Extending Y-STR Loci in Portugal for Forensic and Population Studies: The PowerPlex® Y23 Experience

Renato André Pereira Salazar
Mestrado em Genética Forense
Departamento de Biologia
2013

Orientadora
Professora Doutora Maria João Prata Martins Ribeiro, Professora Associada c/ Agregação na Faculdade de Ciências da Universidade do Porto e Investigadora no Instituto de Patologia e Imunologia Molecular da Universidade do Porto.

Co-orientadora
Cíntia Alves, Diretora do Laboratório de Investigação de Parentescos e Identificação Genética, Instituto de Patologia e Imunologia Molecular da Universidade do Porto.

Instituto de Patologia e Imunologia Molecular da Universidade do Porto

Todas as correções determinadas pelo júri, e só essas, foram efetuadas.

O Presidente do Júri,

Porto, _____/_____/_____

Dissertação de candidatura ao grau de mestre em Genética Forense submetida à Faculdade de Ciências da Universidade do Porto.

O presente trabalho foi desenvolvido no Instituto de Patologia e Imunologia Molecular da Universidade do Porto sob orientação da Professora Doutora Maria João Prata Martins Ribeiro e da Drª Cíntia Alves.

Dissertation for applying to a Master's degree in Forensic Genetics, submitted to the Faculty of Sciences of the University of Porto.

The present work was developed at the Institute of Molecular Pathology and Immunology of the University of Porto under the scientific supervision of Professor Maria João Prata Martins Ribeiro and Dr. Cíntia Alves.

AGRADECIMENTOS

À Professor Maria João Prata, por ter aceitado ser minha orientadora e por toda a ajuda e dedicação que teve para comigo. Mostrou-se sempre disponível para ajudar e as suas correções e sugestões, além de melhorarem este trabalho, tornaram esta experiência numa oportunidade única de aprendizagem.

À Cíntia, por tudo o que me ensinou, nomeadamente a nível prático. Foi a Cíntia que esteve comigo desde o início do trabalho e com a sua paciência, simpatia e dedicação ensinou-me tudo e mais alguma coisa! Não me vou esquecer da confiança que depositou em mim!

Ao Luís pela ajuda com vários programas e pelas sugestões que foi dando ao longo deste trabalho.

Ao Professor António Amorim por me ter aceitado neste Mestrado e portanto ter tornado tudo isto possível.

Ao IPATIMUP por me receber e permitir que este trabalho fosse realizado.

À Célia por todo o apoio e ajuda (e não foi pouca!) que me foi dando ao longo do trabalho e obviamente pela ótima companhia que sempre foi!

Ao Guilherme e à Andreia pela ajuda que me foram dando ao longo deste ano, além da companhia, claro!

A todos os meus colegas de Mestrado e ao Grupo de Genética Populacional.

À minha família e a todos os meus amigos porque sempre me apoiaram e sem eles de certeza que não teria chegado aqui!

ABSTRACT

The specific properties of the Y chromosome make it highly informative not only for tracing human migration and evolution through male lineages but also for forensic studies. While different kind of Y-polymorphisms were proved to be useful in routine forensic casework, up to now short tandem repeats (STRs) have been the most commonly used due to their high levels of diversity when compared to other polymorphisms.

In this study, the recently released PowerPlex® Y23 System (Promega) was evaluated viewing its application in forensic casework in the Portuguese population. The 23 Y-chromosomal STRs included in the kit (DYS576, DYS389I, DYS448, DYS389II, DYS19, DYS391, DYS481, DYS549, DYS533, DYS438, DYS437, DYS570, DYS635, DYS390, DYS439, DYS392, DYS643, DYS393, DYS458, DYS385, DYS456, GATA H4) were typed in 250 samples from unrelated Portuguese males. A total of 236 different haplotypes were found, among which 14 were shared by two individuals. The overall haplotype diversity (HD) was 0.9996. Since the same sample had been previously typed with the AmpFISTR Yfiler™ Amplification Kit (Applied Biosystems), a comparison was undertaken in terms of concordance and HD. Two discrepancies were found between both kits for the loci DYS385 and GATA H4, which were further demonstrated to be caused by a silent allele and a sequence variant, respectively.

It was also assessed which new loci in PowerPlex® Y23 contributed more to HD increment in this sample: by fixing the haplotypes generated with the YFiler™ loci and adding the six new PowerPlex® Y23 loci one by one, the markers that increased HD the most were DYS576 and DYS481. In this sample, the same HD was obtained with or without the DYS643 locus.

Since a group of samples had been previously typed for Y-SNP markers, a haplogroup predictor software was used to evaluate which set of Y-STR markers is currently more efficient in the prediction of haplogroups. The 17 Y-STR set of Yfiler™ showed to be slightly more accurate than the 23 Y-STR set of PowerPlex® Y23.

Population comparisons with several worldwide population data were undertaken through MDS based on FST values, which has illustrated the sharp clustering of three major groups of populations: European, African and Eastern Asian. At the European level, geography was found to account considerably for the pattern of population substructuring, although inconsistencies were not rarely observed in the relationship between geographical and genetic distance. The same analysis based on 17 Y-STR data

has led to similar results, but the F_{ST} values based on this panel of markers were on average higher than produced with the set of PowerPlex® Y23. This finding can possibly be explained by the high levels of diversity and mutability rates of the 6 additional loci, which consequently, when used to extend the number or markers in Yfiler™ might imply some loss in the ability to capture inter-population diversity.

Keywords: PowerPlex Y23, Y-STR, Y Chromosome, Forensic Genetics and Portugal.

RESUMO

As propriedades específicas do cromossoma Y tornam-no útil não só para reconstituir a história da evolução e das migrações humanas pela perspetiva das linhagens paternas, mas também para estudos forenses. Embora não faltem provas de que diferentes tipos de polimorfismos do cromossoma Y podem contribuir muito para a resolução de casos forenses, até agora os *short tandem repeats* (STRs) têm sido os mais usados devido aos elevados níveis de diversidade que os caracterizam quando comparados com outros polimorfismos.

Neste trabalho, o *kit* comercial PowerPlex® Y23 System, recentemente lançado pela Promega, foi avaliado tendo em vista a sua aplicação em casos forenses na população Portuguesa. Os 23 STRs do cromossoma Y incluídos neste *kit* (DYS576, DYS389I, DYS448, DYS389II, DYS19, DYS391, DYS481, DYS549, DYS533, DYS438, DYS437, DYS570, DYS635, DYS390, DYS439, DYS392, DYS643, DYS393, DYS458, DYS385, DYS456, GATA H4) foram tipados em 250 amostras de indivíduos masculinos não aparentados naturais de Portugal. Encontrou-se um total de 236 haplótipos diferentes, dos quais 14 eram partilhados por dois indivíduos. A diversidade haplotípica (HD) registada foi de 0.9996. Como os mesmos indivíduos tinham sido previamente tipados com o AmpFISTR Yfiler™ Amplification Kit (Applied Biosystems), foi feita uma comparação em termos de concordância e HD. Duas discrepâncias foram detetadas entre os resultados de tipagem obtidos com os dois *kits* envolvendo os loci DYS385 e GATA H4, tendo sido possível esclarecer serem resultantes da presença de um alelo silencioso e de uma variação de sequência, respetivamente.

Procurou-se também inferir quais dos novos loci integrados no PowerPlex® Y23 contribuíam mais para o aumento de HD, fixando os haplótipos gerados pelo Yfiler™ e adicionando os seis novos loci um por um. Os marcadores que mais aumentaram a HD foram DYS576 e o DYS481, enquanto que na mesma amostra, o valor de HD não se alterou com a adição do locus DYS643.

Como tinham sido previamente tipados Y-SNPs em parte das amostras, um *software* de previsão de haplogrupos foi usado para avaliar qual o conjunto de STRs que seria mais eficiente na previsão de haplogrupos. Com o conjunto de 17 STRs do Yfiler™ observou-se uma taxa de previsões corretas ligeiramente superior à dada pelo conjunto de 23 STRs do PowerPlex® Y23.

Foram ainda efetuadas comparações populacionais usando informação de várias populações mundiais, nomeadamente através de MDS baseado em valores de

F_{ST}, o que permitiu constatar que as populações se agregavam em três grandes grupos populacionais, bastante bem definidos: Europeus, Africanos e Asiáticos de Leste. Na Europa, a geografia é um fator que contribui consideravelmente para explicar o padrão de subestruturação populacional, embora sejam comuns inconsistências entre distância geográfica e distância genética. Idênticos estudos populacionais baseados em haplótipos do Yfiler™, produziram resultados semelhantes, embora os valores médios de F_{ST} fossem um pouco mais elevados com o Yfiler™ do que os valores obtidos com o PowerPlex® Y23. Esta diferença pode residir nos elevados níveis de diversidade e de taxas de mutação dos 6 loci adicionais do PowerPlex® Y23. Face as estas características, quando os 6 loci são usados para estender o número de marcadores do Yfiler™, tal parece acarretar um decréscimo na capacidade de detetar subestruturação inter-populacional, comparativamente à que oferece o conjunto mais restrito de STRs do Yfiler™.

Palavras-chave: PowerPlex Y23, Y-STR, Cromossoma Y, Genética Forense e Portugal.

INDEX

FIGURES INDEX

TABLES INDEX

ABBREVIATIONS

In order of appearance:

DNA – Deoxyribonucleic Acid
VNTR – Variable Number of Tandem Repeats
RFLP – Restriction Fragment Length Polymorphism
PCR – Polymerase Chain Reaction
STR – Short Tandem Repeat
SWGDAM – Scientific Working Group for DNA Analysis Methods
EDNAP – European DNA Profiling Group
ISFG – International Society for Forensic Genetics
PAR – Pseudoautosomal Region
NRY – Non Recombining Y
MSY – Male-specific Region
Mb – Mega Bases
Yq – Y Chromosome Long Arm
Xq – X Chromosome Long Arm
SNP – Single Nucleotide Polymorphism
UEP – Unique Event Polymorphism
ISOGG – International Society of Genetic Genealogy
Y-SNP – Single Nucleotide Polymorphism on the Y Chromosome
Y-STR – Short Tandem Repeat on the Y Chromosome
bp – Base Pair
GDB – Human Genome Database
ENFSI – European Network of Forensic Science Institutes
YHRD – Y-STR Haplotype Reference Database
ng – Nanogram
µL – Microliter
ºC – Degree Celsius
min – Minute
sec - Second
µM – Micromolar
HD – Haplotype Diversity
MDS – Multidimensional Scaling

1 INTRODUCTION

1.1 Forensic Genetics

Forensic genetics can be defined as the application of genetics (in the sense of a science with the purpose of studying inherited characteristics for the analysis of inter- and intraspecific variations in populations) to the resolution of legal disputes [1].

The consolidation of the field of forensic genetics has been driven by analysis of human genetic variation beginning over a century ago. In 1900 Karl Landsteiner described the ABO blood group system having found out that individuals could be classified into different groups based on their blood type. When later in 1924, Felix Bernstein demonstrated that the system was transmitted according to rules of Mendelian inheritance, soon it became evident that the ABO system could be applicable in solving paternity testing cases and crimes. In the following years numerous other blood groups, the complex HLA system of white blood cells and several polymorphisms in serum proteins or erythrocyte enzymes were characterized and could be analyzed in combination to produce highly discriminatory profiles. That kind of genetic markers constituted a powerful tool but were limited in forensic cases due to the amount of biological material required and to the rapid storage protein degradation [1-3].

In the 1960s and 1970s, developments in molecular biology, including restriction enzymes, Sanger sequencing and Southern blotting, enabled scientists to examine DNA sequences and in 1980 the analysis of the first highly polymorphic locus was reported [2]. Years later, in 1985, Alec Jeffreys described the method of “DNA fingerprinting”. He found that certain genomic regions contained DNA sequences that were repeated over and over again next to each other. Jeffreys also discovered that the number of repeats present in a sample could differ from individual to individual. By developing a technique to examine the length variation of these DNA repeat sequences, he increased the ability to perform human identity tests. These repeat regions became known as VNTRs (variable number of tandem repeats) and the technique used by Jeffreys to examine the length variation of these DNA repeat sequences was based on RFLP (restriction fragment length polymorphism) analysis, which involved the use of restriction enzymes to cut regions of DNA surrounding the VNTRs [2, 4].

A critical development in the history of forensic genetics came with the advent of a technology that can amplify specific regions of DNA, the polymerase chain reaction

(PCR). The PCR was conceptualized by Kary Mullis in 1983 and had a huge impact on disciplines relying much on molecular biology techniques, among which is included forensic genetics. The advent of PCR highly increased the sensitivity of DNA analysis to the point where DNA profiles could be generated from just a few cells or degraded DNA, reduced the time required to produce a profile, and allowed just any polymorphism in the genome to be analyzed [2].

A few years later, around 1989, STRs (short tandem repeats), became the markers of choice for many areas, including forensic genetics where they were recruited since the 1990s and continue today to be the most commonly used genetic markers. The advantages of these markers soon became apparent and a systematic search for the most convenient STRs began [1, 2]. This was followed by a strong investment in standardization of techniques and nomenclature accomplished by groups of forensic scientists such as the Scientific Working Group for DNA Analysis Methods (SWGDAM) in the United States and the European DNA Profiling Group (EDNAP) in Europe plus the active role of the DNA Commission of the International Society for Forensic Genetics (ISFG).

Another important step was the possibility of amplifying multiple STR loci in a single combined multiplex reaction, which in addition to the development of direct detection of amplified products through capillary electrophoresis, the introduction of fluorescent dye-labeled primer technology and the widespread accessibility to DNA sequencers revolutionized the typing strategies in the field of forensic genetics; since then several commercial dye-labeled multiplexes have become available [1].

Around mid-1990s, computer databases containing DNA profiles from crime scene samples, convicted offenders and in some cases persons simply arrested for a crime, have provided law enforcement with the ability to link offenders to their crimes. The establishment of such databases has enabled tens of thousands of crimes — particularly serial crimes by repeat offenders — to be solved around the world [4].

The combination of technical advances, high levels of standardization and quality control have led forensic DNA analysis to be recognized as a robust and reliable forensic tool worldwide [2].

1.2 The Y Chromosome

Structure

The human Y chromosome stands out from all other chromosomes because it is male-specific and is clonally transferred from father to son, except in the two small

telomeric regions that pair and crossover with the X chromosome during meiosis [5]. It represents only 2 % of the human genome and it is approximately 60 Mb in length. As shown in **Fig. 1**, there are two pseudoautosomal regions (PAR1 and PAR2) located at the distal portions of the short and long arms of the chromosome. Since these short regions are homologous of X chromosome sequences, they are responsible for correct pairing between the two sex chromosomes during male meiosis. The majority of the Y chromosome (95%), termed non recombining Y (NRY), does not undergo recombination during meiosis. It is a haploid entity and, therefore, is transmitted intact through paternal lineages [6-10]. Skalestsky *et al.* (2003) renamed the NRY as the male-specific region (MSY) because of evidence of frequent gene conversion or intrachromosomal recombination [11]. Structurally, three domains have been identified in the reference MSY: the euchromatic one spanning approximately 23 Mb, the centromeric region (~1 Mb), and two Yq heterochromatin blocks, the more distal of which extends for about 40 Mb. The latter exhibits a length polymorphism that ultimately accounts for the significant size variation of the Y in the male population [6]. Many sequences in the Y chromosome are highly duplicated either with themselves or with the X chromosome. Three classes of sequences have been characterized in the Y chromosome: X-transposed, X-degenerate, and ampliconic. With a length of 3.4 Mb, the X-transposed sequences are 99% identical to sequences found in Xq21 and do not participate in X-Y crossing over during meiosis. X-degenerate segments of MSY (8.6 Mb) possess up to 96% nucleotide sequence identity to their X-linked homologues. These X-homologous regions can make it challenging to design Y chromosome assays that generate male-specific DNA results. If portions of an X-homologous region of the Y chromosome are examined inadvertently, then female DNA, which possesses two X chromosomes, will be detected. Thus, when testing Y chromosome-specific assays it is important to examine also female DNA to verify that there is little-to-no cross talk with X-homologous regions of the Y chromosome. The ampliconic segments cover about 10.2 Mb of the Y chromosome. Some 60% of these sequences have intrachromosomal identities of 99.9% or greater, so it is very difficult to distinguish these sequences. Another interesting feature of these ampliconic segments is that many of them are palindromes – that is, almost exact duplicate sequences which are inverted with respect to each other's sequence essentially as mirror images [11, 12].

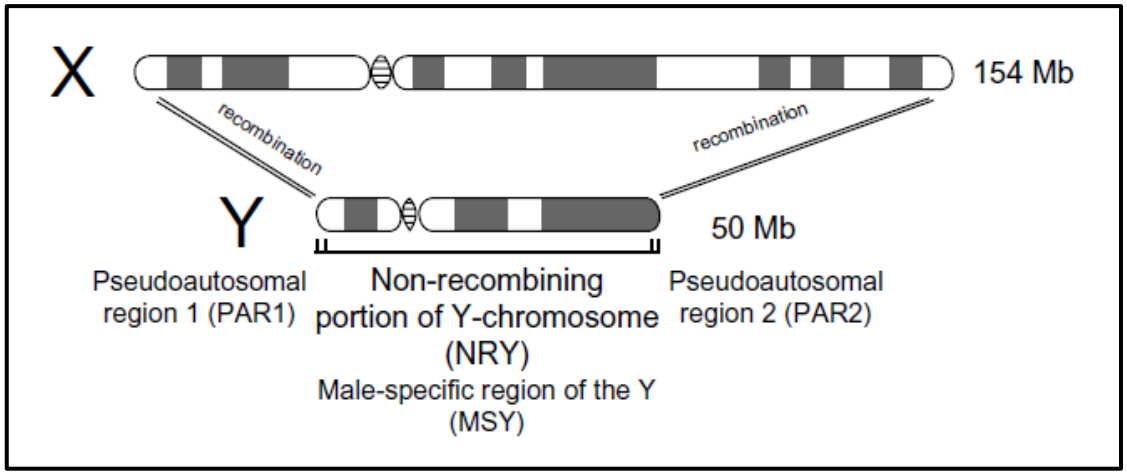

Fig. 1 – Schematics of the sex chromosomes. The tips of Y chromosome (PAR1 and PAR2) recombine with the tips of the X chromosome. The remaining 95% is the MSY region. Adapted from Butler, 2012 [12].

