SAPIENZA
Università di Roma
Facoltà di Scienze Matematiche Fisiche e Naturali

DOTTORATO DI RICERCA
IN GENETICA E BIOLOGIA MOLECOLARE

XXXIV Ciclo
(A.A. 2021/2022)

Investigative leads in forensic genetics: Biogeographical Ancestry (BGA) and Kinship Analyses

Dottorando
CHIARA DELLA ROCCA

Docente guida
Prof. Fulvio Cruciani
Fulvio Cruciani

Tutore
Dr. Filippo Barni

Coordinatore
Prof. Fulvio Cruciani

Table of contents

1 Abstract

The fundamental goal of forensic genetics is personal identification but it is common to obtain inconclusive results from both *direct* and *indirect* STR-profile comparisons. In such cases, it is necessary looking for alternative approaches to generate crucial forensic leads and identify unknown perpetrators.

In this thesis, we propose alternative tools to predict BioGeographical Ancestry (BGA) and to identify males sharing the same Y-haplotype by utilizing multivariate techniques and Rapidly Mutating Y-chromosome short tandem repeats (RM Y-STRs), respectively.

Concerning ethnic inference, we proposed novel statistical approaches (consisting in SLPCA, PLSDA and SVM methods) to group samples into BGA-classes, by testing both autosomal STRs (in African populations) and microhaplotypes (in U.S. populations). The predictive power of such statistics resulted extremely high; in fact, they enhance cluster separation providing misleading classifications for genetically mixed populations only.

As to Y haplotype discrimination improvement, we proved the efficiency in individualization power of RM Y-STRs – reaching a total of 48 markers genotyped – in African populations characterized by high levels of endogamy, patrilinearity and population structuring.

Together, these two innovative approaches converge in demonstrating they represent powerful tools to maximize the information inferable from biological evidence collected at the crime scene.

2 General introduction

2.1 Overview of investigative forensic tools

DNA profiling of biological evidence such as those recovered from crime scenes, mass-disaster areas or missing person investigations is one of the most challenging topics in forensic sciences. When a crime is committed, forensic scientists must follow standardized operative protocols (SOPs) consisting in a hierarchical workflow, to obtain the individual genetic profile. If there are one or more suspects, the profile undergoes a “direct comparison” process, while if there are no suspects it undergoes an “indirect comparison” process, which consists in a database search – in Italy the database is called Banca Dati Nazionale del DNA (Law No. 85 of June 30, 2009 and Presidential Decree No. 87 of April 7, 2016). Both processes could lead to identify the person who left the evidence – match or compatibility – or, unfortunately, results in an “*information gap*”, when no match or no compatibility occurred. Through the years, DNA typing has been more and more employed, exploiting large sets of genetic markers that can be simultaneously analyzed on a single biological sample or trace, even if containing only a few copies of DNA [1, 2], to maximize the information inferable from the genetic profile. While Short Tandem Repeats (STRs) remain the mainstay of forensic analysis [3], several tools have been developed that can be used in addition to well-established techniques [4, 5, 6]. Next Generation Sequencing (NGS) or Massive Parallel Sequencing (MPS) technology could provide a platform that facilitates the use of alternative markers in more laboratories. [7, 8]. Currently, this technology enables genotyping a large number of Short Tandem

Repeats (STRs) loci in addition to an ever-growing number markers such as, for example, autosomal and Y- chromosome Single Nucleotide Polymorphisms (SNPs) and mitochondrial DNA (mtDNA) variants. This could provide additional information as putative age, alleged appearance (visible external characteristics such as hair, eye, and skin color [9, 10]), biogeographical ancestry inference [11, 12, 13] and potential kinship relationship [14].

Among other applications, genetic profile's capability to distinguish biogeographic information among population groups, subgroups and affiliations may have several positive pitfalls in leading investigative activities and, for this reason, has been largely studied and explored in the last decade but, at the same time, represents one of the most challenging applications in forensics because it usually shows a broad area accuracy [13]. BGA inference is based on the concept that individuals belonging to the same population show a genetic similarity. Analogously, for kinship inference the genetic sharing is greater the closer the kinship scenario is. Current approaches for BGA estimation using STRs profiles are usually based on Bayesian methods, which quantify the evidence in terms of likelihood ratio, supporting or not the hypothesis that a certain profile belongs to a specific ethnic group. To date, while autosomal STRs markers are the elective tool for personal identification and kinship inference, they have been poorly employed as Ancestry Informative Markers (AIMs) as STR alleles which are identical by state but not identical by descent occur in different populations, mostly because of recurrent mutation (homoplasy). For this reason, the more evolutionarily stable SNPs, in the biparental and uniparental portions of the genome, are being usually employed to infer the biogeographical ancestry and ethnic origins (generally named as BGA) of individuals [13].

On the other hand, kinship inference is currently performed by comparing autosomal STR profiles of interest and then performing familial searching analyses, sometimes (in case of putative paternally related males) supported by the analysis of Y-chromosome STR [15, 16]. Y-STRs represent an invaluable tool for specific forensic purposes as unbalanced male-female mixture deconvolution, lineage characterization, familial searching, and male suspects exclusion. Nevertheless, the lack of allelic recombination of the male-specific region of the Y chromosome implies that Y-STR haplotypes may be shared by many individuals and, therefore, does not allow individualization to the degree that autosomal markers do [17]. Developments in Y-chromosome STR analysis continue to be carried out [18, 19], as in the case of the identification of new Y-STRs that allow for better discrimination between males sharing the same Y-STRs haplotype [3, 20]. In fact, these limitations have been partially overcome by the identification of 13 Y-STRs characterized by mutation rates higher than 10^{-2} mutation/generation, termed rapidly mutating Y-STRs (RM Y-STRs) [21, 28]. Recently, 12 additional RM Y-STRs were discovered and, together with the previous 13 ones, were proven to significantly refine the male relative differentiation capacity [22, 29]. Different studies demonstrated the high haplotype diversity and discrimination capacity reached by using the RM Y-STRs [23, 24, 25]. Nevertheless, few studies [26, 27] have investigated the potentiality of RM Y-STRs in distinguishing males sharing the same Y-STRs haplotype in regions characterized by high levels of haplotype sharing and strong genetic sub-structuring, as the African continent, due to cultural and social factors (i.e., endogamy, population structure and patrilocality).

These innovative applications may represent a powerful and dynamic tool for investigative and intelligence applications for law enforcement agencies whenever a standard genetic profile is obtained from an unknown DNA donor, and they deserve to be further investigated.

3 Specific introduction: investigative tools considered in this thesis

3.1 Biogeographical ancestry

Among other forensic applications, genetic profile’s capability to allow interference about biogeographic ancestry of unknown persons of interest (BioGeographical Ancestry, BGA) represents the most problematic character because it has broad area accuracy [13]. For these reasons, BGA has been largely explored in the last decade and current approaches using STRs profiles are based on Bayesian methods [30,31], which quantify the evidence in terms of likelihood ratio, supporting or not the hypothesis that a certain profile belongs to a specific ethnic group. Bayesian statistics have been applied to estimate the ethnic affiliation of unknown genetic profiles obtained with autosomal STRs in well-known software such as STRUCTURE [32], the Snipper App suite [33] and PopAffiliator 2 [34]. These approaches perform Bayesian evaluations by inferring the relationships between the allele frequencies of specific populations and the alleles observed in the individuals, which are recognized as part of such populations. This is done by computing the likelihood values of individuals belonging to each of the tested population groups, according to their relative allele frequencies. An advantage of these methodologies is that prior information about the samples can be considered during the advancement of the analysis [35]. In the case of multi-locus genotypes, the power to obtain large amounts of data from a single biological sample requires appropriate statistical strategies to extract as precise information as possible regarding its ancestry. In this context, Multivariate Data Analysis (MDA) techniques may provide useful advantages to infer ethnic affiliation or ancestry of unknown subjects' genetic profiles. These methods may simultaneously perform specific and sensitive discriminations among different groups. Software based on Likelihood Ratios (LR) traditionally involve the comparison of only two alternative hypotheses, while multivariate techniques may efficiently evaluate several population groups together. However, the likelihood-based methods for BGA estimation overcome this issue by computing the likelihood of membership to each of the populations under evaluation [35, 36]. The present thesis provides an alternative approach to the likelihood ratio method that involves Multivariate Data Analysis strategies for the estimation of multiple populations ethnic origin. In fact, we employ multivariate methodologies such as Sparse and Logistic Principal Component Analysis (SL-PCA) [37], Sparse Partial Least Squares-Discriminant Analysis (sPLS-DA) [38, 39, 40] and Support Vector Machines (SVM) [41, 42] on autosomal STRs data sets and on Microhaplotype (MHs) markers data sets. Microhaplotypes are emerging biomarkers of at least two Single Nucleotide Polymorphisms (SNPs) associated in multiple allelic combinations within 300bp. The multi-allelic nature of MHs make them more informative than a Single Nucleotide Polymorphism (SNP) locus and useful for forensically relevant applications, including mixture deconvolution and ancestry inference on massively parallel sequencing platforms [43, 44]. Due to presence of small amplicons and low recombination rate, absence of stutter and preferential amplification, they are promising candidates for biogeographical ancestry (BGA) prediction [45] from both single-source and mixed DNA profiles. In fact, biallelic SNPs are not effective in mixtures, whereas MHs are and this means that a deconvoluted MH profile in single contributors could be useful for predicting the ethnic origin of a minor or major contributor to a mixture. These multivariate techniques were selected as they have proven capable of dealing with the type of genotypic data generated as it can be easily binarized. Our goal was to develop multivariate approaches for the interpretation of DNA profiles to better estimate the biogeographical ancestry information of personal genetic profiles, by building dynamic and flexible models that could be easily modified according to the number of tested populations and the number of markers in the profile and the reference panel. Our multivariate statistics approach may represent a powerful tool for research and investigative purposes.

3.2 Consanguinity

The inference of kinship provides highly accurate information about the familial relationship between two people based on their DNA. These analyses are commonly performed by using Blind Search Analyses (BSAs) and Pedigree construction tool of several conventional software [46] used in forensic genetics or by reconstructing parental lineages using lineage markers, such as mitochondrial DNA (mtDNA) and Y-chromosome markers. Y chromosome STRs (Y-STRs) are widely used in forensic genetics, usually in addition to the autosomal STRs (aSTRs). The holandric inheritance and the lack of recombination imply that Y chromosome haplotypes are usually shared among paternal relatives [16, 17]. Nevertheless, the recent identification of rapidly mutating Y-STR markers (RM Y- STRs) characterized by a mutation rate higher than 10^{-2}/ STR/generation have been proven to be extremely useful in distinguishing among close male relatives [21, 22]. Thus far, the most discriminating Y-STR system commercially available for capillary electrophoresis is represented by the 25 Y-STR multiplex named Yfiler^{TM} Plus PCR amplification kit (ThermoFisher Scientific), which includes six “first generation” RM Y-STRs (five single copy and the two-copy system DYF387S1), while non-commercial multiplex assays, including 13 “first generation” RM Y-STRs [28] and the whole set of 25 RM Y-STRs [29], have been published. In principle, it can be expected that with enough RM Y-STRs available, closely, and especially distantly related men will be distinguishable through observed mutations [16].

However, the performance of multiplexes containing RM Y-STRs (e.g. Yfiler Plus) in populations characterized by high level of endogamy has not been fully investigated so far.

In this thesis, we investigated familial relationship and paternal lineage of 1370 males from African continent, and we assessed the power of those novel genetic markers located on Y-chromosome to improve the discrimination power of male-specific markers in regions characterized by high levels of endogamy. Specifically, we first analyzed the putative kinship relationships among these males using 16 autosomal STRs and the Blind Search Analysis (BSAs) tool of the Familias software [46]. Subsequently, we assessed the male individualization power of “first-generation” RM Y-STRs using the RM-Yplex assay developed by [28] and then using the novel 30 Y-STR markers developed by [29].

4 Materials and Methods

4.1 Autosomal STRs

4.1.1 Sample

Four different population datasets were selected for this study. All the datasets consisted of individual genotypes rather than allele frequencies.

The first dataset consisted of original unpublished genotypes from Northern, sub-Saharan and Eastern African populations analyzed for 16 autosomal STRs loci using the AmpFlSTR® NGM SElect™ PCR Amplification Kit from Thermo Fisher Scientific, i.e. 477 Northern Africans (from Algeria, Egypt, Libya and Morocco), 431 sub-Saharan Africans (from Cameroon and Chad) and 462 Eastern Africans (from Eritrea, Ethiopia, Djibouti and Kenya). All the biological samples included in this dataset were randomly collected from informed people. Despite efforts to avoid the inclusion of relatives during the sampling process, the presence of related males in the sample could not be excluded due to the unavailability of genealogical information. For each subject, the ethnic identity was assessed by self-identification. This study ethically complies with the ISFG guidelines for the publication of genetic population data [47] and was formally approved by the "Reparto Carabinieri Investigazioni Scientifiche di Roma".

The second dataset was extracted from the NIST U.S. population database and consisted of genotypic data for U.S. African-American (N = 342), Asian (N = 97) and Caucasian (N = 361). For this dataset, the following 24 markers were selected: D1S1656, D2S441, D2S1338, D3S1358, D5S818, D6S1043, D7S820, D8S1179, D10S1248, D12S391, D13S317, D16S539, D18S51, D19S433, D21S11, D22S1045, CSF1PO, FGA, Penta D, Penta E, SE33, TH01, TPOX, vWA. Markers F13A01, F13B, FESFPS, LPL and Penta C, which are present in the NIST U.S. population database, were not considered in this study since they are usually not included in commercially available autosomal STR amplification kits commonly used in forensic laboratories.

The third dataset comprised two central Asian populations genotyped for 15 autosomal STRs loci using the AmpFlSTR® IdentifilerTM PCR Amplification Kit panel from (Thermo Fisher Scientific) (D3S1358, vWA, FGA, D8S1179, D21S11, D18S51, D5S818, D13S317, D7S820, TH01, TPOX, CSF1PO, D19S433, D2S1338, D16S539), i.e. 65 unrelated Afghan and 103 Iraqi (mainly from central and southern Iraq provinces).

The fourth dataset comprised two populations genotyped for 16 autosomal STRs loci using the AmpFlSTR® NGM SElectTMPCR Amplification Kit (Thermo Fisher Scientific) (D10S1248, vWA, D16S539, D2S1338, D8S1179, D2S11, D18S51, D22S1045, D19S433, TH01, FGA, D2S441, D3S1358, D1S1656, D12S391, SE33), i.e. 209 unrelated Italian individuals, and 287 Eastern Europeans (223 Romanian and 64 Moldavian subjects).

For each dataset, we evaluated the inter-population genetic differentiation using the F_{ST} statistics (an index of the co-ancestry for randomly chosen alleles within the same subpopulation relative to the entire population) to have a convenient metrics to objectively measure genetic differentiation among populations when estimating BGA of individuals belonging to such populations. F_{ST} values were obtained using the software STRAF v. 1.0.5 [48].

4.1.2 Autosomal STRs DNA Typing

DNA samples from the first dataset were extracted either from blood using a standard phenol-chloroform protocol or from saliva or cell lines using the EZ1&2 DNA Investigator Kit (Qiagen) on a BioRobot EZ1® Advanced XL Workstation (Qiagen). DNA quantification was performed using Quantifiler® Trio DNA Quantification Kit (Thermo Fisher Scientific) and/or QubitTM 4 Fluorometer (Thermo Fisher Scientific). Multiplex amplification of 16 autosomal STRs (D10S1248, vWA, D16S539, D2S1338, D8S1179, D21S11, D18S51, D22S1045, D19S433, TH01, FGA, D2S441, D3S1358, D1S1656, D12S391, SE33) was performed using the AmpFlSTR® NGM SElect™ PCR Amplification Kit (Thermo Fisher Scientific) and 1 ng of genomic DNA, according to manufacturer's protocol. PCR conditions were set up on an Applied Biosystem® Veriti™ 96-Well Thermal Cycler (Thermo Fisher Scientific). Amplified DNAs were then electrophoresed on the 24-capillary Applied Biosystems® 3500XL Genetic Analyzer (Thermo Fisher Scientific), and the fragment analysis was performed throughout GeneMapper® ID-X v.1.6 (Thermo Fisher Scientific). The authors followed ISFG recommendations and internal protocols complying with the requirement ISO17025 for the polymorphism analysis and interpretation [49, 50, 51, 52].

