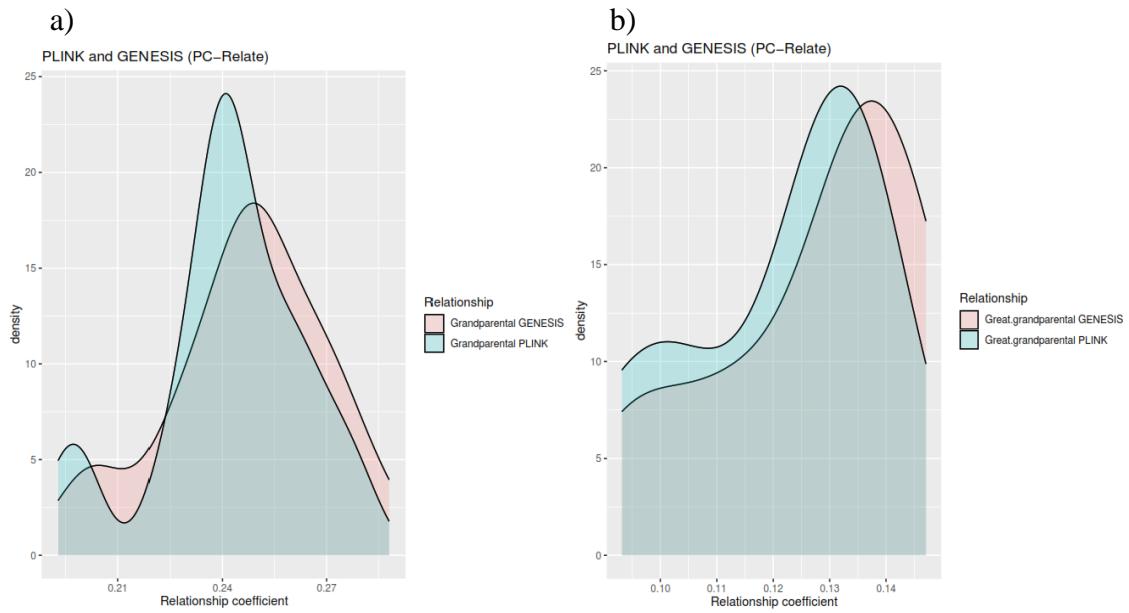
The distribution of the relationship coefficient estimates obtained from PLINK and GENESIS were compared for each relationship type (Figures 2.15 to 2.21). The GENESIS output based on the use of a training set (unrelated samples from the dataset) is not shown here. It is possible to notice, comparing the distribution of observed relationship coefficients, that PO is slightly underestimated by PLINK and over-estimated by GENESIS (the expected  $r$  is 0.5). The mean values for the relationship coefficients estimated in this way by PLINK and GENESIS for each relationship type are summarised in Table 2.6.



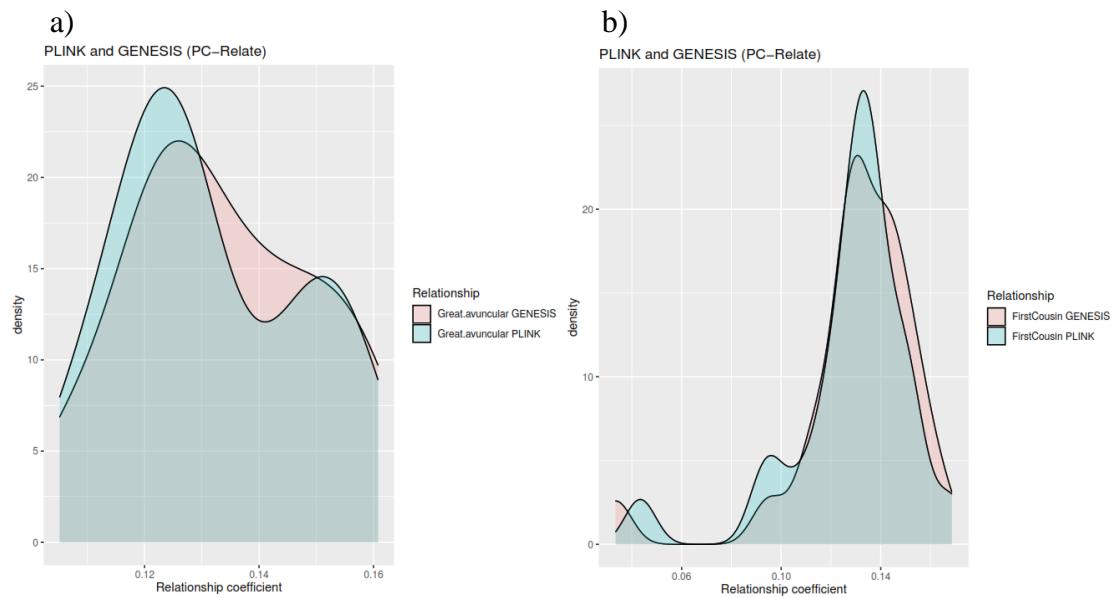
**Figure 2.15 Comparison of relationship coefficient distribution for (a) parent-offspring (PO) and (b) full siblings (FS) relationship based on GENESIS (pink) and PLINK (blue). There are 46 PO pairs, and 13 FS.**



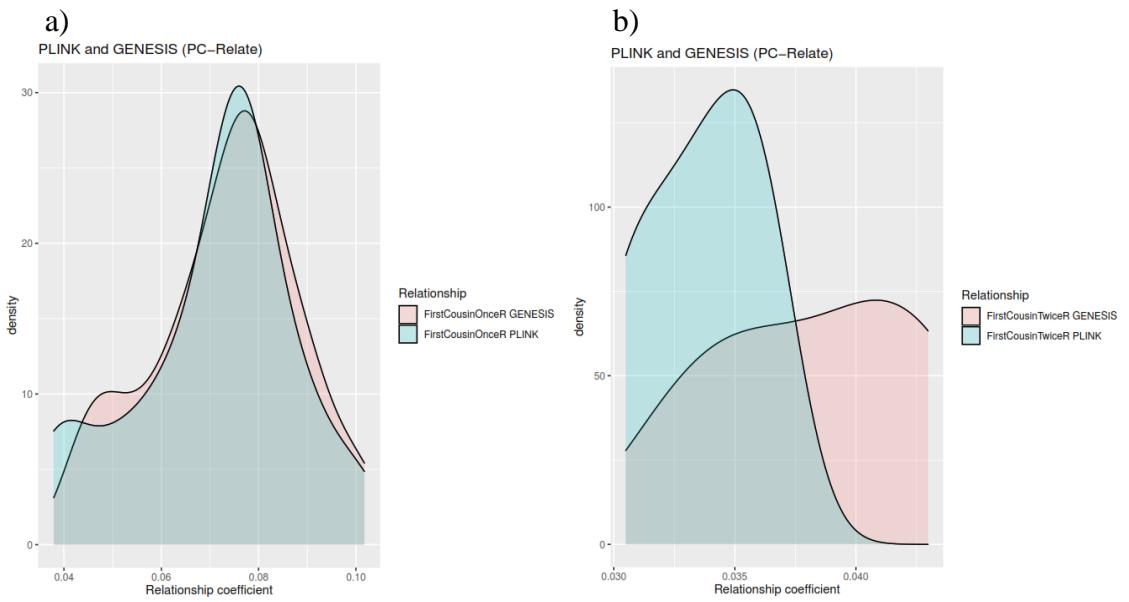
**Figure 2.16 Comparison of relationship coefficient distribution for (a) half-siblings (HS) and (b) avuncular relationship based on GENESIS (pink) and PLINK (blue). There are 5 HS pairs, 34 avuncular.**



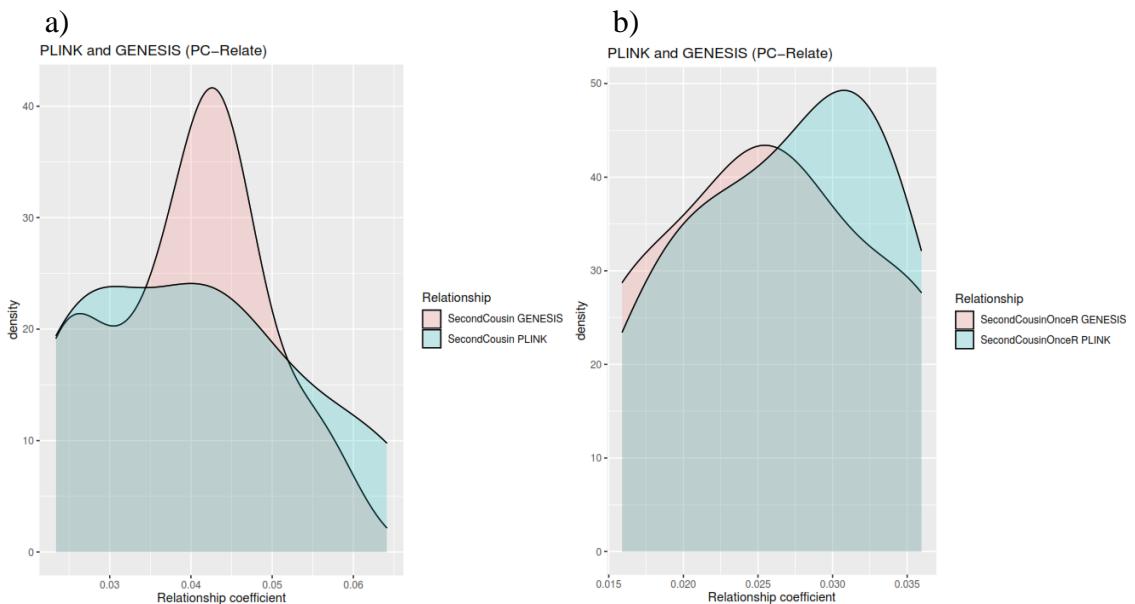
**Figure 2.17 Comparison of relationship coefficient distribution for (a) grandparental and (b) great- grandparental relationship based on GENESIS (pink) and PLINK (blue). There are 27 grandparental and 6 great-grandparental pairs.**



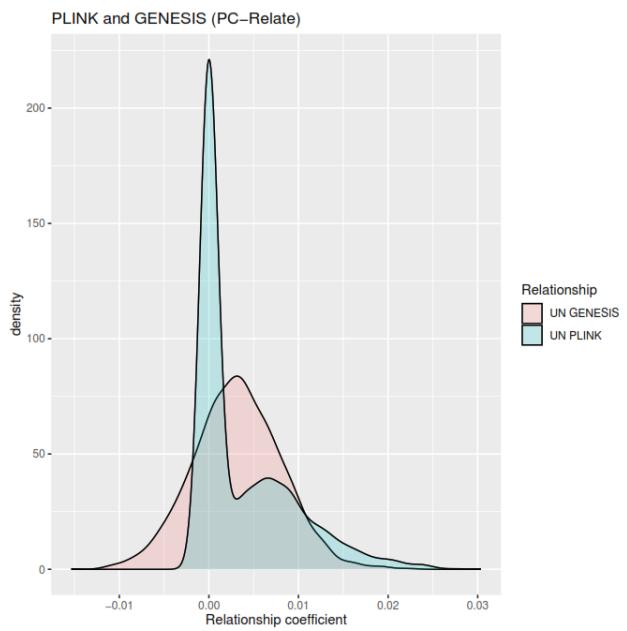
**Figure 2.18 Comparison of relationship coefficient distribution for (a) great-avuncular and (b) first cousin relationship based on GENESIS (pink) and PLINK (blue). There are 18 great-avuncular and 24 first cousin pairs.**



**Figure 2.19 Comparison of relationship coefficient distribution for (a) first cousin once and (b) twice removed relationship based on GENESIS (pink) and PLINK (blue).**  
There are 32 first cousin once and 4 twice removed relationship pairs.



**Figure 2.20 Comparison of relationship coefficient distribution for (a) second cousin and (b) second cousin once removed relationship based on GENESIS (pink) and PLINK (blue).**  
There are 14 second cousin and 4 second cousin once removed pairs.



**Figure 2.21 Comparison of relationship coefficient distribution for unrelated individuals based on GENESIS (pink) and PLINK (blue).**

There are 71 UN pairs.

**Table 2.6 Relationship coefficient estimated by PLINK and GENESIS.**

This table shows the expected relationship coefficient followed by the mean of the coefficient estimated for each relationship type by PLINK and GENESIS.

Kin relationship class	Relationship coefficient	PLINK		GENESIS	
		Mean	Standard deviation	Mean	Standard deviation
<b>Parent-offspring</b>	0.5	0.4987413	0.00241804	0.50432387	0.00392996
<b>Full-sibs</b>	0.5	0.50663077	0.03188632	0.51171226	0.03097783
<b>Half-sibs</b>	0.25	0.25814	0.02476334	0.26342256	0.02763591
<b>Grandparental</b>	0.25	0.24241111	0.02172277	0.24687079	0.02262032
<b>Avuncular</b>	0.25	0.24754706	0.04657517	0.25039196	0.04769485
<b>Great-grandparental</b>	0.125	0.12126667	0.01599379	0.1335117	0.01628008
<b>Great-avuncular</b>	0.125	0.13148333	0.01770318	0.12712892	0.01919608
<b>First cousin</b>	0.125	0.12812083	0.0250252	0.13048741	0.02612626
<b>First cousin once removed</b>	0.0625	0.07105625	0.01649221	0.07319066	0.01510455
<b>First cousin twice removed</b>	0.03125	0.033675	0.00255653	0.03834189	0.00446097
<b>Second cousin</b>	0.03125	0.04039286	0.01333212	0.03925751	0.01013495
<b>Second cousin once removed</b>	0.015625	0.026725	0.00686458	0.02580232	0.00839745
<b>Unrelated</b>	0	0.00377668	0.00540258	0.0031942	0.00494156

### 2.3.4.2 Pedigree reconstruction

After estimating the relatedness level among the samples, more detailed information on the pedigree structure was obtained using PRIMUS. In the first automatic stages of data processing (“prePRIMUS” using PLINK v1.9), an unrelated set of 23 individuals was estimated and merged with the full HapMap3 panel (overlapping with but not identical to the 1000 GP panel), obtaining a dataset of 388,737 variants and 970 individuals (464 males, 506 females). PCA analysis was performed using both the European French (CEU) and European Italian (TSI) subgroups as combined reference population (automatically selected).

When the degree relatedness cutoff was set to the default 3 (i.e. considering relationships up to 3<sup>rd</sup> degree, as defined by PRIMUS), larger and more complex families (families 4 and 7) were not processed, and there were problems with the reconstruction of three other families (1, 2, and 6):

- Family 4 includes 17 individuals (34, including not-genotyped individuals) and a total of 132 relationships (the most distant being second cousin), with about 50% of missing samples (as two untyped individuals do not contribute to the pedigree structure, Figure 2.22 a). With a 3<sup>rd</sup>-degree cutoff, the output pedigree closest to the true pedigree reports untrue relationships among the untyped individuals.
- Family 7 includes 13 individuals (29, including not-genotyped individuals) and a total of 132 relationships (the most distant being second cousin once removed), with about 55.2% of missing samples (Figure 2.23 b). With the 3<sup>rd</sup>-degree cutoff, the true pedigree is the 4<sup>th</sup> output pedigree.
- For families 1 and 2, the true pedigree was not represented by the first network in the PRIMUS outputs (3<sup>rd</sup> and 16<sup>th</sup> network, respectively).
- In family 6, two generations of untyped individuals were not adjusted (adding missing individuals) by PRIMUS, causing wrong directionality in a parent-offspring (PO) couple (individuals 4 and 6).

Choosing a more stringent degree cut-off (from 3rd to 2nd degree), reconstructions for all pedigrees were obtained (Table 2.7; Appendix 2d):

- Among the large families, the correct pedigree for family 4 is not the first pedigree with highest likelihood, but the second (Figure 2.22 b), and family 7 is partially reconstructed, the pedigree being divided into two (correct) sub-networks (Figure 2.23).
- Family 1 is correctly reconstructed and the true pedigree appears listed as the first output.
- Family 2, even if presenting a structure very similar to that of family 1, is incomplete, as individual 4 is not included in the network.
- The output for family 6 presents the same directionality issue as previously, but since 3<sup>rd</sup> degree is the maximum relationship considered, individual 8 is dropped out from the network. In fact, since individual F6P7, who was the connection

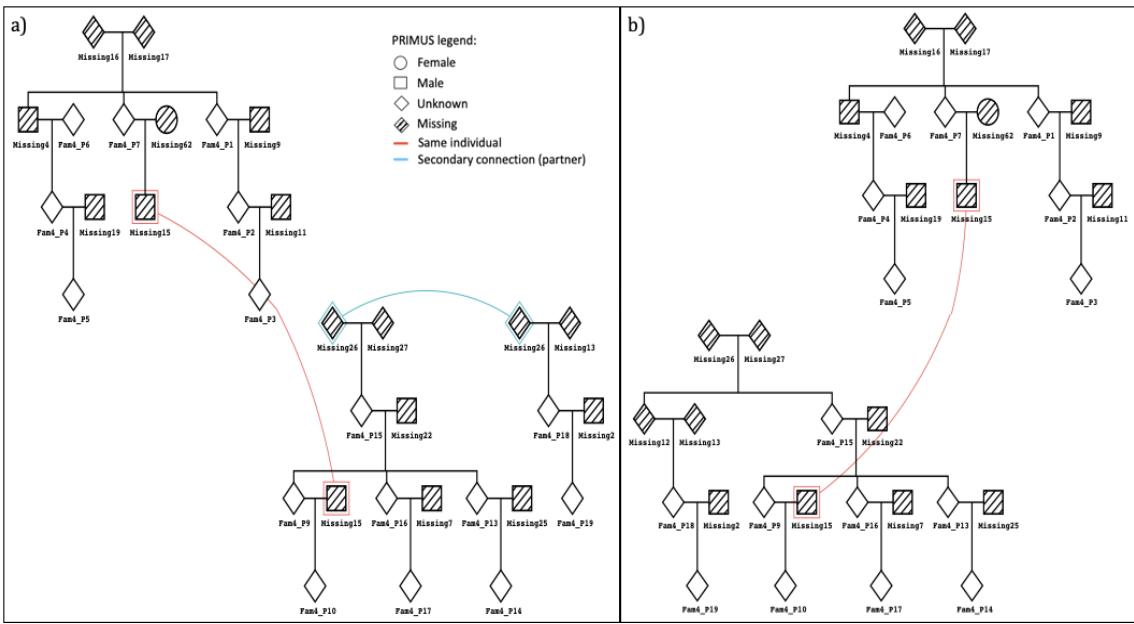
between individuals F6P8 and the rest of the family, is excluded from the analysis, PRIMUS cannot add F6P8 in the network. This creates some additional issues, and PRIMUS is unable to distinguish the directionality of this parent-offspring relationship.

Missingness had an impact on pedigree reconstruction. In family 6 (Figure 2.4 a), there are two generations missing (individual 7 being excluded) and individual 8 is considered as the most distantly related individual (appearing in most unrelated sub-sets). Without this information, PRIMUS cannot assign directionality to the parent-offspring relation of individual 6 and 4. In family 2, individuals 1 and 2 share less genetic information than expected with individual 4, who is often included in the maximum unrelated individuals set, and is, therefore, considered unrelated and excluded. In family 4, the kin relationship between individual 15 and 18 is avuncular, but, since the expected sharing is the same for the half-sibling relationship as well, PRIMUS selects the relationship with less missing samples (half-sib) and so does not include the missing generation required for the true pedigree. Individual 18 and 10, as well as individuals 2 and 5 are first cousins once removed, but present higher IBD sharing than expected, explaining the higher inter-relatedness proposed by the PRIMUS reconstructions.

**Table 2.7 PRIMUS networks reconstruction summary.**

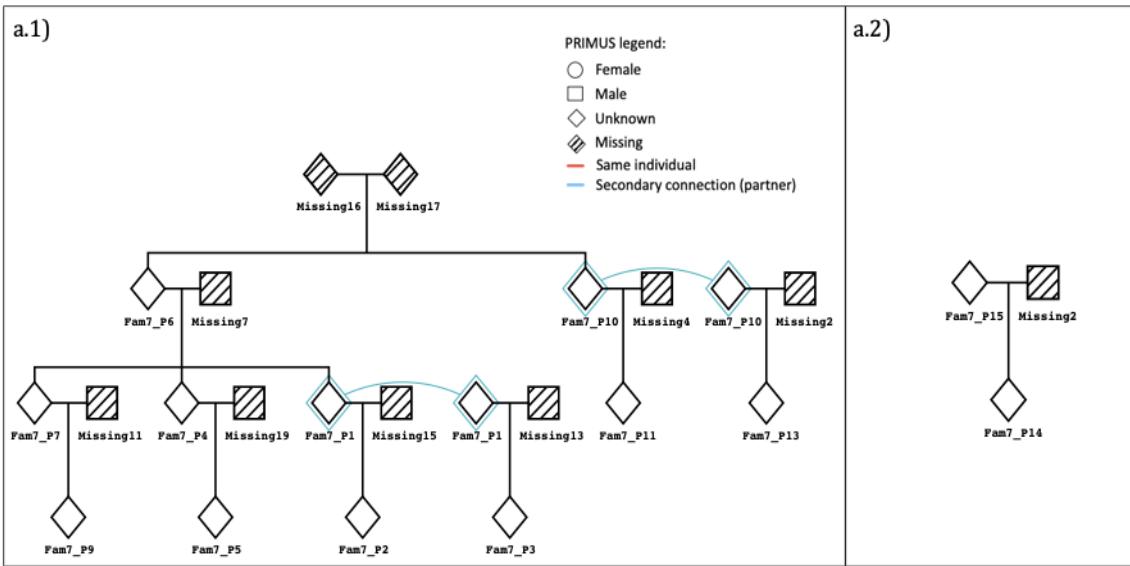
This table shows the number of networks created for each group, which of them gives the highest scores and which samples and how many are included in each network.

<b>Network name</b>	<b>Number of reconstructed pedigrees</b>	<b>Number of samples included by PRIMUS</b>	<b>Samples IDs</b>
<b>1</b>	1	5	F1P6, F1P9, F1P4, F1P1, F1P2
<b>2</b>	1	4	F2P2, F2P1, F2P6, F2P5
<b>3</b>	3	7	F3P1, F3P3, F3P8, F3P5, F3P11, F3P10, F3P2
<b>4</b>	54	16	F4P19, F4P4, F4P16, F4P17, F4P2, F4P3, F4P18, F4P10, F4P7, F4P5, F4P6, F4P13, F4P9, F4P14, F4P15, F4P1
<b>5</b>	1	8	F5P8, F5P6, F5P1, F5P4, F5P5, F5P7, F5P3, F5P2
<b>6</b>	6	6	F6P6, F6P3, F6P5, F6P4, F6P2, F6P1
<b>7</b>	1	11	F7P13, F7P11, F7P1, F7P4, F7P7, F7P10, F7P9, F7P3, F7P2, F7P6, F7P5
<b>8</b>	2	2	F7P14, F7P15
<b>9</b>	8	5	F8P5, F8P1, F8P4, F8P3, F8P7



**Figure 2.22 PRIMUS output for family 4 (relatedness cutoff at 2nd degree).**

Males are shown as squares, females as circles and individuals of unknown sex as rhombi; missingness is represented by a shaded figure, and the same person belonging to different familiar nuclei is highlighted in red. Untyped individuals are named as “Missing” followed by a number; typed individuals are named as “Fam” followed by the family number, “P” and individual number (e.g. individual 1 from family 1 will be “Fam1\_P1”). a) This pedigree is the first output (highest score), but does not completely and correctly fit the true pedigree of family 4. b) The PRIMUS output (second highest scoring pedigree) that corresponds to the true pedigree (Figure 2.4).



**Figure 2.23 PRIMUS output for family 7 (relatedness cutoff at 2nd degree).**

Males are shown as squares, females as circles and individuals of unknown sex as rhombi; missingness is represented by a hatched figure; the same person belonging to different familiar nuclei is highlighted in red, while an individual with different partners is highlighted in blue. Untyped individuals are named as “Missing” followed by a number; typed individuals are named as “Fam” followed by the family number, “P” and individual number (e.g. individual 1 from family 1 will be “Fam1\_P1”). a) Family 7 has been sub-divided by PRIMUS into two outputs (a.1 and a.2).

Because missingness of samples in the family network structure influences the PRIMUS output, another family dataset was created including two poor-quality samples that were previously excluded due to failed QC (i.e. individual 3 from family 2, and individual 7 from family 6). For family 6, the first output still presents directionality issues (parent-offspring are connected, but assigning the wrong generation) and it is unable to connect individual 8 due to high levels of missingness; output three is the closest to the true pedigree but a generation with no typed individuals is still (unsurprisingly) not included. In Family 2, individuals 3 and 4 are still excluded: individual 3 does not even appear in the unrelated sets produced by PRIMUS.

Including a file with hypothesised ages of individuals (taking the generation that each individual belongs to, and assigning all within the same generation the same age, namely, 70, 50, 30, 20), and cutoff of 2nd degree, the directionality issue in family 6 could be solved (not shown here).

In summary, pedigree reconstruction via PRIMUS based on the autosomal SNP data could be carried out with some limitations due to the level of missing individuals in pedigrees: this led to assigning some samples to the wrong generation and splitting pedigrees into smaller familial nuclei.

#### **2.3.4.3 Inclusion of uniparentally-inherited and sex-linked markers**

In principle, the inclusion of MSY, mtDNA and X-chromosome data can help in an analysis when it is not possible to reach a conclusive output with only autosomal markers: SNPs within all three of these loci are included in the SNP chip typed here. In the analysed dataset, there are no ambiguous relationships that cannot be solved through autosomal markers only - in particular, patrilines and matrilines within the pedigrees are short. However, it is useful to consider how these loci can be analysed and under what circumstances they could contribute to resolving pedigree structures.

**MSY SNP analysis.** Using the tool yHaplo, Y-chromosome haplogroups were identified in 28 males in the German families: of 1506 Y-SNPs on the chip, 529 were in the ISOGG phylogeny and therefore usable by yHaplo. Haplogroups were consistent with the family reconstruction based on autosomal markers and with the known pedigree structure (Figure 2.4). Y haplogroups are highly structured geographically, so it is worth asking whether the observed haplogroups were typical of a German population; to do this, the results were compared with published data on two German samples (Mecklenburg, n= 131; western Bavaria, n= 218) (Rębała et al. 2013). Since the most derived Y-SNP identified by yHaplo was not always typed in the published study, it was necessary to use the ISOGG phylogenetic tree (<https://isogg.org/tree/>) and phylotreeY to identify equivalent clades, or to consider haplogroups at a lower resolution (van Oven et al. 2014). The haplogroups found in the German pedigrees, and their frequencies in the published study (Rębała et al. 2013), are reported in Table 2.8. All haplogroups found here were also seen in the published German data, although some at very low frequency.

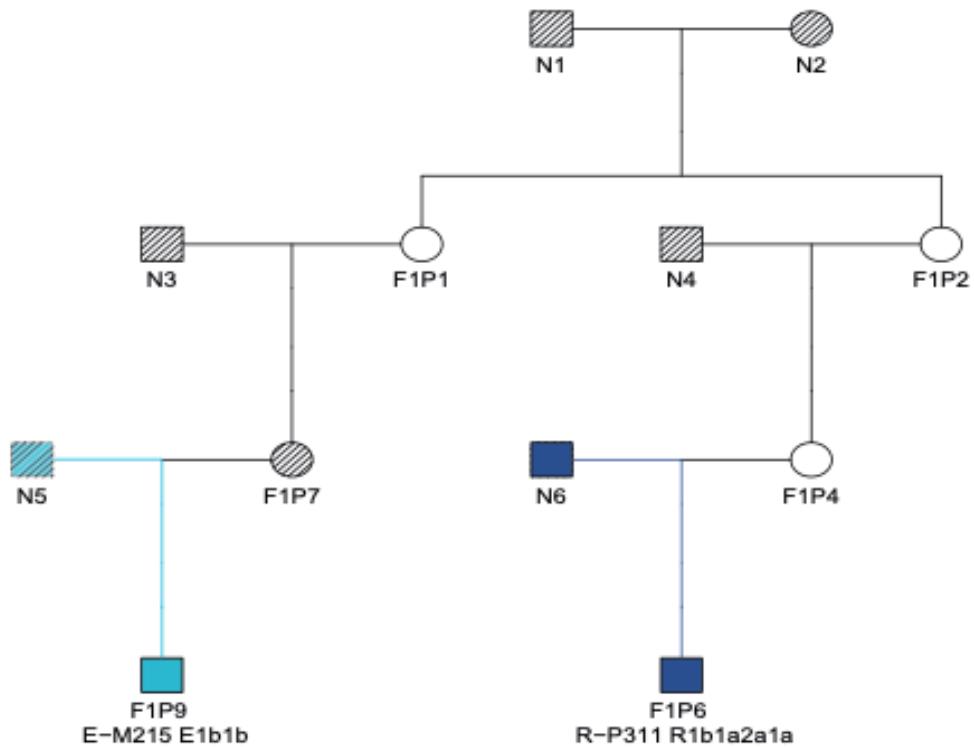
**Table 2.8 Y haplogroups in the family set.**

The haplogroups were identified by yHaplo. The population frequency is based on the set of 349 German samples (Mecklenburg and Bavaria) from (Rębała et al. 2013), and required some adjustment of haplogroup resolution to allow equivalent categories to be defined.

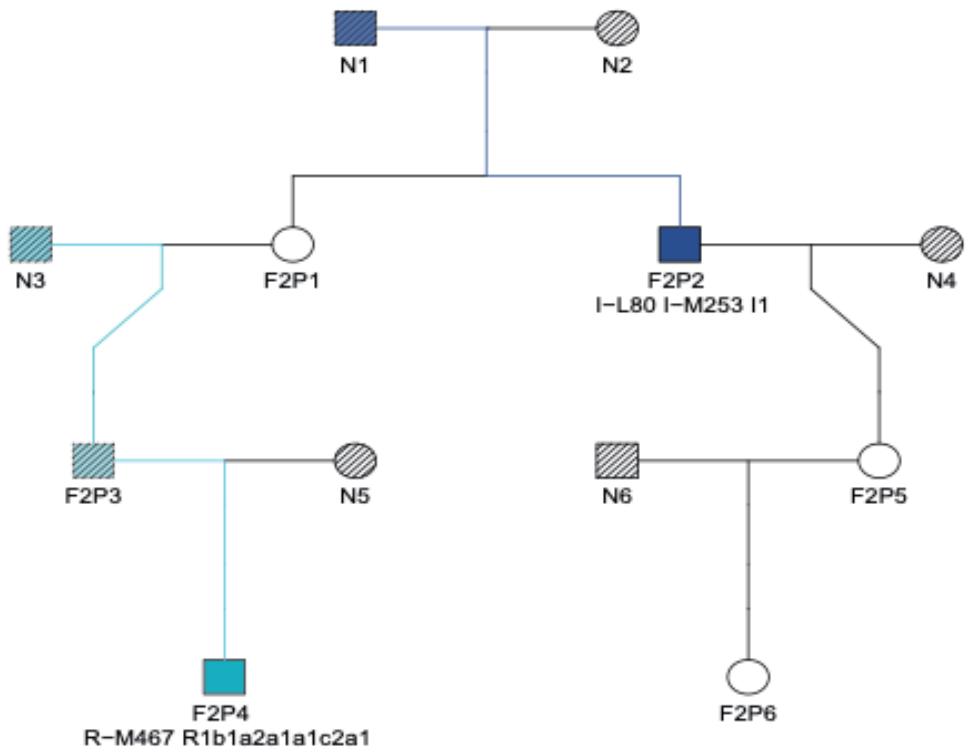
Haplogroup from yHaplo	Total count (no. of independent occurrences)	Equivalent hg in Rebala et al. (2013)	Frequency in Rebala et al. (2013)
E-M215 R1b1a2a1a	1 (1)	E-M35	0.072
R-P311 R1b1a2a1a	8 (4)	R-P311	0.438
I-L80 I-M253 I1	7 (4)	I-M253	0.060
R-M467 R1b1a2a1a1c2a1	1 (1)	R-U198)	0.003
G-L43 G2a2b2a1a1b1a1	1 (1)	G-P15	0.026
R-L23 R1b1a2a	1 (1)	R-L23 *	0.023
T-M70 T1a	5 (2)	T-M70	0.009
R-M417 R1a1a1	4 (2)	R-M17	0.129

Although not helpful in these particular pedigrees, the Y chromosome may prove useful, for example, in deficiency paternity testing cases, and when there is a need to distinguish the maternal and paternal kinship for two males. The Y haplogroup inheritance in the eight analysed families is reported in Figure 2.24. For example, in family 6, if individual 2 was missing for testing and the paternity of individual 1 was questioned (refer to Figure 2.24), it would be possible to test the grandfather (individual 3) to support a possible kin relationship according to the Y haplogroup. As another example, in family 3, it would be possible to clearly confirm that individuals 11 and 12 are half-siblings and not full siblings. Evidential strength in comparisons such as these also depends on the population frequency of the Y haplogroups in question, and in practice Y haplogroups are more powerful for exclusions than inclusions (Jobling et al. 1997).

a)

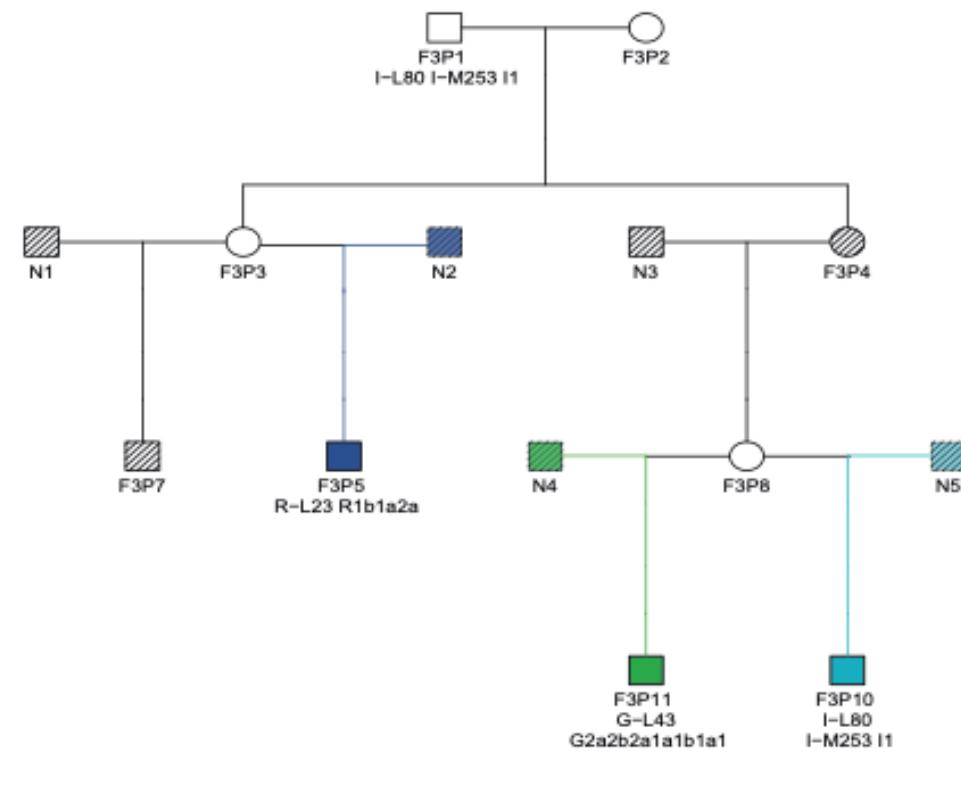
**Family 1**

b)

**Family 2**

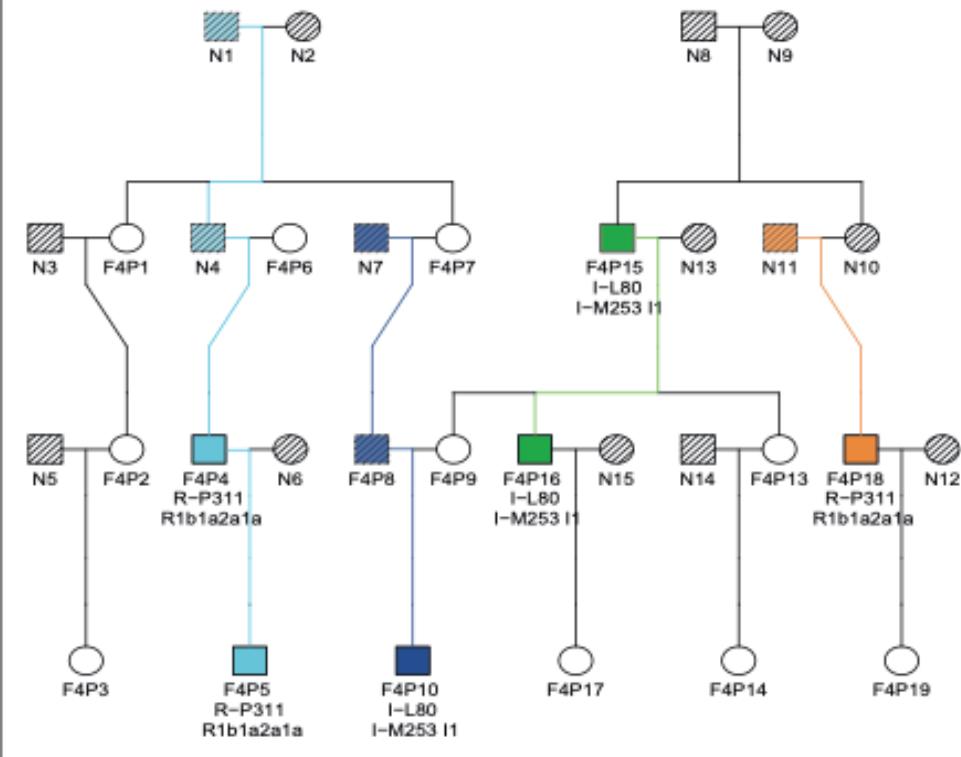
c)

Family 3



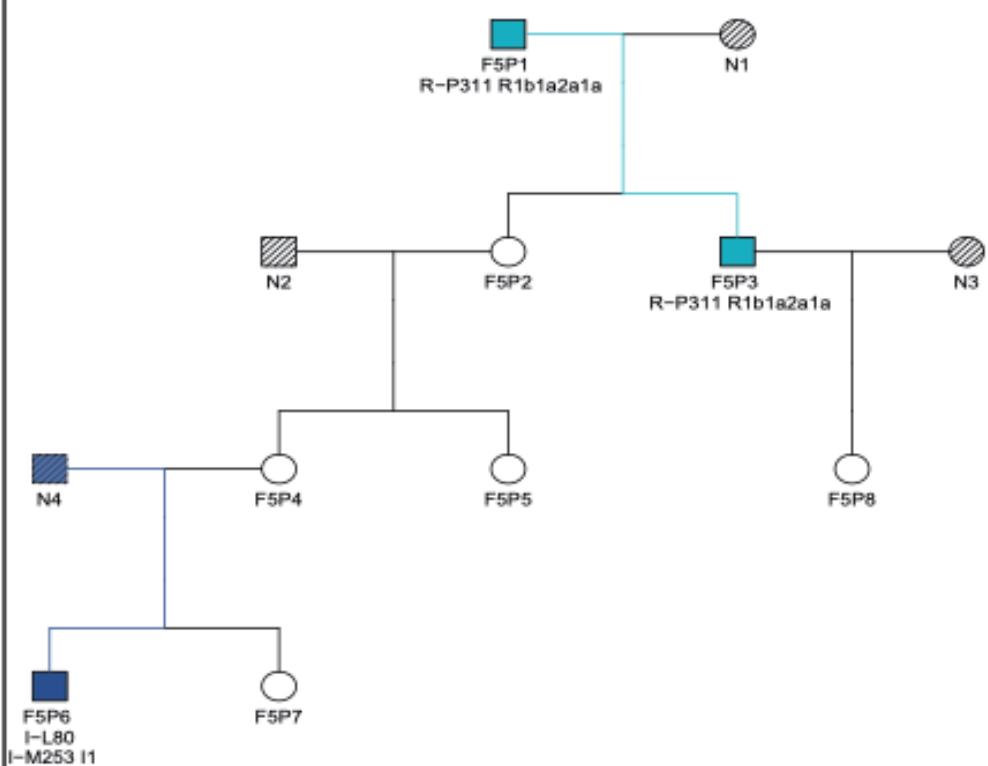
d)

Family 4



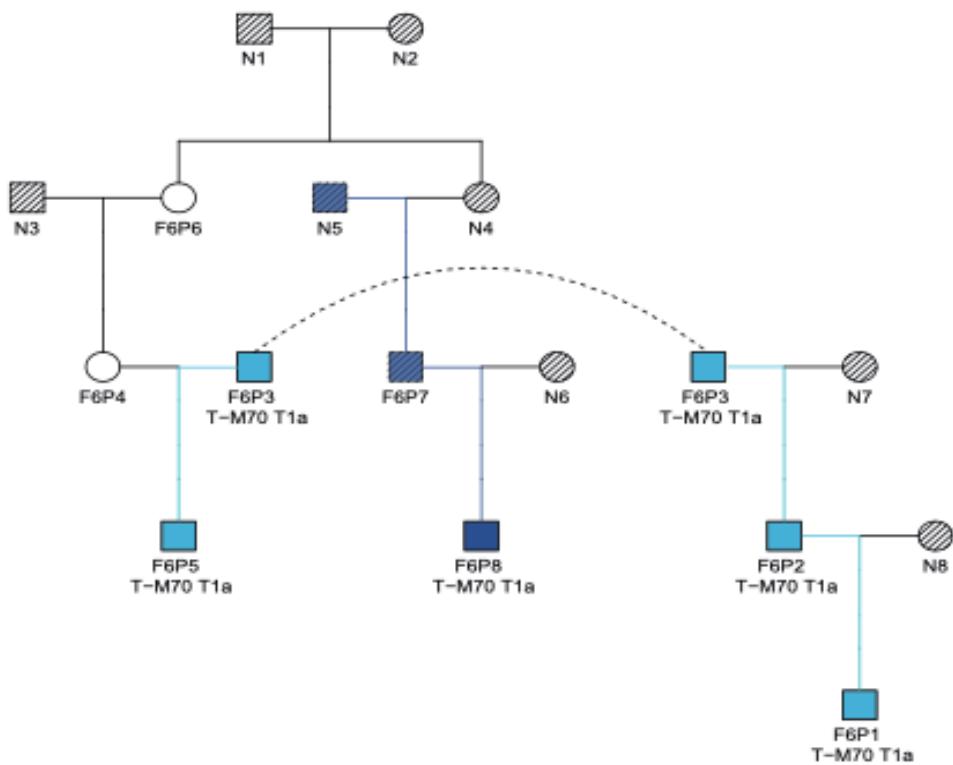
e)

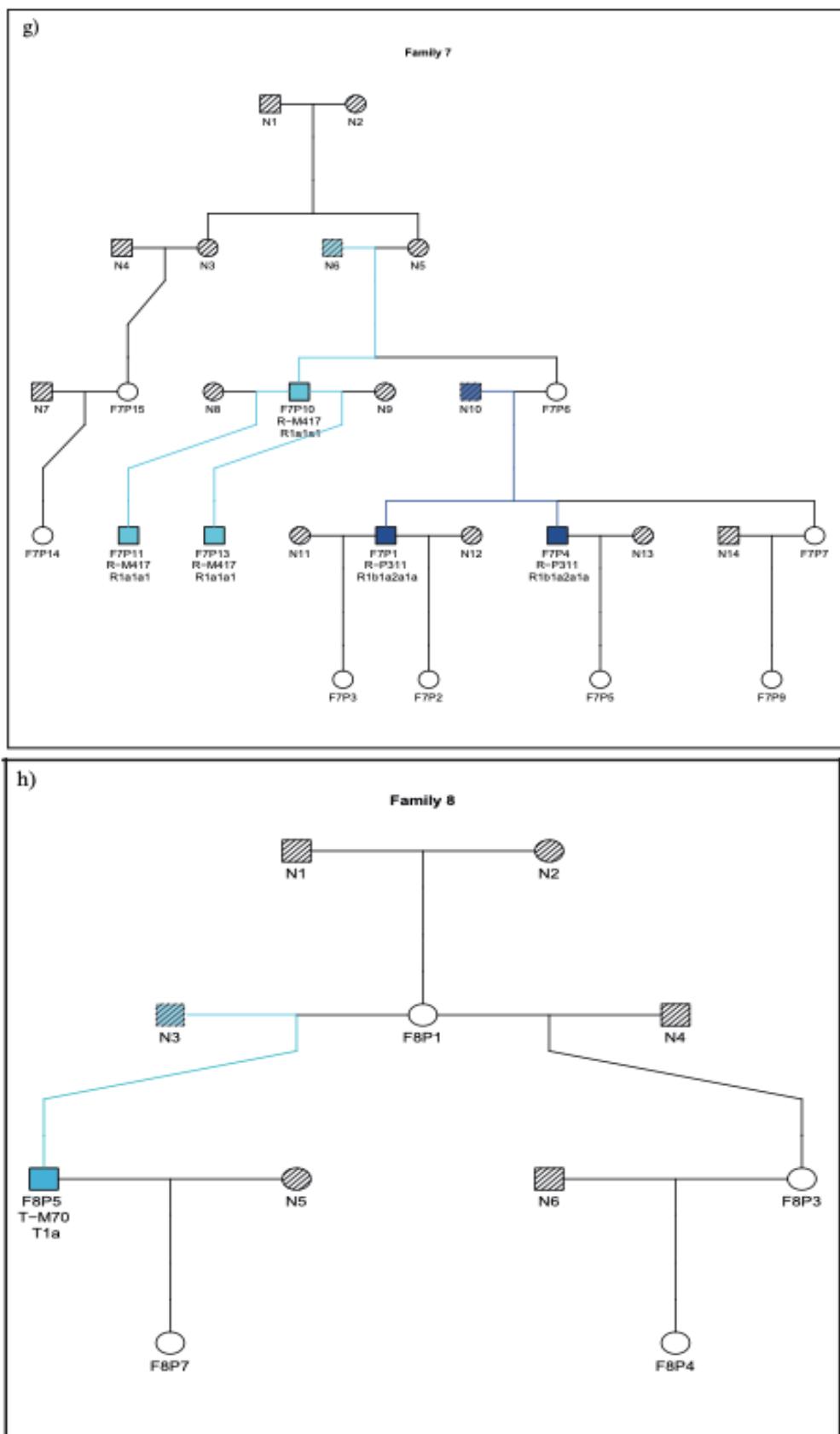
## Family 5



f)

## Family 6





**Figure 2.24 Inheritance of Y haplogroups in the eight families analysed.**

Haplogroups identified through yHaplotype are reported under each individual. The available, uninterrupted paternal lines are highlighted in different colours. The fact that many male founders were not available for testing is clear.

**mtDNA SNP analysis.** Using Haplogrep 2, mtDNA haplogroups were determined from SNP chip data in 66 individuals in the German pedigrees (Figure 2.25). For seven of the eight pedigrees, patterns of mtDNA inheritance are as expected given the known pedigree structures. However, mtDNA haplogroups in one family show patterns that are inconsistent with the known relationships. The haplogroup of individual 7 in family 4 (H26a1b) is different from that of her sister, individual 1 (H2a2a1): this individual (7) carries the variant 16390A, which is not present in the family female line (individual 1, 2 and 3). Also in family 4, there is incompatibility between a mother (individual 6; H1e1a4) and son (individual 4; R8a1a1d). Individual 4 carries the variant 16390A as well as 15326G! (sharing only 709A with individual 6). This variant is also observed in six individuals within family 3, where there are haplotypes that carry it on more than one haplogroup background. It is not among the top 15 hypervariable variants (van Oven 2015b), but it does lie in the control region, where mutation rate is generally elevated about ten-fold above the rest of the mitochondrial genome (Soares et al. 2009), and in these samples it appears to be a recurrent mutation. An alternative explanation would be uncertainty in SNP calling at this site. Notably, if the haplogroup calls here were taken at face value, they would exclude some relationships that are supported by autosomal data and other information on the pedigrees.

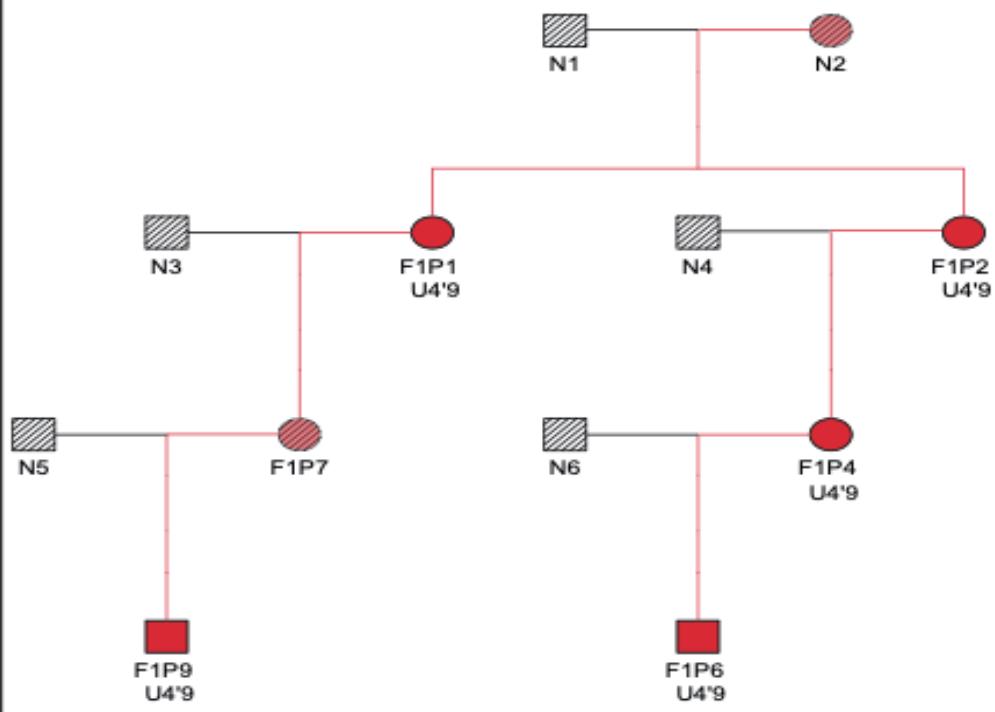
Lineages of mtDNA are not as highly structured as those of the Y chromosome in Europe (Batini et al. 2017; Richards et al. 2000). Nonetheless, we can compare the haplogroups observed here with a European dataset (total n= 260; Serbian, n= 20; Bavarian, n= 20; Irish, n= 20; French (CEU), n= 20; Norwegian, n= 20; Frisian, n= 20; Danish, n= 20; Greek, n= 20; Hungarian, n= 20; Orcadian, n= 20; Italian (TSI), n= 20; Basque, n= 20; English, n= 20; Irish, n= 20; Batini et al. 2017) (Table 2.9).

**Table 2.9 Table listing the haplogroups in the family.**

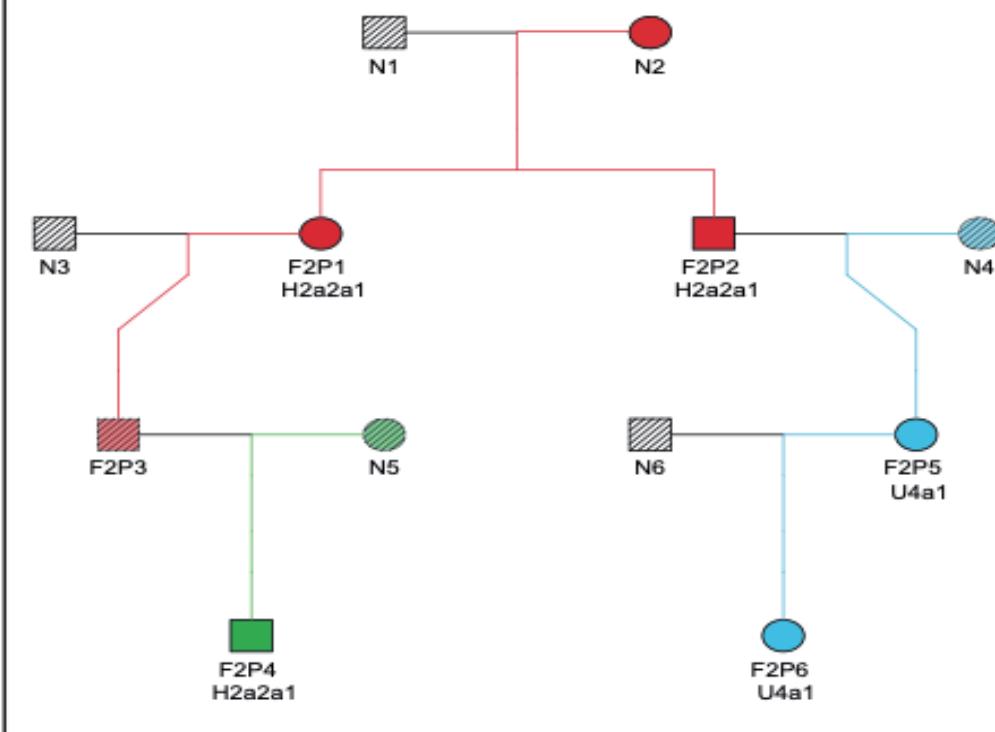
Occurrence of the haplogroup was compared to Batini et al. 2017, which included European populations.

Haplogroup	Total count (no. of independent occurrences)	Branch	Equivalent haplogroup in Batini et al. (2017)	Frequency in Batini et al. (2017)
H2a2a1	19 (6)	H2a2a	H2a2a1a, H2a2a1	0.015
K1a1c	8 (1)	K1a	K1a	0.012
U4 <sup>9</sup>	7 (3)	U4 <sup>9</sup>	U4c1a, U4a2, U4a, U4b1a1a1, U4a2a	0.031
U3a	6 (1)		Not observed	0
H6a1b2e	6 (1)	H6a	H6a1b4, H6a1a	0.012
J1c8	5 (1)		J1c8a, J1c	0.008
X2	4 (1)	X2	X2c1, X2b, X2b+226, X2b8	0.027
T1a1b	3 (1)	T1a	T1a1, T1a	0.023
U4a1	2 (1)	U4a	U4a	0.008
R8a1a1d	1 (1)	R8	Not observed	0
H1e1a4	1 (1)	H1e	H1e1a, H1e2d, H1e	0.015
H26a1b	1 (1)		Not observed	0
J1c	1 (1)	J1c	J1c2q, J1c2e1, J1c5a1, J1c4, J1c1b, J1c3a1, J1c8a, J1c5, J1c7a, J1c, J1c3e1, J1c1b2, J1c3e2, J1c2l, J1c3g, J1c3b1a, J1c3, J1c2e2	0.088
J	1 (1)	J	J1c2q, J1c2e1, J1c5a1, J1c4, J1c1b, J1c3a1, J1c8a, J1c5, J1c7a, J1c, J2a1a1a3, J1c3e1, J1c1b2, J2a2b2, J1c3e2, J1c2l, J1b1a1a, J1c3g, J1c3b1a, J2b1c, J1b2, J1c3, J1c2e2	0.108
T2b7	1 (1)	T2	T2b, T2	0.027

a)

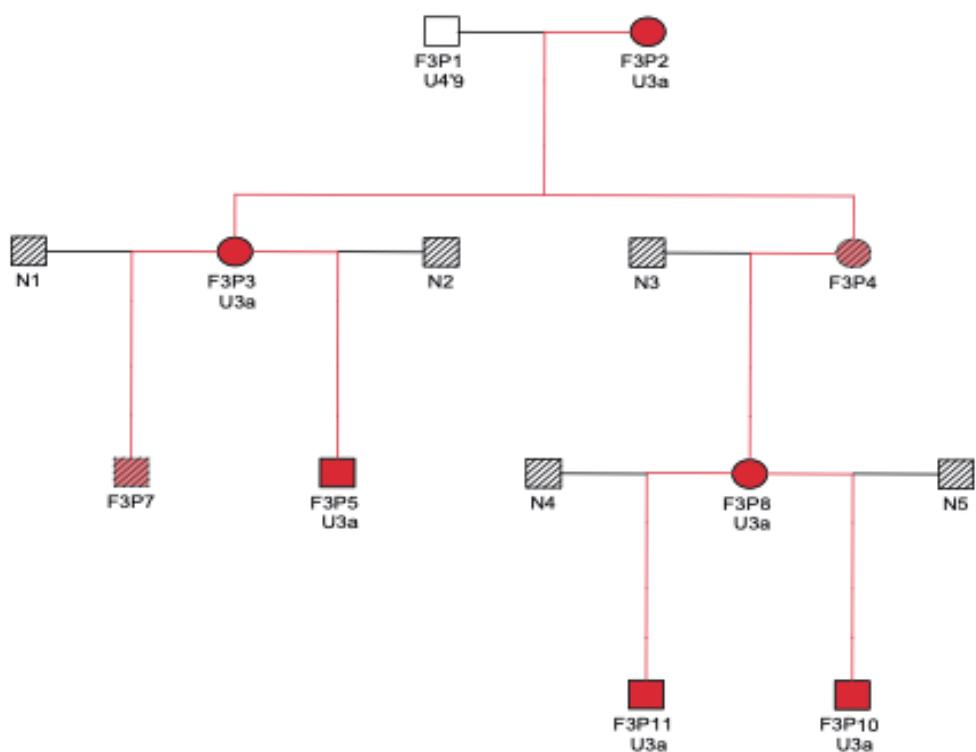
**Family 1**

b)

**Family 2**

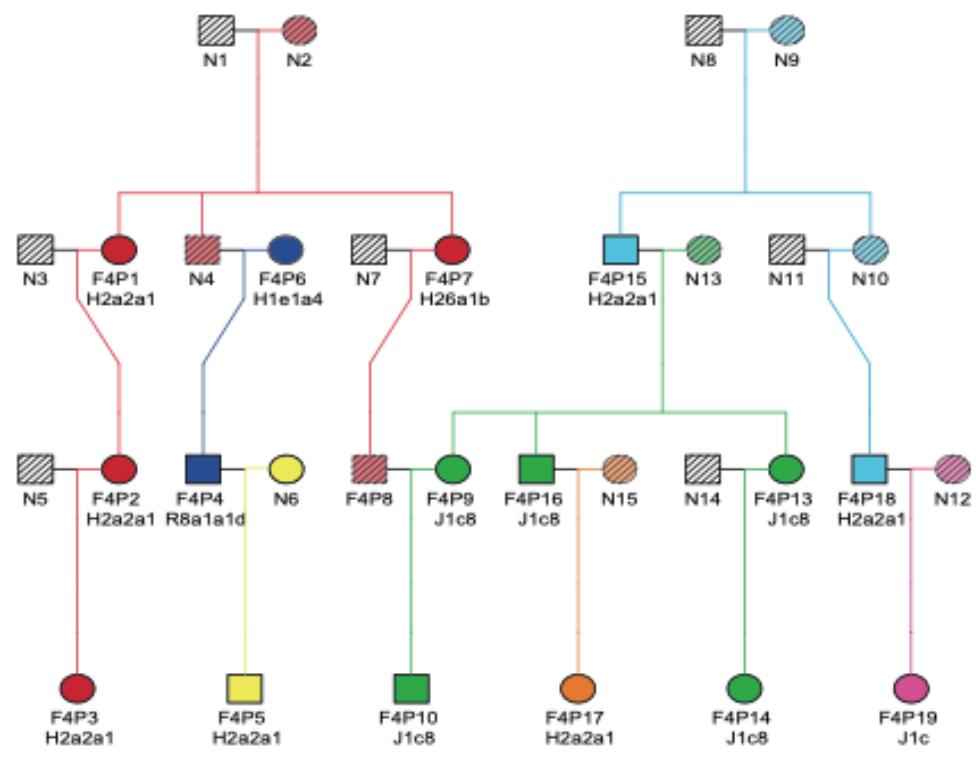
c)

Family 3



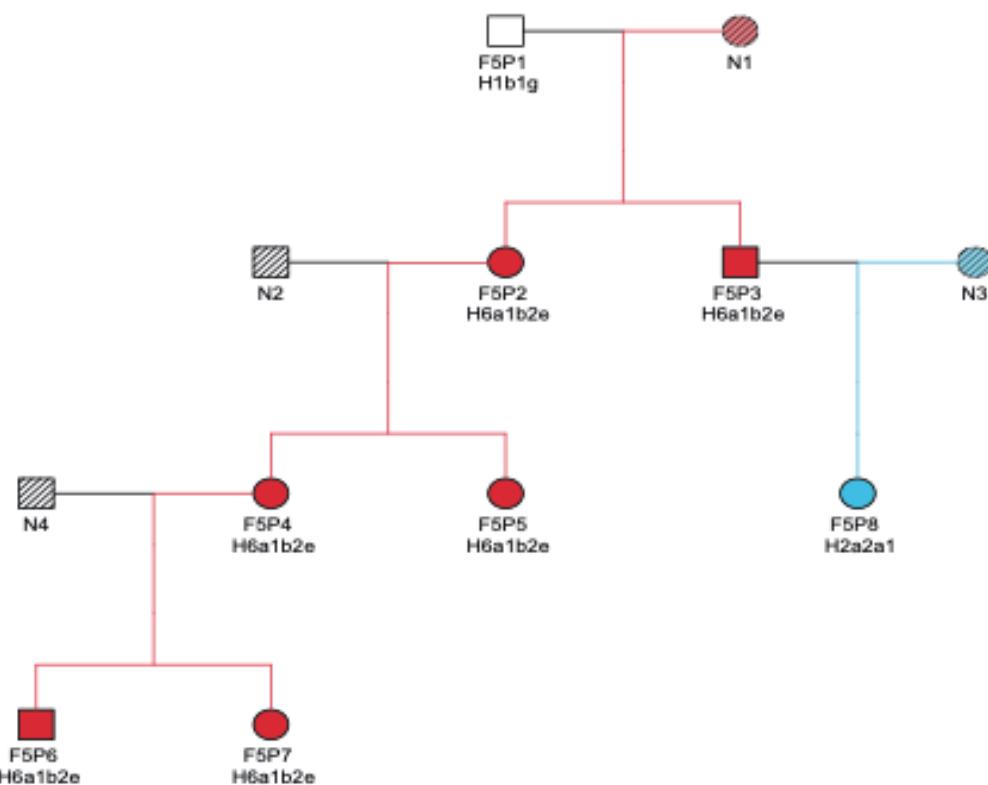
d)

Family 4



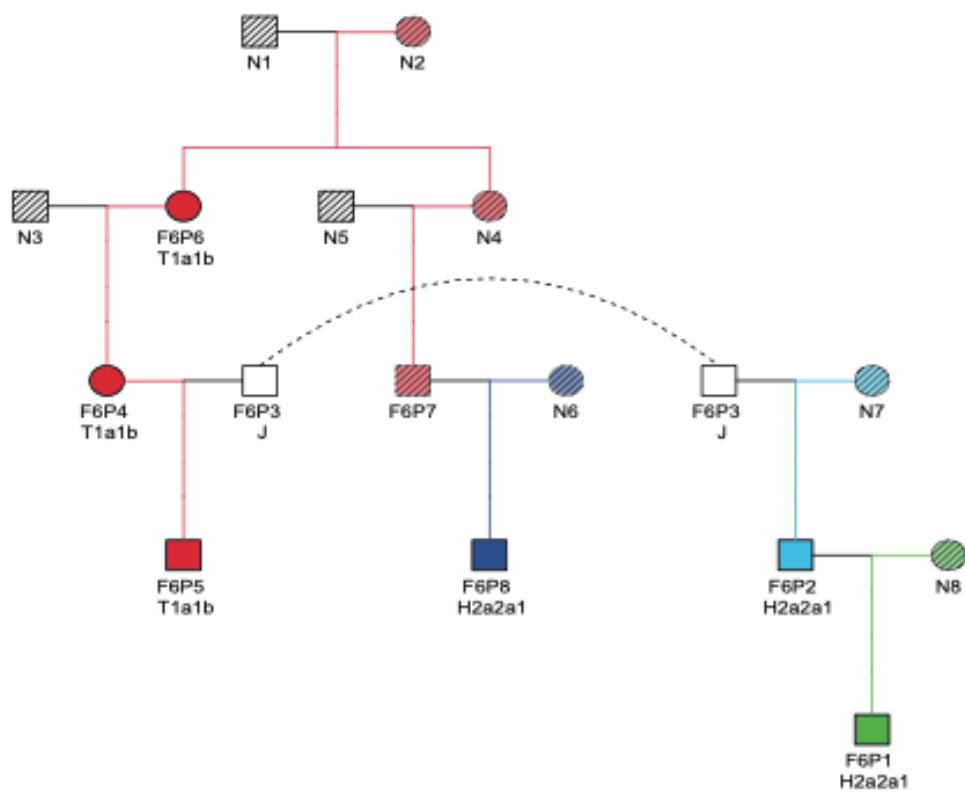
e)

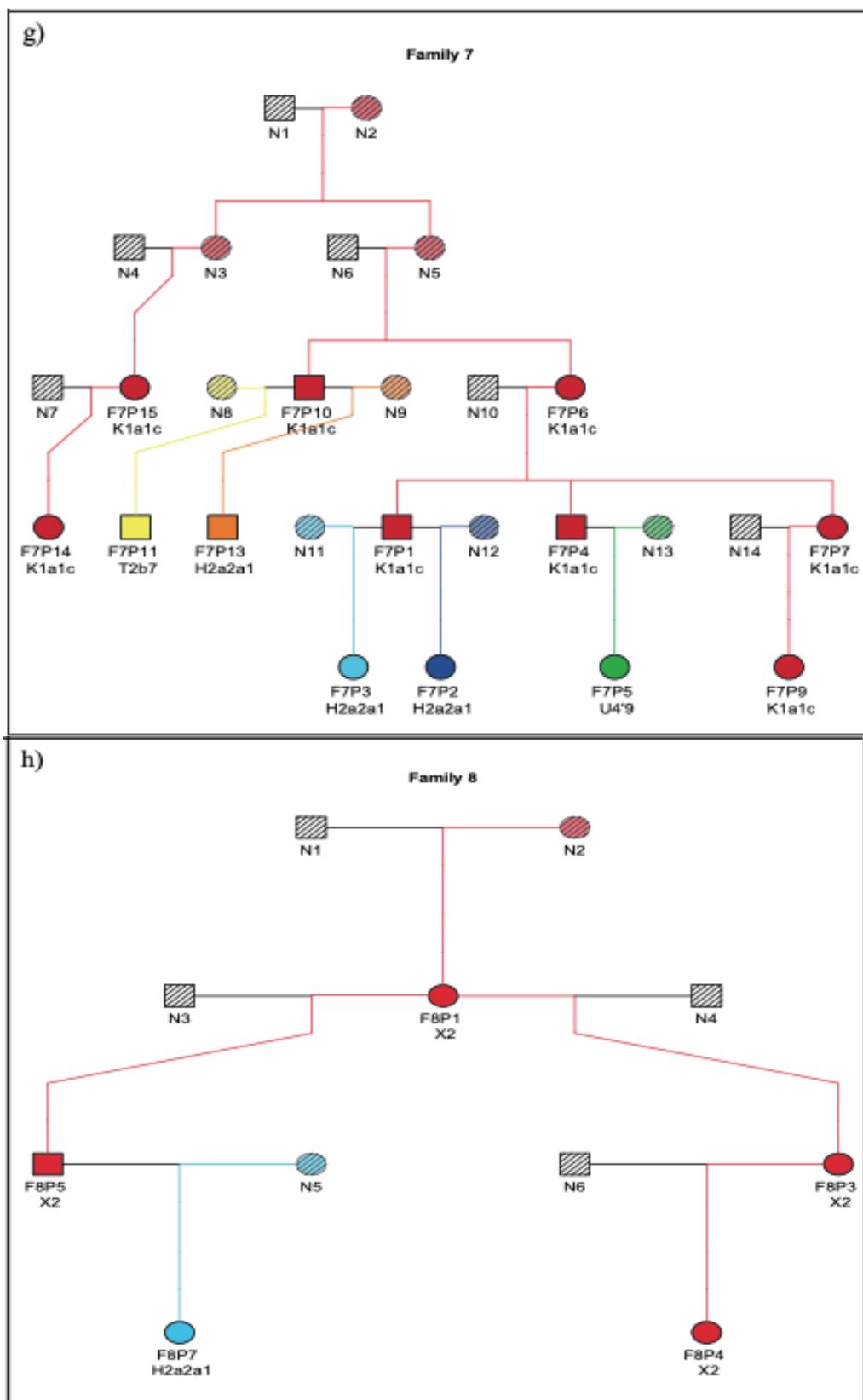
## Family 5



f)

## Family 6





**Figure 2.25 Mitochondrial DNA haplogroups of the families analysed.**

Different maternal lines are highlighted in different colours, the haplogroup estimated by HaploGrep is written under each individual. Note the inconsistencies in Family 4, involving individuals 1, 7, 6 and 4.

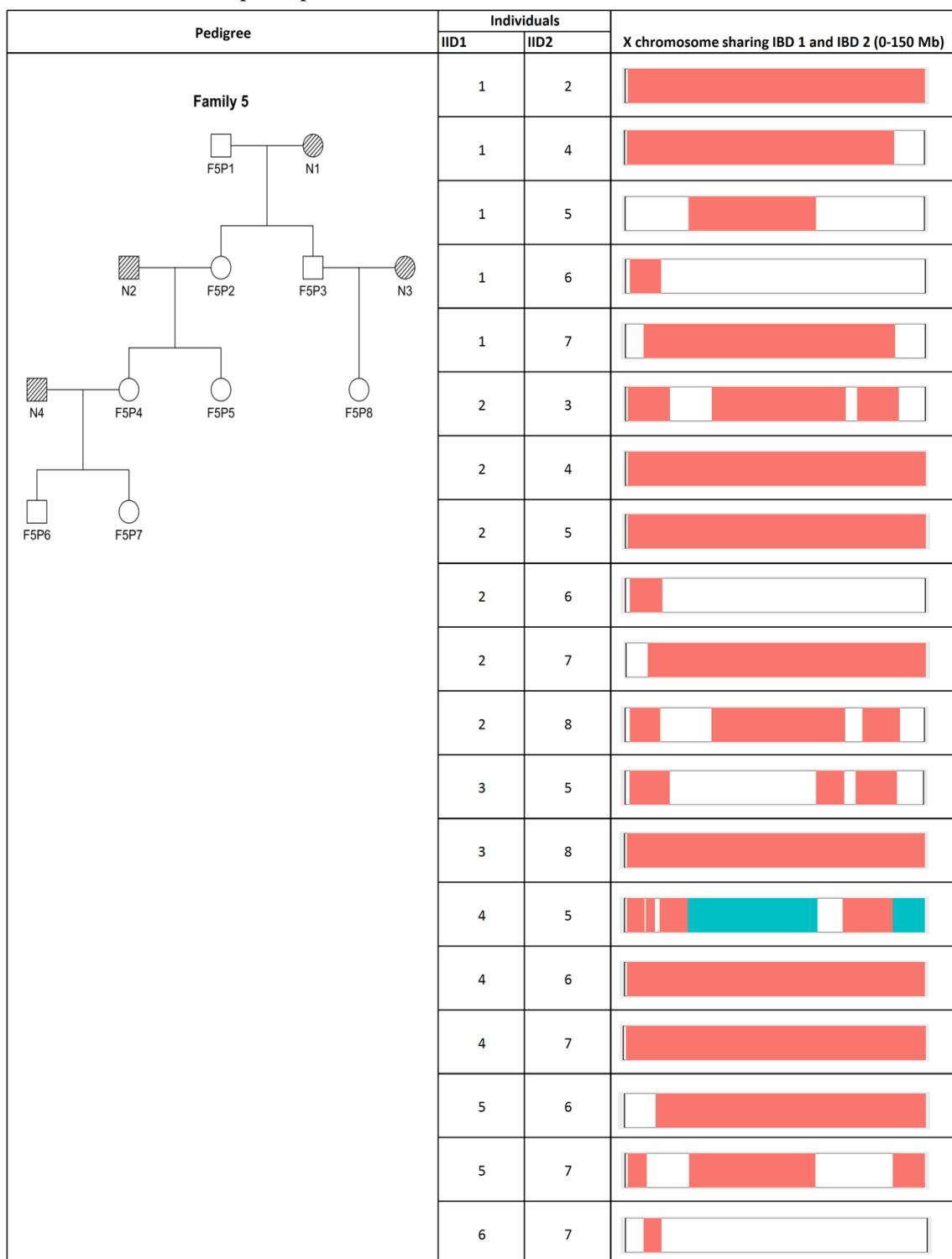
**X-chromosome SNP analysis.** Kinship analyses that aim to identify close and distant relatives may benefit from the inclusion of X chromosome markers, as IBD segments may be preserved over a greater number of generations due to recombination occurring less frequently than in autosomes (Henden et al. 2016). Analysis on the X chromosome may help, for example, if the paternal and maternal line of two sisters is questioned: if the hypothesis that they are the offspring of the same father is true, the two sisters would share at least one allele IBD for each X-chromosomal marker (Kling et al. 2015a). Also, X chromosome markers could be used to confirm that two females are maternal cousins and not unrelated.

The pairwise IBD states are influenced by the sex of the individuals, since less than four alleles can be considered when two males are involved, but when female pairs are considered, IBD states derived from the X chromosome and the autosome are the compatible (Vigeland 2020).

The X SNPs used underwent the same QC as in Section 2.2.3.1 and was pruned for LD in PLINK (window size of 200, step of 100), as higher level of LD are expected in the X (Pritchard and Przeworski 2001; Tomas et al. 2008), obtaining a set of 1612 SNPs out of the total of 22,927 X SNPs. The segments smaller than 7 cM were excluded from further analysis (ISOGG 2020, [https://isogg.org/wiki/Autosomal\\_DNA\\_match\\_thresholds](https://isogg.org/wiki/Autosomal_DNA_match_thresholds) ).

**Table 2.10 Segment sharing on the X chromosome for sample pairs in family 5, based on 1612 SNPs.**

Pedigrees are reported and segment sharing calculated by XIBD on the X chromosome is plotted: the picture represents chromosome X, with segments highlighted in light blue if shared IBD status is 2 between the related pairs, pink if IBD is 1.

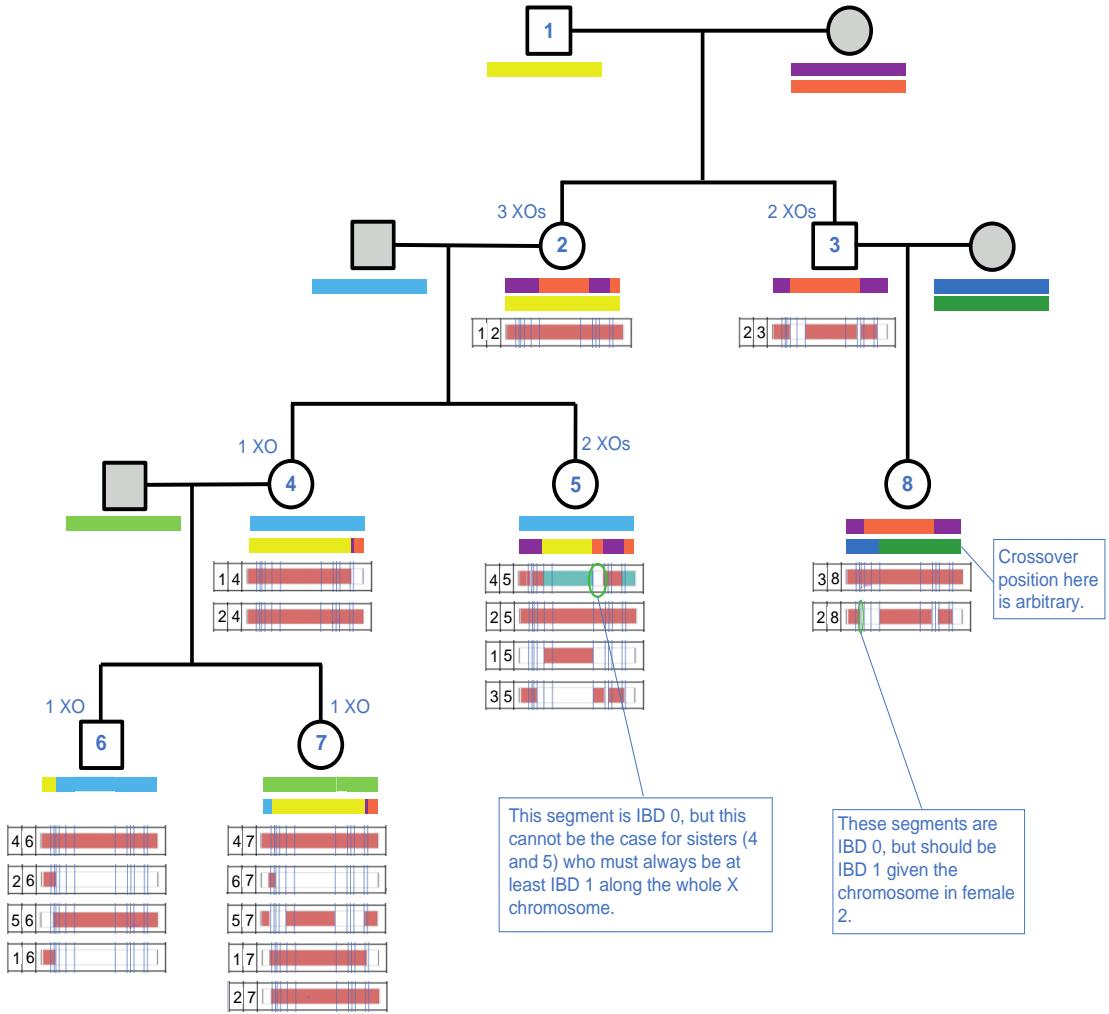


Pinto et al. (Pinto et al. 2012) describe the transmission of the X, highlighting that the pattern of inheritance among pedigrees depends on the number of females linking each generation, and that the X-chromosomal link between generations is broken by a father-son connection within a pedigree. All observations were consistent with these considerations, and with the autosomal data and the known information about the pedigrees.