Origin

X–Y homology in the PAR and the preponderance of XY shared genes support the evolutionary mechanism that the mammalian X and Y originated from a pair of autosomes (**Fig. 2**). Differentiation of the proto-Y began when it acquired a male determining locus, then there was accumulation of other male-advantage genes in a region across which recombination was suppressed. Lack of recombination resulted in progressive degradation because selection no longer acted upon a single gene, but rather the entire MSY. Under these conditions the Y degraded because of higher variation, drift and inefficient selection. The Y chromosome seems to be subject to far more mutation, deletion and insertion than the rest of the genome. This can be attributed to the fact that the Y must spend every generation in the hostile environment of the testis, whereas the presence of X and autosomal chromosomes in testis is only a third or half as often. The “hostility” of testis is due to the large amount of divisions needed to make a sperm, to the oxidative environment and to the lack of repair enzymes in the sperm. In addition, the repetitive structure of the Y chromosome makes deletions very frequent [10, 13-15].

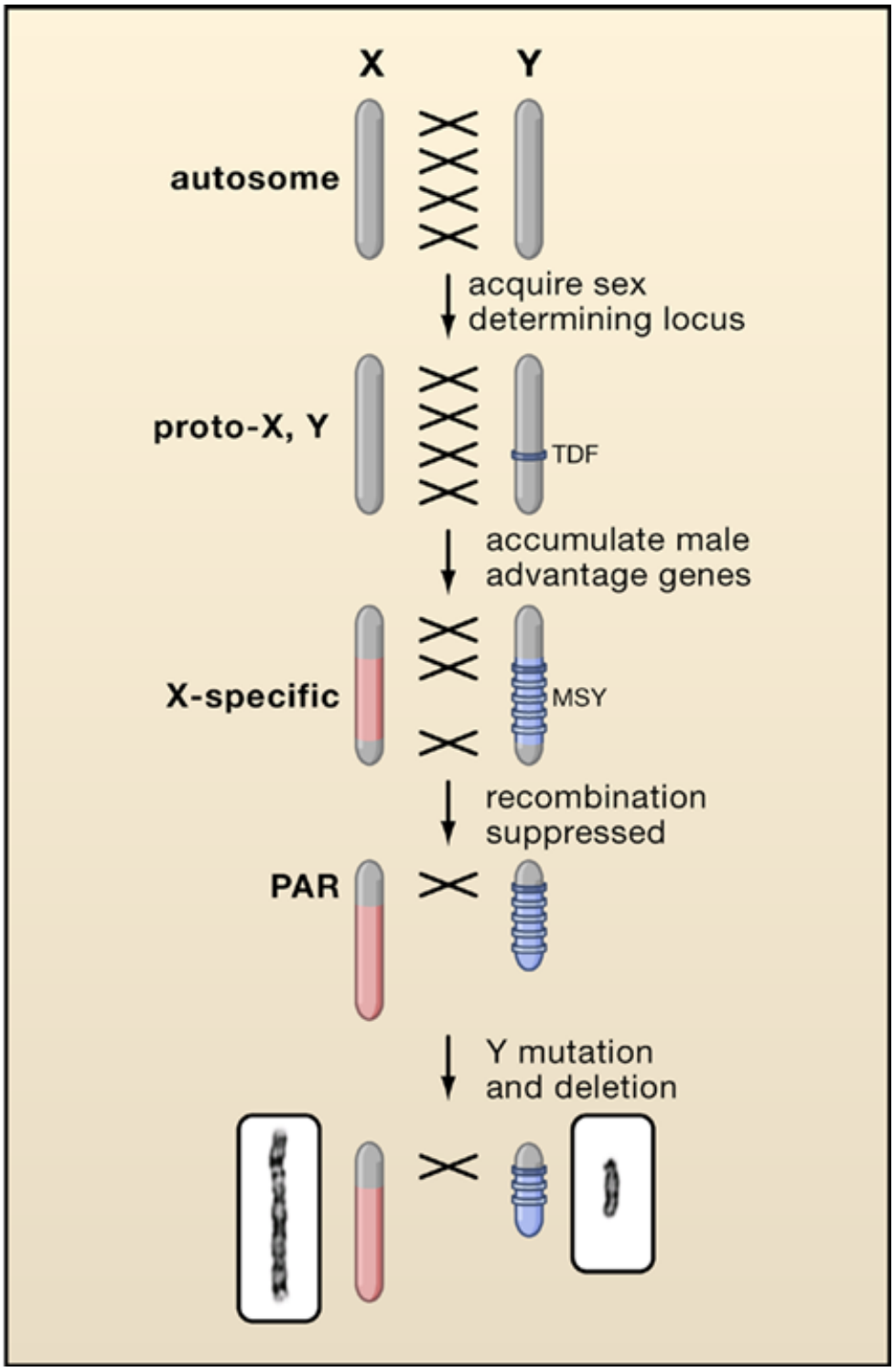

Fig. 2 – Mammalian sex chromosomes origin and evolution from an autosomal pair. Extracted from Graves, 2006 [10].

Forensic Applications

The Y chromosome is valuable in forensic DNA testing because it is found only in males. Since a vast majority of crimes concerning DNA evidence, particularly sexual assaults, involve males as the perpetrator, tests designed to examine restrictively the Y chromosome can be valuable. This way, informative results can be obtained in some cases where analysis of autosomal markers are limited by the evidence, such as high levels of female DNA in the presence of minor amounts of male DNA (**Fig. 3**). In addition, the number of individuals involved in a "gang rape" may be easier to decipher than with complicated autosomal STR mixtures [12, 16].

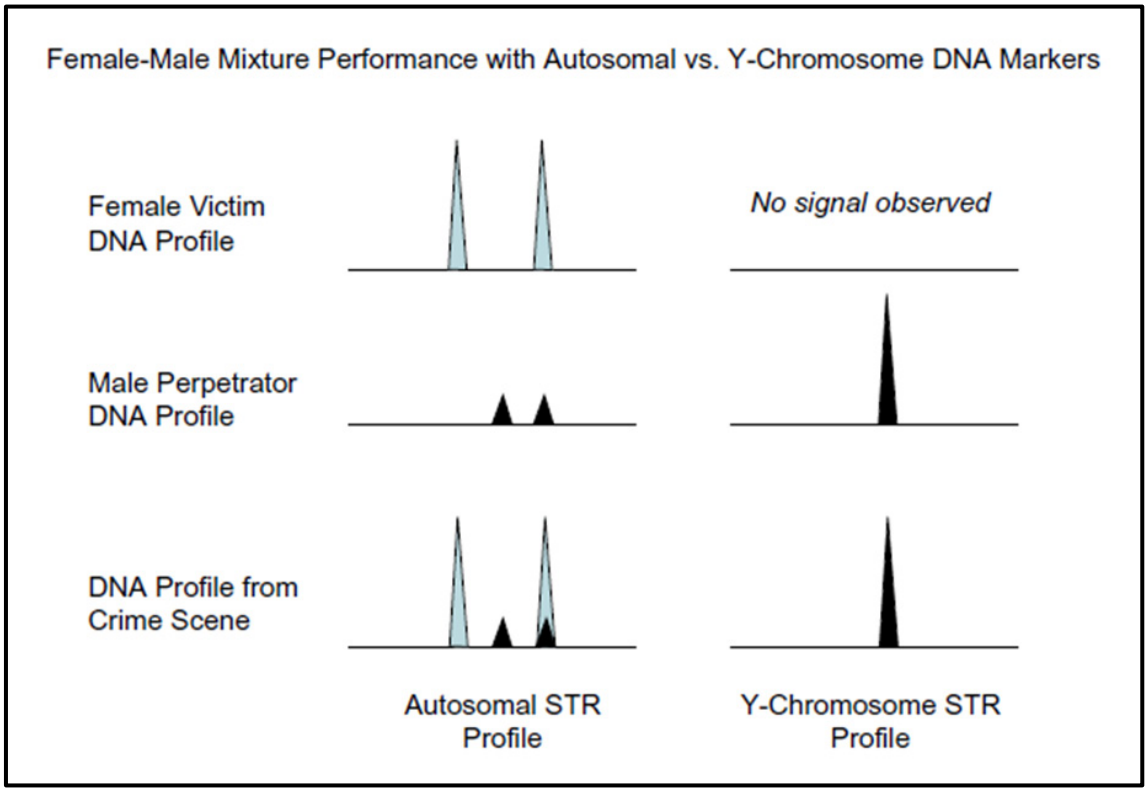

Fig. 3 – Types of autosomal or Y-STR profiles that might be observed with sexual assault evidence where mixtures of high amounts of female DNA may mask the STR profile of the perpetrator. Y-STR testing permits isolation of the male component without having to perform a differential lysis. Extracted from Butler, 2012 [12].

What makes the Y chromosome so special and so helpful to forensic DNA testing can also be its major limitation. Almost the entire Y chromosome is transferred from father to son without recombination. Random mutations are the only source of variation between males from the same paternal lineage. Thus, while exclusions based on results in Y chromosome DNA testing can aid forensic investigations, a match between a

suspect and an evidence sample only means that the individual in question could have contributed to the forensic stain, as could every individual of the same paternal lineage.

The presence of relatives having the same chromosome, on the other hand, can be fundamental in cases of missing persons and mass disaster victim identification because it expands the number of possible reference samples. Deficient paternity tests, meaning those with samples unavailable from the putative father, can also benefit from Y chromosome markers.

Y chromosome testing has become useful for making inferences on human migration and other population genetics issues as well in a number of evolutionary studies because the lack of recombination enables comparison of male individuals separated by large periods of time. Historical research has also been performed using Y chromosome testing as well as genealogical investigations because surnames are most usually transmitted through the paternal lineage and so Y chromosomal information can help make links where the paper trail is limited [2, 8, 9, 12, 17].

Genetic Markers

DNA markers used to examine Y chromosome diversity can be divided in two categories: bi-allelic loci and multi-allelic loci. Results provided by bi-allelic markers are commonly assumed to define haplogroups while those from multi-allelic markers define haplotypes [12, 18].

Bi-allelic markers include single nucleotide polymorphisms (SNPs) and *Alu* element insertions. These markers are sometimes referred as unique event polymorphisms (UEPs) because of their low mutation rates (~10^{-8} to ~10^{-9} per generation) [12]. The simplest type of polymorphism, a genetic variation with a minor allele at frequency ≥ 1% in a population is the SNP. SNPs arise when mutations occur in the cells undergoing DNA replication during meiosis, changing a nucleotide by another. They normally have just two alleles, for example one allele with a guanine and one with an adenine (**Fig. 4**).

Currently, more than 600 bi-allelic Y chromosome markers have been characterized. Many of them were used to construct the updated version of the NRY-haplogroup tree, which represents a phylogenetic reconstruction of the integrated binary polymorphisms. The first phylogenetic chart was published by the Y Chromosome Consortium in 2002 [19]. In 2005, the ISOGG (International Society of Genetic Genealogy) group was formed to create a website which could be updated to keep pace with the rapid developments of the field and can be consulted on

http://www.isogg.org/tree/ [20]. Since the bi-allelic Y markers define haplogroups that are characterized by strong geographical structure, their study has been intensively used in molecular anthropology for evolutionary research. However, they also revealed to be important tools in a variety of other areas, including DNA forensics, although in the last field, because of their low discrimination power, these markers have an important disadvantage against the STRs in routine casework. To match approximately the power of a set of 13 to 15 STR loci, around 40 to 60 SNP are necessary, yet these calculations are based on autosomal markers [2, 12, 19, 21-24].

Fig. 4 – A single nucleotide polymorphism. Two alleles differ at one position indicated by the star. Extracted from Goodwin , 2007 [2].

The most commonly studied multi-allelic markers of the Y include two minisatellites (MSY1 and MSY2) and several hundred short tandem repeat (Y-STR) markers. These multi-allelic loci can be used to differentiate Y chromosome haplotypes with fairly high resolution due to their higher mutation rates (6 x 10^{-2} to 9 x 10^{-2} per generation for minisatellites and 2.1 x 10^{-3} for Y-STRs) [12, 25-28].

STRs are currently the most commonly analyzed genetic polymorphism in forensic genetics. They have a core repeat unit of between 1 and 6 bp and the number of repeats typically range from 50 to 300 bp. The majority of the loci that are used in forensic genetics are tetranucleotide repeats, which have a four base pair repeat motif (**Fig. 5**) [2].

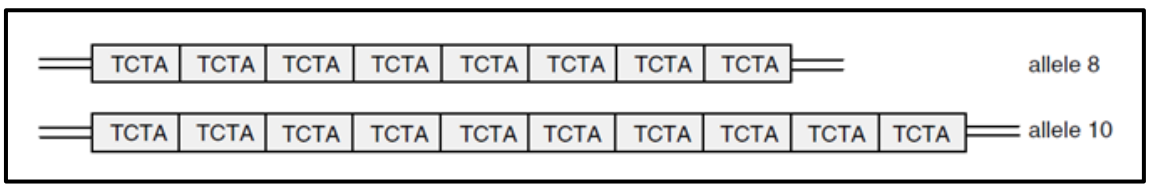

Fig. 5 – A short tandem repeat. The alleles are named according to the number of repeats they contain. Extracted from Goodwin, 2007 [2].

STRs satisfy all the requirements for a forensic marker: they are robust, since they lead to successful analysis of a wide range of biological material even in non-optimal conditions; the results generated in different laboratories are easily comparable; they are highly discriminatory, especially when analyzing a large number of loci simultaneously; they are very sensitive, requiring only a few cells for a successful analysis; and it is relatively cheap and easy to generate STR profiles [2]. In a certain sense, the properties that distinguish Y-STRs from autosomal markers are those that make the first useful for human identification. Since the forensic Y-STR markers are located on the nonrecombining region of the Y chromosome, they produce a haplotype profile when amplified from male DNA samples. These are the properties that make possible the specific applications of the Y chromosome in forensics, as was already mentioned [29, 30].

The number of Y-STR loci available for use in human identity testing has increased dramatically since the turn of the century, accompanying the growing availability of the human genome sequence. By the end of 1900s only a handful of Y-STR markers were characterized and available for use. At the beginning of 2002, about 30 Y-STRs were commonly recruited by researchers. Since that time more than 400 new Y-STRs have been deposited in the Human Genome Database (GDB). Although this repository of DNA marker information is no longer available online, much information regarding Y-STRs, such as nucleotide sequence, can be found on GenBank Database, available at http://www.ncbi.nlm.nih.gov/ [12, 25, 31].

In 1997, even with a limited number of loci on hand, a core set was selected that continue to be named as “minimal haplotype” loci. It is defined by the single copy Y-STR loci DYS19, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393 and DYS385. In 2003, the SWGDAM recommended the addition of the STRs DYS438 and DYS439 to the minimal haplotype. Although other Y-STRs may be added to forensic databases as their further increase in the power of discrimination is demonstrated, and many are becoming part of commercially available kits, the original minimal haplotype loci and SWGDAM recommended Y-STRs are likely to dominate human identity applications in the coming years [12, 32-34].

1.3 Commercial Y-STR Kits

Forensic routine rely heavily on commercially available kits to perform DNA testing. Most laboratories do not have time or resources to design primers, optimize PCR multiplexes, and control the quality of primer synthesis. The convenience of using ready-

made kits is also augmented by the fact that widely used primer sets and conditions allow improved opportunities for sharing data between laboratories without fear of failing the detection of silent alleles. Thus, many laboratories were reluctant to move into Y-STR typing until Y-STR kits were available [12].

Another major advantage in having commercial Y-STR kits is the availability of common allelic ladders. These allelic ladders provide consistent currency that aids in quality assurance of results as well as compatibility of data going into DNA databases. Since various kits might differ concerning the alleles present in their ladders the ability to reliably call a rare allele is greater with the kit which has the ladder with the broader range [35].

Although several kits have been released (**Table 1**), PowerPlex® Y and Yfiler™ are the most widely used. But the recent release of the PowerPlex® Y23 may change this situation in the near future [12].