4.2 Y chromosome STRs

4.2.1 Sample

Samples belonging to the first dataset (1370 males from Northern, sub-Saharan and Eastern African populations previously analyzed for aSTRs) were subsequently analyzed for 27 Y-chromosome STRs loci using the Yfiler^{TM} Plus PCR Amplification Kit (Thermo Fisher Scientific). This study ethically complies with the ISFG guidelines for publication of genetic data [15, 17] and was formally approved by the “Reparto Carabinieri Investigazioni Scientifiche di Roma" and by the “Sapienza Università di Roma" Ethical committee (Document number 2755/15).

4.2.2 27 Y-STRs DNA typing

DNAs were multiplex amplified for 27 Y-STRs (DYS576, DYS389I, DYS635, DYS389II, DYS627, DYS460, DYS458, DYS19, GATAH4, DYS448, DYS391, DYS456, DYS390, DYS438, DTS392, DYS518, DYS570, DYS437, DYS385, DYS449, DYS393, DYS439, DYS481, DYS387S1, DYS533) using the Yfiler^{TM} Plus PCR Amplification Kit (Thermo Fisher Scientific).

Amplification was performed on an Applied Biosystem® Veriti^{TM} 96-Well Thermal Cycler (Thermo Fisher Scientific) according to the manufacturer’s protocol utilizing 1 ng of genomic DNA. Amplified DNAs were then electrophoresed on the 24-capillary Applied Biosystems® 3500xL Genetic Analyzer (Thermo Fisher Scientific) and the fragment analysis was performed with GeneMapper® ID-X software v.1.4 (Thermo Fisher Scientific).

Haplotype data were submitted to the Y-chromosomal haplotype reference database (www.yhrd. org) [53] (accession numbers YA003983, YA 004045, YA 004198 – YA 004207, YA 004351 – YA 004356, YA 004668 – YA 004669). The contributors successfully passed the quality control test.

4.2.3 13 RM Y-STRs DNA typing

Among 1370 male samples from the first dataset previously analyzed with the Yfiler Plus kit, 240 individuals were reported to share 100 distinct Y-STR haplotypes. All those 240 males were analyzed for 13 Rapidly Mutating Y-STRs (RM Y-STRs) – DYF387S1, DYS399S1, DYS403S1a/b, DYF404S1, DYS449, DYS518, DYS526a/b, DYS547, DYS570, DYS576, DYS612, DYS626 and DYS627 – described by [20]. Multiplex amplification of the 13 RM Y-STRs was performed using the 13-locus RM-YPlex assay described in [28] and 1 ng of genomic DNA. Amplification was set up on an Applied Biosystem® Veriti™ 96-Well Thermal Cycler (Thermo Fisher Scientific) and, subsequently, amplified DNAs were electrophoresed on the 8-capillary Applied Biosystems® 3500 Genetic Analyzer (Thermo Fisher Scientific). Then, the fragment analysis was performed with GeneMapper® *ID-X* v.1.6 (Thermo Fisher Scientific).

4.2.4 24 RM Y-STRs + 6 FM Y-STRs

Among 240 male samples from the first dataset previously analyzed with the Yfiler Plus and RM YPlex assays, 107 individuals were reported to share 50 distinct Y-STR haplotypes. All those 107 males were analyzed for further 24 Rapidly Mutating Y-STRs (RM Y-STRs) and 6 Fast Mutating Y-STRs (FM Y-STRs) using the RMplex assay described in [29]. The PCR reactions of the 30 RM Y-STRs were performed on an Applied Biosystem® VeritiTM 96-Well Thermal Cycler (Thermo Fisher Scientific) in two different multiplexes:

- ✓ 1° multiplex made up of 16 Y-STRs: DYF393S1, DYS627, DYS570, DYS713, DYS526b, DYF1000, DYS518, DYS1003, DYS1012, DYS1005, DYS101, DYS1007, DYR88, DYF404S1, DYF387S1, DYS1013
- ✓ 2° multiplex made up of 14 Y-STRs: DYS712, DYS711, DYS626, DYF399S1, DYS449, DYS724, DYS547, DYS576, DYS612, DYF1002, DYF1001, DYF404S1a, DYS442, DYF403S1b

Every multiplex reaction was amplified with the same PCR protocol, according to the suggested protocol [29], as follow: 94 °C for 10 min, 10 cycles of 94 °C for 30 s, 65-1 °C every cycle for 60 s and 72 °C for 60 s, followed by 25 cycles of 94 °C for 30 s, 50 °C for 30 s and 72 °C for 60 s with a final extension step of 60 °C for 45 min. After amplification, 1 μL of the PCR product was mixed with 9 μL of Hi-Di formamide (Thermo Fisher Scientific Inc.) and with 0.3 μL of ILS600 size standard (Promega Corporation). This mixture was incubated at 95 ℃ for 3 minutes and rapidly cooled on ice for 5 minutes. Capillary

electrophoresis was performed on an ABI 3500XL Genetic Analyzer (Thermo Fisher Scientific Inc.) The resulting electropherograms were analyzed using GeneMapper® ID-X v.1.6 (Thermo Fisher Scientific).

4.3 Microhaplotypes

4.3.1 Sample

Five different populations were selected for this study. Four populations represent the four major United States groups and are composed of 88 Afro-American (AA), 114 European American (EA), 102 Southwest

Hispanic (His), and 43 East-Asian American (EAA), respectively. The fifth population comprised 129 admixed individuals (ADMIX); precisely subjects that have the mother and the father coming from 2 different populations or individuals that belong to a genetically admixed population (for example as Puerto Rican, Dominican, American Indian, Vietnamese, Cuban, Mexican, Jewish and St. Lucia).

4.3.2 Next Generation Sequencing with Ion S5 Technology

NGS data for the 347 subjects belonging to the four US population datasets were already available from [45] while the 129 Admixed individuals were genotyped using a 74 MH bioassay on the Ion S5$^{\mathrm{TM}}$ System sequencing platform [44].

Preparation of DNA libraries was manually performed in half- reaction volume using the Precision ID Library (Thermo Fisher Scientific) kit according to the manufacturer's protocol and as outlined below.

Amplification of DNA targets was performed using 5 µL of 74plex MH primer mix assay (primers were pooled equimolar) or 5 µL Precision ID GlobalFiler$^{\mathrm{TM}}$ NGS STR Panel v2, 2 µL of 5X Ion AmpliSeq$^{\mathrm{TM}}$ Mix from the Precision ID Library (Thermo Fisher Scientific) kit, pre-quantified reference and mixed samples at 1 ng and 10 ng, and nuclease-free water. Thermal cycling was performed on the GeneAmp® 9700 System (Thermo Fisher Scientific) using the following PCR amplification conditions: enzyme activation for 2 min at 99 ◦C, amplification for 21 cycles for MHs and 23 cycles for STR Panel v2, denaturation for 15 s at 99 ◦C, annealing/extension for 4 min at 60 ◦C, and hold at 4 ◦C.

To partially digest the ends of MH and STR amplicons, these were treated with 1 μL FuPa Reagent and incubated for 10 min at 50 ◦C, 10 min at 55 ◦C, 20 min at 60 ◦C, and held up to 1 h at 10 ◦C.

Ion P1 Adaptors and Xpress™ Barcodes were ligated to the FuPa digested amplicons. 2 μL of Switch solution, 1 μL of diluted Ion Xpress™ Barcode (barcodes 1–96) and P1 Adapter mix and 11 μL digested PCR reaction were mixed and incubated for 30 min at 22 ◦C, 10 min at 72 ◦C for the panel of MHs and at 68 ◦ C for the Precision ID GlobalFiler™ NGS STR Panel v2 and held up to 1 h at 10 ◦C. Each DNA library was barcoded with a distinct barcode number to enable library pooling. Each library was purified with 1.5X Agencourt® AMPure® XP reagent (Beckman Coulter, FL, USA), as recommended by the manufacturer.

Each DNA library was diluted down to 1:100 and quantitated on the Applied Biosystems 7500 Real-Time PCR System (Thermo Fisher Scientific) following the protocols outlined in the Ion AmpliSeq™ Library Preparation for Human Identification Applications User Guide and on the Ion Library TaqMan™ Quantitation (Thermo Fisher Scientific) Kit [54]. After quantification, libraries were pooled in equimolar amounts, diluted down to approximately 60 pM, as recommended by the manufacturer and run on the 7500 Real-time PCR machine using the Ion Library TaqMan™ Quantitation Kit. Results were analyzed using the HID Real-Time PCR Analysis Software v. 1.2 (Thermo Fisher Scientific). Barcoded DNA libraries were then pooled equivolume. A maximum of 20 DNA libraries was loaded per chip to maximize the read depth for the mixed samples. To verify the correct input amount of pooled library, 1:10 and 1:100 dilutions of the same pool library were prepared. The library-pool was re-quantified to make sure to load the expected library amount into the Ion Chef system. A total of 25 μL of approximately 30 pM and 50 pM library-pool for MHs and STRs, respectively, was loaded into the Ion Chef^{TM} (Thermo Fisher Scientific) system for templating (i.e., ion sphere^{TM} particles enrichment reaction), as recommended by the manufacturer.

The Ion Chef^{TM} system (Thermo Fisher Scientific) was run for template preparation using the Ion S5^{TM} Precision ID Chef & Sequencing Kit. The process involves emulsion PCR (library amplification), ion sphere^{TM} particles recovery and enrichment (carrying target sequence template), and chip loading. Final chip loading involves 25 μL of each equimolar – pooled library along with required reagents and consumables following the manufacturer's recommendation. For the emulsion PCR process, 27-PCR-cycle default protocol was used for STR mixtures and the 45-PCR-cycle default protocol for MH mixtures, as indicated by the manufacturer. Templated/enriched ion sphere^{TM} particles were loaded on the Ion 530^{TM} chip that contains > 30 million wells while sequencing was performed on the Ion S5TM platform, which allows ~400 bp read length. Cartridge reagents, wash and sequencing solutions, and 2 Ion530TM chip were loaded on the instrument and sequencing flows were set to "650" and "850" for STR and MH assay sequencing reactions respectively, as recommended by the manufacture.

Sequencing data were processed using the Ion Torrent Suite Software v. 5.10.0. For MHs, the TVC Microhaplotyper plugin v. 8.1 (Thermo Fisher Scientific) was used to analyze the sequencing reads of each library and output display. The plugin was run to locate the MH regions within Homo sapiens GRCh37/hg 19 reference genome by unaligned BAM files for each barcoded library and a panel of target and hotspot

BED file (mh74_targets and mh74_hotspot) to genotype MH loci and generate output files. These include the TVC Microhaplotyper plugin v 8.1 report detailing information on genotype, coverage, allele sequence and coverage plot for each MH locus and two TXT files generated per each batch result: one filtered TXT file including marker ID, number of alleles and allele sequence and one unfiltered TXT file displaying marker ID, allele coverage (coverage minus and plus).

4.4 Statistical analyses:

4.4.1 Kinship analyses

Kinship for the first autosomal STR dataset consisted of 1370 genotypes from Northern, sub-Saharan and Eastern African populations was assessed using the Blind Search Analysis (BSA) module of the Familias software v. 3.2.8 [46] to reveal putative presence of close relatives and provide, for each alleged kinship, a Likelihood Ratio (LR) value.

We assumed a stepwise mutation model with a mutation rate of 0.001 and the mutational range fixed to 0.1. The fixation index (F_{ST}) value and the typing error rate were set at 0.03 and 0.001, respectively, while drop-in and drop-out were assumed to be absent. The direct-match, parent-child, siblings, second-degree relatives, cousins, and second cousins’ relationships were investigated. For each pair of subjects, the alleged relationship having the highest likelihood ratio (*LR*) value was assumed to be the right one. Since kinship analysis relies on population allele frequencies at the denominator of the LR, we used the Afro-American validated autosomal allele frequency database published in [55] for all the pairs of subjects.

10,000 simulation tests were made for each population/kinship scenario to establish LR thresholds. FST= 0.03 and �= 2 x 10^{-3}. Relationships above an LR threshold of 10^{2} were considered as strongly supported [56, 57], inferred kinship between pairs of subjects with 1<LR<100 were considered as moderate or weakly supported, while pairs of subjects with LR<1 were classified as unrelated.

4.4.2 Y-STR pairwise comparison and mutational analysis

We compared RM Y-STR haplotypes of males sharing the same Yfiler Plus haplotype (100 shared haplotypes, 240 males). The 6 RM Y-STRs overlapping between Yfiler Plus and RM-YPlex (i.e., DYS570, DYS576, DYS518, DYS627, DYF387S1, and DYS449) were checked to assess genotyping consistency between the two PCR assays, while the 7 “first generation” RM Y-STR (DYF399S1, DYF403S1a/b, DYF404S1, DYS526a/b, DYS547, DYS612, and DYS626) which were not included in the Yfiler Plus multiplex, were used to evaluate their power in discriminating between pairs of males. Moreover, the proportion of pairs of males differentiated by the 7 RM Y-STRs was assessed, considering the inferred kinship degree (if any). For each of 100 groups of males, the number of mutations occurring at each locus and for each group of males sharing a Yfiler Plus haplotype was counted under the most conservative scenario (i.e., allowing for single multi-repeat mutations rather than multiple single-repeat mutations). The “phylogenetic” relationships among males belonging to the three major clusters (H12, H21, and H69) were depicted through the UPGMA clustering method using the Manhattan genetic distance as implemented in PAST v. 4.09 software [58].

4.4.3 Forensic parameters

Forensic parameters were calculated for the dataset consisted of Y chromosome STRs from Northern, sub-Saharan and Eastern African populations were calculated to evaluate the discrimination power achieved with Yfiler (17 loci) and Yfiler Plus (25 loci) compared to the Yfiler Plus supplemented with seven [28] and twelve [29] additional RM Y-STRs loci (32 loci and 37 respectively).

Y chromosome haplotype sharing was evaluated using the "profile comparison" function of GeneMapper ID-X®. Then, the number of distinct shared and unique Y chromosome haplotypes was counted, where the number of distinct haplotypes corresponds to the sum of unique and shared haplotypes. Discrimination capacity (*DC*) was calculated as the ratio between the number of distinct haplotypes and the total number of chromosomes in the dataset, while the proportion of matching haplotypes (*MH*) was calculated as the ratio between the number of males sharing a Y-haplotype and the total number of chromosomes.

4.4.4 Bio Geographical Ancestry prediction by Multivariate Statistical Analyses

Bio geographical ancestry (BGA) prediction was performed for all the individuals belonging to the four autosomal datasets and for the five microhaplotype dataset selected for this study.

Multivariate models were built on the autosomal profiles, where each STR profile was converted into a row of zeros and ones by means of an in- house code developed in the R software (version 1.1.463) [59, 60] statistical environment. In details, for all the tested individuals, a value equal to 1 was reported for the alleles x and y (where x is equal to y in case of homozygosity) recorded for a specific marker Z, while a value equal to 0 was reported for the other n available alleles of the previously cited marker Z. Consequently, the STRs DNA profile of each individual was converted into a series of zeros and ones (i.e., binary dataset). Since the matrices obtained by using such computational approach turned to show many zeros as compared to the number of ones, sparse algorithms had to be considered when calculating the multivariate models.