Family 5 is described here as an example: Table 2.10 shows the pairwise X-chromosome sharing IBD, and Figure 2.26 shows these pairwise relationships on the pedigree itself, and the inferred positions of crossovers in X-chromosome meiosis in the females.

As expected, the mother - offspring pairs (F5P2 and F5P4, F5P5; F5P4 and F5P6, F5P7) share IBD1 segments along all chromosome X. In sibling pair F5P6 and F5P7, the crossover in F5P4 heavily influences the amount they share IBD. In a grandparental relationship, the female-male F5P2 and F5P6 pair share very little DNA on the X due to crossover in individual F5P4. Due to the male F5P3, the female F5P8 shares no segments with the grand-father F5P1.

There are two apparent anomalies in the sharing IBD given the pedigree structure and the inferred crossover behaviour. The full sibling pair (F5P4 and F5P5) shows expected IBD2 and IBD1 sharing, with the exception of one segment of 0.1 cM that is expected to be IBD1, but shows IBD0. This is due to the exclusion of short segments during QC (an aspect highlighted above and that needs to be taken into account in interpreting these analyses). Similarly, F5P8 and F5P2 show a small (~7 cM) unexpected gap, which has the same explanation.



**Figure 2.26 Patterns of segment sharing within family 5 and inferred crossovers.**

The pedigree is shown with individuals numbered minimally (1 to 8). Below each individual is schematically shown one (male) or two (female) X chromosomes as bars. Colours of segments trace inheritance and crossovers, with ancestral (founding) colours chosen arbitrarily. Below the X-chromosome bars are shown the pairwise IBD relationships taken from Table 2.10, with fine vertical lines showing crossover positions deduced from the table via visual alignment. Two anomalous segments are highlighted. XO: crossover.

## 2.4 Discussion

This Chapter investigated genome-wide data on eight families to explore the information that can be obtained from autosomal, uniparentally-inherited and sex-linked markers for kinship determination. The results obtained from a large number of SNPs provide a framework for evaluating the massively parallel sequencing (MPS) data that will be considered in the next chapter.

### Pedigree reconstruction – performance and alternative approaches

The autosomal SNP data were analysed with methods that consider IBD sharing among individuals to provide relationship coefficients between pairs (Section 2.3.4.1) and to reconstruct pedigrees (Section 2.3.4.2) in order to identify distant relationships (up to second cousin and second cousin once removed, as available from the samples analysed here). The results are discussed in the light of the constraints of the collection design: the limited number of typed founders have an impact on the analyses, especially for the pedigree reconstruction method.

PRIMUS proved to be a powerful tool, being able to correctly reconstruct most of the tested families, however there were some issues due to the complexity of some pedigrees and the amount of missingness. The most distant relationships were difficult to identify. It appears that the variability of IBD proportions and (unsurprisingly) the missingness of individuals who would act as essential links to the family network affected PRIMUS pedigree reconstruction. Samples that can affect the structure the most are those that create a gap larger than 3rd degree relationship in the family tree, as variance in IBD sharing increases as well as the number of possible pedigrees matching the genetic information, raising the number of possible relationships (as shown in Staples 2014). Nonetheless, PRIMUS does not appear to assign equal likelihoods to multiple networks. Complex relationships, as in this case, will not be recognised and attention is restricted to one of the possible relationship types allowed. This may signal that pairwise estimation may still be the best option in such cases.

The number of genetic markers, admixture and population substructure, together with appropriate reference minor allele frequencies can affect the quality of pairwise IBD

estimates. Other methods are able to consider these factors, such as PLINK, KING, or REAP (as Staples et al. 2014 suggest). A different method such as CLAPPER (Ko and Nielsen 2017) may overcome issues regarding IBD estimate accuracy, using pairwise likelihood values instead of relationship assignments, but is not specific for many close relationships as it tries to target relatives from first-cousin; ERSA (Huff et al. 2011) and BEAGLE (Browning and Browning 2007) are able to reconstruct more distant relationships, based on segment sharing (considering phasing as well).

Pedigree reconstruction methods offer family tree structure information compatible with the genetic data and pairwise estimations, which may be more informative than pairwise relationship information. However, most methods for pedigree inference are limited to one or two generations, as in parentage inference methods or in sib-ship clusters, and/or rely on prior information regarding pedigrees, or cannot handle Linkage Disequilibrium among the variants or cannot reconstruct half-sibling relationships (for example IPED He et al. 2013; Ko and Nielsen 2017; Staples et al. 2014). Some methods, even when able to consider multiple generations, can only use either a limited number of markers (FRANz; Riester et al. 2009), or must include every member in the pedigree (as in Cowell 2009; Cussens et al. 2013), or cannot analyse polygamous pedigrees (IPED, He et al. 2013; Ko and Nielsen 2017). Approaches for both inbred and outbred pedigrees have been proposed (e.g. constructing inbred pedigrees, CIP, and constructing outbred pedigrees, COP, Kirkpatrick et al. 2011).

However, some limitations remain (Staples et al. 2014): often, these tools assume monogamy (Kirkpatrick et al. 2011), can be used only for small pedigrees or with limited numbers of markers, may not be able to accommodate inconsistencies due to pairwise estimation (Stankovich et al. 2005) or missing nodes, or are based on strong assumptions (e.g. COP/CIP, Kirkpatrick et al. 2011; IPED, He et al. 2013; and PREPARE, Shem-Tov and Halperin 2014) assume that all genotyped individuals are in the same generation, which requires prior knowledge of the pedigree structure) or arbitrary assumptions (e.g. recombination models, or linkage equilibrium among variants) (Staples et al. 2014). According to (Staples et al. 2014), PREPARE (Shem-Tov and Halperin 2014) and IPED2 (He et al. 2013) is unable to reconstruct most of the pedigrees containing one half-sibling relationship and multiple generations, while PRIMUS identifies most of them. In addition, IPED2 needs pedigree structures information and PREPARE needs to know

some information *a priori*, as for example which individuals are in the same generation, unlike PRIMUS.

Overall, then, the choice of appropriate method and adjustments relies greatly on the samples' characteristics, and every method has its advantages and disadvantages.

**Inclusion of uniparentally-inherited and sex-linked markers.** Generally, Y, X-chromosome and mtDNA are ignored when SNP chip data are analysed, thus ignoring the potential of adding information to the kinship estimation. There are some factors to consider in interpreting these data:

- There is no consensus on which markers to include on a chip, and this technical issue can limit further analysis.
- For Y and mtDNA, the high frequencies of some haplogroups make these variants useful for exclusion, but not for inclusion. Because haplogroup spectra are highly population-specific, the degree of informativeness of a given chip will depend on the population being considered. The resolution of haplogroups is variable, and some included are monomorphic, or show apparent heterozygotes, so cannot be used;
- Recurrent mutations and variants in hypervariable regions need to be considered when interpreting mtDNA results, especially where they indicate apparent exclusions: as we can see from our data, if we considered only the mtDNA haplogroups, a real relationship would have been excluded (section 2.3.3.3). There is a need to find balance in mtDNA resolution: considering the main clades, the informativeness of the haplogroup may not be sufficient in a kinship testing and the possible informativeness of the full set of variants is lost, however, a very high resolution needs to take into account possible mutation and apparent discrepancies in the family line;
- Interpretation of results on the X chromosome is complicated by high LD, and the artefactual exclusion of small segments due to setting a centiMorgan cutoff.

This Chapter offers guidance on the advantages and disadvantages of kinship analysis based on IBD estimation and pedigree reconstruction methods for SNP chips that include autosomal, X, Y and mtDNA markers. The analysis workflow is illustrated step-by-step, from QC to population substructure detection, evaluating each step with its technical

limits, and to kinship determination of a real-world family dataset. Applying these tools to the analysed dataset allowed the consideration of their empirical limits and characteristics, and showed what output to expect, how distant relationships are identified and what the limits are. Information that will help in processing SNP data for kinship determination in real-world samples is summarised in a form of step-by-step guidance in Box 1, offering suggestions on how to proceed in the analyses.

Further analyses may assess the applicability of these techniques in a more diverse dataset to test the methods in the presence of more divergence in allele frequencies, for example in cases of admixture; analysis in inbred pedigrees would also be illuminating.

**Box 1. Overview of the proposed guidance for determining kinship relatedness using genome-wide SNP data**

**Step 1. Consider the appropriate allele frequencies.**

- a) Allele frequencies are generally based on the founders present in the dataset.
- b) If the number of founders and/or unrelated individuals in the dataset is small, consider a reference dataset for estimating the allele frequencies necessary for the IBD calculations.
- c) Step (b) entails a PCA to confirm the population of origin and possible presence of population subgroups. Note that inclusion of the related samples will affect the PCA.
- d) Choose the representative population: this can be used as external training set or merged to the related individuals dataset.

**Step 2. Autosome focused analysis.** There are several different approaches, here two are suggested.

- e) Consider what type of information your application needs: pairwise kin relationships or family tree information.
- f) *IBD based methods.* The objective is to find the level of relatedness between pairs of individuals (relationship coefficient):
  - Generally, first cousins can be identified, but more distant relationships may show too much variation from the expected relationship coefficient and/or low relationship coefficient;

- Consider relationships with the same IBD (Step 3 may overcome this issue in some cases);
  - A visualisation method can be used to identify the relationship groups (IBD plot).
- g) *Pedigree reconstruction methods.* The objective is to find the family tree structure:
- Distant relationships can be identified, however the complexity of the family tree may have a negative impact;
  - Need to consider the validity of the IBD estimates used as input;
  - A wrong directionality may be assigned to a relationship (i.e. the parent-offspring relationship is identified, but the parent is included as offspring or vice versa); this can be easily solved by adding external information (e.g. age)
  - The proportion of untyped individuals (missingness) has a negative impact; the IBD measurement may be the most accurate option (f).

**Step 3. Uniparentally-inherited and sex-linked marker analysis.** Consider the structure of the pedigree analysed and the sex of the individuals involved: these markers may offer additional information and/or confirm some relationships.

Consider that these analyses can exclude a relationship hypothesis, but cannot include it.

- h) *Y chromosome.* The objective is to determine the haplogroups passed down through the paternal line:
- The haplogroups' main clades can be reported: this low resolution may not always help in a kinship determination setting;
- i) *Mitochondrial DNA.* The objective is to determine the haplogroups passed down through the maternal line:
- Need to consider if the hypervariable region is included: recurrent mutations may lead to apparent discrepancies in the family line;
  - Need to consider if the SNP chip is intended for this type of kinship analysis: insufficient coverage may lead to discrepancies in the family line, as denser marker are more suitable for this type of analysis.
- j) *X chromosome.* The objective is to determine the amount of segment sharing:
- Consider that the analyses are complicated by high LD present in the X: pruning or additional LD information are needed;
  - Excluding short segments (less than 7 cM) is advisable, however consider that this will affect the segment sharing: the output needs to be interpreted with care.

# **Chapter 3: Kinship estimation in a European population using massively-parallel sequencing data on STRs and SNPs**

Data derived from genome-wide SNP chips provide a high degree of discriminatory power, but are generally impractical to generate for the sample types encountered in forensic casework, which are often low in quality and quantity, or mixed. An alternative may be the use of massively-parallel sequencing (MPS), which is becoming a powerful tool in forensic settings. For example, the ForenSeq™ DNA Signature Prep Kit (Verogen, San Diego, CA) (Illumina 2016) can simultaneously analyse up to 230 markers (Table 3.1) chosen for their forensic application, including autosomal SNPs and autosomal, X- and Y-chromosomal STRs, via PCR followed by sequencing. Its high sensitivity, accuracy and precision has been demonstrated by several studies (Li et al. 2019).

**Table 3.1 ForenSeq DNA Signature Prep Kit (Verogen, San Diego, CA) markers.**

Two different primer plexes, A and B, can be used, containing different numbers of markers. The Plex B includes 56 ancestry and 24 phenotypic SNPs (for eye and hair pigmentation), of which two SNPs overlap.

<b>Feature/Type of marker</b>	<b>Number of Markers</b>	<b>Amplicon Size Range (bp)</b>	<b>Plex</b>
Global Autosomal STRs	27	61–467	A/B
Y-STRs	24	119–390	A/B
X-STRs	7	157–462	A/B
Identity SNPs	94	63–231	A/B
Phenotypic SNPs	24	73–227	B
Biogeographical Ancestry SNPs	56	67–200	B

The use of MPS has revealed numerous advantages, and has been recently applied in forensic laboratories, as well as in real casework (Delest et al. 2020). Among these advantages is the ability to analyse larger numbers of STRs than is possible via capillary electrophoresis (CE), since there is no restriction due to overlapping sizes or colour channels. In addition, there is the possibility of analysing markers of different types (Delest et al. 2020). Sequence variation in the STR repeat arrays and flanking regions increases discriminatory power (Delest et al. 2020; Devesse et al. 2020; Gettings et al. 2018a; Gettings et al. 2016; Khubrani et al. 2019; Li et al. 2019; Novroski et al. 2016; Phillips et al. 2018b). The addition of targeted SNPs also offers significant advantages for forensic testing. Because they can be amplified on very short fragments, they may help with challenging samples (degraded or low template DNA), and in sufficient numbers (94 SNPs for ID are included in the ForenSeq system) can be used for human identification. SNPs chosen because they show large allele frequency differences between populations can be used for ancestry determination (56 in the ForenSeq kit). The obstacles posed by CE, which may only target a limited number of SNPs, are handled well by MPS technology (Amorim and Pereira 2005b; King et al. 2018).

In Chapter 2, kinship determination was undertaken based on genome-wide SNP data from a set of eight families. This Chapter examines the possibilities of kinship analysis when using MPS data, specifically from the ForenSeq™ DNA Signature Prep Kit (Verogen, San Diego, CA) (Illumina 2016). The same set of families as used in Chapter 2 was sequenced for the full set of ForenSeq markers, allowing a comparison to the known pedigree information of samples and to the conclusions drawn from genome-wide SNP data analysis.

Further information on the library preparation and sequencing protocol is reported here. The raw sequence data are automatically analysed by the built-in platform Universal Analysis Software (UAS, Verogen). However, other tools may be used. Here, STRait Razor v3 was used for external validation, as well as a more traditional pipeline based on BWA (Li and Durbin 2009), a software package for aligning sequences to a reference genome, and samtools (Li et al. 2009).

### **3.1.1 Data analysis on the Universal Analysis Software (UAS) and statistics**

The UAS was used to assess cluster density and for sequencing QC (phasing and pre-phasing) as recommended by the manufacturers.

The UAS platform provides reports on the sequence data in two sections, one for Plex A and one for Plex B, for both STRs and SNPs. For STRs, the platform shows the genotypes, flagged loci (e.g. for imbalance and stutter), coverage and locus-specific information (allele names, sequences, and read numbers). Alleles are called if their sequence reaches a preset interpretation threshold (or stochastic threshold): this threshold (set at 30 reads by the UAS) should ensure that reads above it are not affected by stochastic effects, limiting the chance of drop-outs. The sequences under the interpretation threshold and above the analytical threshold (10 reads) are flagged and need to be examined by the user. The reads under the analytical threshold are not reported by the UAS.

The UAS produces three reports (sample level and project level genotype, for a single sample and across all samples, respectively, and flanking region across all samples) with genotypes, flanking region information and depth of coverage (DoC), which are easy-to-read reports for the forensic practitioner; however, these cannot be used as input for bioinformatic tools, and therefore the following tools were also used.

### **3.1.2 Summary of the features of STRait Razor v3**

STRait Razor (Woerner et al. 2017a) is a bioinformatic tool that identifies and characterises sequence- and length-based variants in MPS data from specific sequencing platforms, composed of a Perl script for haplotype calling and an Excel-based workbook for the annotation step. Haplotypes are identified through the use of anchor sequences (indexing strategy for 5' and 3') that match substrings in the sequence data (with a limited allowance for mismatch of one base substitution). The Excel workbook shows length- and sequence-based genotypes, locus haplotype data and nomenclature based on a nomenclature database (containing ~2700 haplotypes; alleles not in the database are named “Novel”) in line with ISFG guidance (Parson et al. 2016).

### **3.1.3 Summary of the features of Familias**

Familias v3 is a software for calculating probabilities and likelihoods in kinship cases based on DNA data and relationship hypothesis (Egeland et al. 2000; Kling et al. 2014b). Information on reference allele frequency is necessary, and mutation as well as drop-in/out, and silent allele can be handled.

### **3.1.4 Summary of the features of the forrel package**

The R package forrel v1.3.0 (ped suite, <https://github.com/magnusdv/forrel>) (Egeland et al. 2014; Kling et al. 2017; Vigeland and Egeland 2019) offers tools for pedigree analysis for forensic applications: calculations of likelihood ratios, IBD coefficient estimation, modelling of background inbreeding and mutations, power analysis and marker simulation.

### **3.1.5 Aims of this Chapter**

The ForenSeq Signature Prep DNA kit's purpose is the identification of unknown individuals thanks to the combination of traditional forensic STR markers, identity SNPs, ancestry-informative and phenotypic (for eye and hair colour) SNPs (Primer Mix B). This project aims to apply the information that can be provided by this cutting-edge MPS forensic kit to answer questions regarding kinship deduction, while proposing an efficient analysis pipeline for doing this as well. Note that the reason for including phenotypic and ancestry SNPs here was to increase the overall marker numbers, rather than to attempt prediction of pigmentation and ancestry in the family samples. Results from the ForenSeq analysis will be evaluated through comparison to the known pedigree information of samples and to the output obtained from genome-wide SNP data analysis (Chapter 2).

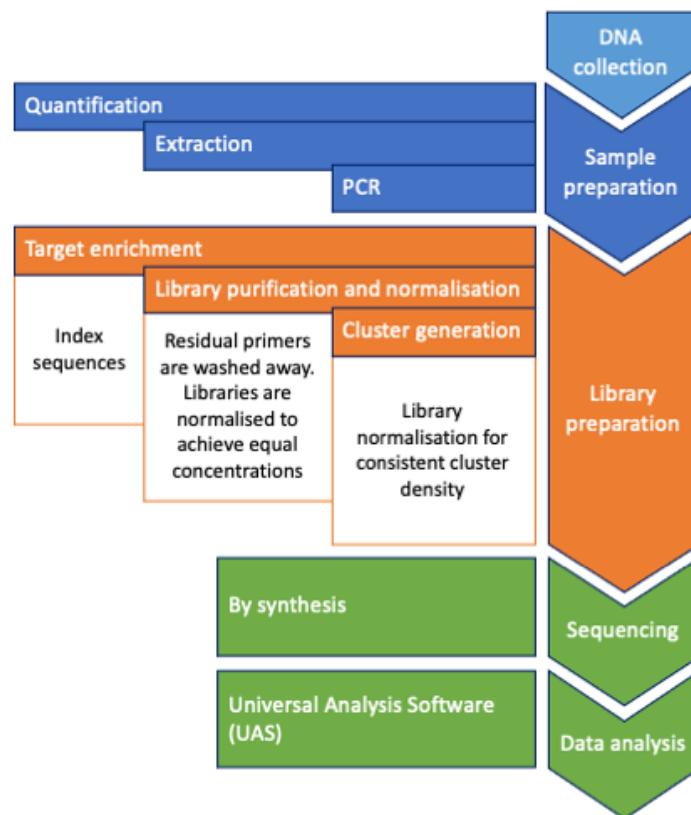
## **3.2 Methods**

### **3.2.1 DNA samples and sequencing**

The samples are described in detail in Chapter 2. DNA from 72 related individuals of European origin (Germany) forming 8 families were quantified using a NanoDrop 2000c spectrophotometer (ThermoFisher Scientific), and sequenced according to the

manufacturer's protocols using the ForenSeq DNA Signature Prep Kit (Verogen, San Diego, CA). Amplicons generated using Primer Mix B (including the phenotypic and biogeographic ancestry SNPs for a total of 172 SNPs and 58 STRs) were sequenced in 3 runs (30 samples per run plus 2 controls).

### Library preparation, normalisation and pooling



**Figure 3.1 Massively parallel sequencing (MPS) workflow.**

The chart shows the necessary steps, from DNA collection to analyses of the sequences. Here, the focus is on the Library preparation step.

Primer Mix B targets 27 autosomal STRs, amelogenin, 24 Y STRs, 7 X STRs, 94 identity informative SNPs (based on the SNPforID, Sanchez et al. 2006; and the iiSNP panel, Pakstis et al. 2010), 54 ancestry SNPs (based on Kidd et al. 2014) and 22 phenotypic SNPs (based on the HiRisPlex, Walsh et al. 2014), with two SNPs in both the ancestry and phenotypic sets.

In this first step, it is important to quantify the sequencing-ready molecules in the samples to obtain a balanced coverage, as in the subsequent clustering step high amounts of DNA may reduce resolution, and low amounts of DNA can lead to sparse clustering. Various sample sources can be used (e.g. extracted genomic DNA, saliva, samples on FTA cards after wash steps), and around 1-2 ng of input undergoes two PCR cycles: PCR1 adds primers with tags, which allows ligation to unique indexed adapters added in PCR2.

### **Target enrichment and library purification**

Two sequences are fundamental to this step: index sequences, which identify sample' amplicons, and adapter sequences, which allow the hybridisation to the flow-cell surface (a glass slide with oligos-coated channels) as they are complementary to the immobilised flow-cell oligos. Residual primers are “washed” away through a purification step.

### **Cluster generation**

Libraries are normalised using magnetic beads to ensure their balanced representation and equal concentration of each library, independently from their original yield. This step aims to achieve consistent cluster density.

### **Sequencing**

Then, libraries to be sequenced together on the same flow cell are pooled, denatured and diluted and loaded on the reagent cartridge for sequencing.

## **3.2.2 Reference dataset**

The Human Genome Diversity Panel (Cann et al. 2002a) was used as reference dataset for allele frequencies (Phillips et al. 2018a). More information on the reference frequencies and description of sequence data is presented in Section 3.3.7.

## **3.2.3 Simulated data**

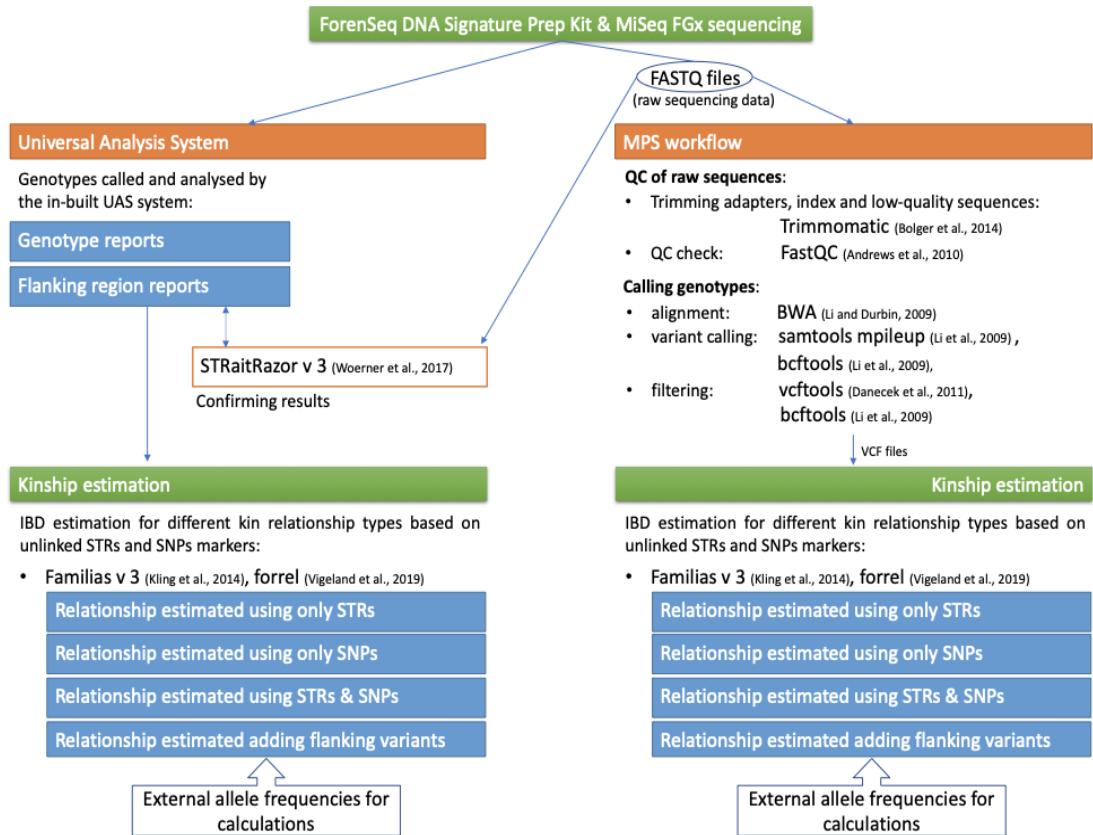
Data in section 3.3.6.1 were simulated using the package *forrel* v1.3.0 (ped suite, <https://github.com/magnusdv/forrel>). The *forrel* package offers the possibility of performing both conditional (based on genotypes to “condition on”) and unconditional simulations. The latter are based on gene dropping, where random alleles are chosen for the founders (parents) based on the provided allele frequencies and each parent-offspring

transmission respects Mendelian inheritance. Unconditional simulations, with the same seed (12345), are used here.

### 3.2.4 Data analysis

The workflow for data analysis included two main steps including genotype calling and kinship analysis:

- 1.a. Sequencing data were processed on the ForenSeq Universal Analysis Software (UAS, Verogen), producing reports with genotypes and flanking region information. Raw data (FASTQ files) were also checked using STRait Razor v3 (Woerner et al. 2017b).
- 1.b. Using raw sequences (FASTQ files), a MPS workflow was applied based on BWA (Li and Durbin 2009), a software package for aligning sequences to a reference genome, and samtools (Li et al. 2009) a tool for manipulating files in Sequence Alignment/Map (SAM) and Variant Call Format (VCF) format.
2. Kin relationship among samples was estimated by applying IBD estimation based on unlinked STR and SNP markers using Familias v3 (Kling et al. 2014b) and forrel (<https://github.com/magnusdv/forrel>).



**Figure 3.2 Workflow of the sequencing analysis.**

Samples are sequenced using the ForenSeq Kit, and genotype data is called and analysed either using the UAS platform (left) or a MPS workflow based on BWA alignment algorithm and samtools (right). Then kinship analysis is based on the Familias v3 tool and forrel package. The necessary allele frequencies are obtained from (Cann et al. 2002a; Phillips et al. 2018a).

### 3.2.4.1 Data analysis workflow: Universal Analysis Software (UAS)

Preliminary data analysis was performed using the ForenSeq™ Universal Analysis Software (UAS), with default analytical (AT) and interpretation thresholds (IT), of 10 and 30 reads respectively, and default Stutter Filter and Intra-Locus Balance. Some alleles were below IT but above AT and were manually called considering percentage balance between them for heterozygotes (the lower-read-number allele is required to be at least 50% of the higher).

The following STRs (and some target SNPs) are reported on the reverse strand compared to the human genome reference in the UAS reports: D1S1656, D2S1338, D5S818, D6S1043, D7S820, D19S433, CSF1PO, FGA, Penta E and vWA. This was accounted for

in external data comparisons and reporting. Results are exported as Microsoft Office Excel reports, containing both genotypes and flanking region information (UAS v1.3).

### **3.2.4.2 Data analysis workflow: STRait Razor**

STRait Razor v3.0 was used for data analysis on the extracted FASTQ files and results were compared to UAS output. Coverage threshold was set to 5 reads. The sequences flagged as “NovelSeq” were further explored in STRSeq (BioProject, Gettings et al. 2017; Gettings et al. 2017b) by BLAST, accessed on 15/04/2021.

### **3.2.4.3 Data analysis workflow: variant calling workflow**

Fastqc (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>) was used for assessing read quality. FASTQ files for each sample were exported from the MiSeq FGx and the following workflow to call genotypes was used.

FASTQ files were aligned to a custom reference genome (based on NCBI hg19), targeting the regions containing ForenSeq markers and additional padding of 400 bp at each end (the phenotypic SNPs on chromosome 16 are considered as one cluster) to include flanking regions, using the Burrows-Wheeler Alignment tool BWA (Maximal Exact Match [mem] algorithm, which searches for the longest substring of sequence that exactly matches to the reference) (Li and Durbin 2009). Although this technology offers Paired-End reads, because of the significant disparity of read lengths between the two pairs generated by the ForenSeq kit (~351 and ~31 bases respectively), only the first read can be used for further analysis. The process involves several steps that take the raw FASTQ files of each sample and filter them, convert to BAM format and then, when variants are identified, into VCF files. Overrepresented sequences in the FASTQ files were investigated further through BLAST (Altschul et al. 1990) and in files of known Illumina sequences (<https://raw.githubusercontent.com/NCBI-Hackathons/OnlineAdapterDatabase/master/datasources/illumina-adapter-sequences1000000002694-01.txt>, [https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry\\_documentation/experiment-design/illumina-adapter-sequences-1000000002694-10.pdf](https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry_documentation/experiment-design/illumina-adapter-sequences-1000000002694-10.pdf)).

This was used to create a custom reference FASTA file for the following clipping step. The Illumina sequences file (TruSeq3) for Single End reads was used (Trimmomatic v0.36), adding the found adapter

sequences from the previous step. Table 3.2 describes each step up to variant calling. After these procedures, rsIDs were assigned to the SNP variants found and PCA performed (on PLINK files obtained from the VCF).

**Table 3.2 MPS variant calling workflow.**

Each step (from raw data to variant calling) and tools used are described.

# step	Analysis step	Tools	versi on	Description	Variables	Ref.
1	Fastq files quality assessment	FastQC	0.11.5	Quality control check of raw sequence data	-	Babraham Bioinformatics, <a href="https://www.bioinformatics.babraham.ac.uk/projects/fastqc/">https://www.bioinformatics.babraham.ac.uk/projects/fastqc/</a>
2	Fastq files quality step	Trimmomatic	0.36	Adapter and low quality bases clipping	SE -phred33 ILLUMINACLIP:TruSeqAll_adapters_PE_T H.fa:2:30:10:1:true LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36	(Bolger et al. 2014) ( <a href="http://www.usadellab.org/cms/?page=trimmomatic">http://www.usadellab.org/cms/?page=trimmomatic</a> )
3	Fastq files quality assessment after trimming	FastQC	0.11.5	Quality control check of raw sequence data	-	Babraham Bioinformatics ( <a href="https://www.bioinformatics.babraham.ac.uk/projects/fastqc/">https://www.bioinformatics.babraham.ac.uk/projects/fastqc/</a> )
4	Alignment of sequences to reference	Burrows-Wheeler Transform algorithm (BWA)	0.7.17	A custom reference based on hg19 was used	-mem	(Li and Durbin 2009) ( <a href="http://bio-bwa.sourceforge.net/">http://bio-bwa.sourceforge.net/</a> )
5	coverage	samtools	1.9	counting UNLIMITED DoC in unrefined (non indel realigned ) BAM against the reference	mpileup -AB -Q 20 -q 50 -d10000000	(Li et al. 2009) ( <a href="http://www.htslib.org/">http://www.htslib.org/</a> )
6	sort and index	samtools		Indexing is a common search technique	sort -o index	(Li et al. 2009) ( <a href="http://www.htslib.org/">http://www.htslib.org/</a> )

# step	Analysis step	Tools	version	Description	Variables	Ref.
7	add read group	picard	2.6.0	Validate sam and bam files; add identifying read groups in multi-sample VCF file	ValidateSamFile AddOrReplaceReadGroups	Broad Institute ( <a href="https://broadinstitute.github.io/picard/">https://broadinstitute.github.io/picard/</a> )
8	sort and index again	samtools	1.9	Indexing is a common search technique	sort -o index	(Li et al. 2009) ( <a href="http://www.htslib.org/">http://www.htslib.org/</a> )
9	variant calling	samtools, bcftools	1.9	minimum base quality (-Q) is set to 20; minimum map quality (-q) is set to 50; maximum number of reads (-d); the mapping quality is encoded as ASCII characters in the output (-s); model for multiallelic and rare-variant calling (-m); comma separated format fields in the output (-f)	samtools mpileup -u -Q 20 -q 50 -g -d 10000000 -s bcftools call -m -f GQ	(Li et al. 2009) ( <a href="http://www.htslib.org/">http://www.htslib.org/</a> ; <a href="http://samtools.github.io/bcftools/bcftools.html">http://samtools.github.io/bcftools/bcftools.html</a> )
10	filter variants	vcf-tools	0.1.14	thresholds are set for the minimum read depth (d=20; default is 2), and SNP to be filtered in a region (default of w=10 bp) around a gap	vcf-annotate -f d=20/w=10	(Danecek et al. 2011) ( <a href="https://vcf-tools.github.io/index.html">https://vcf-tools.github.io/index.html</a> )
11	include only sites that have passed filter	bcftools	1.9	sites that passed the filters imposed are classified as “PASS”	view -f PASS	(Li et al. 2009) ( <a href="http://samtools.github.io/bcftools/bcftools.html">http://samtools.github.io/bcftools/bcftools.html</a> )

#### **3.2.4.4 Kinship determination**

IBD estimation was performed using unlinked STR markers and a combination of STR and SNP markers via Familias v3 (Kling et al. 2014b), using the “blind search” approach, which allows a comparison of all samples. IBD coefficients were also estimated using the R package forrel v1.3.0 (Egeland et al. 2014; Kling et al. 2017; Vigeland and Egeland 2019).

An attempt was also made to estimate IBD using KING (Manichaikul et al. 2010), a tool for kinship estimation which can handle genotype and sequencing data, considering underlying population structure. However, the number of markers generated by the ForenSeq analysis is too low to give results using this tool (data not shown).

### **3.3 Results**

In this section the sequence data are described, reporting the overall quality of the runs. The sequence variation of the targeted markers is then described considering separately the allele length and the sequence variation in the repeat region. These analyses show an increase of sequence variation: isoalleles carry alleles that are indistinguishable by length, but can be distinguished by sequence analysis. Finally, kinship information is reconstructed considering only the length information and both length and flanking sequence, combining the available marker types (STRs, identity, ancestry and phenotypic SNPs).

#### **3.3.1 Sequence data quality**

Samples were sequenced using Primer Mix B (including the phenotypic and biogeographic ancestry SNPs) in 3 runs (30 samples per run plus 2 controls). Table 3.3 reports some summary statistics).

Due to issues with the quality of the DNA, seven samples were excluded from analysis (individual 3 from family 2, individual 4 from family 3 and individual 8 from family 4 [all also excluded from SNP chip data analysis, see Chapter 2, section 2.3.2], individual 11 from family 3, individual 6 from family 4, individual 14 from family 4, individual 13

from family 7; see Figure 2.4, Chapter 2), leaving 65 samples out of the 72 available for analysis. Ten identity SNPs that had more than 50% across-sample missingness (alleles not called by the UAS) were excluded: rs2269355 (34 missing samples out of 65), rs1523537 (34 out of 65), rs1031825 (43), rs10488710 (44), rs1493232 (46), rs1357617 (50), rs1015250 (57), rs7041158 (53), rs826472 (62), and rs1736442 (63). One phenotypic and one ancestry SNP were excluded from the analyses: rs1393350 (59 missing out of 65), rs3811801 (56). The phenotypic SNPs rs1805006 and rs201326893\_Y152OCH were also excluded as SNP calls and hence allele frequency information was not available in the reference dataset. In addition, the reference dataset does not include calls for three other markers (rs999428518, rs6076682, rs1021952013) that were called in the identity and ancestry SNP amplicon flanking regions via the MPS workflow. N29insA\_rs312262906 is included in the analysis, but under a different name (rs796296176).

**Table 3.3 Summary of the three runs performed on the MiSeq FGx (ForenSeq) of 70 samples.**

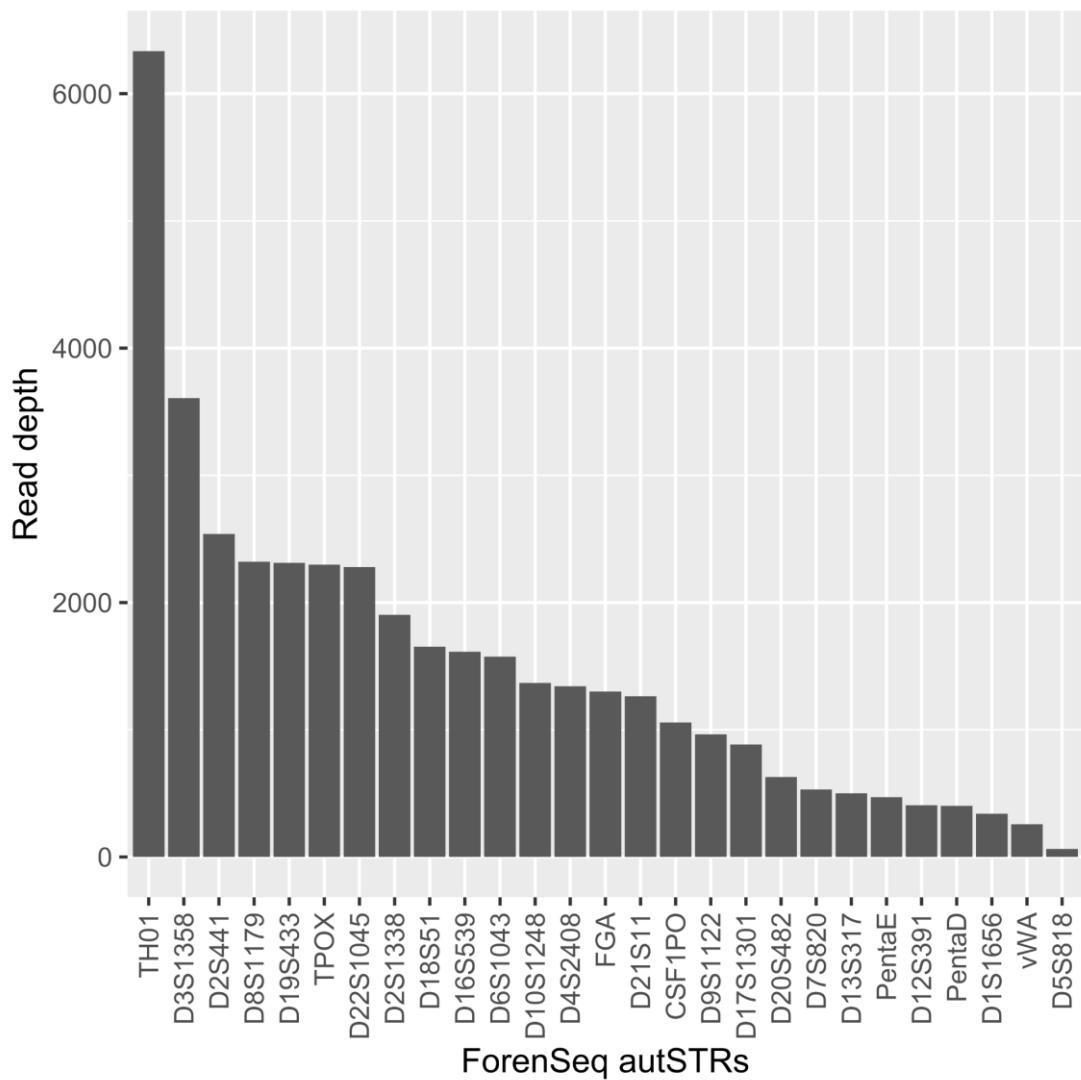
Run	Cluster density	Cluster passing filter	Phasing	Pre-phasing
1	641 k/mm <sup>2</sup>	95.23%	0.213%	0.106%
2	718 k/mm <sup>2</sup>	95.01%	0.16%	0.148%
3	711 k/mm <sup>2</sup>	96.12%	0.185%	0.144%

The sequences were processed through an in-house pipeline to retain only those with  $\geq 30$  reads; heterozygosity balance was set at 50% for STRs and 30% for SNPs. Additional checks were performed for loci known to show high imbalance: D22S1045 (according to manufacturer's protocol), PentaE, PentaD, D1S1656 (only for one sample), D5S818 (excluded from analysis).

### 3.3.2 Autosomal STR alleles and their sequence diversity

Here, sequence-based STR alleles are described, and below the difference in allele diversity between length- (CE equivalent) and sequence-based classifications is considered. Note that there are no independent CE data for the families, so concordance between CE- and MPS-based allele calls is not addressed here.

The DNA Signature Prep Kit Primer Mix B gave an average depth of coverage (DoC) per locus per individual of 1497 reads for the 27 autosomal STRs. Nine individuals had no calls, or sub-threshold calls, at D5S818. One individual (F2P4) had no reads at Penta E.



**Figure 3.3 Read depth of 27 autosomal STRs (ForenSeq markers) across 66 samples.**

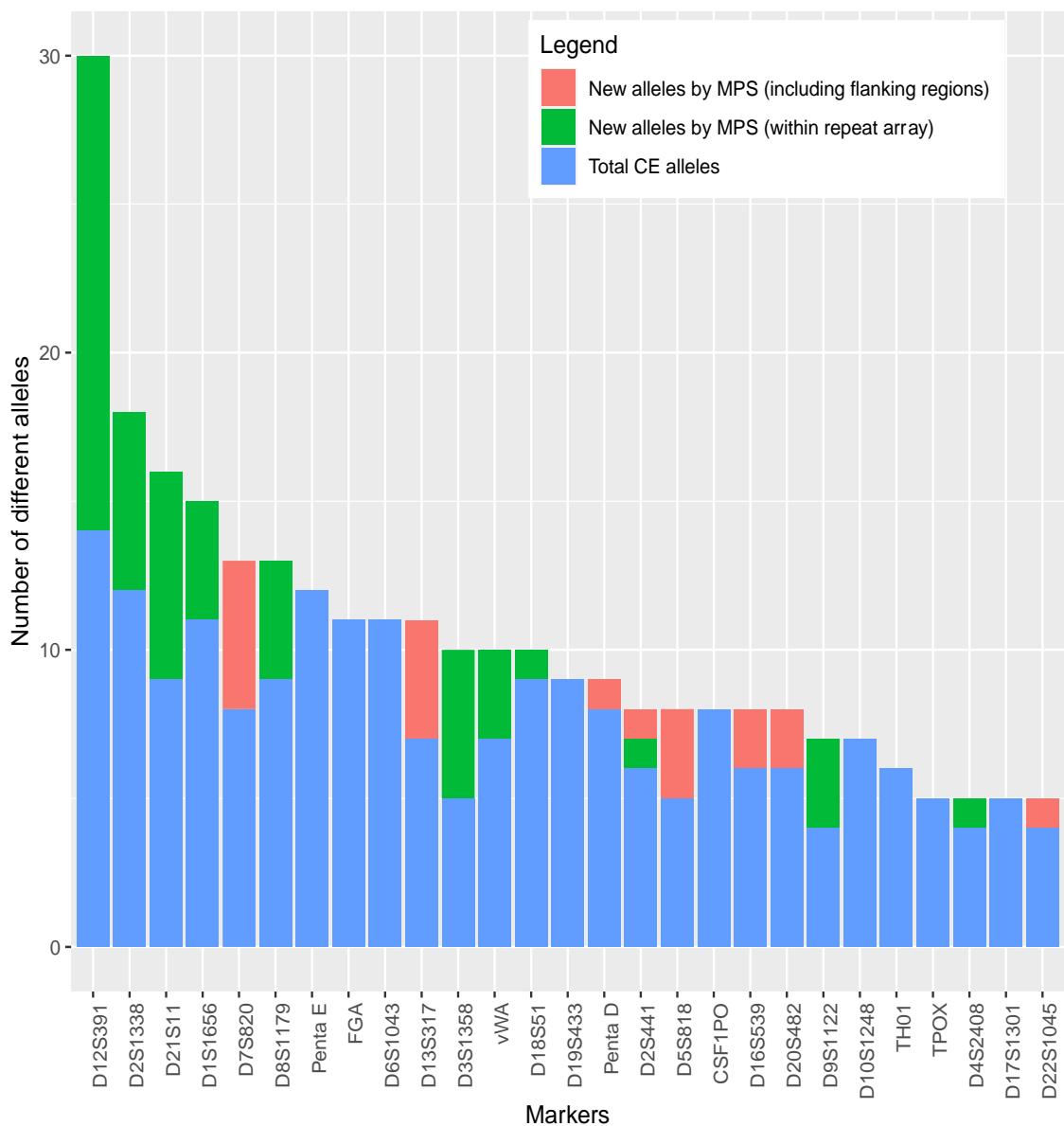
Defining truly novel alleles in MPS analysis is difficult because there is no live database, but as is conventional the observed alleles were compared with the STRSeq Bioproject database. On this basis, four novel alleles were found in the family data. The reality of these novel alleles is supported by the fact that they are all shared IBD between pairs of individuals known to be related: F8FP4 and F8P3, F2P2 and F2P5, and F7P4 and F7P5 are parent-offspring; F7P4 and F7P7 are full siblings.

**Table 3.4 Autosomal STR sequences reported as “Novel” by STRait Razor v3.**

Only the sequence at CSF1PO has no correspondence in STRSeq.

Locus	Allele (CE equivalent)	samples	Novel sequence nomenclature
CSF1PO	10.3	F8FP4, F8P3	CSF1PO [CE 10.3]-GRCh38-Chr5-150076318-150076389 [ATCT]5 ATC [ATCT]5
D12S391	23	F2P2, F2P5	D12S391 [CE 23]-GRCh38-Chr12-12296981-12297168 [AGAT]12 [AGAC]10 AGAT
	18	F7P4, F7P5, F7P7	D12S391 [CE 18]-GRCh38-Chr12-12296981-12297168 [AGAT]11 [AGAC]7
D22S1045	15	F7P14, F7P15	D22S1045 [CE 15]-GRCh38-Chr22-37140181-37140357 [ATT]12 ACT [ATT]2

As is generally the case, considering the sequences of STRs increases the observed diversity of alleles compared to a length-based consideration (Figure 3.4). The locus showing the greatest increase in allele diversity is D12S391, which has also been observed in other studies (e.g. Khubrani et al. 2019; Kim et al. 2017; Lareu et al. 1996). Eight loci show no increase in allele diversity on sequencing (D17S1301, D19S433, D6S1043, D9S1122, FGA, PentaE, TH01, TPOX). Note that, since the analysed individuals here are related, many alleles are identical by descent, so these are not independent occurrences of variants; for this reason, no statistical analysis of the increase in diversity or of match probabilities is carried out.



**Figure 3.4 Histogram showing the number of length alleles identified (Total CE alleles) and sequence alleles in the repeat region and flanking region.**

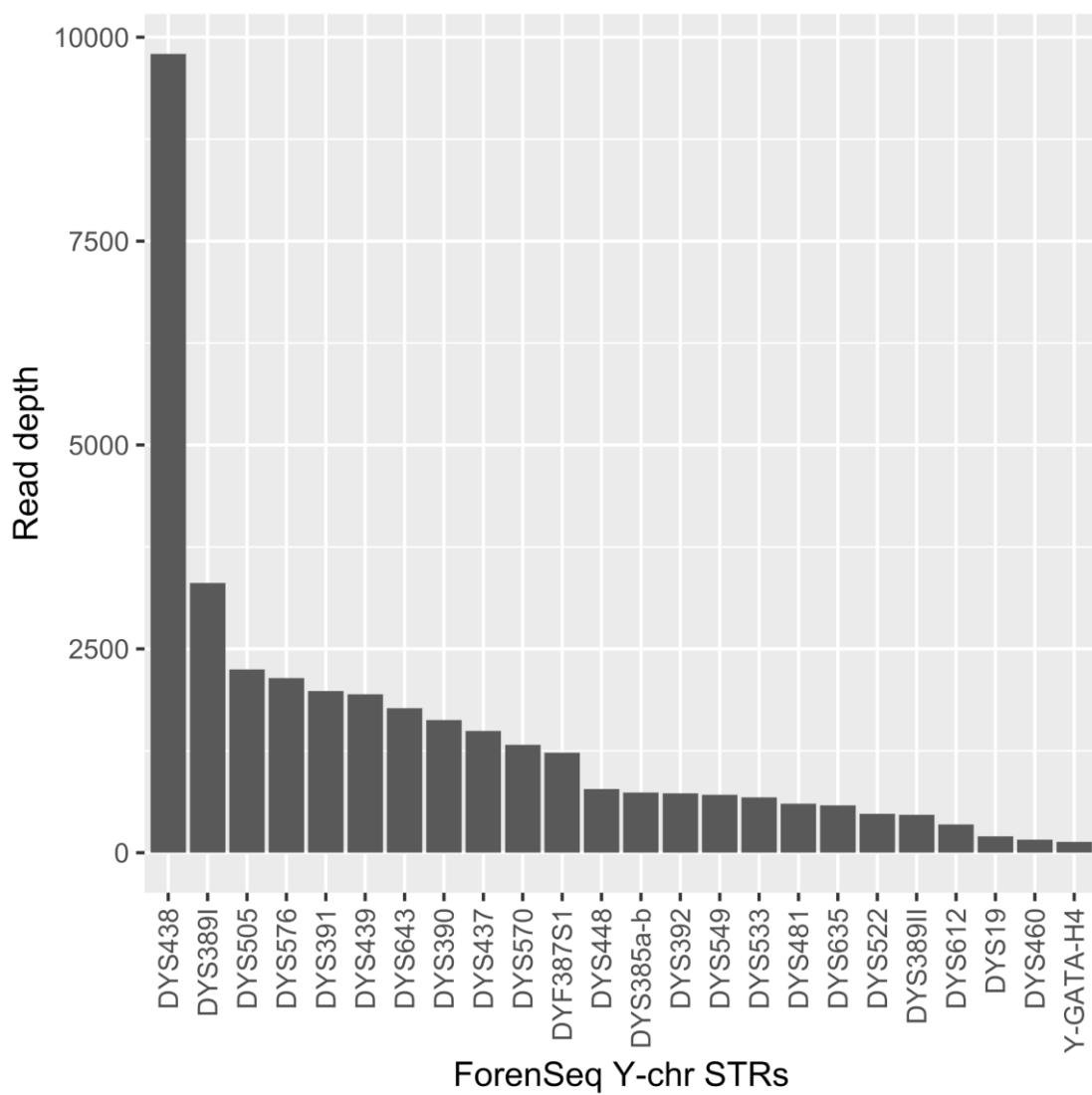
This shows the expected increase in allele diversity when sequence, rather than length is considered. Note that all individuals are included here, so many alleles are identical by descent.

As expected, the dataset contains examples of length homozygotes that are resolved by sequencing. The thirteen allele pairs found are described in Appendix 3a. For example, individual F2P1 is homozygous for allele 22 at D12S391, considering allele length alone, but sequencing shows two different underlying structures that are balanced in DoC ([AGAT]13 [AGAC]9 and [AGAT]13 [AGAC]8 AGAT).

### 3.3.3 Y STR alleles and their sequence diversity

An average depth of coverage (DoC) per locus per male of 1483 reads was obtained for the 24 Y-STRs. UAS and STRait Razor are generally consistent in the way they call Y-STR alleles, but at DYS612 the repeat array is considered differently: STRait Razor includes a 5' [CCT]5 CTT motif in the repeat count, while UAS does not, so CE-based allele calls are six repeat units shorter in the UAS output. Most STRs showed the expected number of copies and behaved consistently in the dataset. Exceptions were as follows:

- One male (F2P4) gave no calls at three loci (DYS19, DYS448, DYS460). These STRs are non-contiguous in the reference sequence (Hanson and Ballantyne 2006), suggesting that their absence is more likely to indicate amplification problems rather than a segmental deletion of Y-chromosomal DNA.
- The Y-STR DYF387S1 shows low sequence quality at the 3' end, which appears to introduce artifactual variants. This has been checked observing related pairs (expected to be IBD for the Y-chromosome) that show this high variation for length alleles 34-38 that were classified as "Novel". For this reason, ~7 bp is excluded from the end of the DYF387S1 sequences in further analyses.
- DYS612: Two samples (F7P10, F7P11) show two alleles (32 and 33). This STR has only recently been introduced into Y-chromosome typing and little published data exist; this literature gives no evidence for allele duplications.



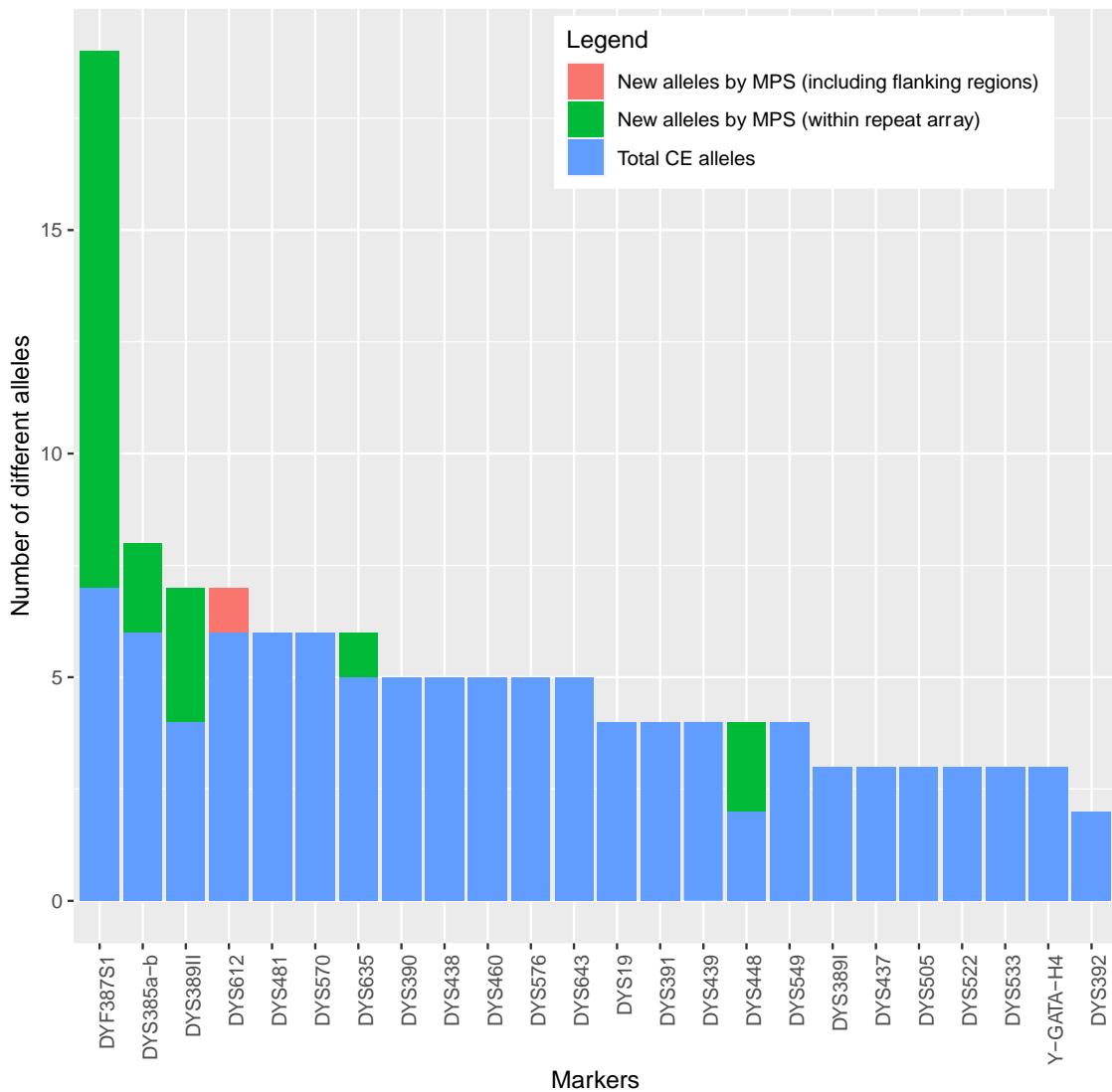
**Figure 3.5 Read depth of 24 Y- STRs (ForenSeq markers) across 28 male samples.**

Three novel alleles were found in the family data (Table 3.5). The “Novel” allele at DYS460 appears in the parent offspring pair (F4P4 and F4P5).

**Table 3.6 Y STR sequences reported as “Novel” by STRait Razor v3.**  
DYS460 has correspondence in STRSeq

Locus	Allele (CE equivalent)	samples	Novel sequence nomenclature
DYF387S1	35	F3P5	DYF387S1 [CE 35]-GRCh38-ChrY-23785347-23785500 [AAAG]2 GTAG [GAAG]4 [AAAG]2 GAAG [AAAG]2 [GAAG]9 [AAAG]13
DYS385 a/b	14	F3P7	DYS385 [CE 14]-GRCh38-ChrY-18639701-18639898 [TTTC]14
DYS460	13	F4P4, F4P5	DYS460 [CE 13]-GRCh38-ChrY:18888869-18889046 TCTG [TCTA]12, [CTAT]13
DYS448	20	F1P9	DYS448 [CE 20]-GRCh38-ChrY:22218904-22219083 [AGAGAT]12 [ATAGAG]2 [AGATAG]3 ATAGAT AGAGAA [AGAGAT]7 AGAAAT

Figure 3.6 shows the increased observed diversity of alleles of STR sequences compared to the length-based alleles. Six loci show an increase in allele diversity on sequencing (DYF387S1, DYS385a,b, DYS389II, DYS612, DYS635, DYS448), while eighteen loci do not. The locus showing the greatest increase in allele diversity is DYF387S1, which has also been observed in other studies (Khubrani et al. 2020; Li et al. 2021). Note that, since the analysed individuals here are related, many alleles are identical by descent, so these are not independent occurrences of variants; for the same reason, no statistical analysis of the increase in diversity or of match probabilities is carried out.



**Figure 3.6 Histogram showing the number of length alleles identified (Total CE alleles) and sequence alleles in the repeat region and flanking region.**

This shows the expected increase in allele diversity when sequence, rather than length is considered. Note that all individuals are included here, so many alleles are identical by descent.

The Y-chromosome is haploid, and single-copy STRs are expected to give a single amplified allele copy. However, a number of loci are duplicated on the chromosome, either constitutively (DYS385a,b; DYF387S1) or sporadically. When a single-length allele is observed for constitutively duplicated loci, these are referred to as homoallellic combinations (Balaresque et al. 2014), and can sometimes be resolved by MPS data. Examples of such alleles found in the dataset are described in Table 3.7. A duplication at DYS448 has been previously observed, linked to the haplogroup E1 (Balaresque et al. 2008): individual F1P9 shows two copies of allele 20 but with different sequences, and his Y chromosome belongs to E-M215 E1b1.

**Table 3.7 Isoalleles resolved by sequencing.**

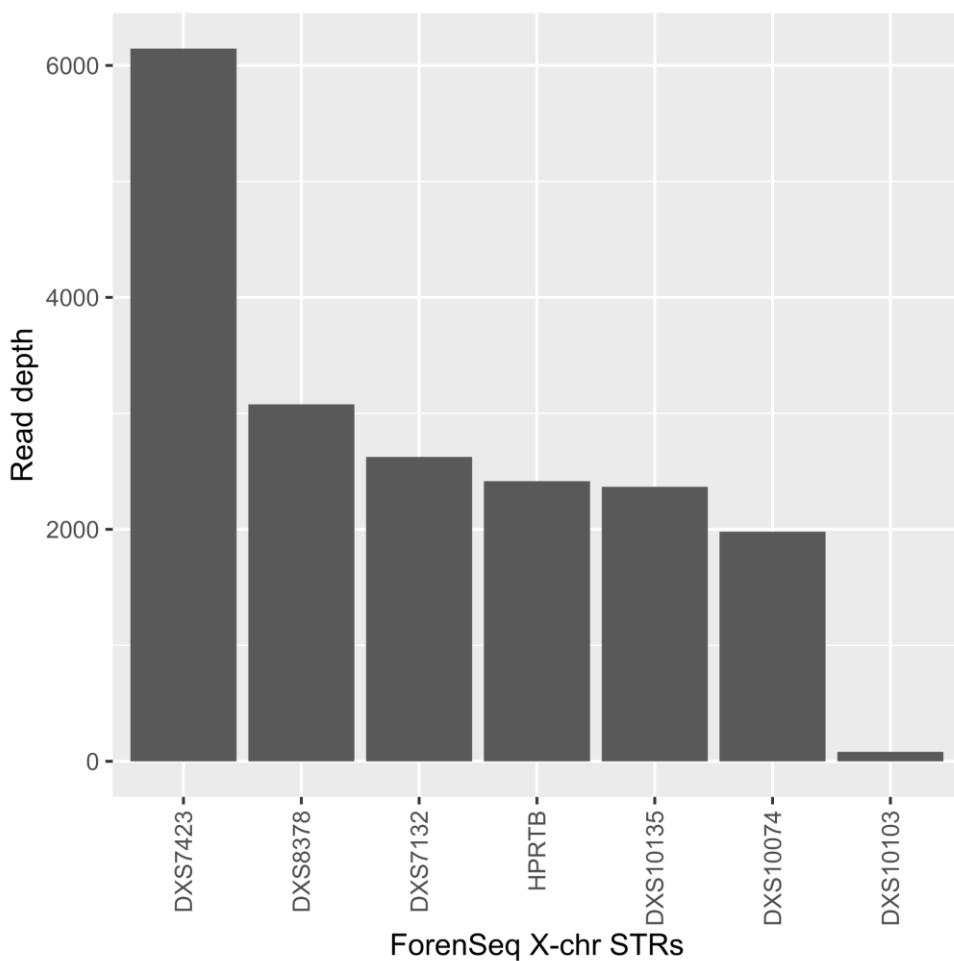
Length homozygotes reported with their nomenclature here.

Locus	Sample	Allele	DoC	Nomenclature
DYS448	F1P9	20	392	DYS448 [CE 20]-GRCh38-ChrY:22218904-22219083 [AGAGAT]12 [ATAGAG]2 [AGATAG]3 ATAGAT AGAGAA [AGAGAT]7 AGAAAT
			427	DYS448 [CE 20]-GRCh38-ChrY:22218904-22219083 [AGAGAT]12 [ATAGAG]2 [AGATAG]3 ATAGAT AGAGAA [AGAGAT]8
DYS612	F2P4	31	50	DYS612 [CE 31]-GRCh38-ChrY-13640705-13640861 [CCT]5 CTT [TCT]4 CCT [TCT]26 rs1603488779-A
			75	DYS612 [CE 31]-GRCh38-ChrY-13640705-13640861 [CCT]5 CTT [TCT]4 CCT [TCT]26
	F4P5	31	93	DYS612 [CE 31]-GRCh38-ChrY-13640705-13640861 [CCT]5 CTT [TCT]4 CCT [TCT]26 rs1603488779-A
			109	DYS612 [CE 31]-GRCh38-ChrY-13640705-13640861 [CCT]5 CTT [TCT]4 CCT [TCT]26
DYF387 S1	Fam3P5	35	1462	DYF387S1 [CE 35]-GRCh38-ChrY-23785347-23785500 [AAAG]3 GTAG [GAAG]4 [AAAG]2 GAAG [AAAG]2 [GAAG]10 [AAAG]12
			1496	DYF387S1 [CE 35]-GRCh38-ChrY-23785347-23785500 [AAAG]3 GTAG [GAAG]4 [AAAG]2 GAAG [AAAG]2 [GAAG]9 [AAAG]13
	Fam5P6	37	1067	DYF387S1 [CE 37]-GRCh38-ChrY-23785347-23785500 [AAAG]3 GTAG [GAAG]4 [AAAG]2 GAAG [AAAG]2 [GAAG]9 [AAAG]15
			1075	DYF387S1 [CE 37]-GRCh38-ChrY-23785347-23785500 [AAAG]3 GTAG [GAAG]4 [AAAG]2 GAAG [AAAG]2 [GAAG]10 [AAAG]14
	Fam4P15	38	795	DYF387S1 [CE 38]-GRCh38-ChrY-23785347-23785500 [AAAG]3 GTAG [GAAG]4 [AAAG]2 GAAG [AAAG]2 [GAAG]9 [AAAG]16
			832	DYF387S1 [CE 38]-GRCh38-ChrY-23785347-23785500 [AAAG]3 GTAG [GAAG]4 [AAAG]2 GAAG [AAAG]2 [GAAG]10 [AAAG]15
	Fam4P16		1107	DYF387S1 [CE 38]-GRCh38-ChrY-23785347-23785500 [AAAG]3 GTAG [GAAG]4 [AAAG]2 GAAG [AAAG]2 [GAAG]9 [AAAG]16
			1116	DYF387S1 [CE 38]-GRCh38-ChrY-23785347-23785500 [AAAG]3 GTAG [GAAG]4 [AAAG]2 GAAG [AAAG]2 [GAAG]10 [AAAG]15

### 3.3.4 X-STR allele description

An average depth of coverage (DoC) per locus per individual of 2680 reads was obtained for the 7 X-STRs for 28 males and 37 females (102 X-chromosomes). DXS10103 gave the lowest read coverage, with three sample drop-outs (F3P7, F8P5, F6P7; under the 30-read threshold). This was found in other studies as well (Hollard et al. 2019; Khubrani et al. 2020; Köcher et al. 2018; Sharma et al. 2020). Also DXS8378 showed one sample drop-out (F2P4, <30 reads).

There was only one length homozygote resolved by sequence data: individual F2P6 is homozygous for the length allele 25 at DXS10135, but shows two different sequences ([AAGA]3 GAAA GGA [AAGA]21 AAAG with 2228 reads and [AAGA]3 GAAA GGA [AAGA]17 [AAGG]3 AAGA AAAG with 2512 reads).



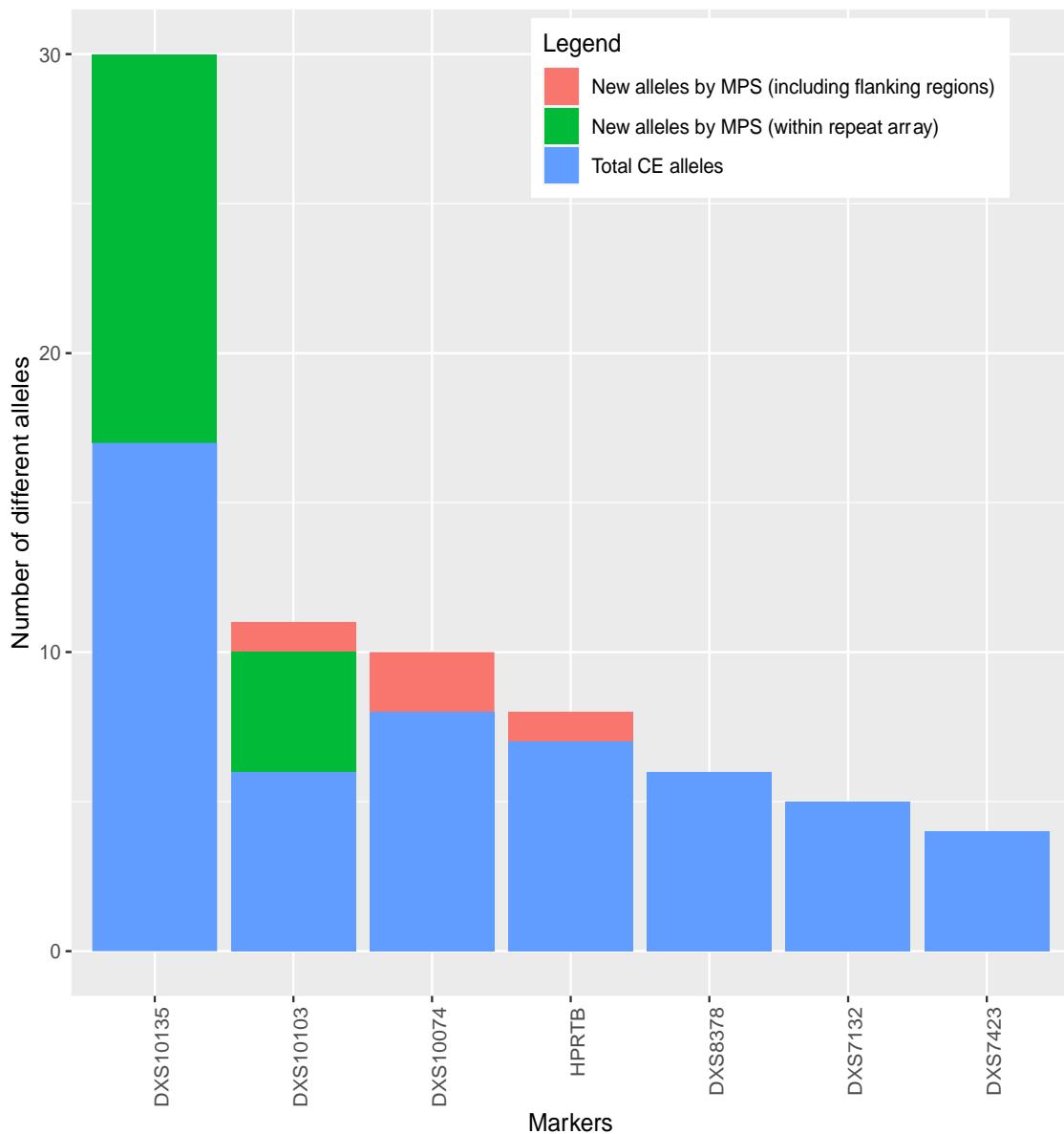
**Figure 3.7 Read depth of 7 X-STRs across 65 samples (28 males, 37 females).**

Three novel alleles were found in the family data (Table 3.8). The alleles at DXS10103 may be IBD (F8P1 and F8P7 are grandmother and granddaughter).

**Table 3.8 X-STR sequences reported as “Novel” by STRait Razor v3 with no correspondence in STRSeq.**

Locus	Allele (CE equivalent)	Sample	Novel sequence nomenclature
DXS10103	17	F8P1, F8P7	DXS10103 [CE 17]-GRCh38-ChrX-134284939-134285040 [TAGA]2 CTGA CAGA [TAGA]9 [CAGA]4 CAGA
DXS10135	22	F4P17	DXS10135 [CE 22]-GRCh38-ChrX-9338302-9338453 [AAGA]3 GAAA GGA [AAGA]11 [AAGG]2 [AAGA]3 AAGG AAGA AAAG
HPRTB	11	F7P10	HPRTB [CE 11]-GRCh38-ChrX-134481429-134481588 [ATCT]11 rs774347605-G

The observed allele diversity is higher considering the sequence than length-based alleles (Figure 3.8). The locus showing the greatest increase in allele diversity is DXS10135, which has also been observed in other studies (Deng et al. 2017; Li et al. 2021; Nagai and Bunai 2011; Sumita and Whittle 2009). Three loci show no increase in allele diversity on sequencing (DXS8378, DXS7132, DXS7423). Note that, since the analysed individuals here are related, many alleles are identical by descent, so these are not independent occurrences of variants; for the same reason, no statistical analysis of the increase in diversity or of match probabilities is carried out.

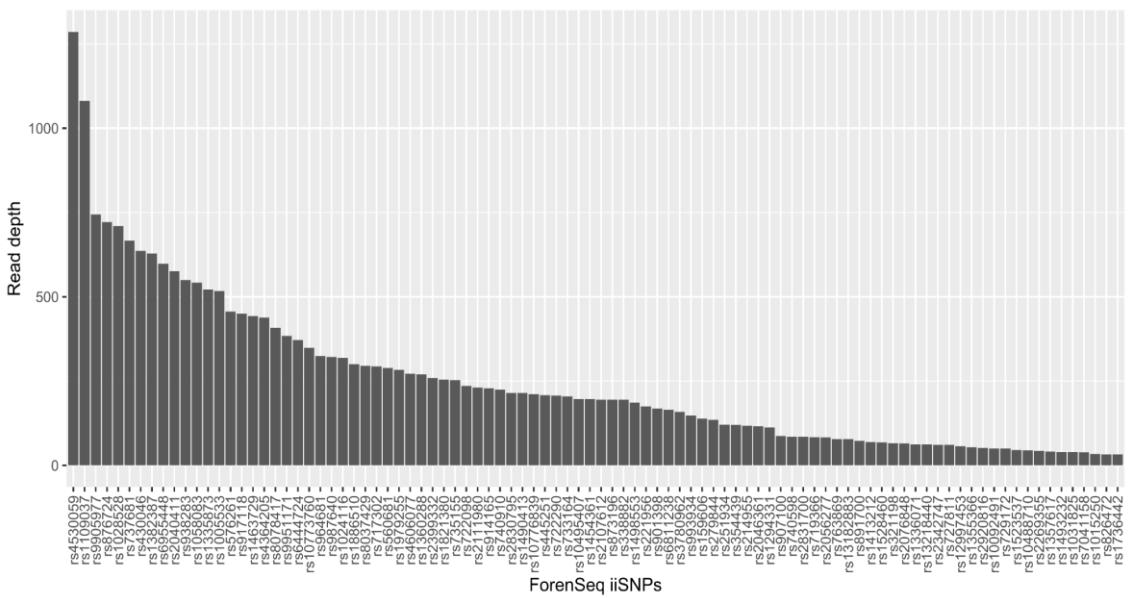


**Figure 3.8 Histogram showing the number of length alleles identified (Total CE alleles) and sequence alleles in the repeat region and flanking region.**

This shows the expected increase in allele diversity when sequence, rather than length is considered. Note that all individuals are included here, so many alleles are identical by descent.

### 3.3.5 SNP description

An average depth of coverage (DoC) per locus per individual of 274 reads was obtained for the 94 identity SNPs. The ten identity SNP markers excluded because of high missingness had low DoC: rs1736442 (33 reads), rs826472 (33 reads), rs1015250 (34), rs7041158 (39), rs1031825 (39), rs1493232 (40), rs1357617 (41), rs2269355 (43), rs10488710 (44), rs1523537 (45).

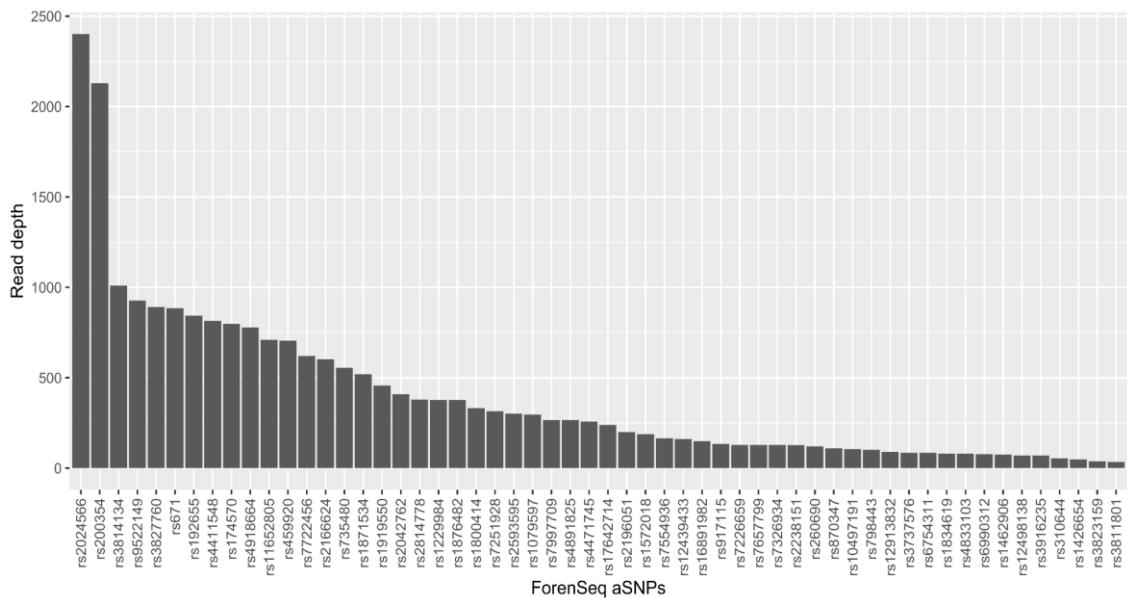


**Figure 3.9 Read depths of 94 ForenSeq identity SNPs.**

The average DoC per locus per individual was 417 reads for the ancestry SNPs (aSNPs) and 554 reads for the phenotypic SNPs (pSNPs) (based on STRait Razor output). Among the aSNPs, rs3811801 was called for one sample only (F7P14, 33 DoC) and is excluded from the analysis. Other SNPs with high drop-out are: rs3823159 (missing in 50 samples out of 65, 38 DoC), rs1426654 (24 out of 65, 48 DoC), rs310644 (18, 54 DoC), rs870347 (13, 109 DoC), rs1834619 and rs12498138 (5, 81 and 70 DoC respectively), rs4833103 (4, 81 DoC), rs6754311 and rs6990312 and rs3916235 (3, with 85 and 77 and 69 DoC respectively), rs10497191 and rs12439433 (2, with 105 and 160 DoC respectively), rs1462906 (F6P4, 75 DoC), rs7226659 (F1P7, 129 DoC).