Kit name (Source)	Release Date	Loci Amplified
Y-Plex™ 6 (ReliaGene Technologies)	2001	DYS393, DYS19, DYS389II, DYS390, DYS391, DYS385
Y-Plex™ 5 (ReliaGene Technologies)	2002	DYS389I, DYS389II, DYS439, DYS438, DYS392
genRES® DYSplex-1 (Serac)	2002	DYS390, DYS39I, DYS385, Amelogenin, DYS5389I/II
genRES® DYSplex-2 (Serac)	2002	DYS392, DYS393, DYS19, DYS389I/II
Y-Plex™ 12 (ReliaGene Technologies)	2003	DYS392, DYS390, DYS385, DYS393, DYS389I, DYS391, DYS389II, Amelogenin, DYS19, DYS439, DYS438
PowerPlex® Y (Promega Corporation)	2003	DYS391, DYS389I, DYS439, DYS389II, DYS438, DYS437, DYS19, DYS392, DYS393, DYS390, DYS385
MenPlex® Argus Y-MU (Biotype)	2004	DYS393, DYS390, DYS385, DYS391, DYS19, DYS389I, DYS392, DYS389II
Yfiler™ (Applied Biosystems)	2004	DYS456, DYS389I, DYS390, DYS389II, DYS458, DYS19, DYS385, DYS393, DYS391, DYS439, DYS635, DYS392, H4, DYS437, DYS438, DYS448

Kit name (Source)	Release Date	Loci Amplified
Investigator® Argus Y-12 QS (Qiagen)	2010	DYS19, DYS385, DYS389I, DYS389II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439
PowerPlex® Y23 (Promega Corporation)	2012	DYS576, DYS389I, DYS448, DYS389II, DYS19, DYS391, DYS481, DYS549, DYS533, DYS438, DYS437, DYS570, DYS635, DYS390, DYS439, DYS392, DYS643, DYS393, DYS458, DYS385, DYS456, H4

Table 1 – Commercial Y-STR kits. Currently only Powerplex® Y, Yfiler™, Investigator® Argus Y-12 QS and PowerPlex® Y23 are available. Adapted from Butler, 2005 [12, 35, 36].

Validation

Before the introduction of a kit in a forensic lab setting, it needs to be submitted to a process of validation. According to Butler [37], validation refers to the process of demonstrating that a laboratory procedure is robust, reliable and reproducible. A robust method is one in which successful results are obtained a high percentage of the time and few, if any, samples need to be repeated. A reliable method refers to one in which the obtained results are accurate and correctly reflect the sample being tested. A reproducible method means that the same or very similar results are obtained each time a sample is tested. There are generally considered to be two stages of validation, i) developmental and ii) internal validation. Developmental validation is usually performed by the commercial kit manufacturers and large laboratories, while internal validation is performed by an individual laboratory when a new method is introduced. Despite some efforts of the European Network of Forensic Science Institutes (ENFSI) or the ISFG, specific and standardized guidelines for the validation process are still missing. The most accepted guidelines are the SWGDAM recommendations [37].

i) The developmental validation studies listed in the SWGDAM Revised Validation Guidelines, include the characterization of the genetic marker, species specificity, sensitivity studies, stability studies, reproducibility, case-type samples, mixture studies, population studies, precision and accuracy studies, as well as PCR based studies [38-40].

ii) Forensic DNA laboratories conduct internal validation studies as part of becoming “validated”. These studies demonstrate that DNA typing results can be consistently and accurately obtained in the specific laboratory environment where the testing is performed. Typical studies for an internal validation include reproducibility, precision measurements for sizing alleles, sensitivity, mixture studies and non-probative casework samples. These studies should be performed using at least 50 samples [38-41].

After a procedure has been successfully implemented for use with forensic casework, proficiency tests are performed on a regular basis to demonstrate successful application of the technique over time by qualified analysts. In addition, new materials and instruments need to be evaluated over time through a quality control process involving a performance check on the validated procedure. These performance checks consist in verifying that the instrument or reagents are working properly, monitoring and assessment of control samples and internal size standard run with each test or set of samples [38].

1.4 Concordance studies

As mentioned before, human STR genotyping for population and forensic studies is mainly performed using commercial kits. Although from different manufacturers, they share many STR loci, most of which are amplified with distinct primer sets [42]. The use of different primer sets might lead to discordant STR allele calls due to different reasons, among which is the presence of silent alleles. In fact silent alleles have been often “discovered” by the observation of different typing results when using kits with different primer sets as, for instance, it happened in a concordance study that combined data of 600 samples for the autosomal locus vWA [2, 38, 43]. The cause of discordant calls after amplification with different primer sets resides in sequence polymorphisms. Variations can occur within or around STR repeat regions, in three locations relative to the primer binding sites: within the repeat region, in the flanking region, or in the primer binding region. However, if a base pair change occurs in the DNA template at the PCR primer binding region, the hybridization of the primer can be disrupted, resulting in a failure to amplify during PCR due to primer hybridization problems (**Fig. 6**) [38]. This is the main source of silent alleles.

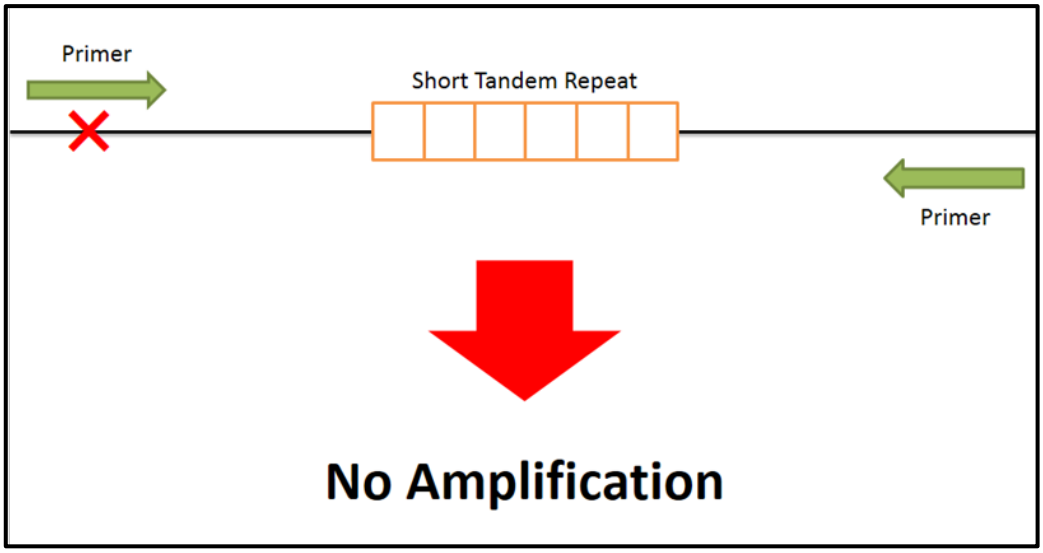

Fig. 6 – Silent allele. A mutation (red cross) at the primer binding region prevents the hybridization of the primer resulting in a failure to amplify during PCR.

On the other hand, apparent silent alleles might arise if there is a deletion within the amplified region that causes the resulting smaller-size amplicon to reside at a position in a smaller locus that differs in two multiplex kits (**Fig. 7**). In this case, the use of different commercial kits can lead to potential nomenclature mistakes. Because of the primers' design differences, the extra peak would appear in a different marker for each kit, although the absence of the allele would be detected in the same locus. This situation would result in two different profiles for each kit, one with two alleles in one locus and the other with two alleles in a different locus, but both with the same locus without an allele [44].

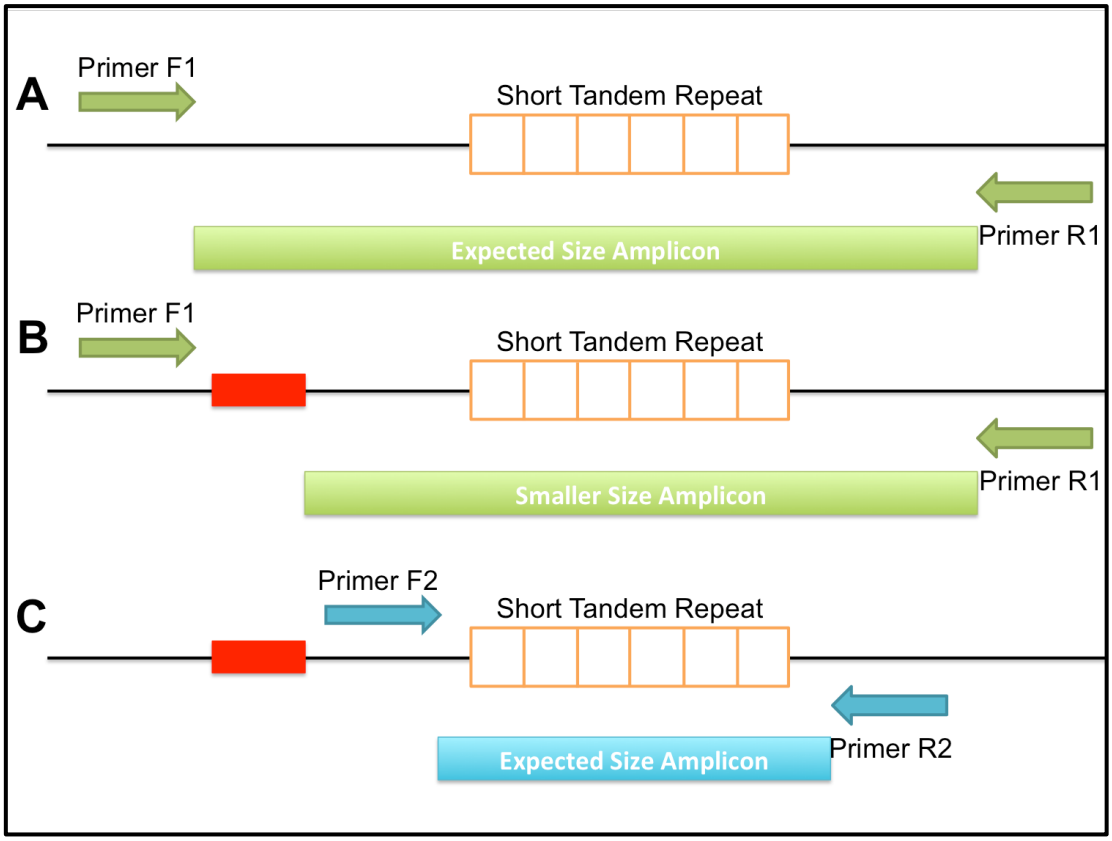

Fig. 7 – Apparent silent allele. A deletion (red rectangle) within the amplified region results in a smaller amplicon (B) than expected (A). Depending on the size of the deletion, this amplicon can reside at a position in a smaller locus. An amplification using a different set of primers which range does not include the deleted zone, results in an amplicon with the expected size for this set of primers (C).

Several concordance studies have been conducted in the past few years, including one very recently performed by Davis and co-workers (2013) with a Prototype PowerPlex® Y23 kit and Yfiler™ kit. In this study, all the typing results for the common Y-STR loci were concordant, and both kits allowed the identification of 13 silent alleles at DYS448 loci [44]. Concordance studies are also important between the various STR kits to assess the level of potential allele dropout, which besides could be caused by silent alleles, can also be observed when low quantity of input DNA is used, resulting in the failure to amplify one or more alleles in the sample (typically the ones with a longer product size), and by allele sizes outside of the normal calling range for a particular locus, resulting in it not being detected [38, 45-47].

1.5 Y-STR Haplotype Databases

There are several Y-STR databases available online. Most of them can be divided in two groups according to the main aim: the genetic genealogical databases and the forensic databases.

The former, such as Ysearch (http://www.ysearch.org/) and SMGF (http://www.smgf.org/), contain Y-STR haplotype information based on different sets of loci from males gathered by genetic genealogy companies, in order to make genealogical researches. Thus, the haplotypes in these genealogy databases are associated with specific individuals and family names.

Forensic databases, such as YHRD or US Y-STR Database (United States population specific), contain data from samples of anonymous and unrelated males that can be used to estimate the frequency of specified Y-STR haplotypes [12].

YHRD

In 2000, it became available online what is now the most widely used database in forensic and other population genetics researches. Created by Lutz Roewer and co-workers, the Y-STR Haplotype Reference Database (YHRD) (http://www.yhrd.org/) was designed to strore Y chromosome haplotypes from worldwide populations. As of March 2013, it contained 112 005 haplotypes from 834 different populations, out of which 5 301 had already full profiles for all the 23 loci present in PowerPlex®Y23 kit. This database was launched with two main objectives: the generation of reliable Y-STR haplotype frequency estimates for Y-STR haplotypes to be used in the quantitative assessment of matches in forensic and kinship casework, and the assessment of male population stratification among world-wide populations as far as reflected by Y-STR haplotype frequency distributions [48].

To this end, a large and steadily growing number of diagnostic and research laboratories have joined in a collaborative effort to collect population data and to create a large reference database. All institutions contributing to this project must participate in an obligate quality control exercise. This is an interactive database that allows the user to search for Y-STR and Y-SNP typed haplotypes and haplogroups in various formats and within specified national databases and metapopulations [48-50].

1.6 Y-STR Results Interpretation

Since the Y chromosome is transmitted unchanged from father to son unless a mutation occurs, the genetic information it contains is consequently shared by all individuals related through the paternal line. The lack of recombination between Y chromosome markers means that Y-STR results have to be combined into a haplotype for searching available databases as well as for estimating the frequency of a particular haplotype. Consequently, the observation of a match with Y-STRs does not have the same power of discrimination and weight in court as an autosomal STR match would.[12].

There are basically three possible interpretations resulting from comparing Y-STR haplotypes produced from two different samples, for simplicity referred to as question and reference samples: exclusion, because the Y-STR profiles are different and could not have originated from the same source; inconclusive, when there is insufficient data to render an interpretation or when ambiguous results were obtained; failure to exclude, where the Y-STR haplotype results from the question and reference samples are the same and could have originated from the same source [12].

When the question and reference samples do not match, then Y-STR typing is helpful in demonstrating the exclusion. However, when the profiles match, the significance of the match has to be carefully assessed. The first step is to estimate frequencies of the Y-STR haplotypes in the population of interest. The simplest method, known as the counting method, consists in reporting the frequency of the Y-STR haplotype in the population. The figure quoted is entirely dependent upon the size of the database and is normally based on searches in databases that are constructed for the major ethnic groups represented within individual countries, although comparisons can also be made to the combined data in the databases. So, for example, a match can be reported as “the haplotype has been observed once in 300 USA Caucasian individuals [2]. A confidence interval can be applied to this method to correct for database size and sample variation. The confidence interval equation commonly used over the last years assumes a normal distribution. The upper bound on the 95% confidence interval that can be applied to a profile’s frequency would be given by $p + 1.96\sqrt{(p)(1-p)}/N$, where p is determined from the number of observations in a database of N profiles [12, 51].

When there is no match in the haplotype database, the upper bound on the confidence interval is $1 - \alpha^{1/N}$, where α is the confidence coefficient (0.05 for a 95% confidence interval) and N is the number of individuals in the used database [12].

The interpretation of Y-STR results can be quite problematic. This is caused mainly by the patrilineal inheritance and clustering of male family members in relatively small geographic areas. This geographical clustering of male relatives coupled with the limited size of the haplotype frequency databases (many haplotypes are seen only once) makes the estimation of profile frequencies hazardous. An alternative method for assessing the significance of a match is to use a likelihood ratio and to incorporate population subdivisions with the increased potential for common co-ancestry. Regardless of the method used to calculate the matching frequency, when presenting the results of Y-STR analysis, there is a need to clearly have in mind and inform how the use of Y-STR typing varies from that of autosomal markers and that there might be other males in the population with the same Y-STR haplotype [2, 12, 52].

1.7 PowerPlex® Y23

In July 12, 2012, Promega Corporation announced the release of the PowerPlex® Y23 kit [53]. This new kit contains several new features such as amplification time roughly cut in half, more sensitivity and specificity, but what makes it potentially more useful is the number of loci that it combines, 23 (**Fig. 8**). It allows co-amplification of the loci recommended in the European Minimal Haplotype, the two additional loci recommended by SWGDAM and contains all the 17 loci present in the Yfiler™ kit as well as 6 additional loci [36, 54].

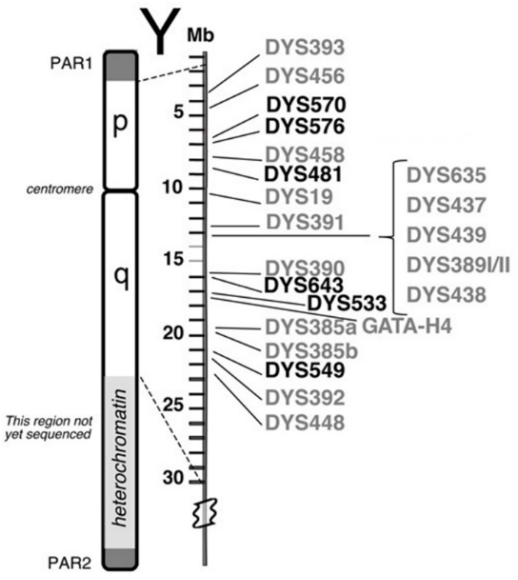

Fig. 8 – Relative positions of PowerPlex® Y23 loci. The six new loci are shown in bold font. Adapted from Butler, 2012 [55].

These six additional loci (DYS570, DYS576, DYS481, DYS643, DYS533, and DYS549) are currently well mapped and their mutation rates have already been studied [36, 56-60]. Although all the six STRs are highly informative, three of them stand out, DYS570, DYS576 and DYS481. The first two are tetranucleotide that were recently identified as rapidly mutating Y-STR loci, both with mutation rates > 1 x 10^{-2} and the latter, a trinucleotide repeat marker, also showed high values of mutation rate and variability in previous studies [60-62].

The inclusion of additional loci with high gene diversity, increases the power of the Y-STR set and reduces the chance of an adventitious haplotype match. While in overall the PowerPlex® Y23 kit holds the potential to increase discrimination in forensic work, the integration of two rapid mutating Y-STR loci may increase importantly the ability to distinguish between close male relatives. Besides, the simple structure and mutational properties of the six additional loci makes them also very useful for population studies [36, 56-58].