SL-PCA, sPLS-DA and SVM multivariate techniques were employed to obtain reliable models for the estimation of the BGA information of unknown genetic profiles. Multivariate modelling and calculations were carried out in R (version 3.6.0) [59, 60]. The following functions and R packages were used to build in-house R code for computing the different models: sparse logistic Principal Component Analysis [37], mixOmics [61] and e1071 [62]. In-house developed codes will be available to the readers upon requests to the authors.

Initially, SL-PCA was utilized as an exploratory analysis tool to verify the capabilities of multivariate statistics in recognizing specific pattern regarding the biogeographical origins of the individuals based on their STR profiles, especially when dealing with binary data (as reported above). PCA, here employed in the *sparse* and *logistic* version reported in [37], is one of the most exploited techniques in the field of multivariate statistics; it allows to graphically represent the information contained into large data matrices by providing useful visual re-presentations of data distributions, similarity trends, classes and outliers [63]. In practice, PCA evaluates the original data collected for several "objects" (i.e., the encoded individuals), by re-modelling them within new Cartesian diagrams. The new axes of these diagrams represent the Principal Components (PCs), defined as a linear combination of the original variables to make them reciprocally orthogonal.

After the preliminary evaluation of SL-PCA modelling, sPLS-DA and SVM models were applied, to assess their predictive capabilities in blind inference of the ethnic affiliation of DNA profiles. sPLS-DA is the *sparse* version of the combination of Partial Least Squares (PLS) and Discriminant Analysis (DA) techniques [40, 64, 65]. In practice, sPLS- regression finds the factors that capture the greatest amount of variance in predictor variables by simultaneously modelling those X predictors that optimally correlate the responses of the Y matrix. Briefly, the PLS algorithm indicates that the Y responses are proportional to the first principal component – named as Latent Variable (LV) – except for some residuals; then, residuals turn proportional to the second LV, except for new residuals, *etc*. Afterwards, the slopes of the regression line – named as PLS weights – are calculated as residual regression coefficients and indicates the direction of the first LV. The variables/predictors are not usually independent and PLS may provide a bilinear projection model, plus some residuals. Because of that, PLS admits that some X-data are not correlated to Y-responses; these data can represent noise or redundancy, thus indicating that PLS tolerates noisy or redundant data, unlike other regression methodologies. On the other hand, LDA is a supervised classification method whose goal is to discriminate different classes of objects by evaluating the optimal boundaries among them. Originally developed by Fisher [38], LDA allows discriminating objects of different classes by examining the probability distributions of the classes to which the objects may belong. Accordingly, each object is classified in the specific class which shows the highest score in terms of probability. Graphically, the probability distributions are expressed as ellipses at different probability levels for each class under examination. These ellipses are respectively tangential to a point that is located half- way among the class centers and a straight delimiter is adopted as a boundary to separate the ellipses and, consequently, the different classes. LDA provides a linear function of the variables and maximizes the ratio between the variances of each class; weights are adopted to provide the best classification of the objects so that LDA can select the direction achieving the maximum separation among the given classes.

Finally, SVM is a Multivariate Data Analysis (also known as Machine Learning) methodology usually adopted for pattern recognition tasks. Very concisely, this methodology was developed by Vapnik [42] with the aim to provide a decision rule in terms of a special type of hyper-planes, defined as “optimal separating hyperplanes” and known as “delimiter” or “margin” [41], capable of recognizing and discriminating the objects of different sets or classes. The delimiter is optimized as the distance between the separating decision boundaries (hyperplanes) and the closest objects to these hyperplanes, which are defined as support vectors. As reported by Vapnik [42], SVM techniques map the objects matrix X into a high-dimensional space called “feature space”; then linear or nonlinear functions (such as kernels) may be adopted to build an optimal separating hyperplane in this space.

All the multivariate models were assembled adopting the 70 % of the available data as training set and the remaining 30 % of data was employed as evaluation set. Repeated double cross-validation procedures were performed by applying a venetian blind design and a number of data splits equal to 5 (i.e., 80 % of the available data of the training set was employed to build the models), in accordance with [66]. Finally, sensitivity and specificity parameters were calculated for all the sPLS-DA and SVM models, as follows: (i) sensitivity is equal to the proportion of individuals belonging to a specific bio-geographical origin that are correctly identified as such, while (ii) specificity is equal to the proportion of individuals belonging to another bio-geographical origin (with reference to the one that is considered by the model) and that are correctly identified as such.

5 Results

5.1 Autosomal STRs:

5.1.1 Allele Frequencies Database

Allele frequencies databases are empirically determined from sets of randomly selected human samples genotyped for autosomal STRs. They represent the basis of population studies in forensic genetics. The reliability and the accuracy of the data are largely based on the responsibility of the individual contributing research groups and centralized quality control and data curation is essential to minimize error [47].

Allele frequencies for 16 STR loci included in NGM SElected™ kit (D3S1358, vWA, D16S539, D2S1338, D8S1179, D21S11, D18S51, D19S433, TH01, FGA, SE33, D10S1248, D22S1045, D2S441, D1S1656, D12S391 and Amelogenin) were determined in African population dataset (including 1370 subjects, see materials and methods) subdivided by the three main regions (Northern Africa, N=477; Sub-Saharan Africa, N=431; and Eastern Africa, N=462). The three datasets have been submitted to online publication [47, 67 **STRidER dataset reference STR000291**] and are reported in Supplementary Table 1.

5.1.2 Multivariate statistics

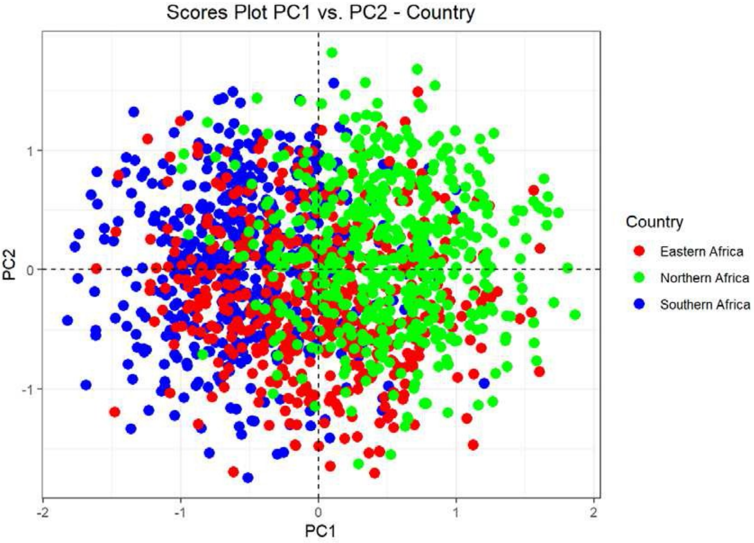
SL-PCA, sPLS-DA and SVM multivariate techniques were employed to obtain reliable models for the estimation of the BGA information of unknown genetic profiles. Multivariate modelling and calculations were carried out in R (version 3.6.0) and RStudio (version 1.1.463) [59. 60] and results are listed below.

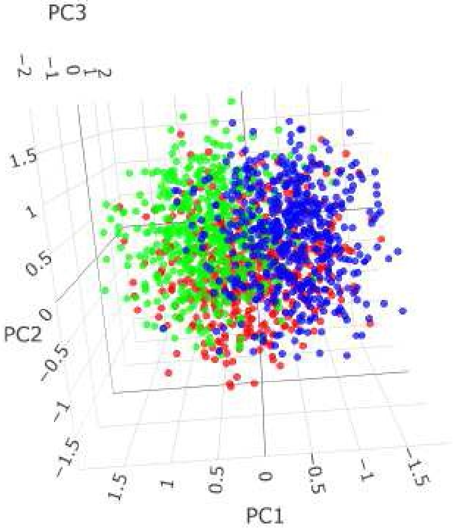
SL-PCA analysis: SL-PCA was first exploited to rapidly investigate the main features in the dataset. As expected, good separation was not observed for the SL-PCA comparison involving the Northern, the Eastern and the Sub-Saharan African individuals (Fig. 1). Among the three datasets, the best separation was observed between Northern and sub-Saharan samples and this may be due to the fact that the Sahara Desert acted as a strong geographic barrier to gene flow between the cited populations in the last five thousand years [68]. In contrast, the East African samples show full overlap with both North African and sub-Saharan African datasets, confirming the extensive gene flow existing between these areas.

In summary, this traditional multivariate procedure allowed us to observe the pertinence of more advanced multivariate statistics in assessing and recognizing the biogeographical ancestry information by evaluating the autosomal STRs DNA profiles, only. Although a fair degree of separation was observed, there is still an important overlap between all three African regions [71, 72].

This could be explained because whenever the populations to be compared showed quite similar STR allele frequencies they would be expected to return unsatisfactory results. For example, when F_{ST} value is higher than the conventional 0.001, as in those cases, we sought to assay more sophisticated and classification-like multivariate models (such as sPLS-DA and SVM techniques, described in the following) to possibly obtain satisfactory separations between the populations and hence better chances of individual assignment.

a)





b)

Fig. 1. SL-PCA PC1 vs. PC2 PCA 2D (a) and 3D (b) Scores Plot. for sub-Saharan (blue) vs Northern-African (green) vs Eastern African (red) subjects.

sPLS-DA analysis: Based on results provided by PCA modelling, sPLS-DA was applied to the same experimental sets to develop useful discrimination models (Fig. 2). The predictive models were evaluated in terms of Root Mean Square Error in Cross-Validation (RMSECV) [69], i.e. the lower the RMSECV value, the higher the discrimination power of the model. Moreover, the number of LVs was determined through the evaluation of further quality parameters such as the Predictive Residual Error Sum of Square (PRESS), Q-residuals, Hotelling's T2, Leverages and Y-Studentized residuals [69]. Sensitivity and specificity values were calculated too and are reported in the ROC Curve (Fig. 3).

By applying this model, the three datasets showed a better discrimination than PCA one, especially for Northern and Sub-Saharan samples which differ mainly along LV1, in agreement with the relatively high inter-population genetic diversity observed (FST values). The AUC values obtained for the first component are equal to 0.516, 0.881 and 0.865, for the second component are equal to 0.813, 0.908 and 0.892, while for the third component are equal to 0.830, 0.934 and 0.895 for Eastern, Northern and Sub-Saharan samples, respectively. These data allow us to affirm that sPLS-DA improves separation of autosomal STR profiles in respect of PCA and might represent a useful tool for improving the routine estimation of the BGA information.

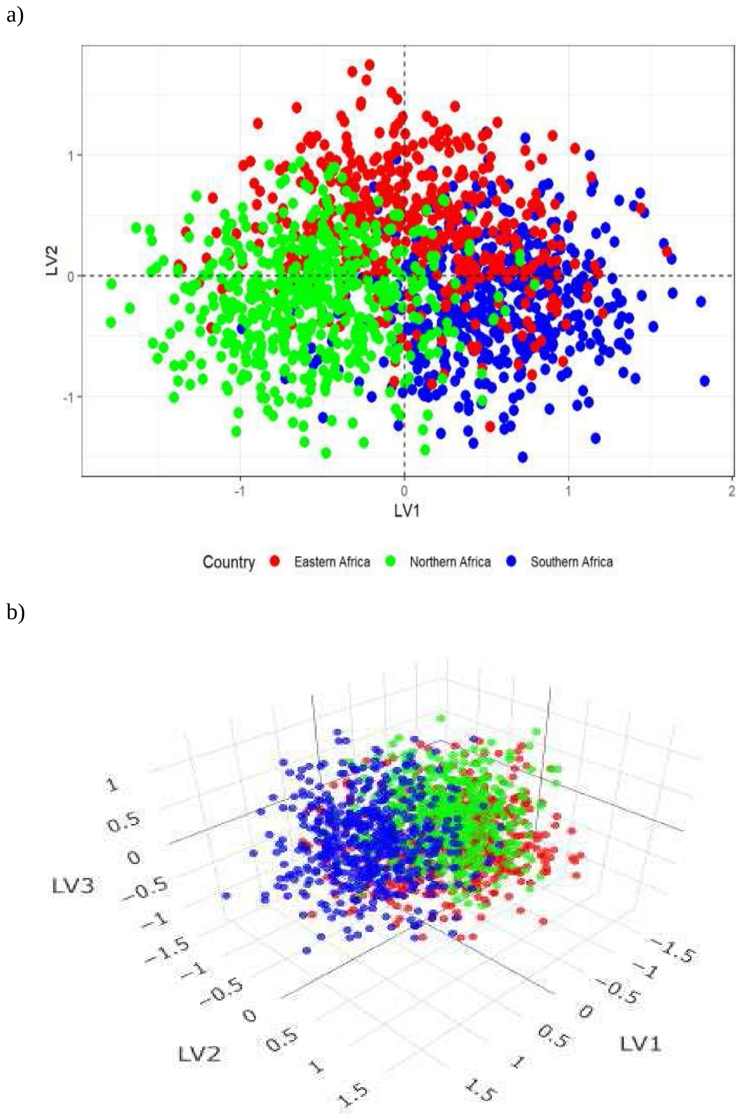


Fig. 2. PLS-DA LV1 vs. PLS-DA LV2 2D (a) and 3D (b) Scores Plot for sub-Saharan (blue) vs Northern-African (green) vs Eastern African (red) subjects.

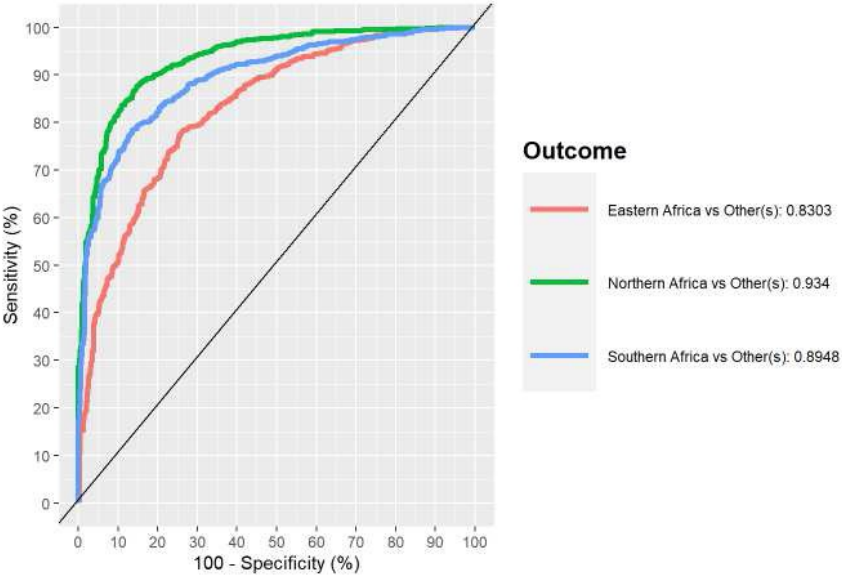


Fig. 3. Receiver Operating Characteristic (ROC) curves outcomes from PLSDA analyses using 1, 2 and 3 components for sub-Saharan (blue) vs Northern-African (green) vs Eastern African (red) subjects.

SVM analysis: SVM was applied and the corresponding sensitivity and specificity values are reported in the resulting ROC curves (Fig. 4). Among the 407 Eastern Africans, 334 (0.82) were correctly assigned, while 30 and 43 were misclassified as Northern and Sub-Saharan samples, respectively. Concerning Northern Africa, 366 out of 375 (0.90) subjects were correctly assigned, while 31 and 11 were misassigned as Eastern Africans and Sub-Saharan Africans, respectively. Finally, 358 out of 398 (0.90) sub-Saharan Africans were correctly assigned, while 35 and 5 were misassigned as Eastern and Northern Africans, respectively.

Consequently, SVM turned out to be a very powerful model, with high specificity and sensitivity values for these ethnic groups, thus proving the reliability of multivariate statistics to extract BGA information from autosomal STRs DNA genetic profiles.