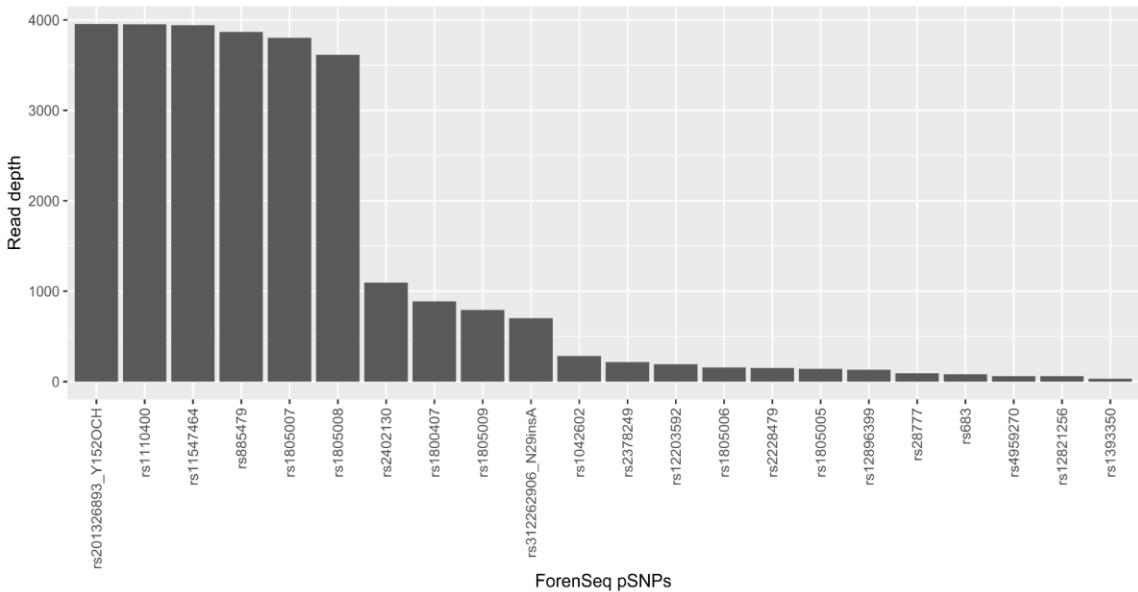
Four samples had high missingness: F6P4 had no calls at rs6754311, rs12498138, rs3811801, rs4833103, rs3823159, rs1462906, rs1426654, rs3916235, rs310644; F6P7 had no calls at rs1834619, rs6754311, rs12498138, rs3811801, rs3823159, rs6990312, rs12439433, rs1426654; F3P7 had no calls at rs10497191, rs1834619, rs6754311, rs3811801, rs4833103, rs3823159, rs6990312, rs1426654, rs310644; F1P9 had no calls at rs6990312, rs12439433, rs260690, rs16891982, rs917115, rs1462906, rs2238151, rs12913832.

There are four monomorphic SNPs: rs3827760, rs1871534, rs671, and rs1800414 are fixed in the dataset. Other markers have high frequency for only one allele: rs2814778 (found in only one individual), rs7326934 (found in only one individual), rs3916235 (found in only one individual), rs1876482 (found in two individuals), rs3814134 (found in two individuals), rs3737576 (found in three individuals), rs7997709 (found in three individuals), rs7226659 (found in three individuals), rs1462906 (found in five individuals).



**Figure 3.10 Read depths of 56 ForenSeq ancestry SNPs.**

Among the pSNPs, rs1393350 (33 DoC) was called for only one sample (F7P9) and is excluded from the analysis. Other SNPs with high drop-out are: rs12821256 (missing in 16 samples out of 65, 58 DoC), rs683 (13, 79 DoC), rs4959270 (5, 58.8 DoC), rs28777 (4, 90.26 DoC). Four samples had some SNP missingness: F1P7 (rs28777, rs4959270, rs683, rs1393350, rs12821256), F1P9 (rs28777, rs683, rs1393350, rs12821256). Only one sample (F5P7) showed the A insertion at N29insA. rs28777 has high frequency for only one allele (found in only four individuals).



**Figure 3.11 Read depths of 22 ForenSeq phenotypic SNPs.**

### SNP allele description: flanking region

It was possible to detect variation outside the target region (see Appendix 3d):

- 79 amplicons (with 69 in identity SNP regions, 7 in ancestry SNP regions, 3 in phenotypic SNP regions), contain additional variants;
- ten variants were found outside the UAS reported sequence (Table 3.9);
- there is evidence of crossover or back mutation in only two cases, at rs2830795 (identity SNP), where all four allelic combinations are observed (**TAA**, **CAA**, **TAG**, **CAG**), or at rs1109037 (**ACGGGG**, **ACGGGA**, **GCGGGA**, **GCAGGG**); there is evidence of this phenomenon in other research (Khubrani et al. 2019).

### 3.3.6 Custom variant calling workflow results

FastQC was used to check the quality of the raw FASTQ files reporting an average number of raw reads per sample of 247532 (average 34.45% of GC content). Adapters and low-quality bases were excluded with an average of reads survival rate of 247140 (~99%). The quality of bases improved, while the amount of overrepresented sequences deemed adaptors decreased. The overrepresented sequences left are taken care of by the mapping tool. The average number of reads is 247000 (average 40.12% of GC content).

After alignment, the mean depth of coverage per locus per individual (samtools coverage, across 10854 aligned positions) was 67. Around 179 variants for each sample were called.

The custom variant calling workflow was used to call SNPs as some are not reliably called by the UAS (Khurani et al. 2019). Among the called flanking variants (Table 3.9), two were called by the UAS as well (rs1390470, rs300773), and two have no rsID (GRCh37 chr16:5,605,961, chr16:5,605,970). It is important to highlight that all other flanking variants reported in Table 3.9 are found outside the sequence considered by both UAS and STRait Razor. Among these variants, rs300773 and rs876724 are in linkage disequilibrium ( $r^2$  0.1062), rs2342748 and rs2342747 (1), and rs1869434 and rs4606077 (1). rs6076682 and rs1021952013 are not in the 1KGP panel.

**Table 3.9 Variants called in the flanking region of ForenSeq SNPs through the MPS workflow.**

\* only 3 samples have this variant; \*\* only 1 sample has this variant; + found in UAS as well. Note that the variants uniquely called via the MPS workflow are outside the UAS and STRait Razor reported sequence.

Flanking rsID	type	Chr:start-end (bp) hg19	Target rsID	Target pos. (bp)	Flanking pos. (bp)	DoC	marker region
rs381840	identity	16:78016651-78017451	rs430046	78017051	78017077	138.7	
rs1390470+	identity	6:165044934 - 165045734	rs727811	165045334	165045290	197.4	
? *	identity	16:5605797-56065972	rs729172	56061972	5605961	148.7	
rs999428518*					5605965	145.7	
? *					5605970	150.3	
rs300773+	identity	2:114574-115374	rs876724	114974	115035	211.9	
rs1005534	identity	20:39486710-39487510	rs1005533	39487110	39487218	194.2	
rs6076682	identity	20:4447083-4447883	rs1031825	444748325	4447420	116.6	
rs75920625	ancestry	8:145639281-145640081	rs1871534	145639681	145639654	58.9	
rs2342748	identity	16:5868300-5869100	rs2342747	5868700	5868729	126.3	
rs73514221					5868743	45.6	
rs12453170	ancestry	17:53568484-53569284	rs4471745	53568884	53568855	89.4	
rs1869434	identity	8:144656354-144657154	rs4606077	144656754	144656765	208.2	
rs1021952013*	ancestry	10:94920665-94921465	rs4918664	94921065	94921393	172.3	

### 3.3.7 Reference data description (HGDP)

To interpret the ForenSeq data in pedigrees, appropriate reference information is required to provide allele frequencies. Published population data that include all ForenSeq loci are based on US populations (Gettings et al. 2018b). It was desirable to use a more geographically appropriate reference, and therefore data from the Human Genome Diversity Project Panel (HGDP) (Cann et al. 2002b) were used here instead. To further investigate the impact of the allele frequencies choice, Appendix 3c shows some tests on the use of different population-based allele frequencies on a set of 16 forensic markers on the real-world family dataset.

The HGDP samples were previously sequenced using the ForenSeq kit (Plex A), and information on the STRs have been published (Phillips et al. 2018a). Raw FASTQ files were received from Chris Phillips for retrieving allelic frequencies of STRs and identity SNPs for the European metapopulation (total = 138, with males = 85, females = 53; Russian = 25; French = 28; French Basque = 23; Orcadian = 15; North Italian = 11; Tuscan = 8; Sardinian = 28). FASTQ files were processed using STRait Razor v3. The sequences were accepted if they reached the minimum number of 30 reads and minimum heterozygote balance of 50% for STRs and 30% for SNPs, with an in-house script. As the sequence data did not have an HGDP ID assigned (Christopher Phillips, University of Santiago de Compostela, personal communication), the STR profiles were compared against the known CE profiles (with correct HGDP IDs) from the same samples to retrieve the correct HGDP sample names (Phillips et al. 2018a). The sequences flagged as “NovelSeq” were further examined by comparing against the STRSeq database. Then allele frequencies were calculated for the European sample, considering the length allele and assigning a code to the sequence-based alleles.

Reference allele frequencies were compiled from these reference data. Sequences observed fewer than five times (from 0 to 5) were assigned the frequency of  $5/(2N)$ , where N is the number of individuals in the population database (National Research Council 1996). Sequences present uniquely in the family dataset and not in the reference dataset (unobserved allele) were included following the same rule.

### Description of sequence data

There are forty-nine sequences that were flagged as “NovelSeq” by the STRait Razor tool (see section 3.1.2), and a comprehensive table is in Appendix 3b. These were searched against the BLAST database, STRSeq BioProject, (Gettings et al. 2017; Gettings et al. 2017b):

- at D13S317, the reference shows a CTAT run followed by [CAAT]2; whereas the identified sequence shows a [CTAT]15 followed by a single CAAT: this change of the first reference CAAT is denoted as a SNP, rs9546005 (the sequence nomenclature is D13S317 15 [CTAT]15 rs9546005 A/T);
- at D6S1043, BLAST fails to highlight a T/G base change at 91740182 (changing the nomenclature from D6S1043 11 [ATCT]11 to D6S1043 11 [ATCT]11 rs1770275992-G);
- at D16S539, a NovelSeq is defined 12.1 by STRait Razor and 12 by BLAST with a variant (rs1728369) at position 86352607 (GRCh38, chromosome 16), falling outside the sequence (86352664-86352781 bp): this is considered as A insertion whereas rs1728369 is an A/C transversion polymorphism at the adjacent base;
- more discrepancies are seen at D19S433.

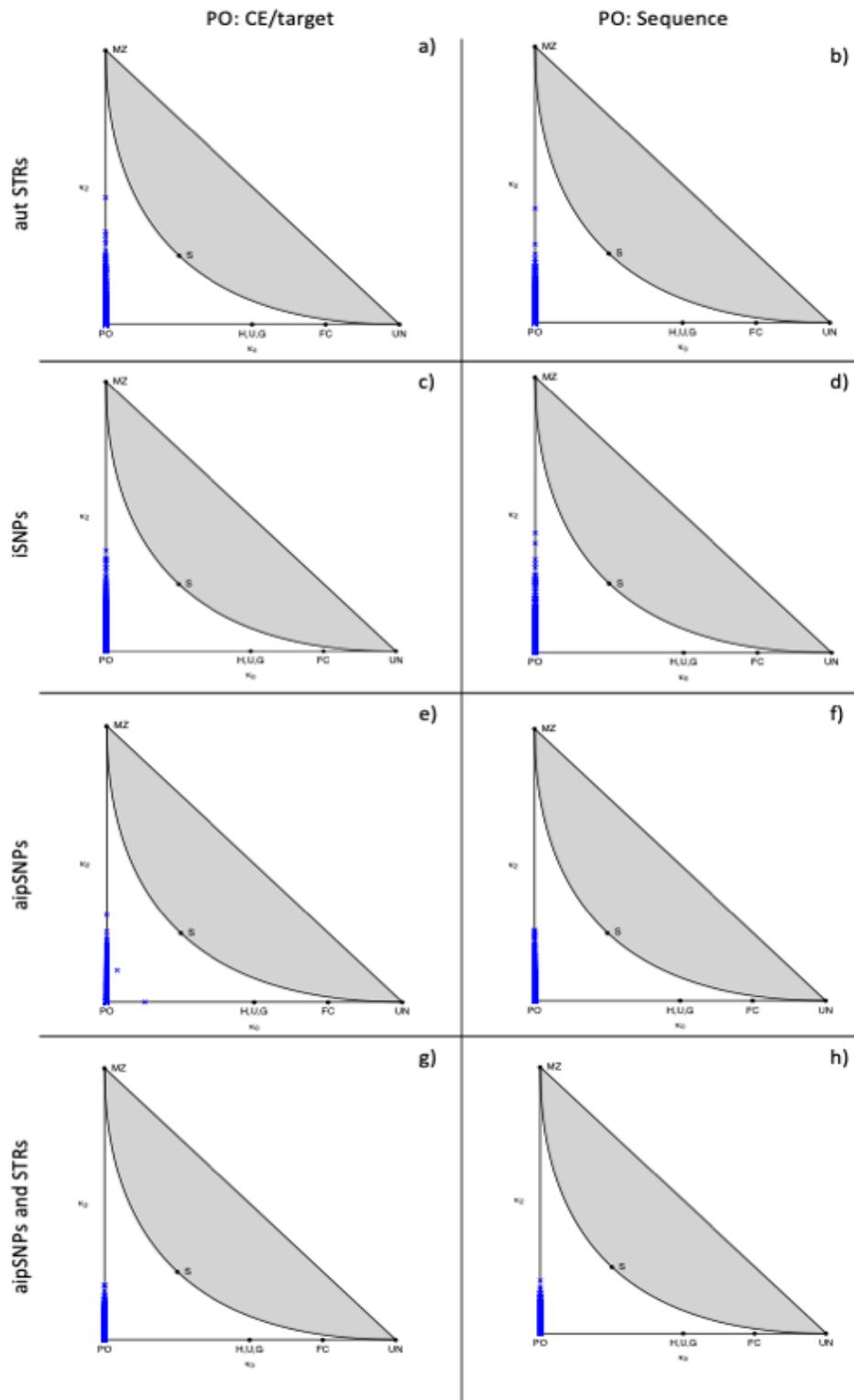
### 3.3.8 Evaluating the power of the markers in kinship determination

Traditional forensic profiles, based on the CE-based length of STR alleles, are the standard input for Familias, a tool for relationship deduction, alongside the markers’ allele frequencies in an appropriate reference population. In order to retain the additional information (e.g. isoallelic variants) offered by MPS data, the different STR sequences were coded, and the SNPs were included as if they were biallelic STRs.

The “blind search” option in Familias was used, allowing all samples to be compared simultaneously, and providing likelihood ratios, IBS state, kinship coefficient and percentage of allele sharing for each different relationship hypothesis (data not shown). Only a few relationship types are considered in this analysis by Familias (Parent-Offspring, Sibship, Half-sibship, Cousin, Second cousin), but IBS and kinship coefficients were used to improve the kinship deduction for different relationship types and more distant relationships.

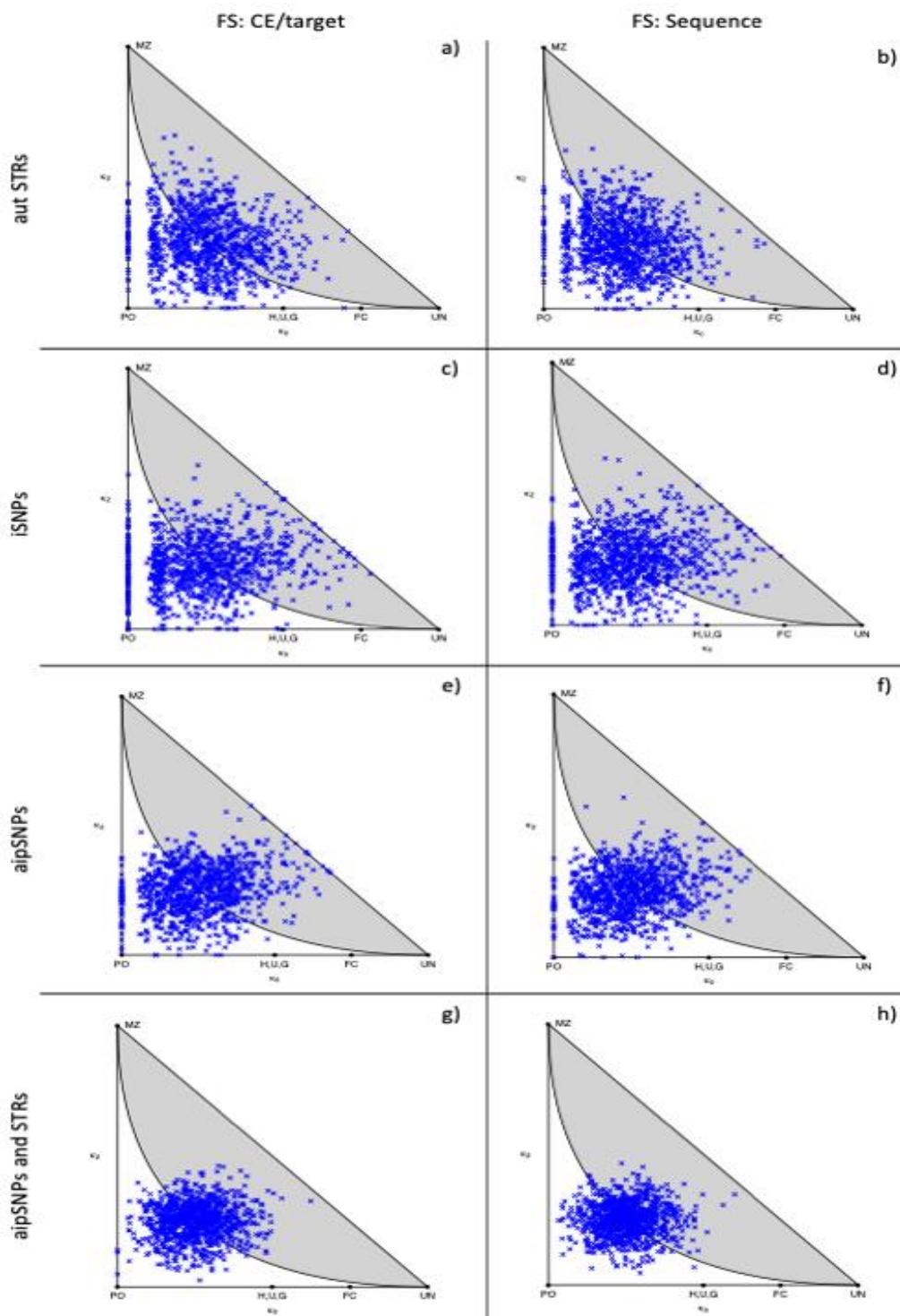
### **3.3.8.1 Simulations via forrel show the expected performance of ForenSeq markers in kinship coefficient estimation**

Prior to analysing the obtained data, the use of simulated data may help in understanding the application of the ForenSeq markers in kinship determination, showing what degree of variation is expected in using these markers for each kin relationship. Simulations (seed set to 12345) were carried out in R, using the forrel package: 1000 pairs with a specified relationship were created assigning a combination of ForenSeq markers (autosomal STRs and SNPs), with allele frequencies obtained from the European HGDP metapopulation panel (excluding the Adygei population), and the IBD coefficients were calculated and plotted on the IBD triangle as shown in Figures 3.12-16. A general trend among all plots is that adding more markers creates the most significant improvement in the observed pattern/clustering. The differences between length/target-based plots and sequence-based plots are minimal (e.g. Figure 3.13 a and b), but when sequence-based data are considered and more markers are included, it is possible to observe less variation on the estimated IBD values. This has an impact on all investigated relationships, obtaining more precise results when considering sequence STRs, identity, ancestry and phenotypic SNPs as well as variants in the flanking region (Figure 3.12-16 h). As expected, more distant relationships are clustering closer to values typical of unrelated pairs (Figure 3.15-16).



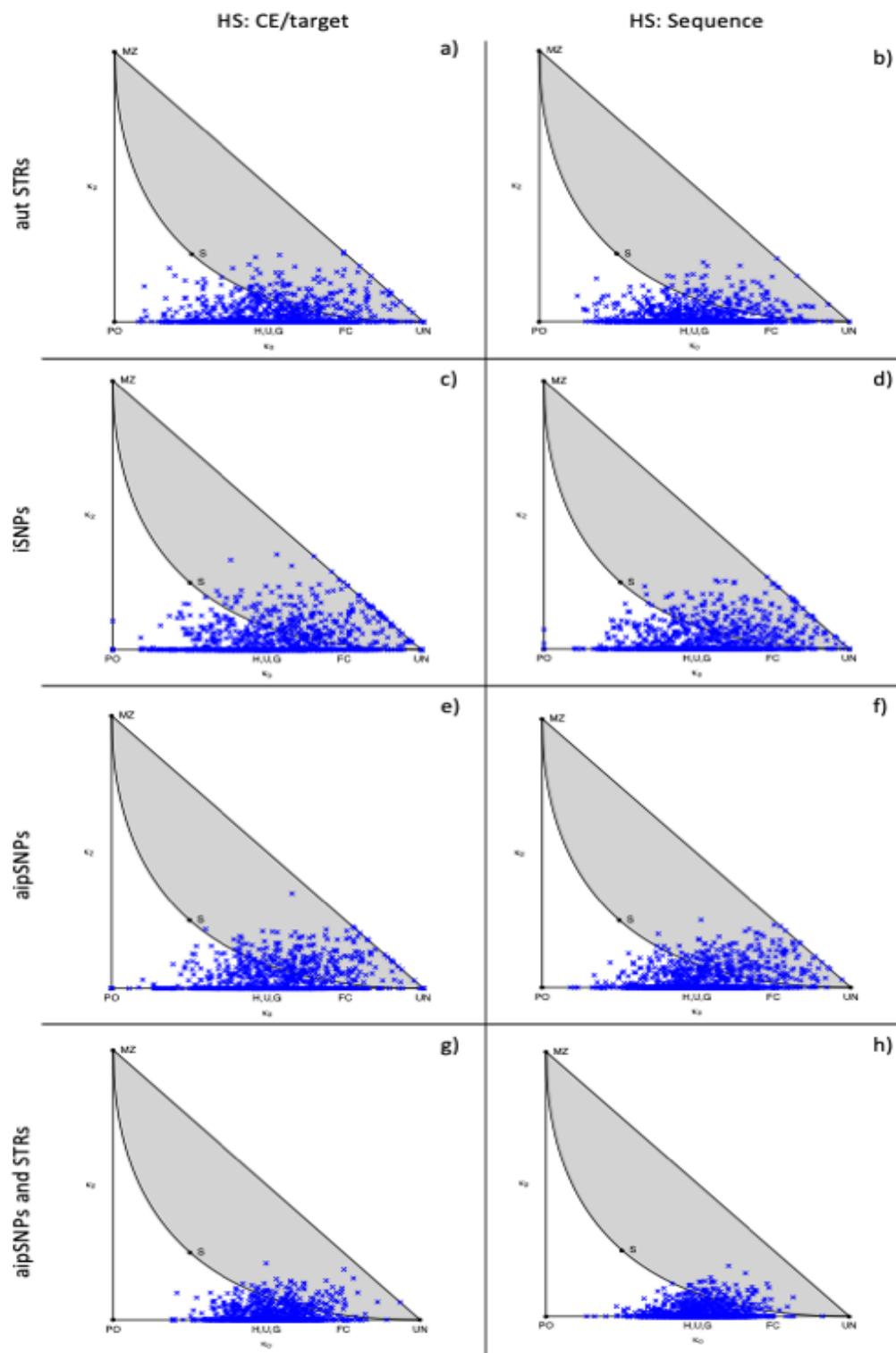
**Figure 3.12 Simulation of 1000 parent-offspring (PO) pairs.**

The parent-offspring relationship is simulated using the R package forrel, 1000 pairs of individuals are simulated (seed = 12345) considering (a) 27 autosomal length-based STRs; (b) 27 autosomal sequence-based STRs; (c) target identity SNPs; (d) identity SNPs (target and flanking); (e) target ancestry, identity and phenotypic SNPs; (f) ancestry, identity and phenotypic SNPs (target and flanking); (g) target ancestry, identity and phenotypic SNPs and length-based STRs (CE); (h) ancestry, identity and phenotypic SNPs (target and flanking) and autosomal sequence-based STRs.



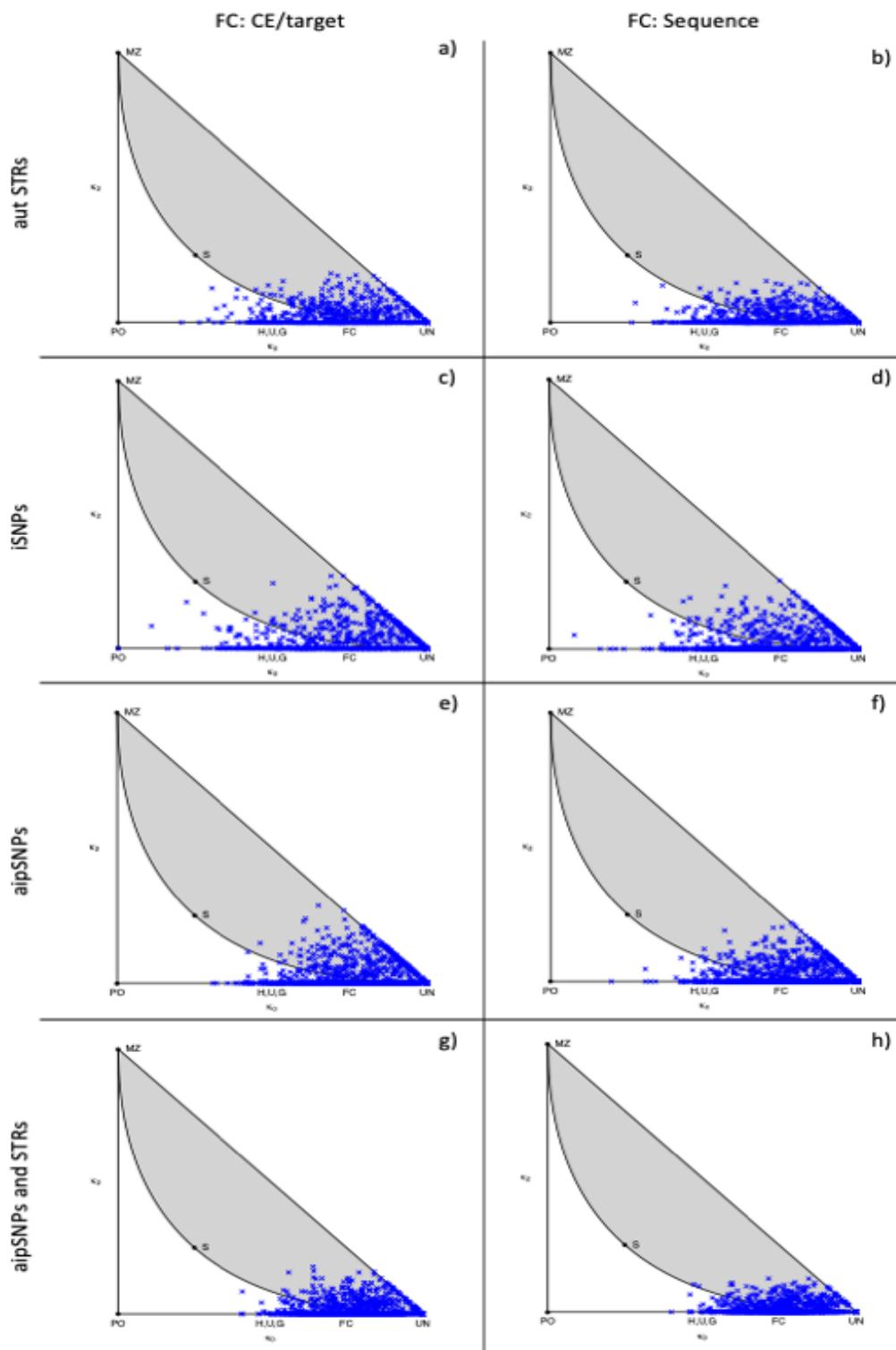
**Figure 3.13 Simulation of 1000 full siblings (FS) pairs.**

The full sibling relationship is simulated using the R package *forrel*, 1000 pairs of individuals are simulated (seed = 12345) considering (a) 27 autosomal length-based STRs; (b) 27 autosomal sequence-based STRs; (c) target identity SNPs; (d) identity SNPs (target and flanking); (e) target ancestry, identity and phenotypic SNPs; (f) ancestry, identity and phenotypic SNPs (target and flanking); (g) target ancestry, identity and phenotypic SNPs and length-based STRs (CE); (h) ancestry, identity and phenotypic SNPs (target and flanking) and autosomal sequence-based STRs.



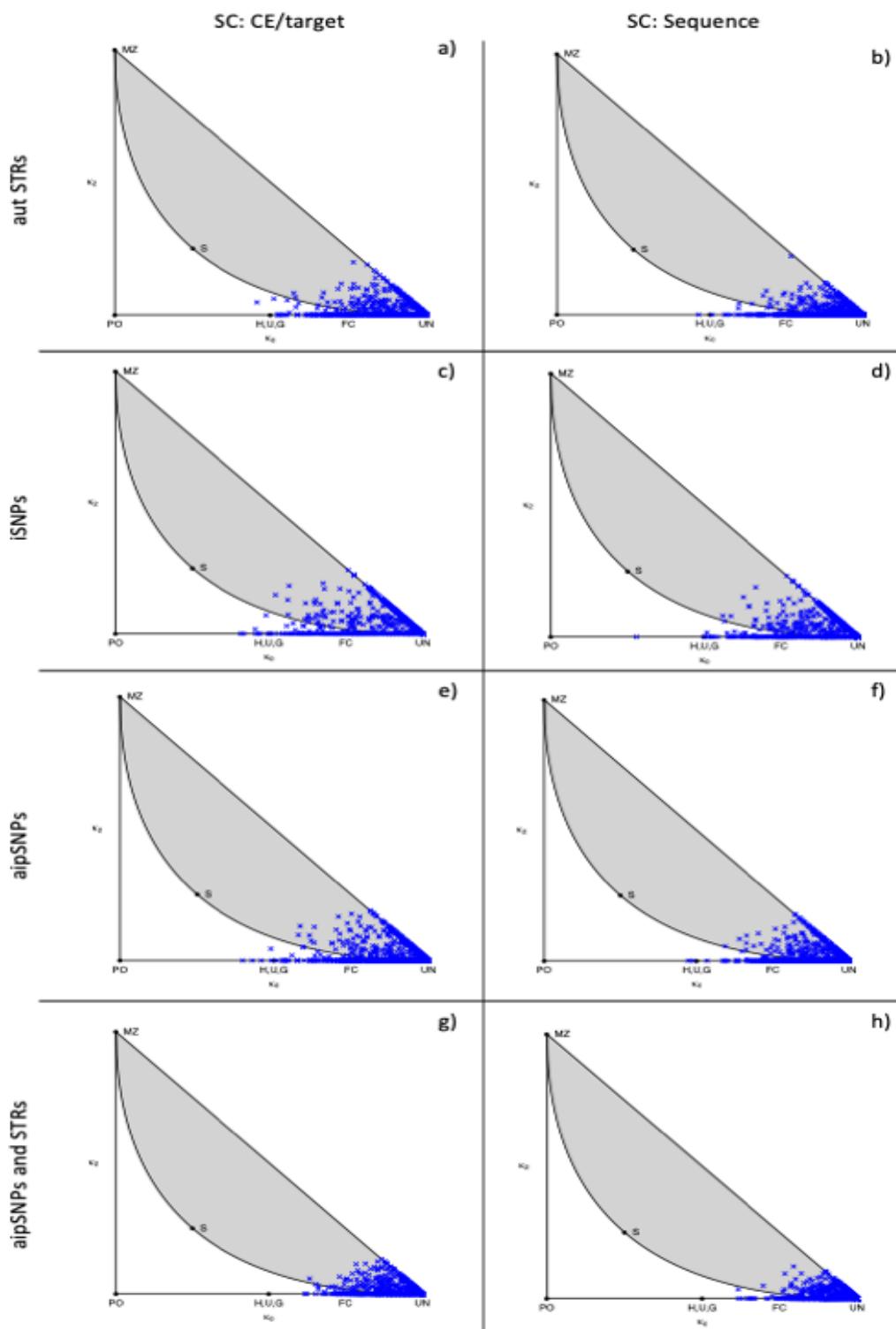
**Figure 3.14 Simulation of 1000 half siblings (HS) pairs.**

The half sibling relationship is simulated using the R package forrel, 1000 pairs of individuals are simulated (seed = 12345) considering (a) 27 autosomal length-based STRs; (b) 27 autosomal sequence-based STRs; (c) target identity SNPs; (d) identity SNPs (target and flanking); (e) target ancestry, identity and phenotypic SNPs; (f) ancestry, identity and phenotypic SNPs (target and flanking); (g) target ancestry, identity and phenotypic SNPs and length-based STRs (CE); (h) ancestry, identity and phenotypic SNPs (target and flanking) and autosomal sequence-based STRs.



**Figure 3.15 Simulation of 1000 first cousin (FC) pairs.**

The first cousin relationship is simulated using the R package forrel, 1000 pairs of individuals are simulated (seed = 12345) considering (a) 27 autosomal length-based STRs; (b) 27 autosomal sequence-based STRs; (c) target identity SNPs; (d) identity SNPs (target and flanking); (e) target ancestry, identity and phenotypic SNPs; (f) ancestry, identity and phenotypic SNPs (target and flanking); (g) target ancestry, identity and phenotypic SNPs and length-based STRs (CE); (h) ancestry, identity and phenotypic SNPs (target and flanking) and autosomal sequence-based STRs.

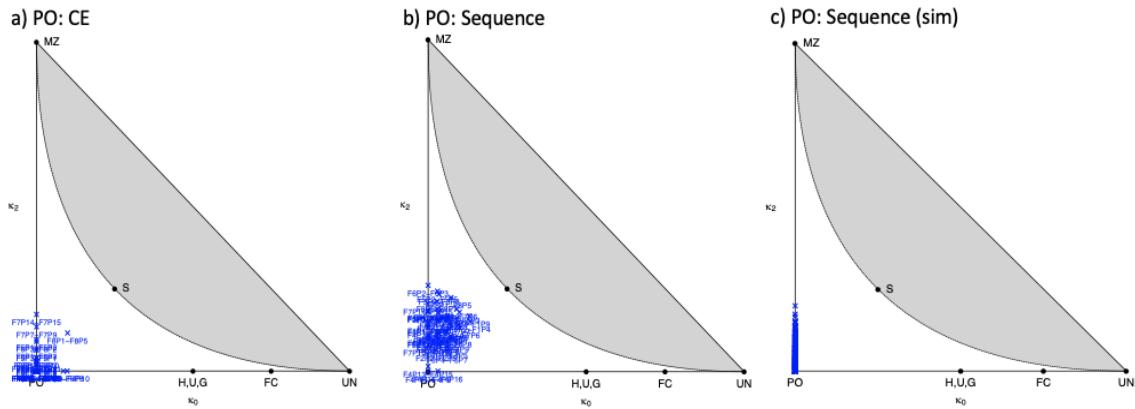


**Figure 3.16 Simulation of 1000 second cousin (SC) pairs.**

The second cousin relationship is simulated using the R package *forrel*, 1000 pairs of individuals are simulated (*seed* = 12345) considering (a) 27 autosomal length-based STRs; (b) 27 autosomal sequence-based STRs; (c) target identity SNPs; (d) identity SNPs (target and flanking); (e) target ancestry, identity and phenotypic SNPs; (f) ancestry, identity and phenotypic SNPs (target and flanking); (g) target ancestry, identity and phenotypic SNPs and length-based STRs (CE); (h) ancestry, identity and phenotypic SNPs (target and flanking) and autosomal sequence-based STRs.

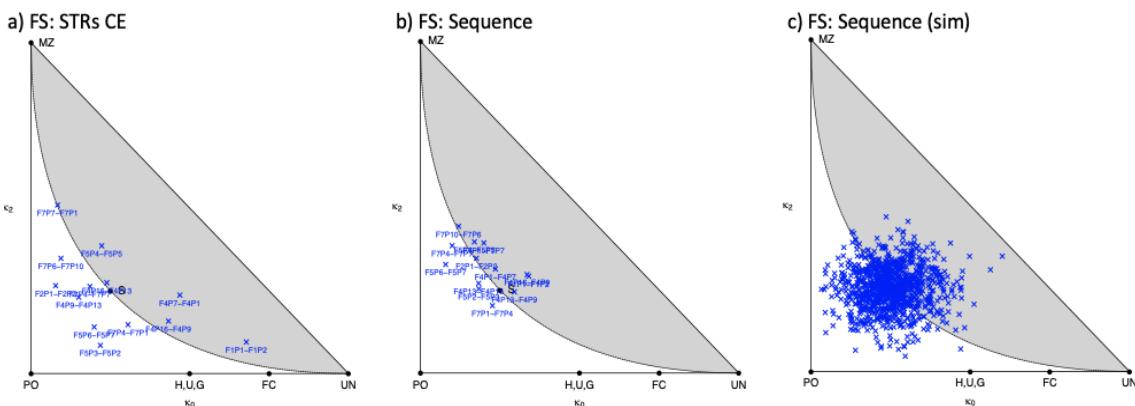
### 3.3.8.2 Determining kinship among real-world samples

The IBD coefficients for the MPS data of the family dataset were calculated using forrel. Figure 3.17-21 show the coefficient values (IBD0 against IBD2) for five relationships, comparing these results with the values obtained from simulations (Section 3.3.6.1). The higher number of markers appears to improve estimation for close relationships.



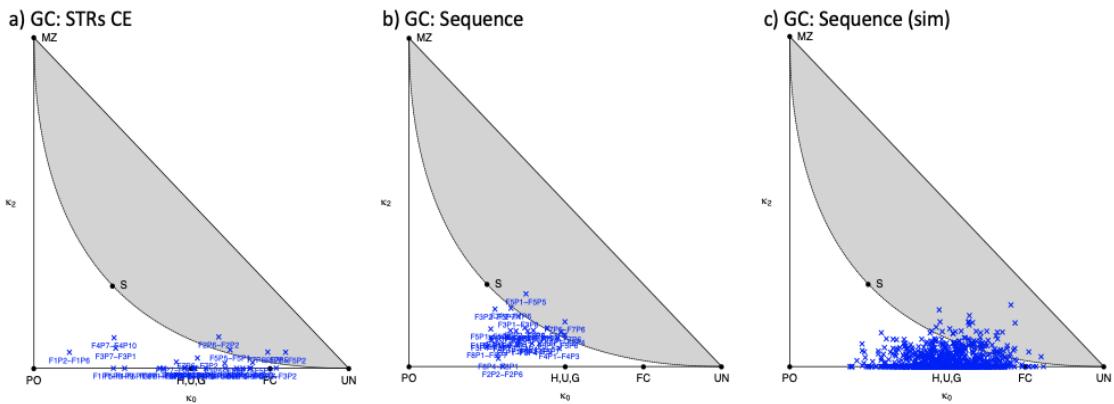
**Figure 3.17 Scatterplot of the 0 IBD ( $k_0$ ) proportion and 2 IBD ( $k_2$ ) for the parent-offspring (PO) relationship.**

The family data are analysed using the R package forrel considering (a) 27 autosomal length-based STRs; (b) ancestry, identity and phenotypic SNPs (target and flanking) and autosomal sequence-based STRs; and (c) simulated data for ancestry, identity and phenotypic SNPs (target and flanking) and autosomal sequence-based STRs.



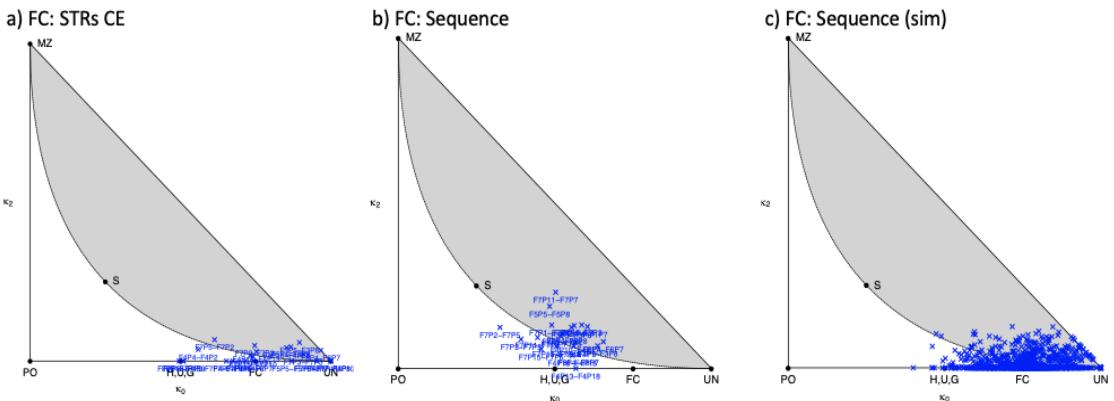
**Figure 3.18 Scatterplot of the 0 IBD ( $k_0$ ) proportion and 2 IBD ( $k_2$ ) for the full sibling (FS) relationship.**

The family data are analysed using the R package forrel considering (a) 27 autosomal length-based STRs; (b) ancestry, identity and phenotypic SNPs (target and flanking) and autosomal sequence-based STRs; and (c) simulated data for ancestry, identity and phenotypic SNPs (target and flanking) and autosomal sequence-based STRs.



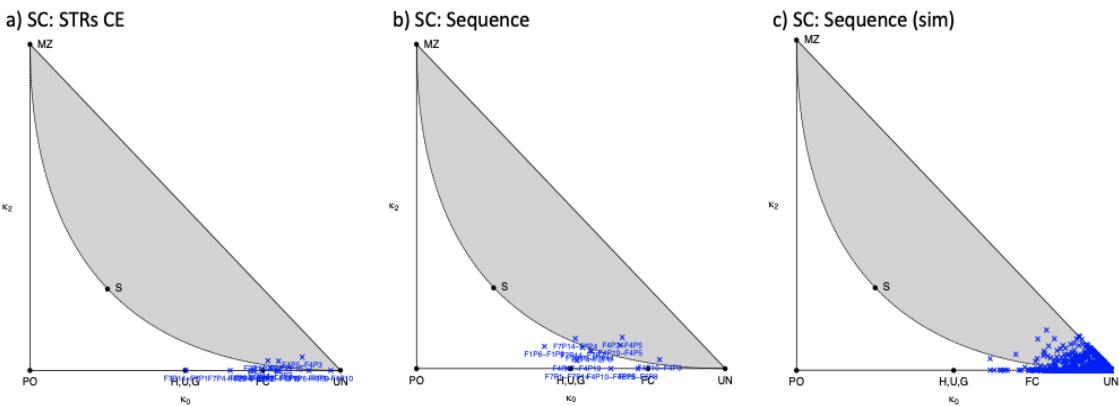
**Figure 3.19 Scatterplot of the 0 IBD ( $k_0$ ) proportion and 2 IBD ( $k_2$ ) for the grandparental (GC) relationship.**

The family data are analysed using the R package forrel considering (a) 27 autosomal length-based STRs; (b) ancestry, identity and phenotypic SNPs (target and flanking) and autosomal sequence-based STRs; and (c) simulated data for ancestry, identity and phenotypic SNPs (target and flanking) and autosomal sequence-based STRs.



**Figure 3.20 Scatterplot of the 0 IBD ( $k_0$ ) proportion and 2 IBD ( $k_2$ ) for the first cousin (FC) relationship.**

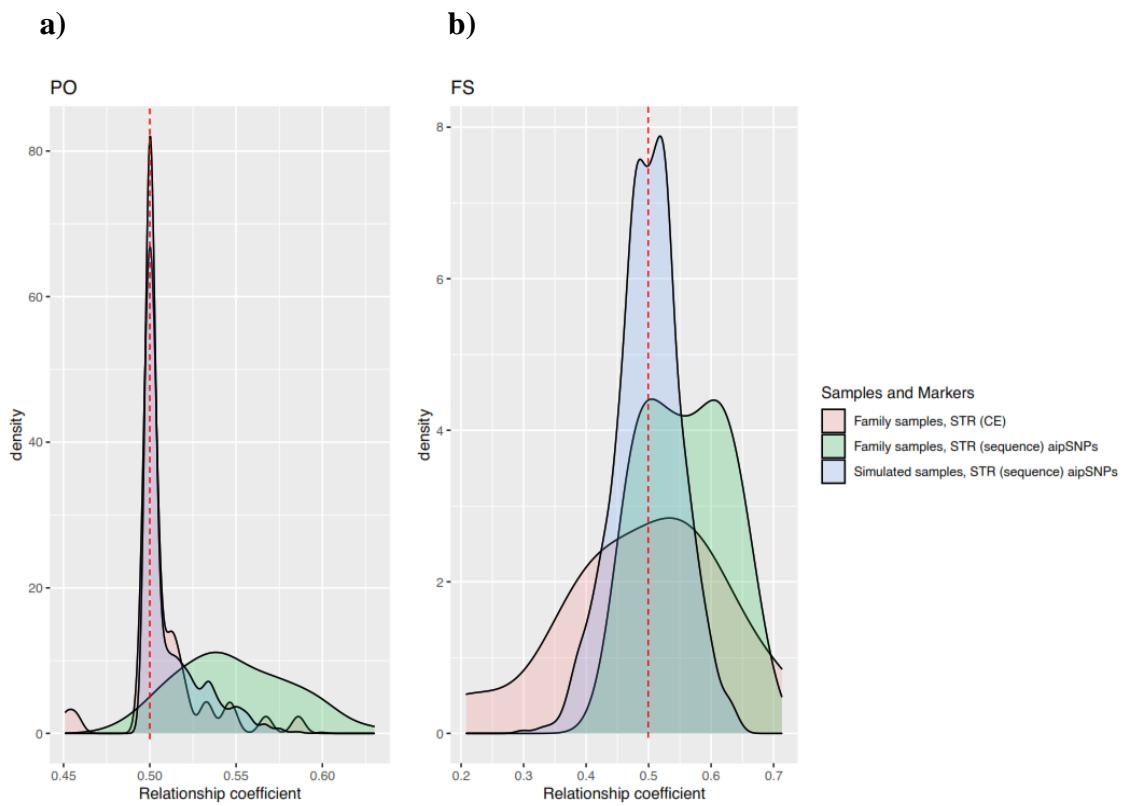
The family data are analysed using the R package forrel considering (a) 27 autosomal length-based STRs; (b) ancestry, identity and phenotypic SNPs (target and flanking) and autosomal sequence-based STRs; and (c) simulated data for ancestry, identity and phenotypic SNPs (target and flanking) and autosomal sequence-based STRs.



**Figure 3.21 Scatterplot of the 0 IBD ( $k_0$ ) proportion and 2 IBD ( $k_2$ ) for the second cousin (SC) relationship.**

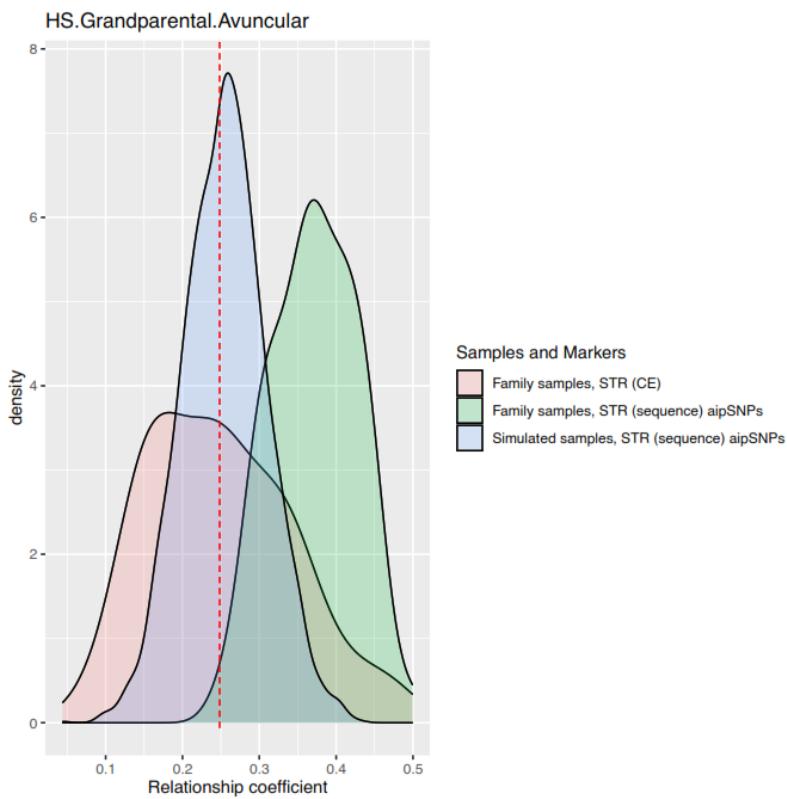
The family data are analysed using the R package *forrel* considering (a) 27 autosomal length-based STRs; (b) ancestry, identity and phenotypic SNPs (target and flanking) and autosomal sequence-based STRs; and (c) simulated data for ancestry, identity and phenotypic SNPs (target and flanking) and autosomal sequence-based STRs.

This has been further investigated comparing the distribution of the coefficient of relationship obtained from simulated data on sequence STRs, SNPs and flanking region variants against the real-world family data, considering the STRs (CE) or the sequence STRs, SNPs and flanking region variants (Figure 3.22-24). Some relationships appear overestimated when considering the real-world family data with the full set of markers (sequence STRs, SNPs and flanking region variants): for example, the expected value for parent-offspring is 0.5, however data distribution is closer to 0.54 (Figure 3.22 a)



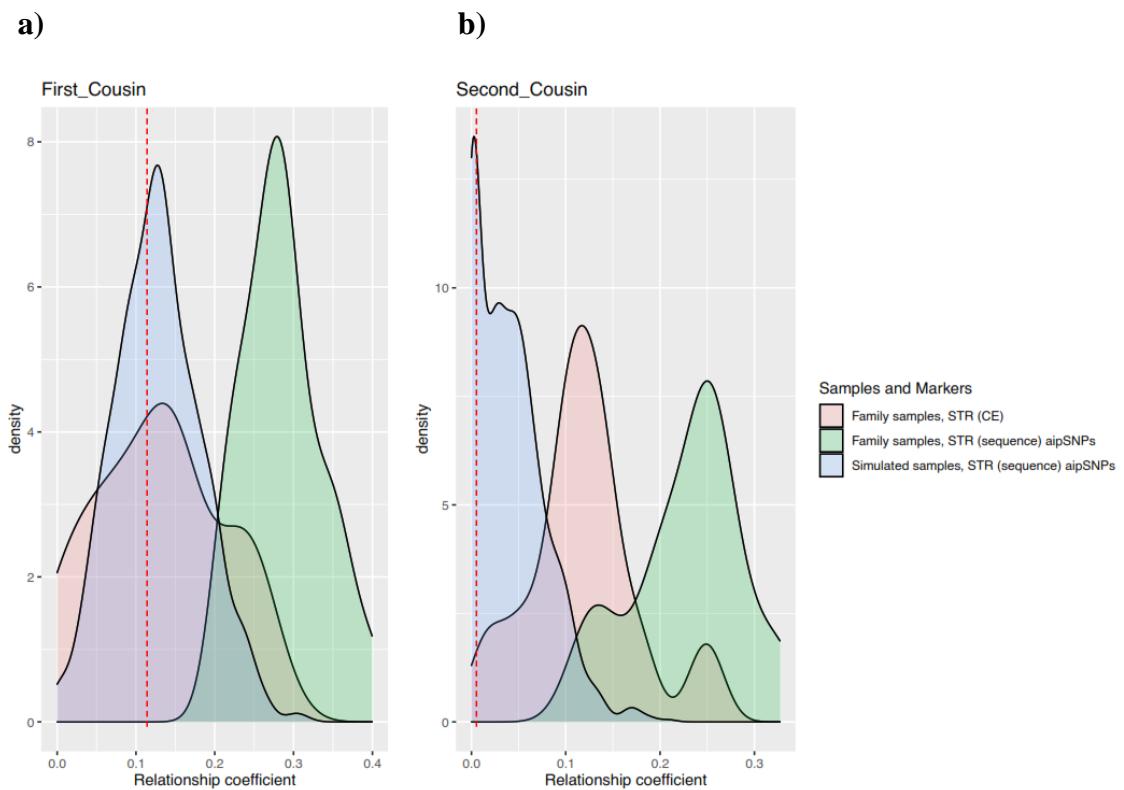
**Figure 3.22 Comparison of relationship coefficient distribution based on STRs (sequence) for (a) parent-offspring (PO) and (b) full siblings (FS) relationship based on real-world family data (pink) and 1000 simulated related pairs (blue).**

Real-world family data include 46 PO pairs, and 13 FS. The red dashed line highlights the expected coefficient of relationship ( $r = 0.5$ ).



**Figure 3.23 Comparison of relationship coefficient distribution based on STRs (sequence) for half-siblings (HS), grandparental and avuncular relationship based on real-world family data (pink) and 1000 simulated related pairs (blue).**

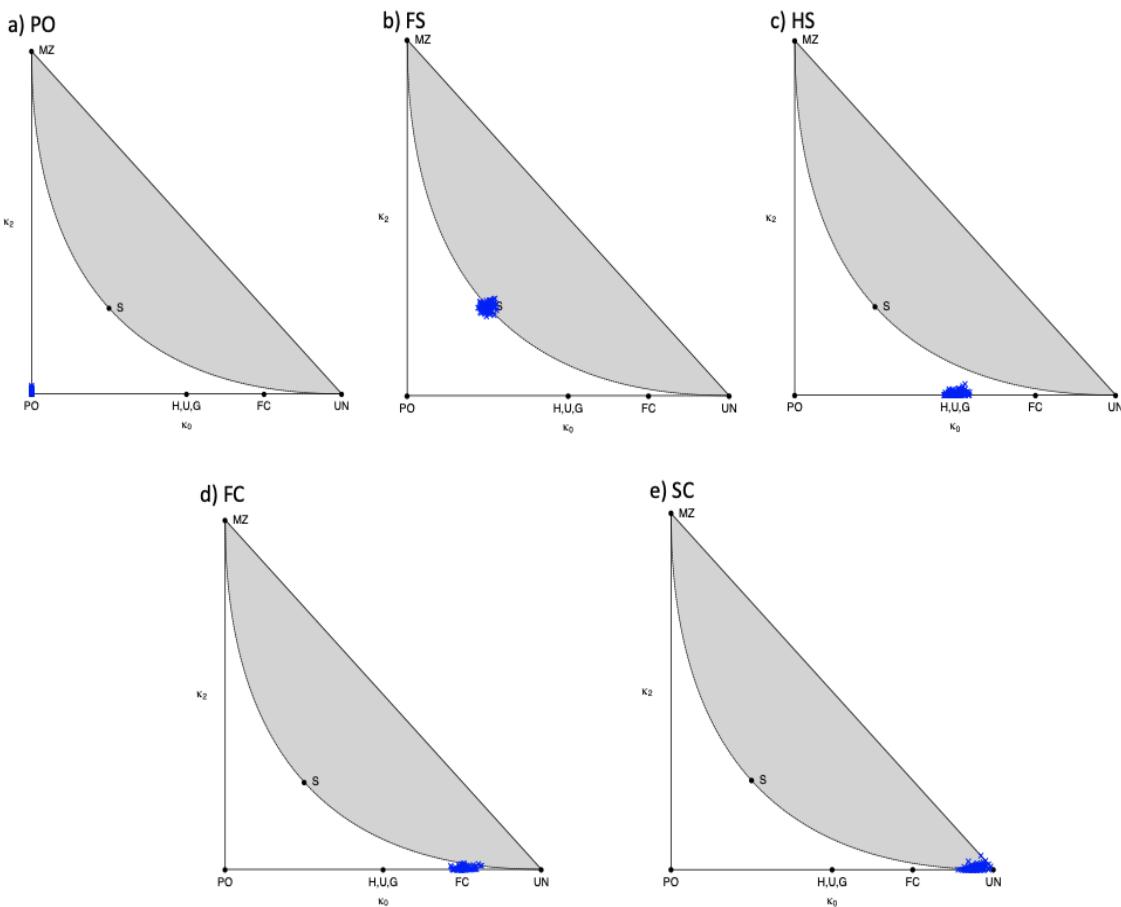
Real-world family data include 4 HS pairs, 27 grandparental, 33 avuncular. The red dashed line highlights the expected coefficient of relationship ( $r = 0.25$ ).



**Figure 3.24 Comparison of relationship coefficient distribution based on STRs (sequence) for (a) first and (b) second cousin relationship based on real-world family data (pink) and 1000 simulated related pairs (blue).**

Real-world family data include 22 first cousin and 12 second cousin pairs. The red dashed line highlights the expected coefficient of relationship ( $r = 0.125$  and  $r = 0.03125$ ).

Stimulated by the recent surge of interest in kinship and genetic genealogical analysis, the company Verogen has offered a new kit focusing on kinship determination. The ForenSeq Kintelligence kit (Verogen) targets a total of 10,230 SNPs: 9867 “kinship” SNPs, 106 X-SNPs, 85 Y-SNPs, and the standard ForenSeq kit identity, ancestry and phenotypic SNPs (94, 54 and 22 with 2 ancestry/phenotypic markers). According to Verogen, these markers were chosen after multiple decision-making steps, starting from the markers available in GEDmatch SNP data, and retaining those that were common with commercial methods, were unlinked to medical traits and without population-biased frequencies, and showed minimal genetic linkage (The ForenSeq Kintelligence workflow webinar, <https://verogen.com/webinars/the-forenseq-kintelligence-workflow/>). Through simulations (as described above), the application of these markers was explored here in a kinship determination setting (Figure 3.25). It is possible to notice that the chosen markers and their numbers reduce the variance in the estimation and, therefore, relationship clusters appear much more defined and closer to the expected values.



**Figure 3.25 Scatterplot of the 0 IBD ( $k_0$ ) proportion and 2 IBD ( $k_2$ ) for five relationships based on the 9867 kinship SNPs (ForenSeq Kintelligence).**

Data were simulated for 100 related pairs using the R package forrel. (a) PO: parent-offspring; (b) FS: full sibling; (c) HS: half-sibling; (d) first cousin; (e) second cousin.

## 3.4 Discussion

This Chapter explored the use of MPS data based on the ForenSeq kit and its application in kinship determination. The ForenSeq kit offers a set of up to 230 markers (Plex B), combining both STRs and SNPs. The advantages of using this kit in forensic practice have been already considered (Peng et al. 2020; Köcher et al. 2018; Jäger et al. 2017; Silvia et al. 2017), however, information on its application to kinship testing is limited (Zhang et al. 2020; Li et al. 2019). Nonetheless its features allow consideration of which set and combination of markers may be more suitable for IBD-based relatedness determination.

The ancestry and phenotypic SNPs show lower diversity in the families compared to the identity SNPs. No statistical test was performed, because of the small number of unrelated individuals. Then, simulated family data were investigated to explore the markers' performance in an "ideal" situation, where the expected variation in estimates (due to stochastic behaviour of recombination and segregation) is limited. Instead, the real-world family data show a larger variation in the IBD estimates, following the expected patterns of IBD sharing for each relationship type. As the IBD scatterplots show (Figure 3.17-21), increasing the number of markers improves most estimates, especially for close relationships, offering a clearer clustering of samples towards the expected IBD values compared to estimations based on only one type of marker (i.e. STRs [CE]) without considering the variants in the flanking region. Although the variance is reduced (i.e. most samples have estimated values closer to the expected IBD values), relatedness generally appears to be overestimated (Figure 3.22-24). A possible reason is the sensitivity of this type of analysis to the reference allele frequencies. The available reference is based on 138 individuals from European metapopulation that, overall, may show different characteristics compared to a dataset of only Northern European (German) samples. The limited number of unrelated samples in the investigated dataset does not allow the use of population statistics (e.g. heterozygosity,  $Fst$ ) to further investigate this phenomenon (see Appendix 3c).

It was possible to demonstrate the advantages and disadvantages of the use of targeted MPS data in a forensic kinship setting for obtaining investigative lead, focusing specifically on the Verogen ForenSeq kit. This Chapter further explored this investigative possibility comparing both simulated and real-world family data, with SNP data on the same samples for confirming the results. To include the most recent advances, data were also simulated on the latest kinship-focused kit from Verogen, the Kintelligence kit.

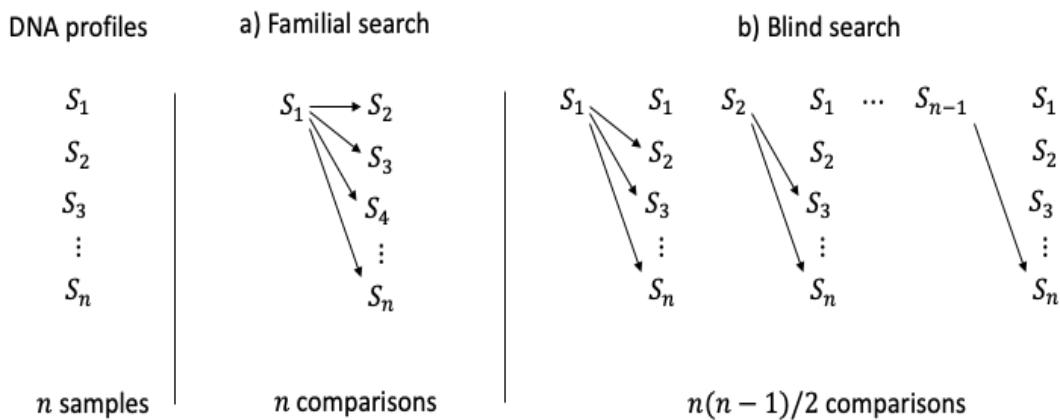
# Chapter 4: Strategies for pairwise searches in forensic kinship analysis

The work described in this Chapter is partly due to collaborative work with Prof Thore Egeland and Dr Hilde Kjelgaard Brustad (NMBU, Oslo, Norway) carried out during an International Society for Forensic Genetics (ISFG) Short-term fellowship funded research visit to Oslo and now published in *Forensic Science International: Genetics*.

Some sections reported in this Chapter are included in the published paper (Kjelgaard Brustad et al. 2021):

- Case 1-2 have major contributions from H. Kjelgaard Brustad;
- Case 3-4 have major contributions from M. Colucci;
- section 4.2.3 is described in the paper;
- section 4.3.4 (Case 4) further explores the analysis of half siblings with inbred founder;
- in section 4.3.5 (Case 5), the deficiency case scenario in paternity testing is not included in the paper, but original material.

Estimating kinship between pairs of individuals is central to many forensic investigations. A widely used approach is familial DNA search, which involves comparing an unknown forensic profile,  $S$ , to each of the known profiles,  $S_1 \dots S_n$ , in a criminal database (Figure 4.1 a) and typically focuses on close kin relationships, such as parent-offspring or sibling. In some cases, however, more complex analysis involving multiple individuals must be carried out. Typical scenarios include missing persons cases, and mass fatality incidents (MFIs) which can be the result of accidental catastrophes (e.g. air crashes with a list of known victims), natural disasters (e.g. tsunamis, where the number of victims is unknown) and terrorism-related events. The aim in a MFI case is to link a DNA sample from the scene to a putative victim (e.g. an individual reported missing since the event) and is known as disaster victim identification (DVI). As these cases involve unidentified DNA samples, a first step in the investigation is to screen the data for related samples. This initial step is referred to as a *blind search* (Egeland et al. 2015). Besides DVI applications (Bertoglio et al. 2020; Olivieri et al. 2018), blind search strategies have also been used in studies of mass graves of archaeological relevance (Palomo-Díez et al. 2018; Palomo-Díez et al. 2015; Parsons et al. 2019), in Y-STR haplotype sharing studies (Della Rocca et al. 2019), and as a tool for removing duplicates and closely related individuals from a population sample prior to calculating allele frequencies to avoid bias in the estimations.



**Figure 4.1 Different database searches are presented here.**

Kinship analysis may be performed as familial search or as blind search, involving different pairwise comparison approaches. Given a dataset of  $n$  individuals, a) a familial search focuses on one sample, searching for related individuals between databases, and compares it against each sample in a reference dataset, obtaining  $n$  pairwise comparisons for one queried relationship hypothesis; while b) a blind search performs all possible pairwise comparisons combinations, searching for related individuals within databases and obtaining  $n(n-1)/2$  pairwise comparisons for one queried relationship hypothesis. Direct search (not shown) looks for direct matches between or within databases.

In a blind search, pairwise comparisons are performed among all available samples, and a likelihood ratio (LR) is typically calculated for each specified relationship,  $H_1$ , versus the “null” hypothesis of no relationship,  $H_0$ . The LRs summarise the statistical DNA evidence. For a set of  $n$  genotyped samples, there are  $N = n(n - 1)/2$  sample pairs (Figure 4.1 b), and for each of these the LR is computed. These computations may then be repeated for different choices of  $H_1$ . Compared to a familial search, where the number of comparisons is proportional to the size of the database, a blind search involves a larger number of comparisons that increases quadratically with the sample size and the number of relationships to be tested. For example, familial search in a database of 100 individuals gives 100 comparisons for one relationship hypothesis, while for a blind search, 4950 computations would be performed. This has wide-ranging implications both for computation and interpretation, which are addressed in this Chapter.

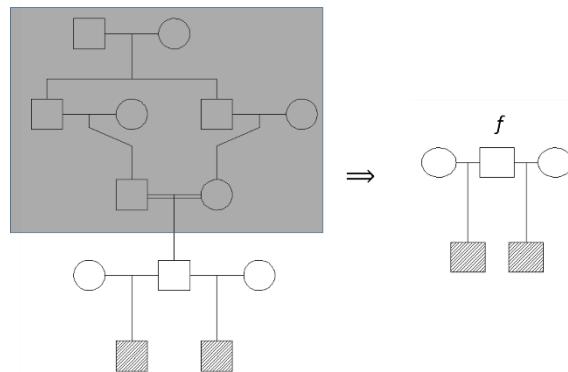
It is always possible to obtain false positives when searching for relatives in large databases using any approach, but spurious matches are far more likely in a blind search because of the large number of comparisons being considered. The probability of such

false positives also increases as more distant relationships are searched for. A method for weighting the detected matches by the probability of obtaining a false positive would be useful for appropriate targeting of investigative resources, and such weighting will become even more relevant as new technologies to detect more distant relationships mature, e.g. MPS (Korneliussen 2015; Li et al. 2019; Waples et al. 2019; Zhang et al. 2020), SNP panels (Gill 2001), and genetic genealogy (Kling and Tillmar 2019). This is a classical multiple testing problem and can be addressed by controlling the false positive rate (FPR) or the familywise error rate (FWER) (Storey 2003).

Current applications of blind search are limited to fairly simple outbred pedigree structures connecting the two individuals of interest. For example, Familias (Egeland et al. 2015), a freely available kinship software package, facilitates blind search and allows the addition of reference samples provided the hypotheses to be considered are selected from a set of five predefined pedigree relationships: parent-offspring, sibling, half-sibling, first cousin and second cousin (Egeland et al. 2000; Kling et al. 2014a). Real scenarios may entail more complex pedigree structures, possibly including founder inbreeding. A wider range of real-world cases can be accommodated by taking a parametric approach to the likelihood computation. In the parametric approach, the hypothesised relationship is represented by a set of relatedness parameters, such as identity-by-descent (IBD) coefficients, rather than as a pedigree structure (i.e. the hypothesis of being parent-offspring does not need to be spelled out and it is specified as  $\kappa = (0,1,0)$ ). Two alleles are said to be *identical by descent* if they originate from the same ancestor for the pedigree we choose to consider, and their IBD status can be expressed by the coefficients  $\kappa = (\kappa_0, \kappa_1, \kappa_2)$ , where  $\kappa_i$ , for  $i = 0,1,2$ , is the probability that the two individuals share  $i$  alleles IBD at a locus (Cotterman 1940). These parameters are defined on a continuous scale and represent all possible outbred pairwise relationships, although not necessarily uniquely, as the same  $\kappa$  probabilities may define different pedigree relationships. For example,  $\kappa = (0.5,0.5,0)$  define grandparent-grandchild and half-sibs, as well as an avuncular pedigree relationship.

Importantly, this approach allows the consideration of founder inbreeding, assuming that the two individuals themselves are not inbred. By assigning a coefficient of inbreeding to one or more of the founders, the background relatedness can be modelled and the IBD coefficients can still be defined (Vigeland and Egeland 2019). For example, if it is

suspected that two individuals are paternal half-siblings and the paternal grandparents are first cousins (Figure 4.2) the common father has an inbreeding coefficient  $f = \frac{1}{16}$  which changes the IBD coefficients for the half sibling relationship from  $\kappa = (0.5, 0.5, 0)$  to  $\kappa = (0.469, 0.531, 0)$ . This can be accommodated in a parametric likelihood calculation.



**Figure 4.2 Figure showing the concept of inbred founders and coefficients of inbreeding.**  
This pedigree shows a first cousin mating: this can be accommodated in the analysis just by assigning a coefficient of inbreeding ( $f$ ) to the inbred individual (in this case,  $f = 0.0625$ ).

Generally, a frequentist approach to a blind search is used, evaluating LR values and looking for values above a pre-specified threshold: the LR quantifies how much more likely it is that a set of genetic data is explained by one hypothesis  $H_1$  than by an alternative hypothesis  $H_0$ ; if the LR exceeds a given threshold,  $t$ , it is typically said that there is strong support for  $H_1$  over  $H_0$  (Kruijver 2015). A Bayesian approach, which calculates posterior probabilities for a set of relationships, is also possible. This has the advantage of not being restricted to comparing two simple hypotheses and comes closer to answering the question that is usually of interest, i.e. which of these relationships is the most probable (Hepler and Weir 2008; Kling et al. 2012; Sheehan and Egeland 2007; Taroni et al. 2014). In a Bayesian approach, when proposing different hypotheses to investigate, we must first assign a prior probability,  $\pi_i$ , that each hypothesis  $i$  is true. These probabilities are set before considering the genetic data and are based on external, non-genetic and possibly subjective information. Prior information could rule out some possibilities (e.g. two individuals of the same age cannot possibly be parent-offspring, whatever the data might suggest). Choosing reasonable prior probabilities is generally

difficult and often a flat prior is used whereby  $\pi_i = \frac{1}{k}$ , for  $k$  competing hypotheses. This makes the strong assumption that all hypothesised relationships are equally likely *a priori*. The specified prior probabilities are combined with the likelihoods to obtain posterior probabilities for a set of relationships which assess how likely each one is relative to the others in the set. This is different from the LR which only compares two alternatives. For a single pair of samples from a blind search and a set of  $k + 1$  hypotheses ( $H_0, \dots, H_k$ ) with prior probabilities  $(\pi_0, \dots, \pi_k)$ , Bayes' theorem yields posterior probabilities

$$P(H_i | data) = \frac{P(data | H_i) \pi_i}{\sum_{j=0}^k P(data | H_j) \pi_j},$$

which can be conveniently expressed as

$$P(H_i | data) = \frac{LR_i \pi_i}{\sum_{j=0}^k LR_j \pi_j},$$

when the LRs have already been computed from the blind search and provided they all have the same alternative in the denominator (Egeland et al. 2015). More informative priors can contribute additional information and this will be reflected in the posterior probabilities. For example, the relationships that are defined by the same IBD coefficients (e.g. grandparent-grandchild, half-siblings, avuncular) would have identical likelihoods and hence would be indistinguishable in the traditional frequentist setting. Age information can easily be incorporated into the Bayesian approach and may yield different posterior probabilities.

#### 4.1.2 Overview of commonly used methods

##### Assumptions underlying LR calculations

All the above considerations and approaches are valid under certain assumptions. Firstly, the population is assumed to be in Hardy-Weinberg Equilibrium (HWE) and in Linkage Equilibrium (LE). Note that adding inbred founders to a pedigree does not violate either of these, since HWE and LE are properties of the population, while founder inbreeding only affects the individuals in a specified pedigree. Secondly, mutations are ignored as the mutation rates are usually small, and the errors induced by ignoring mutation in likelihood calculations are typically negligible (Egeland et al. 2017). However, for a

parent-offspring (PO) relationship i.e.  $\kappa = (0,1,0)$ , the likelihood will be zero if the two samples have genotypes at any locus that are incompatible with this hypothesis. For this special case, there is a simple formulation of the likelihood that incorporates mutation (see Egeland et al., 2015), and it is applied in this Chapter as an extended stepwise mutation model, with mutation rates of  $10^{-3}$  and  $5 \times 10^{-6}$  (integer and non-integer alleles, respectively) (Simonsson 2016). Finally, allele drop-ins and drop-outs, null alleles and genotyping errors are also ignored.

### Properties underlying LR calculations

When performing a hypothesis test by computing a LR, it can be useful to consider the performance this test can give. The following properties are commonly used as a measure of this.

The *False Positive Rate (FNR)* is defined by

$$FPR = P(LR \geq t | H_0),$$

and give the probability of falsely claiming the alternative hypotheses given that the null hypothesis is true.

The *False Negative Rate (FNR)* and *True Positive Rate (TPR)* are defined as

$$FNR = P(LR < t | H_1) \text{ and } TPR = P(LR \geq t | H_1),$$

and FNR is 1-TPR. The FNR expresses the probability of not claiming the alternative hypothesis, given that the alternative hypothesis is true, and conversely, the TPR expresses the probability of claiming the alternative hypothesis given that it is true. In other words, these parameters serve as a measure of the ability to not detect and to detect the true alternative hypothesis, respectively.

The above rates are determined by the hypotheses considered, the number of loci, the properties of each locus and the LR threshold  $t$ . As the probability distribution of the LR is not generally known (further considerations are given in Kruijver 2015), these rates may be estimated through simulations. For example, it is possible to consider this scenario to estimate FPR: genetic data are simulated on many independent pairs of unrelated individuals, the LR for  $H_1$  against  $H_0$  is computed and then it is possible to count how

many of these LRs are above a threshold. The number obtained, divided by the total number of pairs, serves as an estimate of the FPR. In general, it is desirable to have a small FPR. If  $FPR=0.00001$ , then for each 100,000 simulations, one false positive is expected. Therefore, performing 1000 or even 10,000 simulations to estimate FPR would lead to no false positives, and the estimated FPR would, by the above method, be zero. Kruijver (2015) describes several methods for estimating small probabilities. One of these is *importance sampling*. Details are given in Appendix 4a. In Section 4.2, measures of the above rates specifically for a blind search are introduced (lower-case letters are used to highlight parameters for blind search scenarios, e.g.  $fpr$  denotes the false positive rate for a blind search).

### 4.1.3 Aims of this Chapter

In this Chapter, a more rigorous approach to blind search than current implementations is proposed, in which search performance can be evaluated in terms of the number and rate of false positives. The parametric representation of the likelihood is used for both autosomal and X-chromosomal markers, enabling the consideration of complex relationships. For a specified pedigree connecting an outbred pair of individuals, founder inbreeding can be incorporated via an inbreeding coefficient reflecting background relatedness. A Bayesian approach is implemented, also based on the parametric likelihood calculation, which can be adopted alone or in conjunction with a standard frequentist approach.

The Chapter is structured as follows: first, an approach to evaluate the performance of blind search using the concept of family-wise error rate (FWER) and false discovery rate (FDR) is described, considering the problem of multiple testing. Then, the use of blind search is illustrated by five examples based on both real and simulated data, which demonstrate the performance evaluation workflow, application to DVI cases, the consideration of inbreeding and the inclusion of X-chromosomal markers. The analyses have been performed using an enhanced implementation, available as an R package.

## 4.2 Materials and Methods

The above-described methods are not new in the forensic literature. This section describes the measures to evaluate the performance of a blind search, which have not previously been addressed. The topic is motivated by the potential high number of pairwise comparisons in a blind search, and the fact that these comparisons cannot be assumed to be independent of each other, which makes the results different from a familial search.