PowerPlex® Y23 is a 5-dye system (**Fig. 9**). Fragments included in the internal lane standard are detected in the orange channel and are labeled with CC5 (CC5 Internal Lane Standard 500 Y23).

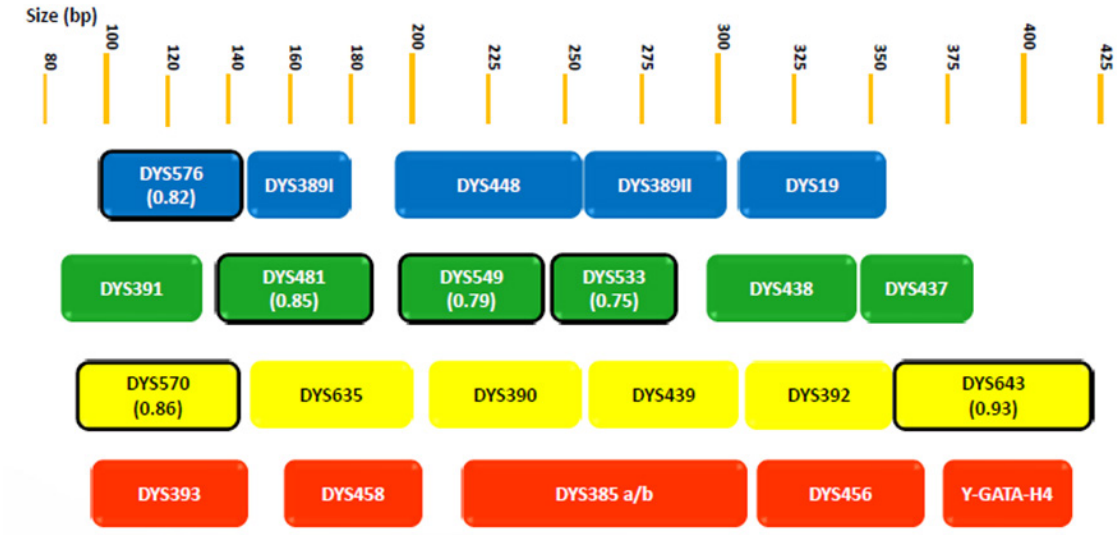

Fig. 9 – PowerPlex® Y23 loci and its relative size range and dye labels. The six new loci are bolded. Extracted from Promega Corporation, 2012 [63].

Apart from adding six more polymorphic Y-STR loci, the PowerPlex® Y23 kit includes more alleles in each allelic ladder. Across the loci in common with Yfiler™, there

are 66 additional alleles present in the PowerPlex® Y23 allelic ladders. These additional alleles can be helpful in appropriately designating rare small or large Y-STR alleles [55].

PowerPlex® Y23: Current Situation

The PowerPlex® Y23 kit is currently taking the first steps to achieve mainstream use in forensic and population genetics laboratories. Because it was recently released, the published data available about the entire set of markers included in the kit is limited, but a collaboration between YHRD and Promega is currently accelerating the accumulation of data. In fact, the YHRD database, as of November 2012, that is 4 months after the commercialization of the kit, has already available 2 144 PowerPlex® Y23 haplotypes in a total of 105 498 haplotypes [48].

Even before the release of the PowerPlex® Y23 kit, Davis and co-workers used a prototype version of the kit in a concordance study. They typed 951 unrelated males from samples of six different populations (three Native Alaskan populations and three major United States populations) with both Prototype PowerPlex® Y23 and Yfiler™ kits. The results obtained showed that for the 17 loci in common between the two kits, allele discordance was rare. Moreover, when accounting for apparent silent alleles the results were entirely concordant. They also observed a few novel duplications and silent alleles at the additional six loci of the PowerPlex® Y23 kit, namely with respect to the DYS643 locus. The study also alerted for the importance of the broader range of the allelic ladder present in the PowerPlex® Y23 kit. Of the 951 samples analyzed, an allele called outside of the bin set designated by Yfiler™ had to be manually called 20 times. All of these alleles were contained within the bin sets of the Prototype PowerPlex® Y23 kit and no alleles were observed to fall outside of the bin set provided by Prototype PowerPlex® Y23. Dissection of the prototype version further demonstrated that while in general augmenting the haplotype diversity (HD) in each population, the combined six loci approached the power of discrimination of the 17 loci in the Yfiler™ kit, thus providing increased power for discriminating male individuals [44].

So far, the Portuguese population has not yet been studied with the PowerPlex® Y23. Still, there are several haplotypes from Portuguese samples in the YHRD database. As for northern Portugal, it already contains 650 SWGDAM haplotypes, of which 85 were also typed for the Yfiler™ kit [48, 64]. Regarding central Portugal, YHRD has deposited 875 Minimal haplotypes, of which 616 are also SWGDAM haplotypes. From those, 386 were typed with the Yfiler™ kit [48, 65, 66]. From southern Portugal, 192 SWGDAM haplotypes are available in the YHRD database, of which 80 were typed with the Yfiler™

kit [48, 64]. Azores and Madeira archipelagos are also represented in the YHRD database, the former with 68 Minimal haplotypes available and the latter with 99 Minimal haplotypes [48, 65, 67]. Thus, the YHRD has 1884 Minimal haplotypes from Portuguese samples, of which 1458 are SWGDAM haplotypes. From those, 551 are Yfiler™ haplotypes.

2 AIMS

In this work, a sample of Portuguese males was characterized for the set of Y-STRs in PowerPlex® Y23 kit, aiming at achieving four main goals, which are:

- Contribute to updating the haplotype information contained in YHRD, adding data on 23 loci haplotypes for the Portuguese population, which were not yet available;

- Calculate, in the Portuguese sample, several parameters of population and forensic relevance provided by the PowerPlex® Y23 STRs;

- Evaluate the increase in effectiveness of the extended set of markers comparatively to that reached with the previously used kit, Yfiler™;

- Perform population analysis with the available data for 1) the entire set of Y-STRs and for 2) the restricted set Yfiler™, and then compare the ability of both kits to infer the pattern of male inferred affinities between human populations.

3 MATERIAL & METHODS

3.1 Sampling and DNA Extraction

A total of 250 unrelated males from the Portuguese population were analyzed. They included individuals living in Northern, Central and Southern mainland Portugal, as well as in Azores and Madeira archipelagos. For all of them, data had already been produced regarding the 17 Y-STRs included in the Yfiler™ kit and the corresponding haplotypes deposited in the YHRD database [64].

DNA had been previously extracted from blood stains or buccal swabs using a standard Chelex® 100 (Bio-Rad) procedure [68].

3.2 Genotyping

Y-STR typing was performed using the PowerPlex® Y23 System (Promega). This kit allows simultaneous analysis of 23 loci, combining the 17 loci already available in the Yfiler™ kit (DYS19, DYS385, DYS389I/II, DYS390, DYS391, DYS392, DYS393, DYS437, DYS438, DYS439, DYS448, DYS456, DYS458, DYS635, and Y-GATA-H4) plus six new Y-STR loci (DYS481, DYS533, DYS549, DYS570, DYS576 and DYS643) (**Table 2**) [36].

Marker name	Allelic range	Repeat motif	Mutation rate (95% CI)
DYS576	11-23	AAAG	1.43×10^{-2} [60]
DYS389I	9-17	(TCTG) (TCTA) (TCTG)	2.52×10^{-3} [48]
DYS448	14-24	AGAGAT	1.57×10^{-3} [48]
DYS389II	24-35	(TCTA)	3.64×10^{-3} [48]
DYS19	9-19	TAGA	2.30×10^{-3} [48]
DYS391	5-16	TCTA	2.60×10^{-3} [48]
DYS481	17-32	CTT	4.97×10^{-3} [60]
DYS549	7-17	GATA	4.55×10^{-3} [60]
DYS533	7-17	ATCT	5.01×10^{-3} [60]
DYS438	6-16	TTTTC	3.06×10^{-4} [48]
DYS437	11-18	TCTA	1.23×10^{-3} [48]
DYS570	10-25	TTTC	1.24×10^{-2} [60]
DYS635	15-28	TSTA compound	3.47×10^{-3} [48]

Marker name	Allelic range	Repeat motif	Mutation rate (95% CI)
DYS390	17-29	(TCTA) (TCTG)	2.10×10^{-3} [48]
DYS439	6-17	AGAT	5.21×10^{-3} [48]
DYS392	4-20	TAT	4.12×10^{-4} [48]
DYS643	6-17	CTTTT	1.50×10^{-3} [60]
DYS393	7-18	AGAT	1.05×10^{-3} [48]
DYS458	10-24	GAAA	6.44×10^{-3} [48]
DYS385	7-28	GAAA	2.13×10^{-3} [48]
DYS456	11-23	AGAT	4.24×10^{-3} [48]
Y-GATA-H4	8-18	TAGA	2.43×10^{-3} [48]

Table 2 – Characteristics of the 23 Y-STR loci amplified with the PowerPlex® Y23 System. These characteristics include allelic range (defined by the PowerPlex® Y23 System allelic ladder**)**, repeat motif and mutation rate [48, 60, 69].

The PCR amplifications were generally performed according to the instructions provided by the manufacturer, although adaptations were done according to the laboratory's procedures (**Table 3** and **Table 4**).

PCR Reaction	
Reagents	**Volume per reaction (µL)**
PowerPlex® Y23 5x Master Mix	2
PowerPlex® Y23 10x Primer Pair Mix	1
Water	6
DNA (±0.5 ng/µL)	1
Final Volume	10

Table 3 – PCR reaction conditions for STR amplification with PowerPlex® Y23 System.

		PCR Program	
		Temperature (ºC)	**Time**
Initial Denaturation		96 ºC	2 min
28 cycles	Denaturation	94 ºC	10 sec
	Annealing	61 ºC	1 min
	Extension	72 ºC	30 sec
Final Extension		60 ºC	20 min
Hold		4 ºC	∞

Table 4 – PCR program conditions for STR amplification with PowerPlex® Y23 System.

PCR products were separated and detected on an ABI PRISM® 3100 Genetic Analyzer, using recommended running conditions, and the results were analyzed with GeneMapper® v4.0 software. Allele designation was in accordance with the bins and panels provided by the manufacturer.

3.3 Concordance Study

When the typing results with PowerPlex® Y23 were discordant from those previously obtained with the Yfiler™ kit, the first step was to retype samples with both multiplex kits in order to confirm discrepancies. Regarding Yfiler™, the PCR amplifications were also performed according to the instructions provided by the manufacturer, although again adaptations were done according to the laboratory's procedures (**Table 5** and **Table 6**).

PCR products were separated and detected on an ABI PRISM® 3100 Genetic Analyzer, using recommended conditions, and the results were analyzed through GeneMapper® v4.0 software. Allele designation was in accordance with the bins and panels provided by the manufacturer.

PCR Reaction	
Reagents	**Volume per reaction (µL)**
Reaction Mix	3,68
Primer Mix	2
AmpliTaq Gold	0,32
Water	3
DNA (±0.5 ng/µL)	1
Final Volume	10

Table 5 - PCR reaction conditions for STR amplification with Yfiler™ kit.

		PCR Program	
		Temperature (ºC)	**Time**
Initial Denaturation		95 ºC	11 min
28 cycles	Denaturation	94 ºC	1 min
	Annealing	61 ºC	1 min
	Extension	72 ºC	1 min
Final Extension		60 ºC	80 min
Hold		4 ºC	∞

Table 6 - PCR program conditions for STR amplification with Yfiler™ kit.

Since after this procedure the observation of discrepant results in two samples was confirmed, to assess the reasons underlying the profile differences between both multiplex kits, the problematic loci were submitted to sequencing.

In one sample, the discrepancy was in the DYS385 locus. To amplify this locus we used the forward primer 5'-AGCATGGGTGACAGAGCTA-3' and the reverse primer 5'- TGGGATGCTAGGTAAAGCTG -3'. The other discrepancy was found in the GATA H4 locus. To amplify this locus, we used the forward primer 5'-GAGACCTAAGCAGAGATGTTGGTTTTC-3' and the reverse primer 5'-CCTCTGATGGTGAAGTAATGGAATTAGA-3. In both cases, the amplification was performed using the QIAGEN Multiplex PCR Kit according to the instructions provided by the manufacturer, but adapted to the laboratory's procedures (**Table 7** and **Table 8**).

PCR Reaction	
Reagents	**Volume per reaction (µL)**
QIAGEN Multiplex PCR Kit, Master Mix 2x	5
Primer Mix (2 µM)	1
Water	3
DNA (±0.5 ng/µL)	1
Final Volume	10

Table 7 – PCR reaction conditions for singleplex amplification of DYS385 and GATA H4 STR loci.

		PCR Program	
		Temperature (ºC)	**Time**
Initial Denaturation		95 ºC	15 min
30 cycles	Denaturation	94 ºC	30 sec
	Annealing	58 ºC	90 sec
	Extension	72 ºC	1 min
Final Extension		60 ºC	45 min
Hold		4 ºC	∞

Table 8 – PCR program conditions for singleplex amplification of DYS385 and GATA H4 STR loci.

The quality of the PCR products was checked after polyacrylamide gel electrophoresis (T 9%, C 5%), using a discontinuous buffer system. The DNA fragments were visualized by the silver staining method [70, 71].

In the case of the DYS385, the PCR fragments had to be separated prior to sequencing. After gel electrophoresis (as previously described), the fragments were eluted from the gel with TE buffer by freezing for 30 minutes and heating at 60 ˚C for 15 minutes, 3 times. The extracted alleles were then re-amplified using the same conditions as described before [71].

As a preparatory step for the sequencing reaction, the amplified products were submitted to an enzymatic purification performed with the addition of 2µl of ExoAp (Exonuclease I + FastAP Thermosensitive Alkaline Phosphatase) to 5µl of the amplified product. This mix was then submitted to 37ºC during 15 minutes followed by 85ºC during 15 minutes, in a thermocycler.

The sequencing reaction was performed using the BigDye® Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems) generally according to instructions of the manufacturer, with some adaptations according to the laboratory's procedure (**Table 9** and **Table 10**).

Sequencing Reaction	
Reagents	**Volume per reaction (µL)**
BigDye Terminator Ready Reaction Mix (2.5x)	1
Big Dye Sequencing Buffer (1x)	1
Primer (5 µM)	0,5
Water	1
PCR product	1,5
Final Volume	5

Table 9 – PCR conditions for sequencing.

		PCR Program	
		Temperature (ºC)	**Time**
Initial Denaturation		96 ºC	2 min
35 cycles	Denaturation	96 ºC	15 sec
	Annealing	50 ºC	9 sec
	Extension	60 ºC	2 min
Final Extension		60 ºC	10 min
Hold		4 ºC	∞

Table 10 – PCR program conditions for sequencing.

The purification of the sequencing products was performed using Sephadex™ G50 (GE Healthcare) columns [72].

The products were then run on the ABI PRISM® 3100 Genetic Analyzer and analyzed through Sequence Scanner v1.0 software.

3.4 Data Analysis

Estimates of haplotype diversity (HD) were obtained with the HapYDive software (http://www.ipatimup.pt/app/; [73]). The same software was used to evaluate which are the best loci out of those in a battery of Y-STRs for HD increment. In this study it was assessed the increase in efficiency provided by each of the new loci contained in PowerPlex® Y23 comparatively to that offered by the combined 17 Y-STRs in Yfiler™.

Gene diversity was calculated using Arlequin v3.5.1.2 Software [74].

Population comparisons were performed through F_{ST}, which were computed also using Arlequin v3.5.1.2. A graphical representation of the resulting F_{ST} values was performed with Arlequin v3.5.1.2 and R v3.0.1 software. The data used in the comparative analyses was from 86 worldwide populations, having been provided by the laboratories that participated in the YHRD/Promega collaboration. All the information regarding the populations can be consulted in **Supplementary Table 1**. A multidimensional scaling of the F_{ST} was performed using IBM SPSS Statistics 21 software [75].

The Haplogroup Predictor web based software [76] was used to evaluate which set of STR markers is more efficient to infer haplogroup allocation. For that, the Yfiler™ and PowePlex® Y23 profiles of those samples with haplogroups already identified through Y-SNP analysis (N=91) were inserted in the Haplogroup Predictor and the accuracy of haplogroup assignment given by both sets of markers was compared.

4 RESULTS & DISCUSSION

The multiplex amplifications with the PowerPlex® Y23 kit were successfully performed in all samples leading to obtain a total of 250 haplotypes defined by the 23 Y-STRs. **Fig. 10** illustrates one of the obtained genetic profiles, while the entire set of raw haplotypic data can be consulted in **Supplementary Table 2**.