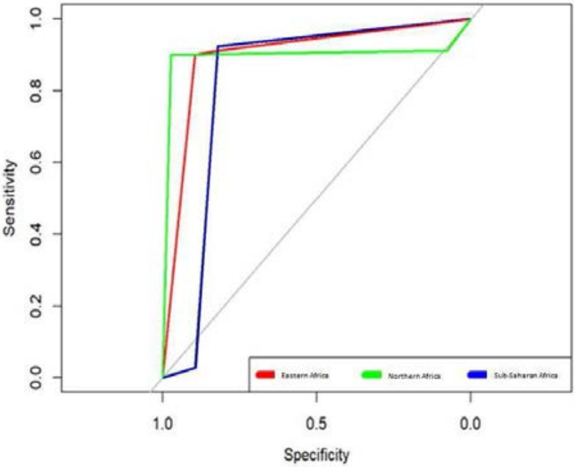
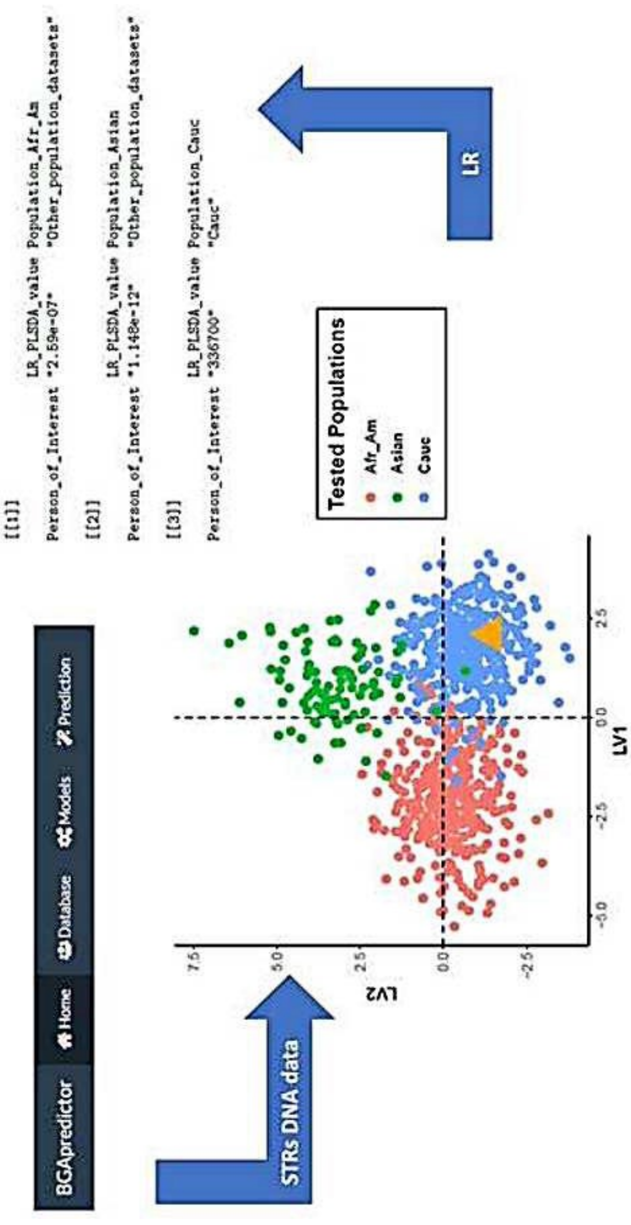


Figure 4 Receiver Operating Characteristic (ROC) curves outcomes from SVM analyses for sub-Saharan (blue) vs Northern-African (green) vs Eastern African (red) subjects.

5.1.3 Development of a new BGA predictor software

We are developing an open-source, freely available and user-friendly R Shiny app with an intuitive graphical user interface named "*BGApredictor*" (https://bgapredictor.shinyapps.io/BGApredictor/) [70] to let the forensic community take practical and operational advantage from these new multivariate statistical approaches for BGA estimation, overcoming the limits in using R software and packages. We are testing and validating the software over the datasets of autosomal STRs markers from our recently published studies [67, 71, 72] to made easily usable by any minimally trained analyst, but these preliminary results are not included in this thesis. Multivariate models are initially calculated on the target populations that have been imported into the R Shiny app. In case the genetic profile of a person of interest (i.e. POI) has been uploaded, the multivariate models are calculated using the matching loci between the POI and the populations under exam (Fig. 5). At the end, the user can save any displayed plot. As soon as we have completed testing and validation, we will proceed with manuscript submission [70].



```
[[1]]
                   LR_PLSDA_value Population_Afr_Am
Person_of_Interest "2.59e-07"     "Other_population_datasets"

[[2]]
                   LR_PLSDA_value Population_Asian
Person_of_Interest "1.148e-12"    "Other_population_datasets"

[[3]]
                   LR_PLSDA_value Population_Cauc
Person_of_Interest "336700"       "Cauc"
```

Fig. 5. Workflow of BGA prediction software.

5.2 Y chromosome STRs:

5.2.1 Haplotype Database

a) The complete list of 1370 haplotypes resulting from the analysis of 25 Y-chromosome STRs using the Yfiler™ Plus PCR Amplification Kit is reported in Supplementary Table 2 along with the SNP haplogroup previously defined and here refined for homogeneity across studies [68, 73, 74]. As for autosomal STRs, we pooled the African sample into three datasets (EA, SA and NA) based on ancestry of the individuals sampled and, to a minor extent, on their "ethnic/linguistic affiliation". This categorization better describes the genetic clustering of observed Y-chromosome patterns than categories such as "nation" or "geography."

Then, the three datasets were submitted to the YHRD (Y-chromosomal Haplotype Reference Database, https://yhrd.org) [53] under the accession numbers **YA004351-YA004356** (release R57 for Northern Africa); **YA004198-YA004207** (release R52 for Eastern Africa) and **YA004668-YA004669** (release R63 for Sub-Saharan Africa). The contributors successfully passed the quality control test.

Among the 1370 subjects analyzed, 240 were found to share 100 Y-STR haplotypes and were genotyped using the 13 RM Y-STRs multiplex PCR system described in [28].

b) The complete list of 240 haplotypes resulting from the analysis of 13 “first generation[1]” RM Y-STRs with the RM-YPlex [28] is reported in Supplementary Table 3.

   Again, 107 out of 240 males kept sharing 50 Y-STR haplotypes and were furthermore genotyped using the additional RM Y-STRs described in [29].

c) The complete list of 107 haplotypes resulting from the analysis of 30 “second generation”[2] RM Y-STRs using the RMPlex [29] is reported in Supplementary Table 4.

5.2.2 Power of discrimination analyses

5.2.2.1 RM-YPlex assay

We observed fully concordant genotyping results between Yfiler™ Plus and RM-YPlex for the 6 RM Y-STRs which are included in both multiplexes (i.e., DYS570, DYS576, DYS518, DYS627, DYF387S1 and DYS449).

Comparing males sharing the Yfiler™ haplotypes, we observed a total of 126 mutations at seven RM Y-STRs (DYS399S1, DYS403S1a/b, DYF404S1, DYS526a/b, DYS547, DYS612, and DYS626), most of which involved a single repeat (Table 1). The number of observed mutations is significantly related (r= 0.93, p=0.0008) to the mutation rates that have been recently reported in the mutation rate update by Neuhuber [27]. Overall, two markers, DYF399S1 (43 mutations) and DYF403S1a (25 mutations), accounted for more than half (54.0%) of

[1] 7 out of 13 are novel rapidly mutating markers, the remaining 6 are already comprised in Yfiler Plus multiplex.

[2] 16 out of 30 are novel rapidly mutatin markers, the remaining 14 are already comprised in RM-YPlex multiplex.

the observed mutations. These two markers have been consistently reported as those with the higher mutation rates among the first-generation RM Y-STRs in populations from different geographic areas [20, 25, 27, 75, 76]. On the opposite side of the mutational range, we observed a low number of mutations for DYS403S1b (2 mutations) and DYS626 (5 mutations). Consistently, both markers have recently been downgraded from RM Y-STRs to “fast-mutating” microsatellites (mutation rates 5 × 10-3 – 1 × 10-2). Besides this general agreement between the number of observed mutations and previously reported mutation rates, we observed an increased mutability for DYS612, which resulted to be the second most mutable locus in both northern and eastern Africa (9 and 7 mutations observed, respectively, Table 1). Since the observed mutations occurred on several different SNP-defined chromosomal backgrounds, reasons for such an apparent difference in relative mutation rates seem unrelated to haplogroup affiliation and/or increased allele length. Since DYS612 is a complex trinucleotide repeat with interruptions in the repeat motif [(CCT)5(CTT)1(TCT)4(CCT)1(TCT)n], homogenizing mutations predisposing to higher mutation rates of longer homogeneous repeat tracts could be a possible explanation that requires further investigations.

	DYF399S1	*DYF403S1a*	*DYF403S1b*	*DYF404S1*	*DYS526a*	*DYS526b*	*DYS547*	*DYS612*	*DYS626*	*Total*
Eastern Africa	19	7	2	2	1	6	3	7	2	47
Northern Africa	11	8	0	2	2	4	2	9	1	38
Sub-Saharan Africa	13	10	0	3	0	1	3	6	2	36
All Africa*	43 (9)	25 (2)	2 (1)	7	3	11	8	22 (3)	5	126 (15)

Table 1. Mutational events observed for 7 (first generation) RM Y-STRs.

** Only the seven “first generation” RM Y-STRs not included in Yfiler Plus are considered.*

The total number of multi-repeat mutations (or multiple single step mutations) is reported within brackets.

The analysis of the 13 RM Y-STRs here analyzed allowed us to improve the discrimination power substantially, with respect to the 25 Yfiler™ Plus markers (Table 3 and Figure 4). Specifically, the number of shared haplotypes decreased from 100 to 51 and the number of males sharing a haplotype decreased from 240 (17.5% of the total) to 109 (8.0%). Overall, the number of distinct haplotypes increased from 1230 to 1312. The discrimination capacity correspondingly increased from 0.898 to 0.958 (Table 3) while the proportion of males sharing a haplotype decreased from 17.5% to 8.0%. These differences in discrimination capacity among Y-STR multiplexes became also more apparent when the Yfiler™ system based on 16 conventional Y-STRs loci is considered.

5.2.2.2 RMPlex assay

Subsequently, we deepened with the investigation of the RM markers included in the RMPlex assay (Table 2) [29]. Here, we observed concordant genotyping results between most of the markers already comprised in Yfiler™ Plus and RM-YPlex assays (i.e., DYF387S1, DYF399S1, DYF404S1, DYS449, DYS518, DYS526b, DYS547, DYS570, DYS576, DYS626 and DYS627) with the exception of DYS612 and DYF403s1 markers, where we observed differences in the allele calling.

Concerning DYS612 marker, the alleles in Yfiler™ Plus and RM-YPlex assays are always called with 6 repeats more than in the RMPlex assay, because of a change in the nomenclature at this complex locus. Differently, for DYF403s1 locus, by using the previous RM-YPlex multiplex, we consistently missed the xx.1 interallele, which is clearly present in all the profiles obtained with the novel multiplex. This systematic difference is due to a new design for the primers used for this multi-copy STR.

Nevertheless, using the additional sixteen RM Y-STRs (DYF1000, DYF1001, DYF1002, DYF393S1, DYR88, DYS442, DYS711, DYS712, DYS713, DYS724. DYS1003, DYS1005, DYS1007, DYS1010, DYS1012 and DYS1013) we achieved the highest level of haplotype's discrimination (Table 3).

We observed a total of 66 independent mutations involving all the loci analyzed with the exception of DYS442 and DYS1013. This is not unexpected for the DYS442 marker, which is categorized as fast mutating (FM) instead of rapidly mutating, because of its relatively low mutation rate (equal to 7.4 x 10^{-3}). In contrast, we were surprised not to observe any mutation on DYS1013 because it was recently upgraded from fast-mutating [22] to rapidly-mutating Y-STR [27] as its mutation rate was recalculated as 10.8 x 10^{-3} instead of the previously calculated 9.9 x10^{-3}.

We observed a positive correlation between the mutation rate and the number of mutations that occurred at each marker (r = 0.729, P <0.001). The highest number of mutations (9 and 10 mutations, respectively) was observed in the two loci having the highest mutation rates, DYF1001 (48 x 10^{-3}) and DYF1000 (35.9 x 10^{-3}). The nature of these microsatellites also needs to be taken into consideration, as both consist of complex tetra-nucleotide repeats, making them more prone to mutation occurrence. On the other hand, the lowest number of mutations was observed in the two loci also having the lowest mutation rate, the DYS1013 (0 mutations) and the DYS1003 (1 mutation).

The only marker not perfectly in line with this straight correlation between mutation rate and number of observed mutations was DYS1012. However, the difficulty in interpreting the results for this marker (primers used in PCR also pair to portions of autosomal chromosomes, leading to the formation of aspecific products) may have resulted in either an underestimate of the number of mutations in the present study or an overestimate of the mutation rate in previous studies..

In total, 60 out of 66 mutations observed (91%) involved a single repeat, while 6 (10%) involved two or more repeats.

	DYF 1001	***DYS 724***	***DYF 1000***	***DYS 712***	***DYS 711***	***DYR 88***	***DYS 1007***	***DYF 1002***	***DYS 1012***	***DYS 1010***	***DYS 713***	***DYS 1003***	***DYS 1013***	***DYS 1005***	***DYS 442***	***DYS 393S1***	***Total***
Eastern Africa	***4***	***1***	***5 (1)***	***2***	***3 (1)***	***5***	***2 (1)***	***1***	***0***	***1***	***1***	***0***	***0***	***2***	***0***	***3***	***30***
Northern Africa	***4***	***6***	***2***	***3***	***2 (2)***	***1***	***1***	***2***	***0***	***2***	***2***	***1***	***0***	***0***	***0***	***1***	***27***
Sub-Saharan Africa	***1***	***0***	***3 (1)***	***0***	***0***	***1***	***1***	***2***	***1***	***0***	***0***	***0***	***0***	***0***	***0***	***0***	***9***
All Africa	***9***	***7***	***10 (2)***	***5***	***5 (3)***	***7***	***4 (1)***	***5***	***1***	***3***	***3***	***1***	***0***	***2***	***0***	***4***	***66(6)***

Table 2. Mutational events observed for 16 (second generation) RM Y-STRs.

** Only the sixteen "second generation" RM Y-STRs not included in Yfiler Plus and RMY YPlex assay are considered.*

The total number of multi-repeat mutations (or multiple single step mutations) is reported within brackets.

The discrimination capacity (DC) obtained with the "second generation" Y-STRs increased from 0.958 to 0.983, while the proportion of matching haplotypes decreased from 0.08 to 0.03. The number of distinct haplotypes turned out to be 84: among them 64 (76%) were unique and only 20 (24%) were shared between subjects (from a minimum of two to a maximum of five). Overall, we were able to distinguish 1346 distinct haplotypes and only 43 males were still found to share the same Y-STRs profile using 48 Y-STRs (Table 3 and Figure 6).

These findings further highlight the capability of RM Y-STRs to distinguish males even in sub-structured populations as African ones, but at the same time call for the discovery and testing of additional RM Y-STRs to fully differentiate Y haplotypes.

	Yfiler™		*Yfiler™ Plus*		*Yfiler™ Plus+ RM Yplex*		*Yfiler™ Plus + RMPlex*	
	(15 Y-STRs)		*(24 Y-STRs)*		*(32 Y-STRs)*		*(48 Y-STRs)*	
	DC	*MH*	*DC*	*MH*	*DC*	*MH*	*DC*	*MH*
Eastern Africa	0.773	0.338	0.887	0.190	0.958	0.078	0.987	0.026
Northern Africa	0.765	0.371	0.908	0.170	0.960	0.078	0.990	0.021
Sub-Saharan Africa	0.817	0.283	0.898	0.165	0.956	0.079	0.972	0.049
Overall	***0.784***	***0.332***	***0.898***	***0.175***	**0.958**	**0.078**	**0.983**	**0.031**

Table 3. Forensic indexes for 16 Y-STR markers (Yfiler™), 25 Y-STR markers (Yfiler™ Plus), 32 Y-STR markers and 48 Y-STRs.