### 4.2.1 Evaluating the performance of a blind search

Here the measures that can be used to evaluate and possibly optimise a blind search are presented and discussed. For a pair of hypotheses  $H_0$  and  $H_1$ , the likelihood ratio considered is

$$LR_i = \frac{P(data_i|H_1)}{P(data_i|H_0)}, i = 1, \dots, N$$

where  $data_i$  denotes the genotype data for the two individuals involved in the  $i^{th}$  comparison. A blind search involving  $N = \frac{n(n-1)}{2}$  pairwise comparisons results in likelihood ratios,  $LR_1, \dots, LR_N$ . The generic hypotheses are  $H_0$ : «the two individuals are unrelated» and  $H_1$ : «the individuals are related as described by a pedigree or a set of IBD coefficients». For specified thresholds  $t_0 < t_1$ ,  $LR_i < t_0$  lends support for  $H_0$  while  $LR_i \geq t_1$  supports  $H_1$ . More data may be required to draw any conclusions from values  $t_0 \leq LR_i < t_1$  (Tillmar and Mostad 2014). For simplicity, it is assumed  $t_0 = t_1 = t$ , so that a conclusion can always be drawn.

Let  $W$  be the number of LR values below  $t$  and let  $R$  be the number greater than  $t$ . If the true hypothesis is known in each case, as would be the case for simulated data, the result of the search can be summarised as shown in Table 4.1. In practice, the truth will not be known and it is only possible to observe the numbers  $W$  and  $R$  but Table 4.1 is useful for designing and evaluating a blind search.

**Table 4.1 Summary of statistics for a blind search.**

Only  $W$ , the number of LRs below  $t$ , and  $R$ , the number of LRs above  $t$ , is observed. Adapted from (Benjamini and Hochberg 1995).

	<b>Claim <math>H_0</math></b>	<b>Claim <math>H_1</math></b>	<b>Total</b>
<b><math>H_0</math> True</b>	$TN$	$FP$	$N_0$
<b><math>H_1</math> True</b>	$FN$	$TP$	$N_1$
<b>Total</b>	$W$	$R$	$N$

Assume that the only true alternatives for the comparisons are either the given relationships or unrelated, such that  $N_0 + N_1 = N$ . The number of type I errors or false positives is  $FP$  while the number of false negatives is  $FN$ . The ideal desirable values are  $FP = 0$  and  $FN = 0$ . However, this is not realistic. For a sufficiently large threshold  $t$ ,  $H_0$  will never be rejected, and there will be no false positives, i.e.  $FP = 0$ . Similarly, there will be no false negatives,  $FN = 0$ , for a sufficiently small threshold. The challenge is to make both  $FP$  and  $FN$  acceptably small. The sum  $FP + FN$  measures the total error from a blind search.

To evaluate the performance of the blind search it can be repeated  $N_{sim}$  times using simulated data. The result for a given LR threshold can be summarised as in Table 4.2.

**Table 4.2 Example summary of 100 simulations of a (hypothetical) blind search with 20 individuals.**

The individuals consist of three sibling pairs and the remaining being unrelated, so  $N_0 = 187$  and  $N_1 = 3$ . Definitions of TN, FP, FN, TP are in Table 4.1 above.

<b>Sim No</b>	<b>TN</b>	<b>FP</b>	<b>FN</b>	<b>TP</b>
1	187	0	0	3
2	186	1	0	3
:	:	:	:	:
100	187	0	1	2

Based on the above table, it is possible to estimate  $fdr$ ,  $fpr$ ,  $tpr$  and  $tnr$  for this blind search by computing their average values. Due to the repeated simulations, it is also possible to investigate the distribution of the estimates. For instance, for simulation 100 in Table 4.2, the  $fpr$  and  $fnr$  are estimated as  $fpr_{100} = \frac{FP}{N_0} = \frac{0}{187}$ ,  $tpr = \frac{TP}{N_1} = \frac{2}{3}$  and

$fnr_{100} = \frac{FN}{N_1} = \frac{1}{3}$ . Recall that FDR =  $FP/(FP+TP)$ , such that  $fdr_{100} = 0$  in this case. An estimate of the total error would be  $FP + FN = 1$ . Based on all simulations, the average  $fpr$  will be  $fpr = (\frac{0}{187} + \frac{1}{187} + \dots + \frac{0}{187})/100$  and similar for the other statistics.

The relationship between  $fpr$  and  $tpr$  is often visualised by a Receiver Operating Characteristic (ROC) curve (Fawcett 2006). The  $tpr$  for different values of the LR threshold is then plotted as a function of the corresponding  $fpr$  at the same thresholds. This shows how the ability to correctly declare the alternative hypothesis is related to the ability to falsely declare the alternative hypothesis.

#### 4.2.2 The problem of multiple testing

Even if the probability of a false positive is very small for a single pairwise comparison, the fact that there are so many tests in a blind search could lead to a substantial probability of at least one false positive. Approaches to analyze and control these false positives in a multiple testing setting have to be applied. The Family Wise Error Rate (FWER) and the False Discovery Rate (FDR) are often used for this purpose.

FWER is defined as the probability of obtaining at least one false positive out of  $N$  tests (Tamhane et al. 1996). For  $N$  independent tests,

$$FWER = P(FP \geq 1) = 1 - (1 - FPR)^N,$$

where the false positive rate (FPR) is assumed to be the same for each test. The pairwise tests in a blind search are not independent and so we have the conservative Bonferroni bound

$$FWER \leq N \cdot FPR = \frac{n(n-1)}{2} FPR.$$

Thus, to obtain a FWER below a given values, say 0.05, a threshold  $t$  should be chosen so that  $FPR = 2(0.05)/(n(n-1))$  for a fixed sample size  $n$ . The FWER does not consider the number of false positives, but rather ensures that the probability that false positives will be present is controlled.

The FDR is defined as the expected ratio of false positives among all the rejected null hypotheses, i.e.  $FDR = E\left(\frac{FP}{R}\right)$  (Benjamini and Hochberg 1995). Benjamini and Hochberg propose a method to control this parameter, based on *p*-values from each test. As the probability distribution of the LR is unknown, exact *p*-values cannot be obtained. Kruijver et al. (Kruijver 2015) also argue the *p*-values should not be used to draw conclusions in LR kinship testing.

### 4.2.3 The likelihood ratio for X-chromosomal markers

X-chromosomal markers are increasingly used in forensic applications to supplement or replace autosomal markers for some specific cases (Pinto and Gusmao 2011). One such example is shown in Figure 4.10. The females B and C are paternal half sibs while C and D are maternal half sibs. The distinction between maternal and paternal is captured by X-chromosomal markers but not by autosomal markers. The paternal half siblings share an allele IBD inherited from their father. The Jacquard coefficients (see section 1.5.1.1, Chapter 1) and the likelihood calculation can be modified to apply for independent X-chromosomal markers (in-house code). Obviously, the sex of the individuals in the pair matters. For example, there are only two possibilities, or two states, for a pair of males: either they share an allele IBD or they do not. Since the number of unlinked markers on the X chromosome is limited, linkage and linkage disequilibrium become an issue (Kling et al. 2015b). This issue is ignored in Case 5 below, where data on the X chromosome are considered. However, relevant findings that take LD into account can be checked using the freely available software FamLinkX (Kling et al. 2015a).

### 4.2.4 Sample description

All samples are assumed to be filtered for direct matches (e.g. duplicates, or victim-victim matches), a process that requires modelling of allele drop-ins and drop-outs, null alleles and genotyping errors.

#### 4.2.4.1 Real family data

DNA samples tested are described in Chapter 3 (Section 3.2.1). For the purposes of this chapter, only the length-based genotypes (CE) from 27 autosomal STRs contained in Plex B of the ForenSeq kit were considered (D1S1656, TPOX, D2S441, D2S1338, D3S1358,

D4S2408, FGA, D5S818, CSF1PO, D6S1043, D7S820, D8S1179, D9S1122, D10S1248, TH01, vWA, D12S391, D13S317, PentaE, D16S539, D17S1301, D18S51, D19S433, D20S482, D21S11, PentaD and D22S1045). Allele frequencies are based on the European dataset in PopSTR (1000 Genomes Project, <http://spsmart.cesga.es/popstr.php>) and downloaded from the Familias website (<https://familias.no/download>).

#### 4.2.4.2 Simulated data

In the Results section, three examples based on simulated data are provided. Cases 1 is based on the 35 STRs and their allele frequencies from the Norwegian allele frequencies database on the Familias website ([https://familias.name/Familias\\_databases/](https://familias.name/Familias_databases/), for further details see Appendix 4b). Case 2 generates data based on the set of loci described above. For Case 4, marker data are simulated based on common commercial kits currently used in DVI cases, such as the PowerPlex® ESX 17 system (Promega, USA), using 16 STRs (D3S1358, D8S1179, D18S51, D21S11, FGA, TH01, vWA, D2S441, D10S1248, D22S1045, D1S1656, D12S391, D2S1338, D16S539, D19S433, SE33) (Ziętkiewicz et al. 2012). Here, the data are simulated using the R package *forrel* v1.0.1.9 (Vigeland and Egeland 2019). Allele frequencies for Case 2 to 4 are based on the European dataset in PopSTR (1000 Genomes Project, <http://spsmart.cesga.es/popstr.php>) and downloaded from the Familias website (<https://familias.no/download>). More information is given in Appendix 4b. In Case 5, the simulations are based on the 12 X-chromosomal STR markers included in the Investigator 284 Argus X-12 kit, with frequencies from (García et al. 2019).

#### 4.2.5 Implementation

The blind search and LR computations were performed using code in R (R Core Team 2014) built on the R libraries *pedtools*, *ribd*, *forrel* and *pedmut* developed by Magnus Dehli Vigeland, freely available from CRAN. This is an implementation of the parametric approach to the likelihood function, computing likelihoods for a series of relationships and converting to LRs and posterior probabilities. The advantage of this implementation is that relationships can be denoted by IBD coefficients. The software Familias has the functionality to perform blind search, but only based on a set of predefined relationships. An advantage of Familias is that it is user friendly, i.e., coding knowledge is not needed.

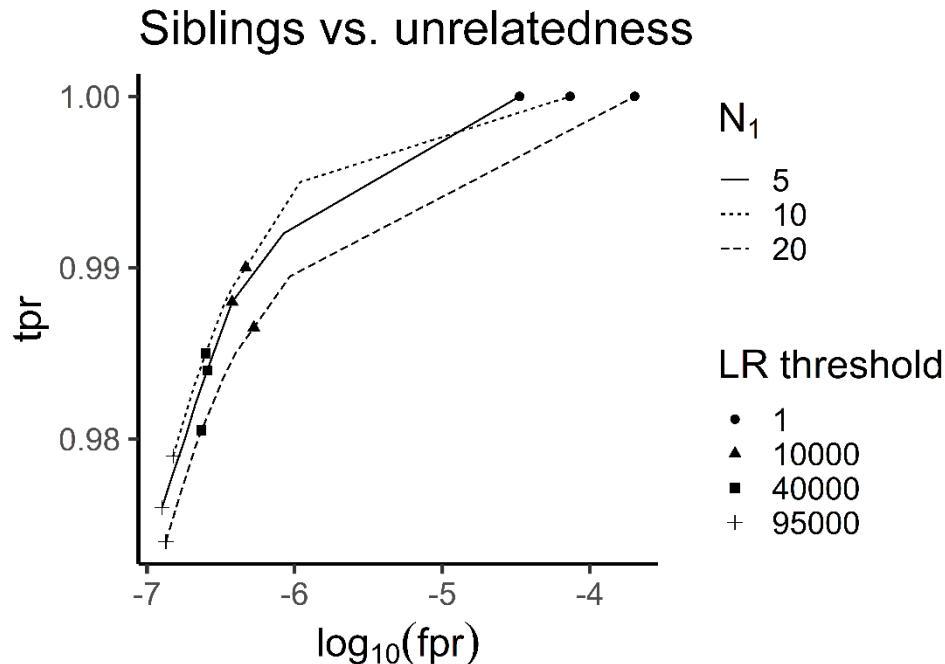
## 4.3 Results

This section explores five cases. In the first case, the performance of the blind search is considered. The second case demonstrates a typical blind search in a DVI setting, based on simulated data, while the third case uses real data. Case four shows how inbreeding can be included. The final case demonstrates a blind search on X-chromosomal markers with simulated and real data.

### 4.3.1 Case 1: Performance of the blind search

Consider a blind search among  $n = 40$  individuals, which gives a total of 780 pairwise comparisons. Three scenarios are specified that differ in the number of true sibling pairs ( $N_1$ ), with  $N_1 = 5, 10, \text{ and } 20$  true sibling relationships. Three different blind searches, one for each value of  $N_1$ , are performed. For each of these three searches, 100 runs are carried out on simulated data to evaluate the performance, and average values computed over these 100 simulations, setting different LR thresholds.

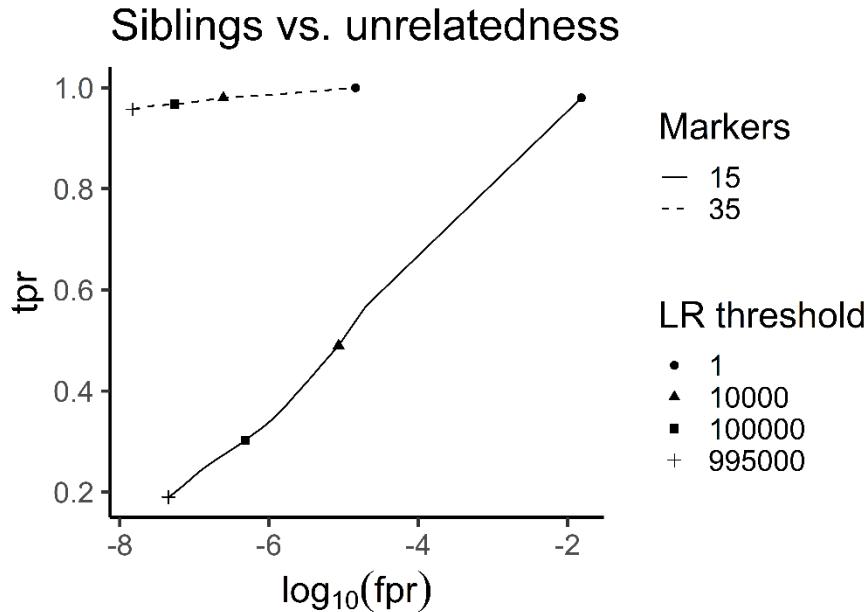
The ROC curves for the estimated false positive rate  $fpr$  and  $tpr$  are shown in Figure 4.3. Even though the  $fpr$  becomes vanishingly small, the ability to correctly declare the alternative hypothesis stays high, close to 1. The rates seem to be fairly constant for different numbers of related pairs in the search.



**Figure 4.3** ROC curve for a blind search among 40 individuals, with  $N_1 = 5, 10$  and 20 of the pairwise comparisons in the search being true siblings.

The rate  $tpr$  is plotted as a function of  $fpr$  (log). Each point on the curve corresponds to an LR threshold  $t$ .

It is common practice for an initial default set of markers to be used in the analysis. If it turns out that this default set of markers is insufficient to perform the analysis, additional markers are often added. Consider a blind search among  $n = 20$  individuals,  $N_1 = 4$  of the pairwise comparisons being siblings and the remaining  $N_0 = 186$  comparisons being unrelated. Estimates of  $fpr$  and  $tpr$ , presented by a ROC curve, for sets of 15 and 35 markers are shown in Figure 4.4. When using 35 markers, the  $tpr$  is close to unaffected when requiring a low  $fpr$ . This is not the case when only 15 markers are analyzed. For a logarithmic value of  $fpr$  of about  $-7$ , the  $tpr$  is as low as 0.2. At about  $(fpr) = -3$ , the  $tpr$  has increased to 0.8, still not reaching the lowest values of  $tpr$  for 35 markers.



**Figure 4.4 ROC curve from 100 simulations of a blind search with 4 sibling pairs and a total of 20 individuals.**

The dashed and solid line show results when 15 and 35 markers are evaluated by the LR.

In the above case, the false positive rates have been estimated by importance sampling. This approach considers the true positives in the search to estimate the false positives. See Appendix 4a for details. An analysis where the number of related pairs is kept constant, but the number of individuals  $n$  in the blind search is changed, will therefore give the same estimates of the  $tpr$  or the  $fpr$  for different  $n$ . Estimation of  $fdr$  is also hard to perform, as the number of false positives is highly likely to be zero in each simulation, and therefore the  $fdr$  will be estimated as 0.

### 4.3.2 Case 2: Blind search in DVI cases

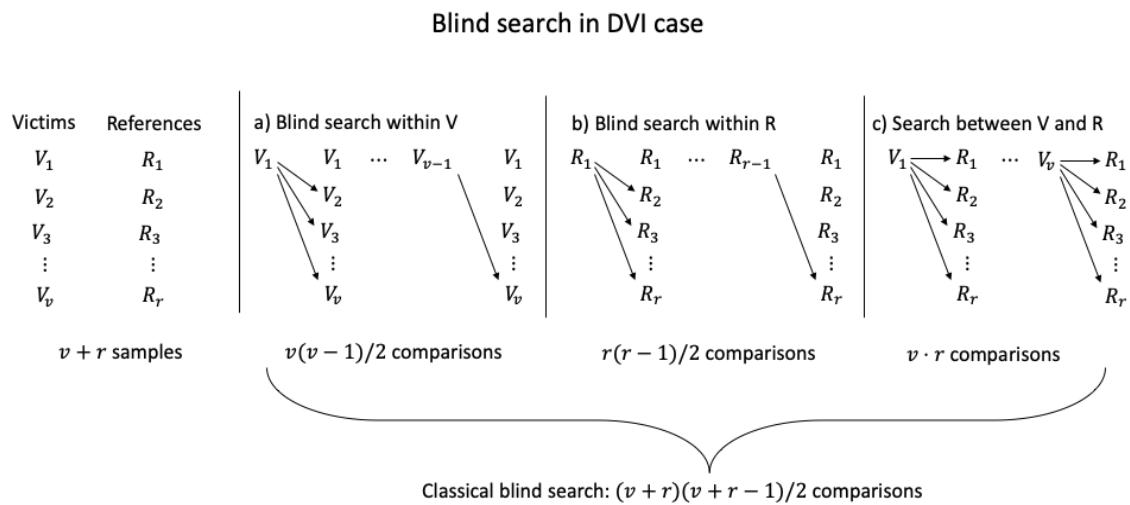
The purpose of this example is to demonstrate how a search can be performed in connection with a common DVI (Disaster Victim Identification) problem. Assume DNA profiles are available from deceased victims,  $V = (V_1, \dots, V_k)$ . To identify the victims, profiles from potential relatives,  $R = (R_1, \dots, R_l)$  have been obtained. In this initial phase there is no access to pedigree information indicating potential relationships within  $V$  or  $R$  or between  $R$  and  $V$ . Even if such relationships were known, it is worthwhile to do a first screening as described below. The first steps are (Figure 4.5):

1. Carry out a blind search within  $V$  and  $R$  separately, to identify possible

relationships between pairs of profiles in  $V$  and possible relationships between pairs of profiles in  $R$ ;

2. Carry out a familial search of  $V$  against each profile of  $R$ .

The two steps above can be performed simultaneously in a single blind search among the profiles of both  $V$  and  $R$ .



**Figure 4.5 Pairwise comparisons performed in a blind search as in a disaster victim identification (DVI) scenario.**

The searches within the victims' dataset and reference dataset and between the two are performed simultaneously when performing blind search.

Simulated data were used to represent a set of 45 DNA profiles from a mass grave (victims) and 57 reference profiles from individuals missing a family member (the missing family members being the victim profiles). A set of 23 loci is used (Appendix 4b).

A short excerpt of the result of a blind search among these profiles, for a LR threshold of 10 000 is shown in Table 4.3. The matches are ordered in victim-victim, reference-reference and victim-reference pairs. By performing a blind search among both reference and victim profiles the two above steps are simultaneously covered.

**Table 4.3 Examples of Likelihood Ratios from of blind search including both reference and victim profiles.**

LR threshold of 10000.

ID1	ID2	H1: PO	H2: SI
Victim-Victim matches			
V03	V40	$9.9 \cdot 10^{11}$	$4.97 \cdot 10^{10}$
V13	V25	$4.24 \cdot 10^8$	$1.91 \cdot 10^6$
Reference-Reference matches			
R13	R14	$2.38 \cdot 10^{10}$	$5.97 \cdot 10^{10}$
R51	R52	$1.02 \cdot 10^7$	
Victim – Reference matches			
V06	R27	$1.49 \cdot 10^6$	
V13	R02	$4.24 \cdot 10^8$	$1.91 \cdot 10^6$

Some pairs have matches for only one alternative hypothesis. When there are matches for several proposed relationships, further analysis needs to be performed. As an example, consider the sample pair V03-V40. Both hypotheses have high LR values, and the LR comparing PO against SI is  $LR\left(\frac{PO}{S}\right) = 9.9 \cdot 10^{11} / 4.97 \cdot 10^{10} \approx 20$ . It is not obvious which relationship to decide on considering the set threshold of 10000, unless additional information is included (e.g. age).

### 4.3.3 Case 3: A blind search using real data

This case considers data at 27 autosomal markers (section 4.2.4.1) on  $n = 65$  individuals from eight European families. The pedigree relationships among these individuals are known, so the performance of the analysis can be evaluated. There are  $N = 2080$  pairwise comparisons in this blind search. Using the upper bound in Equation (2), if FWER is to be  $<0.05$ , an FPR is required that is  $\leq 2.4 \cdot 10^{-5}$ .

**Table 4.4 Alternative hypotheses analysed in the blind search.**

This table shows the hypotheses considered in this example (H), the associated IBD coefficients ( $\kappa$ ) with corresppective pedigree relationship.

IBD coefficients		Relationship
$H_1$	$\kappa = (0,1,0)$	Parent offspring (PO)
$H_2$	$\kappa = (0.25,0.5,0.25)$	Siblings (S)
$H_3$	$\kappa = (0.5,0.5,0)$	Half-sibling/grandparent-grandchild/avuncular (H/GC/A)
$H_4$	$\kappa = (0.75,0.25,0)$	First cousins (FC)
$H_5$	$\kappa = (0.9375,0.0625,0)$	Second cousins (SC)

We define the null hypothesis  $H_0$  as unrelated (UN) so  $\kappa = (1,0,0)$ . Five alternative hypotheses are proposed and shown in Table 4.4. Traditionally, this would require five different blind searches to be performed, but the implemented code can conduct them all simultaneously. LR values are obtained and classified according to whether they are above or below a given threshold. Table 4.5 shows a summary of the blind search for an arbitrary LR threshold of 100,000.

For a given LR threshold, estimated performance rates differ for the different relationships considered. More distant relationships will in general give a lower LR. This means that the  $fpr$  will decrease the more distant the considered relationship is, but so will the  $tpr$ .

Among the 2080 sample pairs, there are 46 with a PO relationship. Of these, all 46 are above the LR threshold. In addition, there are 18 sample pairs that have a LR higher than 100,000, even though their relationship is not PO. Of the 64 H/GC/A relationships present, only two of them are claimed as true, but 40 other relationships are falsely claimed as H/GC/A.

**Table 4.5 Summary of the blind search, for a LR threshold of 100,000.**

$H_1$ : PO	Claim $H_0$	Claim $H_1$	Total
$H_0$ true	1866	1	1867
$H_1$ true	0	46	46
Other true	150	17	167
Total	2016	64	2080

$H_1$ : S	Claim $H_0$	Claim $H_1$	Total
$H_0$ true	1867	0	1867
$H_1$ true	3	10	13
Other true	173	27	200
Total	2043	37	2080

$H_1$ : H/GC/A	Claim $H_0$	Claim $H_1$	Total
$H_0$ true	1867	0	1867
$H_1$ true	62	2	64
Other true	109	40	149
Total	2038	42	2080

$H_1$ : FC	Claim $H_0$	Claim $H_1$	Total
$H_0$ true	1867	0	1867
$H_1$ true	21	0	21
Other true	190	2	192
Total	2078	2	2080

$H_1$ : SC	Claim $H_0$	Claim $H_1$	Total
$H_0$ true	1867	0	1867
$H_1$ true	12	0	12
Other true	201	0	201
Total	2080	0	2080

Table 4.6 shows the LRs for four selected pairwise comparisons in the blind search. For the first pair (F1P1-F2P2), the LRs for all five hypotheses are below the threshold of 100,000. None of the hypothesised relationships seems more likely than unrelated, so it

seems reasonable to conclude  $H_0$ , which is, in fact, correct. For the second pair (F1P1-F1P7) the PO relationship is the only one with a LR above the threshold. Again, the correct relationship (PO) is affirmed. For the F4P10-F4P7 pair, the correct relationship is grandparental (H/GC/A), but both the hypotheses of PO and GC are above the threshold. The LR between these two relationships is  $LR\left(\frac{PO}{GC}\right) = 3.88 \cdot 10^7 / 1.09 \cdot 10^5 \approx 356$  and a clear conclusion cannot be drawn, considering a threshold of 100,000. There is further ambiguity for the last pairing: the true relationship is PO, but the LR values are higher than the threshold for all the hypotheses considered, apart from second cousin (SC). The LR between PO and S is about 2000. This may not be a sufficiently high threshold to support the PO hypothesis, if considering the set threshold of 100,000.

**Table 4.6 Results for seven samples from a blind search on a dataset of 65 individuals, genotyped for 27 autosomal STRs.**

The hypotheses tested are: parent-offspring (PO), full sibling (S), half-sibling/grandparental/avuncular (H/GC/A), first cousin (FC) and second cousin (SC) against the null hypothesis of unrelated (UN). The final column shows the true relationships from the known pedigree.

ID1	ID2	PO	S	H/GC/A	FC	SC	TRUTH
F1P1	F2P 2	4.5370e-03	2.3981e-04	9.4148e-01	2.8835	1.6145	UN
F1P1 7	F1P	<b>7.9913e+08</b>	5.4720e+04	7.4546e+04	6.3951e+02	6.4831	PO
F4P10 7	F4P	<b>3.8765e+07</b>	6.2358e+04	<b>1.0928e+05</b>	2.6216e+03	17.0927	H/GC/ A
F4P1 2	F4P	<b>7.9500e+12</b>	<b>3.9614e+09</b>	<b>1.2560e+09</b>	<b>1.5837e+06</b>	361.3275	PO

Next, the posterior probabilities for the blind search performed above are computed, and the four pairings from Table 4.6 considered. A flat prior is assigned to all five specified relationships and the null hypothesis UN so  $\pi_i = \frac{1}{6}, i \in (1, \dots, 6)$ . The posterior probabilities are computed for every sample pair separately. The results are given in Table 4.7 with the highest posterior probability for each sample pair shown in bold. In the LR analysis, the first pair (F1P1-F2P2) has low LR values for all hypotheses. This is mirrored in the Bayesian analysis which ranks all six possibilities, but assigns a posterior probability of 0.4474 to the best of these (FC). For the other three pairs, posterior probabilities of ~1 are assigned to the PO relationship. This is not correct for the third pair (GC), who were also not uniquely identified in the frequentist setting. For the final

pairing, the frequentist approach identified many possibilities, but the posterior probabilities highlight the PO hypothesis as the most probable.

**Table 4.7 Posterior probabilities from the LRs (Table 4.7) of the blind search for seven samples on a dataset of 65 individuals, genotyped for 27 autosomal STRs.**

A flat prior is used.

ID1	ID2	PO	S	H/GC/A	FC	SC	UN	TRUTH
F1P1	F2P	7.0403e-2	3.7200e-05	0.1461	<b>0.4474</b>	0.2505	0.1551	UN
		04						
F1P1	F1P	<b>0.9998</b>	6.8500e-05	9.3300e-05	8.0000e-07	8.1100e-09	1.2500e-09	PO
		7						
F4P1	F4P	<b>0.9955</b>	0.0016	0.0028	6.7300e-05	4.3900e-07	2.5700e-08	H/GC/A
0	7							
F4P1	F4P	<b>0.9993</b>	5.8237e-04	1.5797e-04	1.9900e-07	4.5400e-11	1.2600e-13	PO
	2							

If additional information is known, this can be included in the calculations of the posterior probabilities. For example, age information is known for the individuals in the first and third pairings (F1P1-F2P2 and F4P10-F4P7). This additional information helps in ruling out a PO relationship, which now has a prior probability of 0 while the five other options have each a prior probability of  $\frac{1}{5}$ . The corresponding posterior probabilities, computed from the LR values in Table 4.6, are shown in Table 4.8. The pairing F4P10-F4P7 has now a posterior probability of 0.6270 for a H/GC/A relationship. Without additional information, it is not possible to distinguish between H, A and GC as the true relationship.

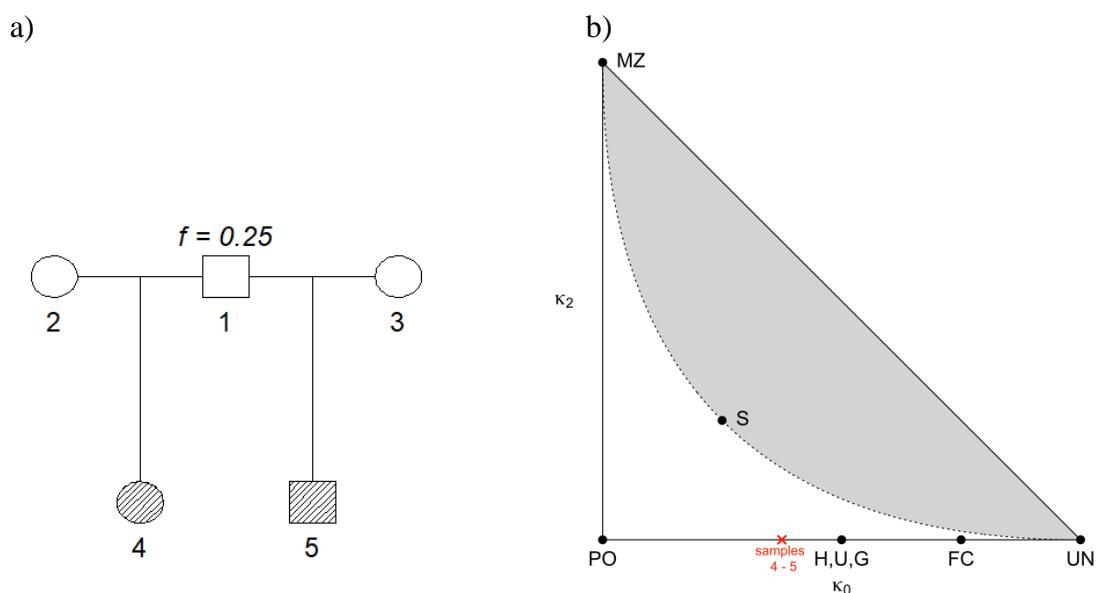
**Table 4.8 Posterior probabilities with priors from the LRs (Table 4.7) of the blind search for seven samples on a dataset of 65 individuals, genotyped for 27 autosomal STRs.**

For pairs F1P1-F2P2 and F4P10-F4P7, the parent-offspring hypothesis has an assigned prior of 0, while the other hypotheses have equal priors of 1/5. The tested hypotheses for the other pairs (F1P1-F1P7, F4P1-F4P2) have priors of 1/6.

ID1	ID2	PO	S	H/GC/A	FC	SC	UN	TRUTH
F1P	F2P	0	3.7239e-05	0.1462	<b>0.4478</b>	0.2507	0.1553	UN
1	2							
F1P	F1P	<b>0.9998</b>	6.8464e-05	9.3269e-05	8.0013e-07	8.1114e-09	1.2512e-09	PO
1	7							
F4P	F4P	0	0.3578	<b>0.6270</b>	0.0150	9.8079e-05	5.7380e-06	H/GC/A
10	7							
F4P	F4P	<b>0.9993</b>	4.9797e-04	1.5788e-04	1.9907e-07	4.5420e-11	1.2570e-13	PO
1	2							

#### 4.3.4 Case 4: Simulations incorporating founder inbreeding

A situation is now considered in which a blind search has been performed and two individuals, labelled 4 and 5 in the pedigree of Figure 4.6, have been singled out in order to consider the possibility of a more complex relationship between them. The true relationship is paternal half siblings (H), where the common father, individual 1, has an extreme inbreeding coefficient of  $f = 0.25$ . The IBD coefficients for this relationship are  $\kappa = (0.25, 0.75, 0)$  and so the expected genetic relationship has moved from the point of half siblings, with  $\kappa = (0.50, 0.50, 0)$ , towards PO, with  $\kappa = (0, 1, 0)$ .

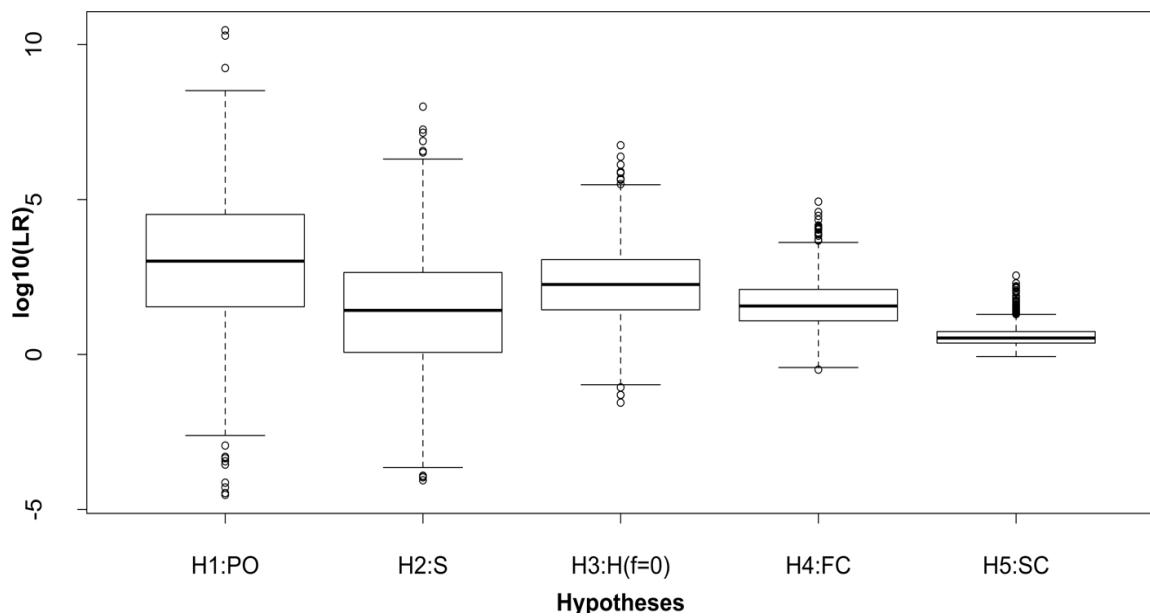


**Figure 4.6** This pedigree (a) represents an inbred family where individual 1 is the offspring of close relatives (siblings) corresponding to an inbreeding coefficient ( $f$ ) of 0.25.

Only the shaded individuals (numbered 4 and 5) are genotyped and their relationship is further investigated after an initial blind search step. b) This plot shows IBD 0 against IBD 2, causing each relationship type to cluster closer to the expected IBD sharing: the individuals 4 and 5 are shown to have higher IBD values than a half-sibship relationship would have (H,U,G).

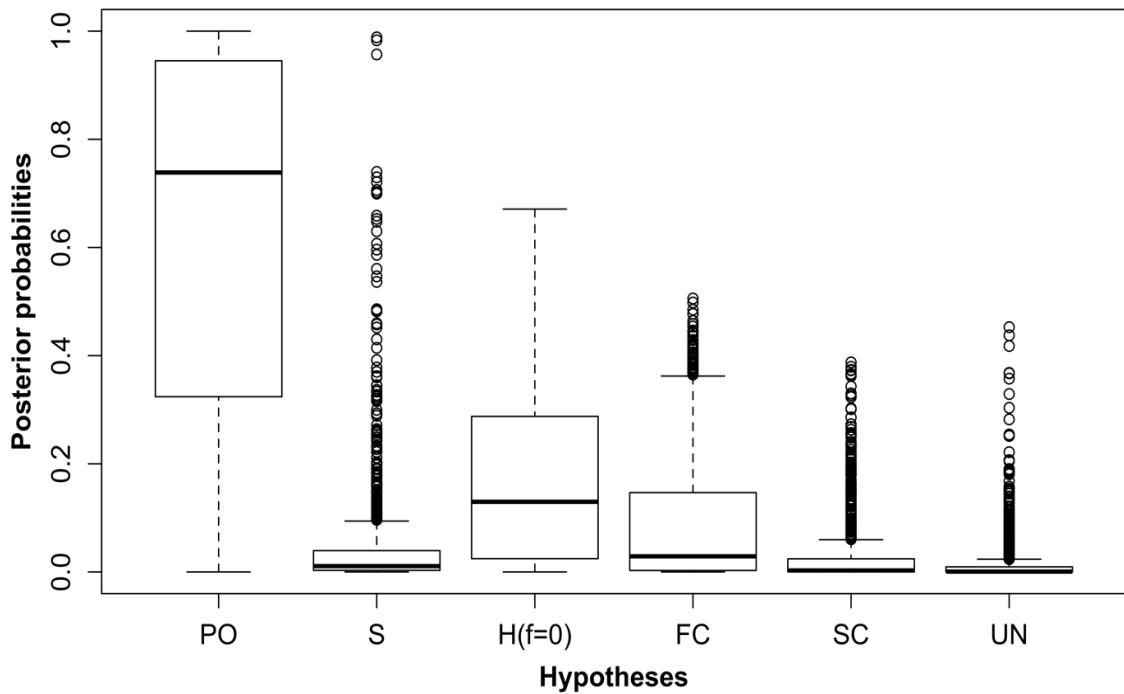
Data are simulated 1000 times on the pedigree in Figure 4.6 for the 16 forensic autosomal STRs described earlier (section 4.2.4.2). LRs for the five relationships specified in Table 4.5, each versus the alternative of unrelated, are obtained. Posterior probabilities for all six options using a flat prior are also computed. Boxplots of the LR values and the posteriors are shown in Figure 4.7 and Figure 4.8 respectively. All of the hypotheses considered are incorrect and both plots show much variability in the values. LRs are all below the (arbitrary) threshold of 100,000 and none of the proposed relationships stands

out from the others. The posterior probabilities are generally higher for the PO relationship, but the range is too large for this to be conclusive.



**Figure 4.7 Boxplots showing  $\log_{10}$  of LRs values for the 1000 half-sib pairings (4 and 5 as in Figure 4.6).**

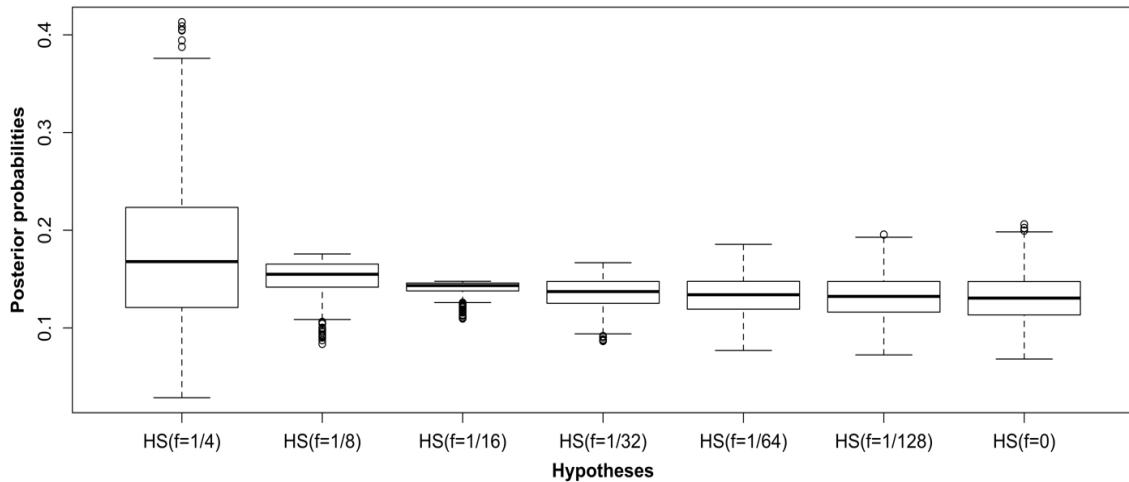
Five relationship hypotheses ( $H_1$  to  $H_5$ ) have been tested against the hypothesis of being unrelated ( $H_0$ ). No inbreeding coefficient has been included in the parametric LR ( $f=0$ ), even if the pedigree founder is inbred ( $f=0.25$ ).



**Figure 4.8 Boxplots showing Posterior probability values for 1000 half-sib pairings (4 and 5 as in Figure 4.6).**

Five relationship hypotheses ( $H_1$  to  $H_5$ ) have been tested against the hypothesis of being unrelated ( $H_0$ ) as shown in Table 4.9. No inbreeding coefficient has been included in the parametric LR ( $f=0$ ), even if the pedigree founder is inbred ( $f=0.25$ ).

Importantly, the proposed approach does not restrict us to simple relationship hypotheses. For example, if the targeted individuals are suspected to be half-siblings and the possibility of founder inbreeding needs to be investigated, the focus of the analysis can be restricted to a set of half-sib relationships with different levels of founder inbreeding, including the outbred option. Figure 4.9 shows the posterior probabilities for 1000 simulated pairs comparing different inbred half-sibling relationships and outbred half-sibship. These values are noticeably much lower than those above (Figure 4.8) because it is a much more difficult identification problem: this approach is trying to distinguish between very close relationships and it is clear that these 16 STRs do not offer the required discriminatory power.



**Figure 4.9** Boxplots showing Posterior probability values (flat prior) for the investigated individuals (4 and 5 as in Figure 4.5).

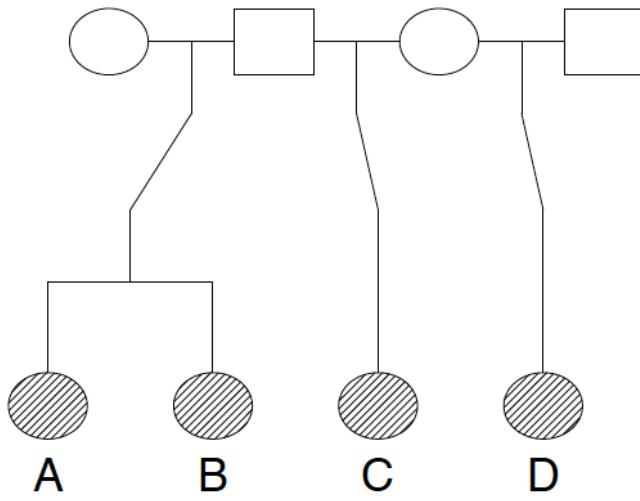
Various inbreeding coefficients are assigned to the tested hypotheses, while the null hypothesis has an inbreeding coefficient of zero ( $f=0$ ).

### 4.3.5 Case 5: Blind search with X-chromosomal markers

Because a male has only one X-chromosome, paternal half-sisters (HSP) must inherit the same X-chromosome from their common father. Their second X chromosomes, inherited from their respective mothers, are not IBD (since their mothers are unrelated), and hence, the IBD coefficients for a HSP relationship are  $\kappa = (0; 1; 0)$ . The IBD coefficients for maternal half siblings (HSM), whether considering X-chromosomal or autosomal markers, are  $\kappa = (0.5; 0.5; 0)$ . In the following example, we show with simulated data how X-chromosomal markers can distinguish between HSP and HSM.

Genotypes for 12 X-chromosomal STRs were simulated for the shaded individuals in Figure 4.10. Genotypes are simulated for each locus independently, by gene-dropping through the pedigree structure. More specifically, genotypes are sampled for the founders of the pedigree (the parents) according to the allele population frequencies and passed down through the pedigree. Only the resulting genotypes of the offspring are kept for the applications in this example. Table 4.9 presents the average posterior probabilities over 100 simulations, for the relationships PO, S, HSP, HSM and UN, for the six possible comparisons between the individuals A, B, C and D. A flat prior  $i = 1=5$  for  $i = 0; \dots; 4$  is assumed. The evidence in favour of C-D being HSM, shown in bold in Table 4.10,

could not be obtained using autosomal markers. Since a flat prior is being used, the LR comparing maternal to paternal half sibs is identical to the posterior probability ratio,  $0.81327/0.01916 = 42.4$ . This value may not be decisive on its own, but supplements other evidence. Note that HSP cannot be distinguished from PO using X-chromosomal markers alone as the row for the comparison A-C confirms. Age information, autosomal marker data or other non-DNA data may solve such cases.



**Figure 4.10 Pedigree connecting the individuals of the analysis in Section 5.5.**

Marker data are simulated for the four daughters to demonstrate blind search with X-chromosomal markers.

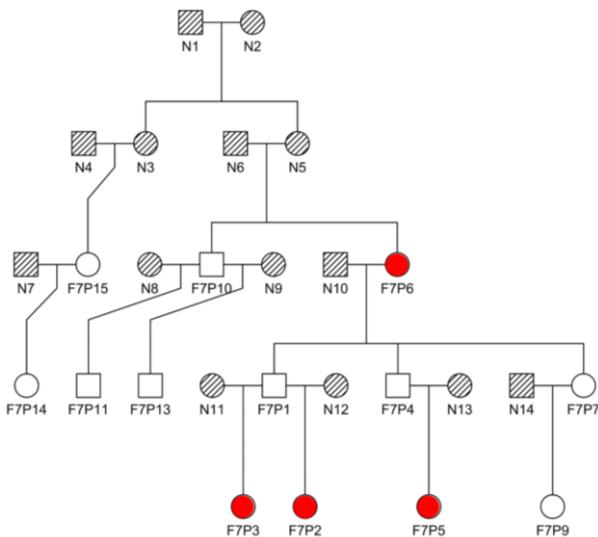
**Table 4.9 Posterior probabilities averaged over 100 simulations for the comparisons between the four daughters in Figure 4.10.**

Samples	PO	S	HSP	HSM	UN
A-B	0.039	0.921	0.039	0.001	0.000
A-C	0.475	0.039	0.475	0.012	$2 \cdot 10^{-5}$
A-D	0.000	0.000	0.000	0.154	0.846
B-C	0.471	0.045	0.471	0.012	$2 \cdot 10^{-5}$
B-D	0.000	0.000	0.000	0.146	0.854
C-D	0.019	0.001	0.019	<b>0.813</b>	0.147

The X chromosome has also applications in complex paternity testing (as explained in Chapter 2, Section 2.3.4.3), when, for example, the putative father is unavailable and the paternal grandmother of the tested child becomes a key figure (Szibor 2007). To recreate

this type of deficiency case scenario, five real families were considered (grandmother - grandchildren pairs, ignoring data on the father). The analyses were based on real data (Chapter 3, Section 3.3) using the seven ForenSeq kit X-STRs and European allele frequencies (Diegoli et al. 2011). Several relationships were considered (PO, S, HSM, HSP, GC, FC, UN), and LR and posterior probability values indicated that the markers may not be sufficient for obtaining a clear conclusion. Tables 4.10a-b show posterior probability results for family 7 (Figure 4.11) as an example, considering seven and five relationships, respectively.

Also, half-sibling pairs were simulated from the real genotypes (1000 simulations, alleles frequencies based on García et al. 2019), but results based on only 7 markers were not conclusive (data not shown).



**Figure 4.11 Pedigree of family 7.**

The individuals highlighted in red are involved in the analysis.

**Table 4.10 a Posterior probabilities for the comparisons between the three grandmother - grandchildren pairs in Figure 4.11 considering seven relationships.**

The seven relationships considered are: parent-offspring (PO), full siblings (S), maternal half siblings (HSM), paternal half sibling (HSP), grandparental (GC), first cousin (FC) and unrelated (UN).

Samples	PO	S	HSM	HSP	GC	FC	UN
F7P6 - F7P2	<b>0.21712907</b>	0.06106755	<b>0.21712907</b>	0.14249095	0.14249095	0.14249095	0.07720145
F7P6 - F7P3	<b>0.22755106</b>	0.01599968	<b>0.22755106</b>	0.14933038	0.14933038	0.14933038	0.08090704
F7P6 - F7P5	0	0	0	<b>0.2595156</b>	<b>0.2595156</b>	<b>0.2595156</b>	0.2214533

**Table 4.10 b Posterior probabilities for the comparisons between the three grandmother - grandchildren pairs in Figure 4.11 considering five relationships.**

The five relationships considered are: parent-offspring (PO), full siblings (S), grandparental (GC), first cousin (FC) and unrelated (UN).

Samples	PO	S	GC	FC	UN
F7P6 - F7P2	<b>0.33906287</b>	0.09536143	0.22251001	0.22251001	0.12055569
F7P6 - F7P3	<b>0.36518101</b>	0.02567679	0.23965003	0.23965003	0.12984214
F7P6 - F7P5	0	0	<b>0.3504673</b>	<b>0.3504673</b>	0.2990654

## 4.4 Discussion

In complex cases involving many allegedly related individuals, it is difficult to propose and examine multiple kinship/relationship hypotheses. Common approaches are based on LR calculations considering two opposing hypotheses. In order to allow for multiple searches among samples, a blind search should be considered.

This Chapter proposes an implementation of the blind search approach and presents strategies on how to analyse its performance and how to interpret results from the search. There are many functionalities, as exemplified in Section 4.3.

In a blind search, pairwise kinship testing is performed for all combinations of samples, resulting in a large number of likelihood ratios (LRs) to evaluate. The performance of the search can be evaluated by the probability of obtaining false positives among these pairwise comparisons. The probability of a false positive for each comparison is usually small, but as we evaluate many sample pairs, this cannot be ignored. The Family Wise Error Rate is described as the probability of at least one false positive among all the pairwise comparisons. An upper limit for this probability depends on the number of samples in the blind search and the false positive rate (FPR). If we require the FWER to be below a given value, the FPR has to be adjusted by setting the appropriate LR threshold. A high LR threshold decreases the FPR, but affects the ability to detect the true relationship. The possibility of linking the hypothesis, markers and LR threshold to the expected number of false positives is a significant addition, and it is necessary when considering a large number of pairwise comparisons. How these parameters are related to each other depends on the hypotheses and the loci evaluated in the blind search.

Two approaches for evaluating the results from a blind search are presented. The traditional approach considers the LR values, in light of an appropriate threshold. The search can be repeated for different alternative hypotheses, also considering posterior probabilities. Table 4.9 shows results (posterior probabilities) from a blind search for selected sample pairs. It allowed the comparison of several hypotheses at the same time, giving a simple overview of the relationships among the pairs. Close relationships (i.e. PO) were easily identified, however, as expected, more distant relationships needed further analysis, and this implementation easily accommodates them. Some restrictions and assumptions are worth noting for the computations. The main focus of the Chapter is on issues related to multiple testing, therefore as data are assumed to be pre-filtered for identical matches, allele drop-ins and drop-outs, null alleles and genotyping errors are not discussed and Hardy-Weinberg equilibrium and independence between the loci are assumed, similarly to other applications of forensic genetics. Mutations are only modelled for the parent-offspring relationship hypothesis.

Additional information can be included when considering the posterior probabilities of a set of hypotheses. Prior probabilities, possibly containing additional information to the analysis, are included and the posterior probabilities reflect how likely a hypothesis is

compared to the other hypotheses analysed. As is shown in Case 3, the impact of prior information can be significant, and can help in identifying the most probable hypothesis.

Existing software for performing blind searches, such as Familias, have a predefined set of hypotheses. This implementation enables the user to define any outbred relationship, offering increased flexibility to accommodate real-world scenarios. It also allows the definition of pairwise relationships where a founder is inbred (as long as the two individuals of interest remain outbred). This is shown in Case 4, where outbred relationships with founder inbreeding are defined as alternative hypotheses.

# **Chapter 5: Estimating biogeographical ancestry and phenotypes in a Portuguese-West African admixed population via massively parallel sequencing**

In the previous Chapters, the use of SNP chip and MPS data (ForenSeq DNA Signature Prep kit) was evaluated, and it was shown that additional SNPs may help in overcoming some limitations of STR profiling for kinship analysis and victim identification. Additional information that can be gained solely from DNA samples is the inference of bio-geographic ancestry (BGA) and externally visible characteristics (EVCs). These can both provide intelligence to assist in the identification of an individual.

The ForenSeq DNA Signature Prep Kit includes a set of 56 SNPs designed to predict BGA, and 22 SNPs for the prediction of eye and hair colour. This Chapter assesses the performance of these predictive systems, with a focus on an admixed population from Cape Verde.

## **5.1.1 Ancestry inference: overview and methods**

The study of a sample's biogeographic ancestry (BGA) is of relevance to several fields (Kopelman et al. 2015) such as population genetics to interpret genetic variation and evolutionary history (Rosenberg et al. 2002); association studies, to identify possible structure-induced genotype-phenotype associations (Pritchard and Donnelly 2001), personalised and medical genomics (Conomos et al. 2015b; De la Cruz and Raska 2014); and conservation genetics and molecular ecology for interpreting gene flow and managing natural populations (Kopelman et al. 2015). BGA assignment is becoming of increased interest in forensic settings, possibly offering investigative leads on unknown individuals (Hopkins et al. 2019).

If population structure exists in a species, such as humans, then the typing and statistical analysis of any large unbiased set of markers will detect it. This is illustrated by a seminal

study of the Human Genome Diversity Panel (Cann et al. 2002a) with 377 STRs (Rosenberg et al. 2002) which identified six major geographically differentiated genetic clusters in the absence of prior information on individual origins. Similar results were obtained on the same samples with 650,000 SNPs (Li et al. 2008). In a forensic setting typing so many markers are not generally practicable, so instead, a set of ancestry-informative markers, AIMs (Shriver et al. 1997) is usually applied. Such sets consist of SNPs (or other binary markers) characterised by significant allele frequency differences between or among populations, generally at the continental level. Table 5.1 lists a series of such AIMs sets. Among these is the set included in the ForenSeq kit; it is notable here that two of the SNPs included in the ForenSeq ancestry-informative category are also phenotypic (pigmentation) SNPs, and this highlights the non-independence of human ancestry and EVCs (Figure 5.2).

**Table 5.1 Overview of ancestry-informative marker (AIM) sets.**

Tool	No. of markers	Description	Reference
DNAprint Ancestry-by-DNA forensic set	178 SNPs	For individual ancestry / admixture proportions from 4 continental populations (Europeans, W. Africans, Indigenous Americans, and E. Asians), based on 100 unrelated individuals (Coriell cell repository, <a href="http://ccr.coriell.org">http://ccr.coriell.org</a> , and (Bonilla et al. 2005))	(Halder et al. 2008)
forensic ancestry panel	47 SNPs	For major continental groups (Africa, Asia, Eurasia, Americas). Based on 74 cell lines (Y Chromosome Consortium panel) from 6 regions; verified in 919 individuals from CEPH Human Genome Diversity Panel (HGDP-CEPH)	(Kersbergen et al. 2009)
Latin American Cancer Epidemiology (LACE) group's American population analysis panel	445 SNPs	Optimised to estimate proportions of Latin American ancestry, based on 953 individuals (African, European, Native American)	(Galanter et al. 2012)
SNPforID 34-plex forensic ancestry test	34 SNPs	Based on 709 individuals from the reference collection of NIST (National Institute of Standards and Technology), representative of the 4 main U.S. population groups: 261 Caucasians, 258 African Americans, 140 Hispanics, 50 E. Asians (self-declared). 66 populations from 1KGP and HGDP-CEPH panels used as reference data.	(Fondevila et al. 2013)
FrogKB	55 SNPs	Based on 3884 individuals (73 populations), and additionally tested on 2969 individuals (52 populations)	(Kidd et al. 2014; Pakstis et al. 2010)
EUROFORGEN Global ancestry-informative SNP panel	128 SNPs	Divergent across 5 population groups (Africans, Europeans, E. Asians, Native Americans, Oceanians). Differentiation between European, Middle East and S. Asian ancestries is less defined in order to introduce Oceanian differentiation.	(Phillips et al. 2014)
Global AIMs Nano	31 SNPs	31-plex SNaPshot assay using EUROFORGEN Global AIM-SNP set, based on 1KGP and HGDP-CEPH. Aims to differentiate between Africans, Europeans, E Asians, Oceanians, Native Americans.	(De la Puente et al. 2016)
microhaplotype panel	65 SNP microhaplotypes	Based on 96 populations (5667 individuals)	(Bulbul et al. 2018)
ForenSeq DNA Signature Prep kit	54 ancestry-informative SNPs and 2 phenotypic SNPs	Based on (Kidd et al. 2014)	Verogen

The populations used to identify AIMs via allele frequency measurement are generally preselected indigenous groups. In practice, however, many individuals have ancestry from more than one such group, i.e. are admixed, so it is important to be able to assess this in addition to assigning to a single source population when this applies. There are a number of different ways to evaluate population structure and to consider admixture in genetic data. Principal Component Analysis, PCA, (Patterson et al. 2006; Price et al. 2006) is one of the most commonly used approaches, and is often applied to high-density single nucleotide polymorphism (SNP) genotyping data (Conomos et al. 2015b). Individuals are projected on a two-dimensional plot as points, and labelled according to their population of origin, after which their differentiation or clustering can be assessed visually; multi-dimensional scaling, MDS, (Purcell et al. 2007) is used in a similar way. Individuals who have admixture from two parental populations are expected to lie on a cline between parental clusters, with the position reflecting the degree of admixture (Patterson et al. 2006). An alternative approach is to use model-based techniques such as STRUCTURE, (Pritchard et al. 2000), FRAPPE (Tang et al. 2005) and ADMIXTURE (Alexander et al. 2009), which resolves multilocus data on a set of individuals into well-differentiated clusters, based on the assumptions that loci are in linkage equilibrium, and that populations are in Hardy-Weinberg equilibrium. These approaches require the user to specify K, the number of clusters. Many studies (e.g. Li et al. 2008; Rosenberg et al. 2002) present and interpret the results for several values of K, though the “best” value of K can be sought either by attempting to maximise the probability of K given the genotype data, or, taking individual genotype probabilities, on minimising the prediction error in a cross-validation approach. The outputs of these clustering algorithms are usually represented as plots in which each individual is represented by a line, divided into coloured segments corresponding to ancestry contributions from the K clusters. In such representations, admixed individuals can be identified and the proportions estimated from the lengths of segments.

In this Chapter, both PCA (undertaken using PLINK, described in Chapter 2, and Eigensoft [smartpca]) and the model-based clustering method ADMIXTURE are applied to the assessment of population structure and admixture.

## 5.1.2 Phenotype prediction

The prediction of externally visible characteristics (EVCs) from DNA evidence is of relevance for criminal investigations as it may act as a “biological witness” (Walsh et al. 2011), offering a physical profile of an unknown individual solely from a biological sample. This information may help in developing new investigative leads, in producing profiles of missing persons or disaster victims, and in confirming eyewitness accounts. Most of the work done so far in this field has focused on pigmentation phenotypes.

Skin colour variation in indigenous populations shows a highly non-random geographical pattern, with darker colours prevalent in tropical latitudes, and paler colours in more northerly regions (Jablonski and Chaplin 2000), although there is also variation among indigenous populations within regions such as Africa (Crawford et al. 2017). Many factors have been proposed to explain skin colour variation: natural selection acting on variable exposure to UV radiation and the need to maintain vitamin D synthesis or, within the same latitude and UV exposure range, assortative mating, drift, and epistasis (Katsara and Nothnagel 2019; Martin et al. 2017). Sex, age and hormones also play a role (Park and Lee 2005). Visible skin pigmentation has been classified and assessed in several ways:

- Skin phototyping (SPT) classification methods (e.g. the six Fitzpatrick phototypes), which are based on subjective classifications and participant questionnaires;
- Quantitative estimation of absorbance characteristic of human skin using a reflectance spectrophotometer (which provides erythema and melanin indices, the E-index and M-Index, respectively).

There is still a lack of full understanding of pigmentation, which is influenced by many genes and their interactions (Edwards et al. 2010), as well as environmental exposure and acclimatisation (tanning). Also, most studies are limited to data from populations of European origin, while the great variability in admixed populations and in populations of African origins has not been fully explored and is not fully represented by an African-American reference population. This bias may affect prediction models based on small panels of markers derived from European studies.

Eye colour variation is almost restricted to individuals of European origin, showing a gradient of colours from brown, which is considered to be the ancestral human phenotype, to intermediate (e.g. green) and blue (Walsh et al. 2012) due to melanin and melanosomes in the outer layer of the iris. Genome-wide association (GWAS) and candidate gene studies suggested several genes linked to the extremes of brown and blue human eye colour variation (Liu et al. 2009; Walsh et al. 2012); however, all these studies were based on indigenous populations of European origin. Variants in *HERC2* and *OCA2* are the most relevant for eye colour:

- rs12913832 (*HERC2*) may provide a regulatory effect to the neighbouring *OCA2* gene and explains most blue - brown eye colour variation (Liu et al. 2009);
- Valenzuela et al. suggest three SNPs that explained 76% of eye colour variation, rs12913832 *HERC2*, rs16891982 *SLC45A2* and rs1426654 *SLC24A5* (Valenzuela et al. 2010);
- rs885479 is almost fixed in Africans, shows considerably low variation in Europeans but displays high variation in Asians;
- rs6119471 no allelic variation in Europeans and Asians, but shows variation in Africans, which is in contrast with what is known on global distribution of eye colour variation; this SNP is not a good predictor for eye colour (Walsh et al. 2012).

Hair colour, like eye colour, shows a high degree of variation among European populations (Harrison 1973; Mengel-From et al. 2009; Shekar et al. 2008; Sulem et al. 2007); the usefulness of prediction is limited by the ease of altering hair colour artificially, and by phenomena such as greying, balding and darkening of the blond hair type between childhood and adulthood in some individuals.

A widely used prediction tool is HIrisPlex-S (Erasmus Medical Center, <https://hirisplex.erasmusmc.nl/>). This prediction model is based on multinomial logistic regression (Liu et al. 2009) and uses a set of 41 SNPs, designed to predict colour for eye, hair, and skin (Table 5.1). Some markers have more weight than others and the model can handle a certain amount of missingness. The HIrisPlex-S system is based on studies carried on European populations, but it has been tested on worldwide samples (HGDP-CEPH H952, the true phenotypes are not known, but assumed in accordance to expected

global pigmentation distribution, so not representing a true validation) (Walsh et al. 2014).

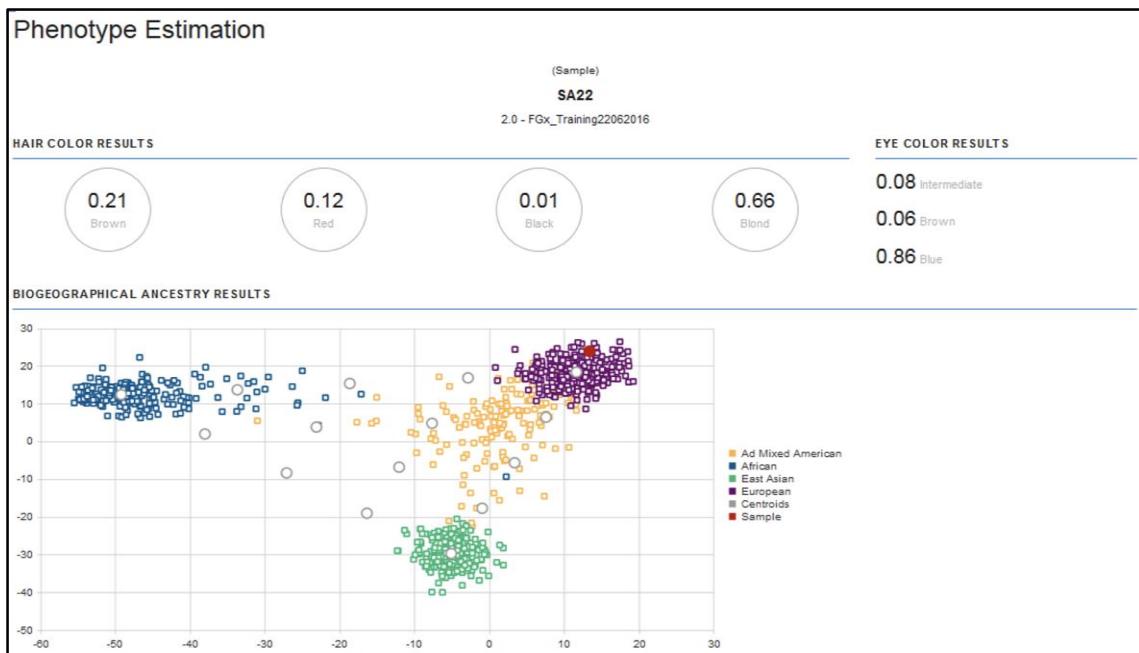
**Table 5.1 The HIrisPlex-S system: the model integrated markers to include eye, hair as well as skin colour prediction.**

Platform	N. of markers	markers	Prediction	Reference
IrisPlex	6	rs12913832 ( <i>HERC2</i> ), rs1800407 ( <i>OCA2</i> ), rs12896399 ( <i>SLC24A4</i> ), rs16891982 ( <i>SLC45A2</i> [ <i>MATP</i> ]), rs1393350 ( <i>TYR</i> ) and rs12203592 ( <i>IRF4</i> )	eye colour	(Liu et al. 2009)
HIrisPlex	18	1 INDEL polymorphism N29insA; 10 SNPs from the <i>MC1R</i> gene, rs11547464, rs885479, rs1805008, rs1805005, rs1805006, rs1805007, rs1805009, <i>Y152OCH</i> , rs2228479; rs1110400, rs28777 ( <i>SLC45A2</i> ), rs12821256 ( <i>KITLG</i> ), rs4959270 ( <i>EXOC2</i> ), rs1042602 ( <i>TYR</i> ), rs2402130 ( <i>SLC24A4</i> ), rs2378249 ( <i>ASIP/ PIGU</i> ), rs683 ( <i>TYRP1</i> )	eye and hair	(Walsh et al. 2013)
HIrisPlex-S	36 (16 genes)	19 of the 24 HIrisPlex SNPs	eye, hair and skin	(Chaitanya et al. 2018; Walsh et al. 2017)

### 5.1.3 ForenSeq kit: approach to BGA and pigmentation estimation

The ForenSeq DNA Signature Prep. kit provides a PCR primer mix (DNA primer mix B) that targets 230 amplicons, including 56 SNPs (aSNPs) for biogeographical ancestry determination and 22 phenotypic SNPs (pSNPs), in addition to standard forensic markers for individual identification. The supplied analysis interface (Universal Analysis System, UAS) outputs a result for biogeographical ancestry prediction based on Principal Component Analysis. The PCA is automatically created plotting the sample together with the 1000 Genomes Project panel (Phase 3) as reference data, and the sample can be compared to the clusters representing the main meta-populations (African, East Asian, European, “Ad Mixed American”). An example of such a PCA is shown in Figure 5.1. For pigmentation, the UAS includes a predictive tool for eye and hair colour based on the previous versions of the HIrisPlex (excluding skin colour), providing the probabilities of various pigmentation phenotypes (Figure 5.1). Table 5.2 lists the variants and the relevant genes for the HIrisPlex, and those that are used in the ForenSeq system.

These approaches to BGA and pigmentation estimation may be valid and useful for indigenous individuals, but admixed individuals may provide less informative results. In PCA, individuals may fall within the generalised “Ad Mixed American” cluster. This could be misleading, as the African-American population that has been used as reference is not representative of many admixed individuals, and of other African-admixed populations. In the pigmentation analysis, the simple output may perform poorly for admixed individuals where epistatic effects may play a role in the phenotypes, in ways that are not included in the model.



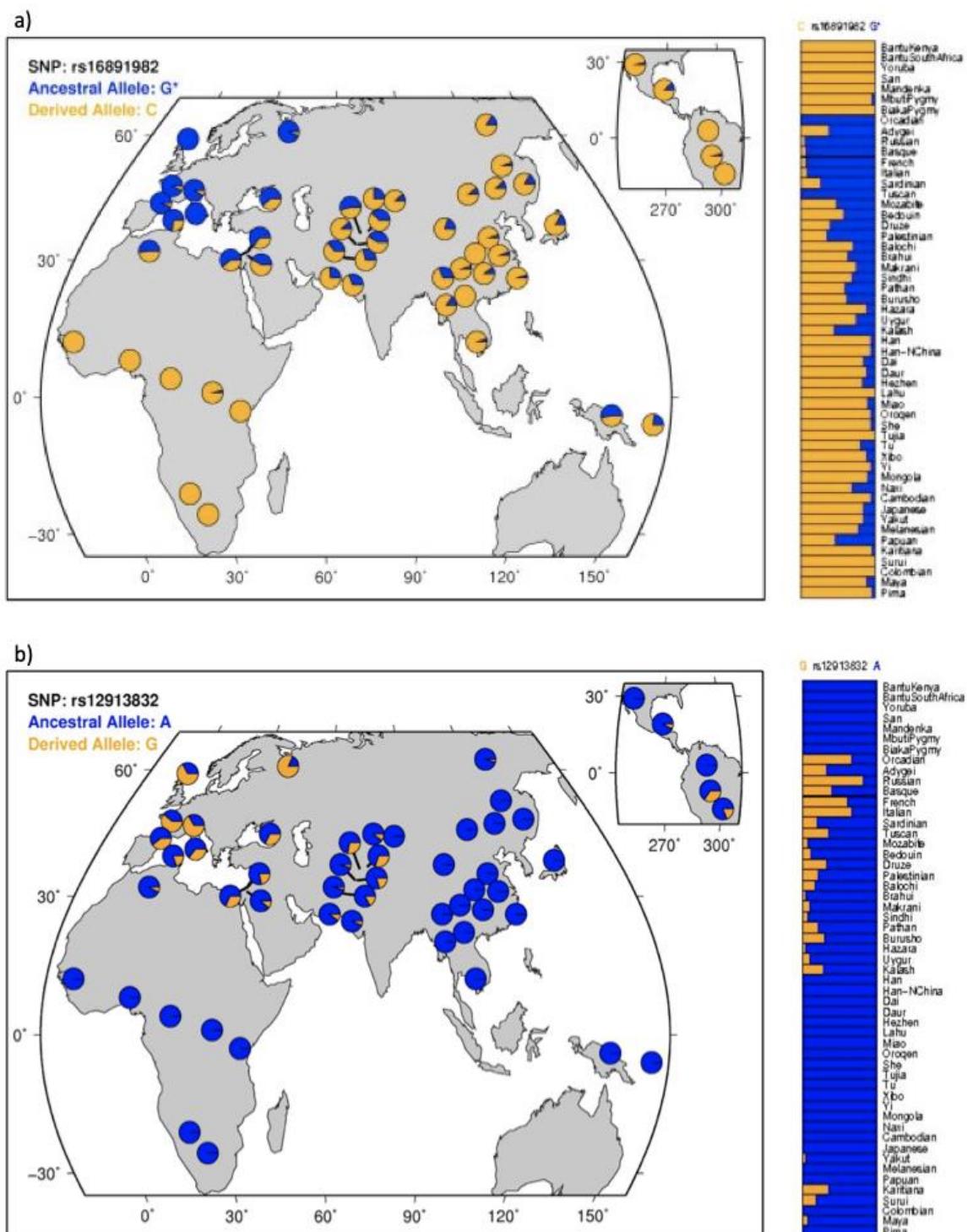
**Figure 5.1 UAS Phenotype and Biogeographic Ancestry (BGA) estimation, showing probabilities of various pigmentation phenotypes and ancestry prediction.**

Results for an anonymous test sample are shown. The pigmentation phenotypes are divided into hair colour and eye colour results, with a predictive value associated with each colour (the higher the value, the higher the probability of being the true pigmentation). The PCA informs on the BGA results: the 1000 Genome Project populations (reference dataset) are coloured squares in the plot, while centroids are empty circles; the queried sample is a red point.

**Table 5.2 List of 24 variants used for phenotypic prediction in the ForenSeq system.**  
 These phenotypic SNPs are all part of the HIrisPlex-S. Two variants (rs12913832, rs16891982) are also part of the ancestry SNPs set.

Gene	rsID	target pigment	h19 position	h38 position
<i>SLC45A2</i>	rs28777	hair	chr5:33,958,959-33,958,959	chr5:33958854-33958854
<i>IRF4</i>	rs12203592	eye	chr6:396,321-396,321	chr6:396321-396321
<i>LOC105374875</i>	rs4959270	hair	chr6:457,748-457,748	chr6:457748-457748
<i>TYRP1</i>	rs683	hair	chr9:12,709,305-12,709,305	chr9:12709305-12709305
<i>TYR</i>	rs1042602	hair	chr11:88,911,696-88,911,696	chr11:89178528-89178528
	rs1393350	eye	chr11:89,011,046-89,011,046	chr11:89277878-89277878
<i>KITLG</i>	rs12821256	hair	chr12:89,328,335-89,328,335	chr12:88934558-88934558
<i>LOC105370627</i>	rs12896399	eye	chr14:92,773,663-92,773,663	chr14:92307319-92307319
<i>SLC24A4</i>	rs2402130	hair	chr14:92,801,203-92,801,203	chr14:92334859-92334859
<i>OCA2</i>	rs1800407	eye	chr15:28,230,318-28,230,318	chr15:27985172-27985172
<i>MC1R</i>	N29insA	hair	chr16:89,985,752-89,985,753	chr16:89919344-89919345
	rs1110400	hair	chr16:89,986,130-89,986,130	chr16:89919722-89919722
	rs11547464	hair	chr16:89,986,091-89,986,091	chr16:89919683-89919683
	rs1805005	hair	chr16:89,986,091-89,986,091	chr16:89919683-89919683
	rs1805006	hair	chr16:89,985,918-89,985,918	chr16:89919510-89919510
	rs1805007	hair	chr16:89,986,117-89,986,117	chr16:89919709-89919709
	rs1805008	hair	chr16:89,986,144-89,986,144	chr16:89919736-89919736
<i>TUBB3</i>	rs1805009	hair	chr16:89,986,546-89,986,546	chr16:89920138-89920138
<i>MC1R</i>	rs201326893_Y152OC H	hair	chr16:89,986,122-89,986,122	chr16:89919714-89919714
	rs2228479	hair	chr16:89,985,940-89,985,940	chr16:89919532-89919532
	rs885479	hair	chr16:89,986,154-89,986,154	chr16:89919746-89919746
<i>PIGU</i>	rs2378249	hair	chr20:33,218,090-33,218,090	chr20:34630286-34630286
<i>HERC2</i>	rs12913832*	eye	chr15:28,365,618-28,365,618	chr15:28120472-28120472
<i>SLC45A2</i>	rs16891982*	eye	chr5:33,951,693-33,951,693	chr5:33951588-33951588

\* these phenotypic SNPs are also considered as ancestry SNPs in the ForenSeq kit



**Figure 5.2 HGDP selection browser maps for two phenotypic and ancestry SNPs.**

The variants rs12913832 (in *HERC2*) and rs16891982 (in *SLC45A2*) contribute to both ancestry and phenotypic prediction, thus showing how closely linked these factors are. They are part of standard AIMs sets and also included in the ForenSeq kit as both ancestry and phenotypic SNPs.

### **5.1.4 Aims of this Chapter**

This Chapter aims to explore the ancestry prediction approach of the Illumina analysis interface (Universal Analysis System, UAS), based on the ForenSeq DNA Signature Prep Kit (Verogen, San Diego, CA) markers, and to compare this with alternative model-based approaches, including both markers explicitly designed for ancestry estimation, and others within the multiplex. It evaluates ForenSeq's ancestry and phenotype prediction, together with the HIrisPlex-S system (Erasmus Medical Center, <https://hirisplex.erasmusmc.nl/>) model on samples of European-African admixed origins (Cape Verde), for whom measured phenotypes and ancestry estimates from genome-wide SNP chip data are available. The samples are described in (Beleza et al. 2013), and the subset analysed here was chosen to cover a wide range of ancestral admixture and pigmentation phenotypes.

## **5.2 Materials and methods**

### **5.2.1 Sample description**

DNA samples from 30 individuals from Cape Verde with known ancestry proportions based on genome-wide data genotyped (Illumina Infinium HD Human1M-Duo Bead array, data from (Beleza et al. 2013) and measured skin, hair and eye colour phenotypes were sequenced using the ForenSeq DNA Signature Prep Kit (Verogen, San Diego, CA).

These individuals were chosen according to their diverse geographic ancestry within Cape Verde (8 from the island of Santiago, 8 from Fogo, 14 from the cluster of São Vicente, Boa Vista, Santo Antão, São Nicolau and North-West cluster), their range of genome-wide African ancestry proportions, and their known phenotypes (eye and skin colour) in order to survey the maximum range possible. Further description of samples can be found in (Beleza et al. 2013). Hair colour is not considered here, because of the prevalence of the dyeing or shaving of head hair (S Beleza, University of Leicester, personal communication). The true phenotypes of individuals for eye and skin colour are known, based on quantitative measurements: a modified melanin index (MM-Index) for skin pigmentation was calculated as the square root of Melanin Index of upper inner arms

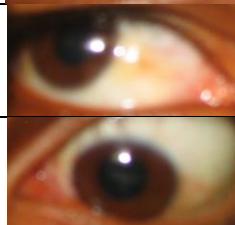
measured using the DSMII ColorMeter spectrophotometer (Cortex Technology, Denmark) and a T-Index for eye colour was obtained through an automated algorithm that quantitatively measures colour based on the Red, Green, and Blue (RGB) values from digital photographs (Beleza et al. 2013). Higher values indicate lighter colours for eye pigmentation (i.e. blue, with a suggested T-Index cut-off at 0.25) and lower values indicate darker colours for skin pigmentation (S. Beleza, University of Leicester, personal communication). Ancestry proportions based on genome-wide data are also known (Beleza et al. 2013).