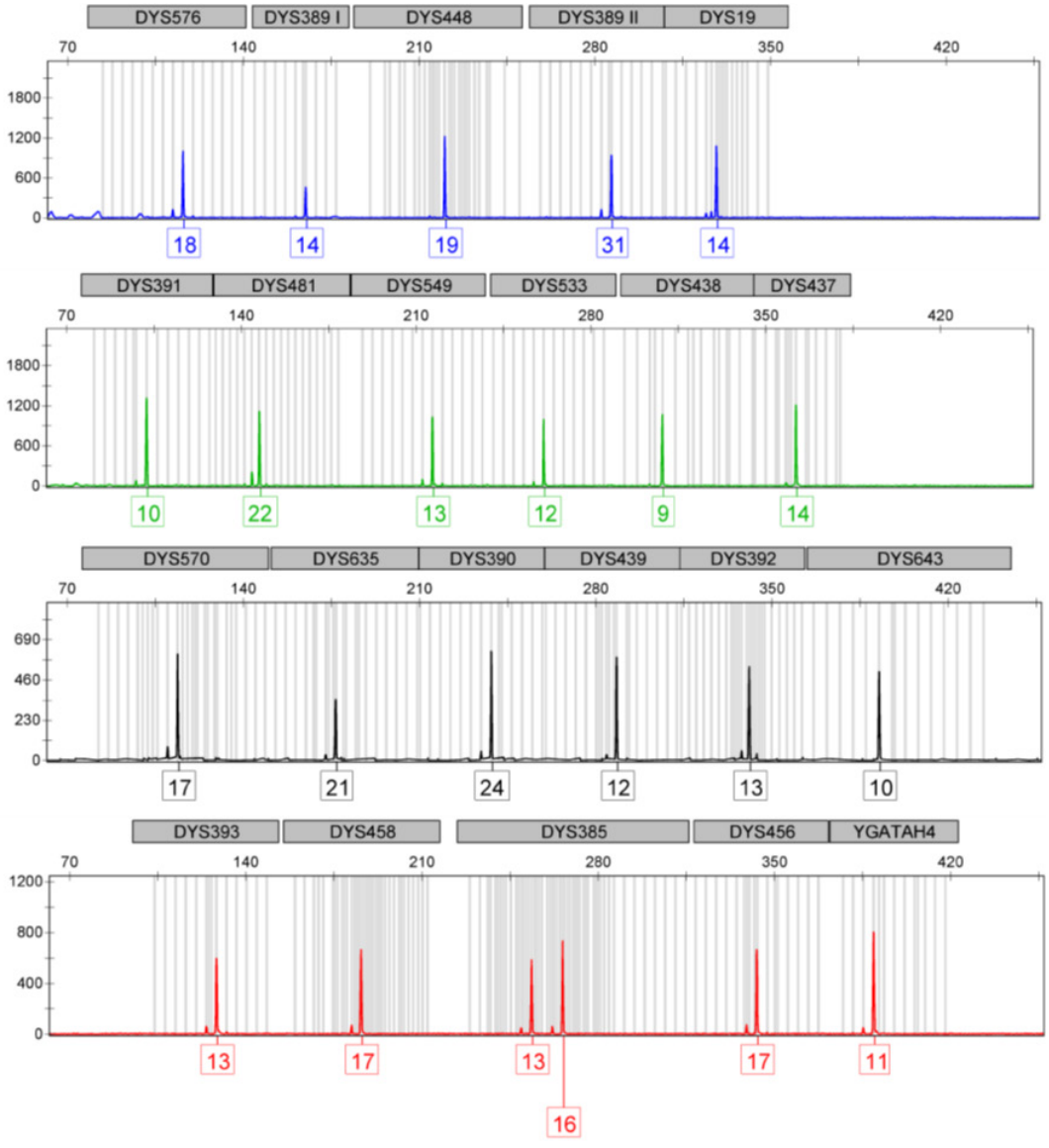

Fig. 10 – Example of an electropherogram obtained with the GeneMapper® v4.0 software.

4.1 Discrepancies with Yfiler™ Results

Since the samples used on this project had already been typed with the Yfiler™ kit, the genetic profiles derived from both kits were compared. Differences were only detected in two samples involving in each a unique, but distinct, Y-STR. Differences were confirmed after further amplifications with both kits.

GATA H4

One of the two discrepant cases involved the GATA H4 locus. When amplified with PowerPlex® Y23 kit, no peak appeared in the assumed range of the GATA H4 marker whereas an extra peak appeared in the range of its closest marker, DYS456, denoting the possibility that the peak actually belonged to the GATA H4 locus if the amplified region encompassing this locus contained a deletion. When amplified with Yfiler™ kit, the same sample was typed as 11 for the GATA H4 locus (**Fig. 11**).

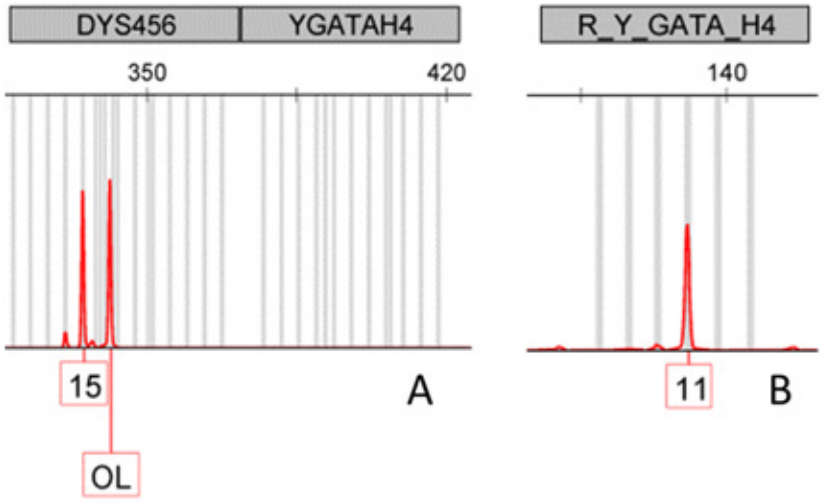

Fig. 11 – Electropherograms. Same sample amplified with PowerPlex® Y23 System (A) and with Yfiler™ kit (B).

The results obtained by sequencing the amplicon for GATA H4 locus in the sample (**Fig. 12**) confirmed the hypothesis. A 48 bp deletion was detected in a region located outside the DNA stretch amplified with the Yfiler™ primers but inside the range amplified with the PowerPlex® Y23 primers, explaining therefore the discordant results obtained with both kits.

gagacctaagcagagatgttggttttctgatacacattgatactttcagcacatcac
ttgtatcctaggaatcatcattaaaatgttatgctgaggagaatttccaaattta[a
gatagatagatagatctatagatagataggtaggtaggtagatagatagatagatag
atagatagatagatagatagatagat]agaatggatagattagatggatga[ataga
tagatagatagatacatagatag]gtgatttatcctgttaagttgtttaacaagtgg
gctatgtaaaatttactaatatttaaacataagtagtttgtagattttcttattta
ttctaattccattacttcaccatcagagg

Fig. 12 – Amplicon with 368 bp (deletion not accounted) for allele 11 of GATA H4. The primers used are in red and underlined, the repeat region of the GATA H4.1 locus is in green and bold, the repeat region of the GATA H4.2 locus is in blue and bold (it is partially highlighted in red because of the deletion), and the deleted area (48 bp) is highlighted in red.

DYS385

The other discrepancy was demonstrated to be a classic case of a silent allele. When amplified with Yfiler™ kit, a sample was typed as 12-15 for the DYS385 locus, whilst with PowerPlex® Y23 kit it was apparently 12-12. However, when re-amplified using an annealing temperature of 54˚C, a 12-15 profile, although quite unbalanced, was obtained as is shown in **Fig. 13**. This strongly indicated that a silent allele could be responsible for the non-concordant genetic profile.

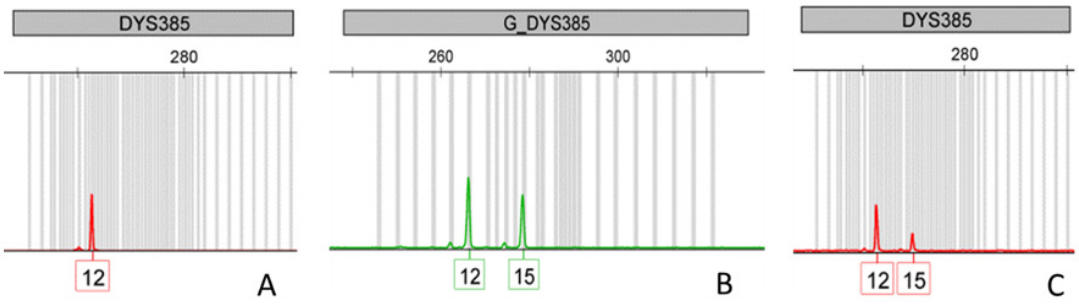

Fig. 13 – Electropherograms. Same sample amplified with PowerPlex® Y23 System (A), amplified with the Yfiler™ kit (B), and then amplified with the PowerPlex® Y23 System with an annealing temperature of 54˚C (C).

The sequencing results confirmed that in fact a silent allele was involved, explaining the apparent dissimilar results. Although the primers sequences included in the PowerPlex® Y23 kit are unknown, a mutation was identified in a region very likely located in the binding site of the forward primer for DYS385 used in that kit. As shown in

Fig. 14, a thymine was found instead of an adenine and this is probably the cause of the non-amplification of the allele 15 in this sample with the PowerPlex® Y23 kit.

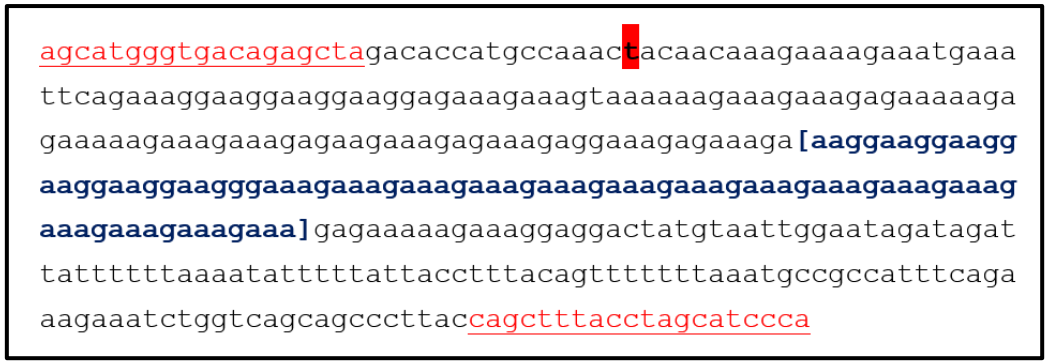

Fig. 14 – Amplicon with 385 bp for allele 15 of DYS385. The primers are in red and underlined, the repeat region is in blue and bold and the transversion from an adenine to a thymine is highlighted in red.

In all procedures, sequencing included, a positive control was used, which further pinpointed an unexpected result in the sequencing data. Sequencing of the positive control (2800M Control DNA-Promega) revealed the presence of a transversion in the repeat region. **Fig. 15** shows the presence of a thymine instead of a guanine in one of the repetitive motifs. Although the detected variation does not influence the Y-STR typing results, it illustrates that sequence variations might occur within the repeat region of an STR, as mentioned before in point 1.4 of the Introduction.

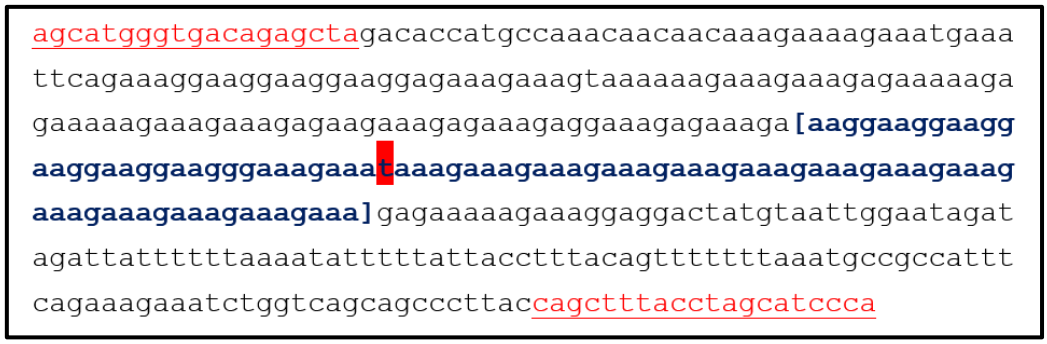

Fig. 15 – Amplicon with 389 bp for allele 16 of DYS385 obtained from 2800M Control DNA-Promega. The primers are in red and underlined, the repeat region is in blue and bold and the transversion from a guanine to a thymine is highlighted in red.

Returning to the non-concordant results, none of the causes here demonstrated has yet been described.

To date, only two published studies have reported sequence variants with PowerPlex® Y23 System.

Coble and co-workers were able to identify several silent alleles using this kit. They detected silent alleles in the DYS389I, DYS389II, DYS439 and DYS448, but in their case the results were concordant with the Yfiler™ kit, with which no amplification was also achieved on the same loci [77].

Davis and co-workers observed silent alleles in the DYS643 locus (exclusive to the PowerPlex® Y23 System) and in the DYS448 locus, the latter being concordant with the results obtained with the Yfiler™ kit. Davis *et al.* also pointed out to a case in the DYS448 locus, identified in a Hispanic individual, very similar to the one described in this work for the GATA H4 locus. With Yfiler™, the DYS448 locus with the deletion was observed in the electropherogram region occupied by the DYS437 locus, whereas here, with PowerPlex® Y23 System, it was observed in the DYS576 locus range [44]. These differences occur due to the different configurations of the markers in each multiplex kit, since the order of markers in the electropherograms depends on the kit used.

4.2 Forensic Data

In the sample of 250 Portuguese males a total of 236 different haplotypes were found, among which 14 haplotypes were shared each of them by two individuals. Since this sample had been previously typed with the Yfiler™ kit, a comparison was also undertaken in terms of HD. The overall HD of the set of STRs in the PowerPlex® Y23 reached 0.9996, being higher than the value obtained with the markers of Yfiler™ kit (0.9993).

By fixing the haplotypes generated with the Yfiler™ loci and adding the six new PowerPlex® Y23 loci one by one, we assessed the influence of each of the six new loci in HD increment. In our sample, the markers that contribute more to increase HD are DYS576 and DYS481. The former is considered a rapid mutating locus (1.43×10^{-2}) and the latter is a tri-nucleotide repeat motif locus with a high mutation rate (4.97×10^{-3}) [60]. Each of these two loci increased HD from 0.9993 to 0.9994, and the number of unique haplotypes from 230 to 232. On the other hand, one of the additional markers, DYS643, did not contribute in our sample to any augment in discrimination capacity since the same HD was obtained with or without it. This is not surprising since this locus has a relatively low mutation rate (1.50×10^{-3}).

Focusing individually on each of the loci integrating the set of 23 Y-STRs (**Table 11**), the locus revealing the highest gene diversity in our Portuguese sample was

DYS570 followed by DYS576, which are two of the six new loci introduced in the PowerPlex® Y23 system and both are considered rapidly mutating loci [60]. DYS576 was one of the new loci contributing more to increase HD after fixing the battery of Yfiler™ loci.

In the lower range of levels of diversity, DYS393 was the locus characterized by the lowest value. But amongst the 6 supplementary loci contained in PowerPlex® Y23, the less diverse was DYS533, followed by DYS549, this being one of the new loci not accounting to increment haplotype diversity.

Locus	Gene Diversity
DYS576	0.77934
DYS389I	0.50599
DYS448	0.70304
DYS389II	0.52909
DYS19	0.60671
DYS391	0.56752
DYS481	0.72845
DYS549	0.67332
DYS533	0.58625
DYS438	0.66101
DYS437	0.57745
DYS570	0.78490
DYS635	0.60871
DYS390	0.66972
DYS439	0.67955
DYS392	0.58390
DYS643	0.67997
DYS393	0.48389
DYS458	0.76520
DYS385	0.86294
DYS456	0.69375
GATA H4	0.58034
Mean	0.64939 ± 0.08060

Table 11 – Gene diversity locus by locus.

Despite the fact that the two loci with the highest value of gene diversity (excluding DYS385, which is a duplicated locus) also have the highest values of mutation rate, the relation between both parameters is far from being linear as is shown in **Fig. 16**.

What is also noticeable in **Fig. 16** is that 5 of the 6 new loci of the PowerPlex® Y23 kit have gene diversity levels higher than the average number for this set of 23 Y-STRs. Only locus DYS533 has a gene diversity level below the average, which makes this locus the logical choice if we had to eliminate one of the new loci from the set.

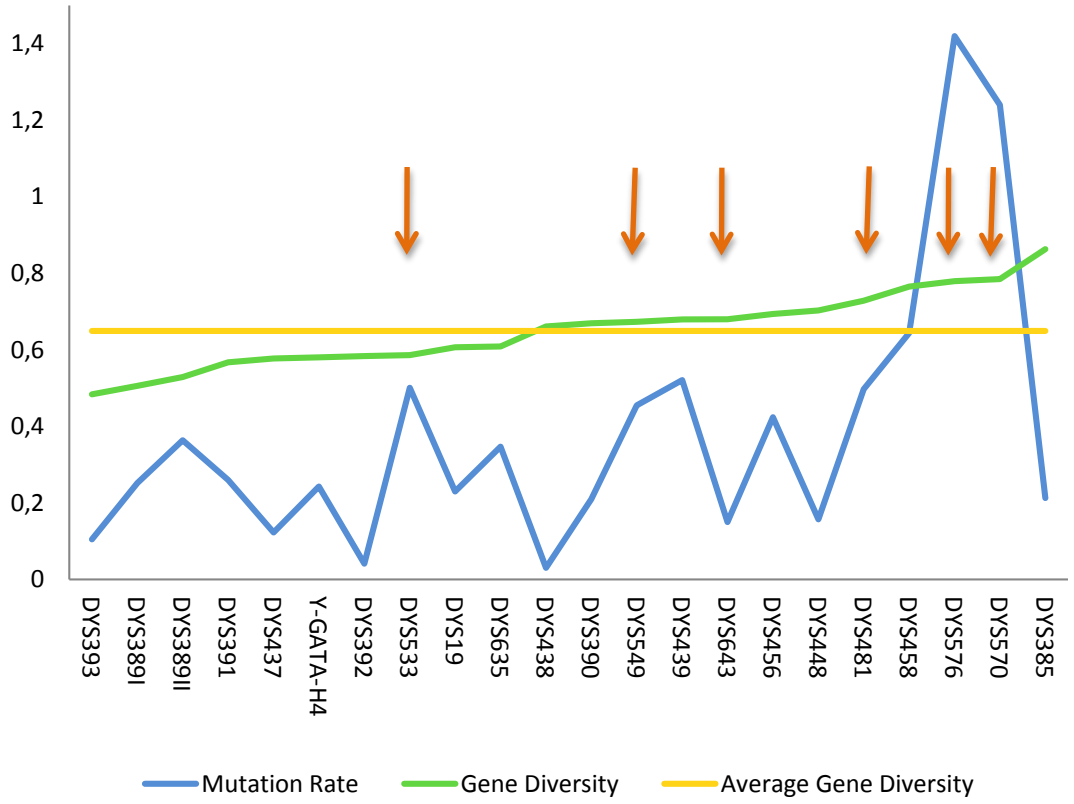

Fig. 16 – Genetic diversity vs. Mutation rate locus by locus. The mutation rate values were multiplied by 100 to be in the same order of magnitude of the genetic diversity values and therefore be easily compared in the same plot. The average gene diversity value is 0.64939. The red arrows mark the 6 new loci of the PowerPlex® Y23.