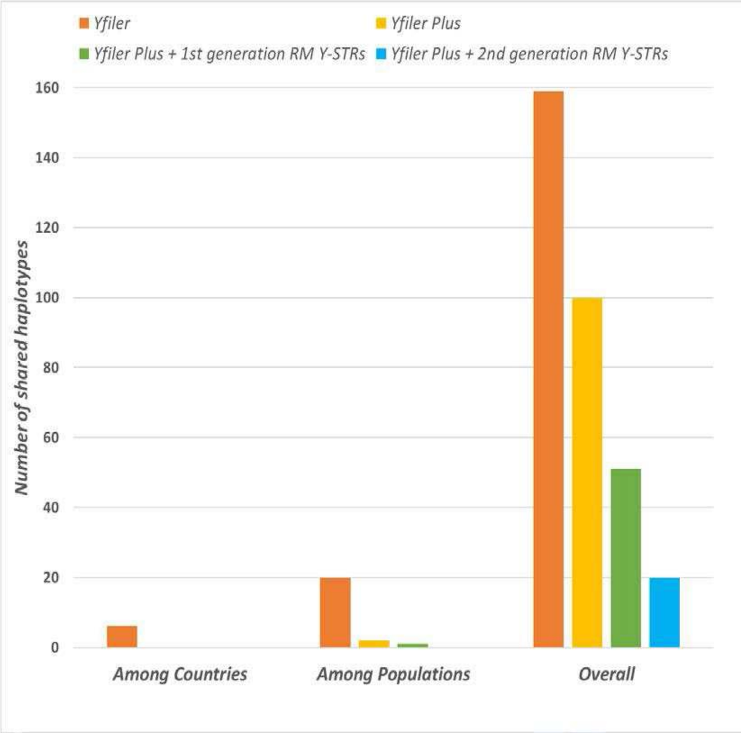


Figure 6. Discrimination power improvement using different RM Y-STR multiplexes. Number of shared haplotypes by using Yfiler™, Yfiler™ Plusand Yfiler™ Plus PCR + RMYPlex and Yfiler™ Plus PCR + RMPlex in 1370 African males (44 populations and 10 countries).

5.2.3 Estimated kinship relationships and Likelihood Ratio threshold values

Knowledge of the degree of kinship among members of a population sample is relevant in forensic analyses. For example, the construction of a DNA database should avoid relatives to guarantee Hardy-Weinberg equilibrium (HWE) among loci. Therefore, after autosomal DNA typing, described in the previous paragraphs, blind search analyses (BSAs) of the *Familias* software v.3.2.8 [46] were performed among pairs of males sharing the same Yfiler™ Plus haplotype. Note that, for each shared haplotype, the number of pairwise comparisons corresponds to *n (n-1)/2*, where *n* is the number of subjects sharing that haplotype. The total number of pairwise comparisons for each haplotype ranges from 1 (for haplotypes shared by two males) to 36(a single haplotype shared by 9 males) for a total number of 228 pairwise comparisons, resulting from 240 males sharing 100 haplotypes. For each pairwise comparison, Supplementary Table 5 shows the most likely alleged relationship (if any) and, for LR values > 1, the corresponding LR, the estimated inbreeding coefficient, the proportion of shared alleles, and proportion of loci sharing 0, 1, or 2 alleles. It should be noted that using these approaches, grandparent-grandchild (GP), avuncular (AV), and half-sibling (HS) pairs, couldn't be distinguished, since all of them are second-degree relatives sharing 25 % of their autosomal genome. This is even though members of these kinds of pairs are separated by a different amount of meiosis along the paternal lineage (GP and HS, two meiosis, AV three meiosis). Moreover, it should be noted that, using a relatively low number of autosomal STRs, a high rate of false positives (i.e., pairs of unrelated subjects inferred to be related) and false negatives (i.e., pairs of related subjects inferred to be unrelated) is expected, especially among putative cousins and, to a lesser extent, second-degree relatives [77]. The simulation performed on the LR distribution under different kinship scenarios is in line with this prediction, with false negative rates (averaged across the three regions) for LR > 1 corresponding to 0.48, 0.29 and 0.12 for second cousins, first cousins and second- degree relatives, respectively. Taking these caveats into account, and using the verbal scale for reporting the value of observed LRs proposed by [56], that is in line with the results of our simulations on the LR distribution (Supplementary Table 6), we found a total of 44 pairs whose kinship was strongly supported, with LR > 100 and positive predictive values (PPV) higher than 0.997 (15 parent-child, 22 siblings and 7 s-degree relatives), 26 moderately supported related pairs, with LR values in the range 10–100 (two pairs of siblings, PPV ≥ 0.997; 24 second-degree relatives, PPV ≥ 0.975), 63 weakly supported related pairs, with LR values in the range 1–10 (13 second-degree relatives, PPV ≥ 0.902; 36 first cousins, PPV ≥ 0.740 and 14 second cousins, PPV ≥0.561) and 94 putatively unrelated pairs with LR values < 1 (Table 4 and Supplementary Tables 5–7). A single direct match was also observed, strongly suggestive of a couple of monozygotic twins (LR = 4 × 10^{25}) or a sample duplicate.

The inferred kinship relationships were concordant in 85.1 % of the comparisons (194/228 pairwise comparisons) using region-specific or NIST allele frequencies database. Most discordant cases were characterized by LR values < 5 and affected cousins and second cousins (Supplementary Table 5). Overall, these results showed that our kinship analysis is robust enough concerning the geographic specificity of the allele frequency database used.

In addition, it is necessary to clarify that the presence of a relatively high number of closely related subjects in our global African sample is not unexpected. Most of the sampling fieldwork was performed in rural areas, where, because of patrilocality, small villages are mainly inhabited by related males belonging to the same ethnic group. In fact, all the African males that share a Y-STR haplotype were from the same country and shared the same binary haplogroup (Supplementary Table 1). With only two exceptions, haplotype-sharing males were also from the same ethnic group. These two exceptions deserve further consideration. In the first case, it regards the direct-match comparison, describe above. Thus, barring possible mistakes, such as tube duplication or exchange, these two males should belong indeed to the same ethnic group. In the second case, two haplotype-sharing males from Cameroon resulted to be a father-son pair (LR = 12155), a finding that also suggested an ethnic group misassignment during the sampling phase.

So, haplotype sharing between males in our sample set seems to be indicative (at the very least) of common ethnic affiliation, thus representing a relevant investigative lead. The relatively low number of second-degree (or closer) relatives identified reveals that close relatedness explains only a small proportion of the Y- STR haplotype sharing observed.

					Kinship Scenarios			
		Related *(LR>1)*	*Unrelated* *(LR<1)*	*Total*	*Parent-Child*	*Siblings*	*2nd degree*	*1st/2nd Cousins*
YfilerPlus + RM Yplex (32 Y-STRs)	***Differentiated***	78	79	157	3	10	25	40
	%	58.21%	84.04%	68.86%	20%	40%	56.82%	80%
	Undifferentiated	56	15	71	12	15	19	10
	%	41.79%	15.96%	31.14%	80%	60%	43.18%	20%
	Overall	134	94	228	15	25	44	50
YfilerPlus + RMPlex (48 Y-STRs)	***Differentiated***	105	93	198	9	15	32	49
	%	78,36%	98,94%	86,84%	60%	60%	72,72%	98%
	Undifferentiated	28	1	29	6	10	12	1
	%	20,89%	1,06%	12,71%	40%	40%	27,27%	2%
	Overall	134	94	228	15	25	44	50

Table 4. Number and proportion of pairs of males differentiated for different degrees of relatedness.

5.2.4 Male lineage resolution determination

By considering, the kinship inferred through BSA, we observed, as expected, a negative correlation between the degree of relatedness and ability of the “first”, primarily, and the “second”, secondly, RM Y-STR to discriminate among pairs of males (Table 4).

The analyses performed with the RM Y-Plex assay [28] were able to distinguish 78 out of 134 (58.2%) pairs of related males (LR > 1), in comparison to 79 out of 94 (84.0%) pairs of unrelated males (�2, p < 10^{-4}). Among related pairs, the proportion of discriminated pairs increased from 20.0% to 40.0% for father-son and sibling, and from 56.8% to 80.0% for second-degree relatives and cousins, respectively. The partial inability to discriminate between closely related males using the 13 RM Y-STRs markers is well exemplified by the results obtained from the three most numerous groups of males sharing a Yfiler™ Plus haplotype (haplotypes H21, H12 and H69, consisting of 6, 8 and 9 individuals, respectively) (Figure 7). Males belonging to haplotype H21 resulted to be all closely related and none of them was distinct by the 7 additional RM Y-STRs. In contrast, males sharing haplotypes H12 or H69, which were found to be unrelated or distantly related, were completely (H12) or mostly (69) distinguished by the analysis of additional Y-STRs.

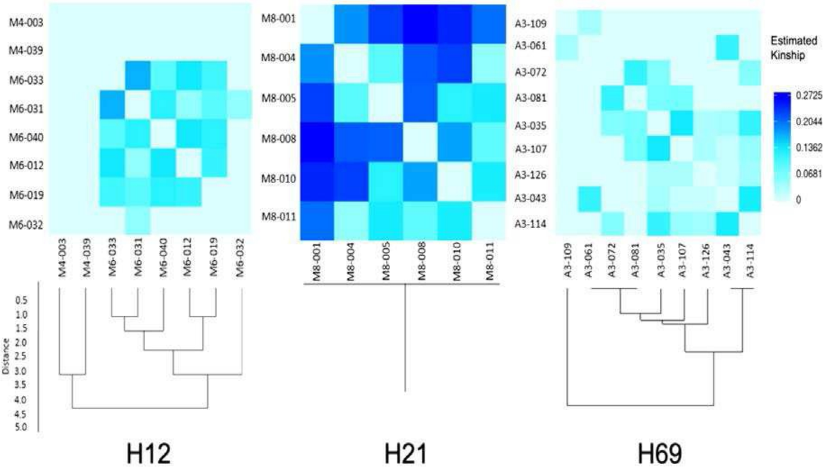


Fig. 7. Relationships among males sharing Yfiler™ Plus haplotypes H12, H21 and H69. (A) Relationships based on the number of mutations

observed at 7 additional RM Y-STRs, depicted through an UPGMA phylogenetic tree. (B) Heatmaps depicting relationships based on kinship are inferred through BSA.

Subsequently, the analyses performed with the RMPlex assay [29] were able to distinguish 27 out of the remaining 55 (48%) pairs of related males (LR > 1), in comparison to 14 out of 15 (93.0%) pairs of unrelated males (�2, p < 10-4).

In this case, the proportion of discriminated pairs increased to 60.0% for father-son, to 64 % for siblings, to 79.5% for second-degree relatives and to 88.0% for cousins.

This demonstrates the ability of these novel marker in further distinguishing male haplotypes and the major capability in solving male pedigrees using a higher number of genetic markers.

5.3 Microhaplotypes

5.3.1 Multivariate statistics

SL-PCA analysis: SL-PCA was exploited to investigate the main features in the dataset. While EAA, EA and AA individuals formed three well distinct clusters (Fig. 8), both Hispanic and “Admixed” datasets showed some degree of overlapping with other populations, likely as a consequence of ancient or recent admixture events.

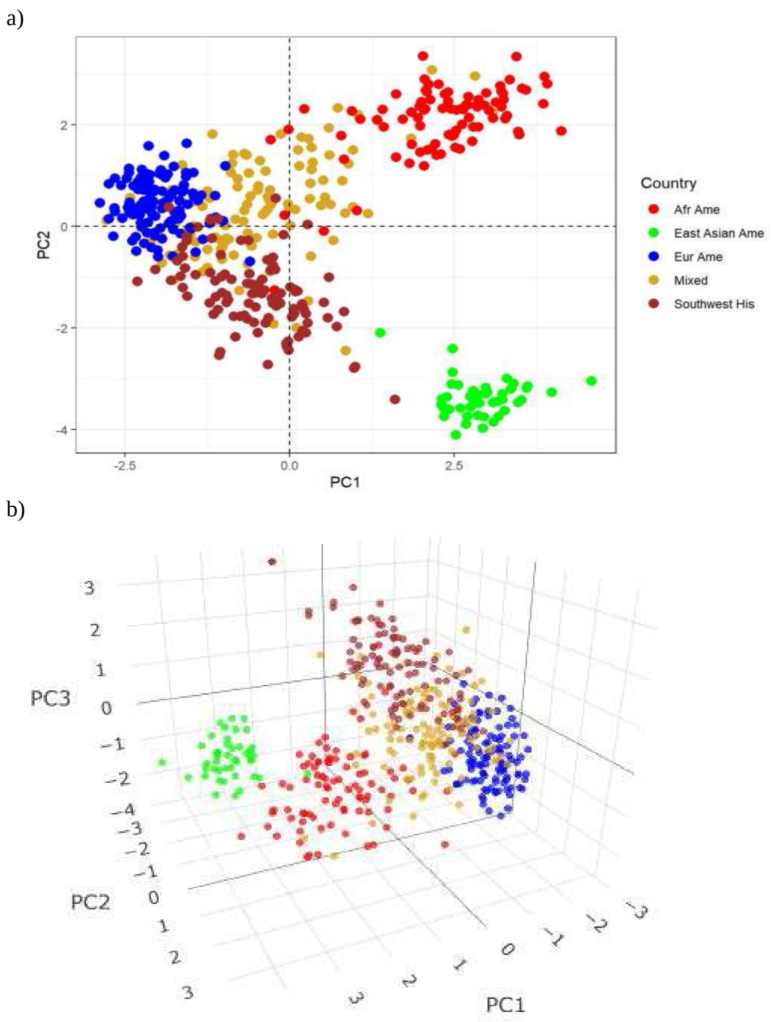


Fig. 8. SL-PCA PC1 vs. PC2 PCA 2D (a) and 3D (b) Scores Plot for the five US populations tested.

sPLS-DA model: Based on results provided by PCA modelling, sPLS-DA was applied to the same experimental sets to develop useful discrimination models. The predictive models were evaluated in terms of Root Mean Square Error in Cross-Validation (RMSECV) [69], and the number of LVs was determined through the evaluation of further quality parameters such as the Predictive Residual Error Sum of Square (PRESS), Q-residuals, Hotelling's T2, Leverages and Y-Studentized residuals [69].

The PLS-DA approach performs a supervised classification aimed to well discriminate all the five population datasets (Fig. 9). Sensitivity and specificity values were calculated too and the outcomes from ROC curves are reported in Fig. 10. These data allow us to affirm that sPLS-DA might represent a useful tool for improving the routine estimation of the BGA information of MHs with respect to SL-PCA analysis.