**Table 5.3 Summary of information on the 30 Cape Verdean samples.**

This table reports the ancestry proportions (European and African), the Modified Melanin index for skin pigmentation and the T-index for eye pigmentation. It is ordered from higher to lower T-index.

Sample	Island	Ancestry proportions		Modified Melanin Index	T-index	Eye picture
		European proportion	African proportion			
CVSA34	Santo Antao	0.41704	0.58296	45.15	0.33097935	
CVSA111	Santo Antao	0.552665	0.447335	47.12	0.29812619	
CVST800	Santiago	0.252788	0.747212	48.89	0.28293361	
CVBV12	Sao Vicente	0.356494	0.643506	47.19	0.27998478	
CVF323	Fogo	0.42033	0.57967	46.91	0.27457122	
CVBV75	Boa Vista	0.359017	0.640983	54.52	0.25928612	
CVF303	Fogo	0.593917	0.406083	41.87	0.2578066	
CVSA99	Santo Antao	0.587525	0.412475	32.76	0.24401089	
CVSN112	Sao Vicente	0.413189	0.586811	43.29	0.19271047	

<b>CVST357</b>	Fogo	0.439044	0.560956	64.45	0.18443629	
<b>CVST202</b>	Fogo	0.493036	0.506964	59.86	0.13330693	
<b>CVST511</b>	Fogo	0.662044	0.337956	40.35	0.11200486	
<b>CVSN68</b>	Sao Nicolau	0.493015	0.506985	54.15	0.10413404	
<b>CVST285</b>	Santiago	0.196127	0.803873	64.03	0.09774783	
<b>CVST561</b>	NWcluster	0.51568	0.48432	36.37	0.09774783	
<b>CVST505</b>	Santiago	0.525776	0.474224	52.28	0.08989502	
<b>CVSV156</b>	Boa Vista	0.453929	0.546071	39.12	0.08891381	
<b>CVSN41</b>	Santo Antao	0.362489	0.637511	63.01	0.08449898	
<b>CVST813</b>	Santiago	0.254273	0.745727	88.36	0.08106558	
<b>CVST459</b>	Fogo	0.445599	0.554401	51.42	0.07861313	
<b>CVSA133</b>	Santo Antao	0.391241	0.608759	44.26	0.07812263	
<b>CVF105</b>	Fogo	0.525887	0.474113	43.84	0.07174518	
<b>CVST207</b>	Santiago	0.27882	0.72118	66.48	0.07125452	
<b>CVST310</b>	Santiago	0.329349	0.670651	73.83	0.06487417	
<b>CVF26</b>	Fogo	0.422417	0.577583	72.19	0.06389228	

<b>CVSV63</b>	Santo Antao	0.458842	0.541158	63.79	0.05554291	
<b>CVST284</b>	Santiago	0.299601	0.700399	56.68	0.05406896	
<b>CVSN119</b>	Sao Nicolau	0.437491	0.562509	52.2	0.04571551	
<b>CVSA44</b>	Santo Antao	0.420086	0.579914	50.21	0.03883991	
<b>CVST349</b>	Santiago	0.226183	0.773817	70.25	0.01058958	

### 5.2.2 Sample processing and sequencing

From blood spots on FTA paper (Whatman), 10 punches (Whatman Harris Uni-Core™ punch) of 2 mm diameter for each sample underwent a Chelex-based extraction, adapting the methodology from (Hailemariam et al. 2017). The Qubit® 2.0 fluorometer (Thermo Fisher Scientific) with the Qubit® dsDNA HS kit was used to verify the quantity of double-stranded DNA. Sequencing using the MiSeq ForenSeq Forensic Genomics System is described in Section 3.2.1. It was performed in one run, including 30 samples and 2 controls.

### 5.2.3 Data analysis

The genotype calling workflow is described in Section 3.2.4, and includes:

- processing sequencing data on the MiSeq Illumina platform, UAS, with genotypes and flanking region reports;
- checking raw data (FASTQ files) and variant calling using STRaitRazor v 3 (Woerner et al. 2017b) (thresholds were lowered to 2 reads and manually checked);
- applying a MPS workflow (using raw data), based on BWA (Li and Durbin 2009) and samtools (Li et al. 2009): cleaned raw reads were aligned to a custom reference (GRCh 37, UCSC); the minimum read depth in calling variants was set to 20; VCF variants were queried in dbSNP (build 151).

### 5.2.4 Biogeographical ancestry inference

The samples' biogeographical ancestry was assessed using: PCA utilising Eigensoft smartpca v7.2.1; PLINK v1.90b6.3 (Purcell et al. 2007), and ADMIXTURE (Alexander et al. 2009). Components determined through model-based approaches (i.e. ADMIXTURE) were plotted via Cluster Markov Packager Across K (CLUMPAK, Kopelman et al. 2015), available at <http://clumpak.tau.ac.il/index.html>, a post-processing online tool. These analyses were performed using both genome-wide SNP data and MPS data, with different combinations of markers. The available markers from data generated here are ancestry SNPs, identity SNPs, phenotypic SNPs, and autosomal STRs as targeted by the ForenSeq kit and SNPs found in the flanking regions of the targeted markers (Table 5.4). dbSNP (build 151) was used to identify the rsIDs of SNPs found in the flanking regions.

**Table 5.4 Number of SNPs and STRs available in the set of 30 Cape Verdean samples.**

The ForenSeq target markers as well as those variants identified in the amplicon and flanking regions are considered here. One variant found by the UAS in the identity SNP flanking region was also found via the MPS workflow. To note, the UAS highlights also the variants that show the reference allele. The information on STR amplicon regions are still subject to final checks and were not included in the analyses. na = not applicable.

	ForenSeq SNP			ForenSeq STRs	Total SNPs available
	Ancestry	Identity	Phenotypic		
No. of markers in the kit	56	94	22	58	172
No. of SNPs found in flanking regions (MPS workflow)	3	10*	0	27	212
No. of SNPs found in flanking regions (UAS)	14 (and 1 deletion)	41	6	na	234
Total	74	144	28	85	

\* 2 are not present in the 1000 Genomes Project Phase 3 dataset

## **Principal Component Analysis**

Eigensoft smartpca v7.2.1 was used to perform PCA. Individuals from 1000 Genomes Project Phase 3 (The 1000 Genomes Project Consortium et al. 2015) were included as reference datasets (for a total of 1911 individuals): 503 individuals from the European metapopulation (British [GBR], n = 91; French [CEU], n = 99; Finnish [FIN], n = 99; Iberian [IBS], n = 107; Toscani [TSI], n = 107), 504 from the East Asian group (Dai Chinese [CDX], n = 93; Han Chinese [CHB], n = 103; Japanese [JPT], n = 104; Kin Vietnamese [KHV], n = 99; Southern Han Chinese [CHS], n = 105), 96 African Caribbean [ACB], 504 African (Esan [ESN], n = 99; Gambian Mandinka [GWD], n = 113; Luhya [LWK], n = 99; Mende [MSL], n = 85; Yoruba [YRI], n = 108), 49 African Ancestry [ASW], and 255 American individuals who were defined as “Admixed” (Colombian [CLM], n = 93; Mexican Ancestry [MXL], n = 58; Puerto Rican [PUR], n = 104). The datasets were sub-setted to include only the same markers as in the Cape Verde dataset.

## **Model-based ancestry determination**

The chosen number of population components ( $K$ ) took into account the known settlement history of the islands and the cross-validation error estimates were considered for each  $K$  (from 1 to 5, Appendix 5b). The obtained ancestry coefficients (Q-matrices) are then used as input in CLUMPAK for visualisation.

### **5.2.5 Phenotype analysis**

Pigmentation phenotypes were reported and predicted using two different methods: the UAS phenotypic report, which is based on genotypes for 24 markers of the HIrisPlex (Walsh et al. 2014) and is produced only if all markers are typed and above a 30-read threshold; and the HIrisPlex-S webtool (Chaitanya et al. 2018, available at <https://hirisplexerasmusmc.nl/> accessed in May 2019), which has the additional potential of predicting skin colour. Genotypes obtained from sequencing were converted to the input format for the HIrisPlex-S web platform through an in-house script and the predicted colour was accepted if the prediction value was above 0.5 (accepted ranges vary from  $\geq 0.5$  to  $\geq 0.9$ , Walsh et al. 2012).

## 5.3 Results

### 5.3.1 Sequencing: depth of coverage, missingness, concordance with SNP chip data

#### Universal analysis software output

Sequence data were processed on the UAS platform, and manually checked (alleles under threshold, unbalanced markers, depth of coverage [DoC]): new thresholds were set, lowering them to 20 reads for homozygous SNPs and to 10 for heterozygous SNPs. STR sequences were checked for presence of stutter. These thresholds are similar to those found in the literature (Khurani et al. 2019). Positive control had 59/59 STRs typed and 167/172 SNPs typed (issues with this control are also reported by Sharma et al. 2017; Sharma et al. 2019).

Some markers were flagged by the UAS: D22S1045 (imbalance), rs1736442 (dropout), DY5385a-b (+1 allele). There were 2 ancestry SNPs (rs310644 in 12 samples, rs3811801 in 10 samples) and 4 phenotypic SNPs (rs1393350 in 8 samples, rs1805005 in 3 samples, and the group rs1805005, rs1805006 and rs2228479 in 5 samples) under threshold. According to Walsh et al. (2014), *MCIR* markers are more prone to drop-out than other markers, generally due to low DNA input.

The genotypes obtained were compared to the SNP chip data to provide concordance with previous studies (Beleza et al. 2013); based on 37 iSNPs, 99.8% concordance was found: two SNPs in two samples were called heterozygous in the SNP chip and homozygous by the UAS (rs733164 for CVST349, with DoC of 41; rs1523537 for CVST207, with DoC of 36). Two SNPs in two samples were called only by the UAS (rs4606077 for CVST813, with 121 DoC; rs2107612 for CVST202, with 526 DoC). To support the calling of alleles, output was analysed using an independent tool, STRait Razor (Woerner et al. 2017b).

These processing steps highlighted some under-performing SNPs, which have a significant impact on the prediction output report of the UAS. The identity SNP rs1736442 is excluded from the analyses. rs12913832 is also a key SNP for eye and hair pigmentation. Other research (Frégeau 2021) highlights that many phenotypic SNPs present low DoC (rs12821256, rs28777 and rs2378249 for hair colour; rs12203592 and rs12913832 for both hair and eye), as well as some ancestry SNPs (rs310644, rs10497191,

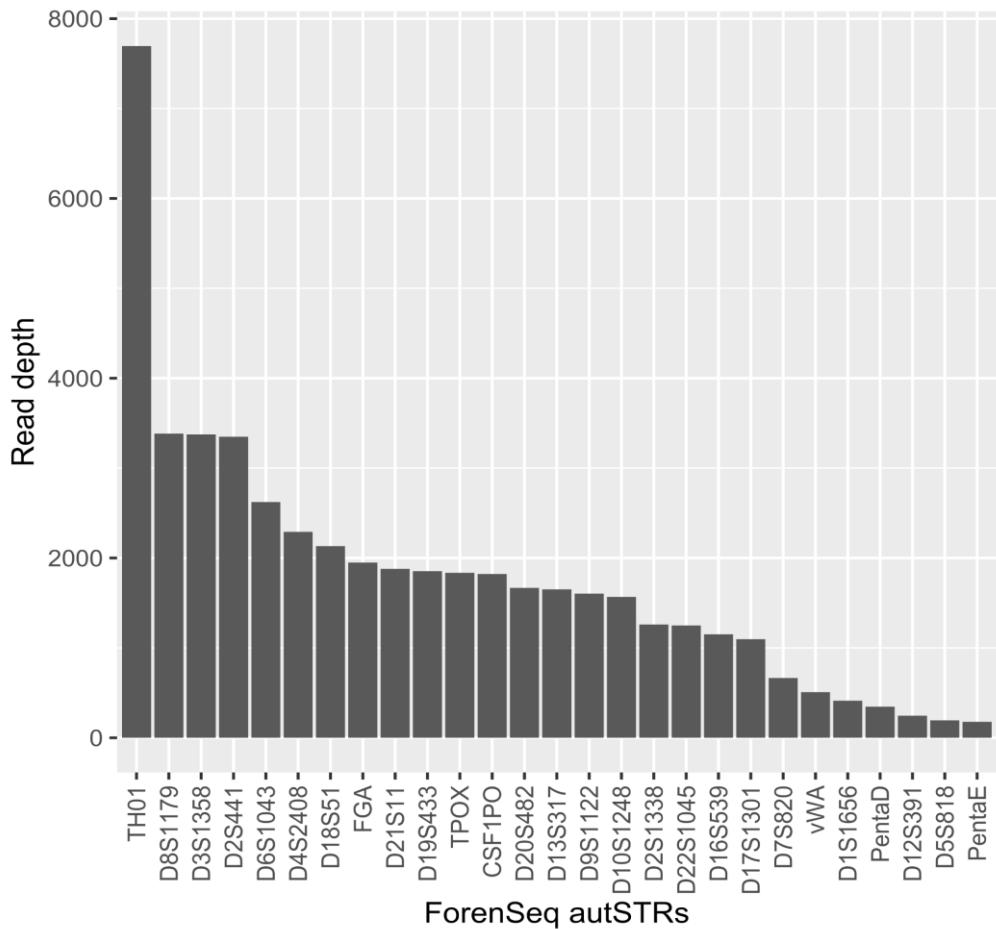
rs3916235, rs3737576 and rs1834619). Frégeau (2021) and Sharma et al. (2019) found ~50-fold average read count difference between the lowest and highest DoC (for rs310644 and rs2024566, respectively). According to them, there is a correlation between low read depth, amplicon length and AT-content (and its effect on PCR efficiency).

**Table 5.5 Description of seven ForenSeq Plex B under-performing markers, ordered according to the number of samples affected.**

Int. thresh: interpretation threshold issue (an allele between the analytical and interpretation thresholds).

Under-performing marker rsID	Chr:position (GRCh37)	Type of marker	issue	Number of samples with issue
rs1736442	chr18:55,225,777	identity	not called	27
rs826472	chr10:2,406,631	identity	int. thresh. and not called	17 (of which 6 not called)
rs310644	chr20:62,159,504	ancestry	int. thresh. and imbalance	11
rs3811801	chr4:100,244,319	ancestry	int. thresh. and not called	10 (of which 1 not called)
rs1805005	16:89,985,844	phenotypic	int. thresh. and not called	8 (of which 3 not called)
rs1393350	chr11:89,011,046	phenotypic	int. thresh. and not called	8 (of which 1 not called)
rs3823159	chr6:136,482,727	ancestry	int. thresh.	7
rs1805006	chr16:89,985,918	phenotypic	int. thresh. and not called	5 (of which 3 not called)
rs2228479	chr16:89,985,940	phenotypic	int. thresh. and not called	5 (of which 3 not called)

As well as SNPs, autosomal, Y- and X-STRs were sequenced in the Cape Verde samples. Alleles are reported in Appendix 5e, but these markers are not used in the ancestry prediction approaches described here, because equivalent data in the reference 1000 Genomes Project samples are unavailable. This point will be revisited in the Discussion to this Chapter. Some information on read depth and isometric alleles is reported here (Figure 5.3).



**Figure 5.3 Read depth of 27 autosomal ForenSeq STRs.**

### MPS workflow output

The variants for the Cape Verdean samples were called through an MPS workflow (described in Chapter 3). Of the ForenSeq kit's 172 targeted SNP markers, 170 could be successfully called (the identity SNP rs1736442 is not called for many samples and rs1805005, a phenotypic SNP, is not called in any sample). Eleven additional SNPs were called in the ForenSeq SNP regions (Table 5.6).

LD is non-existent among the ForenSeq markers, with exception of phenotypic SNPs, given that 10/24 SNPs lie in *MC1R*, two in *SLC45A2*, and two in *TYR* (more information on LD is in Table 5.7) and some markers found in amplicon flanking regions. ADMIXTURE does not explicitly account for LD and excluding LD is difficult, especially for recently admixed populations, but it is possible to mitigate its effect

excluding makers according to their correlation coefficient or by including markers that are distant from each other. In the analysis undertaken here, LD is not considered.

**Table 5.6 Linkage among the clusters of phenotypic SNPs.**

Based on Table 5.2, the SNPs in three genes are reported here. This table shows the two markers (Variant1 and Variant 2), their minor allele frequency (MAF) based on 1KGP (dbSNP and [http://grch37.ensembl.org/Homo\\_sapiens/Tools/AlleleFrequency](http://grch37.ensembl.org/Homo_sapiens/Tools/AlleleFrequency)), and Linkage Disequilibrium (based on Gambian in Western Division [GWD] and Iberian [IBS] populations, 1KGP <https://ldlink.nci.nih.gov/?tab=home>). LE: Linkage equilibrium.

Gene	Variant 1			Variant 2			$r^2$ (GWD)	$r^2$ (IBS)
	rsID	Chr:pos. (GRCh37)	MAF GWD; IBS	rsID	Chr:pos.	MAF GWD; IBS		
TYR	rs1042602	chr11: 88,911,696	0; 0.3925	rs1393350	chr11: 89,011,046	0; 0.285	LE	0.2576
MC1R	rs312262906	chr16: 89,985,752	0; 0	rs885479	chr16: 89,986,154	0; 0.0187	LE	LE
	rs312262906	chr16: 89,985,752	0; 0	rs1805006	chr16: 89,985,918	0; 0.0093	LE	LE
	rs1805006	chr16: 89,985,918	0; 0.0093	rs2228479	chr16: 89,985,940	0; 0.0514	LE	0.0005
	rs11547464	chr16: 89,986,091	0; 0.028	rs1805005	chr16: 89,985,844	0; 0.1542	LE	0.0053
	rs1805005	chr16: 89,985,844	0; 0.1542	rs1805007	chr16: 89,986,117	0; 0.0327	LE	0.0062
	rs1805007	chr16: 89,986,117	0; 0.0327	rs201326893	chr16: 89,986,122	0; 0	LE	LE
	rs201326893	chr16: 89,986,122	0; 0	rs1110400	chr16: 89,986,130	0; 0.0234	LE	LE
	rs1110400	chr16: 89,986,130	0; 0.0234	rs1805008	chr16: 89,986,144	0; 0.093	LE	0.0002
SLC45A2	rs28777	chr5: 33,958,959	0.1327; 0.0738	rs16891982	chr5: 33,951,693	029 0.0044; 0.8178	0.029	0.6479

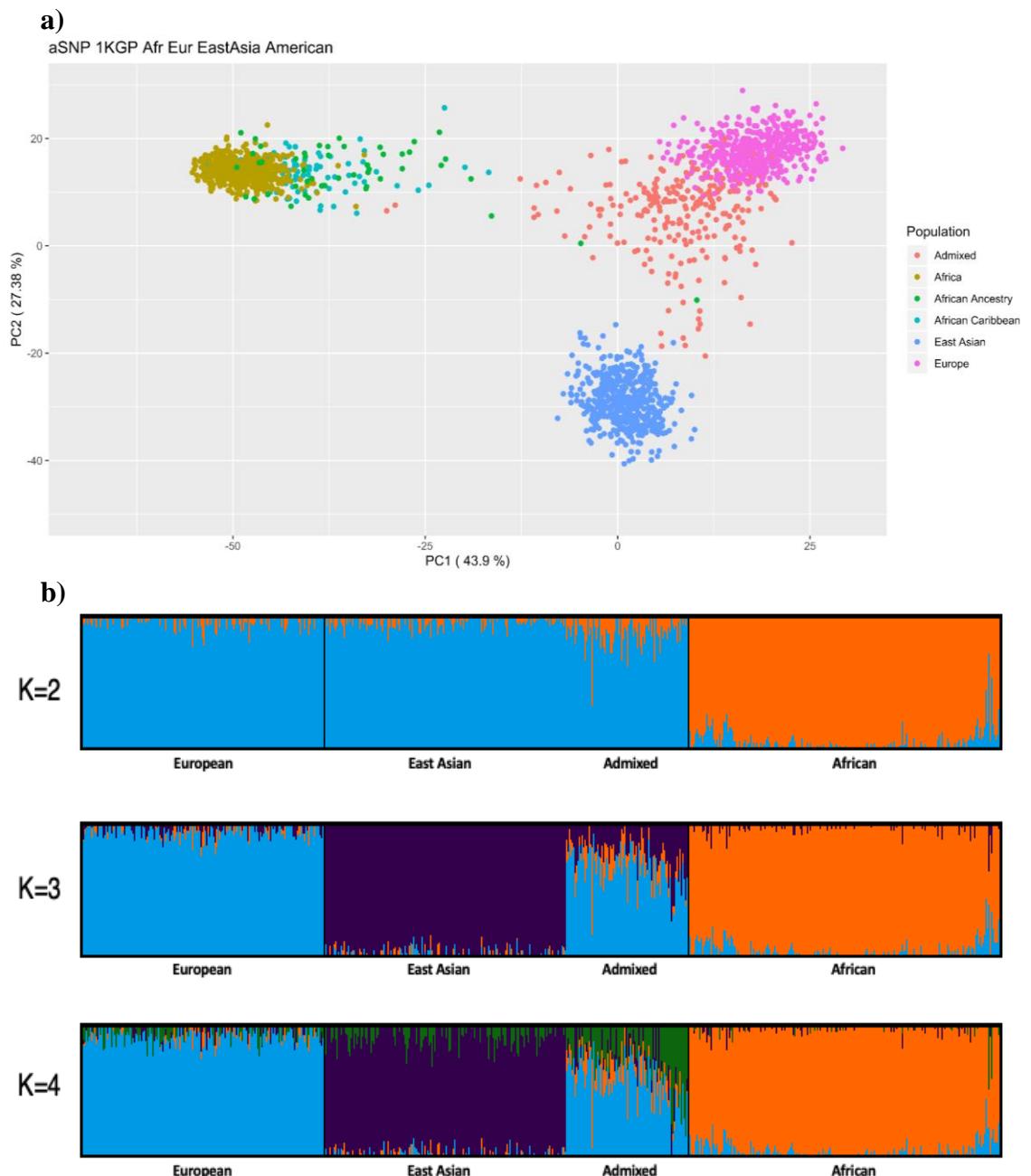
**Table 5.7 Summary information on markers called in the ForenSeq SNP regions through the MPS workflow.**

This table describes the position of the called SNP, its minor allele frequency (MAF) based on 1KGP (dbSNP and [http://grch37.ensembl.org/Homo\\_sapiens/Tools/AlleleFrequency](http://grch37.ensembl.org/Homo_sapiens/Tools/AlleleFrequency)), target SNP, and Linkage Disequilibrium between the two markers (based on Gambian in Western Division [GWD] and Iberian [IBS] populations, 1KGP <https://ldlink.nci.nih.gov/?tab=home>). LE: Linkage equilibrium.

Flanking SNP	Chr:pos. (GRCh37)	MAF based on Africans (and GWD)	MAF based on Europeans (and IBS)	Target: rsID (pos. GRCh37)	r <sup>2</sup> (GWD)	r <sup>2</sup> (IBS)
rs1005534	chr20:39,487,218	0.1445 (0.15)	0.0805 (0.08)	identity: rs1005533 (chr20:39,487,110)	0.1932	0.0757 (LE)
rs116227288	chr2:158,667,240	0.0885 (0.10)	0.0010 (0)	ancestry: rs10497191 (chr2:158,667,217)	0.0082 (LE)	LE
rs76875728	chr9:137,417,276	0.0045 (0)	0.0040 (0)	identity: rs10776839 chr 9:137417308)	LE	LE
rs75920625	chr8:145,639,654	0.1490 (0.16)	0.0030 (0.01)	ancestry: rs1871534 (chr8:145,639,681)	0.042 (LE)	0.0771 (LE)
rs112167443	chr21:43,606,944	0.0023 (0.01)	0.0865 (0.14)	identity: rs221956 (chr21:43,606,997)	0.0247 (LE)	0.3028
rs73514221	chr16:5,868,743	0.1218 (0.08)	0.0000 (0.00)	identity: rs2342747 (chr16:5,868,700)	0.069 (LE)	LE
rs381840	chr16:78,017,077	0.9508 (0.98)	1.0000 (1.00)	identity: rs430046 (chr16:78,017,051)	0.0124 (LE)	LE
rs12453170	chr17:53,568,855	0.3336 (0.31)	0.3897 (0.4)	ancestry: rs4471745 (chr17:53,568,884)	0.0312 (LE)	0.0399 (LE)
rs1390470	chr6:165,045,290	0.9796 (0.98)	1.0000 (1.00)	identity: rs727811 (chr6:165045334)	0.0082 (LE)	LE
rs16991914	chr22:33,559,555	0.1989 (0.17)	0.0020 (0.00)	identity: rs987640 (chr22:33,559,508)	0.1493	0.0029 (LE)
rs999428518	chr16:33,559,555	0	0	identity: rs729172 (chr16:5,606,197)	not in 1KGP reference panel	not in 1KGP reference panel

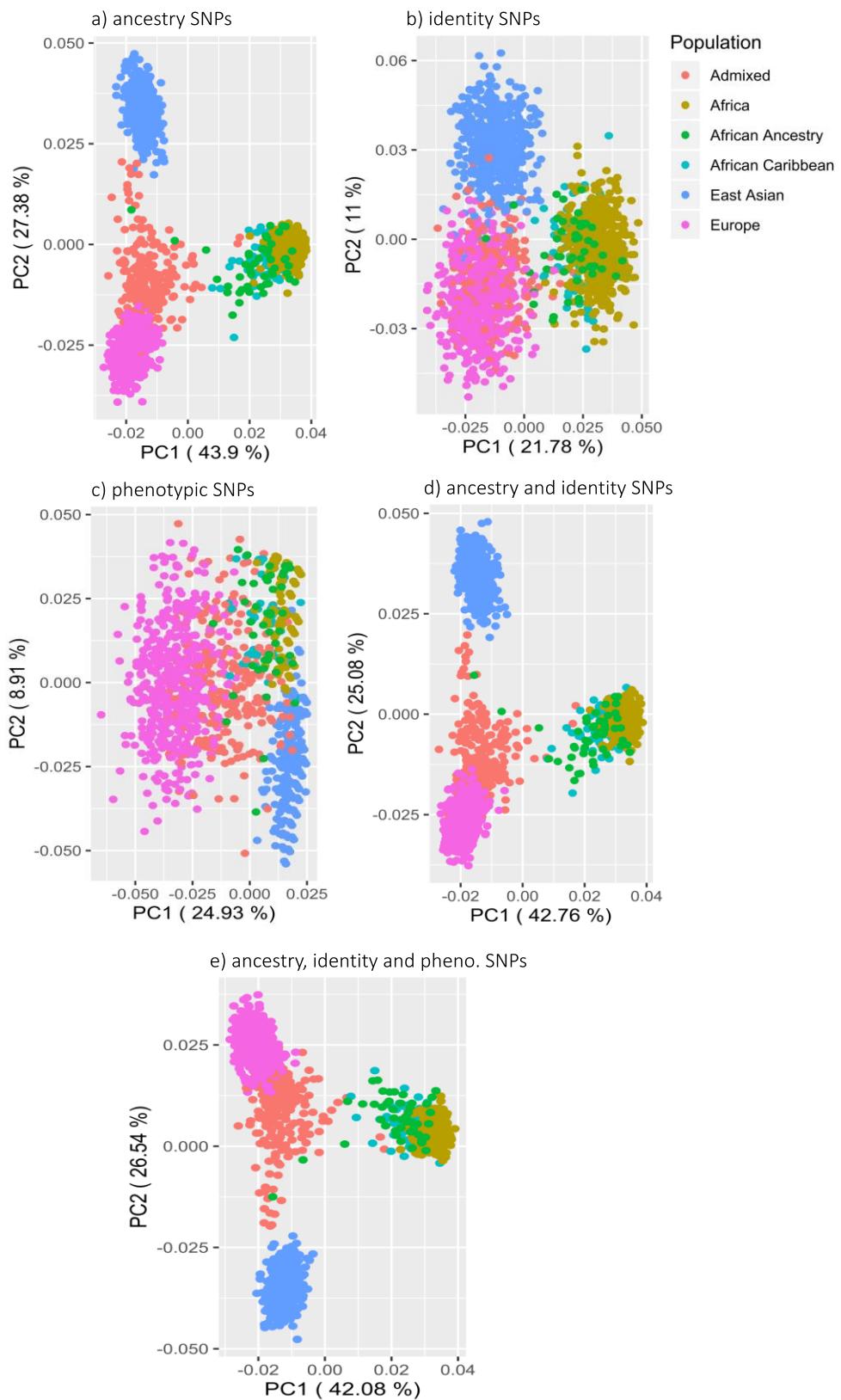
### **5.3.2 Ancestry prediction of 1000 Genomes Project samples: considerations on ForenSeq SNPs and UAS output**

As described before (section 5.1.3, and Chapter 3, section 3.2.1), the ForenSeq kit includes 56 ancestry SNPs used for biogeographical ancestry prediction by the UAS platform: a PCA based on 1KGP samples as reference is plotted. Figure 5.4 shows a PCA, modified to reproduce the UAS output: the 1KGP samples were used and the 56 ancestry SNPs. This is followed by ADMIXTURE analysis, which offers information on the ancestral proportions. The best value of  $K$  for 1KGP is 3 (Appendix 5b).



**Figure 5.4 Analysis of ancestry in African, European, East Asian and Admixed metapopulations from the 1000 Genomes Project using the ForenSeq ancestry SNPs (n= 56).**  
 (a) PCA plot reproducing the UAS approach; (b) ADMIXTURE plots for  $K = 2-4$ . Peruvians are not included in the “Admixed” group. The African group in the ADMIXTURE plot includes African, African Ancestry and African Caribbean samples highlighted in the PCA.

The following PCA plots (Figure 5.5) show the behaviour of the reference samples when including additional markers. These plots were not modified to reproduce the UAS PCA format, so the orientation of the axes is different, but the information content is the same.



**Figure 5.5 PCA plots of African, European, East Asian and Admixed metapopulations from 1000 Genomes Projects using different sets of ForenSeq SNPs.**

(a) ancestry SNP (n=56); (b) identity SNPs (n= 91); (c) phenotypic SNPs (n= 23); (d) ancestry and identity SNPs (n= 147); (e) ancestry, identity and phenotypic SNPs (n= 168), if considering the flanking, there are 179 SNPs (not shown here). Peruvians are not included in the “Admixed” group.

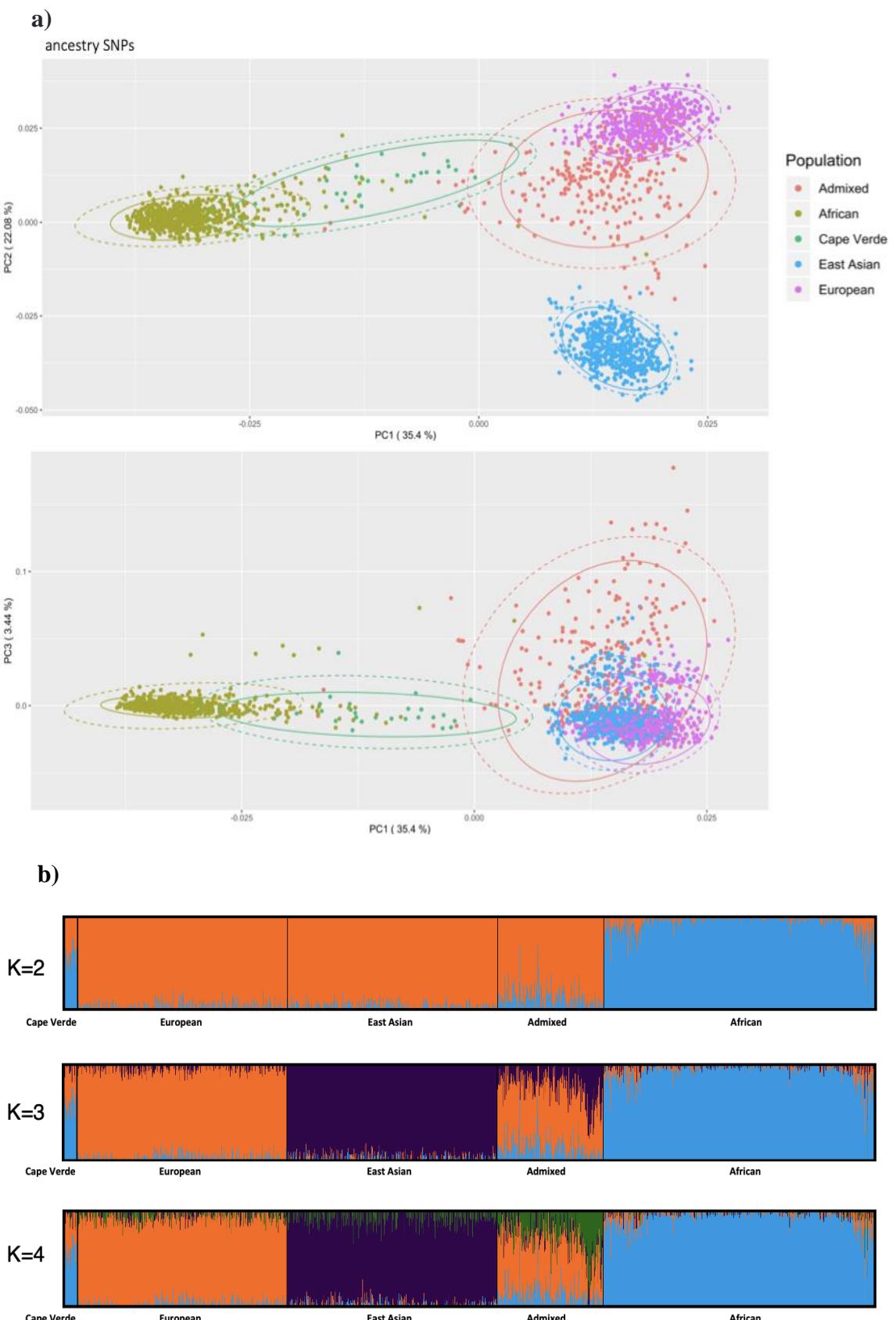
Comparing the other plots with that for aSNPs, a number of features emerge. Despite being ascertained for individual identification, the iSNPs display marked population structure, albeit with less differentiation between clusters. In the pSNPs plot the differentiation of Africans vs non-Africans is poor, and the major (PC1) differentiation observed is of Europeans vs non-Europeans. This likely reflects the Eurocentric bias of the pSNPs in the kit. Combining the SNPs (as aSNPs plus iSNPs, or as all SNPs) somewhat improves cluster definition.

### **5.3.3 Ancestry prediction of an admixed population: considerations on ForenSeq kit and UAS output**

The previous section showed the output from analysis of reference populations of known ancestral composition. Here, a set of 30 admixed individuals from Cape Verde is analysed.

The following plots show the output of Eigensoft (smartpca) and ADMIXTURE analysis. The variants called in the flanking regions were also considered for inclusion, but this involved only a small increase in the total number of available markers (for example, increasing the number of markers in ancestry SNP amplicons from 56 to 59). These additional variants therefore add little information and are not included in the PCA and ADMIXTURE analyses. Figure 5.6 shows the results considering the aSNPs. The best  $K$  in ADMIXTURE was 3, having the lowest cross-validation error compared to other  $K$  values.

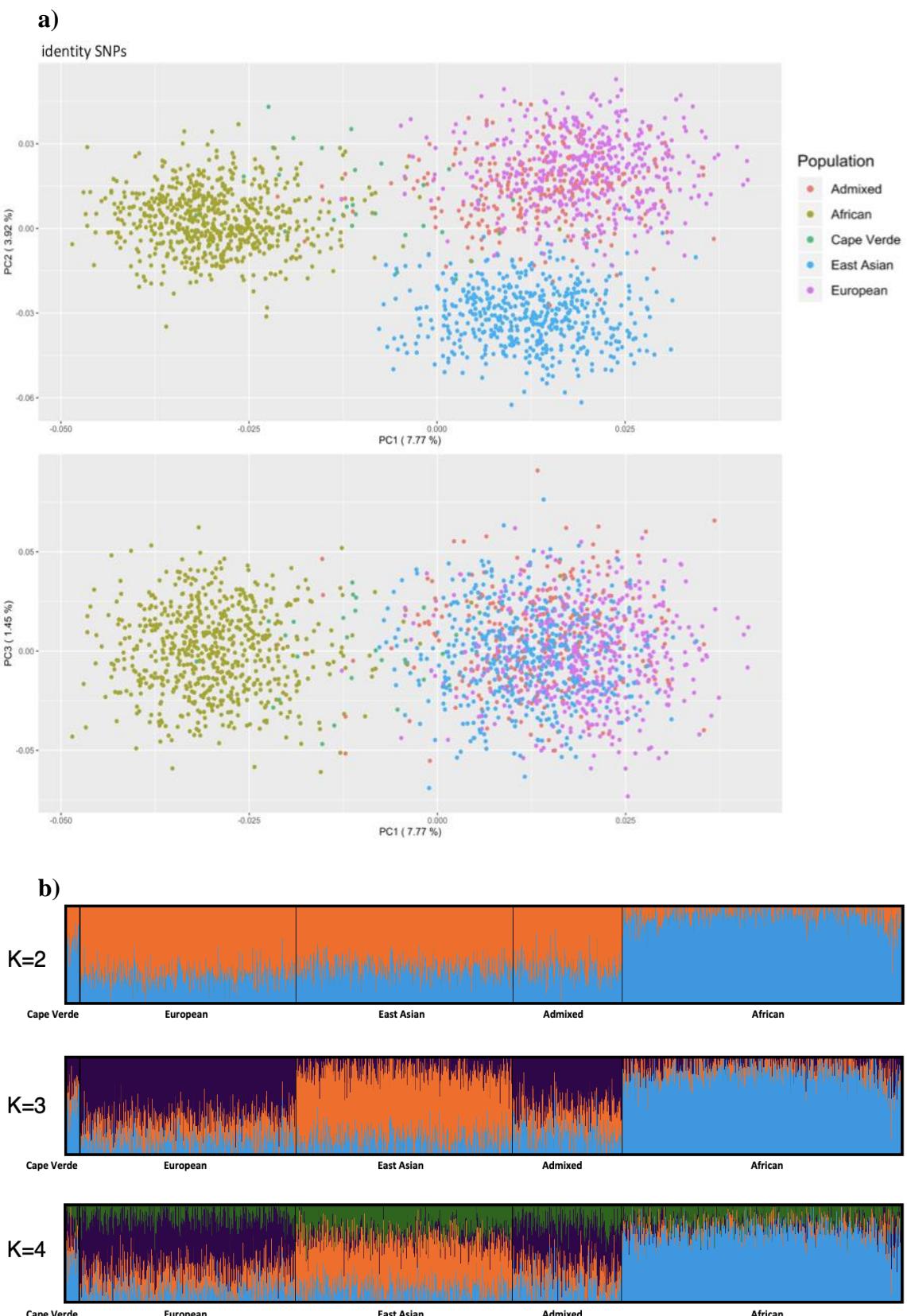
The following ADMIXTURE analyses demonstrate how the UAS choice of reference may limit the conclusions on sample ancestry. ADMIXTURE allows better exploration of the ancestry proportions, going beyond clustering determined by a limited number of reference populations. Overall, the Cape Verdean samples behave consistently throughout the analyses based on different marker subsets.



**Figure 5.6 Analysis of ancestry in Cape Verdean samples and reference set based on African, European, East Asian and Admixed metapopulations from the 1000 Genomes Project using the ForenSeq ancestry SNPs (n= 56).**

(a) PCA plot; (b) ADMIXTURE plots for  $K = 2-4$ ; best  $K = 3$ .

The analysis based on identity SNP markers (Figure 5.7) shows a broadly similar pattern at the population level to that based on the ancestry SNPs (particularly at  $K = 2$  and 3): the African, European and East Asian clusters are differentiated, and the Admixed and Cape Verdean groups are dispersed between them, overlapping their source populations (African-European), though as expected this pattern is much “noisier”. At  $K=4$ , a green component appears in all populations. This component, which is highest in the Admixed group, may reflect Native American ancestry; note that no known Native American group is included in this ADMIXTURE analysis.

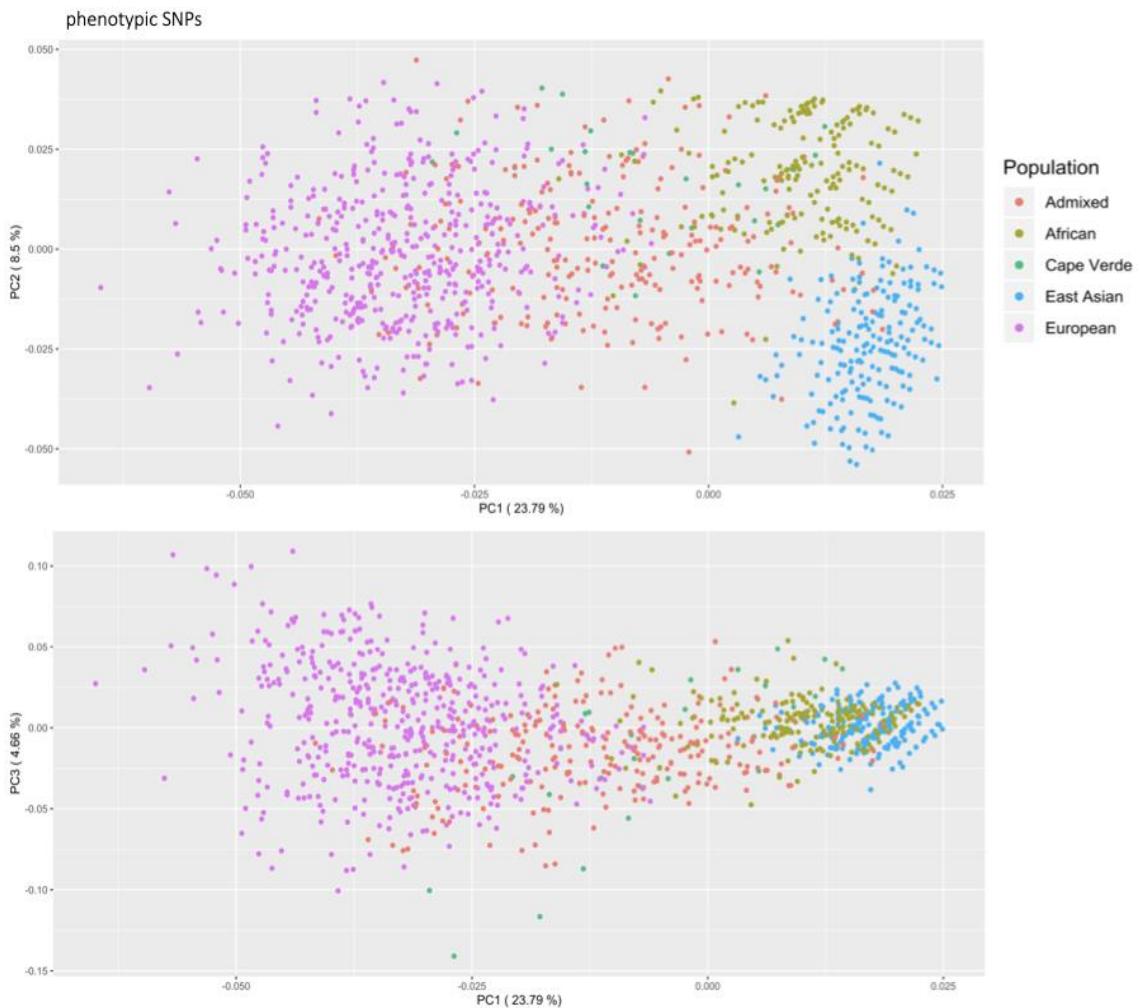


**Figure 5.7 Analysis of ancestry in Cape Verdean samples and reference set based on African, European, East Asian and Admixed metapopulations from the 1000 Genomes Project using the ForenSeq identity SNPs (n= 93).**

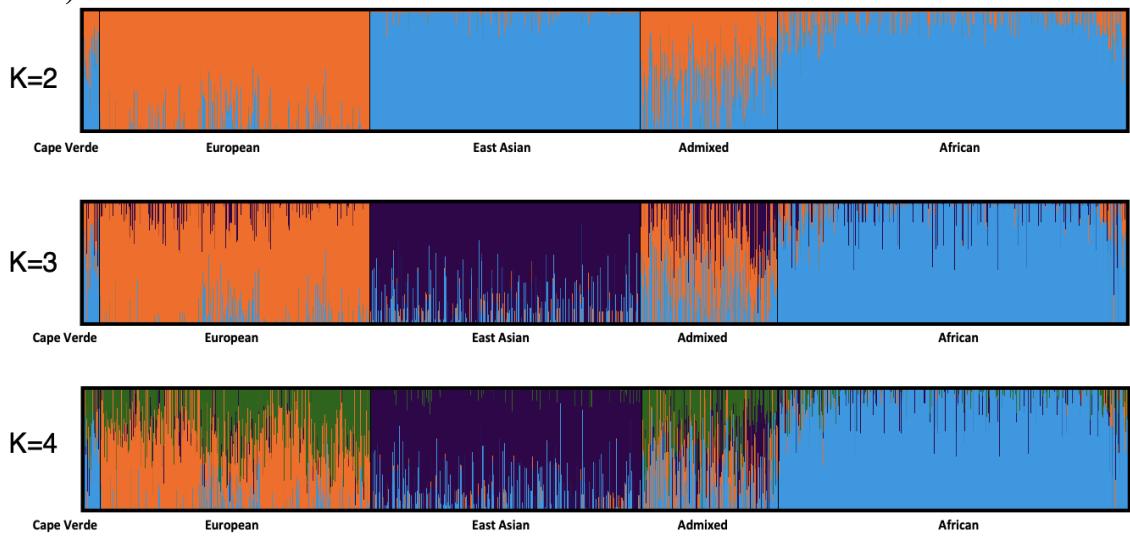
(a) PCA plot; (b) ADMIXTURE plots for  $K = 2-4$ ; best  $K = 3$ .

The markers provide a good African and non- African discrimination in PC1, but, for example, the phenotypic set (Figure 5.8) provides misleading results at  $K = 2$  for African against East Asian ancestry (as also seen in Figure 5.5), confirming that the best choice of  $K$  is 3. Adding these markers to a larger set does not have a strong impact, however, by themselves, these SNPs perform poorly: this may be due to the way the SNPs were selected, they may be uninformative for some categories (i.e. eye pigmentation in Asia or Africa) as they are specifically Eurocentric, and many relevant pigmentation SNPs are not present (e.g. skin colour is not well captured). The architecture of these SNPs is complex (see Discussion). It is possible to notice that the larger the set of markers ( $n = 171$ ), the clearer the differentiation (Figure 5.10); however, the iSNPs and pSNPs (Figure 5.7; 5.8), because of relatively low interpopulation allele frequency differences, contribute less to differentiation than the aSNPs (Figures 5.6).

a)

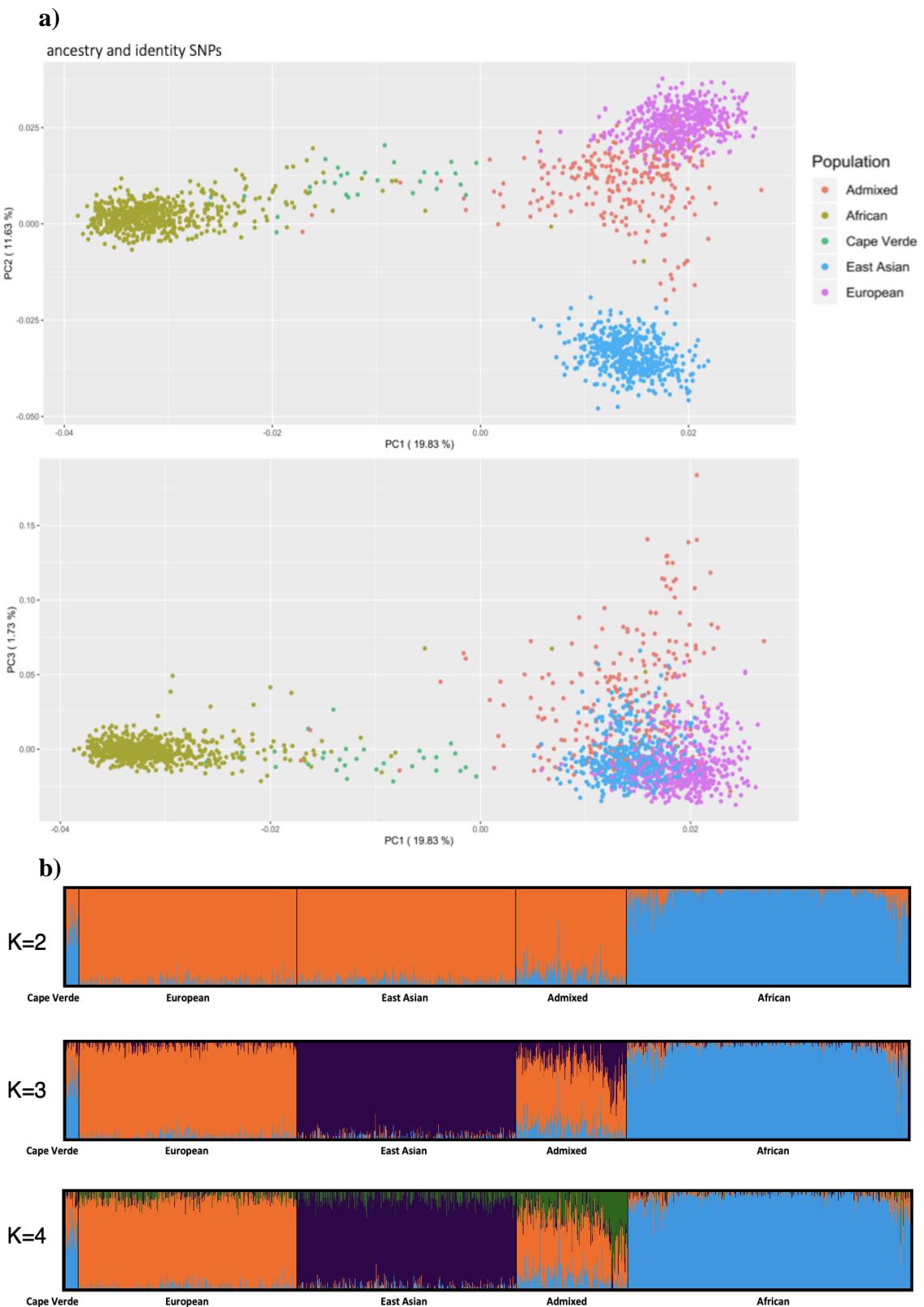


b)



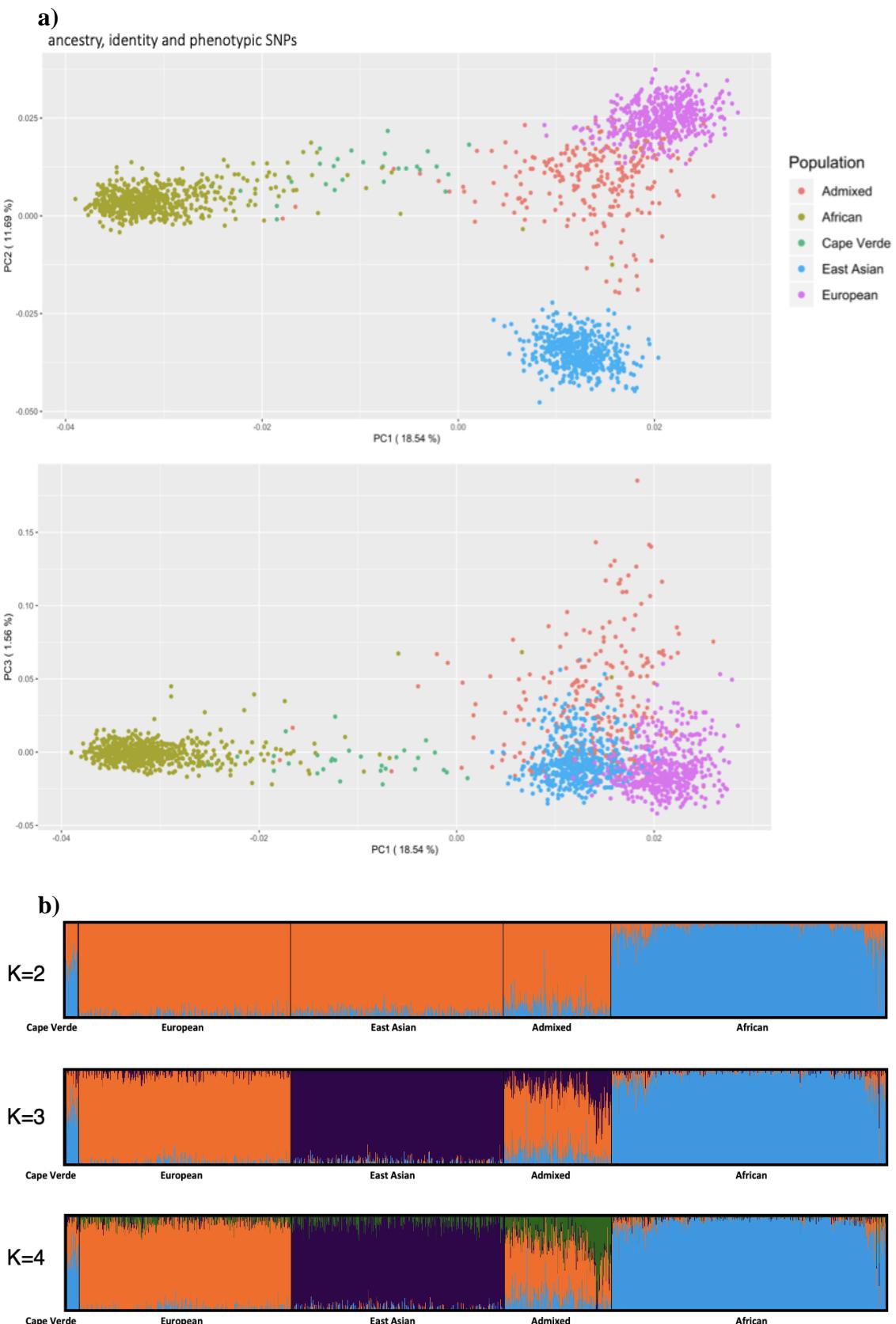
**Figure 5.8 Analysis of ancestry in Cape Verdean samples and reference set based on African, European, East Asian and Admixed metapopulations from the 1000 Genomes Project using the ForenSeq phenotypic SNPs (n= 24).**

(a) PCA plot; (b) ADMIXTURE plots for  $K = 2-4$ ; best  $K = 3$ .



**Figure 5.9 Analysis of ancestry in Cape Verdean samples and reference set based on African, European, East Asian and Admixed metapopulations from the 1000 Genomes Project using the ForenSeq ancestry and identity SNPs ( $n = 149$ ).**

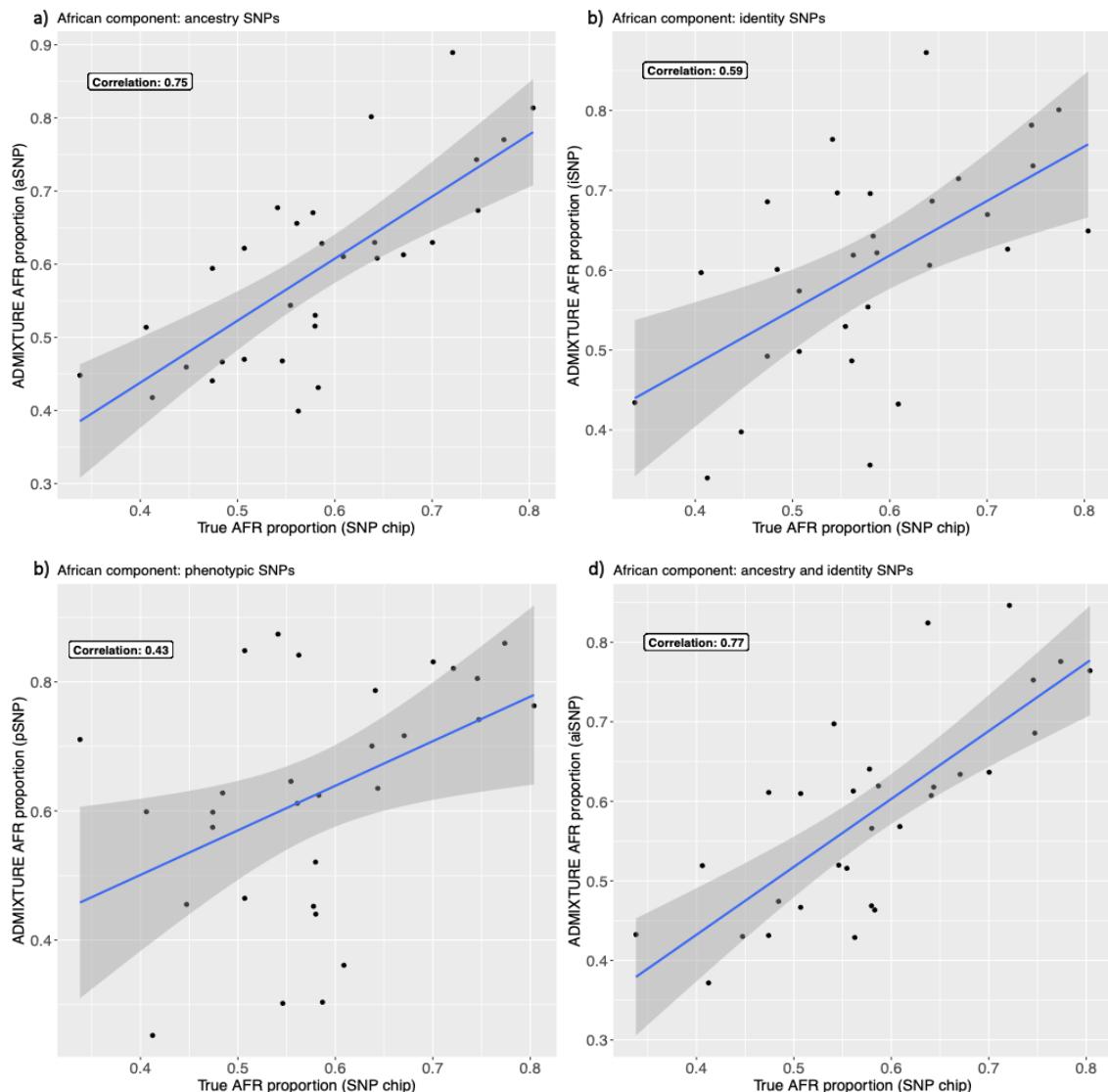
(a) PCA plot; (b) ADMIXTURE plots for  $K = 2-4$ ; best  $K = 3$ .

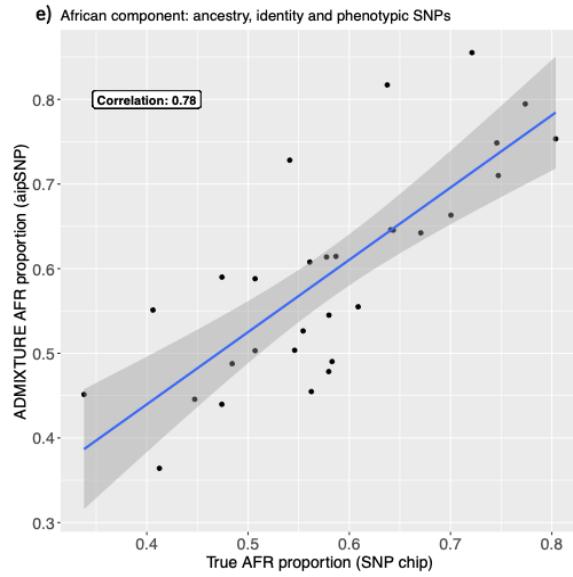


**Figure 5.10 Analysis of ancestry in Cape Verdean samples and reference set based on African, European, East Asian and Admixed metapopulations from the 1000 Genomes Project using the ForenSeq ancestry, identity and phenotypic SNPs (n = 171).**  
 (a) PCA plot; (b) ADMIXTURE plots for  $K = 2-4$ ; best  $K = 3$ .

## Comparing the estimated ancestry components

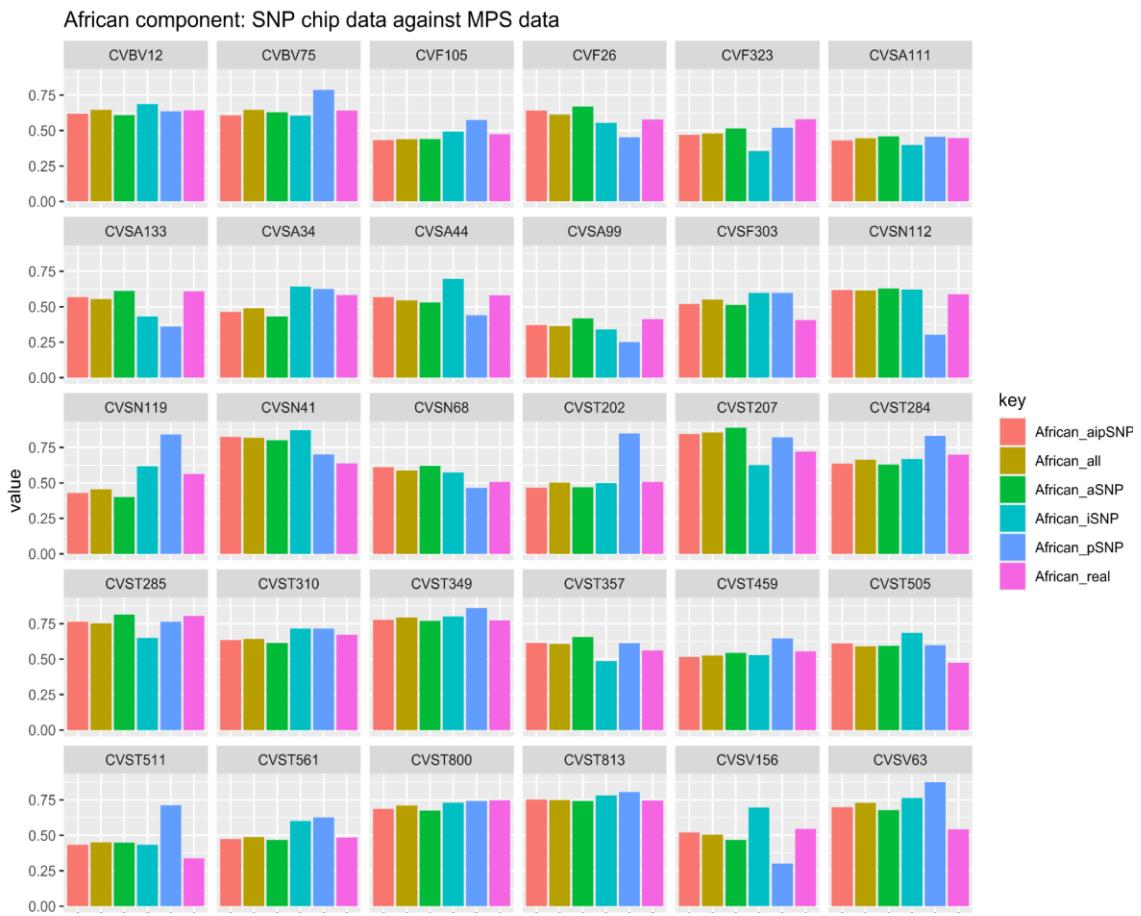
The number of SNPs available allows an exploration of which marker combination may offer a more detailed ancestry composition. Considering the ancestry proportions that were obtained from the original study (Beleza et al. 2013) as the “true” ancestral composition of each sample, the results obtained with five different marker subsets (based on ADMIXTURE) can be compared (Figure 5.11 and Figure 5.12). The subsets are: ancestry SNPs (56), identity SNPs (93), phenotypic SNPs (24), ancestry and identity SNPs (149), and all SNPs (173). These analyses are based on a two-population reference set (1KGP European, n = 503; African, n = 649, including 96 African Caribbean and 49 African ancestry), informed by the documented history of Cape Verde.





**Figure 5.11 Correlation between the African ancestry proportion based on the SNP chip and ADMIXTURE using different combinations of markers.**

(a) Ancestry SNP ( $r = 0.75$ ); (b) identity SNP ( $r = 0.59$ ); (c) phenotypic SNP ( $r = 0.43$ ); (d) ancestry and identity SNPs ( $r = 0.77$ ); (e) ancestry, identity and phenotypic SNP ( $r = 0.78$ ).

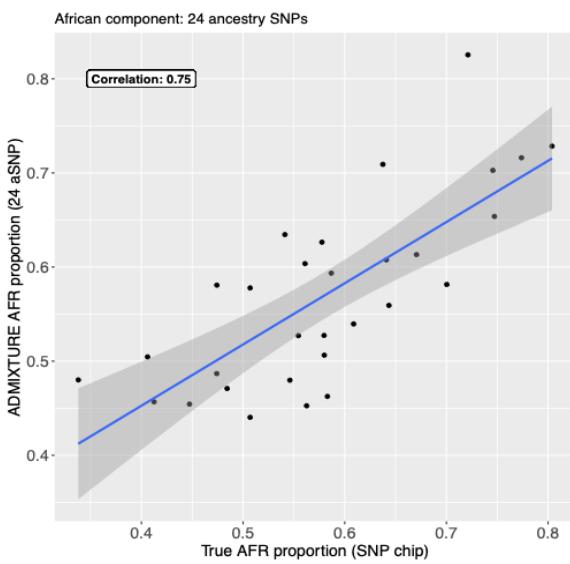


**Figure 5.12 Bar plot comparing individual African ancestry proportions as estimated from different sets of markers.**

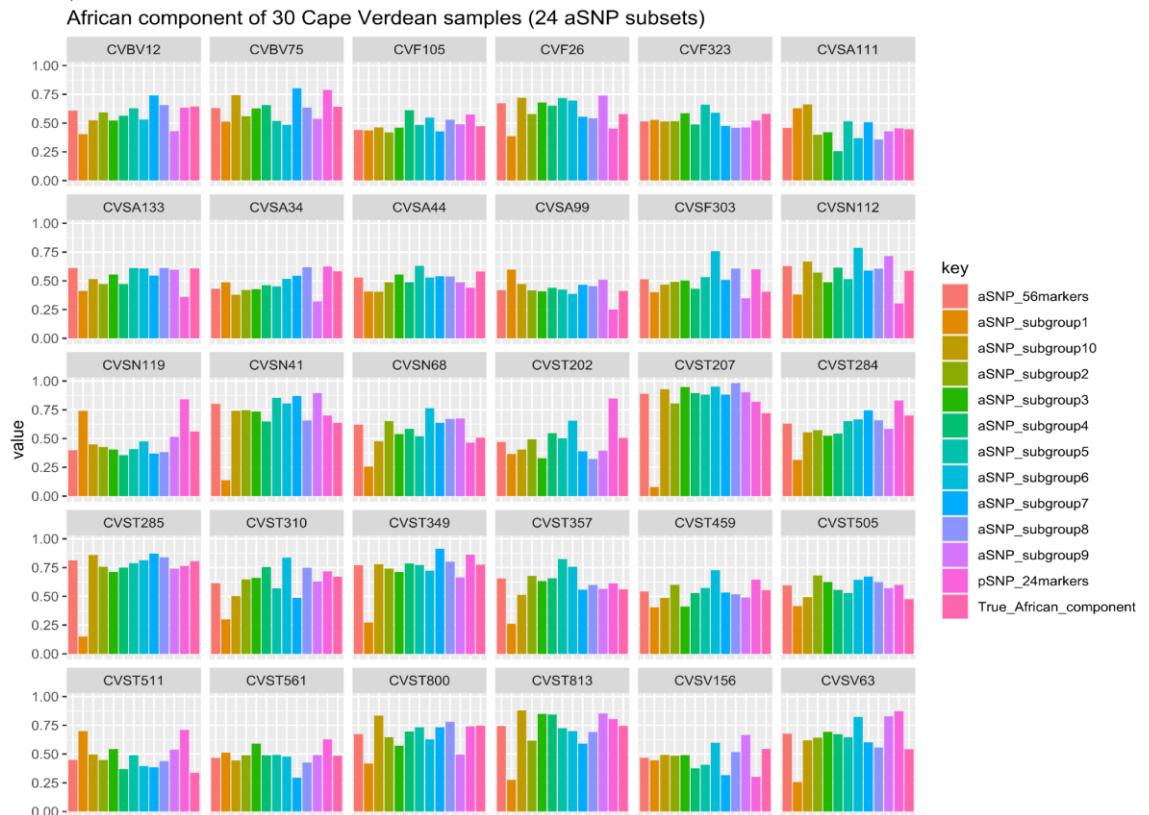
This plot summarises the predicted ancestry components using the SNP chip (Beleza et al. 2013) (considered as the “true” value), ancestry SNPs, identity SNPs, phenotypic SNPs, and a combination of the three (aipSNPs) estimated via ADMIXTURE for each sample.

It is noticeable that the output based on phenotypic SNPs has the weakest correlation with the true ancestry proportions based on SNP chip data (Figure 5.11). In principle, this might simply be due to the low number of markers used (24 pSNPs). To test this, ancestral components were calculated on smaller sets of 24 randomly selected aSNPs from the total set of 56 (Figure 5.13): the correlation between these random subsets and the “true” ancestry proportions remains good, so the low number of pSNPs is not likely to be responsible for their weak ancestry correlation. The most likely explanation is the relatively weak differentiation of this set of SNPs between African and European ancestry, as seen in Figure 5.8.

a)



b)



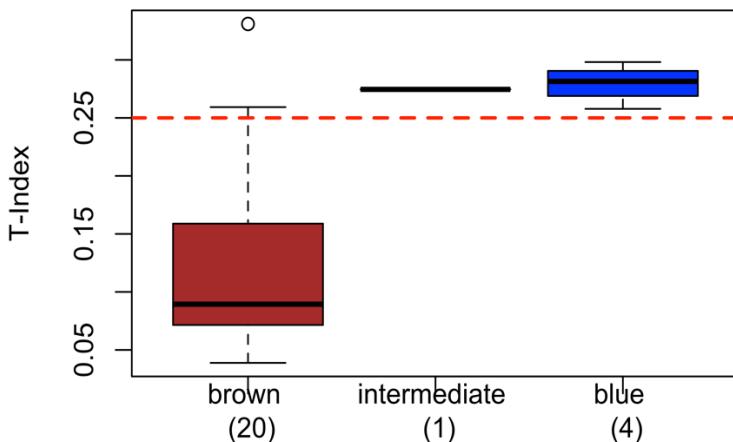
**Figure 5.13 Individual African ancestral proportions based on the average from ADMIXTURE output using random subsets of 24 ancestry SNPs (out of 56).**

(a) Correlation between the African ancestry proportion based on the SNP chip and the average from ADMIXTURE output: ten random subsets of 24 ancestry SNPs were used to explore the impact of a small number of markers on the analysis;  $r = 0.75$ . (b) This bar plot compares the samples' African ancestry proportion based on 56 ancestry SNPs against the proportion calculated on 10 subgroups of 24 ancestry SNPs, 24 phenotypic SNPs and the "true" African component based on the SNP chip data (Beleza et al. 2013); this plot intends to check the effect of low number of markers on ancestry proportion calculation to explore the output based on the limited number of phenotypic SNPs available.

### **5.3.5 UAS platform phenotypic and ancestry prediction**

UAS prediction outputs a probability value for each qualitative category (e.g. “blue”, “intermediate”, “brown” for eye colour), and the category with the highest probability value is the phenotype to be accepted (the cut-off generally accepted is 0.5, Walsh et al. 2011). The marker set for this model targets eye and hair colour (based on the HirisPlex-S), but includes some markers generally used for predicting skin colour (rs1800414 and rs1426654), not yet implemented in the UAS. In order to obtain a phenotype prediction all markers must be called. Instead, the missingness of certain markers does not limit the HirisPlex-S, where the prediction accuracy is expressed by the area under the receiver-operating characteristic curve (AUC) and the loss in AUC when partial profiles are considered (see manual, <https://hirisplex.erasmusmc.nl/pdf/hirisplex.erasmusmc.nl.pdf> ; Walsh et al. 2012). This situation may lead users of the kit to turn to HIrisPlex-S when no results are obtained by the UAS. As HIrisPlex-S also provides information for skin colour, it is necessary to evaluate the prediction results and to consider the expected limits and usefulness of the results. In this section, eye pigmentation is the main focus, but some considerations on skin pigmentation are also given.

The categorical pigmentation (e.g. “blue”, “intermediate”, “brown”) predicted by the UAS was plotted against the individuals’ actual pigmentation measure (e.g. T-Index), as in Figure 5.14. The real phenotype for eye colour is expressed as a “T-Index”, where the cut-off between brown and blue colour is at 0.25 (with blue colours corresponding to higher T-Index values). There are 5 individuals with no UAS phenotypic report (Table 5.8).



**Figure 5.14 The qualitative classification of the phenotype assigned by the UAS model for 25 Cape Verdean individuals compared to the real phenotype (the measure of eye pigmentation, T-Index).**

Cut-off between brown and blue is set at 0.25. Predicted brown = 20, intermediate = 1, blue = 4 (correct number of samples with high T-Index is 7).

**Table 5.8 Samples with no UAS phenotypic prediction.**

This table shows the reason why no UAS phenotypic report was obtained, how many SNPs were flagged, and what phenotypic prediction would be obtained if the SNPs were manually called. Interm. = intermediate colour.

Sample	total DoC	Reason	pSNP (flagged)	aSNP (flagged)	iSNP (flagged)	under threshold (eye prediction)	under threshold (hair prediction )
<b>CVST285</b>	352845	3 <i>MCR1</i> not called	na	1	5	na	na
<b>CVST349</b>	109819	3 <i>MCR1</i> not called and rs1393350 under threshold (17 reads)	2	5	21	interm. 0.01, brown 0.99, blue 0	brown 0.33, read 0, black 0.65, blond 0.03
<b>CVST310</b>	237500	rs1805006 under threshold (reads 19)	2	2	7	interm. 0.01, brown 0.99, blue 0	brown 0.39, red 0, black 0.59, blond 0.02
<b>CVSN41</b>	306858	3 pSNP under threshold (all 15 reads)	na	1	10	na	na
<b>CVF26</b>	279791	rs1393350 not called	na	1	7	na	na

Greater attention was given to the outlier (Table 5.9), and ancestry proportions were checked. According to the SNP chip data, this individual (from Santo Antão) shows similar proportions of African and European ancestry (African component = 0.58, European component = 0.42), supported by the ADMIXTURE output based on the smaller number of ForenSeq ancestry SNPs (African component = 0.43, European component = 0.57) and when using all ForenSeq markers (African component = 0.46, European component = 0.54). The ForenSeq plex does not include markers for skin colour, like the 15 markers in the HIrisPlex-S; this may indicate that the skin colour markers are important for eye colour and vice versa, impacting on the overall prediction of pigmentation (i.e. SNPs in *SLC45A2* are used for eye, hair and skin colour prediction). The individual with intermediate eye colour (Table 5.10) is from Fogo and shows a lower proportion of European ancestry (SNP chip - based African component = 0.58 and European component = 0.42; ForenSeq ancestry SNPs based African component = 0.52, and European component = 0.49; ForenSeq SNPs based African component = 0.47 and European component = 0.53). More considerations on link between phenotype and admixed ancestry are given in the Discussion section.

**Table 5.9 Information on eye colour prediction for the outlier sample (CVSA34).**

From the UAS phenotypic prediction (Figure 5.14), this sample appears to be an outlier. High T-index indicates blue pigment. High prediction values for the pigmentation category are in bold. na = not applicable.