4.3 Haplogroup Predictor

Several approaches have been developed for predicting the Y-chromosome haplogroup from a set of Y-STR markers, among which is that implemented in Haplogroup Predictor.

To compare the prediction accuracy between haplotypes defined by the PowerPlex® Y23 STRs and those defined by Yfiler™ markers, we selected 91 of the 250 samples used in this study for which we already knew their actual haplogroups from previous studies with Y-SNPs [78], and then we tested their less (the 17 STRs in Yfiler™) and more (17 plus 5 new STRs in PowerPlex® Y23) extended haplotypic profiles in Haplogroup Predictor. The DYS549 marker data was not used because it is not supported by the Haplogroup Predictor program.

The results obtained (**Supplementary Table 3**) showed that for 80 samples, the predictions with haplotypic data from the two kits were concordant, and both fitted the haplogroups previously determined with Y-SNPs; moreover, in general the predictions were associated to high probability values.

In 6 cases, the predictions obtained with Yfiler™-haplotypes were correct whereas they turned out wrong with the extended haplotypes provided by PowerPlex® Y23. Of note, however, that all these discrepancies involved correct haplogroups (deduced by Y-SNP typing and well predicted with Yfiler™-haplotypes) that are low-frequent or moderately represented in Western European populations (J2, T1a, N1). Plus, 3 of these 6 cases were misclassifications as G2 of lineages known to be J2. Although haplogroups G2 and J2 are not phylogenetically very distant from each other, the level of relatedness seems to be distinguishable enough by the proper set of STRs , as in fact Yfiler™ did [20].

In 5 cases the results were wrong for both multiplex kits, but 4 of them can be explained by the fact that the Haplogroup Predictor doesn't consider haplogroups with scarce phylogenetic resolution, as was the case of types K and P, in which the referred 4 samples were classified based on SNP analysis (**Fig. 17**).

No instance occurred of wrong haplogroup prediction with Yfiler™-haplotypes becoming correctly ascertained after considering the extended PowerPlex® Y23.

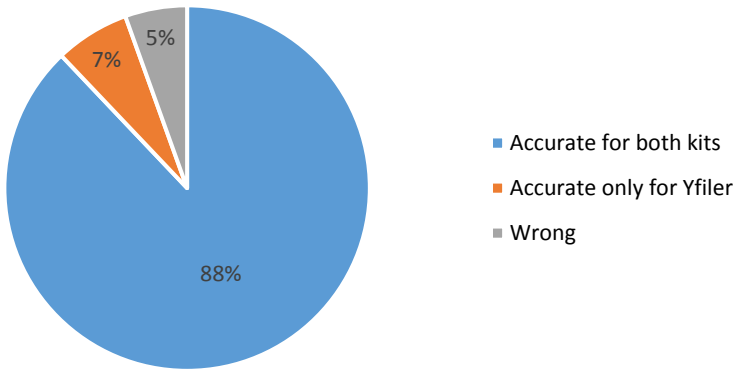

Fig. 17 – Graphic distribution for haplogroup prediction results obtained with Yfiler™ and PowerPlex® Y23 kits.

As a whole, these findings indicate that currently the set of Yfiler™ markers is better alone to infer Y-chromosome haplogroups than complemented with the additional markers contained in PowerPlex® Y23. This may be likely due to the still very incipient coverage of data for the new markers available on the Haplogroup Predictor database. The lack of enough data can introduce strong bias in the predictions, especially for haplogroups that are relatively uncommon in populations, for which the sample sizes needed to obtain a reasonable picture of their STR profiles were probably not yet achieved for the additional markers in PowerPlex® Y23, since only now they are beginning to be more widely studied.

The present-day situation, that is the ability of predicting more accurately a haplogorup using Yfiler™-haplotypes, which provides a shorter Y-haplotype compared to PowerPlex® Y23, expectedly will tend to change in the near future as long as the number of samples available in databases increase for the less studied Y-STRs in PowerPlex® Y23.

4.4 Population Comparisons

The population comparisons were performed using haplotypic data from 86 populations dispersed across the world (**Fig. 18**). To avoid misclassification of populations according to geographical ancestry, we have considered as “admixed” all American populations that are not native from America plus the African British and Asian British populations. More information on all populations such as the number of individuals or the groups they were integrated for the analyses performed, can be consulted in **Supplementary Table 1**.

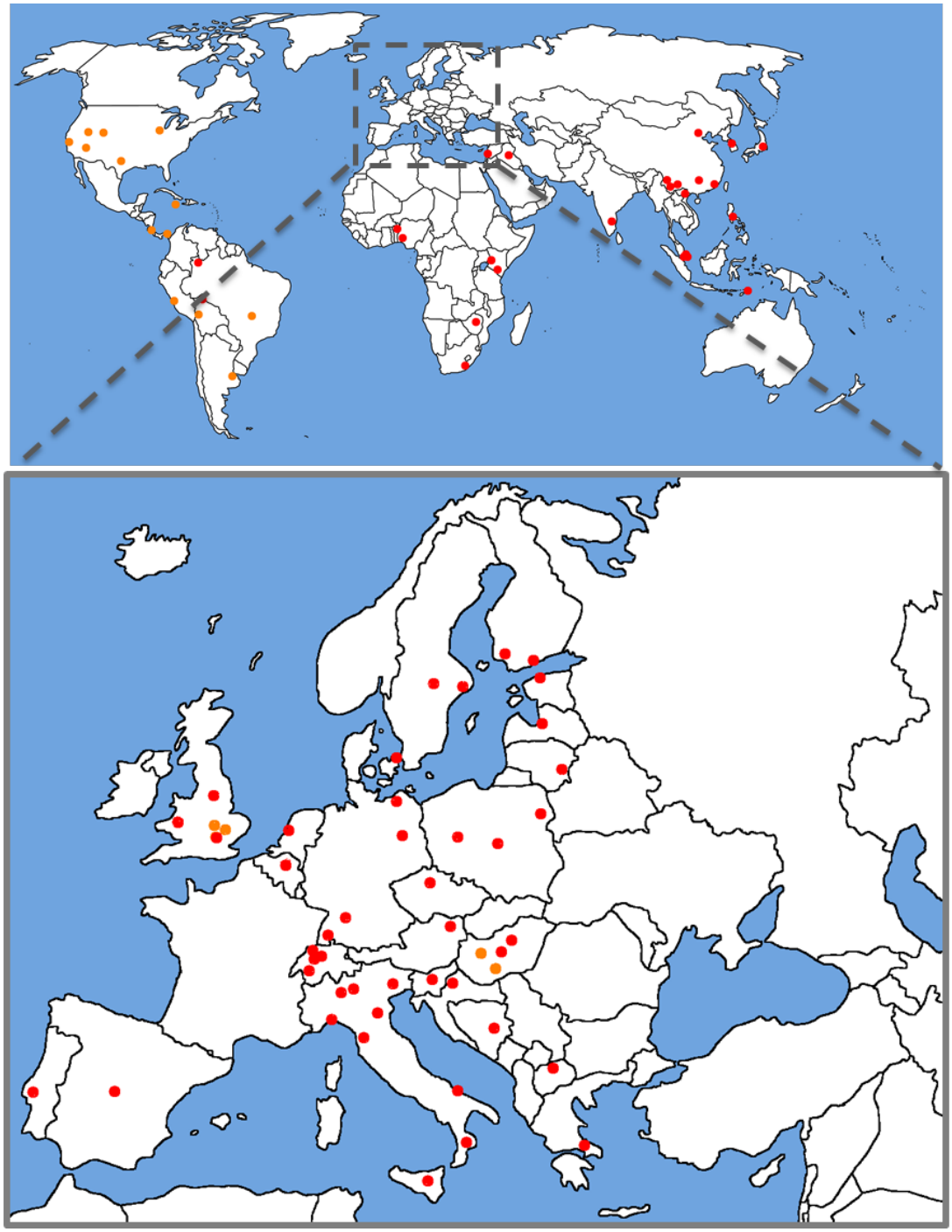

Fig. 18 – All 86 worldwide populations used and its geographical location. Each red dot represents a population. Each orange dot represents an admixed population or, in the case of the dots in Hungary, Romani populations.

F_{ST} distances were chosen to be herein presented over R_{ST} values due to the higher loss of information that R_{ST} implicates. In fact, R_{ST} analysis was performed, which determines that the data of DYS385 locus had to be eliminated, as had all data respecting to non-consensual variant alleles. Besides, no major difference was observed

in the overall pattern of relationships between populations using R_{ST} based analyses comparatively to those conducted with F_{ST} (results not shown).

The results obtained with the computations of the F_{ST} pairwise differences between 86 worldwide populations (typed with PowerPlex® Y23), including our population, are graphically represented in **Fig. 19**.

All F_{ST} values obtained for every analysis and its respective p-value can be consulted at https://www.dropbox.com/s/btkak5xrtzg95kp/Tables%20FST%20distances.xlsx .

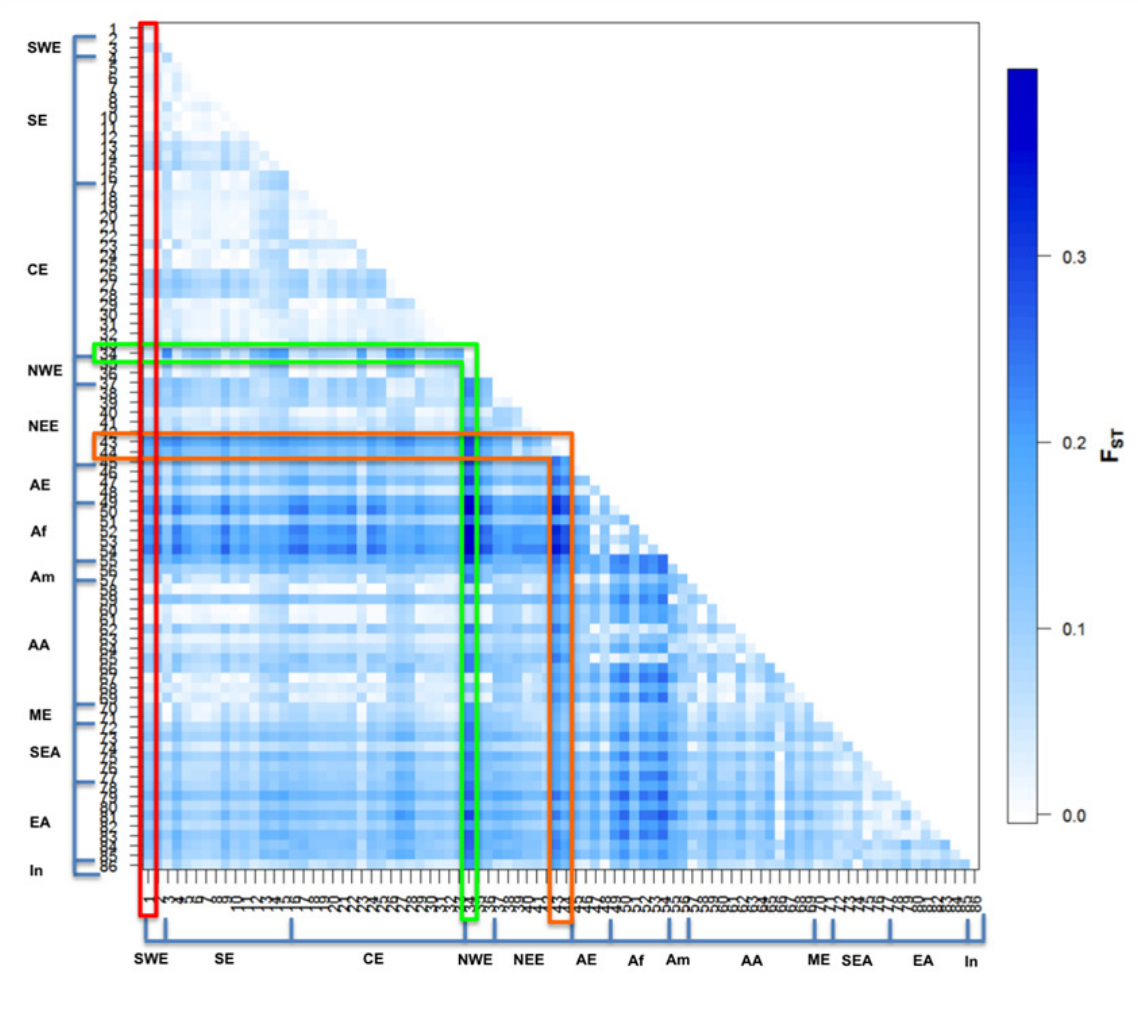

Fig. 19 – Graphical representation of the pairwise F_{ST} distance between 86 Worldwide populations. F_{ST} values are based on the 23 Y-STR haplotypes present in PowerPlex® Y23**.** Each number corresponds to a population. All populations, its respective numbers and groups can be consulted in **Supplementary Table 1**. SWE: Southwestern Europe; SE: Southern Europe; CE: Central Europe; NEW: Northwestern Europe; NEE: Northeastern Europe; AE: Admixed European; Af: African; Am: Native American; AA: Admixed American; ME: Middle East; SEA: Southeastern Asia; EA: Eastern Asia; In: India. The Portuguese population is highlighted in red, the Welsh in green and the Finnish in orange.

Based on haplotypes defined by the 23 Y-STRs the F_{ST} distances between worldwide populations immediately afford a broad panorama on the global structure of human genetic diversity.

Among European populations, levels of genetic differentiation were relatively low. Accordingly, many of the F_{ST} distances were statistically non-significant. The African populations, as expected, immediately stand out from the rest, since in general the biggest F_{ST} values were found in pairwise comparisons involving an African population. The Asian populations are notoriously distant from the others, but F_{ST} values are clearly lower than those obtained for the African populations. The Native American populations are distant from the others, presenting F_{ST} values in the same order as the Asian populations. Excluding the comparisons within Europe and involving admixed populations, almost every F_{ST} value obtained was statistically significant, even after applying the Bonferroni correction for multiple tests.

Average F_{ST} values between all the continents were calculated. The average F_{ST} between European and African populations was 0.16680; between European and Asian populations, it was 0.09916; between the European and the Native American populations, average F_{ST} was 0.11490; the medium F_{ST} value between African and Asian populations was 0.14973; between African and Native American populations was 0.19977; finally, the average F_{ST} value between Asian and the Native American populations was 0.11997. All of these distances were significant. The highest distances were observed in the comparisons involving African populations, which was not surprising. The lowest inter-continental differentiation was registered between European and Asian populations.

Unexpectedly, however, two European populations stand out: the Welsh and Finnish populations, which generally present very high levels of F_{ST} values not only in the worldwide context, but even, and more remarkably, in the panorama of other European populations, especially the Finnish show a clearly unusual high differentiation. Although surprising, given the levels of differentiation captured by the set of PowerPlex® Y23 STRs, results pointing towards a strong differentiation of the Finnish population have already been published, both based on SNPs and Y-STRs [79-81]. This can be explained by the demographic history of the Finns, which was strongly shaped by several factors: a limited number of founders; isolation, due to geographic and language barriers; and numerous population bottlenecks, that despite the involvement of Finland in several wars, were probably mainly caused by epidemics and famines [81-83]. Regarding the Welsh population, there is also some published data supporting our results. Weale *et al.*, using Y-STRs and Y-SNPs, provided evidence indicating the presence of a strong

genetic barrier between Central England and Northern Wales [84]. Soon after, however, the study of Capelli *et al.* suggested that the transition between Wales and England was somewhat gradual, not sustaining the findings of Weale *et al.* [85]. Some historical facts can explain the differentiation of the Welsh population: the Anglo-Saxon settlements and culture appeared throughout England but did not extend to North Wales; continued conflicts between the Welsh and the Anglo-Saxon Kingdoms and the construction of Offa's Dyke (an earthwork barrier 240 km long between Wales and England); the linguistic, cultural and political separation of Wales and England lasted at least until 1282 [84].

Even so, the strong differentiation revealed by the Welsh and Finnish samples must be interpreted with caution, since the lack of knowledge about the criteria used for obtaining both samples hinders for the moment to explore adequately the level of differentiation they show.

A multidimensional scaling (MDS) performed with the F_{ST} distances clearly revealed the outlier position of the Welsh and Finnish, even in the context of worldwide populations. For this reason the two samples were excluded from the MDS analysis whose plot is presented in **Fig. 20**.