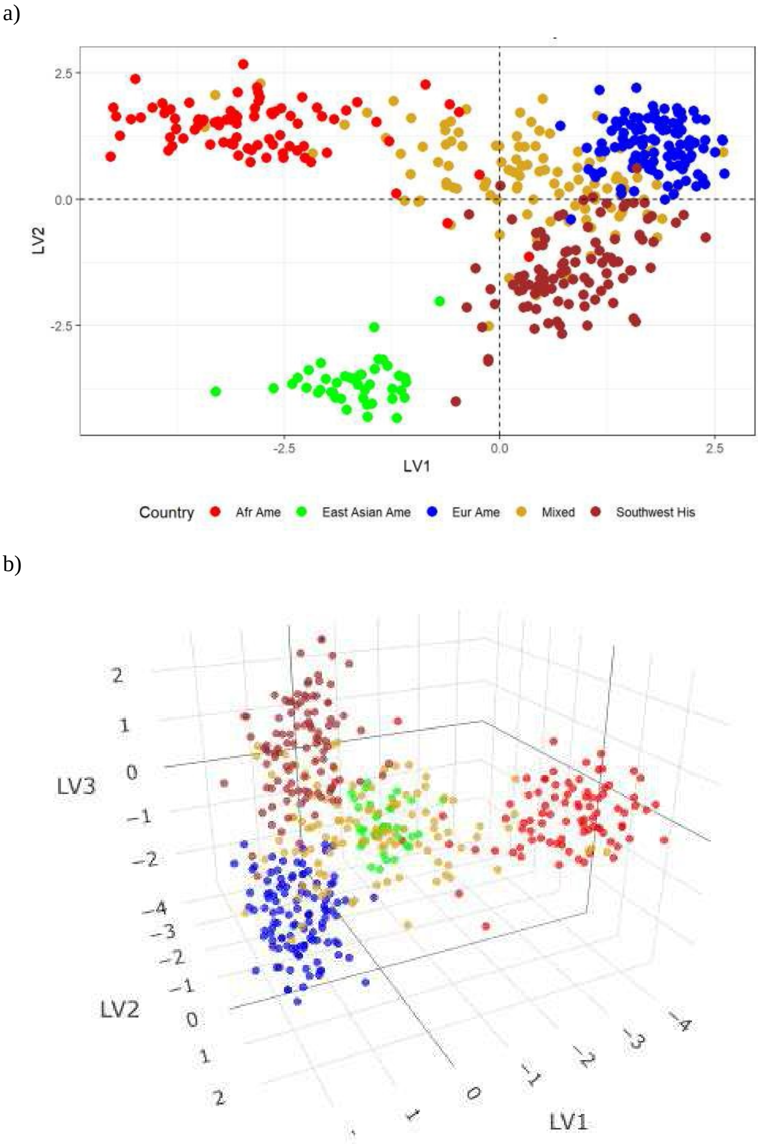


Fig. 9. sPLS-DA LV1 vs LV2 2D (a) and 3D (b) Scores Plot for the five US populations tested.

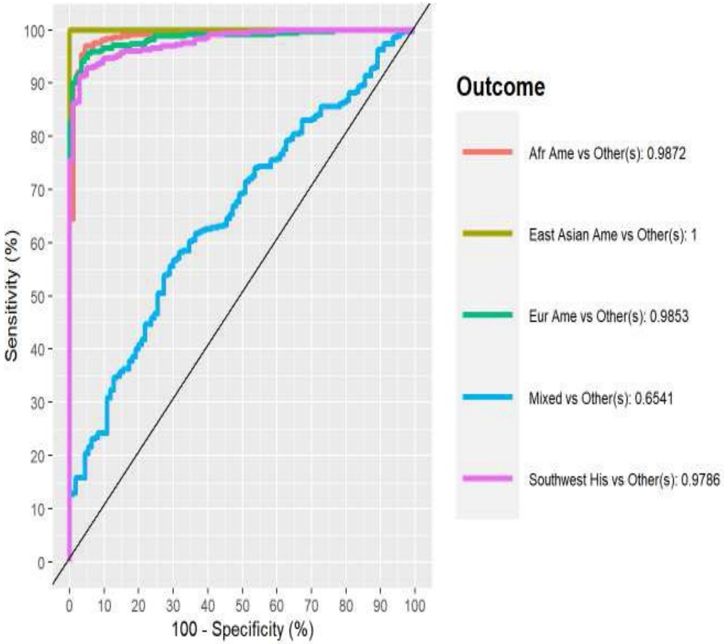


Fig. 10. Receiver Operating Characteristic (ROC) curves outcomes from PLSDA analyses for the five US populations tested.

SVM model: SVM was applied and the corresponding sensitivity and specificity values resulted to be 100% for all the five dataset (Fig. 11). This approach turned out to be a very powerful technique for the evaluation of MH data for all the 5 US populations, providing a 100 % accuracy on the tested sample sets and no misclassifications.

Consequently, SVM revealed to be a very powerful model, with high specificity and sensitivity values, for these ethnic groups, thus proving once again the reliability of multivariate statistics to extract BGA information from microhaplotype data.

In fact, compared to autosomal STRs, microhaplotypes have shown a better cluster separation, supporting their potential role as ancestry informative markers.

Overall, both PLSDA and SVM approaches significantly improved ancestry inference, by enhancing the separation of the five population clusters, providing robust classifications, yielding high sensitivity and specificity models capable of discriminating the populations investigated.

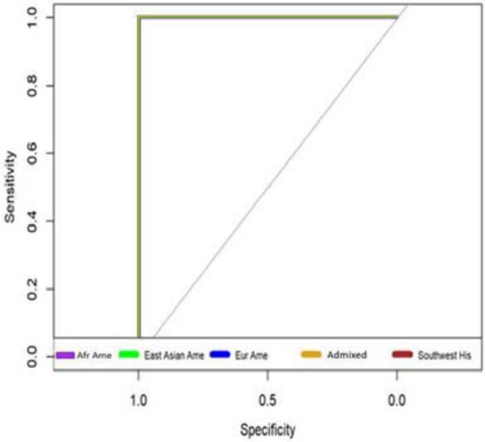


Figure 11. Sensitivity vs specificity plot for SVM results for the five US populations tested.

5.3.2 Ethnic affiliation prediction

PLSDA results were used to assess the accuracy of ethnic affiliation of four tested individuals – one for each of the 4 main population groups in US (i.e. AA, EAA, EA and HIS) in terms of Likelihood Ratio (LR). We excluded Admixed because of the peculiarity of this dataset, in fact, as described in Materials and Methods, it is composed of 129 individuals that belong to Puerto Rican, Dominican, American Indian, Vietnamese, Cuban, Mexican, Jewish or St. Lucia populations.

The resulting LR values provide an indication of how much more likely it is to observe the MH profile of interest if it originated from the test population at the numerator than if it originated from the other three populations at the denominator. In each case, the highest LR was observed in correspondence of the correct affiliation, confirming the accuracy in BGA prediction (Table 4). Specifically, LR = 10^{99}, LR = 10^{62}, LR = 10^{4}, and LR = 2,3 were obtained for Afro-American, East Asian American, European American and Southeast Hispanic, respectively. As expected, higher LR values were observed for well-genetically defined populations – as Afro-American and East Asian American – while the lowest values from Southwest Hispanic, which represent the most genetically admixed among the four populations tested.

Person of Interest	*Likelihood Ratio*	*Population*
Afro American	**9,3 x 10^{99}**	**Afro American**
	6,8 x 10^{-214}	European American
	1,2 x 10^{-35}	East Asian American
	2,3 x 10^{-67}	Southwest Hispanic
East Asian American	1,6 x 10^{-94}	Afro American
	1,5 x 10^{-127}	European American
	3,1 x 10^{62}	**East Asian American**
	4,4 x 10^{-54}	Southwest Hispanic
European American	2,6 x 10^{-9}	Afro American
	3,3 x 10^{3}	**European American**
	5,5 x 10^{-55}	East Asian American
	6,5 x 10^{-2}	Southwest Hispanic
Southwest Hispanic	1,8 x 10^{-7}	Afro American
	1,7 x 10^{-1}	European American
	9,6 x 10^{-25}	East Asian American
	2,3	**Southwest Hispanic**

Table 5. LR values for BGA affiliation of four tested individual (one for each of the main US population). In bold the highest LR and the relative population affiliation obtained.

6 Discussion

6.1 Probative value in ancestry estimation: from classical to multivariate statistical analyses on short tandem repeats (STRs) and microhaplotypes (MHs) data

BioGeographical Ancestry (BGA) has been defined as the *heritable component of "race" or heritage, which is relevant on any scale of resolution* [78, 79]. Nowadays, the inference of the BGA of a person or trace relies on three ingredients: (1) a reference database of DNA samples including ethnic information; (2) a set of loci, which segregate dependent on geographical location, i.e., a set of so-called Ancestry Informative Markers (AIMs) and (3) a statistical clustering method [80].

In the present proof-of-concept study, we proposed an alternative statistical method to classical analyses to improve the estimation of BGA, focusing on the above-mentioned points 2 e 3.

With reference to the set of loci, we tested the ability in ancestry inference of both short tandem repeats (STRs) and microhaplotypes (MHs). STRs represent the *golden standard*s for personal identification so they are routinely used by the international scientific community in forensic caseworks, already included in analytical protocols (usually in line with the international standard requirement for testing and calibration laboratories; i.e., ISO IEC 17025) and their relatively validated allele frequency databases are already available. In contrast, MHs are *novel markers* in forensic routines, composed of two- or more single-nucleotide polymorphisms (SNPs) within 300 bp length. The use of these two different kinds of *unconventional* markers for BGA estimation has some advantages. STRs profiles, for example, may already be available, as forensic experts have conducted classical DNA typing analysis with unsatisfactory results (i.e., no direct match, neither indirect match searching in national and international databases). The advantage of microhaplotypes is found in caseworks where the primary source sample is particularly complex (low template DNA, degraded DNA or mixtures), such as for unidentified human remains (UHRs) found in advanced state of decomposition or putrefaction or samples collected from outdoor environment in extreme condition (for example, extremely high temperatures, humidity, unfavorable atmospheric conditions, proliferation of mold, bacteria or fungi, etc.). In such cases, it is necessary to choose shorter (in bp length) markers – such as SNPs or, indeed, MHs – to maximize the yield of genetic DNA typing and to further enhance the deconvolution capabilities of mixed-DNA source samples, providing additional forensically useful information on the contributor(s) detected.

With reference to the clustering method, we proposed novel approaches based on multivariate techniques to group samples into BGA-classes.

In fact, as explained in Alladio et al. [72], although PCA analysis allows to assign an individual to his/her population of origin through a visual, intuitive, and easy to interpret approach, it does not provide significant divergence between populations, and obviously, it cannot be used alone in forensic context because it does not provide an accurate statistical estimate of the weight of the evidence. PLS-DA was then applied to develop more reliable discrimination models to classify the variables and, as a result, it turned noteworthy.

As a matter of fact, the results presented in this thesis allowed to observe the pertinence of new multivariate statistics in assessing and recognizing the biogeographical ancestry information by evaluating both the autosomal STRs and microhaplotype DNA profiles.

As expected, the only cases in which the new multivariate statistics returned unsatisfactory results were those in which the populations to be compared had fairly similar allele frequencies as consequence of recent admixture or common ancestry.

It is well known that aSTR, because of their very high mutation rate, are not the markers of choice for ancestry inference. Notwithstanding , good results were obtained using a limited number of STRs in the African scenario when the new multivariate statistics were used: the AUC values between 0.5 and 0.9 for Eastern Africa and greater than 0.9 for Northern and Sub-Saharan populations, suggest an excellent capacity of discrimination and outstanding discrimination, respectively.

With respect to MH-profiles results, unsatisfactory results were observed for Admixed population only, while the other four US population tested (i.e., AA, EA, EAA and HIS) appeared to be well discriminated.

Unlike the results obtained in [22] using conventional PCA statistics, the application of the more sophisticated PLSDA was determinant in reaching a satisfying level of ancestry inference. In fact, PLSDA assay we tested revealed a good separation of Southwest Hispanic with respect to the other three main US groups (i.e., AA, EA and EAA).

The only substantial overlap we observed is restricted to Admixed, who share the highest level of ancestry with Afro Americans, European American and Southwest Hispanic. This was confirmed by the AUC values, which range between 0.5 and 0.6 for Admixed, between 0.6 and

0.9 for Southwest Hispanic and greater than 0.9 for the remaining populations.

We then sought to assay the more sophisticated technique of Support Vector Machine (SVM) to achieve more satisfactory separations in both STRs and MHs scenarios.

Similarly, for the two types of genetic data, the numerical results for the performance of the model (in terms of ROC curves) showed the best separation among all the populations. Specifically, the ROC curves outcomes for STRs panel turned out to be higher than 0.8 for Eastern while greater than 0.9 for Northern and Sub-Saharan Africa; while for MHs panel equal to 0.7 for Admixed and greater than 0.9 for the remaining US populations. These results show that the SVM is the best classification assay as it allows obtaining an even more excellent separation among the population tested and assessing the group affiliation of the examined DNA profiles, with a high degree of confidence.

All together these data demonstrate the ability of multivariate statistics approaches to predict the population affiliation from both autosomal STR and microhaplotype genetic profiles. The predictive power of such multivariate techniques turned extremely high – in fact, they correctly classify individuals from different ethnic groups by enhancing cluster separation and providing no misleading classifications – indicating that the adoption of multivariate models may represent a powerful and useful tool for the investigative authorities to ease their decision processes when estimating the BGA of individuals. Obviously, classification efficiency is higher for more genetically differentiated populations, whereas in the case of a profile of an individual of admixed ancestry, the risk of the profile being rejected in constituent populations increases [82].

Future perspectives include the application of these multivariate strategies to discriminate even more locally restricted populations, and further research studies are already planned and will be performed using Next-Generation Sequencing (NGS)/Massive Parallel Sequencing (MPS). The idea is to merge our data and later combine it with other forensic genetic markers, such as Y-STRs and SNPs, to achieve even finer resolution in ethnic prediction.

As a matter of fact, in a judicial context the probative value of ancestry inference is extremely high and needs further investigation. First of all, the range of application of ancestry prediction analysis is extremely broad, as it is possible to infer a subject's ethnicity from any biological sample found at a crime scene, during mass disasters or missing person investigations. In addition, achieving true DNA-based racial profiling provides additional, often essential, information that can narrow the field of suspects, enabling concrete support for classical investigations. This support is greater the finer the resolution of the statistical analysis, confirming the need to adopt multivariate techniques in the forensic routine of the near future.

6.2 Y markers: from paternal lineage inference to personal identification.

Y-chromosome STR analysis has become very popular in forensic practices for male lineage characterization, unbalanced male-female mixture deconvolution, estimation of the number of contributors in mixed samples and exclusion of male suspects [15, 17, 83]. The relatively low discrimination power of conventional Y-STR multiplexes, due to linkage disequilibrium (*LD*) among polymorphic loci, has been partially overcome by the introduction of rapidly mutating Y microsatellites (*RM Y-STRs*) with mutation rates exceeding 1 × 10^{-2}/generation. In previous studies [68, 73, 74], we reported an unexpectedly high level of haplotype sharing among African males by using the Yfiler Plus PCR Amplification kit that is the most powerful commercially available system including 19 conventional Y-STRs and 6 RM Y-STRs.

In the present study we analyzed for autosomal and Y-chromosome STRs 1370 males from northern, eastern and central Africa. Actually, the peculiarity of the populations tested – characterized by high levels of endogamy and sub-structuring – makes them particularly suitable for these studies.

Firstly, we found out 240 subjects sharing 100 Y-STR haplotypes and secondly, throughout Blind Search Analyses (BSA) and Simulation test tools of Familias Software [46], we highlighted the hidden familial relationships and demonstrated that the discrimination failure obtained in previous studies [68, 73, 74] was only partially due to close relatedness among males. Specifically, we found a total of 44 pairs whose kinship was strongly supported (with LR > 100 and PPV > 0.997), 26 moderately supported related pairs (with LR values in the range 10–100 and PPV ≥ 0.997), 63 weakly supported related pairs (with LR values in the range 1–10 and PPV ≥ 0.902) and 94 putatively unrelated pairs (LR values < 1) [84]. The relatively low number of second-degree (or closer) relatives identified reveals that close relatedness explains only a small proportion of the Y- STR haplotype sharing observed in our sample set.