Tool	Eye colour			Picture
	blue	intermediate	brown	
UAS output	0.38	0.10	<b>0.51</b>	
HIrisPlex-S output	0.33774438	0.11480387	<b>0.54745174</b>	
Real value (T-Index)	0.33097935	na	na	

**Table 5.10 Information on eye colour prediction for the sample with intermediate eye colour (CVF323).**

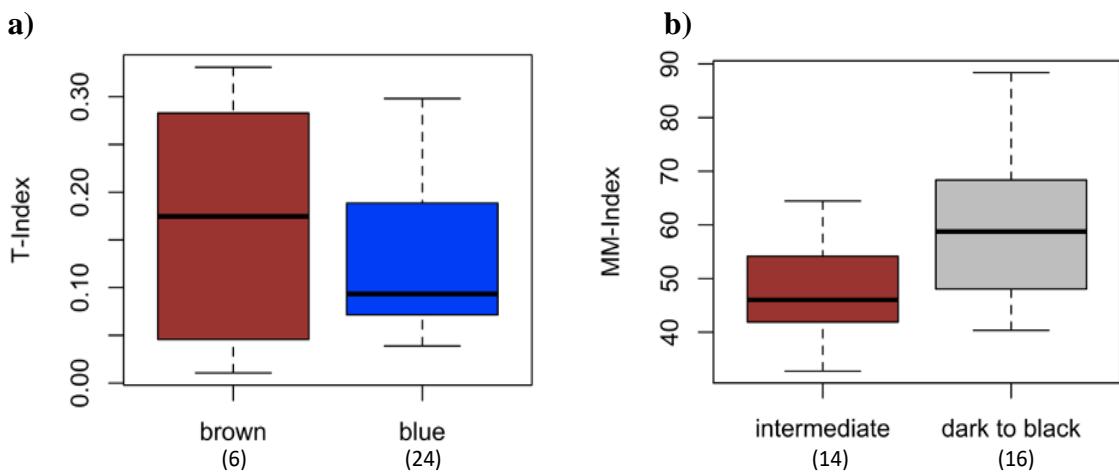
From the UAS phenotypic prediction (Figure 5.14), this is the only sample with intermediate colour predicted. High T-index indicates blue pigment. High prediction values for the pigmentation category are in bold. na = not applicable.

Tool	Eye colour			Picture
	blue	intermediate	brown	
UAS output	0.04	<b>0.85</b>	0.11	
HIrisPlex-S output	<b>0.73059021</b>	0.1359141	0.13349569	
Real value (T-Index)	0.27457122	na	na	

### 5.3.5 HIrisPlex-S web-tool phenotypic prediction

Initially, phenotypic prediction output from the HIrisPlex-S did not fully match the UAS output: intermediate colours for the eye are not predicted (Figure 5.15a). It is known that the intermediate category has the lowest level of accuracy in prediction, as the genes/SNPs associated with this trait are not believed to have been found yet (Walsh and Kayser 2016) or because of the absence of expressed eumelanin at different quantities within irises that is not easily captured by the IrisPlex SNPs alone (Walsh et al. 2014). As expected, the output for skin colour (based on a limited number of markers) does not predict other colour ranges (i.e. intermediate). According to the manual, missing markers *SLC24A5* rs1426654 and *OCA2* rs1800414/*MCIR* rs3212355 have a severe impact on identifying intermediate colour from dark/ dark-to-black colour (especially with samples of Asian ancestry due to proposed convergent evolution) (Endicott 2013; Norton et al. 2007).

The individuals with no UAS phenotypic prediction were subsequently excluded from HIrisPlex-S analysis (data not shown), but no changes were seen in the distribution. Also, two underperforming ancestry SNPs were excluded (data not shown), with no significant impact on eye colour prediction; however, some changes can be seen in skin colour prediction (overview given in Table 5.11).

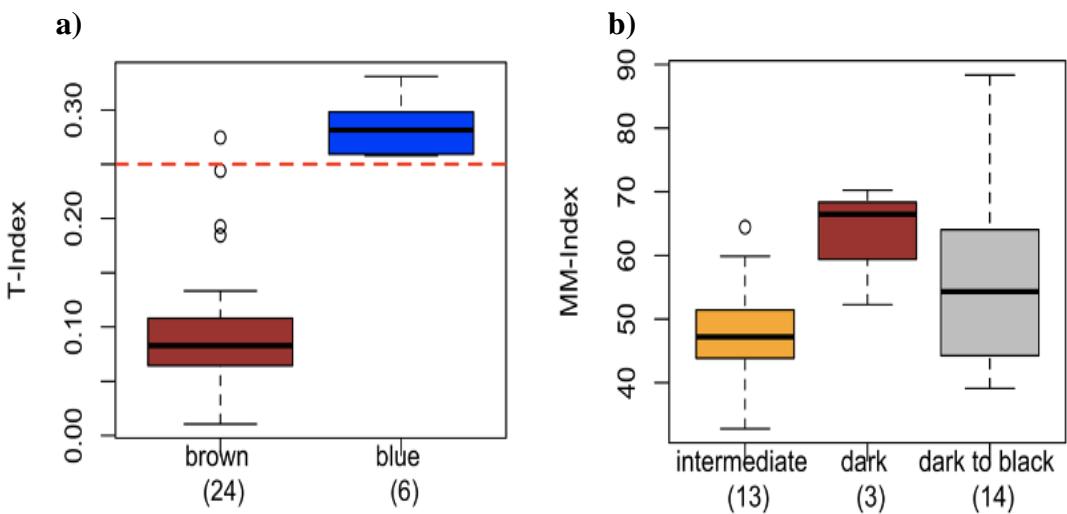


**Figure 5.15 The qualitative classification of the phenotype assigned by the HirisPlex model for 30 Cape Verdean individuals compared to the real phenotype.**

(a) the cut-off between brown and blue is set at 0.25 (T-Index); blue = 24 (of which 4 are truly blue), brown = 6 (of which 3 are truly brown); (b) higher values of modified melanin index (MM\_Index) correspond to darker skin tones, intermediate = 14, dark to black = 16.

Assessing further the differences in the basis of the two models, it appears that the UAS predictive tool is based on an older version of the HIrisPlex-S model (Schneider et al. 2019), where different alleles are specified as input for the model, leading to a better performance (Figure 5.16). The original model (IrisPlex, Liu et al. 2009, based on an interactive document) is based on different minor alleles for marker entries compared to the later HIrisPlex-S (web-tool). Using the old minor alleles legend for eye colour, the result improved: the only error in prediction correspond to one sample whose real eye colour is blue, but the IrisPlex estimates it as brown and the UAS as intermediate. The six SNPs for eye colour and the genotype responsible for each pigmentation prediction are reported in Figure 5.17, showing world frequency distributions and the Cape Verdean samples' genotypes.

Then, this older version of the minor alleles set was used in estimating hair colour and skin colour: there is improvement in hair colour prediction, and it is also possible to predict the dark colour (which is the colour between intermediate and black) for skin colour prediction.

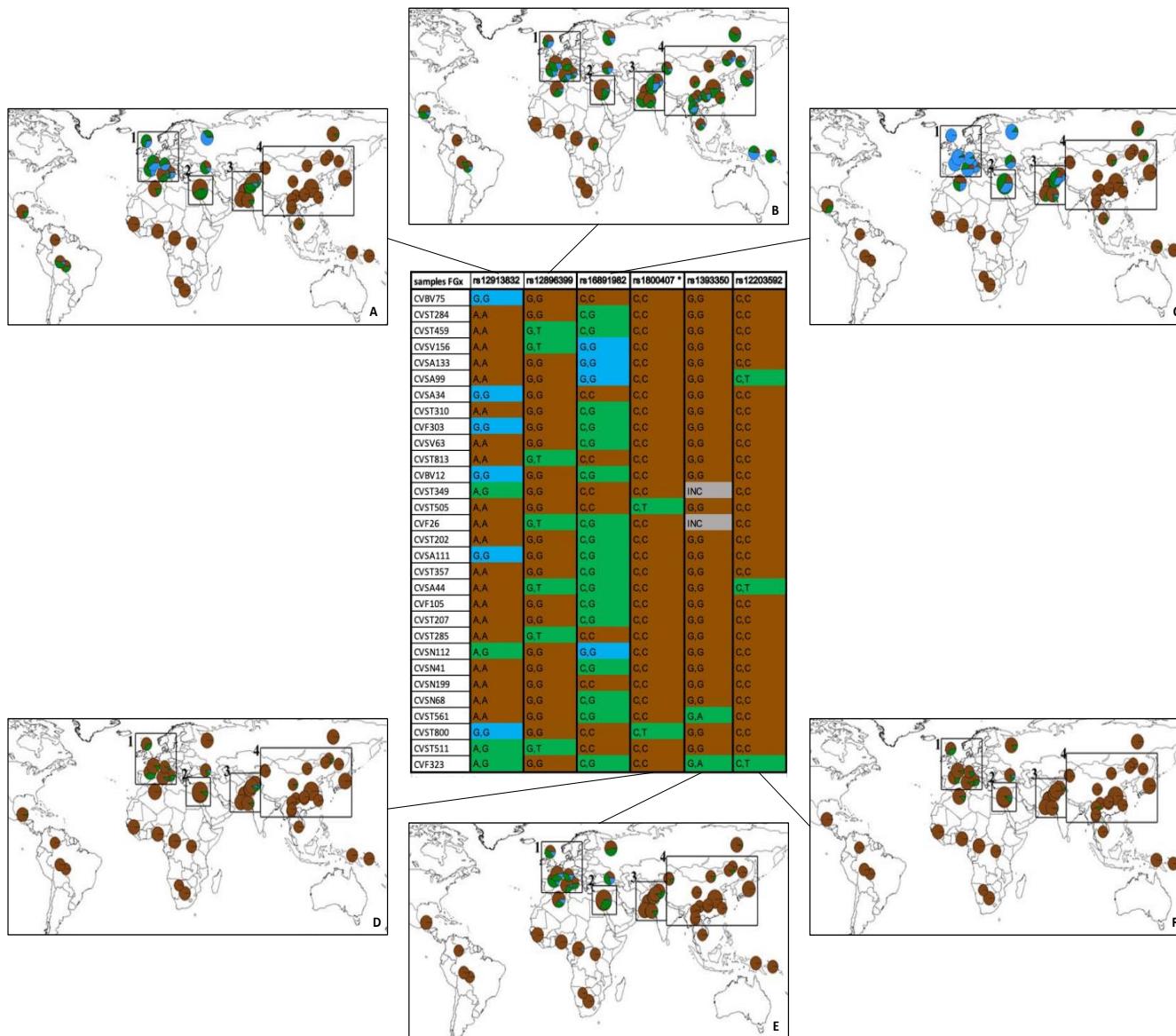


**Figure 5.16** The qualitative classification of the phenotype assigned by the HIrisPlex model was compared to the real phenotype using an older version of the model.

(a) the cut-off between brown and blue is set at 0.25 (T-Index), (b) higher values of modified melanin index (MM\_Index) correspond to darker skin tones (Beleza et al. 2013). Inconsistencies in the predictive output may be linked to several factors: the predictive model is based on European markers and validated using a world-population training set with no known phenotypes.

**Table 5.11** Number of samples that have pigment predicted for eye and skin colour based on UAS and HIrisPlex-S. na: not available.

Platform	Eye			Skin	
	Blue	Intermediate	Brown	Intermediate	Dark to Black
UAS	4	1	20	na	na
HIrisPlex-S	24 (including 4 with missing UAS profile)	na	6 (including 1 with missing UAS profile)	14	16
HIrisPlex-S without 2 aSNPs	24 (including 4 with missing UAS profile)	na	6 (including 1 with missing UAS profile)	23	7



**Figure 5.17 Genotypes of the 30 Cape Verdean samples at the 6 loci for eye colour prediction in the HIrisPlex-S with global map of frequency of the respective pigmentation.**

The 6 loci are: (A) rs12913832, (B) rs1800407, (C) rs12896399, (D) rs16891982, (E) rs1393350 and (F) rs12203592. The genotypes at these loci are associated to a colour for the eye, according to the HIrisPlex-S model (e.g. at rs12913832, the genotype GG corresponds to blue eye pigmentation). Adapted from: Walsh et al. 2011.

Further considerations must be given to:

- the underlying genetic architecture of this admixed population, which is very different from the (European) population used as training set for these models (which may explain the inaccuracy of some categories and especially of the skin colour prediction, as a limited number of SNPs is available in the ForenSeq kit);
- the possibility that small SNP panels used by these prediction models may not be enough to ensure accuracy and low false positive rates, negatively affecting, for example, the prediction of intermediate colours;
- the value of phenotypic prediction as an investigative lead.

These considerations are based on the analysis of the European-African admixed population of Cape Verde. To ask about performance in non-admixed populations, the HIrisplex-S was used on the 1KGP populations of Finnish (FIN, n = 99) and Yoruba (YRI, n = 108), using 23 phenotypic SNPs. One SNP for hair colour is missing (rs312262906), and 17 SNPs are missing for skin colour (rs3114908, rs1800414, rs10756819, rs2238289, rs17128291, rs6497292, rs1129038, rs1667394, rs1126809, rs1470608, rs1426654, rs6119471, rs1545397, rs6059655, rs12441727, rs3212355, rs8051733).

All YRI samples were predicted with brown eyes, (dark) black hair and dark-to-black skin colour as expected. Among the FIN samples, 18 are predicted to have brown eyes while the rest have blue eyes. For hair and skin colour there is a need to consider the pale/dark category (for hair) and the second category with highest probability score (according to the manual), better including nuances. The blond (43 blond, 34 blond/dark blond) and brown (1 brown, 8 brown/dark brown, 10 dark brown/ black) colours are the most common, with only 3 samples with red hair; 9 samples have low probability values with both brown and blond showing similar values. Intermediate colour was the most common: 56 intermediate colour (of which 8 have values very close to the “pale” category predicted value), 39 pale-intermediate and 4 pale. Essentially, then, the predictions based on the ForenSeq pSNPs perform as expected in these two non-admixed populations, which underlines the issues associated with accurate prediction in an admixed sample.

**Table 5.12 HIrisPlex-S results on two non-admixed populations from 1KGP based on 23 ForenSeq phenotypic SNPs.**

YRI = Yoruba (African), FIN = Finnish (European). The pale-intermediate category (a lighter intermediate colour) is assigned here: the first and second colour with highest probability values have both a strong impact on the predicted pigmentation according to the manual (<https://hirisplex.erasmusmc.nl/pdf/hirisplex.erasmusmc.nl.pdf>); this may be highlighted in a different way in Chaitanya et al. 2018.

<b>Category</b>	<b>Pigment</b>	<b>Populations</b>	
		<b>YRI (n = 108)</b>	<b>FIN (n = 99)</b>
eye colour	blue	0	81
	brown	108	18
hair colour	blond	0	43
	red	0	3
	blond/dark blond	0	34
	brown	0	1
	brown/dark brown	0	8
	dark brown/ black	0	10
	black	108	0
skin colour	pale	0	4
	pale-intermediate	0	39
	intermediate	0	56
	dark-to-black	108	na

## **5.4 Discussion**

This Chapter offers an evaluation of current methods used for biogeographical ancestry and phenotypic deduction in forensic applications. Considering realistic scenarios, the main focus is on an admixed population and the additional complexity this brings to the analysis. The ForenSeq kit provides both BGA and phenotypic prediction, and this Chapter demonstrates what caveats need to be considered and what alternative methods can be used.

### **5.4.1 Utility of ForenSeq and the UAS in evaluating ancestry**

The ForenSeq kit and the associated UAS functions within a limited framework of pre-selected reference populations that may not be suitable for some investigative scenarios. Published data suggest that UAS predictions are unlikely to be suitable for populations not already included in the reference set (England and Harbison 2019), such as Polynesian (England and Harbison 2019), Malaysian groups (Ramani et al. 2017); Middle Eastern (Almohammed et al. 2017); Native American (Wendt et al. 2016) as well as South Asian, South American and North African populations (Hussing et al. 2015). This raises the question of whether adding groups or ancestry informative SNPs ascertained in more diverse populations to the current UAS reference dataset may help (England and Harbison 2019).

The UAS displays the results of aSNP analysis as a PCA plot. Work in this chapter indicates that alternative approaches (such as ADMIXTURE) may offer more relevant information, considering the ancestral proportions instead of cluster affinity, which asks the user to make a more arbitrary interpretation. The results (Section 5.3.3) show increased precision when using ancestry markers and the full set of available markers (ancestry, identity and phenotypic SNPs). There is no particular advisable number of markers needed in ADMIXTURE, this depends on the genetic differentiation of the analysed populations (and inversely proportional to the genetic distance [Fst] among the populations (Alexander et al. 2009; Patterson et al. 2006)). Alternative marker panels (Table 5.1) may also be considered for improving ancestry determination (it is important to highlight that none of these panels include STRs and EVC markers, England and Harbison 2019).

As part of this work, STRUCTURE (Pritchard et al. 2000), a model-based technique that divides samples into clusters based on genetic similarity, to determine ancestry components/proportions, and SNIPPER (<http://mathgene.usc.es/snipper/index.php>, C. Phillips 2007), a web portal for rapid ancestry assignment via likelihood calculations, were used, but the results are not shown here. Results from STRUCTURE do not differ substantially from those of ADMIXTURE, and SNIPPER performs generally poorly with admixed populations and fails to offer a meaningful value for admixture proportions.

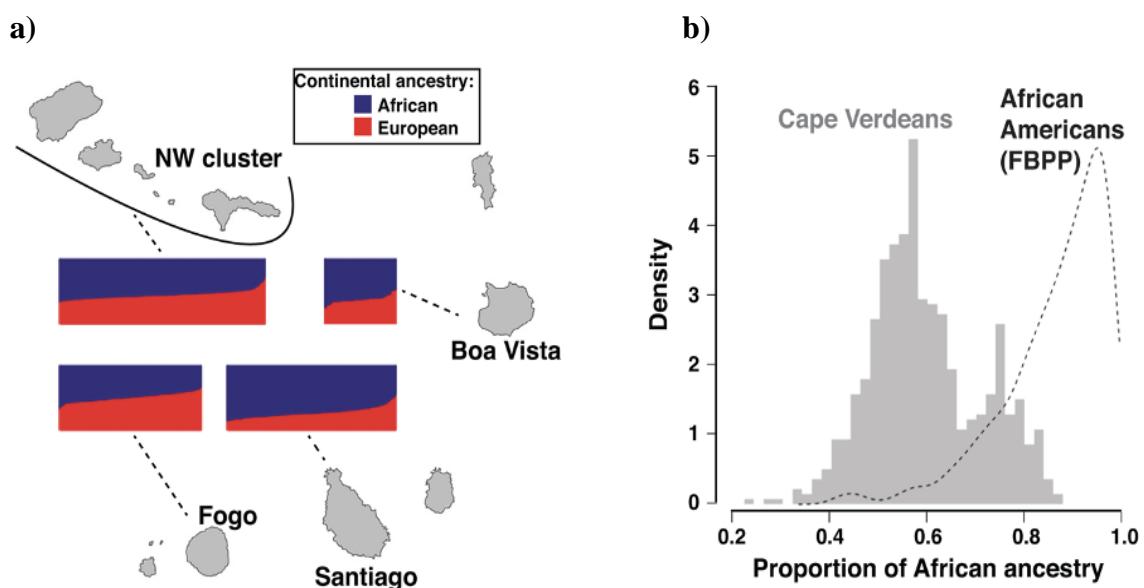
### **5.4.2 Utility of ForenSeq and the UAS in evaluating pigmentation phenotypes**

In this Chapter, the UAS prediction of phenotypes was compared with the publicly available and widely used HIrisPlex-S web tool. This indicated that the UAS is based on an older version of HIrisPlex-S. For example, some colour categories for hair are not yet introduced in the UAS (e.g. light hair, dark hair). The same SNPs are considered in both Figure 5.14 (UAS based) and Figure 5.15 (HIrisPlex-S web tool based), however, a different result is obtained, apparently due to differences in the way that alleles are defined. This could be misleading if operators decide to use HIrisPlex-S without knowing this. The UAS wrongly assigned the eye colour to one sample (brown instead of blue, considering that, overall, three samples have no phenotypic report due to missing SNPs), compared to the HIrisPlex-S that wrongly assigned the blue category to 20 samples (including two samples with no phenotypic report) and identifying correctly only three samples with brown category and four samples out of seven for the blue category.

Intermediate pigmentation is known to be hard to predict (Liu et al. 2009; Walsh 2013). This was also found also by Sharma and co-workers (Sharma et al. 2019), in a study based on 266 individuals, including mixed populations of Hispanic origins, with qualitative description of phenotypes and self-reported ancestry. Issues were encountered in predicting intermediate eye colour (76 individuals), and ancestry of individuals from South Asia and of admixed origins. They also found rs1393350 and rs12913832 (which has an impact on eye, hair colour prediction as well as ancestry prediction) to be consistently underperforming. They conclude by suggesting the addition of rs6119471 for better eye colour prediction.

### 5.4.3 Consideration on ancestral admixture and phenotype

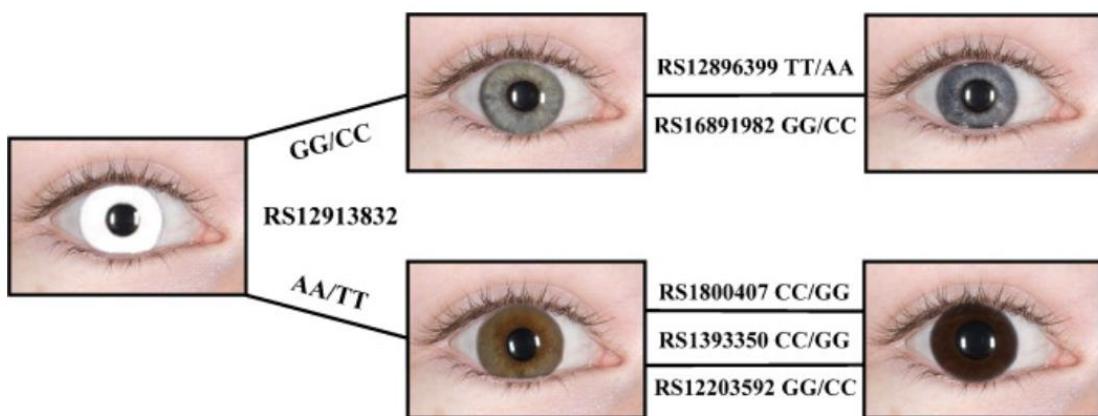
The effects of the loci on phenotype may vary among admixed and non-admixed populations. Beleza et al. (2013) highlights that the Cape Verde population shows a different genetic architecture underlying the eye - skin colour correlation compared to other populations. It appears that individual genomic ancestry has little effect on eye pigmentation (with *HERC2* [*OCA2*] and *SLC24A5* having the major impact), whereas it has a strong effect on skin pigmentation, even more significant than all skin colour loci combined (*APBA2* [*OCA2*], *GRM5-TYR*, and *SLC24A5*). Previous analysis of admixture proportions on these samples (Beleza et al. 2013) based on genome-wide data and utilising the tool frappe (Tang et al. 2005) showed an African component range of 23.5-87.9% (median of 58%), with significant differences among islands (Figure 5.18). Possible explanations for this general ancestry effect may be linked to non-genetic factors or to the influence of many genes with small effects. This needs consideration in phenotype prediction, as dense SNP data may be more informative than limited sets of targeted SNPs (Beleza et al. 2013).



**Figure 5.18 Ancestry proportions of Cape Verdean samples obtained using frappe (supervised analysis, K=2, with HapMap CEU and YRI as reference populations):**  
 (a) bar plots of ancestral proportions across islands (population total of 520 individuals); (b) plot comparing the distribution of 685 Cape Verdean samples' African ancestry against the kernel density curve of African ancestry in 802 African American samples (Family Blood Pressure Program). Source: Beleza et al. 2013.

It is also important to consider that the choice of markers in the phenotypic prediction plex is based on research undertaken in European populations. Combining individuals with different ancestry to identify pigmentation predictive SNPs is not advised (effects of BGA cannot be distinguished between those of phenotype) (Valenzuela et al. 2010), and in practice methods have been developed from European data and then validated with non-European samples (Walsh et al. 2012).

Models for eye colour prediction (e.g. HIrisPlex-S) are heavily based on the genotype at rs12913832 (*OCA2*), which is considered to be the marker with highest impact, determining the differentiation between brown pigmentation (rs12913832:AA or rs12913832:GA, ancestral allele being A, for dark pigmentation) and blue (rs12913832:GG). Figure 5.19 shows how the decision-making process of the model proceeds in assigning the eye colour based on the genotype at rs12913832. However, this needs to be interpreted with caution: some individuals do not follow this rule, as it was found that rs12913832:GA may unexpectedly result in blue or intermediate pigmentation, and rs12913832:GG can be linked to brown eyes contrary to the general expectations (Andersen et al. 2013 ; Andersen et al. 2016). Andersen and co-workers (2016) suggest that this behaviour may be explained by rs74653330:T (p.Ala481Thr) or rs121918166:G (p.Val443Ile), which have an effect not only on rs12913832:A and its consequent phenotype, but also on skin pigmentation in North Europeans.



**Figure 5.19 Workflow for determining eye pigmentation based on the 6 IrisPlex SNPs.**

In this model, rs12913832 has the biggest impact and the alleles at this locus determine the distinction between blue (GG/CC) or brown (AA/TT) eye colour. The next steps better determine the light or dark pigmentation. Source: Walsh et al. 2011.

Ancestry and phenotype show complex associations in admixed populations. (Kim et al. 2021) explores different mating models to explain how and at which speed the connection between traits and ancestry is eliminated through generations. The random mating model shows that the correlation decouples after the first generation and is below 0.5 in 6 generations (near zero after 20 generations); this process is longer with positive assortative mating by phenotype (below 0.5 in 11 generations) and by admixture (10 generations). Notably, the assortative mating model shows higher probability of mating within source populations than the random mating model (therefore retaining higher correlations for more generations). Also, it is important to consider the number of traits under examination (the more the traits, the more similar the pattern for all models) and the frequency of the traits (without fixed differences, the models again show similar patterns). The authors also consider a less stringent assumption on mate recognition and a model with single admixture event, and highlight that they assumed no sex bias, no stochasticity during genetic transmission nor fitness effects. Dominance, epistasis, environmental effects, and heritability were not included. This study concludes that traits such as pigmentation may say very little regarding the genetic ancestry of individuals (Kim et al. 2021).

# Chapter 6: Discussion

## 6.1 Summary of results

This project explored the use of SNP chip and targeted massively parallel sequencing data in forensic and genealogical applications, focusing on indirect information that can be obtained, such as kinship information, biogeographical ancestry and externally visible characteristics (EVCs), which might have intelligence value. The possibilities with current data and techniques were evaluated, offering a possible pipeline and highlighting the aspects that need to be considered at each step and how to integrate them in the analysis.

In Chapter 2, SNP chip data generated in eight families (65 individuals) of North European origin (Germany) were analysed in order to evaluate the reconstruction of family relationships. The data were processed for quality, and possible population substructure was assessed. This confirmed the overall homogeneity of the data, while also revealing a possible sample outlier, and showing the impact of family clustering in PCA analysis. This highlights the necessity of considering the appropriateness of reference allele frequencies used for subsequent analyses and their impact. The methods chosen for estimating kinship were based on IBD (calculating the probability of sharing zero, one or two alleles IBD) and network analysis to reconstruct pedigree structures. The family data allowed the exploration of various relationships, and the density of the SNP data offered the opportunity of considering distant relationships (most distant being second cousin once removed), however, the more distant ones (from first cousin once removed and beyond) were scarcely distinguishable from unrelated pairs. The pedigree reconstruction method offered an informative insight into the structure of the family trees, albeit affected by missing family members relevant for the tree structure; these cases may still be considered using the IBD approach. The possibility of using X-chromosome data through exploring IBD segment sharing, and Y- and mtDNA haplogroups, can also help in resolving more complex questions (e.g. relationships that share the same IBD coefficients such as half-siblings, grandparental and avuncular relationships; distinguishing between paternal and maternal half-siblings). The impact of LD in the data was also discussed.

This Chapter offers a step-by-step guide on the use of dense SNP data for kinship determination, and it is the bases for evaluating the performance of real-world targeted MPS data.

In Chapter 3, the eight families (from Chapter 2) were sequenced using a forensic-focused MPS kit, the ForenSeq DNA Signature Prep Kit (Verogen, San Diego, CA), which includes a combination of STR and SNP markers. The sequence data were analysed through the UAS platform, Verogen's in-built "black box" system, as well as through STRait Razor v3 and an alternative MPS workflow based on standard variant calling pipelines. The non-UAS methods allowed the discovery of additional sequence variation in the STR arrays and additional variants in the SNP flanking regions (79 SNPs), of which 10 are not called by the UAS. The amplicons of rs2830795 and rs1109037 contained additional polymorphic sites, forming two microhaplotypes. It was possible to explore how this set of markers performs in kinship estimation analyses, using the HGDP panel (European metapopulation) as allele-frequency reference. Combining all available markers helped in reducing the variation of the estimated IBD coefficients (i.e. estimated values are closer to the expected, theoretical coefficients). Simulated data offered an overview of the theoretical outcomes of kinship analysis under controlled conditions and was used to check the performance of the real-world data. This Chapter explored the possible use of targeted MPS data (specifically, the ForenSeq kit) for IBD-based kinsjip analyses.

In Chapter 4 (Kjelgaard Brustad et al. 2021), the focus was on existing forensic applications of kinship testing and familial searching, specifically on searches that involve a large number of individuals and comparisons ("blind search"). A way to assess performance based on Family Wise Error Rate and false positive rate was presented, considering the issues and inter-dependence of multiple testing. This Chapter proposes an implementation (R package available): this involves a Bayesian and parametric approach to Likelihood Ratio calculations, offering increased flexibility in specifying queried relationships for both outbred families and families with inbred founders and additional information through the use of prior data. Both real-world data (from Chapter 3) and simulated data were used.

In Chapter 5, a set of 30 selected individuals of European-African admixed ancestral origins (Cape Verde) was sequenced using the ForenSeq DNA Signature Prep Kit (Verogen, San Diego, CA). The biogeographical ancestry and phenotypic predictive capacity of the UAS was explored, considering its inability to provide an output if one or more markers are untyped. An alternative approach to ancestry prediction was evaluated, comparing the model-based method ADMIXTURE with the traditionally used PCA. The possibility of investigating ancestral components for each individual constitutes a more powerful tool than that currently offered by the UAS. Also, combining all available markers and not only the 56 ancestry SNPs may offer some advantages in better defining ancestral proportions. The phenotypes considered in this Chapter were limited to eye and skin colour. It is important to note that the UAS currently offers no prediction for skin pigmentation, but the widely used web-tool HIrisPlex-S would provide one, alongside eye and hair colour, based on the available markers. Skin and eye colour predictions were therefore considered against the measured phenotypes in the studied individuals. The output brought to attention the genetic complexity of the admixed population and the limits of the predictive tools, usually based on populations of European origins.

### **6.1.1 Limitations, caveats, and future work**

The family data and samples used in Chapters 2 and 3 were collected and provided by collaborators. In retrospect, a different sample collection approach that included a greater number of founders for the family data would have been advantageous: the number of unrelated individuals was limited in this study, excluding the possibility of exploring further population statistics, and considering allele frequencies generated from the dataset itself. The absence of many founders also shaped the type of simulations adopted, excluding conditional simulations based on the real pedigrees. Additional founders would have also reduced the overall sample missingness, simplifying the pedigree reconstruction process. Differently constituted family trees may have also allowed the further evaluation of the use of sex-linked markers, which were not fundamental in solving the available relationships, but, nonetheless posed the question of how to integrate them in the analyses.

Similarly, a more thorough study considering biogeographical ancestry and phenotypes could be performed by increasing the sample size of the admixed dataset and by random

selection from the Cape Verde population sample (Beleza et al. 2013): in Chapter 5 the samples were selected to span the range of ancestry and phenotypic proportions. The current selected sample set was unsuitable for population genetic analysis, which is nonetheless worthwhile.

In the context of kinship determination using SNP chip data, different techniques based on chromosome segment sharing could have been used, posing different questions more closely aligned to the framework of genetic genealogy. However, this would not have allowed ready comparisons with the targeted MPS data, which was a major focus of this project.

A further step in the kinship estimation process would be to apply the proposed workflow from non-admixed family samples to admixed ones. A wider variety of pedigree structure should also be used for better investigating the use of X- and Y-chromosome markers and mtDNA. Even if limited in number, the ForenSeq X- and Y-STRs should be further investigated, considering the limits of linkage groups and linkage disequilibrium (for the X), which are not well explored and cannot be optimally handled by current tools, but may be initially assessed using tools like Merlin (Abecasis et al. 2002a). Alternative analyses that integrate machine learning classification approaches may be performed. For example, (Smith et al. 2021) proposed a method based on a feature selection measure (mutual information, MI) and machine learning classification analysis to evaluate the possibility of combining both the estimation of the number of IBD segments and the pairwise coefficient of relatedness.

The biogeographical predictive procedure could include additional steps, such as the use of the Bayesian Information Criterion (BIC) to assess the best supported model, and the number and nature of clusters obtained. Alternatively, a probabilistic assignment of individuals to each group can be provided as in Bayesian clustering methods where the main limitations (i.e. numbers of variables or alleles must be less than the number of individuals, and LD is not handled) can be overcome by Discriminant Analysis of Principal Components (DAPC, Jombart et al. 2010).

The project considered the impact of linkage disequilibrium (LD) between markers to some extent. Some work has recently been carried out in this area (Li et al. 2019; Mo et

al. 2018), but many questions, such as how to efficiently calculate and incorporate linkage and mutation in the analysis and in IBD estimates, are still left unanswered.

## 6.2 Future directions: genetic genealogy

DTC testing companies began offering options to find the relatives of customers in 2009, with a product from 23andMe (Kling et al. 2021). This feature has subsequently become standard to all major companies: FamilyTreeDNA (FTDNA) in 2010, AncestryDNA in 2012, MyHeritage DNA in 2016, and LivingDNA in 2018. In exploring kinship estimation from genome-wide data, this project should offer an opportunity to reflect on some aspects of family finding approaches and the claims of DTC companies. However, a major difficulty with assessing the veracity of commercial claims is that the methods employed are not generally transparent or peer reviewed. AncestryDNA is alone in publishing a white paper on the subject (<https://support.ancestry.com/s/article/AncestryDNA-White-Papers>); recently, more information on the algorithm used by MyHeritage (which builds on some aspects of the AncestryDNA model) has become available (Petter et al. 2020), although it omits key details (e.g. the parameters used). There are no data available (to the author's knowledge) on the false positive and false negative rates of family finding searches in commercial databases. In this project, analyses based on IBD estimations showed that the sensitivity to reference allele frequencies may be an issue. This may have consequences in commercial testing, for example, considering the impact on segment sharing, if reference data are inappropriate to a tested sample.

DTC genetic data are playing an increasing role in some criminal justice systems (particularly that of the USA) through the burgeoning field of investigative genetic genealogy. Of the large providers, FTDNA allows law-enforcement searches (limited to customers who agreed to this service and “opted-in”). Other searches are mostly carried out via the third-party database GEDmatch. While such searches can be successful, there are many associated ethical and legal problems (Syndercombe Court 2018). Genetic information on an individual is not limited to that individual alone, but includes information on their relatives (Clayton et al. 2019). De Groot and van Beers (2021)

highlight multiple levels of privacy concerns: the sensitivity of the information that can be obtained (in terms of family structures and medically relevant deductions), the security of the stored genetic data and its (unclear) regulation, customer consent (considering that the genetic information also involves the customer's distant genetic family), and the lack of official accreditation for genetic genealogists who access and use the data (de Groot and van Beers 2021). Data breaches have recently affected the main companies and GEDmatch, for example in July 2020, making opted-out and law enforcement profiles available to searches, and a technical issue in January 2021, causing deleted DNA profiles to be restored in GEDmatch (see Edge and Coop 2020; Kling et al. 2021). DTC companies may inform their customers regarding possible law enforcement access to their data, but the methods and conditions of use are not generally clear (de Groot and van Beers 2021).

The UK has been applying STR-based familial searching in serious and violent crimes since 2003 (BFEG 2020). Its pioneer role is partly due to the efficiency of its criminal DNA database (NDNAD): established in 1995, it led to successful individual identification in 731,160 unsolved crimes between April 2001 and March 2020, and supported 33 familial searches between 2018 and 2020. In the past decade, the reported identification success rate of such searching was around 20%. This figure may be improved by the use of investigative genetic genealogy (BFEG 2020) which can find more distant relatives, but apart from the practical issues of DNA quality and quantity, the application of this technique is still controversial and not promoted by the National Police Chiefs Council (Kling et al. 2021). The US legal system shows much more flexibility, as it saw the application of genetic genealogy in solving the first case of this type (see Chapter 1, section 1.5.3.3), and has applied it to around 200 law enforcement investigations (Samuel and Kennett 2020). In 2019, the Interim Policy on Forensic Genetic Genealogical DNA Analysis and Searching was released (US Department of Justice, <https://www.justice.gov/olp/page/file/1204386/download>), proposing the genetic genealogy approach only as a last resort, highlighting inefficiencies and a backlog issue (the “CODIS gap”) in the system in applying more standard techniques like familial searching. In Sweden, investigative genetic genealogy has been used to solve a double-murder cold case in 2020 (Tillmar et al. 2021), and its applications have been considered in other countries as well (for example, Canada, Netherlands and Philippines, Batha 2020; de Groot and van Beers 2021).

These differences in legislation between different countries highlight another important limit is the “transnational” aspect of genetic genealogy, as searches and family trees may include individuals from different countries under different legislations. Moreover, due to the composition of databases, suspects from under-represented parts of the world may not be identified. Although it may be difficult to find global guidelines due to differences in legislation and approaches, privacy and data security should be a priority (BFEG 2020), avoiding unnecessary breaches, and limiting the invasiveness of searches (especially in handling sensitive medical and personal data).

## **6.3 Future directions: MPS approaches**

As the potential and limits of investigative genetic genealogy are being explored, other tools are being developed as alternatives or adjuncts to the traditional STR profiling. This project investigated the use of SNP chip and MPS data, their expected results and limits. Successful SNP chip typing requires DNA of a quality and quantity that is not routinely found in forensic casework. Here, targeted MPS offers an alternative, and provides the necessary sensitivity, though not the large number of markers and the ability to assess segment sharing (as in DTC testing). This project suggested techniques based on IBD probabilities. One point to consider for future application is the absence of a set of robust and diverse allele frequency reference datasets for the MPS. As well as targeted analysis, MPS can be used for whole genome sequencing (WGS) which, as the ancient DNA field demonstrates, can be accomplished with small amounts of damaged DNA. Once WGS data are available, specific SNPs can be extracted and used to query databases such as GEDmatch (Kling et al. 2021; Tillmar et al. 2020). This approach was used in the so-called Buckskin Girl identification case (Kennett 2019).

## **6.4 Future directions: new traits for externally visible characteristic predictions**

This project explored not only kinship relationships, but also the EVCs of eye and skin colour. Other phenotypic traits are receiving attention in forensic research: facial morphology, hair structure and thickness, balding, and freckling, to name a few.

Facial morphology is a genetically-linked, highly variable trait (Liu et al. 2021). There is still bias towards European populations in association studies, and studies on African populations are limited (Cole et al. 2016): 203 signals across 138 genetic loci were found through GWAS in Europeans (White et al. 2020), of which only eight (HOXD cluster, *PAX3*, *TBX3*, *SOX9*, *PAX1*, 4q31.3, 6p21.1, 20q12) were found across all populations. Problems with current studies are the lack of homogeneity in the approaches and techniques used, leading to little overlap in results among them (Liu et al. 2021). This may also be due to possible insufficient power and the effects of the populations' genetic architectures (Liu et al. 2021). Its use in forensics raises privacy and ethical concerns similar to those for genetic genealogy, including the use of live facial recognition (LFR) technology by private and public agencies, and consequent possible unlawful or discriminating targeting based on "protected characteristics" (Equality Act 2010) (BFEG 2020).

Despite the work going on in these areas, the complexity of these traits and their underlying genetic architecture currently limits individual predictability and forensic utility (Schneider et al. 2019). Also, as shown in this project, the commonplace phenomenon of admixture is likely to make prediction even more difficult than it is within a homogeneous population.

# Bibliography

- Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002a) Merlin - rapid analysis of dense genetic maps using sparse gene flow trees. *Nature Genetics* 30: 97–101. doi: <https://doi.org/10.1038/ng786>
- Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002b) Merlin - rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* 30: 97–101. doi: <https://doi.org/10.1038/ng786>
- Abecasis GR, Wigginton JE (2005) Handling marker-marker linkage disequilibrium: pedigree analysis with clustered markers. *Am J Hum Genet* 77: 754–67.
- Al-Khudhair A, Qiu S, Wyse M, Chowdhury S, Cheng X, Bekbolsynov D, Saha-Mandal A, Dutta R, Fedorova L, Fedorov A (2015) Inference of Distant Genetic Relations in Humans Using “1000 Genomes”. *Genome Biol Evol*. 7: 481-492.
- Albrechtsen A, Korneliussen TS, Moltke I, van Overeem Hansen T, Nielsen FC, Nielsen R (2008) Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genetic Epidemiology* 33. doi: <https://doi.org/10.1002/gepi.20378>
- Albrechtsen A, Sand Korneliussen TS, Moltke I, van Overeem Hansen T, Nielsen FC, Nielsen R (2009) Relatedness mapping and tracts of relatedness for genome-wide data in the presence of linkage disequilibrium. *Genetic Epidemiology: The Official Publication of the International Genetic Epidemiology Society* 33: 266–274. doi: <https://doi.org/10.1002/gepi.20378>
- Alexander DH, Novembre J, Lange K (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*. 19: 1655-64. doi: doi: 10.1101/gr.094052.109
- Almalki N, Chow HY, Sharma V, Hart K, Siegel D, Wurmbach E (2017) Systematic assessment of the performance of Illumina’s MiSeq FGxTM forensic genomics system. *Electrophoresis* 38: 846–854. doi: 10.1002/elps.201600511
- Almohammed E, Iyengar A, Ballard D, Devesse L, Hadi S (2017) Evaluation of ForenSeq DNA signature kit for Qatari population. *Forensic Science International: Genetics Supplement Series* 6: e596–e598. doi: <https://doi.org/10.1016/j.fsigss.2017.10.003>
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J. Mol. Biol.* 215: 403–410.
- Amorim A, Pereira L (2005a) Pros and cons in the use of SNPs in forensic kinship investigation: a comparative analysis with STRs. *Forensic Science International* 150: 17–21. doi: <https://doi.org/10.1016/j.forsciint.2004.06.018>
- Amorim A, Pereira L (2005b) Pros and cons in the use of SNPs in forensic kinship investigation: a comparative analysis with STRs. *Forensic Science International* 150: 17–21. doi: doi:10.1016/j.forsciint.2004.06.018
- Andersen JD, Johansen P, Harder S, Christoffersen SR, Delgado MC, Henriksen ST, Nielsen MM, Sørensen E, Ullum H, Hansen T, Dahl AL, Paulsen RR, Børsting C, Morling N (2013 ) Genetic analyses of the human eye colours using a novel objective method for eye colour classification. *Forensic Sci. Int. Genet.* 7: 508–515. doi: 10.1016/j.fsigen.2013.05.003.
- Andersen JD, Pietroni C, Johansen P, Andersen MM, Pereira V, Børsting C, Morling N (2016) Importance of nonsynonymous OCA2 variants in human eye color prediction. *Mol Genet Genomic Med.* 4: 420–430. doi: 10.1002/mgg3.213

- Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT (2010) Data quality control in genetic case-control association studies. *Nat Protoc.* 5: 1564–1573. doi: doi: 10.1038/nprot.2010.116
- Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJ, Staden R, Young IG (1981) Sequence and organization of the human mitochondrial genome. *Nature* 290: 457-65. doi: doi: 10.1038/290457a0
- Andrews RM, Kubacka I, Chinnery PF, Lightowers RN, Turnbull DM, Howell N (1999) Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. *Nat Genet* 2. doi: doi: 10.1038/13779
- Astle W, Balding DJ (2009) Population Structure and Cryptic Relatedness in Genetic Association Studies. *Statistical Science* 24: 451–471. doi: DOI: 10.1214/09-STS307
- Balaresque P, Bowden GR, Parkin EJ, Omran GA, Heyer E, Quintana-Murci L, Roewer L, Stoneking M, Nasidze I, Carvalho-Silva DR, Tyler-Smith C, de Knijff P, Jobling MA (2008) Dynamic nature of the proximal AZFc region of the human Y chromosome: multiple independent deletion and duplication events revealed by microsatellite analysis. *Hum Mutat* 29: 1171-1180. doi: 10.1002/humu.20757
- Balaresque P, King TE, Parkin EJ, Heyer E, Carvalho-Silva D, Kraaijenbrink T, de Knijff P, Tyler-Smith C, Jobling MA (2014) Gene conversion violates the stepwise mutation model for microsatellites in Y-chromosomal palindromic repeats. *Hum Mutat* 35: 609-17. doi: 10.1002/humu.22542
- Ballard D, Winkler-Galicki J, Wesoły J (2020) Massive parallel sequencing in forensics: advantages, issues, technicalities, and prospects. *International Journal of Legal Medicine* 134: 1291–1303. doi: <https://doi.org/10.1007/s00414-020-02294-0>
- Barbujani G, Magagni A, Minch E, Cavalli-Sforza LL (1997) An apportionment of human DNA diversity. *Proceedings of the National Academy of Sciences* 94: 4516-4519. doi: <https://doi.org/10.1073/pnas.94.9.4516>
- Batha E (2020) Could genealogy websites help catch aid worker sex abusers? *Reuters* <https://news.trust.org/item/20200716031354-blleb/> [Accessed 13 Sept. 2021].
- Batini C, Hallast P, Vågene ÅJ, Zadik D, Eriksen HA, Pamjav H, Sajantila A, Wetton JH, Jobling MA (2017) Population resequencing of European mitochondrial genomes highlights sex-bias in Bronze Age demographic expansions. *Scientific Reports*. doi: 10.1038/s41598-017-11307-9
- Beleza S, Johnson NA, Candille SI, Absher DM, Coram MA, Lopes J, Campos J, Araújo II, Anderson TMV, B.J. Nordborg, M. , Correia e Silva AS, M.D. , Rocha J, Barsh GS, Tang H (2013) Genetic Architecture of Skin and Eye Color in an African-European Admixed Population. *PLOS Genetics* 9: e1003372. doi: <https://doi.org/10.1371/journal.pgen.1003372>
- Belle EM, Landry PA, Barbujani G (2006) Origins and evolution of the Europeans' genome: evidence from multiple microsatellite loci. *Proc. Biol. Sci.* 273: 1595–1602.
- Benjamini Y, Hochberg Y (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57: 289-300.
- Bertoglio B, Grignani P, Di Simone P, Polizzi N, De Angelis DC, C., Iadicicco A, Fattorini P, Presciuttini SP, C. (2020) Disaster victim identification by kinship analysis: the Lampedusa October 3rd, 2013 shipwreck. *Forensic Science International: Genetics* 44: 102156. doi: <https://doi.org/10.1016/j.fsigen.2019.102156>

- BFEG (2020) Should we be making use of genetic genealogy to assist in solving crime? A report on the feasibility of such methods in the UK.
- Biasutti R (1941) La razze e i popoli della Terra. Unione Tipografico-Editrice Torinese.
- Bittles AH, Black ML (2010) Evolution in health and medicine Sackler colloquium: Consanguinity, human evolution, and complex diseases. Proc Natl Acad Sci 107: 1779-1786. doi: <https://doi.org/10.1073/pnas.0906079106>
- Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: A flexible trimmer for Illumina Sequence Data. Bioinformatics: btu170.
- Bonilla C, Gutierrez G, Parra EJ, Kline C, Shriver MD (2005) Admixture analysis of a rural population of the state of Guerrero, Mexico. Am J Phys Anthropol 128: 861-869.
- Børsting C, Morling N (2011) Mutations and/or close relatives? Six case work examples where 49 autosomal SNPs were used as supplementary markers. Forensic Science International: Genetics 5: 236-241. doi: doi:10.1016/j.fsigen.2010.02.007
- Boyles AL, Scott WK, Martin ER, Schmidt S, Li YJ, Ashley-Koch A, Bass MP, Schmidt M, Pericak-Vance MA, Speer MC, Hauser ER (2005) Linkage disequilibrium inflates type I error rates in multipoint linkage analysis when parental genotypes are missing. Hum Hered 59: 220-227.
- Brown MD, Glazner CG, Zheng C, Thompson EA (2012) Inferring Coancestry in Population Samples in the Presence of Linkage Disequilibrium. Genetics 190: 1447-1460. doi: <https://doi.org/10.1534/genetics.111.137570>
- Brown WRA (1988) A physical map of the pseudoautosomal region. EMBO J 7: 2377-2385.
- Browning BL, Browning SR (2011a) A fast, powerful method for detecting identity by descent. The American Journal of Human Genetics 88: 173-182. doi: <https://doi.org/10.1016/j.ajhg.2011.01.010>
- Browning BL, Browning SR (2013) Improving the accuracy and efficiency of identity-by-descent detection in population data. Genetics 194: 459-471. doi: <https://doi.org/10.1534/genetics.113.150029>
- Browning BL, Browning SR (2016) Genotype Imputation with Millions of Reference Samples. Am J Hum Genet 98: 116-26. doi: doi: 10.1016/j.ajhg.2015.11.020
- Browning BL, Zhou Y, Browning SR (2018) A one-penny imputed genome from next generation reference panels. Am J Hum Genet 103: 338-348. doi: doi:10.1016/j.ajhg.2018.07.015
- Browning SR, Browning BL (2007) Rapid and accurate haplotype phasing and missing data inference for whole genome association studies by use of localized haplotype clustering. Am J Hum Genet 81: 1084-1097. doi: doi:10.1086/521987
- Browning SR, Browning BL (2010) High-resolution detection of identity by descent in unrelated individuals. Am J Hum Genet 86: 526-539.
- Browning SR, Browning BL (2011b) Haplotype phasing: existing methods and new developments. Nature Reviews Genetics 12: 703-714. doi: <https://doi.org/10.1038/nrg3054>
- Browning SR, Browning BL (2012) Identity by Descent Between Distant Relatives: Detection and Applications. Annu. Rev. Genet. 46: 617-33. doi: 10.1146/annurev-genet-110711-155534
- Browning SRB, B.L. (2012) Identity by Descent Between Distant Relatives: Detection and Applications. Annu. Rev. Genet. 46: 617-33. doi: 10.1146/annurev-genet-110711-155534
- Bruijns B, Tiggelaar RG, H. (2018) Massively parallel sequencing techniques for forensics: A review. Electrophoresis 39: 2642-2654.

- Budowle B, Planz JV, Campbell RS, Eisenberg AJ (2004) Single Nucleotide Polymorphisms and Microarray Technology in Forensic Genetics - Development and Application to Mitochondrial DNA. *Forensic Sci Rev.* 16: 21-36.
- Budowle B, Van Daal A (2008) Forensically relevant SNP classes. *BioTechniques* 44: 603-610.
- Budowle B, Van Daal A (2018) Extracting evidence from forensic DNA analyses: future molecular biology directions. *Biotechniques* 46.
- Bulbul O, Pakstis AJ, Soundararajan U, Gurkan C, Brissenden JE, Roscoe JM, Evsanaa B, Togtokh A, Paschou P, Grigorenko EL, Gurwitz D, Wootton S, Lagace R, Chang J, Speed WC, Kidd KK (2018) Ancestry inference of 96 population samples using microhaplotypes. *International Journal of Legal Medicine* 132: 703-711. doi: <https://doi.org/10.1007/s00414-017-1748-6>
- Butler JM (2005) *Forensic DNA Typing : Biology, Technology, and Genetics of STR Markers*, 2nd Edition. Elsevier Science & Technology.
- Butler JM (2006) Genetics and genomics of core short tandem repeat loci used in human identity testing. *J. Forensic Sci.* 51: 253-265. doi: 0.1111/j.1556-4029.2006.00046.x
- Butler JM (2012) Advanced Topics in Forensic DNA Typing: Methodology - Chapter 6. Capillary Electrophoresis: Principles and Instrumentation. Academic Press: 141-165. doi: <https://doi.org/10.1016/B978-0-12-374513-2.00006-3>
- Butler JM (2015) Chapter 10 - STR Population Data Analysis" from Advanced Topics in Forensic DNA Typing: Interpretation. Elsevier Inc.: 239-279.
- C. Phillips AS, J.J. Sánchez, M. Fondevila, A. Gómez-Tato, J. Álvarez-Dios, M. Calaza, M. Casares de Cal, D. Ballard, M.V. Lareu, A. Carracedo, The SNPforID Consortium (2007) Inferring ancestral origin using a single multiplex assay of ancestry-informative marker SNPs. *Forensic Science International: Genetics* 1: 273-280. doi: doi:10.1016/j.fsigen.2007.06.008
- Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, Chen Z, Chu J, Carcassi C, Contu L, Du R, Excoffier L, Ferrara GB, Friedlaender JS, Groot H, Gurwitz D, Jenkins T, Herrera RJ, Huang X, Kidd J, Kidd KK, Langaney A, Lin AA, Mehdi SQ, Parham P, Piazza A, Pistillo MP, Qian Y, Shu Q, Xu J, Zhu S, Weber JL, Greely HT, Feldman MW, Thomas G, Dausset J, Cavalli-Sforza LL (2002a) A human genome diversity cell line panel. *Science* 296: 261-262. doi: doi:10.1126/science.296.5566.261bpmid:11954565
- Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, Piouffre L, Bodmer J, Bodmer WF, Bonne-Tamir B, Cambon-Thomsen A, Chen Z, Chu J, Carcassi C, Contu L, Du R, Excoffier L, Ferrara GB, Friedlaender JS, Groot H, Gurwitz D, Jenkins T, Herrera RJ, Huang X, Kidd J, Kidd KK, Langaney A, Lin AA, Mehdi SQ, Parham P, Piazza A, Pistillo MP, Qian Y, Shu Q, Xu J, Zhu S, Weber JL, Greely HT, Feldman MW, Thomas G, Dausset J, Cavalli-Sforza LL (2002b) A human genome diversity cell line panel. *Science* 296: 261-262. doi: 10.1126/science.296.5566.261b
- Carroll D (2001) Genetic Recombination. *Encyclopedia of Genetics* - Academic Press: 841-845. doi: <https://doi.org/10.1006/rwgn.2001.0543>
- Chaitanya L, Breslin K, Zuñiga S, Wirken L, Pospiech E, Kukla-Bartoszek M, Sijen T, de Knijff P, Liu F, Branicki W, Kayser M, Walsh S (2018) The HIrisPlex-S system for eye, hair and skin colour prediction from DNA: T Introduction and forensic developmental validation. *Forensic Science International: Genetics* 35: 123-135.

- Chen YC, Pollack S, Hunter DJ, Hirschhorn JN, Kraft P, Price AL (2013) Improved ancestry inference using weights from external reference panels. *Bioinformatics* 29: 1399–1406.
- Cheung CYK (2013) Using Inheritance Vectors to Impute Genotypes and Detect Genotyping Errors. Doctoral Thesis, University of Washington.
- Cho S, Shin ES, Yu HJ, Lee JH, Seo HJ, Kim MY, Lee SD (2017) Set up of cutoff thresholds for kinship determination using SNP loci. *Forensic Science International: Genetics* 29: 1–8. doi: <http://dx.doi.org/10.1016/j.fsigen.2017.03.009>
- Churchill JD, Schmedes SE, King JL, Budowle B (2016) Evaluation of the Illumina® Beta Version ForenSeq™ DNA Signature Prep Kit for use in genetic profiling. *Forensic Sci Int Genet.* 20: 20–29. doi: doi: 10.1016/j.fsigen.2015.09.009
- Clayton EW, Evans BJ, Hazel JW, Rothstein MA (2019) The law of genetic privacy: applications, implications, and limitations. *J Law Biosci* 6: 1–36. doi: doi: 10.1093/jlb/lzv007
- Cole JB, Manyama M, Kimwaga E, Mathayo J, Larson JR, Liberton DK, Lukowiak K, Ferrara TM, Riccardi SL, Li M, Mio W, Prochazkova M, Williams T, Li H, Jones KL, Klein OD, Santorico SA, Hallgrímsson B, Spritz RA (2016) Genomewide Association Study of African Children Identifies Association of SCHIP1 and PDE8A with Facial Size and Shape. *PLoS genetics* 12: e1006174. doi: <https://doi.org/10.1371/journal.pgen.1006174>
- Coleman JRI, Euesden J, Patel H, Folarin AA, Newhouse S, Breen G (2016) Quality Control, Imputation and Analysis of Genome-Wide Genotyping Data From the Illumina HumanCoreExome Microarray. *Briefings in Functional Genomics* 15: 298–304. doi: <https://doi.org/10.1093/bfgp/elv037>
- Conomos MP, Gogarten SM, Brown L, Chen H, Rice K, Sofer T, Thornton T, Yu C (2018) GENESIS: GENetic ESTimation and Inference in Structured samples (GENESIS): Statistical methods for analyzing genetic data from samples with population structure and/or relatedness. R package version 2.12.0.
- Conomos MP, Miller M, Thornton T (2015a) Robust Inference of Population Structure for Ancestry Prediction and Correction of Stratification in the Presence of Relatedness. *Genet Epidemiol.* 39: 276–293. doi: 10.1002/gepi.21896
- Conomos MP, Miller MB, Thornton TA (2015b) Robust Inference of Population Structure for Ancestry Prediction and Correction of Stratification in the Presence of Relatedness. *Genetic Epidemiology* 39: 276–293. doi: doi:10.1002/gepi.21896.
- Conomos MP, Reiner AP, Weir BS, Thornton TA (2016) Model-free Estimation of Recent Genetic Relatedness. *American Journal of Human Genetics* 98: 127–148. doi: 10.1016/j.ajhg.2015.11.022
- Coon CS (1939) The races of Europe. New York, Macmillan Co.
- Cotterman C (1941) Relatives and human genetic analysis. *Sci Mon.* 53: 227–34.
- Cotterman CW (1940) A calculus for statistico-genetics. *Biology.*
- Cowell RG (2009) Efficient maximum likelihood pedigree reconstruction. *Theor. Popul. Biol.* 76: 285–91.
- Crawford NG, Kelly DE, Hansen MEB, Beltrame MH, Fan S, Bowman SL, Jewett E, Ranciaro A, Thompson S, Lo Y, Pfeifer SP, Jensen JD, Campbell MC, Beggs W, Hormozdiari F, Mpoloka SW, Mokone GG, Nyambo T, Meskel DW, Belay G, Haut J, NISC Comparative Sequencing Program, Rothschild H, Zon L, Zhou Y, Kovacs MA, Xu M, Zhang T, Bishop K, Sinclair J, Rivas C, Elliot E, Choi J, Li SA, Hicks B, Burgess S, Abnet C, Watkins-Chow DE, Oceana E, Song YS, Eskin E, Brown KM, Marks MS, Loftus SK, Pavan WJ, Yeager M, Chanock S, Tishkoff

- SA (2017) Loci associated with skin pigmentation identified in African populations. *Science* 358.
- Cussens J, Bartlett M, Jones EM, N.A. S (2013) Maximum Likelihood Pedigree Reconstruction Using Integer Linear Programming. *Genetic Epidemiology* 37: 69-83.
- Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, McVean G, Group GPA (2011) The variant call format and VCFtools. *Bioinformatics* 27: 2156-2158.
- De Andrade M, Ray D, Pereira AC, Soler JP (2015) Global individual ancestry using principal components for family data. *Hum Hered.* 80: 1–11.
- de Groot NF, van Beers BC, Meynen, G. (2021) Commercial DNA tests and police investigations: a broad bioethical perspective. *J Med Ethics*: 1–8. doi: doi:10.1136/medethics-2021-107568
- De la Cruz O, Raska P (2014) Population structure at different minor allele frequency levels. *BMC Proceedings* 8: S55. doi: 10.1186/1753-6561-8-S1-S55
- De la Puente M, Santos C, Fondevila M, Manzo L, Carracedo A, Lareu M, Phillips C (2016) The global AIMs Nano set: A 31-plex SNaPshot assay of ancestry-informative SNPs. *Forensic Science International: Genetics* 22: 81–88. doi: 10.1016/j.fsigen.2016.01.015
- Debus-Sherrill S, Field MB (2017) Understanding Familial DNA Searching: Policies, Procedures, and Potential Impact, Summary Overview. IFC.
- Debus-Sherrill S, Field MB (2019) Familial DNA searching- an emerging forensic investigative tool. *Science & Justice* 59: 20-28.
- Delest A, Godfrin D, Chantrel Y, Ulus A, Vannier J, Faivre M, Hollard C, Laurent FX (2020) Sequenced-based French population data from 169 unrelated individuals with Verogen's ForenSeq DNA signature prep kit. *Forensic Science International: Genetics* 47: 102304. doi: <https://doi.org/10.1016/j.fsigen.2020.102304>
- Della Rocca C, Alladio E, Barni F, Cannone F, D'Atanasio E, Trombetta B, Berti A, Cruciani F (2019) LOW DISCRIMINATION POWER OF THE YFILER™ PLUS PCR AMPLIFICATION KIT IN AFRICAN POPULATIONS. DO WE NEED MORE RM Y-STRs? *Forensic Science International: Genetics Supplement Series* 7: 671-673.
- Deng C, Song F, Li J, Ye Y, Zhang L, Liang W, Luo H, Li Y (2017) Forensic parameters of 19 X-STR polymorphisms in two Chinese populations. *International Journal of Legal Medicine* 131: 975–977. doi: <https://doi.org/10.1007/s00414-017-1538-1>
- Devesse L, Davenport L, Borsuk L, Gettings K, Mason-Buck G, Vallone PM, Syndercombe Court D, Ballard D (2020) Classification of STR allelic variation using massively parallel sequencing and assessment of flanking region power. *Forensic Science International: Genetics* 48: 102356. doi: <https://doi.org/10.1016/j.fsigen.2020.102356>
- Dib C, Faure S, Fizames C, Samson D, Drouot N, Vignal A, Millasseau P, Marc S, Hazan J, Seboun E, Lathrop M, Gyapay G, Morissette J, Weissenbach J (1996) A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature* 380: 152-4.
- Diegoli TM, Linacre A, Vallone PM, Butler JM, Coble MD (2011) Allele frequency distribution of twelve X-chromosomal short tandem repeat markers in four U.S. population groups. *Forensic Science International: Genetics Supplement Series* 3: e481-e483. doi: <https://doi.org/10.1016/j.fsigss.2011.09.102>

- Dixon LA, Murray CM, Archer EJ, Dobbins AE, Koumi P, Gill P (2005) Validation of a 21-locus autosomal SNP multiplex for forensic identification purposes. *Forensic Sci. Int. Genet.* 154: 62–77.
- Dou J, Sun B, Sim X, Hughes JD, Reilly DF, Tai ES, Liu J, Wang C (2017) Estimation of kinship coefficient in structured and admixed populations using sparse sequencing data. *PLoS Genet* 13: e1007021. doi: <https://doi.org/10.1371/journal.pgen.1007021>
- Dulik MC, Zhadanov SI, Osipova LP, Askapuli A, Gau L, Gokcumen O, Rubinstein S, Schurr TG (2012) Mitochondrial DNA and Y chromosome variation provides evidence for a recent common ancestry between Native Americans and Indigenous Altaians. *American Journal of Human Genetics* 90: 229-246. doi: DOI: 10.1016/j.ajhg.2011.12.014
- Edge MD, Algee-Hewitt BFB, Pemberton TJ, Li JZ, Rosenberg NA (2017) Linkage disequilibrium matches forensic genetic records to disjoint genomic marker sets. *Proc Natl Acad Sci USA* 114: 5671-5676. doi: DOI: 10.1073/pnas.1619944114
- Edge MD, Coop G (2020) Attacks on genetic privacy via uploads to genealogical databases. *eLife*: e51810. doi: 10.7554/eLife.51810
- Edwards M, Bigham A, Tan J, Li S, Gozdzik A, Ross K, Jin L, Parra EJ (2010) Association of the OCA2 Polymorphism His615Arg with Melanin Content in East Asian Populations: Further Evidence of Convergent Evolution of Skin Pigmentation. *PLoS Genet* 6: e1000867. doi: <https://doi.org/10.1371/journal.pgen.1000867>
- Egeland T, Kling D, Mostad P (2015) Relationship Inference with Familias and R. *Statistical Methods in Forensic Genetics*. Academic Press.
- Egeland T, Mostad PF, Mevåg B, Stenersen M (2000) Beyond traditional paternity and identification cases: Selecting the most probable pedigree. 110: 47-59. doi: [https://doi.org/10.1016/S0379-0738\(00\)00147-X](https://doi.org/10.1016/S0379-0738(00)00147-X)
- Egeland T, Pinto N, Amorim A (2017) Exact likelihood ratio calculations for pairwise cases. *Forensic Science International: Genetics* 29: 218-224.
- Egeland T, Pinto N, Vigeland MD (2014) A general approach to power calculation for relationship testing. *Forensic Science International: Genetics* 9: 186-190. doi: <https://doi.org/10.1016/j.fsigen.2013.05.001>
- Egeland T, Sooten K (2017) Kinship analysis: properties of likelihood ratios. Seoul, ISFG workshop.
- England R, Harbison S (2019) A review of the method and validation of the MiSeq FGxTM Forensic Genomics Solution. *WIREs Forensic Sci.* e1351. doi: <https://doi.org/10.1002/wfs2.1351>
- Endicott P (2013) Introduction: revisiting the “negrito” hypothesis: a transdisciplinary approach to human prehistory in Southeast Asia. *Hum Biol.* 85: 7–20.
- Fan H, Chu JY (2007) A Brief Review of Short Tandem Repeat Mutation. *Genomics, Proteomics & Bioinformatics* 5: 7-14. doi: [https://doi.org/10.1016/S1672-0229\(07\)60009-6](https://doi.org/10.1016/S1672-0229(07)60009-6)
- Fawcett T (2006) An introduction to ROC analysis. *Pattern recognition letters* 27: 861-874.
- Fedorova L, Qiu S, Dutta R, Fedorov A (2016) Atlas of Cryptic Genetic Relatedness Among 1000 Human Genomes. *Genome Biol Evol.* 8: 777–790.
- Fondevila M, Phillips C, Santos C, Aradas AF, Vallone P, Butler J, Lareu MV, Carracedo A (2013) Revision of the SNPforID 34-plex forensic ancestry test: Assay enhancements, standard reference sample genotypes and extended population

- studies. *Forensic Science International: Genetics* 7: 63–74. doi: <https://doi.org/10.1016/j.fsigen.2012.06.007>
- Frégeau CJ (2021) Validation of the Verogen ForenSeqTM DNA Signature Prep kit/Primer Mix B for phenotypic and biogeographical ancestry predictions using the Micro MiSeq® Flow Cells. *Forensic Science International: Genetics* 53: 102533. doi: <https://doi.org/10.1016/j.fsigen.2021.102533>
- Fungammasan A, Ananda G, Hile SE, Shu-Wei Su M, Sun C, Harris R, Medvedev P, Eckert K, Makova KD (2015) Accurate typing of short tandem repeats from genome-wide sequencing data and its applications. *Genome Research* 25: 736–749.
- Galanter JM, Fernandez-Lopez JC, Gignoux CR, Barnholtz-Sloan J, Fernandez-Rozadilla C, Via M, Hidalgo-Miranda A, Contreras AV, Figueroa LU, Raska P, Jimenez-Sanchez G, Silva Zolezzi I, Torres, M., Ruiz Ponte C, Ruiz Y, Salas A, Nguyen E, Eng C, Borjas L, Zabala W, Barreto G, Rondón González F, Ibarra A, Taboada P, Porras L, Moreno F, Bigham A, Gutierrez G, Brutsaert T, León-Velarde F, Moore LG, Vargas E, Cruz M, Escobedo J, Rodriguez-Santana J, Rodriguez-Cintrón W, Chapela R, Ford JG, Bustamante C, Seminara D, Shriver M, Ziv E, Gonzalez Burchard E, Haile R, Parra E, Carracedo A, Consortium ftL (2012) Development of a Panel of Genome-Wide Ancestry Informative Markers to Study Admixture Throughout the Americas. *PLoS Genet* 8: e1002554. doi: <https://doi.org/10.1371/journal.pgen.1002554>
- García MG, Catanesi CI, Penacino GA, Gusmão L, Pinto N (2019) X-chromosome data for 12 STRs: Towards an Argentinian database of forensic haplotype frequencies. *Forensic Science International: Genetics* 41 E8-E13.
- Gazal S, Sahbatou M, Babron MC, Génin E, Leutenegger AL (2015) High level of inbreeding in final phase of 1000 Genomes Project. *Sci Rep.* 5: 17453.
- Gershaw CJ, Schweighardt AJ, Rourke LC, & Wallace MM (2011) Forensic utilization of familial searches in DNA databases. *Forensic Science International: Genetics* 5: 16-20.
- Gettings B, Borsuk LA, Steffen CR, Kiesler KM, Vallone PM (2018a) Sequence-based U.S. population data for 27 autosomal STR loci. *Forensic Science International: Genetics* 37: 106-115.
- Gettings KB, Aponte RA, Vallone PM, Butler JM (2015) STR allele sequence variation: Current knowledge and future issues. *Forensic Science International: Genetics* 18: 118-130. doi: <https://doi.org/10.1016/j.fsigen.2015.06.005>
- Gettings KB, Borsuk LA, Ballard D, Bodner M, Budowle B, Devesse L, King J, Parson W, Phillips C, Vallone PM (2017) STRSeq: A catalog of sequence diversity at human identification Short Tandem Repeat loci. *Forensic Science International: Genetics* 31: 111-117. doi: <https://doi.org/10.1016/j.fsigen.2017.08.017>
- Gettings KB, Borsuk LA, Steffen CR, Kiesler KM, Vallone PM (2018b) Sequence-based U.S. population data for 27 autosomal STR loci. *Forensic Sci Int Genet* 37: 106-115. doi: doi: 10.1016/j.fsigen.2018.07.013
- Gettings KB, Borsuk LA, Vallone PM (2017b) Performing a BLAST search of the STRSeq BioProject. *Forensic Science International: Genetics* 6: E372-E374. doi: <https://doi.org/10.1016/j.fsigss.2017.09.173>
- Gettings KB, Kiesler KM, Faith SA, Montano E, Baker CH, Young BA, Guerrieri RA, Vallone PM (2016) Sequence variation of 22 autosomal STR loci detected by next generation sequencing. *Forensic Sci. Int. Genet.* 21: 15-21.
- Gill P (2001) An assessment of the utility of single nucleotide polymorphisms (SNPs) for forensic purposes. *Int J Legal Med* 114: 204–210.

- Gjertson DW, Brenner CH, Baur MP, Carracedo A, Guidet F, Luque JA, Lessig R, Mayr WR, Pascali VL, Prinz M, Schneider PM, Morling N (2007a) ISFG: Recommendations on biostatistics in paternity testing. *Forensic Sci Int Genet* 1: 223-31.
- Gjertson DW, Brenner CH, Baur MP, Carracedo A, Guidet F, Luque JA, Lessig R, Mayr WR, Pascali VL, Prinz M, Schneider PM, Morling N (2007b) ISFG: Recommendations on biostatistics in paternity testing. *Forensic Science International: Genetics* 1: 223–231. doi: doi:10.1016/j.fsigen.2007.06.006
- Gogarten SM, Bhangale T, Conomos MP, Laurie CA, McHugh CP, Painter I, Zheng X, Crosslin DR, Levine D, Lumley T, Nelson SC, Rice K, Shen J, Swarnkar R, Weir BS, Laurie CC (2012) GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. *Bioinformatics* 28: 3329-3331. doi: doi: 10.1093/bioinformatics/bts610
- Granja R, Machado H (2019) Ethical Controversies of Familial Searching: The Views of Stakeholders in the United Kingdom and in Poland. *Science, Technology, & Human Values* 20: 1-25. doi: DOI: 10.1177/0162243919828219
- Graves JA (2006) Sex chromosome specialization and degeneration in mammals. *Cell* 124: 901-914. doi: DOI: 10.1016/j.cell.2006.02.024
- Greely HT, Riordan DP, Garrison NA, Mountain JL (2006) Family Ties: The Use of DNA Offender Databases to Catch Offenders' Kin. *Journal of Law, Medicine & Ethics. DNA Fingerprinting & Civil liberties.*
- Gusev A, Lowe JK, Stoffel M, Daly MJ, Altshuler D, Breslow JL, Friedman JM, Pe'er I (2009) Whole population, genome-wide mapping of hidden relatedness. *Genome Res* 19: 318-26. doi: doi: 10.1101/gr.081398.108.
- Hailemariam Z, Ahmed JS, Clausen PH, Nijhof AM (2017) A comparison of DNA extraction protocols from blood spotted on FTA cards for the detection of tick-borne pathogens by Reverse Line Blot hybridization. *Ticks and Tick-borne Diseases* 8: 185-189. doi: <https://doi.org/10.1016/j.ttbdis.2016.10.016>
- Halder I, Shriver M, Thomas M, Fernandez JR, Frudakis T (2008) A panel of ancestry informative markers for estimating individual biogeographical ancestry and admixture from four continents: utility and applications. *HUMAN MUTATION* 29: 648-658. doi: <https://doi.org/10.1002/humu.20695>
- Han L, Abney M (2011) Identity by descent estimation with dense genome-wide genotype data. *Genet Epidemiol* 35: 557-67. doi: DOI: 10.1002/gepi.20606
- Hanson EK, Ballantyne J (2006) Comprehensive annotated STR physical map of the human Y chromosome: Forensic implications. *Leg Med* 8: 110-120. doi: <https://doi.org/10.1016/j.legalmed.2005.10.001>
- Harrison GA (1973) Differences in human pigmentation: measurement, geographic variation and causes. *Jounal of Investigative Dermatology* 60: 418-426. doi: 10.1111/1523-1747.ep12702616
- He D, Wang Z, Han B, Parida L, Eskin E (2013) IPED: Inheritance Path-based Pedigree Reconstruction Algorithm Using Genotype Data. *J Comput Biol.* 20: 780–791.
- Henden L, Wakeham D, Bahlo M (2016) XIBD: software for inferring pairwise identity by descent on the X chromosome. *Bioinformatics* 32: 2389–2391. doi: <https://doi.org/10.1093/bioinformatics/btw124>
- Hepler AB, Weir BS (2008) Object-oriented Bayesian networks for paternity cases with allelic dependencies. *Forensic Sci Int Genet* 2: 166–175.
- Hollard C, Ausset L, Chantrel Y, Jullien S, Clot M, Faivre M, Suzanne É, Pène L, Laurent FX (2019) Automation and developmental validation of the ForenSeq™ DNA Signature Preparation kit for high-throughput analysis in forensic laboratories.