The figure illustrates the sharp clustering of three major groups of populations: European, African and Eastern Asian. The Western Asian populations stand closer to the European than to the Eastern Asian. The two Native American populations are distant from the remaining populations, being also relatively far from each other. The sample from Bolivia (population 59), classified in the group of "admixed" populations due to its acknowledged mixed ancestry, is relatively close to both Native American samples. Other admixed populations occupy disperse positions in the plot reflecting the distinct contributions to their origins.

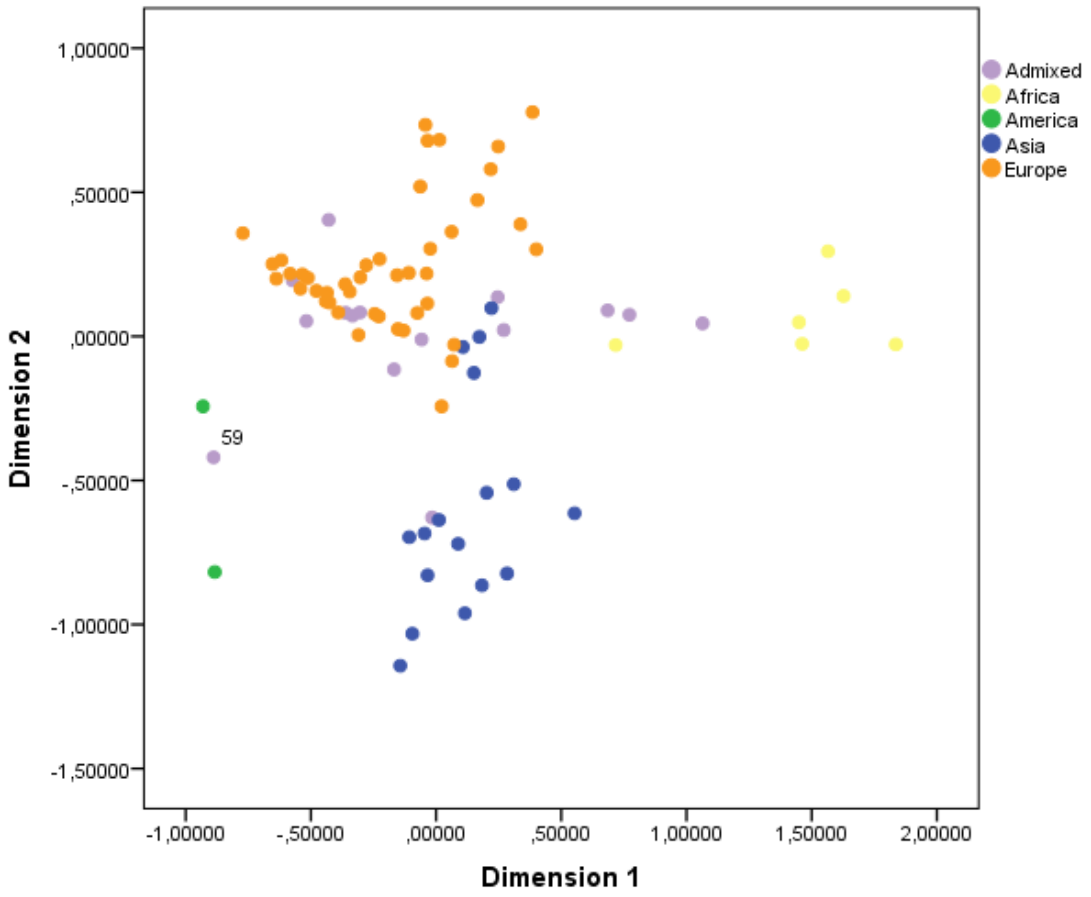

Fig. 20 – MDS plot based on F_{ST} values for 23 Y-STR based haplotypes showing relationships among 86 Worldwide populations. Finnish and Welsh populations were not included in this analysis.

Focusing now restrictively on the pairwise F_{ST} distances between European populations, the obtained values are graphically represented in **Fig. 21**.

Again, the Finnish and the Welsh populations showed the highest pairwise F_{ST} values (almost all of them statistically significant), with the Finnish being seemingly more differentiated than the Welsh.

Apart from these two samples, Northeastern populations compared to others reveal in general considerably high F_{ST} distances (almost all of them significant), as do some populations from Central and Southern Europe like those from Poland, Croatia, Macedonia and Bosnia and Herzegovina. The Roma from Hungary also show to be highly differentiated, which otherwise fits well with what is being demonstrated about the genetic characteristics of Roma from Europe [86-88].

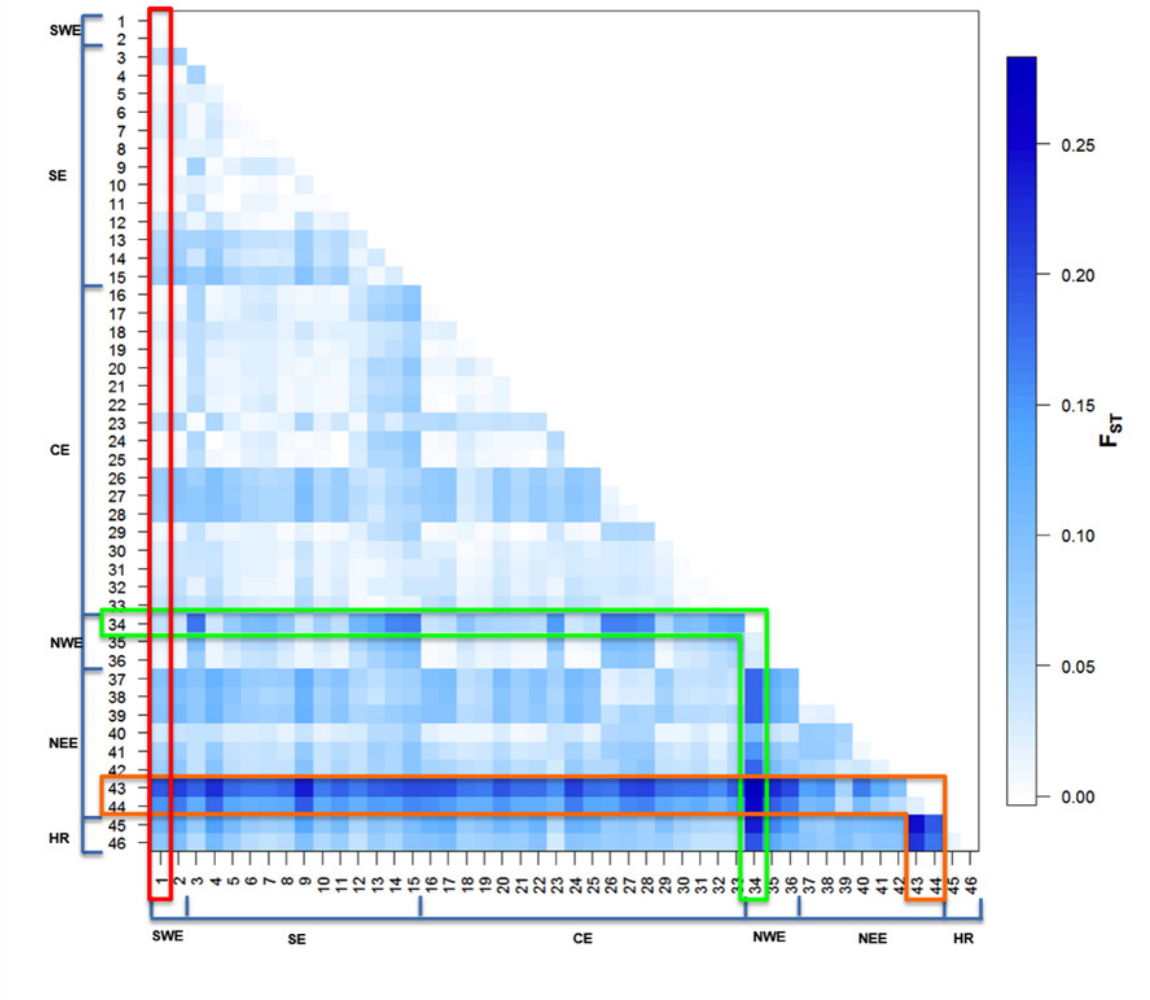

Fig. 21 – Graphical representation of the pairwise F_{ST} distance between 46 European populations. F_{ST} values are based on the 23 Y-STR haplotypes present in PowerPlex® Y23. Each number corresponds to a population. All populations, its respective numbers and groups can be consulted in **Supplementary Table 1**. SWE: Southwestern Europe; SE: Southern Europe; CE: Central Europe; NWE: Northwestern Europe; NEE: Northeastern Europe; HR: Hungary Roma. The Portuguese population is highlighted in red, the Welsh in green and the Finnish in orange.

As for the Portuguese population, it presents pairwise F_{ST} values in the lower range found in Europe, except when compared with the Northeastern and Hungarian Romani populations as well as with some Southern and Central European populations.

Regarding the significance of the F_{ST} values, populations geographically close to each other tend to have non-significant F_{ST} between them, although a few exceptions occur. Populations distant from each other tend to have higher and significant F_{ST} between them. For example, Portugal is not significantly distant neither from Spain nor around half of the Southern and Central European populations.

A multidimensional scaling performed with pairwise F_{ST} between European populations (Finnish and Welsh again excluded), presented in **Fig. 22**, indicates that for the pattern of population substructuring in Europe geography accounts considerably,

although not being rare the observation of inconsistencies, in the relationship between geographical and genetic distance. The dimension 2 of the MDS plot discriminates populations mainly according to latitude: populations with the lowest values for dimension 2 are the Northern, followed by the Central and then the Southern ones. Although not capturing such a clear trend as dimension 2, dimension 1, reflects some relationship with the longitude. The major exceptions to this geographic pattern of population clustering are populations from countries like Bosnia and Herzegovina (population 15) and Croatia (population 13). Those from Calabria (population 3) and Macedonia (population 14) are also not clearly clustered with the remaining Southern European populations, while the populations from Poland (populations 26, 27 and 28) and Basel (population 23) are as well rather distant from the other Central European populations. The Portuguese population is quite well integrated in the group of Southern populations that is positioned quite near most of the populations from Central Europe.

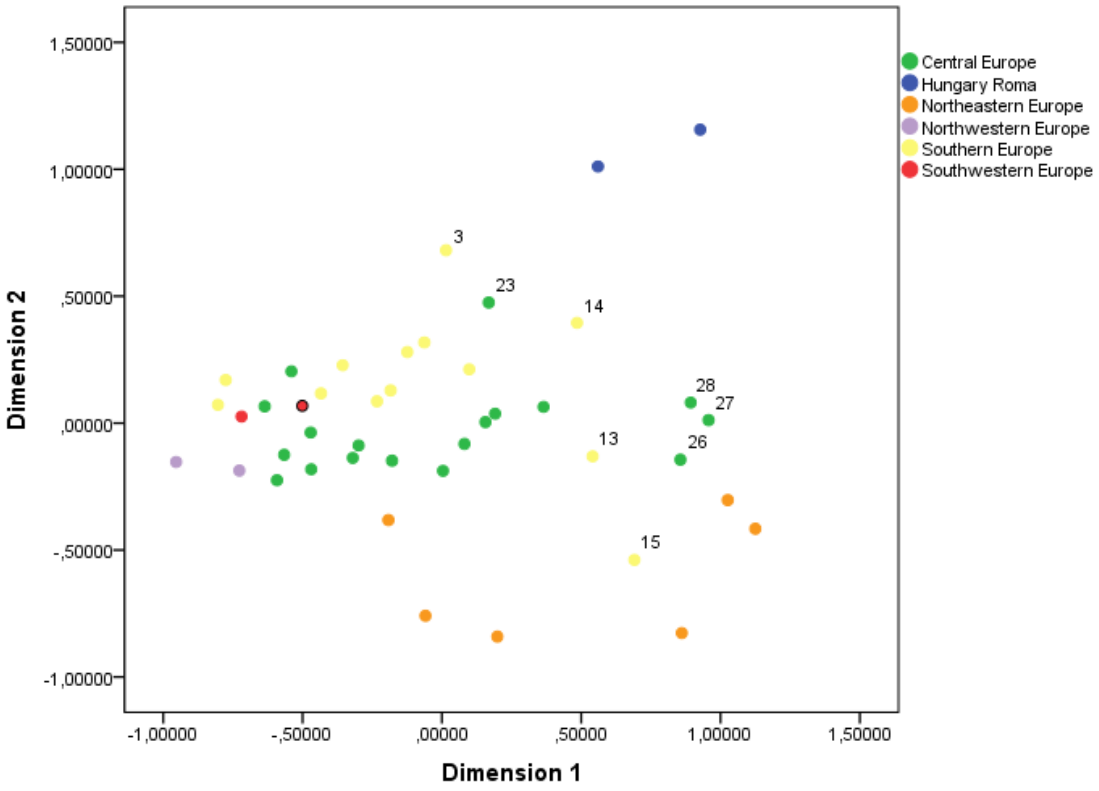

Fig. 22 – MDS plot based on F_{ST} values for 23 Y-STR based haplotypes showing relationships among 46 European populations. The Portuguese population is represented in red with a black border. Finnish and Welsh populations were not included in this analysis. Each number corresponds to a population and they all can be consulted in **Supplementary Table 1**.

To assess better at which extent geography influences Y-chromosome diversity in Europe, we have analyzed the correlation between geographic and genetic distances between the European populations performing a linear regression analysis. The scatter plot with the geographic distances and the F_{ST} values between the European populations is depicted in **Fig. 23**, and the significant R^2 (p-value equals to 0.000) obtained with linear regression indicates that 27.86% of the genetic distance variation between European populations can be explained by differences in geographic distance between them.

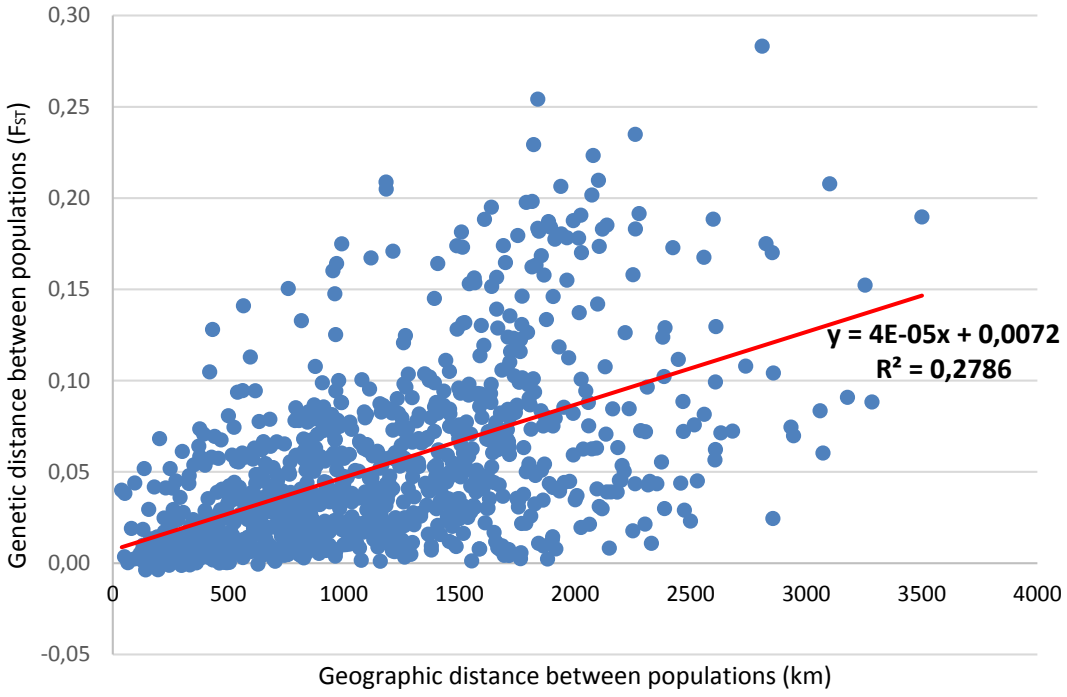

Fig. 23 – Scatter plot and linear regression of the genetic (F_{ST} values based on PowerPlex® Y23 haplotypes) and geographic distances of the European populations. P-value equals to 0.000.

PowerPlex® Y23 vs. Yfiler™

In order to compare the differences between both Powerplex® Y23 and Yfiler™ multiplexes in this kind of population analyses, the pairwise F_{ST} distances for the 86 populations were computed using only the data from the loci present in Yfiler™ kit.

The results obtained for the F_{ST} analyses are graphically represented in **Fig. 24**.