On the contrary, the presence of a relatively high number of closely related subjects is not unexpected. Most of the sampling fieldwork was performed in rural areas, where, because of patrilocality, small villages are mainly inhabited by related males belonging to the same ethnic group. Differences in discrimination power reported for African populations using the Yfiler Plus multiplex could be thus explained, at least partially, by different sampling strategies. In any case, our analysis suggested that the high level of haplotype sharing could not be entirely explained by kinship, since about half of the pairwise comparisons involved unrelated (or distantly related) males [84].

Starting from these results, we deepened the analyses by genotyping the additional seven "*first generation*" RM Y-STR described in [28]. Although we substantially improved the discrimination capacity in these populations we still failed in distinguishing among most related individuals and some putatively unrelated males. The resulting haplotype sharing is restricted to males belonging to the same ethnic group; thus, it seems to indicate (at the very least) of common ethnic affiliation, representing a relevant investigative lead in forensic context. Moreover, to overcome these issues we performed the analysis of the "*second generation*" RM Y-STRs [22, 29]. These additional markers were found to be necessary to advance further toward the full differentiation of males, allowing us to achieve the highest level of discriminatory capacity and the fewest number of matching haplotypes, even in close kinship scenarios and in sub-structured populations such as those in Africa.

We empirically demonstrated the improved differentiation of males sharing Y-haplotypes – close and distant relatives and unrelated too – achieved with 48 Y-STRs genotyped, compared to the current state-of-the-art commercially available tools. In addition, via molecular study of the mutations occurred for each RM locus, we provided further evidence on which loci should be the most relevant to be included in the validated and commercially available forensic multiplexes.

Until additional data from more populations become available, caution shall be placed when identifying more mutating markers– and consequently when applying mutation rate estimates – established in one population, as in ours, to forensic cases involving males suspected of paternal lineage from other populations, especially non-African ones [85].

Overall, our data, converge in demonstrating that RM Y-STRs represent a very powerful forensic tool not only for paternal lineage definition, but also for personal identification purposes in forensic genetics. We also supported the relevance of including additional RM Y-STRs in fully validated and commercially available multiplexes.

7 Conclusions

7.1 The Importance of alternative forensic tools in the investigative stream to narrow down suspects

The fundamental goal of forensic genetics is personal identification or, in other words, sample attribution to associate an item of evidence with some person or persons. The most common scenarios involve a direct comparison, performed between DNA profiles obtained from an evidentiary item and a reference sample collected from the Person of Interest (i.e. PoI), or an indirect comparison, resulting in national or international DNA database searching.

Unfortunately, it is common to obtain inconclusive results from both direct comparisons (STR-profiling fails to produce a DNA-match with known suspects) and indirect comparisons (STR-profiling fails to produce a DNA-match in forensic DNA databases). In such cases, it is desirable to maximize the information inferable from the biological materials found at the crime scene to generate crucial leads to identify unknown perpetrators or to identify unknown human remains.

In such cases, the DNA-based inference of appearance traits, the biogeographical ancestry (BGA), and chronological age allows to narrow the list of putative suspects.

Among these, BGA inference turns out to be the most aleatory variable, characterized by the highest range of uncertainty. The basic idea in ancestry prediction is that any two individuals, including those apparently unrelated, can share short segments of DNA inherited by a distant common ancestor, and these matching segments of DNA shared by two or more people are called identical by descent (IBD).

Obviously, the percentage of genetic sharing is the higher the closest the kinship scenario is.

The first goal of the present proof-of-concept thesis is to present an innovative statistical method for BGA inference. We demonstrated the capability of novel multivariate statistics approaches to predict in detail the population affiliation of PoIs using both autosomal STR profiles and microhaplotypes. Therefore, it can be considered a powerful tool for generating unconventional investigative leads and can be easily implemented in operational settings [91].

Another alternative forensic tool to narrow down potential suspects in the forensic field is represented by the use of Rapidly Mutating Y chromosome STRs (RM Y-STRs).

Since Y-chromosome DNA analysis is important in genetic genealogy and for population genetic purposes such as personal ancestry identification, as well as for the identification of male lineages and inferring paternal genetic ancestry [92 – 98], many papers investigated the individualization potentiality of highly mutating markers located on this chromosome. In general, similarities at Y-chromosome DNA markers indicate shared paternal ancestry of individuals and populations, whereas differences are used to conclude the absence of close paternal relationships [21].

Instead, the genetic typing of RM Y-STR set provides near-complete paternal lineage differentiation in general populations as well as in sub-structured populations with reduced Y-chromosome diversity, due to peculiarities in population history or cultural practices [84]. This results in the reduction of “*adventitious correspondences*” or, in other words, in the inclusion of innocent individuals in investigations due to adventitious Y-STR haplotype matches [21].

The second objective of the present proof-of-concept study was to provide additional support on the identification power of both “*first*” and “*second*” generation RM Y-STRs. Thanks to the results obtained with Blind Search Analyses conducted, we proved that the failure in distinguishing among males sharing the same RM Y-haplotype is limited only to cases of extremely close kinship.

The analyses of these loci have increased the overall haplotype diversity and discrimination power, resulting in lower match probabilities. The increased amount of profile information generated from the additional loci is beneficial for exclusionary purposes too.

Thus, all together our findings support the relevance of including RM Y-STRs in available multiplexes to maximize the possibility of solving patrilineal lineages and to specifically identify the subject of forensic investigative interest.

8 References

1. Gosch, A. and Courts, C. (2019) "On DNA transfer: The lack and difficulty of systematic research and how to do it better," Forensic Science International: Genetics, 40, pp. 24–36. Available at: https://doi.org/10.1016/j.fsigen.2019.01.012.
2. Alaeddini, R., Walsh, S.J. and Abbas, A. (2010) "Forensic implications of genetic analyses from degraded DNA—a review," Forensic Science International: Genetics, 4(3), pp. 148–157. Available at: https://doi.org/10.1016/j.fsigen.2009.09.007.
3. Goodwin, W. (2016) Forensic DNA typing protocols. New York, USA: Humana Press.
4. Kayser, M. (2015) "Forensic DNA phenotyping: Predicting human appearance from crime scene material for investigative purposes," Forensic Science International: Genetics, 18, pp. 33–48. Available at: https://doi.org/10.1016/j.fsigen.2015.02.003.
5. Jordan, D. and Mills, D.E. (2021) "Past, present, and future of DNA typing for analyzing human and non-human forensic samples," Frontiers in Ecology and Evolution, 9. Available at: https://doi.org/10.3389/fevo.2021.646130.
6. Oosthuizen, T. and Howes, L.M. (2022) "The development of forensic DNA analysis: New debates on the issue of fundamental human rights," Forensic Science International: Genetics, 56, p. 102606. Available at: https://doi.org/10.1016/j.fsigen.2021.102606.
7. Chen, M. et al. (2020) "Comparison of ce- and mps-based analyses of forensic markers in a single cell after whole genome amplification," Forensic Science International: Genetics, 45, p. 102211. Available at: https://doi.org/10.1016/j.fsigen.2019.102211.
8. Xu, Q. et al. (2022) "Evaluating the effects of whole genome amplification strategies for amplifying trace DNA using capillary electrophoresis and massive parallel sequencing," Forensic Science International: Genetics, 56, p. 102599. Available at: https://doi.org/10.1016/j.fsigen.2021.102599.

9. Katsara, M.-A. and Nothnagel, M. (2019) “True colors: A literature review on the spatial distribution of eye and hair pigmentation,” Forensic Science International: Genetics, 39, pp. 109–118. Available at: https://doi.org/10.1016/j.fsigen.2019.01.001.
10. Katsara, M.-A. et al. (2021) “Evaluation of supervised machine-learning methods for predicting appearance traits from DNA,” Forensic Science International: Genetics, 53, p. 102507. Available at: https://doi.org/10.1016/j.fsigen.2021.102507.
11. Tvedebrink, T. and Eriksen, P.S. (2019) “Inference of admixed ancestry with ancestry informative markers,” Forensic Science International: Genetics, 42, pp. 147–153. Available at: https://doi.org/10.1016/j.fsigen.2019.06.013.
12. Xavier, C. et al. (2022) “Development and inter-laboratory evaluation of the Visage enhanced tool for appearance and ancestry inference from DNA,” Forensic Science International: Genetics, 61, p. 102779. Available at: https://doi.org/10.1016/j.fsigen.2022.102779.
13. Phillips, C. (2015) “Forensic genetic analysis of bio-geographical ancestry,” Forensic Science International: Genetics, 18, pp. 49–65. Available at: https://doi.org/10.1016/j.fsigen.2015.05.012.
14. Pinto, N. et al. (2019) “Optimizing the information increase through the addition of relatives and genetic markers in identification and kinship cases,” Forensic Science International: Genetics, 40, pp. 210–218. Available at: https://doi.org/10.1016/j.fsigen.2019.02.019.
15. Gusmão, L. et al. (2006) “DNA commission of the International Society of Forensic Genetics (ISFG): An update of the recommendations on the use of Y-strs in forensic analysis,” Forensic Science International, 157(2-3), pp. 187–197. Available at: https://doi.org/10.1016/j.forsciint.2005.04.002.
16. Kayser, M. (2017) “Forensic use of Y-chromosome DNA: A general overview,” Human Genetics, 136(5), pp. 621–635. Available at: https://doi.org/10.1007/s00439-017-1776-9
17. Roewer, L. et al. (2020) “DNA commission of the International Society of Forensic Genetics (ISFG): Recommendations on the interpretation of Y-STR

results in forensic analysis," Forensic Science International: Genetics, 48, p. 102308. Available at: https://doi.org/10.1016/j.fsigen.2020.102308

18. Ballantyne, K.N. et al. (2010) "Mutability of Y-chromosomal microsatellites: Rates, characteristics, molecular bases, and forensic implications," The American Journal of Human Genetics, 87(3), pp. 341–353. Available at: https://doi.org/10.1016/j.ajhg.2010.08.006.
19. Diegoli, T.M. (2015) "Forensic typing of short tandem repeat markers on the X and Y chromosomes," Forensic Science International: Genetics, 18, pp. 140–151. Available at: https://doi.org/10.1016/j.fsigen.2015.03.013.
20. Ballantyne, K.N. et al. (2012) "A new future of forensic Y-Chromosome Analysis: Rapidly mutating Y-strs for differentiating male relatives and paternal lineages," Forensic Science International: Genetics, 6(2), pp. 208–218. Available at: https://doi.org/10.1016/j.fsigen.2011.04.017.
21. Ballantyne, K.N. et al. (2014) "Toward male individualization with rapidly mutating Y-chromosomal short tandem repeats," Human Mutation, 35(8), pp. 1021–1032. Available at: https://doi.org/10.1002/humu.22599.
22. Ralf, A. et al. (2020) "Identification and characterization of novel rapidly mutating y-chromosomal short tandem repeat markers," Human Mutation, 41(9), pp. 1680–1696. Available at: https://doi.org/10.1002/humu.24068.
23. Claerhout, S. et al. (2018) "Determining y-STR mutation rates in deep-routing genealogies: Identification of Haplogroup differences," Forensic Science International: Genetics, 34, pp. 1–10. Available at: https://doi.org/10.1016/j.fsigen.2018.01.005
24. Robino, C. et al. (2015) "Development of an Italian RM Y-STR haplotype database: Results of the 2013 GEFI Collaborative Exercise," Forensic Science International: Genetics, 15, pp. 56–63. Available at: https://doi.org/10.1016/j.fsigen.2014.10.008.
25. Adnan, A. et al. (2016) "Improving empirical evidence on differentiating closely related men with RM Y-strs: A comprehensive pedigree study from Pakistan," Forensic Science International: Genetics, 25, pp. 45–51. Available at: https://doi.org/10.1016/j.fsigen.2016.07.005.
26. Rakha et al., Rakha, A. et al. (2018) "Discriminating power of rapidly mutating Y-strs in deep rooted endogamous pedigrees from Sindhi population

of Pakistan," Legal Medicine, 34, pp. 17–20. Available at: https://doi.org/10.1016/j.legalmed.2018.08.001.

27. Neuhuber, F. et al. (2022) "Improving the differentiation of closely related males by RMPLEX analysis of 30 Y-strs with high mutation rates," Forensic Science International: Genetics, 58, p. 102682. Available at: https://doi.org/10.1016/j.fsigen.2022.102682.
28. Alghafri, R. et al. (2015) "A novel multiplex assay for simultaneously analysing 13 rapidly mutating Y-strs," Forensic Science International: Genetics, 17, pp. 91–98. Available at: https://doi.org/10.1016/j.fsigen.2015.04.004.
29. Ralf, A. et al. (2021) "RMPLEX: An efficient method for analyzing 30 Y-strs with high mutation rates," Forensic Science International: Genetics, 55, p. 102595. Available at: https://doi.org/10.1016/j.fsigen.2021.102595.
30. Brenner, C.H. (2006) "Some mathematical problems in the DNA identification of victims in the 2004 tsunami and similar mass fatalities," Forensic Science International, 157(2-3), pp. 172–180. Available at: https://doi.org/10.1016/j.forsciint.2005.11.003.
31. Brenner, C.H. and Weir, B.S. (2003) "Issues and strategies in the DNA identification of World Trade Center victims," Theoretical Population Biology, 63(3), pp. 173–178. Available at: https://doi.org/10.1016/s0040-5809(03)00008-x.
32. Porras-Hurtado, L. et al. (2013) "An overview of structure: Applications, parameter settings, and supporting software," Frontiers in Genetics, 4. Available at: https://doi.org/10.3389/fgene.2013.00098.
33. Santos, C. et al. (2016) "Inference of ancestry in Forensic Analysis II: Analysis of Genetic Data," Methods in Molecular Biology, pp. 255–285. Available at: https://doi.org/10.1007/978-1-4939-3597-0_19.
34. Pereira, L. et al. (2010) "PopAffiliator: Online Calculator for individual affiliation to a major population group based on 17 autosomal short tandem repeat genotype profile," International Journal of Legal Medicine, 125(5), pp. 629–636. Available at: https://doi.org/10.1007/s00414-010-0472-2.

35. Santos, C. et al. (2016) “Inference of ancestry in Forensic Analysis II: Analysis of Genetic Data,” Methods in Molecular Biology, pp. 255–285. Available at: https://doi.org/10.1007/978-1-4939-3597-0_19.

36. Porras-Hurtado, L. et al. (2013) “An overview of structure: Applications, parameter settings, and supporting software,” Frontiers in Genetics, 4. Available at: https://doi.org/10.3389/fgene.2013.00098

37. Lee, S., et al. (2010) “Sparse logistic principal components analysis for Binary Data,” The Annals of Applied Statistics, 4(3). Available at: https://doi.org/10.1214/10-aoas327.

38. Barker, M. and Rayens, W. (2003) “Partial least squares for discrimination,” Journal of Chemometrics, 17(3), pp. 166–173. Available at: https://doi.org/10.1002/cem.785.

39. Ballabio, D. and Consonni, V. (2013) “Classification tools in chemistry. part 1: Linear Models. PLS-da,” Analytical Methods, 5(16), p. 3790. Available at: https://doi.org/10.1039/c3ay40582f.

40. Lê Cao, K.-A. et al. (2008) “A sparse PLS for variable selection when integrating omics data,” Statistical Applications in Genetics and Molecular Biology, 7(1). Available at: https://doi.org/10.2202/1544-6115.1390

41. Hearst, M.A. et al. (1998) “Support Vector Machines,” IEEE Intelligent Systems and their Applications, 13(4), pp. 18–28. Available at: https://doi.org/10.1109/5254.708428

42. Vapnik, V.N. (2010) The nature of statistical learning theory. New York: Springer.

43. Oldoni, F., Kidd, K.K. and Podini, D. (2019) “Microhaplotypes in forensic genetics,” Forensic Science International: Genetics, 38, pp. 54–69. Available at: https://doi.org/10.1016/j.fsigen.2018.09.009.

44. Oldoni, F. et al. (2020) “Population genetic data of 74 microhaplotypes in four major U.S. population groups,” Forensic Science International: Genetics, 49, p. 102398. Available at: https://doi.org/10.1016/j.fsigen.2020.102398.