- Forensic Science International: Genetics 40: 37-45. doi: <https://doi.org/10.1016/j.fsigen.2019.01.010>
- Hopkins C, Taylor D, Hill K, Henry J (2019) Analysis of the South Australian Aboriginal population using the Global AIMs Nano ancestry test. *Forensic Sci. Int. Genet.* 41: 34–41. doi: <https://doi.org/10.1016/j.fsigen.2019.03.020>
- Hormozdiari F, Joo JWJ, Wadia A, Guan F, Ostrosky R, Sahai A, Eskin E (2014) Privacy preserving protocol for detecting genetic relatives using rare variants. *Bioinformatics* 30: i204-i211.
- Huff CD, Witherspoon DJ, Simonson TS, Xing J, Watkins WS, Zhang Y, Tuohy TM, Neklason DW, Burt RW, Guthery SL, Woodward SR, Jorde LB (2011) Maximum-likelihood estimation of recent shared ancestry (ERSA). *Genome research* 21: 768-774. doi: doi: 10.1101/gr.115972.110
- Hussing C, Borsting C, Mogensen HS, Morling N (2015) Testing of the Illumina (R) ForenSeq (TM) kit. *Forensic Science International:Genetics Supplement Series* 5. doi: <https://doi.org/10.1016/j.fsigss.2015.09.178>
- Illumina (2016) ForenSeq™ Universal Analysis Software Guide. Document #15053876 v01.
- International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million sNPs. *Nature* 449: 851–861. doi: doi: 10.1038/nature06258
- International HapMap Consortium A, D., Donnelly P (2005) A haplotype map of the human genome. *Nature* 437: 1299–1320. doi: <https://doi.org/10.1038/nature04226>
- International human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409: 860–921. doi: <https://doi.org/10.1038/35057062>
- ISOGG (2020) Autosomal DNA match thresholds. International Society of Genetic Genealogy Wiki.
- Jablonski NG, Chaplin G (2000) The evolution of skin coloration. *J Hum Evol* 39: 57-106.
- Jacquard A (1972) Genetic Information Given by a Relative. *Biometrics* 28: 1101-1114. doi: DOI: 10.2307/2528643
- Jäger AC, Alvarez ML, Davis CP, Guzmán E, Han Y, Way L, Walichiewicz P, Silva D, Pham N, Caves G, Bruand J, Schlesinger F, Pond SJK, Varlaro J, Stephens KM, Holt CL (2017) Developmental validation of the MiSeq FGx Forensic Genomics System for Targeted Next Generation Sequencing in Forensic DNA Casework and Database Laboratories. *Forensic Sci Int Genet* 28: 52-70. doi: 10.1016/j.fsigen.2017.01.011
- Jakobsson M, Scholz SW, Scheet P, Gibbs JR, VanLiere JM, Fung HC, Szpiech ZA, Degnan JH, Wang K, Guerreiro R, Bras JM, Schymick JC, Hernandez DG, Traynor BJ, Simon-Sanchez J, Matarin M, Britton A, van de Leemput J, Rafferty I, Bucan M, Cann HM, Hardy JA, Rosenberg NA, Singleton AB (2008) Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* 451: 998–1003. doi: doi: 10.1038/nature06742
- Jeffreys AJ, Brookfield JF, Semeonoff R (1985a) Positive identification of an immigration test-case using human DNA fingerprints. *Nature* 317: 818-9. doi: doi: 10.1038/317818a0
- Jeffreys AJ, Kauppi L, Neumann R (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature genetics* 29.
- Jeffreys AJ, Wilson V, Thein SL (1985b) Individual-specific 'fingerprints' of human DNA. *Nature* 316: 76-79. doi: DOI: 10.1038/316076a0

- Jobling M, Hollox E, Hurles M, Kivisild T, Tyler-Smith C (2014) Human Evolutionary Genetics (2nd Edition). New York: Garland Science.
- Jobling MA, Pandya A, Tyler-Smith C (1997) The Y chromosome in forensic analysis and paternity testing. *Int J Legal Med* 110: 118-124.
- Jobling MA, Tyler-Smith C (2003) The Human Y Chromosome: An Evolutionary Marker Comes of Age. *Nat Rev Genet* 4: 598-612. doi: doi: 10.1038/nrg1124
- Jobling MA, Tyler-Smith C (2017a) Human Y-chromosome Variation in the Genome-Sequencing Era. *Nat Rev Genet* 18: 485-497. doi: doi: 10.1038/nrg.2017.36
- Jobling MA, Tyler-Smith C (2017b) Human Y-chromosome variation in the genome-sequencing era. *Nat Rev Genet* 18: 485-497.
- Jombart T, Devillard S, Balloux F (2010) Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genetics* 11. doi: <http://www.biomedcentral.com/1471-2156/11/94>
- Jurmain R, Kilgore L, Trevathan W, Ciochon RL, Bartelink E (2018) Introduction to Physical Anthropology. Cengage Learning © Cengage, 15th Edition.
- Just RS, Moreno LI, Smerick JB, Irwin JA (2017) Performance and concordance of the ForenSeqTM system for autosomal and Y chromosome short tandem repeat sequencing of reference-type specimens. *Forensic Science International: Genetics* 28: 1-9. doi: <http://dx.doi.org/10.1016/j.fsigen.2017.01.001>
- Katsara MA, Nothnagel M (2019) True colors: A literature review on the spatial distribution of eye and hair T pigmentation. *Forensic Science International: Genetics* 39: 109–118. doi: <https://doi.org/10.1016/j.fsigen.2019.01.001>
- Kauppi L, Jeffreys AJ, Keeney S (2004) Where the crossovers are: recombination distributions in mammals. *Nat Rev Genet* 5: 413-424. doi: DOI: 10.1038/nrg1346
- Kennett D (2019) Using genetic genealogy databases in missing persons cases and to develop suspect leads in violent crimes. *Forensic Science International* 301: 107-117. doi: <https://doi.org/10.1016/j.forsciint.2019.05.016>
- Kersbergen P, van Duijn K, Kloosterman AD, den Dunnen JT, Kayser M, de Knijff P (2009) Developing a set of ancestry-sensitive DNA markers reflecting continental origins of humans. *BMC Genet* 10. doi: <https://doi.org/10.1186/1471-2156-10-69>
- Khubrani YM, Hallast P, Jobling MA, Wetton JH (2019) Massively parallel sequencing of autosomal STRs and identity-informative SNPs highlights consanguinity in Saudi Arabia. *Forensic Science International: Genetics* 43: 102164. doi: <https://doi.org/10.1016/j.fsigen.2019.102164>
- Khurani YM, Jobling MA, Wetton JH (2020) Massively parallel sequencing of sex-chromosomal STRs in Saudi Arabia reveals patrilineage-associated sequence variants. *Forensic Science International: Genetics* 49: 102402. doi: <https://doi.org/10.1016/j.fsigen.2020.102402>
- Kidd KK, Pakstis AJ, Speed WCG, E.L. Kajuna, S.L.B. , Karoma NJ, Kungulilo S, Kim JJ, Lu RB, Odunsi A, Okonofua F, Parnas J, Schulz LO, Zhukova OV, Kidd JR (2006) Developing a SNP panel for forensic identification of individuals. *Forensic Science International* 164: 20–32. doi: doi:10.1016/j.forsciint.2005.11.017
- Kidd KK, Speed WC, Pakstis AJ, Furtado MR, Fang R, Madbouly A, Maiers M, Middha M, Friedlaender FR, Kidd JR (2014) Progress toward an efficient panel of SNPs for ancestry inference. *Forensic Science International: Genetics* 10: 23-32. doi: <https://doi.org/10.1016/j.fsigen.2014.01.002>
- Kim EH, Lee HY, Kwon SY, Lee EY, Yang WI, Shin KJ (2017) Sequence-based diversity of 23 autosomal STR loci in Koreans investigated using an in-house massively parallel sequencing panel. *Forensic Science International: Genetics* 30: 134-140. doi: <https://doi.org/10.1016/j.fsigen.2017.07.001>

- Kim J, Edge MD, Goldberg A, Rosenberg NA (2021) Skin deep: The decoupling of genetic admixture levels from phenotypes that differed between source populations. *Am J Phys Anthropol.* 175: 406–421. doi: DOI: 10.1002/ajpa.24261
- King JL, Churchill JD, Novroski NMM, Zeng X, Warshauer DH, Seah LH, Budowle B (2018) Increasing the discrimination power of ancestry- and identity-informative SNP loci within the ForenSeq™ DNA Signature Prep Kit. *FSIG* 36: 60-76. doi: <https://doi.org/10.1016/j.fsigen.2018.06.005>
- Kirkpatrick B, Li SC, Karp RM, Halperin E (2011) Pedigree reconstruction using identity by descent. *J. Comput. Biol.* 18: 1481-93. doi: doi: 10.1089/cmb.2011.0156
- Kjelgaard Brustad H, Colucci M, Jobling MA, Sheehan NA, Egeland T (2021) Strategies for pairwise searches in forensic kinship analysis. *Forensic Sci Int Genet* 54: 102562. doi: DOI:<https://doi.org/10.1016/j.fsigen.2021.102562>
- Kling D, Dell'Amico B, Tillmar AO (2015a) FamLinkX- implementatio of a general model for likelihood computations for X-chromosomal marker data. *Forensic Science International: Genetics* 17: 1-7.
- Kling D, Egeland T, Tillmar AO (2012) FamLink – A user friendly software for linkage calculations in family genetics. *Forensic Science International: Genetics* 6: 616–620.
- Kling D, Mostad PF, Egeland T (2017) Manual Familias3.
- Kling D, Phillips C, Kennett D, Tillmar A (2021) Investigative genetic genealogy: Current methods, knowledge and practice. *Forensic Science International: Genetics* 52: 102474. doi: <https://doi.org/10.1016/j.fsigen.2021.102474>
- Kling D, Tillmar A (2019) Forensic genealogy—A comparison of methods to infer distant relationships based on dense SNP data. *Forensic Science International: Genetics* 42: 113–124. doi: <https://doi.org/10.1016/j.fsigen.2019.06.019>
- Kling D, Tillmar A, Egeland T (2014a) Familias 3-Extensions and new functionality. *Forensic Science International: Genetics* 13.
- Kling D, Tillmar A, Egeland T, Mostad P (2015b) A general model for likelihood computations of genetic marker data accounting for linkage, linkage disequilibrium, and mutations. *International Journal of Legal Medicine* 129: 943–954. doi: 10.1007/s00414-014-1117-7
- Kling D, Tillmar A, Egeland T, Mostad P (2015c) A General Model for Likelihood Computations of Genetic Marker Data Accounting for Linkage, Linkage Disequilibrium, and Mutations. *Int J Legal Med* 129: 943-54. doi: doi: 10.1007/s00414-014-1117-7
- Kling D, Tillmar AO, Egeland T (2014b) Familias 3 – Extensions and new functionality. *Forensic Science International: Genetics* 13: 121–127.
- Kling D, Welander J, Tillmar A, Skare Ø, Egeland T, Holmlund G (2012b) DNA microarray as a tool in establishing genetic relatedness—Current status and future prospects. *Forensic Science International: Genetics* 6: 322–329. doi: doi:10.1016/j.fsigen.2011.07.007
- Kloss-Brandsta ïter A, Pacher D, Schönherr S, Weissensteiner H, Binna R, Specht G, Kronenberg F (2010) HaploGrep: A Fast and Reliable Algorithm for Automatic Classification of Mitochondrial DNA Haplogroups. *HUMAN MUTATION* 32: 25–32. doi: DOI 10.1002/humu.21382
- Ko A, Nielsen R (2017) Composite likelihood method for inferring local pedigrees. *PLoS Genet* 13: e1006963. doi: <https://doi.org/10.1371/journal.pgen.1006963>
- Köcher S, Müller P, Berger B, Bodner M, Parson W, Roewer L, Willuweit S, Consortium TD (2018) Inter-laboratory validation study of the ForenSeq™ DNA Signature

- Prep Kit. Forensic Science International: Genetics 36: 77-85. doi: <https://doi.org/10.1016/j.fsigen.2018.05.007>
- Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Jonasdottir A, Wong WS, Sigurdsson G, Walters GB, Steinberg S, Helgason H, Thorleifsson G, Gudbjartsson DF, Helgason A, Magnusson OT, Thorsteinsdottir U, Stefansson K (2012) Rate of de novo mutations and the importance of father's age to disease risk. *Nature* 488: 471-475. doi: DOI: 10.1038/nature11396
- Kong A, Gudbjartsson DF, Sainz J, Jonsdottir GM, Gudjonsson SA, Richardsson B, Sigurdardottir S, Barnard J, Hallbeck B, Masson G, Shlien A, Palsson ST, Frigge ML, Thorgeirsson TE, Gulcher JR, Stefansson K (2002) A high-resolution recombination map of the human genome. *Nat Genet* 31: 241-7. doi: doi: 10.1038/ng917
- Kong A, Masson G, Frigge ML, Gylfason A, Zusmanovich P, Thorleifsson G, Olason PI, Ingason A, Steinberg S, Rafnar T, Sulem P, Mouy M, Jonsson F, Thorsteinsdottir U, Gudbjartsson DF, Stefansson H, Stefansson K (2008) Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat. Genet.* 40: 1068–75. doi: doi: 10.1038/ng.216
- Kopelman NM, Mayzel J, Jakobsson M, Rosenberg NA, Mayrose I (2015) CLUMPAK: a program for identifying clustering modes and packaging population structure inferences across K. *Molecular Ecology Resources*: 1179–1191. doi: Molecular Ecology Resources (2015) 15, 1179–1191
- Korneliussen T, and Moltke, I (2015) NgsRelate: a software tool for estimating pairwise relatedness from next-generation sequencing data. *Bioinformatics* 31: 4009–4011. doi: <https://doi.org/10.1093/bioinformatics/btv509>
- Kruijver M (2015) Efficient computations with the likelihood ratio distribution. *Forensic Sci Int Genet.* 14: 116-24. doi: doi:10.1016/j.fsigen.2014.09.018
- Kurbasic A, Hošsjer O (2008) A General Method for Linkage Disequilibrium Correction for Multipoint Linkage and Association. *Genetic Epidemiology* 32: 647–657. doi: DOI: 10.1002/gepi.20339
- Kvaløy K, Galvagni F, Brown WRA (1994) The sequence organization of the long arm pseudoautosomal region of the human sex-chromosomes. *Hum Mol Genet* 3: 771-778.
- Lander ES, Green P (1987) Construction of multilocus genetic linkage maps in humans. *Proc. Natl. Acad. Sci. U.S.A.* 84: 2263–2267.
- Lareu MV, del Carmen Pestoni M, Barros F, Salas A, Carracedo A (1996) Sequence variation of a hypervariable short tandem repeat at the D12S391 locus. *Gene* 182: 151-153. doi: [https://doi.org/10.1016/S0378-1119\(96\)00540-9](https://doi.org/10.1016/S0378-1119(96)00540-9)
- Laurent, F.X., Fischer, A., Oldt, R.F., Kanthaswamy, S., Buckleton, J.S., and Hitchin, S.(2022) Streamlining the decision-making process for international DNA kinship matching using Worldwide allele frequencies and tailored cutoff log10LR thresholds. *Forensic Sci Int Genet.* 57:102634. doi:<https://doi.org/10.1016/j.fsigen.2021.102634>
- Lawson DJ, Falush D (2012) Population Identification Using Genetic Data. *Annual Review of Genomics and Human Genetics* 13: 337-361. doi: <https://doi.org/10.1146/annurev-genom-082410-101510>
- Lewontin RC (1964) The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* 49: 49-67.
- Lewontin RC (1972) The apportionment of human diversity. *Evolutionary Biology*: 381-398.

- Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* 27: 2987-93. doi: <http://www.htslib.org/doc/#publications>
- Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25: 1754–1760. doi: <https://doi.org/10.1093/bioinformatics/btp324>
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP (2009) The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* 25: 2078-9.
- Li H, Zhang C, Song G, Ma K, Cao Y, Zhao X, Yang Q, Xie J (2021) Concordance and characterization of massively parallel sequencing at 58 STRs in a Tibetan population. *Mol Genet Genomic Med.* 9: e1626. doi: 10.1002/mgg3.1626
- Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, Ramachandran S, Cann HM, Barsh GS, Feldman M, Cavalli-Sforza LL, Myers RM (2008) Worldwide Human Relationships Inferred from Genome-Wide Patterns of Variation. *Science* 319: 1100–1104. doi: 10.1126/science.1153717
- Li R, Li H, Peng D, Hao B, Wang Z, Huang E, Wu R, Sun H (2019) Improved pairwise kinship analysis using massively parallel sequencing. *Forensic Science International: Genetics* 38: 77-85. doi: <https://doi.org/10.1016/j.fsigen.2018.10.006>
- Life Technologies (2012) GlobalFiler™ Express PCR Amplification Kit User Guide. Publication Part Number 4477672 Rev. A.
- Liu C, Lee I MK, Naqvi S, Hoskens H, Liu D, White JD, Indencleef K, Matthews H, Eller RJ, Li I J, Mohammed J, Swigut T, Richmond S, Manyama M, Hallgrímsson B, Spritz RA, Feingold E, Marazita ML, Wysocka J, Walsh S, Shriver MD, Claes P, Weinberg SM, Shaffer JR (2021) Genome scans of facial features in East Africans and cross-population comparisons reveal novel associations. *PLoS Genet* 17: e1009695. doi: <https://doi.org/10.1371/journal.pgen.1009695>
- Liu F, van Duijn K, Vingerling JR, Hofman A, Uitterlinden AG, Janssens AC, Kayser M (2009) Eye color and the prediction of complex phenotypes from genotypes. *Current Biology* 19.
- Lynch M, Ritland K (1999) Estimation of Pairwise Relatedness With Molecular Markers. *GENETICS* 152: 1753-1766.
- Manichaikul A, Mychaleckyj JC, Rich SS, Daly K, Sale M, Chen WM (2010) Robust relationship inference in genome-wide association studies. *Bioinformatics* 26: 2867-2873.
- Martin AR, Lin M, Granka JM, Myrick JY, Liu X, Sockell A, Atkinson EG, Werely CJ, Möller M, Sandhu MS, Kingsley DM, Hoal EG, Liu X, Daly MJ, Feldman MW, Gignoux CR, Bustamante CD, Henn BM (2017) An Unexpectedly Complex Architecture for Skin Pigmentation in Africans. *Cell* 171: 1340–1353. doi: <https://doi.org/10.1016/j.cell.2017.11.015>
- McVean G (2009) A Genealogical Interpretation of Principal Components Analysis. *PLoS Genet.* 5: e1000686. doi: doi: 10.1371/journal.pgen.1000686
- Mengel-From J, Wong TH, Morling Nea (2009) Genetic determinants of hair and eye colours in the Scottish and Danish populations. *BMC Genet* 10. doi: <https://doi.org/10.1186/1471-2156-10-88>
- Miller G (2010) Familial DNA testing scores a win in serial killer case. *Science* 329: 262-262. doi: doi:10.1126/science.329.5989.262
- Milligan B (2003) Maximum-Likelihood Estimation of Relatedness. *Genetics Society of America* 163: 1153–1167

- Miyazawa H, Kato M, Awata T, Kohda M, Iwasa H, Koyama N, Tanaka T, Huqun K, S., Okazaki Y, Hagiwara K (2007) Homozygosity haplotype allows a genomewide search for the autosomal segments shared among patients. *Am. J. Hum. Genet.* 80: 1090–1102. doi: <https://doi.org/10.1086/518176>
- Mo SK, Ren ZL, Yang YR, Liu YC, Zhang JJ, Wud HJ, Lia Z, Bo XC, Wang SQ, Yan JY, Ni M (2018) A 472-SNP panel for pairwise kinship testing of second-degree relatives. *Forensic Science International: Genetics* 34: 178–185. doi: <https://doi.org/10.1016/j.fsigen.2018.02.019>
- Moreau R (2012) DNA forensics: In the U.S. states can pass their own familial searching laws. *DNA Forensics: New and Information about DNA Databases*.
- Morimoto, C., Tsujii, H., Manabe, S., Fujimoto, S., Hirai, E., Hamano, Y., and Tamaki, K. (2020) Development of a software for kinship analysis considering linkage and mutation based on a Bayesian network. *Forensic Sci Int Genet.* 47: 102279. doi:<https://doi.org/10.1016/j.fsigen.2020.102279>.
- Mu W, Zhang W (2013) Chapter 8 - Molecular Approaches, Models, and Techniques in Pharmacogenomic Research and Development. *Pharmacogenomics: Challenges and Opportunities in Therapeutic Implementation*, Academic Press: 273-294. doi: <https://doi.org/10.1016/B978-0-12-391918-2.00008-1>
- Nagai A, Bunai Y (2011) Structural polymorphisms at the X-chromosomal short tandem repeat loci DXS10134, DXS10135, DXS10146 and DXS10148. *Forensic Science International: Genetics Supplement Series* 3: e343-e344. doi: <https://doi.org/10.1016/j.fsigss.2011.09.034>
- National Research Council (1996) *The Evaluation of Forensic DNA Evidence*. Washington, DC: The National Academies Press. doi: <https://doi.org/10.17226/5141>
- Norton HL, Kittles RA, Parra E, McKeigue P, Mao X, Cheng K, Canfield VA, Bradley DG, McEvoy BM, Shriver MD (2007) Genetic evidence for the convergent evolution of light skin in Europeans and East Asians. *Mol Biol Evol* 24: 710–22. doi: <https://doi.org/10.1093/molbev/msl203>
- Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, Indap A, King KS, Bergmann S, Nelson MR, Stephens M, Bustamante CD (2008) Genes mirror geography within Europe. *Nature* 456: 98-101.
- Novroski NMM, King JL, Churchill JD, Seah LH, Budowle B (2016) Characterization of genetic sequence variation of 58 STR loci in four major population groups. *Forensic Science International: Genetics* 25: 214-226. doi: <https://doi.org/10.1016/j.fsigen.2016.09.007>
- Olivieri L, Mazzarelli D, Bertoglio B, De Angelis DP, C., Grignani P, Cappella AP, S. Bertuglia, C., Di Simone P, Polizzi N, Iadicicco A, Piscitelli V, Cattaneo C (2018) Challenges in the identification of dead migrants in the Mediterranean: The case study of the Lampedusa shipwreck of October 3rd 2013. *Forensic Science International* 285: 121-128. doi: <https://doi.org/10.1016/j.forsciint.2018.01.029>
- Pakstis AJ, Speed WC, Fang R, Hyland FCL, Furtado MR, Kidd JR, Kidd KK (2010) SNPs for a universal individual identification panel. *Hum Genet* 127: 315-24. doi: doi: 10.1007/s00439-009-0771-1
- Palomo-Díez S, Esparza Arroyo Á, Tirado-Vizcaíno M, Velasco Vázquez J, López-Parra AM, Gomes C, Baeza-Richer C, Arroyo-Pardo E (2018) Kinship analysis and allelic dropout: a forensic approach on an archaeological case. *Annals of human biology* 45: 365-368.
- Palomo-Díez S, López-Parra, A.M., Gomes C, Baeza-Richer, C., Esparza-Arroyo, A., Arroyo-Pardo E (2015) Kinship analysis in mass graves: evaluation of the Blind

- Search tool of the Familias 3.0 Software in critical samples. *Forensic Science International: Genetics Supplement Series* 5: e547–e550. doi: <http://dx.doi.org/10.1016/j.fsigss.2015.09.216>
- Park JH, Lee MH (2005) A Study of Skin Color by Melanin Index According to Site, Gestational Age, Birth Weight and Season of Birth in Korean Neonates. *J Korean Med Sci* 20: 105-8.
- Parson W, Ballard D, Budowle B, Butler JM, Gettings KB, Gill P, Gusmão L, Hares DR, Irwin JA, King JL, de Knijff P, Morling N, Prinzo M, Schneider PM, Van Neste C, Willuweit S, Phillips C (2016) Massively parallel sequencing of forensic STRs: Considerations of the DNA commission of the International Society for Forensic Genetics (ISFG) on minimal nomenclature requirements. *Forensic Science International: Genetics* 22: 54-63.
- Parsons TJ, Huel RML, Bajunović Z, Rizvić A (2019) Large scale DNA identification: The ICMP experience. *Forensic Science International: Genetics* 38: 236-244. doi: <https://doi.org/10.1016/j.fsigen.2018.11.008>
- Patterson N, Price AL, D R (2006) Population structure and eigenanalysis. *PLoS Genet* 2: e190. doi: doi:10.1371/journal.pgen.0020190
- Peng D, Zhang Y, Ren H, Li H, Li R, Shen X, Wang N, Huang E, Wu R, Sun H (2020) Identification of sequence polymorphisms at 58 STRs and 94 iiSNPs in a Tibetan population using massively parallel sequencing. *Nature Scientific Reports* 10: 12225. doi: <https://doi.org/10.1038/s41598-020-69137-1>
- Petter E, Schweiger R, Shahino B, Shor T, Aker M, Almog L, Weissglas-Volkov D, Naveh Y, Navon O, Carmi S, Li JH, Berisa T, Pickrell JK, Erlich Y (2020) Relative matching using low coverage sequencing. *bioRxiv*. doi: <https://doi.org/10.1101/2020.09.09.289322>
- Phillips C (2018) The Golden State Killer investigation and the nascent field of forensic genealogy. *Forensic Science International: Genetics* 36: 186-188. doi: <https://doi.org/10.1016/j.fsigen.2018.07.010>
- Phillips C, Devesse L, Ballard D, van Weert L, de la Puente M, Melis S, Álvarez Iglesias V, Freire-Aradas A, Oldroyd N, Holt C, Syndercombe Court D, Carracedo A, Lareu MV (2018a) Global patterns of STR sequence variation: Sequencing the CEPH human genome diversity panel for 58 forensic STRs using the Illumina ForenSeq DNA Signature Prep Kit. *Electrophoresis* 39: 2708–2724.
- Phillips C, Fernandez-Formoso L, Garcia-Magarinos M, Porras L, Tvedebrink T, Amigo J, Fondevila M, Gomez-Tato A, Alvarez-Dios J, Freire-Aradas A, Gomez-Carballa A, Mosquera-Miguel A, Carracedo A, Lareu MV (2011) Analysis of global variability in 15 established and 5 new European Standard Set (ESS) STRs using the CEPH human genome diversity panel. *Forensic Science International: Genetics* 5: 155–169. doi: doi:10.1016/j.fsigen.2010.02.003
- Phillips C, Gettings KB, King JL, Ballard D, Bodner M, Borsuk L, Parson W (2018b) “The devil’s in the detail”: release of an expanded, enhanced and dynamically revised forensic STR Sequence Guide. *Forensic Sci. Int. Genet.* 34: 162-169.
- Phillips C, Parson W, Lundsberg B, Santos C, Freire-Aradas A, Torres M, Eduardoff M, Børsting C, Johansen P, Fondevila M, Morling N, Schneider P, EUROFORGEN-NoE Consortium, Carracedo A, Lareu MV (2014) Building a forensic ancestry panel from the ground up: The EUROFORGEN global AIM-SNP set. *Forensic Science International: Genetics* 11: 13–25.
- Pinto N, Gusmao L, and Amorim, A. (2011) X-chromosome markers in kinship testing: a generalisation of the IBD approach identifying situations where their contribution is crucial. *Forensic Science International: Genetics* 5: 27- 32.

- Pinto N, Silva PV, Amorim A (2012) A general method to assess the utility of the X-chromosomal markers in kinship testing. *Forensic Science International: Genetics* 6: 198–207. doi:10.1016/j.fsigen.2011.04.014
- Popa, A, Samollow, P, Gautier, C, and Mouchiroud, D (2012) The sex-specific impact of meiotic recombination on nucleotide composition. *Genome biology and evolution*, 4(3): 412–422. doi:<https://doi.org/10.1093/gbe/evs023>
- Powell JE, Visscher PM, Goddard ME (2010) Reconciling the analysis of IBD and IBS in complex trait studies. *Nature Reviews Genetics* 11: 800–805. doi: DOI: 10.1038/nrg2865
- Poznik GD (2016) Identifying Y-chromosome haplogroups in arbitrarily large samples of sequenced or genotyped men. *bioRxiv*. 23andMe. doi: doi: <https://doi.org/10.1101/088716>
- Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet.* 38: 904-9.
- Pritchard JK, Donnelly P (2001) Case-control studies of association in structured or admixed populations. *Theoretical Population Biology* 60: 227–237.
- Pritchard JK, Przeworski M (2001) Linkage Disequilibrium in Humans: Models and Data. *Am. J. Hum. Genet.* 69: 1-14. doi: <https://doi.org/10.1086/321275>
- Pritchard JK, Stephens M, Donnelly P (2000) Inference of Population Structure Using Multilocus Genotype Data. *Genetics* 155: 945–959.
- Promega (2017) PowerPlex® ESI 17 Pro System Technical Manual.
- Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, Maller J, Sklar P, de Bakker PIW, Daly MJ, Sham PC (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. *American Journal of Human Genetics* 81: 559-75. doi: 10.1086/519795
- QIAgen (2016) Investigator® 24plex QS Handbook.
- R Core Team (2014) R: A language and environment for statistical computing (<http://www.R-project.org/>). R Foundation for Statistical Computing, Vienna, Austria.
- Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL (2005) Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. *Proc Natl Acad Sci USA* 102: 15942-15947.
- Ramani A, Wong Y, Tan SZ, Shue BH, Syn C (2017) Ancestry prediction in Singapore population samples using the Illumina ForenSeq kit. *Forensic Science International: Genetics* 31: 171–179. doi: <http://dx.doi.org/10.1016/j.fsigen.2017.08.013>
- Rębała K, Martínez-Cruz B, Tönjes A, Kovacs P, Stumvoll M, Lindner I, Büttner A, Wichmann HE, Siváková D, Soták M, Quintana-Murci L, Szczerkowska Z, Comas D, Consortium tG (2013) Contemporary paternal genetic landscape of Polish and German populations: from early medieval Slavic expansion to post-World War II resettlements. *Eur J Hum Genet.* 2: 415–422. doi: 10.1038/ejhg.2012.190
- Reich D, Patterson N, Campbell D, Tandon A, Mazieres S, Ray N, Parra MV, Rojas W, Duque C, Mesa N, García LF, Triana O, Blair S, Maestre A, Dib JC, Bravi CM, Bailliet G, Corach D, Hünemeier T, Cátila Bortolini M, Salzano FM, Petzl-Erler ML, Acuña-Alonso V, Aguilar-Salinas C, Canizales-Quinteros S, Tusié-Luna T, Riba L, Rodríguez-Cruz M, Lopez-Alarcón M, Coral-Vazquez R, Canto-Cetina T, Silva-Zolezzi I, Fernandez-Lopez JC, Contreras AW, Jimenez-Sanchez G,

- Gómez-Vázquez MJ, Molina J, Carracedo A, Salas A, Gallo C, Poletti G, Witonsky DB, Alkorta-Aranburu G, Sukernik RI, Osipova L, Fedorova SA, Vasquez R, Villena M, Moreau C, Barrantes R, Pauls D, Excoffier L, Bedoya G, Rothhammer F, Dugoujon JM, Larrouy G, Klitz W, Labuda D, Kidd J, Kidd K, Di Rienzo A, Freimer NB, Price AL, Ruiz-Linares A (2012) Reconstructing Native American population history. *Nature* 488: 370–374. doi: <https://doi.org/10.1038/nature11258>
- Relethford JH (1992) Cross-cultural analysis of migration rates: effects of geographic distance and population size. *Am. J. Phys. Anthropol.* 89: 459–466. doi: <https://doi.org/10.1002/ajpa.1330890407>
- Richards M, Macaulay V, Hickey E, Vega E, Sykes B, Guida V, Rengo C, Sellitto D, Cruciani F, Kivisild T, Villems R, Thomas M, Rychkov S, Rychkov O, Rychkov Y, Golge M, Dimitrov D, Hill E, Bradley D, Romano V, Cali F, Vona G, Demaine A, Papiha S, Triantaphyllidis C, Stefanescu G, Hatina J, Belledi M, Di Rienzo A, Novelletto A, Oppenheim A, Norby S, Al-Zaheri N, Santachiara-Benerecetti S, Scozzari R, Torroni A, Bandelt HJ (2000) Tracing European founder lineages in the near eastern mtDNA pool. *Am J Hum Genet* 67: 1251–1276.
- Riester M, Stadler PF, Klemm K (2009) FRANz: reconstruction of wild multi-generation pedigrees. *Bioinformatics* 25: 2134–2139.
- Rosenberg NA, Pritchard JK, Weber JL, Cann HM, Kidd, K.K., Zhivotovsky LA, Feldman MW (2002) Genetic Structure of Human Populations. *Science* 298: 2381. doi: DOI: 10.1126/science.1078311
- Roslin NM, Li W, Paterson AD, Strug LJ (2016) Quality control analysis of the 1000 Genomes Project Omni2.5 genotypes (Abstract/Program #576/F). Presented at the 66th Annual Meeting of The American Society of Human Genetics, October 18–22, 2016, Vancouver, Canada.
- Salmela E, Lappalainen T, Fransson I, Andersen PM, Dahlman-Wright K, Fiebig A, Sistonen P, Savontaus ML, Schreiber S, Kere J, Lahermo P (2008) Genome-Wide Analysis of Single Nucleotide Polymorphisms Uncovers Population Structure in Northern Europe. *PLoS ONE* 3: e3519. doi: <https://doi.org/10.1371/journal.pone.0003519>
- Samuel G, Kennett D (2020) Problematizing consent: searching genetic genealogy databases for law enforcement purposes. *New Genetics and Society*: 1–21. doi: <https://doi.org/10.1080/14636778.2020.1843149>
- Sanchez JJ, Phillips C, Børsting C, Balogh K, Bogus M, Fondevila M, Harrison CD, Musgrave-Brown E, al. e (2006) A multiplex assay with 52 single nucleotide polymorphisms for human identification. *Electrophoresis* 27: 1713–1724.
- Sanchez M, Givry SD, Schiex T (2008) Mendelian Error Detection in Complex Pedigrees Using Weighted Constraint Satisfaction Techniques. *Constraints* 13: 130–154. doi: DOI 10.1007/s10601-007-9029-5
- Schlauch D, Fier H, Lange C (2017) Identification of genetic outliers due to sub-structure and cryptic relationships. *Bioinformatics* 33: 1972–1979.
- Schlötterer C, Tautz D (1992) Slippage synthesis of simple sequence DNA. *Nucl Acids Res* 20: 211–215. doi: DOI: 10.1093/nar/20.2.211
- Schneider PM (2012a) Beyond STRs: The Role of Diallelic Markers in Forensic Genetics. *Transfusion Medicine and Hemotherapy* 39. doi: <https://doi.org/10.1159/000339139>
- Schneider PM (2012b) Beyond STRs: The Role of Diallelic Markers in Forensic Genetics. *Transfus Med Hemother* 39: 176–180. doi: DOI: 10.1159/000339139

- Schneider PM, Prainsack B, Kayser M (2019) The Use of Forensic DNA Phenotyping in Predicting Appearance and Biogeographic Ancestry. *Dtsch Arztbl Int* 51-52: 873-880. doi: 10.3238/arztbl.2019.0873
- Sharma V, Chow HY, Siegel D, Wurmbach E (2017) Qualitative and quantitative assessment of Illumina's forensic STR and SNP kits on MiSeq FGxTM. *PLoS ONE* 12: e0187932. doi: <https://doi.org/10.1371/journal.pone.0187932>
- Sharma V, Jani K, Khosla P, Butler E, Siegel D, Wurmbach E (2019) Evaluation of ForenSeq™ Signature Prep Kit B on predicting eye and hair coloration as well as biogeographical ancestry by using Universal Analysis Software (UAS) and available web-tools. *Electrophoresis* 40: 1353-1364. doi: <https://doi.org/10.1002/elps.201800344>
- Sharma V, van der Plaat DA, Liu Y, Wurmbach E (2020) Analyzing degraded DNA and challenging samples using the ForenSeq™ DNA Signature Prep kit. *Science & Justice* 60: 243-252. doi: <https://doi.org/10.1016/j.scijus.2019.11.004>
- Sheehan NA, Egeland T (2007) Structured Incorporation of Prior Informationin Relationship Identification Problems. *Ann. Hum. Genet.* 71: 501-518. doi: doi: 10.1111/j.1469-1809.2006.00345.x
- Shekar SN, Duffy DL, Frudakis T, Montgomery GW, James MR, Sturm RA, Martin NG (2008) Spectrophotometric methods for quantifying pigmentation in human hair-influence of MC1R genotype and environment. *Photochemistry and Photobiology* 84: 719-726. doi: 10.1111/j.1751-1097.2007.00237.x
- Shem-Tov D, Halperin E (2014) Historical pedigree reconstruction from extant populations using PArtitioning of RElatives (PREPARE). *PLoS Comput. Biol.* 10: e1003610. doi: doi: 10.1371/journal.pcbi.1003610
- Shifman S, Kuypers J, Kokoris M, Yakir B, Darvasi A (2003) Linkage disequilibrium patterns of the human genome across populations. *Human Molecular Genetics* 12: 771-776. doi: <https://doi.org/10.1093/hmg/ddg088>
- Shriver MD, Smith MW, Jin L, Marcini A, Akey JM, Deka R, Ferrell RE (1997) Ethnic-affiliation estimation by use of population-specific DNA markers. *Am J Hum Genet* 60: 957-964.
- Silvia AL, Shugarts N, Smith J (2017) A preliminary assessment of the ForenSeq™ FGx System: next generation sequencing of an STR and SNP multiplex. *Int J Legal Med* 131: 73-86. doi: 10.1007/s00414-016-1457-6
- Simonsson I, and Mostad, P. (2016) Stationary mutation models. *Forensic Science International: Genetics* 23: 217-225.
- Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova R, Ali J, Bieri T, Chinwalla A, Delehaunty A, Delehaunty K, Du H, Fewell G, Fulton L, Fulton R, Graves T, Hou S-F, Latrielle P, Leonard S, Mardis E, Maupin R, McPherson J, Miner T, Nash W, Nguyen C, Ozersky P, Pepin K, Rock S, Rohlfing T, Scott K, Schultz B, Strong C, Tin-Wollam A, Yang S-P, Waterston RH, Wilson RK, Rozen S, Page DC (2003) The male-specific region of the human Y chromosome: a mosaic of discrete sequence classes. *Nature* 423: 825-837.
- Skare Ø, Sheehan N, Egeland T (2009) Identification of distant family relationships. *BIOINFORMATICS* 25: 2376–2382. doi: doi:10.1093/bioinformatics/btp418
- Slatkin M (2008) Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nat Rev Genet.* 9: 477–485. doi: doi: 10.1038/nrg2361
- Slooten K, Meester R (2014) Probabilistic strategies for familial DNA searching. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 63: 361-384.

- Smith J, Qiao Y, Williams AL (2021) Evaluating the utility of identity-by-descent segment numbers for relatedness inference via information theory and classification. bioRxiv. doi: <https://doi.org/10.1101/2021.09.14.460357>
- Soares P, Ermini L, Thomson N, Mormina M, Rito T, Rohl A, Salas A, Oppenheimer S, Macaulay V, and Richards, M.B. (2009) Correcting for purifying selection: an improved human mitochondrial molecular clock. *Am J Hum Genet* 84: 740-759.
- Stankovich J, Bahlo M, Rubio JP, Wilkinson CR, Thomson R, Banks A, Ring M, Foote SJ, Speed TP (2005) Identifying nineteenth century genealogical links from genotypes. *Human Genetics* 117: 188–199.
- Staples J (2014) PRIMUS: Pedigree Reconstruction and Identification of a Maximum Unrelated Set. University of Washington: Thesis.
- Staples J, Ekunwe L, Lange E, Wilson JG, Nickerson DA, Below JE (2016) PRIMUS: improving pedigree reconstruction using mitochondrial and Y haplotypes. *Bioinformatics* 32: 596–598.
- Staples J, Qiao D, Cho MH, Silverman EK, University of Washington Center for Mendelian Genomics N, D.A., Below JE (2014) PRIMUS: Rapid Reconstruction of Pedigrees from Genome-wide Estimates of Identity by Descent. *Am J Hum Genet.* 95: 553–564. doi: doi: 10.1016/j.ajhg.2014.10.005
- Steffens M, Lamina C, Illig T, Bettecken T, Vogler R, Entz P, Suk E, Toliat MR, Klopp N, Caliebe A, König IR, Köhler K, Lüdemann J, Lacava AD, Fimmers R, Lichtner P, Ziegler A, Wolf A, Krawczak M, Nürnberg P, Hampe J, Schreiber S, Meitinger T, Wichmann HE, Roeder K, Wienker TF, Baur MP, Karsten AE, Singh A, Karsten PA, Braun MWH (2006) SNP-Based Analysis of Genetic Substructure in the German Population. *Hum. Hered.* 62: 20–29. doi: DOI: 10.1159/000095850
- Stevens EL, Heckenberg G, Roberson EDO, Baugher JD, Downey TJ, Pevsner J (2011) Inference of Relationships in Population Data Using Identity-by-Descent and Identity-by-State. *PLoS Genet* 7: e1002287. doi: <https://doi.org/10.1371/journal.pgen.1002287>
- Stokes L (2008) Press release: DNA technology to progress more cold cases. The Forensic Science Service 31.
- Storey JD (2003) The positive false discovery rate: a Bayesian interpretation and the q-value. *The Annals of Statistics* 31: 2013-2035.
- Sulem P, Gudbjartsson DF, Stacey SN, Helgason A, Rafnar T, Magnusson KP, Manolescu A, Karason A, Palsson A, Thorleifsson G, Jakobsdottir M, Steinberg S, Pálsson S, Jonasson F, Sigurgeirsson B, Thorisdottir K, Ragnarsson R, Benediktsdottir KR, Aben KK, Kiemeney LA, Olafsson JH, Gulcher J, Kong A, Thorsteinsdottir U, Stefansson K (2007) Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nature Genetics* 39: 1443–1452. doi: <https://doi.org/10.1038/ng.2007.13>
- Sumita DR, Whittle MR (2009) Updated allelic structures of the DXS10135 and DXS10078 STR loci. *Forensic Science International: Genetics Supplement Series* 2: 51-52. doi: <https://doi.org/10.1016/j.fsigss.2009.08.077>
- Sun JX, Helgason A, Masson G, Ebenesersdottir SS, Li H, Mallick S, Gnerre S, Patterson N, Kong A, Reich D, Stefansson K (2012) A direct characterization of human mutation based on microsatellites. *Nat. Genet.* 44: 1161-5. doi: doi: 10.1038/ng.2398
- Syndercombe Court D (2018) Forensic genealogy: Some serious concerns. *Forensic Science International: Genetics* 36: 203-204. doi: <https://doi.org/10.1016/j.fsgen.2018.07.011>

- Szibor R (2007) X-chromosomal markers: Past, present and future. *Forensic Science International: Genetics* 1: 93–99. doi: doi:10.1016/j.fsigen.2007.03.003
- Tamhane AC, Hochberg Y, Dunnett CW (1996) Multiple test procedures for dose finding. *Biometrics* 52: 21-37.
- Tang H, Peng J, Wang P, Risch NJ (2005) Estimation of individual admixture: analytical and study design considerations. *Genet Epidemiol* 28: 289–301. doi: <https://doi.org/10.1002/gepi.20064>
- Taroni F, Biedermann A, Bozza S, Garbolino P, Aitken C (2014) Bayesian networks for probabilistic inference and decision analysis in forensic science. John Wiley & Sons, Incorporated, New York.
- TGPC (2015) A global reference for human genetic variation. *Nature* 526: 68–74. doi: <https://www.nature.com/articles/nature15393>
- Thatte BD, Steel M (2008) Reconstructing pedigrees: A stochastic perspective. *Journal of Theoretical Biology* 251: 440–449.
- The 1000 Genomes Project Consortium (2015) A global reference for human genetic variation. *NATURE* 526. doi: doi:10.1038/nature15393
- The 1000 Genomes Project Consortium A, A., Brooks LD, Durbin RM, Garrison EP, Kang HM, Korbel JO, Marchini JL, McCarthy S, McVean GA, Abecasis GR (2015) A global reference for human genetic variation. *Nature* 526: 68-74.
- Thompson EA (1975) The estimation of pairwise relationships. *Ann. Hum. Genet.* 39: 173–188.
- Thompson EA (1985) Pedigree Analysis in Human Genetics. Johns Hopkins University Press, Baltimore.
- Thompson EA (2000) Statistical inference from genetic data on pedigrees. In NSF-CBMS regional conference series in probability and statistics. Institute of Mathematical Statistics and the American Statistical Association.
- Thompson EA (2008) The IBD process along four chromosomes. *Theoretical population biology* 73: 369-373. doi: <https://doi.org/10.1016/j.tpb.2007.11.011>
- Thornton T (2015) Statistical Methods for Genome-Wide and Sequencing Association Studies of Complex Traits in Related Samples. *Curr Protoc Hum Genet.* 84: 1.28.1–1.28.9. doi: doi: 10.1002/0471142905.hg0128s84
- Thornton T, and Bermejo, JL (2014) Local and Global Ancestry Inference, and Applications to Genetic Association Analysis for Admixed Populations. *Genet Epidemiol.* 38: S5–S12. doi: doi:10.1002/gepi.21819.
- Thornton T, Conomos MP, Sverdlov S, Blue EM, Cheung CYK, Glazner CG, Lewis SM, Wijsman EM (2014) Estimating and adjusting for ancestry admixture in statistical methods for relatedness inference, heritability estimation, and association testing. *BMC Proc.* 8(Suppl 1): S5. doi: 10.1186/1753-6561-8-S1-S5
- Thornton T, Tang H, Hoffmann, T.J., Ochs-Balcom HM, Caan BJ, Risch N (2012) Estimating kinship in admixed populations. *Am J Hum Genet* 91: 122-38. doi: doi: 10.1016/j.ajhg.2012.05.024
- Thornton TA, Bermejo JL (2014) Local and Global Ancestry Inference, and Applications to Genetic Association Analysis for Admixed Populations. *Genet Epidemiol.* 38: S5–S12.
- Tillmar A, Fagerholm SA, Staaf J, Sjölund P, Ansell R (2021) Getting the conclusive lead with investigative genetic genealogy - A successful case study of a 16 year old double murder in Sweden. *Forensic Sci Int Genet* 53: 102525. doi: <https://doi.org/10.1016/j.fsigen.2021.102525>
- Tillmar A, Sjölund P, Lundqvist B, Klippmark T, Älgenäs C, Green H (2020) Whole-genome sequencing of human remains to enable genealogy DNA database

- searches – A case report. *Forensic Science International: Genetics* 46: 102233. doi: <https://doi.org/10.1016/j.fsigen.2020.102233>
- Tillmar AO, Mostad P (2014) Choosing supplementary markers in forensic casework. *Forensic Sci Int Genet*. 13: 128-133. doi: <https://doi.org/10.1016/j.fsigen.2014.06.019>
- Tiret M, Hospital F (2017) Blocks of chromosomes identical by descent in a population: Models and predictions. *PLoS ONE* 12: e0187416. doi: <https://doi.org/10.1371/journal.pone.0187416>
- Tomas C, Sanchez JJ, Barbaro A, Brandt-Casadevall C, Hernandez A, Dhiab MB, Ramon M, Morling N (2008) X-chromosome SNP analyses in 11 human Mediterranean populations show a high overall genetic homogeneity except in North-west Africans (Moroccans). *BMC Evolutionary Biology* 8. doi: <https://doi.org/10.1186/1471-2148-8-75>
- Trombetta B, Cruciani F (2017) Y Chromosome Palindromes and Gene Conversion. *Hum Genet* 136: 605-619. doi: doi: 10.1007/s00439-017-1777-8
- Urquhart A, Kimpton CP, Downes TJ, Gill P (1994) Variation in short tandem repeat sequences--a survey of twelve microsatellite loci for use as forensic identification markers. *Int J Legal Med* 107: 13-20. doi: doi: 10.1007/BF01247268
- Valenzuela RK, Henderson MS, Walsh MH, Garrison NA, Kelch JT, Cohen- Barak O, Erickson DT, Meaney FJ, Walsh JB, Cheng KC, Ito S, Wakamatsu K, Frudakis T, Thomas M, Brilliant MH (2010) Predicting phenotype from genotype: normal pigmentation. *J. Forensic Sci.* 55: 315–322.
- van Oven M (2015a) PhyloTree Build 17: growing the human mitochondrial DNA tree. *Forensic Sci. Int. Genet. Suppl. Ser.* 5: 392–394.
- van Oven M (2015b) PhyloTree Build 17: Growing the human mitochondrial DNA tree. FSIG 5 E392-E394. doi: DOI:<https://doi.org/10.1016/j.fsigss.2015.09.155>
- van Oven M, Van Geystelen A, Kayser M, Decorte R, Larmuseau MH (2014) Seeing the wood for the trees: a minimal reference phylogeny for the human Y chromosome. *Hum Mutat* 35: 187-91.
- Vigeland MD (2020) Relatedness coefficients in pedigrees with inbred founders. *J. Mathematical Biology* 81: 185-207. doi: doi: 10.1007/s00285-020-01505-x
- Vigeland MD, Egeland T (2019) Handling founder inbreeding in forensic kinship analysis. *Forensic Science International: Genetics Supplement Series* 7: 780–781. doi: <https://doi.org/10.1016/j.fsigss.2019.10.175>
- Walsh S (2013) Thesis: DNA Phenotyping: The Prediction of Human Pigmentation Traits from Genetic Data.
- Walsh S, Chaitanya L, Breslin K, Muralidharan C, Bronikowska A, Pospiech E, Koller J, Kovatsi L, Wollstein A, Branicki W, Liu F, Kayser M (2017) Global skin colour prediction from DNA. *Hum Genet* 136: 847–863. doi: 10.1007/s00439-017-1808-5
- Walsh S, Chaitanya L, Clarisse L, Wirken L, Draus-Barini J, Kovatsi L, Maeda H, Ishikawa T, Sijen T, de Knijff P, Branicki W, Liu F, Kayser M (2014) Developmental validation of the HIrisPlex system: DNA-based eye and hair colour prediction for forensic and anthropological usage. *Forensic Science International: Genetics* 9: 150–161.
- Walsh S, Kayser M (2016) A Practical Guide to the HIrisPlex System: Simultaneous Prediction of Eye and Hair Color from DNA. *Methods in molecular biology* 1420: 213-31 doi: [https://doi.org/10.1007/978-1-4939-3597-0\\_17](https://doi.org/10.1007/978-1-4939-3597-0_17)

- Walsh S, Liu F, Ballantyne KN, van Oven M, Lao O, Kayser M (2011) IrisPlex. A sensitive DNA tool for accurate prediction of blue and brown eye colour in the absence of ancestry information. *Forensic Sci. Int. Genet.* 5 170-180.
- Walsh S, Liu F, Wollstein A, Kovatsi L, Ralf A, Kosiniak-Kamysz AB, W., Kayser M (2013) The HIrisPlex system for simultaneous prediction of hair and eye colour from DNA. *Forensic Science International: Genetics* 7: 98–115.
- Walsh S, Wollstein A, Liu F, Chakravarthy U, Rahu M, Selander JH, Soubrane GT, L., Topouzis F, Vingerling JR, .., Vioque J, Fletcher AEB, K.N., Kayser M (2012) DNA-based eye colour prediction across Europe with the IrisPlex system. *Forensic Science International: Genetics* 6: 330–340.
- Wang B, Sverdlov S, Thompson E (2017) Efficient estimation of realized kinship from SNP genotypes. *Genetics* 205: 1063-1078. doi: <https://doi.org/10.1534/genetics.116.197004>
- Wang J (2002) An Estimator for Pairwise Relatedness Using Molecular Markers. *GENETICS* 160: 1203-1215.
- Waples RK, Albrechtsen A, and Moltke I (2019) Allele frequency-free inference of close familial relationships from genotypes or low-depth sequencing data. *Molecular Ecology* 28: 35-48. doi: 10.1111/mec.14954
- Weale ME (2010) Quality Control for Genome-Wide Association Studies. In: Barnes M., Breen G. (eds) Genetic Variation. Methods in Molecular Biology (Methods and Protocols). Humana Press, Totowa, NJ 628. doi: [https://doi.org/10.1007/978-1-60327-367-1\\_19](https://doi.org/10.1007/978-1-60327-367-1_19)
- Weir B, Anderson A, Hepler A (2006) Genetic relatedness analysis: modern data and new challenges. *Nat. Rev. Genet.* 7: 771–780. doi: <https://doi.org/10.1038/nrg1960>
- Weissensteiner H, Pacher D, Kloss-Brandstatter A, Forer L, Specht G, Bandelt HJ, Kronenberg, F., Salas A, Schönherr S (2016) HaploGrep 2: mitochondrial haplogroup classification in the era of high-throughput sequencing. *Nucleic Acids Research* 44. doi: doi: 10.1093/nar/gkw233
- Wendt FR, Churchill JD, Novroski NMM, King JL, Ng J, Oldt RF, McCulloh KL, Weise JA, Glenn Smith D, Kanthaswamy S, Budowle B (2016) Genetic analysis of the Yavapai native Americans from west-Central Arizona using the Illumina MiSeq FGx™ forensic genomics system. *Forensic Science International: Genetics* 24: 18–23. doi: <https://doi.org/10.1016/j.fsigen.2016.05.008>
- Werrett DJ (1997) The National DNA Database. *Forensic Sci Int* 88: 33-42.
- White JD, Indencleef K, Naqvi S, Eller RJ, Roosenboom J, Lee MK, al. e (2020) Insights into the genetic architecture of the human face. *bioRxiv*. doi: <https://doi.org/10.1101/s41588-020-00741-7>
- Wickenheiser RA (2019) Forensic genealogy, bioethics and the Golden State Killer case. *Forensic Science International: Synergy* 1: 114-125. doi: <https://doi.org/10.1016/j.fsisyn.2019.07.003>
- Willem T, Gymrek M, Highnam G, The 1000 Genomes Project Consortium M, D., Erlich Y (2014) The landscape of human STR variation. *Genome Res.* 24: 1894–1904. doi: doi: 10.1101/gr.177774.114
- Woerner A, King JL, Budowle B (2017a) Fast STR allele identification with STRait Razor 3.0. *Forensic Sci Int Genet.* 30: 18-23. doi: doi: 10.1016/j.fsigen.2017.05.008
- Woerner AE, King JL, Budowle B (2017b) Fast STR allele identification with STRait Razor 3.0. *Forensic Science International: Genetics* 30: 18–23. doi: <http://dx.doi.org/10.1016/j.fsigen.2017.05.008>

- Zeng X, King J, Hermanson S, J. P, Storts DR, Budowle B (2015) An evaluation of the PowerSeq™ Auto System: A multiplex short tandem repeat marker kit compatible with massively parallel sequencing. *Forensic Science International: Genetics* 19: 172-179. doi: <https://doi.org/10.1016/j.fsigen.2015.07.015>
- Zhang Q, Zhou Z, Wang L, Quan C, Liu Q, Tang Z, Liu L, Liu Y, Wang S (2020) Pairwise kinship testing with a combination of STR and SNP loci. *Forensic Science International: Genetics*. doi: <https://doi.org/10.1016/j.fsigen.2020.102265>
- Zhang S, Bian Y, Chen A, Zheng H, Gao Y, Hou Y, Li C (2017) Developmental validation of a custom panel including 273 SNPs for forensic application using Ion Torrent PGM. *Forensic Science International: Genetics* 27: 50-57. doi: <https://doi.org/10.1016/j.fsigen.2016.12.003>
- Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS (2012) A high-performance computing toolset for relatedness and principal component analysis of SNP data. *Bioinformatics* 28: 3326–3328. doi: <https://doi.org/10.1093/bioinformatics/bts606>
- Zhou Y, Browning SR, Browning BL (2020) A Fast and Simple Method for Detecting Identity-by-Descent Segments in Large-Scale Data. *Am J Hum Genet* 106: 426-437. doi: doi: 10.1016/j.ajhg.2020.02.010
- Ziętkiewicz E, Witt M, Daca P, Źebracka-Gala J, Goniewicz M, Jarząb B, Witt M (2012) Current genetic methodologies in the identification of disaster victims and in forensic analysis. *Journal of applied genetics* 53: 41-60.

# Electronic Appendices

The supplementary files are available on OneDrive [https://1drv.ms/u/s!Au78\\_aneAZRsgYcJyXD97MyipwI8w?e=gEn4hD](https://1drv.ms/u/s!Au78_aneAZRsgYcJyXD97MyipwI8w?e=gEn4hD), and after the processing of the thesis that will be available on <https://leicester.figshare.com/>.

Note: All appendix material is provided in the above link; Appendix 3a, Appendix 3b, Appendix 4a and Appendix 4c are also provided in text form here.

## Chapter 2: Appendix 2

- Appendix 2a. PCA plots of Family samples and subsets of the reference dataset (based on 1000 Genomes Project).
- Appendix 2b. PCA plot of unrelated individuals (identified via PRIMUS) from the Family samples (Family) and European set (1000 Genomes Project).
- Appendix 2c. PLINK PCA plots of 8 groups of unrelated individuals from the Family samples (Family) and 1000 Genomes Project (European metapopulation, French [CEU], CEU and Iberian [IBS]).
- Appendix 2d. Summary of the output from PRIMUS.

## Chapter 3: Appendix 3

- Appendix 3a. Table of autosomal STR Isoalleles resolved by sequencing.
- Appendix 3b. Table of Autosomal, X-chromosome and Y-chromosome STR sequences reported as “Novel” by STRait Razor v3 with no correspondence in STRSeq.
- Appendix 3c. Tables of Likelihood Ratios and Posterior probabilities of real pedigree using different population allele frequencies to explore the impact of different population-based allele frequencies on a real-world family set.
- Appendix 3d. Tables of Family samples allele description.

## Chapter 4: Appendix 4

- Appendix 4a. Background information: Importance sampling, Family Wise Error Rate.

- Appendix 4b. Tables of the allele frequencies used in Chapter 4.
- Appendix 4c. Kjelgaard Brustad, H., Colucci, M., Jobling, M.A., Sheehan, N.A., and Egeland, T. (2021). "Strategies for pairwise searches in forensic kinship analysis." *Forensic Sci Int Genet* 54: 102562.

## **Chapter 5: Appendix 5**

- Appendix 5a. Protocol: Chelex extraction.
- Appendix 5b. Standard error of the cross-validation error estimate to identify the best  $k$  in ADMIXTURE analysis.
- Appendix 5c. NevGen prediction of 30 Cape Verdean samples.
- Appendix 5d. Ancestry determination based on SNIPPER.
- Appendix 5e. Tables of Cape Verdean samples allele description.
- Appendix 5f. Tables of phenotypic prediction for the Cape Verdean samples.

### Appendix 3a Table of autosomal STR Isoalleles resolved by sequencing

Length homozygotes are reported with their nomenclature here.

Locus	Sample	Allele	DoC	Nomenclature
D12S391	F2P1	22	249	D12S391 [CE 22]-GRCh38-Chr12-12296981-12297168 [AGAT]13 [AGAC]9
			264	D12S391 [CE 22]-GRCh38-Chr12-12296981-12297168 [AGAT]13 [AGAC]8 AGAT
	F7P11	21	383	D12S391 [CE 21]-GRCh38-Chr12-12296981-12297168 [AGAT]13 [AGAC]8
			390	D12S391 [CE 21]-GRCh38-Chr12-12296981-12297168 [AGAT]12 [AGAC]8 AGAT
	F7P14	11	422	D13S317 [CE 11]-GRCh38-Chr13-82147986-82148107 [TATC]11 82148069-T
			783	D13S317 [CE 11]-GRCh38-Chr13-82147986-82148107 [TATC]11
	F7P6	11	380	D13S317 [CE 11]-GRCh38-Chr13-82147986-82148107 [TATC]11 82148069-T
			429	D13S317 [CE 11]-GRCh38-Chr13-82147986-82148107 [TATC]11
	F1P7	11	88	D13S317 [CE 11]-GRCh38-Chr13-82147986-82148107 [TATC]11 82148069-T
			120	D13S317 [CE 11]-GRCh38-Chr13-82147986-82148107 [TATC]11
	F7P5	11	463	D13S317 [CE 11]-GRCh38-Chr13-82147986-82148107 [TATC]11 82148069-T
			746	D13S317 [CE 11]-GRCh38-Chr13-82147986-82148107 [TATC]11
	F7P15	11	210	D13S317 [CE 11]-GRCh38-Chr13-82147986-82148107 [TATC]11 82148069-T
			253	D13S317 [CE 11]-GRCh38-Chr13-82147986-82148107 [TATC]11
	F3P7	12	146	D13S317 [CE 12]-GRCh38-Chr13-82147986-82148107 [TATC]12
			209	D13S317 [CE 12]-GRCh38-Chr13-82147986-82148107 [TATC]12 82148069-T

	F2P6	12	325	D13S317 [CE 12]-GRCh38-Chr13-82147986-82148107 [TATC]12
			404	D13S317 [CE 12]-GRCh38-Chr13-82147986-82148107 [TATC]12 82148069-T
	F3P2	12	148	D13S317 [CE 12]-GRCh38-Chr13-82147986-82148107 [TATC]12
			245	D13S317 [CE 12]-GRCh38-Chr13-82147986-82148107 [TATC]12 82148069-T
D16S539	F4P2	9	1278	D16S539 [CE 9]-GRCh38-Chr16-86352664-86352781 [GATA]9
			1325	D16S539 [CE 9]-GRCh38-Chr16-86352664-86352781 [GATA]9 86352761-C
D20S482	F4P5	14	240	D20S482 [CE 14]-GRCh38-Chr20-4525674-4525771 [AGAT]14 4525680-T
			303	D20S482 [CE 14]-GRCh38-Chr20-4525674-4525771 [AGAT]14
	F7P4	14	298	D20S482 [CE 14]-GRCh38-Chr20-4525674-4525771 [AGAT]14 4525680-T
			397	D20S482 [CE 14]-GRCh38-Chr20-4525674-4525771 [AGAT]14
	F6P6	14	439	D20S482 [CE 14]-GRCh38-Chr20-4525674-4525771 [AGAT]14 4525680-T
			530	D20S482 [CE 14]-GRCh38-Chr20-4525674-4525771 [AGAT]14
	F7P6	14	613	D20S482 [CE 14]-GRCh38-Chr20-4525674-4525771 [AGAT]14 4525680-T
			651	D20S482 [CE 14]-GRCh38-Chr20-4525674-4525771 [AGAT]14
	F2P6	14	554	D20S482 [CE 14]-GRCh38-Chr20-4525674-4525771 [AGAT]14 4525680-T
			620	D20S482 [CE 14]-GRCh38-Chr20-4525674-4525771 [AGAT]14
	F6P7	14	163	D20S482 [CE 14]-GRCh38-Chr20-4525674-4525771 [AGAT]14 4525680-T
			208	D20S482 [CE 14]-GRCh38-Chr20-4525674-4525771 [AGAT]14

	F4P19	14	451	D20S482 [CE 14]-GRCh38-Chr20-4525674-4525771 [AGAT]14 4525680-T
			626	D20S482 [CE 14]-GRCh38-Chr20-4525674-4525771 [AGAT]14
D21S11	F3P7	29	463	D21S11 [CE 29]-GRCh38-Chr21-19181939-19182111 [TCTA]4 [TCTG]6 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]11
			620	D21S11 [CE 29]-GRCh38-Chr21-19181939-19182111 [TCTA]6 [TCTG]5 [TCTA]3 TA [TCTA]3 TCA [TCTA]2 TCCA TA [TCTA]10
D2S1338	F2P2	17	1825	D2S1338 [CE 17]-GRCh38-Chr2-218014856-218014964 [GGAA]12 [GGCA]5
			1969	D2S1338 [CE 17]-GRCh38-Chr2-218014856-218014964 [GGAA]11 [GGCA]6
D2S441	F2P1	11	1836	D2S441 [CE 11]-GRCh38-Chr2-68011918-68012017 [TCTA]11 68011922-A
			1989	D2S441 [CE 11]-GRCh38-Chr2-68011918-68012017 [TCTA]11
D3S1358	F4P7	16	3287	D3S1358 [CE 16]-GRCh38-Chr3-45540691-45540820 TCTA [TCTG]3 [TCTA]12
			3322	D3S1358 [CE 16]-GRCh38-Chr3-45540691-45540820 TCTA [TCTG]2 [TCTA]13
	F6P5	16	2124	D3S1358 [CE 16]-GRCh38-Chr3-45540691-45540820 TCTA [TCTG]3 [TCTA]12
			2140	D3S1358 [CE 16]-GRCh38-Chr3-45540691-45540820 TCTA [TCTG]2 [TCTA]13
	F4P15	17	3045	D3S1358 [CE 17]-GRCh38-Chr3-45540691-45540820 TCTA [TCTG]2 [TCTA]14
			3591	D3S1358 [CE 17]-GRCh38-Chr3-45540691-45540820 TCTA [TCTG]3 [TCTA]13
	F2P6	17	2656	D3S1358 [CE 17]-GRCh38-Chr3-45540691-45540820 TCTA [TCTG]2 [TCTA]14
			3022	D3S1358 [CE 17]-GRCh38-Chr3-45540691-45540820 TCTA [TCTG]3 [TCTA]13
	F4P18	17	2786	D3S1358 [CE 17]-GRCh38-Chr3-45540691-45540820 TCTA [TCTG]2 [TCTA]14

			2831	D3S1358 [CE 17]-GRCh38-Chr3-45540691-45540820 TCTA [TCTG]3 [TCTA]13
F4P1	17	1875	D3S1358 [CE 17]-GRCh38-Chr3-45540691-45540820 TCTA TCTG [TCTA]15	
		2211	D3S1358 [CE 17]-GRCh38-Chr3-45540691-45540820 TCTA [TCTG]3 [TCTA]13	
F7P9	17	3758	D3S1358 [CE 17]-GRCh38-Chr3-45540691-45540820 TCTA [TCTG]3 [TCTA]13	
		4057	D3S1358 [CE 17]-GRCh38-Chr3-45540691-45540820 TCTA [TCTG]2 [TCTA]14	
D5S818	F5P5	12	48	D5S818-12-D5S818 [CE 12]-GRCh38-Chr5-123775543- 123775606 [ATCT]12 123775552-A- TATTATACATCTATCTATCTATCTATCTATCTATCTATCT ATCTATCTATCTATCTATCTTCAAAAT
		69	D5S818-12-D5S818 [CE 12]-GRCh38-Chr5-123775543- 123775606 [ATCT]12 - TATTATACCTCTATCTATCTATCTATCTATCTATCT ATCTATCTATCTATCTATCTTCAAAAT	
	F4P3	12	66	D5S818-12-D5S818 [CE 12]-GRCh38-Chr5-123775543- 123775606 [ATCT]12 - TATTATACCTCTATCTATCTATCTATCTATCTATCT ATCTATCTATCTATCTATCTTCAAAAT
		75	D5S818-12-D5S818 [CE 12]-GRCh38-Chr5-123775543- 123775606 [ATCT]12 123775552-A- TATTATACATCTATCTATCTATCTATCTATCTATCT ATCTATCTATCTATCTATCTTCAAAAT	
D7S820	F6P7	11	613	D7S820 [CE 11]-GRCh38-Chr7-84160191-84160297 [TATC]11 84160204-A
		648	D7S820 [CE 11]-GRCh38-Chr7-84160191-84160297 [TATC]11	
	F7P15	12	651	D7S820 [CE 12]-GRCh38-Chr7-84160191-84160297 [TATC]12
		721	D7S820 [CE 12]-GRCh38-Chr7-84160191-84160297 [TATC]12 84160204-A	
	F4P7	13	379	D7S820 [CE 13]-GRCh38-Chr7-84160191-84160297 [TATC]13 84160204-A
		533	D7S820 [CE 13]-GRCh38-Chr7-84160191-84160297 [TATC]13	

	F4P1	13	335	D7S820 [CE 13]-GRCh38-Chr7-84160191-84160297 [TATC]13
			372	D7S820 [CE 13]-GRCh38-Chr7-84160191-84160297 [TATC]13 84160204-A
D8S1179	F1P4	13	1843	D8S1179 [CE 13]-GRCh38-Chr8-124894867-124894921 [TCTA]13
			2096	D8S1179 [CE 13]-GRCh38-Chr8-124894867-124894921 TCTA TCTG [TCTA]11
	F4P16	13	1904	D8S1179 [CE 13]-GRCh38-Chr8-124894867-124894921 [TCTA]13
			2558	D8S1179 [CE 13]-GRCh38-Chr8-124894867-124894921 TCTA TCTG [TCTA]11
	F5P8	13	2185	D8S1179 [CE 13]-GRCh38-Chr8-124894867-124894921 [TCTA]13
			2232	D8S1179 [CE 13]-GRCh38-Chr8-124894867-124894921 TCTA TCTG [TCTA]11
	F6P8	13	1834	D8S1179 [CE 13]-GRCh38-Chr8-124894867-124894921 [TCTA]13
			2233	D8S1179 [CE 13]-GRCh38-Chr8-124894867-124894921 TCTA TCTG [TCTA]11
	F4P13	13	1457	D8S1179 [CE 13]-GRCh38-Chr8-124894867-124894921 TCTA TCTG [TCTA]11
			1471	D8S1179 [CE 13]-GRCh38-Chr8-124894867-124894921 [TCTA]13
D9S1122	F1P7	11	1962	D8S1179 [CE 13]-GRCh38-Chr8-124894867-124894921 TCTA TCTG [TCTA]11
			2090	D8S1179 [CE 13]-GRCh38-Chr8-124894867-124894921 [TCTA]13
	F7P2	11	645	D9S1122 [CE 11]-GRCh38-Chr9-77073809-77073880 [TAGA]11
			819	D9S1122 [CE 11]-GRCh38-Chr9-77073809-77073880 TAGA TCGA [TAGA]9
			582	D9S1122 [CE 11]-GRCh38-Chr9-77073809-77073880 [TAGA]11
			747	D9S1122 [CE 11]-GRCh38-Chr9-77073809-77073880 TAGA TCGA [TAGA]9

	F4P4	12	670	D9S1122 [CE 12]-GRCh38-Chr9-77073809-77073880 [TAGA]12
			717	D9S1122 [CE 12]-GRCh38-Chr9-77073809-77073880 TAGA TCGA [TAGA]10
F5P3	12	1281	D9S1122 [CE 12]-GRCh38-Chr9-77073809-77073880 TAGA TCGA [TAGA]10	
			1394	D9S1122 [CE 12]-GRCh38-Chr9-77073809-77073880 [TAGA]12
F2P5	12	1119	D9S1122 [CE 12]-GRCh38-Chr9-77073809-77073880 TAGA TCGA [TAGA]10	
			1356	D9S1122 [CE 12]-GRCh38-Chr9-77073809-77073880 [TAGA]12
F5P6	12	764	D9S1122 [CE 12]-GRCh38-Chr9-77073809-77073880 TAGA TCGA [TAGA]10	
			1063	D9S1122 [CE 12]-GRCh38-Chr9-77073809-77073880 [TAGA]12
F6P6	12	445	D9S1122 [CE 12]-GRCh38-Chr9-77073809-77073880 TAGA TCGA [TAGA]10	
			487	D9S1122 [CE 12]-GRCh38-Chr9-77073809-77073880 [TAGA]12
F5P5	12	554	D9S1122 [CE 12]-GRCh38-Chr9-77073809-77073880 TAGA TCGA [TAGA]10	
			864	D9S1122 [CE 12]-GRCh38-Chr9-77073809-77073880 [TAGA]12
F8P3	12	782	D9S1122 [CE 12]-GRCh38-Chr9-77073809-77073880 TAGA TCGA [TAGA]10	
			904	D9S1122 [CE 12]-GRCh38-Chr9-77073809-77073880 [TAGA]12
F3P2	12	625	D9S1122 [CE 12]-GRCh38-Chr9-77073809-77073880 [TAGA]12	
			645	D9S1122 [CE 12]-GRCh38-Chr9-77073809-77073880 TAGA TCGA [TAGA]10
F4P19	12	1197	D9S1122 [CE 12]-GRCh38-Chr9-77073809-77073880 TAGA TCGA [TAGA]10	
			1416	D9S1122 [CE 12]-GRCh38-Chr9-77073809-77073880 [TAGA]12

	F3P3	12	649	D9S1122 [CE 12]-GRCh38-Chr9-77073809-77073880 TAGA TCGA [TAGA]10
			681	D9S1122 [CE 12]-GRCh38-Chr9-77073809-77073880 [TAGA]12
	F5P2	12	667	D9S1122 [CE 12]-GRCh38-Chr9-77073809-77073880 TAGA TCGA [TAGA]10
			690	D9S1122 [CE 12]-GRCh38-Chr9-77073809-77073880 [TAGA]12
	F4P16	13	972	D9S1122 [CE 13]-GRCh38-Chr9-77073809-77073880 TAGA TCGA [TAGA]11
			1035	D9S1122 [CE 13]-GRCh38-Chr9-77073809-77073880 [TAGA]13
vWA	F3P8	16	177	vWA [CE 16]-GRCh38-Chr12-5983950-5984049 [TAGA]12 [CAGA]3 TAGA
			182	vWA [CE 16]-GRCh38-Chr12-5983950-5984049 [TAGA]11 [CAGA]4 TAGA

**Appendix 3b Table of Autosomal, X-chromosome and Y-chromosome STR sequences reported as “Novel” by STRait Razor v3 with no correspondence in STRSeq**

Locus	Sample	Allele (CE equivalent)	Novel sequence nomenclature
D10S1248	HGDP00880	15	TTGAACAAATGAGTGAGAGGAAGGAAGGAAGGAAGGAAGGAA GGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAATGAAGACAATACA ACCAGAGTT
D12S391	HGDP00522, HGDP00885, HGDP01372, HGDP00530, HGDP01169	23	CAGAGAGAAGAACAGGATCAATGGATGCATAGGTAGATAGAT AGATAGATAGATAGATAGATAGATAGATAGATAGACAGA CAGACAGACAGACAGACAGACAGACAGACAGACAGACAGACAG TTATTAGAGGAATTAGCTCAAGTGTATGGAGGCTGAAAATCTCATG ACAGTCCATCTGCAA
			CAGAGAGAAGAACAGGATCAATGGATGCATAGGTAGATAGAT AGATAGATAGATAGATAGATAGATAGATAGATAGACAGA CAGACAGACAGACAGACAGACAGACAGACAGACAGACAG TTATTAGAGGAATTAGCTCAAGTGTATGGAGGCTGAAAATCTCATG ACAGTCCATCTGCAA
			CAGAGAGAAGAACACAGGATCAATGGATGCATAGGTAGATAGAT AGATAGATAGATAGATAGATAGATAGATAGATAGACAGA CAGACAGACAGACAGACAGACAGACAGACAGACAGACAG TATTAGAGGAATTAGCTCAAGTGTATGGAGGCTGAAAATCTCATG CAGTCCATCTGCAA
D16S539	HGDP00533	12.1	TCTCTTCCTAGATCAATAACAGACAGACAGAGGTGGATAGATAGA TAGATAGATAGATAGATAGATAGATAGATAGATAATCATTGAAA GACAAACAGAGATGGATGATAGATAC
D20S482	HGDP01071, HGDP01076	14	AGACACTGAACCAATAATAGATAGATAGATAGATAGATAGATAG ATAGATAGATAGATAGATAGATAGATAGATAGAGATTATTAGAATTGA TT
D21S11	HGDP00531	35.2	AAATATGTGAGTCATTCCCCAAGTGAATTGCCCTATCTATCTATCT ATCTATCTATCTGTCTGTCTGTCTGTCTATCTATCTATCTATCTA TCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCT TATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTAT
D2S1338	HGDP00883, HGDP00522	21	GAGGGAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGA AGGAAGGAAGGGAGGCAGGCAGGCAGGCAGGCAGGCAGGCAGGCC AAGCCATT
	HGDP00881	15	GAGGGAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGCAGGC AGGCAGGCAGGCAGGCAGGCAGGCCAGCCATT
D2S441	HGDP00515	8	CCAGGAACGTGGCTCATCTATGAAAACCTCTATCTATCTATCTA TCTATCTGTCTATCATAACACACCACAGCCACTTA
	HGDP00530	12	CCAGAAACTGTGGCTCATCTATGAAAACCTCTATCTATCTATCTA TCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCT A
D4S2408	HGDP00528	7	CTATGCATCTATCTATCTATCTATCTATCTATCTAATGGTTA
D6S1043	HGDP01157	11	AGATGGCATATTGTGAAATTCCGAGCTTCCATAATTGTATGAGCCACT TCCATAATAAAATCTATCTATCTATCTATCTATCTATCTATCTATCT CTATCTATCTGATCTATCTAATCTATTGATC
D7S820	HGDP00531	14	TATTTAGTGGAGATAAAAAAAATCAATCTGTCTATCTATCTATCTA TCTATCTATCTATCTATCTATCTATCTATCTATCTATCGTTAGTCG TTCTAAACTAT

D9S1122	HGDP00674	11	AGATAACTGTAGATAGATAGATAGATAGATAGATAGATAGATAG ATAGATAGATAGATATTAAAT
FGA	HGDP00879	16	CCAGCAAAAAGAAAGGAAGAAGGAAGGGAGGAGAAAGAAAGAAA GAAAGAAAGAAGAAAGAAGAGAAAAAGAAAGAAAGAAAGAAA
	HGDP01166	20.2	CCAGCAAAAAGAAAGGAAGAAGGAAGGGAGGAGAAAGAAAGAAA GAAAGAAAGAAGAAAGAAGAGAAAGAAAGAGAAAGAAAGAAAA AGAAAGAAAGAAA
PentaD	HGDP00794	13	GATCACTTGAGCCTGGAAGGTCGAAGCTGAAGTGGACATGATCACAC CACTACACTCCAGCCTAGGTGACAGAGCAAGACACCCTCAAGAAC AAAAAAAAGAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAAAGAAA AGAAAAGAAAAGAAAAGAAAAGACAAACGAA
DXS10074	HGDP00672, HGDP01078	16	TGTGTGTGCATGCATACACACACAGAGAGAGAGAGAGAGAGAAAGAA GAAAGAAAGAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAA AAGAAAGGAAGAAAGAAAGGAAGAAAATAGAACAAATCAGTTATAT TCAGTATTTTTAGTATTTCTGTGTCAGCTC
DXS10135	HGDP00665	16.1	AAGAAAGAAAGAGAAAGGAAGAAAGAAAGAAAGAAAGAAAGAAAAG AAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAAG AGAAAAGAGAAAAGAAAAAGAAAAGAAA
	HGDP00800	19	AAGAAAGAAAGAGAAAGGAAGAAAGAAAGAAAGAAAGAAAGAAAAG AAAGAAAGAAAGAAAGAAAGGAAGGAAGGAAGGAAGAAAAGAGAATAGAA AAGAAGAGAAGAGAAAAGAGAAAAGAAAAAGAAAAGAAA
	HGDP00528	20	AAGAAAGAAAGAGAAAGGAAGAAAGAAAGAAAGAAAGAAAGAAAAG AAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGGAAGGAAGAAAAGAGAAT AGAAAAGAAGAGAAGAGAAAAGAGAAAAGAAAAAGAAAAGAAA
	HGDP00885	22	AAGAAAGAAAGAGAAAGGAAGAAAGAAAGAAAGAAAGAAAGAAAAG AAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGGAAGGAAGGA AAGAGAATAGAAAAGAAGAGAAGAGAAAAGAGAAAAGAAAAAGAAA AAGAAA
	HGDP00880	25	AAGAAAGAAAGAGAAAGGAAGAAAGAAAGAAAGAAAGAAAGAAAAG AAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAA AGAAAAGAAAGAAAGGAGAATAGAAAAGAAGAGAAGAGAAAAGAGAA AAGAAAAAAGAAAAGAAA
	HGDP00805	27	AAGAAAGAAAGAGAAAGGAAGAAAGAAAGAAAGAAAGAAAGAAAAG AAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGGAAGGA AGAAAGAAAGAAAGGAAGAAAAGAGAATAGAAAAGAAGAGAAGAGA AAAGAGAAAAGAAAAAGAAAAGAAA
	HGDP00522	29	AAGAAAGAAAGAGAAAGGAAGAAAGAAAGAAAGAAAGAAAGAAAAG AAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAA AGGAAGAAAGAAAGAAAGGAAGAAAAGAGAATAGAAAAGAAGAGAAG AGAAGAGAAAAGAGAAAAGAAAAAGAAAAGAAA
DXS8378	HGDP00885	31	AAGAAAGAAAGAGAAAGGAAGAAAGAAAGAAAGAAAGAAAGAAAAG AAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAA AGAAAGAAAGAAAGAAAGGAAGGAAGGAAGGAAGAAAAGAGAATAG AAAAGAAGAGAAGAGAAAAGAGAAAAGAAAAAGAAAAGAAAAGAAA
		15	AGTGAGCTGAGATGGTGCCTGAACCTCCAGCCTGGGCACAAGAGCG AAACTCCAACTAaaaaATAATAATAATAATAGATAGATAGATAG ATAGATAGATAGATAGATAGATAGATAGATAGATAGATAGTGAC CTGCCAGGAGCAGGGACCCGGGTTGCTTAAGGAGGGTGAACTGT CCCAGGATGGAATGAAACA
DYF387S1	HGDP00528	40	GAAGAAAGAGAAAAAAAGAAAGAAAGGTAGGAAGGAAGGAAGGAAG AAAGAAAGGAAGAAAGAAAGGAAGGAAGGAAGGAAGGAAGGAAGGA AGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAG AAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAAG
	HGDP00528	42	GAAGAAAGAGAAAAAAAGAAAGAAAGGTAGGAAGGAAGGAAGGAAG AAAGAAAGGAAGAAAGAAAGGAAGGAAGGAAGGAAGGAAGGAAGGA AGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAGGAAG AAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAGAAAG
DYS385	HGDP00880, HGDP00882	9	TTCTTTCTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTT CTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTT CTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTT CTTCTTCTTCTGAAATTCTTCTTCTTCTTCTTCTTCTTCTTCTTCTT



	HGDP01358	22	ATCAATCAATGAATGGATAAAGAAAATGTGATAGATAGATAGATAGAT AGATAGATAGATAGATAGATAGATAGATAGATACATAGATAGAT ACATACATAGATAGATAGAGATTCTATGCAAAGTGAGAACCCA
DYS643	HGDP01077	12	TGATTTTGCAGGTGTTCACTGCAAGCCATGCCATCGCTGGTTAAACTACTGTG CCTTTCTTTCTTTCTTTCTTTCTTTACTTTCTTTCTTTCTTTCT TTCTTTCTTTCTTTAAAACCTT
YGATAH4	HGDP01069	14	CTATCTATCTATCTATTCTATCCATCTAAATCTATCCATTCTATCTATCTAT CTATCTATCTATCTATCTATCTATCTATCTATCTATCTATCTACCTACCT ACCTATCTATCTATAGATCTATCTATCTATCT

## Appendix 4a. Background information: Importance sampling, Family Wise Error Rate.

### Importance sampling

Importance sampling is a method that can be used to approximate small probabilities. We first introduce the indicator function,  $I(LR > t)$ , that takes on the value 1 when the LR is above the threshold  $t$  and 0 otherwise. The expected value of the function  $I$  becomes

$$E(I(LR > t)) = 0 \cdot P(LR \leq t) + 1 \cdot P(LR > t) = P(LR > t) = FPR. \quad (2)$$

It is therefore valid to say that  $FPR = E(I(LR > t))$ . Then consider the expression for the expected value in a more general sense. The value of the function  $I$  is dependent on the value of the LR, which is a function of the genotypes  $G$  of the samples. The probability distribution of  $G$  is governed by the relationships that has generated the data. For this consideration, we assume that this relationship is either  $H_P$  or  $H_D$ . Denote by  $X$  the values that  $I$  can take on. We then have

$$E(I(LR > t)) = \sum_j^D X_j \cdot P(G_j | H_D) \approx \frac{1}{D} \sum_{i=1}^D I(LR_i^{H_D} > t)$$

In the last expression, the expected value is estimated by the sample mean of  $I$ . The genotypes  $G$ , and then also  $X$ , are distributed according to  $H_D$ , which is denoted by the superscript of the LR. Then, consider the opposite probability distribution,  $P(G | H_P)$ , where the genotypes are distributed according to  $H_P$ . As long at  $P(G_j | H_D) = 0$  whenever  $P(G_j | H_P) = 0$ , we can write

$$E(I(LR > t)) = \sum_j^D X_j \cdot \frac{P(G_j | H_D)}{P(G_j | H_P)} P(G_j | H_P) \approx \frac{1}{D} \sum_{i=1}^D \frac{I(LR_i^{H_P} > t)}{LR_i^{H_P}}$$

Using this method, the LR is sampled under the wrong hypothesis ( $H_P$ ), instead of the desired hypothesis ( $H_D$ ). The bias this introduces is adjusted for by the weight  $LR_i^{H_P}$ . An estimate of  $FPR$  is then

$$\widehat{FPR} = \frac{1}{D} \sum_{i=1}^D \frac{I(LR_i^{H_P} > t)}{LR_i^{H_P}}$$

## Family Wise Error Rate

FWER is the probability of getting at least one false positive among  $N$  LRs in a blind search,  $P(V \geq 1)$ . If the LRs are assumed to be independent, FWER is given as

$$\begin{aligned} FWER &= P(V \geq 1) = 1 - P(V = 0) \\ &= 1 - P(LR_i \leq t)^N = 1 - (1 - P(LR_i > t))^N = 1 - (1 - FPR)^N. \end{aligned}$$

Here,  $FPR$  is still defined as  $P(LR > t)$  for a specified threshold  $t$ . The independence assumption does not hold for a blind search. In order to find an upper limit for FWER, apply Bonferroni's inequality. Denote by  $A_i$ , the event of getting a false positive, i.e.  $LR_i > t$ . An upper limit for the probability of getting one or more false positives out of  $N$  pairs in a blind search is then

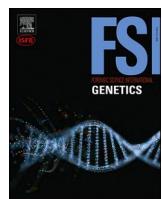
$$P(\bigcup_{i=1}^N A_i) \leq \sum_{i=1}^N P(A_i),$$

which means that

$$FWER \leq N \cdot FPR = \frac{n(n-1)}{2} FPR.$$

**Appendix 4c. Kjelgaard Brustad, H., Colucci, M., Jobling, M.A., Sheehan, N.A., and Egeland, T. (2021). "Strategies for pairwise searches in forensic kinship analysis." Forensic Sci Int Genet 54: 102562.**

(See next page)



## Research paper

## Strategies for pairwise searches in forensic kinship analysis



Hilde Kjelgaard Brustad <sup>a,b,\*</sup>, Margherita Colucci <sup>c</sup>, Mark A. Jobling <sup>c</sup>, Nuala A. Sheehan <sup>d</sup>,  
Thore Egeland <sup>a</sup>

<sup>a</sup> Faculty of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Post box 5003 NMBU, 1432 Aas, Norway

<sup>b</sup> Oslo Centre for Biostatistics and Epidemiology, University of Oslo, Post box 1122 Blindern, 0316 Oslo, Norway

<sup>c</sup> Department of Genetics & Genome Biology, University of Leicester, University Road, Leicester LE1 7RH, UK

<sup>d</sup> Department of Health Sciences, University of Leicester, University Road, Leicester LE1 7RH, UK

## ARTICLE INFO

**Keywords:**  
Kinship testing  
Blind search  
LR thresholds  
Inbred relationships  
X chromosomal markers

## ABSTRACT

Testing kinship between pairs of individuals is central to a wide range of applications. We focus on cases where many tests are done jointly. Typical examples include cases where DNA profiles are available from a burial site, a plane crash or a database of convicted offenders. The task is to determine the relationships between DNA profiles or individuals. Our approach generalises previous methods and implementations in several respects. We model general, possibly inbred, pairwise relationships which is important for non-human applications and in archaeological studies of ancient inbred populations. Furthermore, we do not restrict attention to autosomal markers. Some cases, such as distinguishing between maternal and paternal half siblings, can be solved using X-chromosomal markers. When many tests are done, the risk of errors increases. We address this problem by building on the theory of multiple testing and show how optimal thresholds for tests can be determined. We point out that the likelihood ratios in a blind search may be dependent so multiple testing methods and interpretation need to account for this. In addition, we show how a Bayesian approach can be helpful. Our examples, using simulated and real data, demonstrate the practical importance of the methods and implementation is based on freely available software.