45. Oldoni, F. et al. (2017) “Microhaplotypes for ancestry prediction,” Forensic Science International: Genetics Supplement Series, 6. Available at: https://doi.org/10.1016/j.fsigss.2017.09.209.

46. Kling, D. et al. (2014) “Familias 3 – extensions and new functionality,” Forensic Science International: Genetics, 13, pp. 121–127. Available at: https://doi.org/10.1016/j.fsigen.2014.07.004.

47. Bodner, M. et al. (2016) “Recommendations of the DNA commission of the International Society for Forensic Genetics (ISFG) on quality control of autosomal short tandem repeat allele frequency databasing (strider),” Forensic Science International: Genetics, 24, pp. 97–102. Available at: https://doi.org/10.1016/j.fsigen.2016.06.008.

48. Gouy, A. et al. (2017). STRAF - A convenient online tool for STR data evaluation in forensic genetics. Forensic Science International: Genetics, 30, 148-151.

49. Gill, P. et al. (2012) “DNA commission of the International Society of Forensic Genetics: Recommendations on the evaluation of STR typing results that may include drop-out and/or drop-in using probabilistic methods,” Forensic Science International: Genetics, 6(6), pp. 679–688. Available at: https://doi.org/10.1016/j.fsigen.2012.06.002.

50. Gill, P. et al. (2006) “DNA commission of the International Society of Forensic Genetics: Recommendations on the interpretation of mixtures,” Forensic Science International, 160(2-3), pp. 90–101. Available at: https://doi.org/10.1016/j.forsciint.2006.04.009.

51. Gill, P. et al. (2020) “DNA commission of the International Society for Forensic Genetics: Assessing the value of forensic biological evidence - guidelines highlighting the importance of propositions. part II: Evaluation of biological traces considering activity level propositions,” Forensic Science International: Genetics, 44, p. 102186. Available at: https://doi.org/10.1016/j.fsigen.2019.102186.

52. Gill, P. et al. (2018) “DNA commission of the International Society for Forensic Genetics: Assessing the value of forensic biological evidence - guidelines highlighting the importance of propositions,” Forensic Science International: Genetics, 36, pp. 189–202. Available at: https://doi.org/10.1016/j.fsigen.2018.07.003.

53. Willuweit, S. et al. (2007) “Y chromosome haplotype reference database (YHRD): Update,” Forensic Science International: Genetics, 1(2), pp. 83–87. Available at: https://doi.org/10.1016/j.fsigen.2007.01.017.

54. Ion Library TaqMan™ Quantitation Kit, User Guide. Thermo Fisher Scientific https://assets.thermofisher.com/TFS-Assets/LSG/manuals/MAN0015802_IonLibrary_Taqman_Quantitation_Kit_UG.pdf

55. Hill, C.R. et al. (2013) “U.S. population data for 29 autosomal STR loci,” Forensic Science International: Genetics, 7(3). Available at: https://doi.org/10.1016/j.fsigen.2012.12.004.

56. Marquis, R. et al. (2016) “Discussion on how to implement a verbal scale in a forensic laboratory: Benefits, pitfalls and suggestions to avoid misunderstandings,” Science &amp; Justice, 56(5), pp. 364–370. Available at: https://doi.org/10.1016/j.scijus.2016.05.009.

57. Della Rocca, C. et al. (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(1), pp. 671–673. Available at: https://doi.org/10.1016/j.fsigss.2019.10.133.

58. Cooper, R.A. (2007) “paleontological data analysis. by øyvind&nbsp; Hammer and David&nbsp; Harper. Oxford: Blackwell, 2006. 351 pages. $89.95 paper.,” The Journal of Geology, 115(5), pp. 609–609. Available at: https://doi.org/10.1086/519781.

59. R Core Team, R: a Language and Environment for Statistical Computing, (2014)

60. An introduction to R (2015). Samurai Media.

61. Rohart, F. et al. (2017) “MixOmics: An R package for ‘OMICS feature selection and Multiple Data Integration,” PLOS Computational Biology, 13(11). Available at: https://doi.org/10.1371/journal.pcbi.1005752.

62. Meyer, D. et al. (2020) “Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071)”

63. Bro, R. et al. (2014) “Principal component analysis,” Anal. Methods, 6(9), pp. 2812–2831. Available at: https://doi.org/10.1039/c3ay41907j.

64. Lê Cao. et al. (2011) “Sparse PLS discriminant analysis: Biologically relevant feature selection and graphical displays for multiclass problems,” BMC Bioinformatics, 12(1). Available at: https://doi.org/10.1186/1471-2105-12-253.

65. Wold, S. et al. (2001) “PLS-regression: A basic tool of Chemometrics,” Chemometrics and Intelligent Laboratory Systems, 58(2), pp. 109–130. Available at: https://doi.org/10.1016/s0169-7439(01)00155-1.

66. Filzmoser, P. et al. (2009) “Repeated double cross validation,” Journal of Chemometrics, 23(4), pp. 160–171. Available at: https://doi.org/10.1002/cem.1225.

67. Alladio, E. et al. (2019) “A multivariate statistical approach to for the evaluation of the biogeographical ancestry information from traditional strs,” Forensic Science International: Genetics Supplement Series, 7(1), pp. 253–255. Available at: https://doi.org/10.1016/j.fsigss.2019.09.097.

68. Della Rocca, C. et al. (2020) “Ethnic fragmentation and degree of urbanization strongly affect the discrimination power of Y-STR haplotypes in central Sahel,” Forensic Science International: Genetics, 49, p. 102374. Available at: https://doi.org/10.1016/j.fsigen.2020.102374.

69. Forina, M. et al. (2004) “Selection of useful predictors in multivariate calibration,” Analytical and Bioanalytical Chemistry, 380(3), pp. 397–418. Available at: https://doi.org/10.1007/s00216-004-2768-x.

70. Alladio, E. et al., “BGApredictor – An alternative online tool for the prediction of the biogeographical ancestry information”. Submitted.

71. Alladio, E. et al. (2020) “A multivariate statistical approach for the estimation of the ethnic origin of unknown genetic profiles in Forensic Genetics,” Forensic Science International: Genetics, 45, p. 102209. Available at: https://doi.org/10.1016/j.fsigen.2019.102209.

72. Alladio, E. et al. (2022) “Multivariate Statistical Approach and machine learning for the evaluation of biogeographical ancestry inference in the forensic field,” Scientific Reports, 12(1). Available at: https://doi.org/10.1038/s41598-022-12903-0.

73. Iacovacci, G. et al. (2017) “Forensic data and microvariant sequence characterization of 27 Y-STR loci analyzed in four eastern African

countries," Forensic Science International: Genetics, 27, pp. 123–131. Available at: https://doi.org/10.1016/j.fsigen.2016.12.015.

74. D'Atanasio, E. et al. (2019) "Rapidly mutating Y-strs in rapidly expanding populations: Discrimination power of the YFILER plus multiplex in Northern Africa," Forensic Science International: Genetics, 38, pp. 185–194. Available at: https://doi.org/10.1016/j.fsigen.2018.11.002.
75. Yuan, L. et al. (2018) "Mutation analysis of 13 RM Y-str loci in Han population from Beijing of China," International Journal of Legal Medicine, 133(1), pp. 59–63. Available at: https://doi.org/10.1007/s00414-018-1949-7.
76. Boattini, A. et al. (2019) "Estimating y-STR mutation rates and TMRCA through deep-rooting Italian pedigrees," Scientific Reports, 9(1). Available at: https://doi.org/10.1038/s41598-019-45398-3.
77. Tamura, T. et al. 2015) "Evaluation of Advanced Multiplex short tandem repeat systems in pairwise kinship analysis," Legal Medicine, 17(5), pp. 320–325. Available at: https://doi.org/10.1016/j.legalmed.2015.03.005.
78. STURM, R. (2004) "Eye colour: Portals into pigmentation genes and ancestry," Trends in Genetics, 20(8), pp. 327–332. Available at: https://doi.org/10.1016/j.tig.2004.06.010.
79. Gannett, L. (2014) "Biogeographical ancestry and Race," Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences, 47, pp. 173–184. Available at: https://doi.org/10.1016/j.shpsc.2014.05.017.
80. Pfaffelhuber, P. et al. (2020) "How to choose sets of ancestry informative markers: A supervised feature selection approach," Forensic Science International: Genetics, 46, p. 102259. Available at: https://doi.org/10.1016/j.fsigen.2020.102259.
81. Buckleton, J. et al. (2016) "Population-specific F values for forensic STR markers: A worldwide survey," Forensic Science International: Genetics, 23, pp. 91–100. Available at: https://doi.org/10.1016/j.fsigen.2016.03.004.
82. Tvedebrink, T. et al. (2019) "Inference of admixed ancestry with ancestry informative markers," Forensic Science International: Genetics, 42, pp. 147–153. Available at: https://doi.org/10.1016/j.fsigen.2019.06.013.

83. Gill, P. et al. (2001) “DNA commission of the International Society of Forensic Genetics: Recommendations on forensic analysis using Y-chromosome strs,” Forensic Science International, 124(1), pp. 5–10. Available at: https://doi.org/10.1016/s0379-0738(01)00498-4.
84. Della Rocca, C. et al. (2022) “Improving discrimination capacity through rapidly mutating Y-strs in structured populations from the African continent,” Forensic Science International: Genetics, 61, p. 102755. Available at: https://doi.org/10.1016/j.fsigen.2022.102755.
85. Otagiri, T. et al. (2022) “RMPLEX reveals population differences in RM y-STR mutation rates and provides improved father-son differentiation in Japanese,” Forensic Science International: Genetics, 61, p. 102766. Available at: https://doi.org/10.1016/j.fsigen.2022.102766.
86. Woerner, A.E. et al. (2022) “Optimized variant calling for estimating kinship,” Forensic Science International: Genetics, 61, p. 102785. Available at: https://doi.org/10.1016/j.fsigen.2022.102785.
87. Xavier, C. et al. (2022) “Evaluation of the Visage basic tool for appearance and ancestry inference using ForenSeq® Chemistry on the MISEQ FGX® system,” Forensic Science International: Genetics, 58, p. 102675. Available at: https://doi.org/10.1016/j.fsigen.2022.102675.
88. Samuel, G. and Prainsack, B. (2019) “Shifting ethical boundaries in forensic use of DNA,” Jahrbuch für Wissenschaft und Ethik, 24(1), pp. 155–172. Available at: https://doi.org/10.1515/jwiet-2019-0007.
89. Schneider, P.M. et al. (2019) “The use of forensic DNA phenotyping in predicting appearance and biogeographic ancestry,” Deutsches Ärzteblatt international [Preprint]. Available at: https://doi.org/10.3238/arztebl.2019.0873.
90. Tillmar, A. et al. (2021) “The force panel: An all-in-one SNP marker set for confirming investigative genetic genealogy leads and for general forensic applications,” Genes, 12(12), p. 1968. Available at: https://doi.org/10.3390/genes12121968.
91. Alghafri, R. et al. (2017) “Rapidly mutating Y-STR analyses of compromised forensic samples,” International Journal of Legal Medicine, 132(2), pp. 397–403. Available at: https://doi.org/10.1007/s00414-017-1600-z.

92. Kayser, M. et al. (1997) “Applications of microsatellite-based Y chromosome haplotyping,” Electrophoresis, 18(9), pp. 1602–1607. Available at: https://doi.org/10.1002/elps.1150180920.

93. Underhill, P.A. et al. (2000) “Y chromosome sequence variation and the history of human populations,” Nature Genetics, 26(3), pp. 358–361. Available at: https://doi.org/10.1038/81685.

94. Hammer, M.F. et al. (2001) “Hierarchical patterns of global human Y-Chromosome diversity,” Molecular Biology and Evolution, 18(7), pp. 1189–1203. Available at: https://doi.org/10.1093/oxfordjournals.molbev.a003906.

95. Oota, H. et al. (2001) “Human mtdna and Y-chromosome variation is correlated with matrilocal versus Patrilocal residence,” Nature Genetics, 29(1), pp. 20–21. Available at: https://doi.org/10.1038/ng711.

96. Jobling, M.A. et al. (2003) “The human Y chromosome: An evolutionary marker comes of age,” Nature Reviews Genetics, 4(8), pp. 598–612. Available at: https://doi.org/10.1038/nrg1124.

97. Roewer, L. et al. (2005) “Signature of recent historical events in the European Y-chromosomal STR haplotype distribution,” Human Genetics, 116(4), pp. 279–291. Available at: https://doi.org/10.1007/s00439-004-1201-z.

98. Shi, M. et al. (2011) “Population genetics for Y-chromosomal strs haplotypes of Chinese xibe ethnic group,” Forensic Science International: Genetics, 5(5). Available at: https://doi.org/10.1016/j.fsigen.2010.08.004.

9 List of Publications

- Della Rocca C, Trombetta B, Barni F, D'Atanasio E, Hajiesmaeil M, Berti A, Hadi S, Cruciani F. Improving discrimination capacity through rapidly mutating Y-STRs in structured populations from the African continent. Forensic Sci Int Genet. 2022 Nov;61:102755. doi: 10.1016/j.fsigen.2022.102755. Epub 2022 Aug 4. PMID: 35985094.
- Oldoni F, Della Rocca C, Podini D. Investigation of 74 microhaplotypes for kinship testing in US populations. Forensic Sci Int Genet Supplement Series, Volume 8, 2022, Pages 40-41, ISSN 1875-1768 doi: https://doi.org/10.1016/j.fsigss.2022.09.015.
- Ravasini F, D'Atanasio E, Bonito M, Bonucci B, Della Rocca C, Berti A, Trombetta B, Cruciani F. Sequence Read Depth Analysis of a Monophyletic Cluster of Y Chromosomes Characterized by Structural Rearrangements in the AZFc Region Resulting in DYS448 Deletion and DYF387S1 Duplication. Front Genet. 2021 Apr 16;12:669405. doi: 10.3389/fgene.2021.669405. PMID: 33936180; PMCID: PMC8085532.
- Della Rocca C, Cannone F, D'Atanasio E, Bonito M, Anagnostou P, Russo G, Barni F, Alladio E, Destro-Bisol G, Trombetta B, Berti A, Cruciani F. Ethnic fragmentation and degree of urbanization strongly affect the discrimination power of Y-STR haplotypes in central Sahel. Forensic Sci Int Genet. 2020 Nov;49:102374. doi: 10.1016/j.fsigen.2020.102374. Epub 2020 Aug 21. PMID: 32890883.
- Alladio E, Della Rocca C, Barni F, Dugoujon JM, Garofano P, Semino O, Berti A, Novelletto A, Vincenti M, Cruciani F. A multivariate statistical approach for the estimation of the ethnic origin of unknown genetic profiles in forensic genetics. Forensic Sci Int Genet. 2020 Mar;45:102209. doi: 10.1016/j.fsigen.2019.102209. Epub 2019 Nov 27. PMID: 31812099.
- Alladio E, Della Rocca C, Cruciani F, Vincenti M, Garofano P, Berti A, Barni F. A multivariate statistical approach to for the evaluation of the biogeographical ancestry information from traditional STRs. Forensic Sci

Int Genet Supplement Series. Volume 7, Issue 1, 2019, Pages 253-255, ISSN 1875-1768, doi: https://doi.org/10.1016/j.fsigss.2019.09.097.

- Della Rocca C, Alladio E, Barni F, Cannone F, D’Atanasio E, Trombetta B, Berti A, Cruciani F. Low discrimination power of the YFiler™ Plus PCR amplification kit in african populations. Do we need more RM Y-STRs?. Forensic Sci Int Genet Supplement Series. Volume 7, Issue 1, 2019, Pages 671-673, ISSN 1875-1768, doi: https://doi.org/10.1016/j.fsigss.2019.10.133.