## 1. Introduction

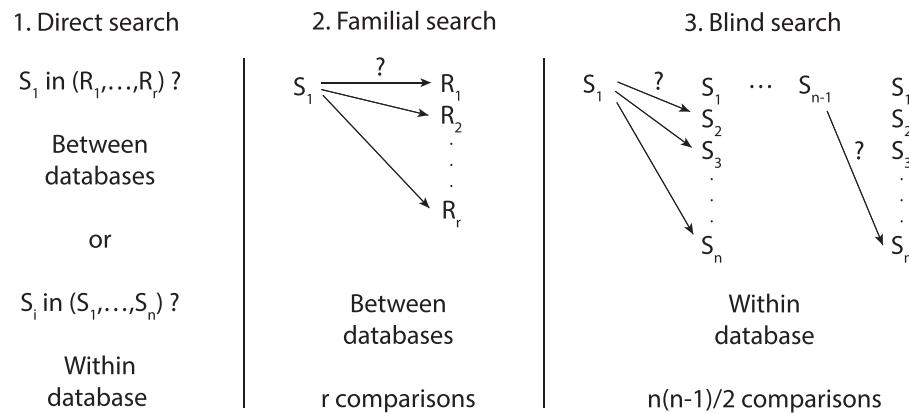
Inferring the relationship between pairs of individuals is central to many forensic applications. Examples include mass fatality incidents, which can be the result of accidental catastrophes like air crashes with a list of known victims [1] or shipwrecks without passenger lists [2,3]. Other applications are natural disasters like tsunamis, where the number of victims is unknown [4] and terrorism-related events [5]. The aim is to link DNA samples from the scene to putative victims (e.g. individuals reported missing since the event) and is known as disaster victim identification (DVI). There are various other important applications like searching for relationships among individuals in mass graves of archaeological relevance [6–8]. We may also check databases collected to estimate population statistics such as allele frequencies. Duplicates and close relatives should be excluded prior to the statistical analysis to avoid biased estimates of allele frequencies [9].

As these cases involve unidentified DNA samples, a first step in the investigation is to screen the data for related samples. This initial step is

referred to as a blind search [10]. It is helpful to first position the topics that we are addressing in the wider context of database searching. Assume that there is a case database of DNA profiles. This could comprise profiles obtained from a crime scene, a disaster site or a burial site. In addition, there may be a reference database of DNA profiles like a national database of convicted offenders. There are various searches that can be performed to detect pairwise relationships as illustrated in Fig. 1: .

1. Search for duplicates, i.e., direct search, performed within or between the databases. If this is done within a database, the objective is to merge identical samples. A search between databases corresponds to the widely discussed database search problem [11].
2. Familial searching involves searching between databases [12]. A selected DNA profile is compared to the profiles of a database with the aim of detecting close kin relationships, such as parent-offspring or sibling rather than a direct match.

\* Corresponding author at: Faculty of Chemistry, Biotechnology and Food Science, Norwegian University of Life Sciences, Post box 5003 NMBU, 1432 Aas, Norway.  
E-mail address: [h.k.brustad@medisin.uio.no](mailto:h.k.brustad@medisin.uio.no) (H.K. Brustad).



**Fig. 1.** Different database searches. 1. Direct search: Search for direct matches between or within databases. 2. Familial search: Search for related individuals between databases. 3. Blind search: search for related individuals within databases.

3. Blind search. This is a search within a database and is the topic of this paper.

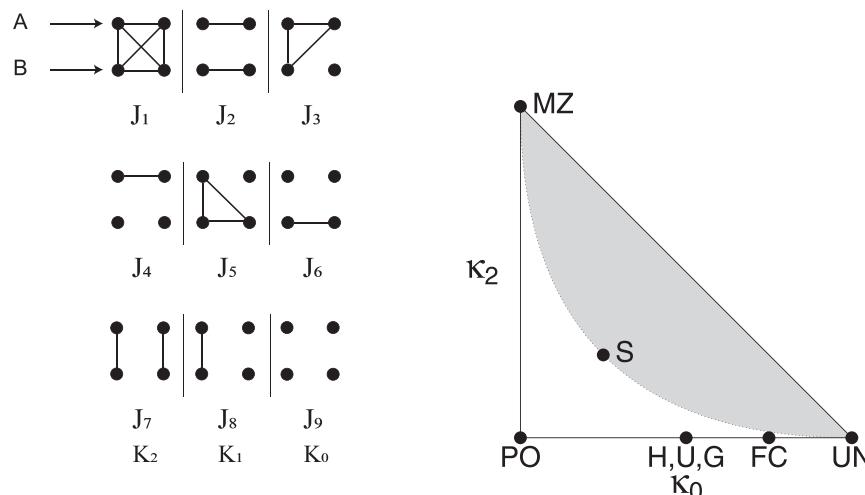
In a blind search, comparisons are performed among all pairs of DNA samples. A likelihood ratio (LR), comparing the relationship specified by  $H_1$  to the one specified by  $H_0$ , is computed for each pair. The LRs summarise the statistical DNA evidence. For pre-specified threshold values  $t_0$  and  $t_1$ , small values of  $LR < t_0$  are often interpreted as supporting  $H_0$ , while large values of  $LR > t_1$  favour  $H_1$ . A blind search typically involves a large number of comparisons. If there are  $n$  profiles in the database, the number of comparisons is  $n(n-1)/2$ , e.g. 4950 comparisons for 100 profiles. The implications of this high number of pairwise comparisons in a blind search are of key concern in this paper. Also, it is not obvious how the thresholds  $t_0$  and  $t_1$  should be specified. Conventional thresholds used in paternity testing, for example, may not apply. The false positive rate  $FPR = P(LR > t_1 | H_0)$  and false negative rate  $FNR = P(LR < t_0 | H_1)$  should both be close to 0. Even if these error rates are small for each comparison, the probability that errors occur when many comparisons are done may be considerable. Determination of thresholds and optimisation of search strategies have been discussed in connection with database searches and familial searching [13]. The classical statistical theory of multiple testing [14] is also relevant.

Current implementations of blind search are limited to fairly simple

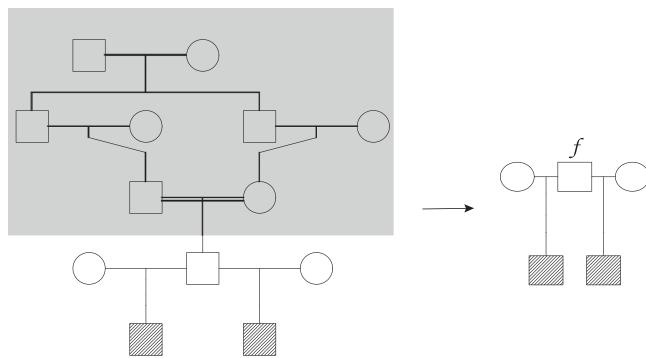
outbred pedigree structures connecting the two individuals of interest. For example, Familias [10], a freely available kinship software package, accommodates parent offspring (PO), sibling (S), half sibling (H), first cousin (FC) and second cousin (SC) [15,16]. We model general pairwise relationship, possibly with inbreeding, using the Jaccard coefficients [17]. By including X-chromosomal markers, some additional relationships can be addressed. For instance, paternal and maternal half sisters can be distinguished.

Prior, non-DNA, information can sometimes be important. For instance, two individuals of the same age cannot possibly constitute a parent-offspring pair even if the DNA profiles suggest otherwise. To formally include prior information, we require a Bayesian approach. In the Bayesian framework we start out with a set of prior probabilities, reflecting our belief in the hypotheses, before considering any genetic data. Our belief in each hypothesis is then updated by incorporating the DNA information. Informative priors can contribute additional information to the genetic data and this will be reflected in the posterior probabilities. A more general prior distribution for pedigrees has been discussed elsewhere [18].

Our paper is structured as follows. We first review the parametric representation of relationships and the corresponding parametric likelihood and likelihood ratio, for both autosomal and X-chromosomal markers. A review of the Bayesian approach to kinship testing is given,



**Fig. 2.** Left: Jacquard states  $J_1, \dots, J_9$ . Dots denote alleles, and lines connect IBD alleles. Right: IBD triangle, with location of some common relationships. Abbreviations: MZ - monozygotic twins, PO - parent offspring, S - full siblings, H - half siblings, U - avuncular, G - grandparent grandchild, FC - first cousins, UN - unrelated.



**Fig. 3.** Figure showing the concept of founder inbreeding, as described in Section 2.1.1. The shaded part showing the first cousin relationship, is modelled by the inbreeding coefficient  $f$ .

before we return to the likelihood ratio and its properties. These properties are incorporated when presenting the theory for evaluating the performance of a blind search. We then introduce the data used in the results section and give a brief description of our implementation. We provide several examples and conclude with a discussion of the challenges and the advantages of the work we present.

## 2. Review of previous results

### 2.1. Relatedness coefficients

Two homologous alleles are *identical by descent* (IBD) if they are identical by state (IBS) and inherited from a common ancestor. IBD is therefore defined with reference to a specified pedigree. The idea is that closely related individuals share more of their genetic material IBD than more distantly related individuals.

The simplest measure of pairwise relationships is the kinship coefficient,  $\varphi$ , defined as the probability that a random allele at a locus from one individual is IBD to a random allele at the same locus from another individual. This is the same as the inbreeding coefficient  $f$  of a child of these two individuals [19].

The Jacquard coefficients [17] provide a description of general pairwise relationships. The four alleles of two individuals are in one of the nine Jacquard states  $J_i$  for  $i = 1, \dots, 9$  (see left panel of Fig. 2). The probability that the alleles at a locus are in the different Jacquard states are given by the Jacquard coefficients,  $\Delta = (\Delta_1, \dots, \Delta_9)$ , where  $\Delta_i = P(J_i)$ . The coefficients sum to one.

The first six Jacquard states model inbreeding in one or both of the individuals. The only possible IBD states for two outbred individuals are  $J_9$ ,  $J_8$  and  $J_7$ , referred to as the IBD states  $K_0$ ,  $K_1$  and  $K_2$ , respectively. Thus, for two outbred individuals, the Jacquard coefficients reduce to the IBD coefficients [20],  $\kappa = (\kappa_0, \kappa_1, \kappa_2)$ , where  $\kappa_i = P(K_i)$ . Since  $\sum_{i=0}^2 \kappa_i = 1$ , the coefficients can be visualised in the IBD triangle, with coordinates  $(\kappa_0, \kappa_2)$ . Fig. 2 shows the IBD triangle with the location of some common pedigree relationships.

Thompson [21] showed that the coefficients satisfy the inequality  $\kappa_1^2 \geq 4\kappa_0\kappa_2$ , which creates an inadmissible region, shown in grey in Fig. 2. This means that it is not possible to construct a pedigree connecting two outbred individuals with IBD coefficients in the inadmissible region.

#### 2.1.1. Relatedness coefficients and founder inbreeding

By assigning a coefficient of inbreeding to one or more of the founders of a pedigree, background relatedness can be modelled [22]. Inbreeding of a pedigree founder (or several founders) affects the genetic relationship between other members of the pedigree [23], but does not necessarily make the pedigree members of interest inbred. For example, if it is suspected that two individuals are paternal half-siblings

**Table 1**

Likelihood of  $\kappa = (\kappa_0, \kappa_1, \kappa_2)$ , when observing genotypes for two individuals, for three unlinked loci, as described in the example of Section 2.3.

	$P(G K_0)$	$P(G K_1)$	$P(G K_2)$	$L(\kappa)$
$G_1 = (ab, ac)$	0.06	0.03	0	$0.06 \cdot \kappa_0 + 0.03 \cdot \kappa_1$
$G_2 = (bc, bb)$	0.011	0.004	0	$0.011 \cdot \kappa_0 + 0.004 \cdot \kappa_1$
$G_3 = (aa, bc)$	0.03	0	0	$0.03 \cdot \kappa_0$

and the paternal grandparents are first cousins, as depicted in Fig. 3, the common father has an inbreeding coefficient  $f = 1/16$ . The IBD coefficients for these half siblings are given by  $\kappa = (0.469, 0.531, 0)$  in contrast with  $\kappa = (0.5, 0.5, 0)$  for the non-inbred setting with  $f = 0$ . It can be shown that there is some finite pedigree with founder inbreeding that corresponds to each admissible point in the IBD triangle [24].

### 2.2. The likelihood function

Our data comprise pairs of DNA profiles, genotyped at  $m$  unlinked loci, i.e.,  $m$  statistically independent loci. For a single pair of individuals, A and B, let  $G_j = (g_{A,j}, g_{B,j})$  denote their respective genotypes at locus  $j$  for  $j = 1, \dots, m$ . The likelihood of  $\Delta$ , i.e., the probability of observing the data  $G = (G_1, \dots, G_m)$  assuming  $\Delta$  to be true, is

$$L(\Delta) = \prod_{j=1}^m \sum_{i=1}^9 \Delta_i P(G_j|J_i). \quad (1)$$

The probabilities  $P(G_j|J_i)$  are given in Table 9 in the Appendix A. For outbred individuals, the likelihood of  $\kappa$  is

$$L(\kappa) = \prod_{j=1}^m \sum_{i=0}^2 \kappa_i P(G_j|K_i), \quad (2)$$

where the probabilities  $P(G_j|K_i)$  for  $i = 0, 1, 2$  correspond to the last three columns of Table 9.

### 2.3. Parametric representation of the likelihood ratio

The likelihood ratio (LR) quantifies how much more likely it is that a set of genetic data is explained by one hypothesis  $H_1$  than by another hypothesis  $H_0$ . In our applications, each hypothesis states a pairwise relationship, expressed by a set of relatedness coefficients  $\Delta$  (or  $\kappa$  for outbred relationships). The LR that compares (1) for two sets of coefficients  $\Delta_1$  and  $\Delta_0$  is

$$LR(\Delta_1, \Delta_0) = \frac{P(G|H_1)}{P(G|H_0)} = \frac{L(\Delta_1)}{L(\Delta_0)}. \quad (3)$$

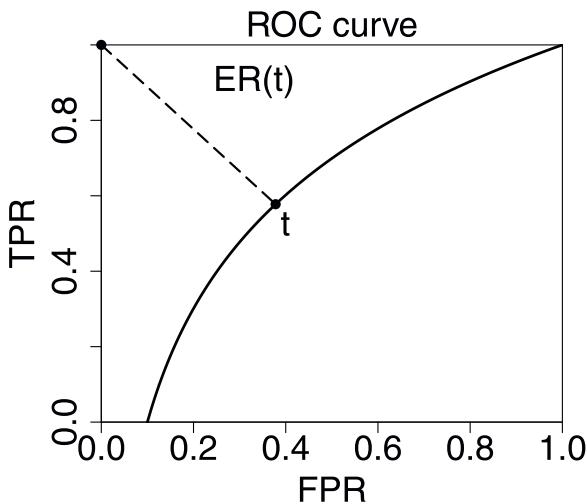
The hypotheses  $H_1$  and  $H_0$  are not necessarily exhaustive, meaning that there may be other hypotheses that better explain the data.

#### 2.3.1. Example

The purpose of this example is merely to illustrate how LRs can be easily computed for different sets of IBD coefficients using the representation in (3).

Consider two individuals genotyped at three loci. Each locus has three alleles  $a$ ,  $b$  and  $c$ , with population frequencies 0.5, 0.3 and 0.2, respectively. The genotypes at each locus are given in the first column of Table 1. The likelihood of  $\kappa$  for each locus are given in the last column.

When comparing siblings,  $\kappa_1 = (0.25, 0.5, 0.25)$  against unrelated (UN)  $\kappa_0 = (1, 0, 0)$ , the LR becomes



**Fig. 4.** Figure showing the concept of a ROC curve with a corresponding threshold. The rate TPR is plotted as a function of FPR. Each point on the curve corresponds to an LR threshold  $t$ . The dashed line shows the Euclidean distance (unweighted) from the optimal point  $(0,1)$  to the ROC curve, given by (7).

$$\begin{aligned} \text{LR}(\kappa_1, \kappa_0) = & \frac{(0.06 \cdot 0.25 + 0.03 \cdot 0.5)}{0.06} \\ & \frac{(0.011 \cdot 0.25 + 0.004 \cdot 0.5)}{0.011} \\ & \frac{(0.03 \cdot 0.25)}{0.03} = 0.054. \end{aligned} \quad (4)$$

Similarly, we find the LR for half siblings (or avuncular or grandparent grandchild),  $\kappa_1 = (0.5, 0.5, 0)$ , against unrelated,  $\kappa_0 = (1, 0, 0)$ , to be 0.256. The probabilities in the middle three columns of Table 1 are independent of the tested relationships.

#### 2.4. Properties of the LR

For specified thresholds  $t_0 < t_1$ , an  $\text{LR} < t_0$  essentially supports  $H_0$ , while an  $\text{LR} \geq t_1$  favours  $H_1$ . More data may be required to make a decision when  $t_0 \leq \text{LR} < t_1$  [25]. For simplicity, we will assume  $t_0 = t_1 = t$ , so that a conclusion can always be drawn.

When  $\text{LR} \geq t$ , but  $H_0$  is true, we have a *false positive* (FP). If  $\text{LR} \geq t$  and  $H_1$  is true, we have a *true positive* (TR). We define the *false positive rate* (FPR) and the *true positive rate* (TPR) as

$$\text{FPR} = P(\text{LR} \geq t | H_0), \quad \text{TPR} = P(\text{LR} \geq t | H_1). \quad (5)$$

The TPR measures the ability to detect the relationship, while the FPR is the probability of falsely declaring a relationship. The relationship between FPR and TPR is often visualised by a receiver operating characteristic (ROC) curve [26]. Fig. 4 in Section 3.3 illustrates the concept of a ROC curve.

#### 2.5. The Bayesian approach to kinship testing

A frequentist approach to evaluating kinship is based on the LR reflecting the probabilities of the data we have observed under two specified hypotheses. An alternative approach is provided by a Bayesian framework.

Instead of just testing one hypothesis  $H_1$  against  $H_0$ , we consider a set of hypotheses  $H_i$ ,  $i = 1, \dots, k$ , each against  $H_0$ . With some prior belief in each hypothesis  $\pi_0, \dots, \pi_k$ , with  $\sum_{i=0}^k \pi_i = 1$ , Bayes' theorem expresses the posterior probability of each hypothesis as

$$P(H_i | \text{data}) = \frac{\text{LR}_i \pi_i}{\sum_{j=0}^k \text{LR}_j \pi_j}, \text{ for } i = 0, \dots, k, \quad (6)$$

where  $\text{LR}_i$  is the likelihood ratio when  $H_i$  is compared against  $H_0$  [10]. In fact, the denominator in the LR cancels out, so (6) actually compares the likelihood of each hypothesis against all the other hypotheses jointly.

Just as for LRs, we cannot infer anything about the true relationship between the individuals as this might not be one of the hypotheses considered. For a flat prior, the posterior probabilities do not add any information to that provided by the genetic data and hence simply scale the relevant likelihoods (or LRs). More informative priors, on the other hand, can contribute additional information and this will be reflected in the posterior probabilities. For example, the three relationships half-sibling (H), avuncular (U) and grandparental (G) all have the same IBD coefficients and identical likelihoods. They are hence indistinguishable in the traditional frequentist setting and in a Bayesian setting using flat priors. Age information can easily be incorporated into the Bayesian approach and may yield different posterior probabilities.

### 3. Methods for blind search

#### 3.1. The likelihood ratio for X-chromosomal markers

X chromosomal markers are increasingly used in forensic applications to supplement or replace autosomal markers for some cases of practical importance [27]. One such example is shown in Fig. 8. The females B and C are *paternal* half sibs while C and D are *maternal* half sibs. The distinction between maternal and paternal is captured by X-chromosomal markers but not by autosomal markers. The paternal half sibs share an allele IBD inherited from their father. The Jacquard coefficients and the likelihood calculation can be modified to cater for independent X-chromosomal markers (details omitted). Obviously, the sex of the individuals in the pair matters. As an example note that there are only two possibilities, or two states, for a pair of males: either they share an allele IBD or they do not.

Since the number of unlinked markers on the X chromosome is limited, linkage and linkage disequilibrium become an issue [28]. We will ignore such dependence in Section 5.5. However, relevant findings that take dependence into account can be checked using the freely available software FamLinkX [29].

#### 3.2. Estimation of FPR and TPR

The true positive and false positive rates are determined by the hypotheses considered, number of loci, properties of each locus and the LR threshold. These rates can be calculated numerically using the algorithm described in [30]. However, this method only works for a small number of markers, say up to 10. In practice, we therefore resort to simulation. We denote estimates of FPR and TPR by  $\widehat{\text{FPR}}$  and  $\widehat{\text{TPR}}$ , respectively.

Typically TPR is close to 1 and FPR close to 0. These values are generally poorly estimated from direct Monte Carlo simulation. For instance, when  $\text{FPR} = 0.00001$ , 1 of 100000 simulations is expected to give a false positive. The conventional number of simulations in the range 100–10000 is therefore likely to return an estimate of 0. Kruijver [30] describes several methods for estimating small probabilities in forensic applications. One of these is importance sampling, which we use to estimate FPR in the results section. Details about importance sampling are given in Appendix B.

#### 3.3. Optimal LR threshold

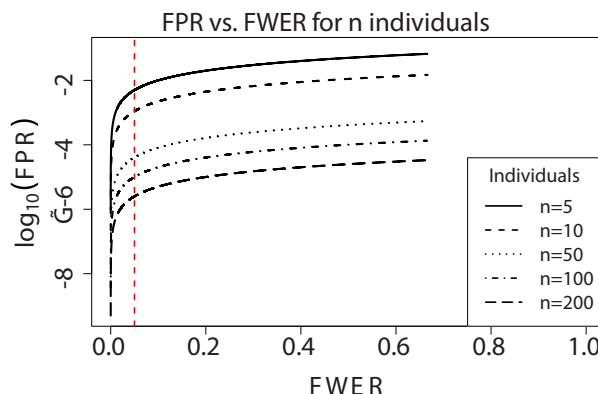
Intuitively we seek a threshold for LR that minimises the number of errors. Several approaches for choosing the optimal threshold have been suggested and compared [31]. We will focus on ROC-based methods. Fig. 4 shows a general ROC curve. Each point on the curve corresponds to a threshold  $t$ , with corresponding values for FPR and TPR. The value  $t$  in the figure corresponds roughly to  $\text{FPR} = 0.4$  and  $\text{TPR} = 0.6$ . The upper left corner corresponds to  $\text{FPR} = 0$  and  $\text{TPR} = 1$  and is therefore

**Table 2**

The statistics of a blind search summarised, as described in Section 3.4. Only  $W_0$ , the number of LRs below  $t$ , and  $W_1$ , the number of LRs above  $t$ , are observed.

	Claim $H_0$	Claim $H_1$	Total
$H_0$ true	TN	FP	$N_0$
$H_1$ true	FN	TP	$N_1$
Total	$W_0$	$W_1$	$N$

Source: Adapted from [32].



**Fig. 5.** Highest acceptable value of FPR, as a function of  $\alpha$ , given by (8) in Section 3.4. Plotted for blind search with 5, 10, 50, 100 and 200 individuals. The vertical line is located at  $\alpha = 0.05$ .

called the optimal point. Consequently, it is reasonable to choose a threshold that minimises the weighted Euclidean distance between the ROC curve and the point (0,1),

$$ER(t) = \sqrt{(w \text{FPR}(t))^2 + (1 - TPR(t))^2}, \quad (7)$$

where  $w \geq 0$ . In our examples the weight  $w = 1$ . It may be that one of the errors, typically a false positive, is more important to avoid than the other, a false negative. The relative importance of errors can be modelled by using other values of  $w$ . Because we do not know the exact values of FPR and TPR, they are replaced by their estimates.

#### 3.4. The problem of multiple testing in blind search

When doing a blind search among  $n$  DNA profiles, we compute one LR comparing two hypotheses,  $H_0$  and  $H_1$ , for each pair of DNA profiles, leaving us with a total of  $N = n(n-1)/2$  LRs,  $LR_1, \dots, LR_N$ . The blind search can be repeated with a different pair of hypotheses, but here we restrict attention to a single blind search with two hypotheses  $H_1$  and  $H_0$  for all pairwise comparisons. If the true hypothesis is known in each case, the result of the search (or any other multiple testing scenario) can be summarised as shown in Table 2. In practice, the truth will only be known for simulated data.

Assume that the only possibilities are the relationship stated by  $H_1$  or  $H_0$ , such that  $N_0 + N_1 = N$ . The number of type I errors or false positives is FP, while the number of false negatives is FN. Ideally, we want  $FP = 0$  and  $FN = 0$ . However, this is not realistic. For a sufficiently large threshold  $t$  we will never reject  $H_0$  and there will be no false positives, i.e.  $FP = 0$ . Similarly, there will be no false negatives,  $FN = 0$  (which means  $TP = N_1$ ), for a sufficiently small threshold. The challenge is to make both FP and FN acceptably small, or equivalently, make FP as small as possible and TP as close to  $N_1$  as possible.

Even if the probability of a false positive is very small for a single pairwise comparison, the fact that there are so many tests in a blind search could lead to a substantial probability of at least one false positive. Approaches to analyse and control these false positives in a

multiple testing setting have to be applied. The Family Wise Error Rate (FWER) [14] is often used for this purpose.

FWER is defined as the probability of getting at least one false positive out of  $N$  tests [32]. Let  $\alpha$  denote the FWER. For  $N$  independent tests,

$$\alpha = P(FP \geq 1) = 1 - (1 - FPR)^N,$$

where the FPR, as defined in (5), is assumed to be the same for each test. As we illustrate in the results section, the pairwise tests in a blind search are not independent and so we use the Bonferroni bound

$$\alpha \leq N \cdot FPR = \frac{n(n-1)}{2} FPR.$$

Thus, to obtain an  $\alpha$  below a given value, we choose a threshold so that  $FPR_\alpha \leq 2\alpha/(n(n-1))$  (8)

for a fixed sample size  $n$ . Fig. 5 plots  $FPR_\alpha$  as a function of  $\alpha$ , for a blind search with 5, 10, 50, 100 and 200 individuals. The red vertical line is located at  $\alpha = 0.05$ .

The aim is to find the threshold  $t$  that minimises  $ER(t)$  given in (7), with the constraint that  $FPR \leq FPR_\alpha$ .

## 4. Data and implementation

### 4.1. Real data and simulations

The DNA profiles evaluated in Sections 5.3 and 5.4 are from 65 individuals of Northern European origin (Germany) forming 8 pedigrees, with a variety of declared kinships up to 7th degree (the number of meioses between the persons of interest [33]). Most founders were not genotyped, and pedigree sizes ranged from 5 to 17, with an average of 9 members per family. Genotyping was done via massively parallel sequencing using the ForenSeq™ DNA Signature Prep kit (Verogen Inc., San Diego, CA, USA) and will be discussed in full elsewhere (M. Colucci, B. Rolf, N. A. Sheehan, M. A. Jobling, in preparation). Samples were collected with informed consent. For the purposes of the current study, we consider only the length-based genotypes from 27 autosomal STRs contained in Plex B of this kit. Allele frequencies are based on the European dataset in PopSTR (<http://spsmart.cesga.es/popstr.php>[34,35]) and downloaded from the Familias website (<https://familias.no/download>).

The performance analysis in Section 5.2, that leads to the blind search in the following section, is based on simulated data assuming the same set of loci as the real data, i.e., the 27 autosomal STR loci described above. This set of STR markers is also used in the simulations for the last example in the results section.

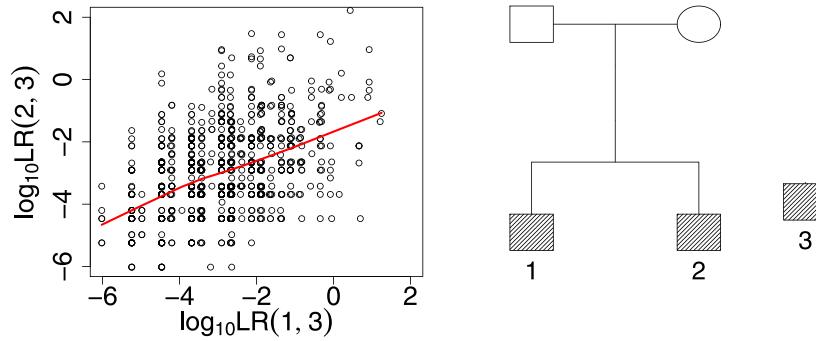
To demonstrate the use of X-chromosomal markers, data are simulated based on 12 X-chromosomal STR markers included in the kit "Investigator Argus X-12", with frequencies taken from an Argentinian database [36]. This is the most widely used kit for forensic applications.

### 4.2. Implementation

The analyses in this paper are all performed using R code developed by the authors. The code is available from the first author on request. The main engine of the code is an implementation of the parametric version of the likelihood function. This efficiently computes likelihoods for a series of relationships and converts to LRs and posterior probabilities. The code builds on the R libraries *pedtools*, *ribd*, *forrel* and *pedmut* developed by Magnus Dehli Vigeland, freely available from CRAN [37].

### 4.3. Assumptions

All the equations above are based on (1) which is only valid under certain assumptions. Firstly, the population is assumed to be in Hardy-



**Fig. 6.** Figure corresponding to the correlation discussion in Section 5.1. Left: Scatter plot of LRs of simulated data for two siblings and an unrelated individual. Red line shows regression line. Right: Pedigree used for simulation of data, identifying the id labels 1, 2 and 3 in left panel.

Weinberg Equilibrium (HWE) and in Linkage Equilibrium (LE). Secondly, mutations are ignored. Mutation rates are usually small, and the errors induced by ignoring them in likelihood calculations are typically negligible [38]. However, for a parent offspring (PO) relationship, i.e.  $\kappa = (0, 1, 0)$ , the likelihood will be zero if the two samples have genotypes at any locus that are incompatible with this hypothesis, e.g.  $g_A = aa$  and  $g_B = bb$ . For this special case, there is a simple formulation of the likelihood that incorporates mutation (see [10]). This likelihood formula is applied throughout the paper when the likelihood of  $\kappa = (0, 1, 0)$  is computed. An extended stepwise mutation model is implemented, with mutation rates of  $10^{-3}$  and  $5 \cdot 10^{-6}$  for mutation of integer and non-integer alleles, respectively and a mutation range of 0.1. For further details on this mutation model, see paper by Simonsson and Mostad [39]. We ignore allele drop-ins and drop-outs, null alleles and genotyping errors.

## 5. Results

The first example shows that the LRs in a blind search are not independent. The second example demonstrates how to evaluate the performance of a blind search such as we present in the third example. We then carry out a blind search on X-chromosomal markers before showing how inbreeding can be accommodated.

### 5.1. Correlation between LRs in a blind search

In this example, we show by simulation a case where the LRs of a blind search are correlated. Consider the pedigree in Fig. 6 and the hypotheses  $H_1$  stating a sibling relationship and  $H_0$  unrelated. Let  $LR_{1,3}$  denote the likelihood ratio when individual 1 is compared to 3 and

define  $LR_{2,3}$  analogously. We use 10 independent loci, each with 10 alleles and equal allele frequencies of 0.1. Note that the LRs are random variables. We simulate 1000 sets of DNA profiles for the three shaded individuals of the pedigree in Fig. 6. The values of  $LR_{1,3}$  and  $LR_{2,3}$  are computed for each simulation. The results are shown in the scatter plot in Fig. 6, the red line denoting a regression line.

The estimated correlation between the logarithmic values of  $LR_{1,3}$  and  $LR_{2,3}$  is 0.484. This shows that the LRs are not independent. In other words, the outcome of different comparisons cannot be interpreted independently if one individual is involved in several comparisons. We elaborate on the implications of this correlation in the discussion.

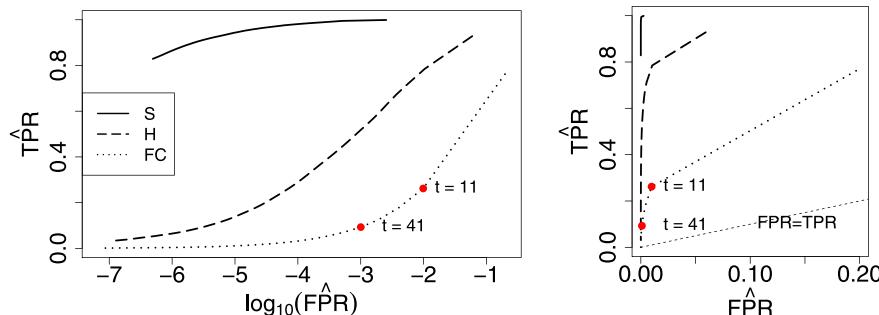
### 5.2. From FWER to choice of LR threshold

In Section 5.3, we carry out a blind search among 65 individuals, genotyped for a set of 27 STR loci. Here, we present the preliminary evaluation required to obtain optimal LR thresholds for that search.

The first step is to decide on an acceptable value of  $\alpha$ . From this value of  $\alpha$  we can decide on an upper limit of the FPR and then the corresponding optimal LR threshold. For a blind search of  $n = 65$  individuals, with the requirement that  $\alpha \leq 0.05$ , Equation (8) gives an upper limit for the false positive rate of  $FPR_{0.05} = 2.404 \cdot 10^{-5}$ .

The next step is to analyse how the FPR and TPR relate to each other for this particular set of markers. This depends on what hypotheses we test in the blind search. In the following example, we consider the hypotheses  $H_1$ : PO,  $H_2$ : S,  $H_3$ : H/U/G and  $H_4$ : FC, all against  $H_0$ : UN. We therefore consider these hypotheses here when estimating FPR and TPR.

Fig. 7 shows ROC curves for the different hypotheses. The values for  $\widehat{FPR}$  and  $\widehat{TPR}$  are estimated from simulated data, as described in Section 3.2. For  $H_1$ : PO, we only obtained estimates of FPR smaller than  $10^{-7}$ ,



**Fig. 7.** ROC curves for the analysis performed in Section 5.2. The hypothesis  $H_1$  stating S, H and FC and  $H_0$  unrelated, using 27 STR markers. ROC curves from simulated data. A threshold of 11 corresponds to an estimated false positive rate of about 0.01 and an estimated true positive rate of about 0.26 for FC. The right figure shows the same estimated ROC curves and the line  $FPR = TPR$ , with an untransformed first axis.

**Table 3**

Optimal thresholds for different relationships, with corresponding  $\widehat{\text{FPR}}$  and  $\widehat{\text{TPR}}$ , for the analysis performed in [Section 5.2](#). For  $\alpha = 0.05$  (left table) and  $\alpha = 0.1$  (right table), for blind search with  $n = 65$  individuals.

$\alpha = 0.05 \Rightarrow \text{FPR}_\alpha = 2.404 \cdot 10^{-5}$			
	$t$	$\widehat{\text{FPR}}$	$\widehat{\text{TPR}}$
PO	65531	$5.439 \cdot 10^{-8}$	1.000
S	771	$2.296 \cdot 10^{-5}$	0.961
H	2501	$2.365 \cdot 10^{-5}$	0.184
FC	421	$1.689 \cdot 10^{-5}$	0.014

$\alpha = 0.1 \Rightarrow \text{FPR}_\alpha = 4.808 \cdot 10^{-5}$			
	$t$	$\widehat{\text{FPR}}$	$\widehat{\text{TPR}}$
PO	65531	$5.439 \cdot 10^{-8}$	1.000
S	311	$4.713 \cdot 10^{-5}$	0.972
H	1551	$4.770 \cdot 10^{-5}$	0.232
FC	251	$4.421 \cdot 10^{-5}$	0.022

**Table 4**

Results of the blind search among  $n = 65$  individuals in [Section 5.3](#) with  $\alpha = 0.05$  where  $N_1$  denotes the total number of pairs in the sample with the tested relationship, TP is the number of these pairs with a LR above the threshold, and FP is the number of unrelated individuals with a LR above the threshold. The last row gives the number of other (differently related) pairs with an LR above the threshold.

	PO	S	H/U/G	FC
$N_1$	46	13	64	21
TP	43	12	12	0
FP	0	1	2	0
H <sub>1</sub> Claimed, other true	4	53	57	59

with a corresponding estimated TPR of 0.999 or higher. This shows that the LR comparing PO to UN is high when the true relationship is PO and low otherwise. Parent offspring and unrelated individuals are easily distinguished as expected and so we have omitted this curve from the graph.

The ROC curves show the estimated properties of a single computation of the LR, for the respective hypotheses, for this specific set of STR markers. The curves do not depend on the number of individuals in the blind search.

The last step in the performance analysis is to identify the optimal threshold, by minimising  $\text{ER}(t)$ , with the constraint  $\widehat{\text{FPR}} \leq \text{FPR}_\alpha$ . The highest optimal thresholds for  $\alpha = 0.05$  and  $\alpha = 0.1$  are listed in [Table 3](#).

### 5.3. Blind search with real data

In this example we do a blind search on the data described in [Section 4.1](#). The data set contains 65 DNA profiles. A blind search among these profiles results in 2080 pairwise comparisons. We want to test the hypotheses  $H_1$ : PO,  $\kappa_1 = (0, 1, 0)$ ,  $H_2$ : S,  $\kappa_2 = (0.25, 0.5, 0.25)$ ,  $H_3$ : H/U/G,  $\kappa_3 = (0.5, 0.5, 0)$ ,  $H_4$ : FC,  $\kappa_4 = (0.9375, 0.0625, 0)$ , against  $H_0$ : UN,  $\kappa_0 = (1, 0, 0)$ . In the previous section, we obtained optimal thresholds for blind searches with these hypotheses ([Table 3](#)). A stepwise mutation model is implemented in the evaluation of PO.

[Table 4](#) summarises the blind searches performed on the real data. This table is possible to construct because we know the true relationship for each pair from the pedigree information. In practice, only the sum of the last three rows (for each relationship) would be known.

For PO, we are left with a list of 47 hits. 43 of these are true PO, while 4 of the 47 hits are pairs of individuals with another relationship. 3 pairs with true PO relationship are not detected. By lowering the threshold, the remaining 3 pairs could have been detected. However, the probability of obtaining false positives increases by decreasing the threshold. For S, only one true sibling pair is not detected and there is only one false

**Table 5**

LR values for seven pairs of the blind search in [Section 5.3](#). Values for H, U and G are the same and shown in the column H/U/G. Values smaller than  $10^{-6}$  are set to 0.

	PO	S	H/U/G	FC	UN	True
1	<b><math>5.181 \cdot 10^{10}</math></b>	<b><math>1.205 \cdot 10^8</math></b>	<b><math>1.825 \cdot 10^7</math></b>	<b><math>5.593 \cdot 10^4</math></b>	1	PO
2	353.460	<b><math>1.544 \cdot 10^8</math></b>	<b><math>3.886 \cdot 10^5</math></b>	<b><math>5.189 \cdot 10^3</math></b>	1	S
3	0	0.681	57.572	20.519	1	H
4	0	$5.017 \cdot 10^{-3}$	4.156	4.0984	1	U
5	0	0.030	13.269	16.916	1	G
6	0	$1.115 \cdot 10^{-4}$	0.163	1.375	1	FC
7	0	$1.821 \cdot 10^{-6}$	0.022	0.349	1	UN

**Table 6**

Posterior probabilities, computed from the LR values of [Table 5](#), when applying a flat prior, i.e.,  $\pi_i = 1/7$  for  $i = 0, \dots, 6$ , as described in [Section 5.4](#). Values for H, U and G are the same and shown in the column H/U/G. Probabilities smaller than  $10^{-6}$  are set to 0.

	PO	S	H/U/G	FC	UN	True
1	<b>0.997</b>	0.002	0.0004	$1.076 \cdot 10^{-6}$	0	PO
2	$2.272 \cdot 10^{-6}$	<b>0.993</b>	0.002	$3.336 \cdot 10^{-5}$	0	S
3	0	0.0003	<b>0.295</b>	0.105	0.005	H
4	0	$2.86 \cdot 10^{-4}$	<b>0.237</b>	0.233	0.057	U
5	0	$5.15 \cdot 10^{-4}$	0.230	<b>0.293</b>	0.017	G
6	0	$3.90 \cdot 10^{-5}$	0.057	<b>0.480</b>	0.349	FC
7	0	$1.29 \cdot 10^{-6}$	0.0162	0.246	<b>0.706</b>	UN

positive. However, the list of hits contains 66 pairs of individuals, 53 of these having another relationship.

We conclude that the summary in [Table 4](#) is consistent with the performance evaluation shown in [Table 3](#). PO can easily be distinguished from UN. The more distant the tested relationship, the lower the power to distinguish it from unrelated. With the obtained optimal thresholds, the number of false positives stays low as desired. For each hypothesis tested, the list of pairs warranting further investigation comprises those in the final row of [Table 4](#), i.e., those who do not have the tested relationship and who are also not unrelated.

### 5.4. Analysis of posterior probabilities

The result of each of the blind searches performed in [Section 5.3](#) is a list of pairs with a LR above the threshold. Some pairs of individuals may appear in several of the lists, while other pairs may not be present in any of the lists. In this example, we turn to Bayesian analysis to further investigate specific pairs.

[Table 5](#) shows LR values for 7 pairs from the above blind search. Values above the LR thresholds are in bold font. The rightmost column gives the true relationship. Only the first two pairs have LR values above the thresholds given in the left table of [Table 3](#) corresponding to  $\alpha = 0.05$ . For pairs 3–7, the LRs are low, some below 1, indicating that a UN relationship is more plausible than the alternative hypothesis.

Next we calculate posterior probabilities to see if it is possible to infer a relationship for the different pairs. LR thresholds are not required for this. [Table 6](#) shows posterior probabilities for the different hypotheses, with flat prior probabilities, i.e.,  $\pi_i = 1/7$  for  $i = 0, \dots, 6$ . The highest probability for each pair is in bold and corresponds to the true relationship for several of the pairs. For example, the LRs in [Table 5](#) comparing S, H/U/G and FC against UN for the second pair were all above the relevant LR thresholds. The posterior probability of S is close to 1, now making it possible to correctly infer this relationship. For pairs 3, 4 and 5, the highest posterior probabilities are just below 0.3. Even though the corresponding relationship is the most probable, a posterior probability of 0.3 is maybe not high enough to allow firm conclusions to be drawn.

The relationships H, U and G are indistinguishable in the parametric framework presented in [Section 2](#). Also posterior probabilities with a flat

**Table 7**

Posterior probabilities with informative priors, as described in Section 5.4. Probabilities smaller than  $10^{-6}$  are set to 0.

	PO	S	H	U	G	FC	UN	True
1	<b>0.999</b>	6.57 · $10^{-4}$	3.06 · $10^{-5}$	2.52 · $10^{-4}$	2.07 · $10^{-4}$	0	0	PO
2	8 · $10^{-6}$	<b>0.988</b>	7.65 · $10^{-4}$	0.006	0.005	5.36 · $10^{-5}$	0	S
3	0	0.001	0.038	<b>0.317</b>	0.259	0.072	0.312	H
4	0	2.94 · $10^{-5}$	0.007	0.062	0.051	0.039	<b>0.841</b>	U
5	0	1.26 · $10^{-4}$	0.017	0.143	0.117	0.116	<b>0.608</b>	G
6	0	0	3.42 · $10^{-4}$	0.003	0.002	0.015	<b>0.979</b>	FC
7	0	0	4.78 · $10^{-5}$	3.95 · $10^{-4}$	3.23 · $10^{-4}$	0.004	<b>0.995</b>	UN

prior as in Table 6 can not differentiate between them. Additional information, preferably objective, needs to be considered.

Suppose now that we have knowledge of how many pairs of the different relationships are present among the DNA profiles. This could be the case in a plane crash with a known passenger list. There are 1867 unrelated pairs, 46 parent-offspring pairs, 13 sibling pairs, 4 half sibling pairs, 33 avuncular pairs, 27 grandparental pairs and 21 first cousin pairs. The remaining 69 pairs have other more distant relationships not investigated here. The prior probabilities are then  $\pi_0 = 0.928$  (UN),  $\pi_1 = 0.023$  (PO),  $\pi_2 = 0.006$  (S),  $\pi_3 = 0.002$  (H),  $\pi_4 = 0.016$  (U),  $\pi_5 = 0.013$  (G) and  $\pi_6 = 0.010$  (FC).

Posterior probabilities using these more informative priors are shown in Table 7. The prior probability of a PO relationship is  $\pi_1 = 0.023$ , i.e., there is a chance of 2.3% that a pair of individuals has a PO relationship. The corresponding posterior probability for the first pair is 0.999. The genetic data give such strong support to PO, that even though the prior probability is low, the posterior probability of this relationship is approximately 1.

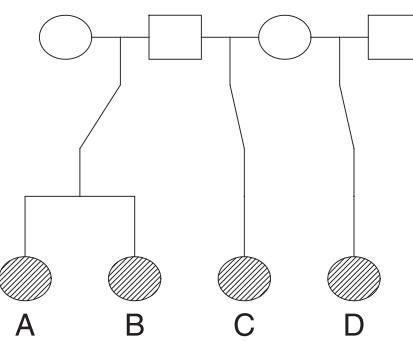
In this blind search (as in most other blind searches), most pairs of individuals are unrelated, making the prior probability of UN close to 1 and the others low. This requires the LRs for the other relationships to be high in order to be supported by the posterior probabilities. For the relationships H/U/G and FC, the LR of the true relationships against UN is typically low. The combination of priors and LRs makes the posterior probability of UN high while the posterior probability of the true relationship remains low.

For this reason, this particular set of prior probabilities, even though objective, does not help us to distinguish between the H, U and G relationships in these data.

### 5.5. Blind search with X-chromosomal markers

Because a male has only one X-chromosome, paternal half sisters (HSP) must inherit the same X-chromosome from their common father. Their second X-chromosomes, inherited from their respective mothers, are not IBD (since their mothers are unrelated), and hence, the IBD coefficients for a HSP relationship are  $\kappa = (0, 1, 0)$ . The IBD coefficients for maternal half siblings (HSM), whether considering X-chromosomal or autosomal markers, are  $\kappa = (0.5, 0.5, 0)$ . In the following example, we show with simulated data how X-chromosomal markers can distinguish between HSP and HSM.

We simulated genotypes for 12 X-chromosomal STR markers, for the shaded individuals in Fig. 8. Genotypes are simulated for each locus independently, by gene-dropping through the pedigree structure. More specifically, genotypes are sampled for the founders of the pedigree (the parents) according to the allele population frequencies and passed down through the pedigree assuming the rules of Mendelian inheritance. Only the resulting genotypes of the offsprings are kept for the applications in this example. Table 8 presents the average posterior probabilities over



**Fig. 8.** Pedigree connecting the individuals of the analysis in Section 5.5. Marker data are simulated for the four daughters to demonstrate blind search with X-chromosomal markers.

100 simulations, for the relationships PO, S, HSP, HSM and UN, for the six possible comparisons between the individuals A, B, C and D. A flat prior  $\pi_i = 1/5$  for  $i = 0, \dots, 4$  is assumed.

The evidence in favour of C-D being HSM, shown in bold in Table 8, could not be obtained using autosomal markers. Since we are using a flat prior, the LR comparing maternal to paternal half sibs can be found from the posterior probability ratio,  $0.81327/0.01916 = 42.4$ . This value may not be decisive on its own, but supplements other evidence. Note that HSP cannot be distinguished from PO using X-chromosomal markers alone as the row for the comparison A-C confirms. Age information, autosomal marker data or other non-DNA data may solve such cases.

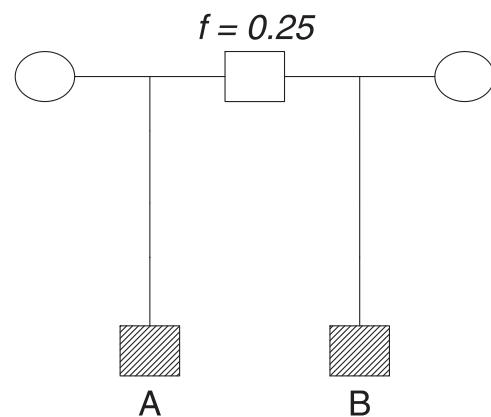
### 5.6. Half siblings with inbred founder

Computations of LRs and posterior probabilities are restricted to a limited set of predefined pedigree relationships in many current software implementations. The parametric form of the LR given in (3) enables us to compute LRs and do blind search for any pairwise

**Table 8**

Posterior probabilities averaged over 100 simulations for the comparisons between the four daughters in Fig. 8.

	PO	S	HSP	HSM	UN
A-B	0.039	0.921	0.039	0.001	0.000
A-C	0.475	0.039	0.475	0.012	$2 \cdot 10^{-5}$
A-D	0.000	0.000	0.000	0.154	0.846
B-C	0.471	0.045	0.471	0.012	$2 \cdot 10^{-5}$
B-D	0.000	0.000	0.000	0.146	0.854
C-D	0.019	0.001	0.019	<b>0.813</b>	0.147



**Fig. 9.** Half sibling pedigree with founder inbreeding assumed in the analysis in Section 5.6.

relationship. In this example we show how background inbreeding can be modelled and how this can be taken into account in the Bayesian framework.

Assume a set of DNA profiles among which we want to do a blind search. The number of profiles is not important. The pedigrees connecting the individuals are unknown, but we know that the individuals come from a population where inbreeding is common. We consider the hypotheses  $H_1$ : PO,  $H_2$ : S,  $H_3$ : H and  $H_4$ : H with founder inbreeding  $f = 0.25$ , all against  $H_0$ : UN. The relationship in  $H_4$  is shown in Fig. 9. Individuals A and B are outbred paternal half siblings, with the father being inbred with an inbreeding coefficient  $f = 0.25$ . This value of  $f$  corresponds to extreme inbreeding where the parents of the father are siblings. The IBD coefficients for the half sibling relationship are  $\kappa = (0.375, 0.625, 0)$ .

We consider one pair with true relationship  $H_4$ . A total of 100 simulations of DNA profiles for this pair is performed. LRs and posterior probabilities, with a flat prior  $\pi_i = 1/5$ ,  $i = 0, \dots, 4$  are computed for each simulation. Mean values of the posterior probabilities for the hypotheses  $H_0, \dots, H_4$ , are:  $\bar{p}_1 = 0.017$ ,  $\bar{p}_2 = 0.094$ ,  $\bar{p}_3 = 0.374$ ,  $\bar{p}_4 = 0.495$  and  $\bar{p}_0 = 0.019$ . It can be seen that the mean posterior probability of hypothesis  $H(f)$  is about 0.5, making it possible to distinguish it from the half sibling relationship without inbreeding.

The coefficient of inbreeding in this example is quite high. Lower values of  $f$  make the pair genetically more similar to half siblings without inbreeding, and distinguishing these relationships becomes harder without additional information. This high degree of inbreeding may be more relevant for non-human applications.

## 6. Discussion

The topic of this paper is blind search, a procedure used to search for pairwise relationships among a set of unidentified DNA profiles. Each pairwise comparison is similar to a kinship test performed, for instance, to resolve a paternity dispute. In the paper, we focus mainly on issues related to multiple testing. For this reason we will not discuss Hardy-Weinberg equilibrium and other assumptions that our applications share with other applications in forensic genetics. For instance, it is not obvious how evidence from different DNA sources like autosomal markers and X-chromosomal markers should be combined. However, this challenge is no different for a blind search than for a kinship test and is therefore not addressed here.

Case workers must decide on how the results of a blind search should be evaluated and reported. The context, or specific application, is obviously not irrelevant. In a DVI application, a false identification is likely to be a more serious error than missing an identification. To account for this, the metric for determining the threshold in (7) allows a weight to be specified which would penalise false identifications. Other applications, such as screening a database for relatives prior to estimating allele frequencies, may not require a weighting for errors. If costs can be specified for the possible errors, optimal decision rules can be derived as explained in Chapter 8.1 of [10]. However, there is hardly ever an objective way to balance the two errors that can occur and so specification of weights or costs may not be a viable option. We have used the unweighted form of the metric throughout. We have only considered binary decisions (corresponding to  $t_0 = t_1$ ) as stated in the beginning of Section 2.4. One could drop this requirement and declare a test to be inconclusive if  $t_0 < LR < t_1$ . In this case, a cost for making no decision would have to be added and (7) modified accordingly.

We only presented one method to determine an optimal threshold based on the distance illustrated in Fig. 4 although several alternatives are available [31]. Results using different approaches were practically identical for the examples we presented and so we chose not to discuss the thresholds based on the other metrics. Furthermore, other approaches to obtain ROC curves could be considered. For instance, there are several ways to smooth ROC curves. It is also possible to provide confidence bands for the ROC curves and study the impact of

assumptions. This has been explored in previous papers [40]. Fig. 6 shows that the LRs from a blind search may be correlated when the same individual is involved in two comparisons. This has several implications. In particular, the results of different comparisons cannot be interpreted independently. Intuitively, we may get a high LR if unrelated individuals A and B happen to share a rare allele. Another individual C, who is a close relative of A, is likely to share this allele IBD with A and so we can also expect a high LR when comparing B and C. Importantly, the methods used to control the overall error rate must allow for dependence and for this reason we used the Bonferroni bound (8) as an upper limit for the FWER. Another frequently used measure to control the overall error rate in multiple testing scenarios is the false discovery rate (FDR) [14]. When controlling the FDR, the outcome of each test is based on p-values. However, conventional significance testing based on p-values are not recommended to evaluate the strength of DNA evidence in forensic genetics [41,42].

Furthermore, a blind search will not necessarily provide a globally consistent ‘solution’ in the sense that the LRs may support impossible combinations of relationships, like one individual having two mothers. An interesting extension to this paper would be to investigate alternative search strategies that may improve the results of a blind search. One strategy could be to do the search sequentially, where hypotheses to be tested depend on the outcome of the previous pairwise comparisons. For instance, if individuals A and B are classified as PO and A and C as PO, then it would be logical to test if B and C are siblings, half siblings or a grandparent-grandchild pair. There are also methods and software for pedigree reconstruction, see Chapter 8 of [37]. Finally, the true relationship may not be among the alternatives considered. This is also true for the Bayesian approach.

A Bayesian interpretation might seem more appropriate than the frequentist alternative for blind search applications than for a kinship case. The alternative, based on the LR, is designed to deal with only two hypotheses. If there are several hypotheses, a reference hypothesis must be specified. The posterior probabilities reported using a Bayesian approach make comparison of several competing hypotheses simpler as they are between 0 and 1. However, as always, a prior is needed for the Bayesian approach and the choice of prior may be crucial. If DNA is of poor quality, leading to few markers being typed, or if the competing hypotheses specify relationships that are very close to each other, conclusions may hinge on the choice of the prior.

An important aspect of this paper is the use of the parametric representation of relationships. This enables us to investigate any admissible pairwise relationship between two outbred individuals. By defining founder inbreeding in a pedigree structure, as shown in Fig. 3, background inbreeding can also be modelled [22]. Rather than proposing specific alternative relationships, we could simply estimate the coefficients describing the relationship. In the outbred case, these estimates can be plotted in the IBD triangle in Fig. 2 which would indicate where these relationships lie in relation to the well known relationships. For instance, pairs with estimates close to (0.25, 0.25) could be classified as siblings.

Throughout, we have restricted attention to pairwise testing. In principle, the blind search can be extended to search for relationship between triplets. However, the parametric approach based on the Jacuard coefficients then becomes impractical. The number of parameters needed to describe the relationship between three individuals increases, from 2 to 15 in the outbred case.

Issues to do with reporting DNA evidence are currently of key interest as evidenced by the so-called “DNA database controversy” (see [11] and references therein). The main message of this paper is that there are also problems related to multiple testing in kinship analyses which cannot be ignored.

## Conflict of interest

The authors declare that they have no conflict of interest.

## Acknowledgements

We thank Burkhard Rolf for DNA samples. Margherita Colucci was supported by an iCASE MIBTP Biotechnology and Biological Sciences Research Council PhD studentship, grant ref. BB/M01116X/1, partnered

by DNA Worldwide, and by a Short Term Fellowship of the International Society for Forensic Genetics (ISFG). We thank NUCLEUS Genomic Services at the University of Leicester for access to DNA sequencing. This research used the SPECTRE High Performance Computing Facility at the University of Leicester for data analysis.

## Appendix A. Table of genotype probabilities

**Table 9**

The conditional probability  $P(G|J_i)$  of a pair of genotypes  $G = (g_A, g_B)$ , given a Jaccard state  $J_i$ . The symbols  $a, b, c$  and  $d$  represent different alleles, with population frequencies  $p_a, p_b, p_c$  and  $p_d$  respectively.

$(g_A, g_B)$	$J_1$	$J_2$	$J_3$	$J_4$	$J_5$	$J_6$	$J_7$	$J_8$	$J_9$
$(aa, aa)$	$p_a$	$p_a^2$	$p_a^2$	$p_a^3$	$p_a^2$	$p_a^3$	$p_a^2$	$p_a^3$	$p_a^4$
$(aa, bb)$	0	$p_a p_b$	0	$p_a p_b^2$	0	$p_a^2 p_b$	0	0	$p_a^2 p_b^2$
$(aa, ab)$	0	0	$p_a p_b$	$2p_a^2 p_b$	0	0	0	$p_a^2 p_b$	$2p_a^3 p_b$
$(aa, bc)$	0	0	0	$2p_a p_b p_c$	0	0	0	0	$2p_a^2 p_b p_c$
$(ab, aa)$	0	0	0	0	$p_a p_b$	$2p_a^2 p_b$	0	$p_a^2 p_b$	$2p_a^3 p_b$
$(bc, aa)$	0	0	0	0	0	$2p_a p_b p_c$	0	0	$2p_a^2 p_b p_c$
$(ab, ab)$	0	0	0	0	0	0	$2p_a p_b$	$p_a p_b(p_a + p_b)$	$4p_a^2 p_b^2$
$(ab, ac)$	0	0	0	0	0	0	0	$p_a p_b p_c$	$4p_a^2 p_b p_c$
$(ab, cd)$	0	0	0	0	0	0	0	0	$4p_a p_b p_c p_d$

## Appendix B. Importance sampling

Importance sampling is a method that can be used to approximate small probabilities. We first introduce the indicator function,

$$I(LR > t) = \begin{cases} 1, & \text{if } LR \geq t, \\ 0, & \text{if } LR < t. \end{cases}$$

The expectation of  $I$  becomes

$$\begin{aligned} E(I(LR \geq t)) &= 0 \cdot P(LR < t) + 1 \cdot P(LR \geq t) \\ &= P(LR \geq t) \\ &= FPR \end{aligned}$$

It is therefore valid to say that  $FPR = E(I(LR \geq t))$ . Then consider the expression for the expected value in a more general sense. The value of the function  $I$  is dependent on the value of the LR, which is a function of the genotypes  $G$  of the DNA profiles. The probability distribution of  $G$  is governed by the relationships that has generated the data. For this consideration, we assume that this relationship is either  $H_0$  or  $H_1$ . Denote by  $X$  the values that  $I$  can take on. We then have

$$E(I(LR \geq t)) = \sum_j X_j \cdot P(G_j | H_0) \approx \frac{1}{N} \sum_{i=1}^N I(LR_i^{H_0} \geq t).$$

In the last expression, the expected value is estimated by the sample mean of  $I$ , from a set of  $N$  simulations. The genotypes  $G$ , and then also  $X$ , are distributed according to  $H_0$ , which is indicated by the superscript of the LR. Then, consider the opposite probability distribution,  $P(G|H_1)$ , where the genotypes are distributed according to  $H_1$ . As long as  $P(G_j|H_0) = 0$  whenever  $P(G_j|H_1) = 0$ , we can write

$$E(I(LR \geq t)) = \sum_j X_j \cdot \frac{P(G_j | H_0)}{P(G_j | H_1)} P(G_j | H_1) \approx \frac{1}{N} \sum_{i=1}^N \frac{I(LR_i^{H_1} \geq t)}{LR_i^{H_1}}.$$

Using this method, the LR is sampled under the wrong hypothesis ( $H_1$ ), instead of the desired hypothesis ( $H_0$ ). The bias this introduces is adjusted for by the weight  $LR_i^{H_1}$ . An estimate of FPR is then

$$\widehat{FPR} = \frac{1}{N} \sum_{i=1}^N \frac{I(LR_i^{H_1} \geq t)}{LR_i^{H_1}}.$$

## References

- [1] B. Olaisen, M. Stenersen, M. M. Identification by DNA analysis of the victims of the August 1996 Spitsbergen civil aircraft disaster, *Nat. Genet.* 15 (1997) 402–405.
- [2] B. Bertoglio, P. Grignani, P. Di Simone, N. Polizzi, D. De Angelis, C. Cattaneo, A. Iadicicco, P. Fattorini, S. Presciuttini, C. Previderè, Disaster victim identification by kinship analysis: the Lampedusa October 3rd, 2013 shipwreck, *Forensic Sci. Int.: Genet.* 44 (2020) 102–156.
- [3] L. Olivieri, D. Mazzarelli, B. Bertoglio, D. De Angelis, C. Previderè, P. Grignani, A. Cappella, S. Presciuttini, C. Bertuglia, P. Di Simone, Challenges in the identification of dead migrants in the Mediterranean: the case study of the Lampedusa shipwreck of October 3rd 2013, *Forensic Sci. Int.* 285 (2018) 121–128.
- [4] C.H. Brenner, Some mathematical problems in the DNA identification of victims in the 2004 tsunami and similar mass fatalities, *Forensic Sci. Int.* 157 (2006) 172–180.
- [5] C.H. Brenner, B.S. Weir, Issues and strategies in the DNA identification of World Trade Center victims, *Theor. Popul. Biol.* 63 (2003) 173–178.
- [6] T.J. Parsons, R.M.L. Huel, Z. Bajunović, A. Rizvić, Large scale DNA identification: the ICMP experience, *Forensic Sci. Int.: Genet.* 38 (2019) 236–244.
- [7] S. Palomo-Díez, A. Esparza-Arroyo, M. Tirado-Vizcaño, J. Velasco Vázquez, A. M. López-Parra, C. Gomes, C. Baeza-Richer, E. Arroyo-Pardo, Kinship analysis and allelic dropout: a forensic approach on an archaeological case, *Ann. Hum. Biol.* 45 (2018) 365–368.
- [8] S. Palomo-Díez, A.M. López-Parra, C. Gomes, C. Baeza-Richer, A. Esparza-Arroyo, E. Arroyo-Pardo, Kinship analysis in mass graves: evaluation of the Blind Search tool of the Familias 3.0 Software in critical samples, *Forensic Sci. Int.: Genet. Suppl. Ser.* 5 (2015) e547–e550.
- [9] T.J. Pemberton, C. Wang, J.Z. Li, N.A. Rosenberg, Inference of unexpected genetic relatedness among individuals in hapmap phase iii, *Am. J. Hum. Genet.* 87 (2010) 457–464.
- [10] T. Egeland, D. Kling, P. Mostad, Relationship Inference with Familias and R: Statistical Methods in Forensic Genetics, Academic Press, 2015.
- [11] G. Storvik, T. Egeland, The DNA database search controversy revisited: bridging the Bayesian-frequentist gap, *Biometrics* 63 (2007) 922–925.
- [12] F.R. Bieber, C.H. Brenner, D. Lazer, Finding criminals through DNA of their relatives, *Science* (2006).
- [13] M. Kruijver, R. Meester, K. Slooten, Optimal strategies for familial searching, *Forensic Sci. Int.: Genet.* 13 (2014) 90–103.
- [14] J.D. Storey, The positive false discovery rate: a Bayesian interpretation and the q-value, *Ann. Stat.* 31 (2003) 2013–2035.
- [15] D. Kling, A.O. Tillmar, T. Egeland, Familias 3-extensions and new functionality, *Forensic Sci. Int.: Genet.* 13 (2014) 121–127.
- [16] T. Egeland, P.F. Mostad, B. Mevåg, M. Stenersen, Beyond traditional paternity and identification cases: selecting the most probable pedigree, *Forensic Sci. Int.* 110 (2000) 47–59.
- [17] A. Jacquard, Genetic information given by a relative, *Biometrics* (1972) 1101–1114.
- [18] N.A. Sheehan, T. Egeland, Structured incorporation of prior information in relationship identification problems, *Ann. Hum. Genet.* 71 (2007) 501–518.
- [19] S. Wright, Coefficients of inbreeding and relationship, *Am. Nat.* 56 (1922) 330–338.
- [20] E.A. Thomson, The estimation of pairwise relationships, *Ann. Hum. Genet.* 39 (1975) 173–188.
- [21] E.A. Thompson, A restriction on the space of genetic relationships, *Ann. Hum. Genet.* 40 (1976) 201–204.
- [22] M.D. Vigeland, T. Egeland, Handling founder inbreeding in forensic kinship analysis, *Forensic Sci. Int.: Genet. Suppl. Ser.* 7 (2019) 780–781.
- [23] H.K. Brustad, T. Egeland, The impact of ignoring inbreeding in pairwise kinship evaluations, *Forensic Sci. Int.: Genet. Suppl. Ser.* 7 (2019) 462–464.
- [24] M.D. Vigeland, Relatedness coefficients in pedigrees with inbred founders, *J. Math. Biol.* 81 (2020) 185–207.
- [25] A.O. Tillmar, P. Mostad, Choosing supplementary markers in forensic casework, *Forensic Sci. Int.: Genet.* 13 (2014) 128–133.
- [26] T. Fawcett, An introduction to ROC analysis, *Pattern Recognit. Lett.* 27 (2006) 861–874.
- [27] N. Pinto, L. Gusmão, A. Amorim, X-chromosome markers in kinship testing: a generalisation of the IBD approach identifying situations where their contribution is crucial, *Forensic Sci. Int.: Genet.* 5 (2011) 27–32.
- [28] D. Kling, A. Tillmar, T. Egeland, P. Mostad, A general model for likelihood computations of genetic marker data accounting for linkage, linkage disequilibrium, and mutations, *Int. J. Leg. Med.* 129 (2015) 943–954.
- [29] D. Kling, B. Dell'Amico, A.O. Tillmar, FamLinkX—implementation of a general model for likelihood computations for X-chromosomal marker data, *Forensic Sci. Int.: Genet.* 17 (2015) 1–7.
- [30] M. Kruijver, Efficient computations with the likelihood ratio distribution, *Forensic Sci. Int.: Genet.* 14 (2015) 116–124.
- [31] M. Rota, L. Antolini, Finding the optimal cut-point for Gaussian and Gamma distributed biomarkers, *Comput. Stat. Data Anal.* 69 (2014) 1–14.
- [32] A.C. Tamhane, Y. Hochberg, C.W. Dunnett, Multiple test procedures for dose finding, *Biometrics* (1996) 21–37.
- [33] M. Steffens, C. Lamina, T. Illig, T. Bettecken, R. Vogler, P. Entz, E. Suk, M.R. Toliat, N. Klopp, A. Caliebe, SNP-based analysis of genetic substructure in the German population, *Hum. Hered.* 62 (2006) 20–29.
- [34] C. Phillips, L. Fernandez-Formoso, M. Garcia-Magarinos, L. Porras, T. Tvedebirk, J. Amigo, M. Fondevila, A. Gomez-Tato, J. Alvarez-Dios, A. Freire-Aradas, Analysis of global variability in 15 established and 5 new european standard set (ess) strts using the ceph human genome diversity panel, *Forensic Sci. Int.: Genet.* 5 (2011) 155–169.
- [35] J. Amigo, C. Phillips, T. Salas, L.F. Formoso, A. Carracedo, M. Lareu, pop.str - an online population frequency browser for established and new forensic strts, *Forensic Sci. Int.: Genet. Suppl. Ser.* 2 (2009) 361–362.
- [36] M. García, C.I. Catanesi, G.A. Penacino, L. Gusmão, N. Pinto, X-chromosome data for 12 STRs: Towards an Argentinian database of forensic haplotype frequencies, *Forensic Sci. Int.: Genet.* 41 (2019) e8–e13.
- [37] M.D. Vigeland, Pedigree Analysis in R, Academic Press, 2021.
- [38] T. Egeland, N. Pinto, A. Amorim, Exact likelihood ratio calculations for pairwise cases, *Forensic Sci. Int.: Genet.* 29 (2017) 218–224.
- [39] I. Simonsson, P. Mostad, Stationary mutation models, *Forensic Sci. Int.: Genet.* 23 (2016) 217–225.
- [40] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J. Sanchez, M. Müller, pROC: an open-source package for R and S+ to analyze and compare ROC curves, *BMC Bioinforma.* 12 (2011) 1–8.
- [41] M. Kruijver, R. Meester, K. Slooten, p-Values should not be used for evaluating the strength of DNA evidence, *Forensic Sci. Int.: Genet.* 16 (2015) 226–231.
- [42] D.W. Gjertson, C.H. Brenner, M.P. Baur, A. Carracedo, F. Guidet, J.A. Luque, R. Lessig, W.R. Mayr, V.L. Pascali, M. Prinz, ISFG: recommendations on biostatistics in paternity testing, *Forensic Sci. Int.: Genet.* 1 (2007) 223–231.