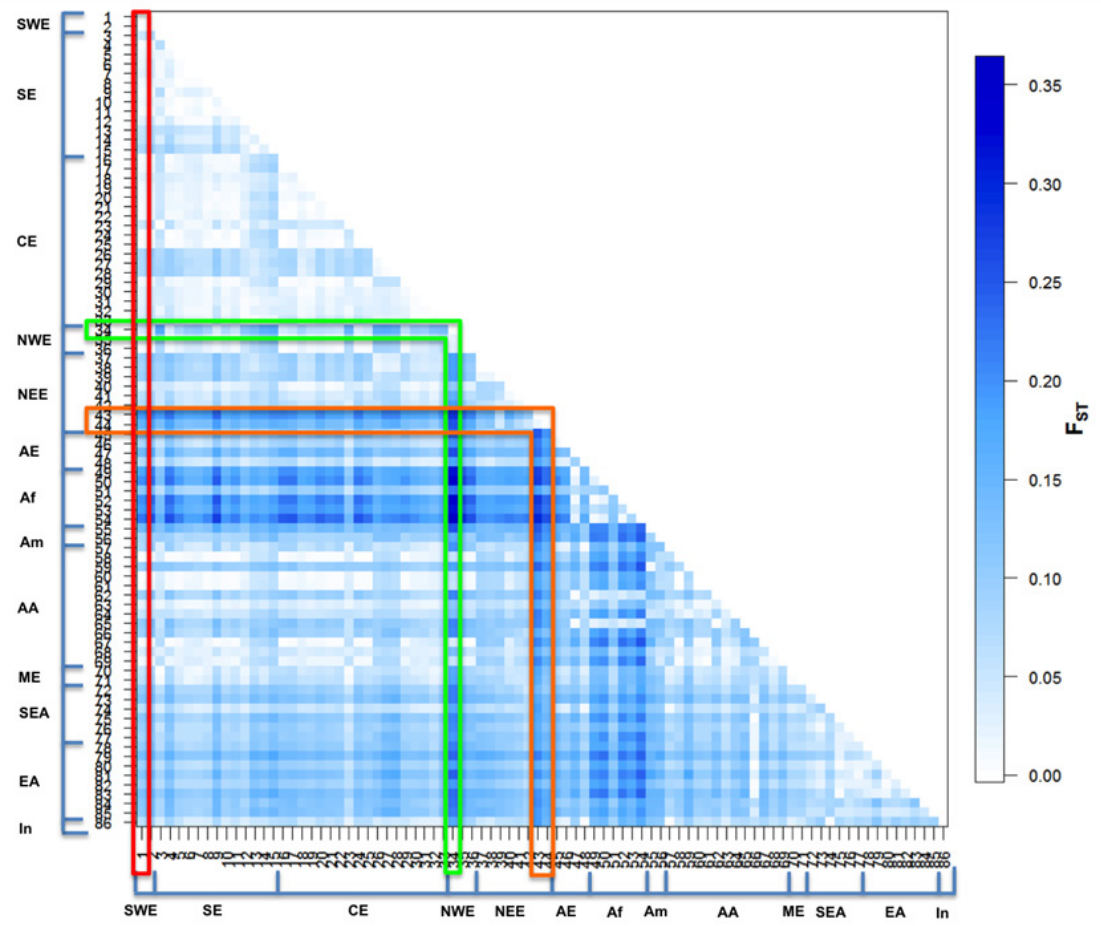

Fig. 24 – Graphical representation of the pairwise F_{ST} distance between 86 Worldwide populations. F_{ST} values are based on the 17 Y-STR haplotypes present in Yfiler™. Each number corresponds to a population. All populations, its respective numbers and groups can be consulted in **Supplementary Table 1**. SWE: Southwestern Europe; SE: Southern Europe; CE: Central Europe; NEW: Northwestern Europe; NEE: Northeastern Europe; AE: Admixed European; Af: African; Am: Native American; AA: Admixed American; ME: Middle East; SEA: Southeastern Asia; EA: Eastern Asia; In: India. The Portuguese population is highlighted in red, the Welsh in green and the Finnish in orange.

The F_{ST} distances computed based on the 17 Y-STR haplotypes present in Yfiler™ are very similar to the ones based on the haplotypes obtained with PowerPlex® Y23. Among European populations, levels of genetic differentiation were relatively low, with several pairwise F_{ST} being non-significant. In **Fig. 24** it is possible to see replicated the findings obtained with PowerPlex® Y23 once the same populations stand out as the most differentiated: Africans in general plus Finnish and Welsh. The relative distances involving Asian populations are identical, as are also similar those involving Native American populations.

In general, average FST values between major groups of human populations tend to be slightly higher when only based upon the 17 Y-STR of Yfiler™ than based on this

set plus the 6 additional markers in PowerPlex® Y23 (**Table 12**). What is probably accounting for this difference is the fact that all the six additional loci are fast mutating and very diverse in most populations, therefore losing some ability to gain inter-population diversity at a level identical to Y-STRs typically less diverse and mutable.

	PowerPlex® Y23	Yfiler™
European – African	0.16680	0.17483
European – Asian	0.09916	0.10818
European – Native American	0.11490	0.13534
African - Asian	0.14973	0.16098
African – Native Ameircan	0.19977	0.21321
Asian – Native American	0.11997	0.13227

Table 12 – Average F_{ST} distances values between European, Asian, African and Native American populations for both PowerPlex® Y23 and Yfiler™ haplotypes.

The multidimensional scaling performed with the F_{ST} distances based on 17 Y-STR haplotypes (**Fig. 25**) is again very similar to the one performed based on 23 Y-STR haplotypes. Although some differences are easily detected, they are usually minor, since in overall the relative position of every population is very similar and the main inferences drawn from the PowerPlex® Y23 based MDS can be drawn from this one. Briefly, three major groups appear well resolved: Europe, Africa and Eastern Asia. Western Asian populations are again close to Europeans whereas Native-American populations also form a distinct coherent cluster.

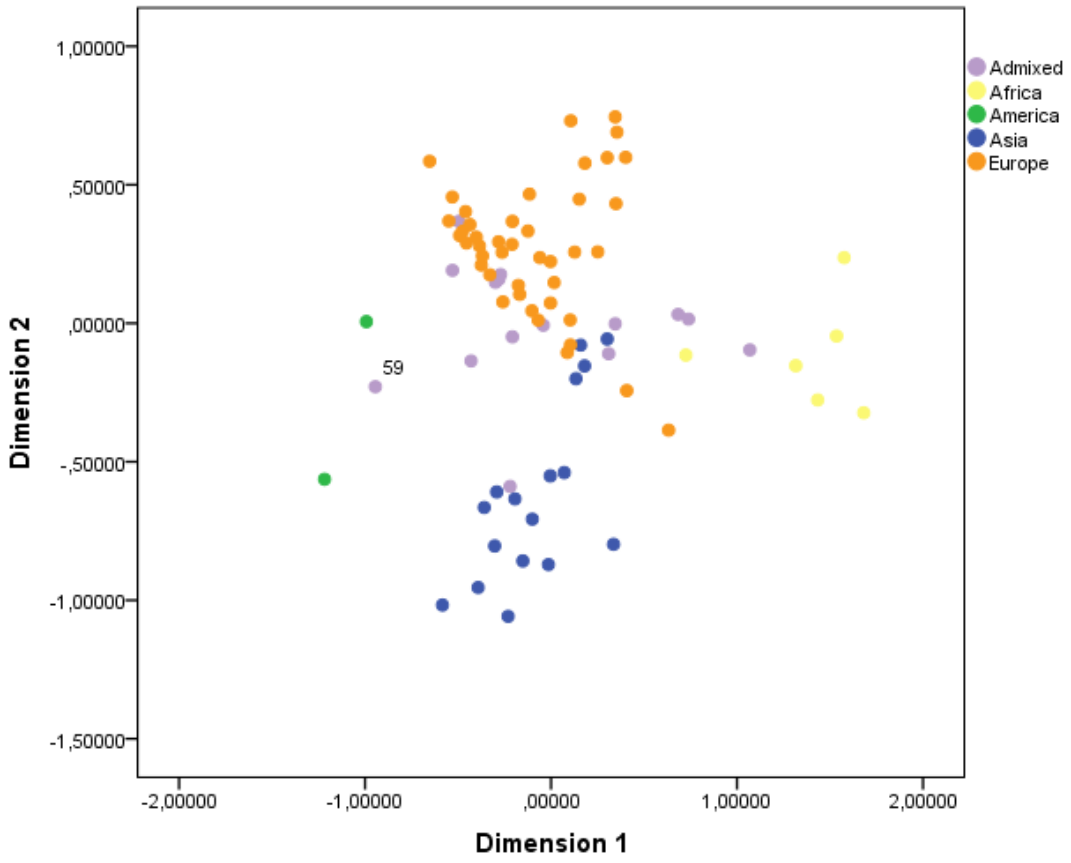

Fig. 25 – MDS plot based on F_{ST} values for 17 Y-STR based haplotypes showing relationships among 86 Worldwide populations. Finnish and Welsh populations were not included in this analysis.

The F_{ST} distances between the European populations using 17 Y-STRs haplotypes are graphically represented in **Fig. 26**.

Also in Europe, the broad pattern of differentiation between populations provided by Yfiler™ is very similar to that based on 23 Y-STR haplotypes. The Finnish and the Welsh populations once again showed to be the most distinct European populations, with all of their pairwise F_{ST} being statistically significant. The Northeastern European populations and Hungarian Romani show again considerable F_{ST} distances (almost all of them being significant), as well as populations from Poland, Croatia, Macedonia and Bosnia and Herzegovina.

Regarding the Portuguese population, the results obtained with 17 Y-STR haplotypes are similar to the ones obtained with 23 Y-STRs haplotypes.

In a few cases, distances between some populations showed to be statistically significant when based on 23 Y-STRs haplotypes while not-significant when based on 17 STRs haplotypes, and vice-versa.

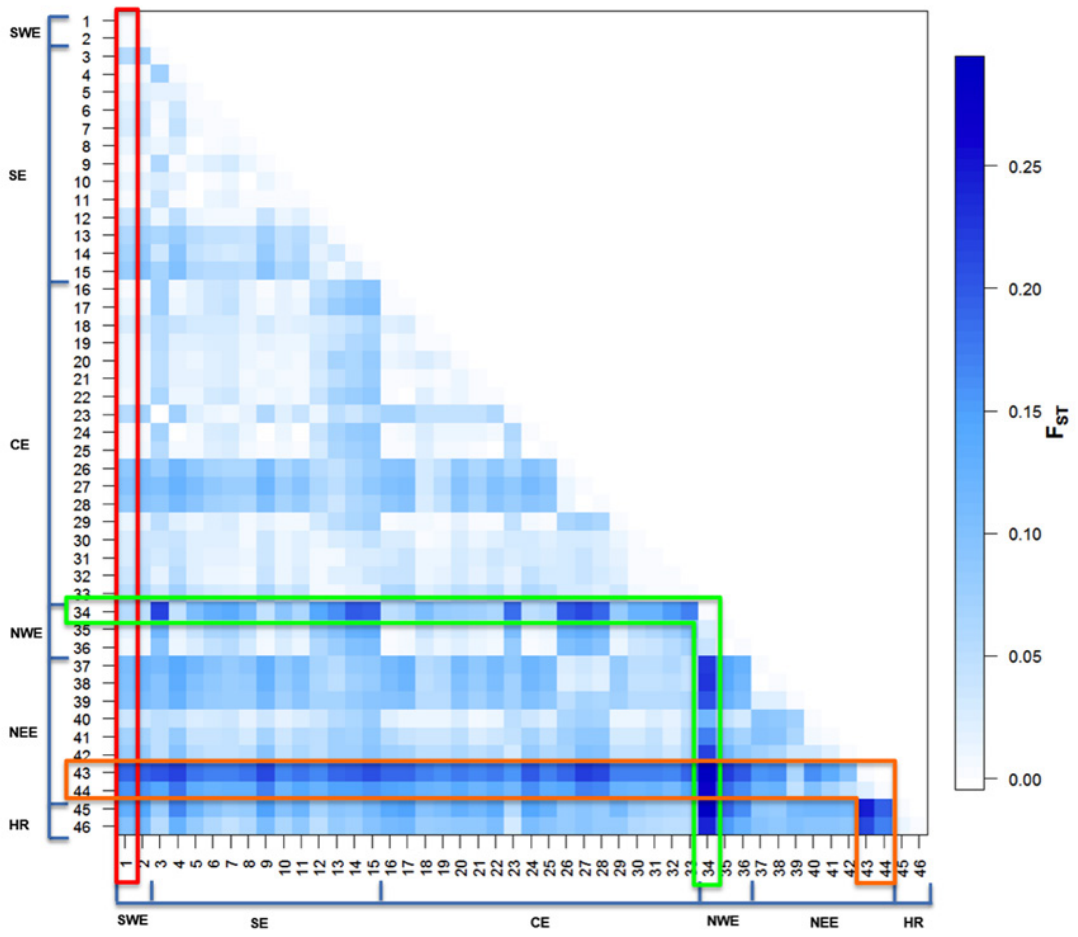

Fig. 26 – Graphical representation of the pairwise F_{ST} distance between 46 European populations. F_{ST} values are based on the 17 Y-STR haplotypes present in Yfiler™. Each number corresponds to a population. All populations, its respective numbers and groups can be consulted in **Supplementary Table 1**. SWE: Southwestern Europe; SE: Southern Europe; CE: Central Europe; NWE: Northwestern Europe; NEE: Northeastern Europe; HR: Hungary Roma. The Portuguese population is highlighted in red, the Welsh in green and the Finnish in orange.

The strong parallelism with PowerPlex®Y23-based results was also patent in the multidimensional scaling performed with the pairwise F_{ST} yielded just from data of the 17 Y-STR haplotypes in European populations (**Fig. 27**).

The dimension 2 of the MDS reflects again much of the inter-population differentiation that varies with latitude, and the relative positions of Northern, Central and Southern populations are maintained as in plot **Fig. 22**. Likewise, dimension 1 also discriminates populations mainly according to variation in longitude, although the position of some populations does not fit so clearly the longitudinal pattern, like those from Croatia (population 13) and Bosnia Herzegovina (population 15). Concerning Bosnia Herzegovina its differentiation in the context of the neighboring populations, is now more evident than in the plot **Fig. 22**.

The populations from Calabria (population 3) and Macedonia (population 14), stand out somehow from the remaining Southern European populations, whereas among Central Europe, the populations from Poland (populations 26, 27 and 28) and Basel (population 23) are those more differentiated from the neighboring populations.

The Portuguese are again close to both Southern and Central European populations.

In summary, the markers added in PowerPlex®Y23 do not appear to alter the broad picture on patterns of male genetic diversity in Europe provided by the 17 STRs in Yfiler™. Possibly, a more detailed analysis focusing on restricted geographical regions will reveal differences in Yfiler™ versus PowerPlex® Y23 in resolving population structure. At a broader perspective, however, the pictures are very similar on population affinities provided by the two multiplexes.

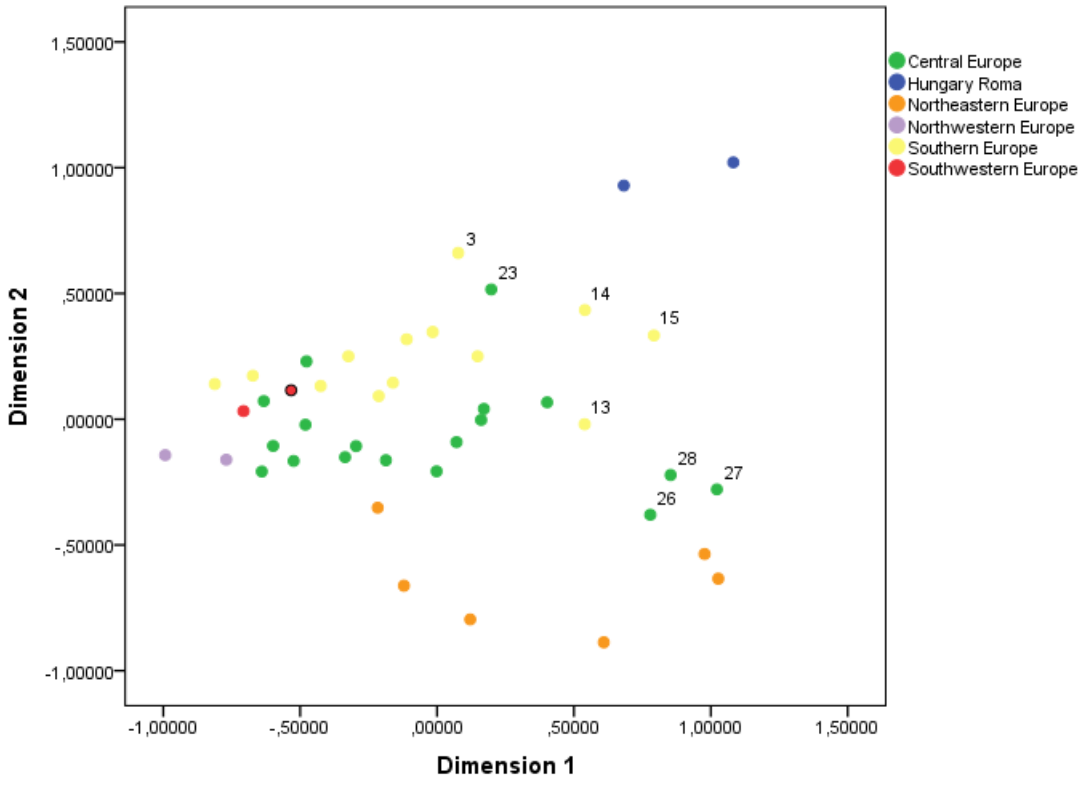

Fig. 27 – MDS plot based on F_{ST} values for 17 Y-STR based haplotypes showing relationships among 46 European populations. The Portuguese population is represented in red with a black border. Finnish and Welsh populations were not included in this analysis. Each number corresponds to a population and they all can be consulted in **Supplementary Table 1**.

A regression analysis between geographic distances and F_{ST} values obtained with Yfiler™ haplotypes (**Fig. 28**) produced this time a significant R^2 (p-value equals to

0.000) of 28.29%, which is slightly superior to the one based on PowerPlex® Y23 haplotypes. In line with the differences in the magnitude of F_{ST}, this analysis also indicates that the set of Yfiler™ markers portrays more geographical structure in Europe than added up with the 6 markers of PowerPlex® Y23.

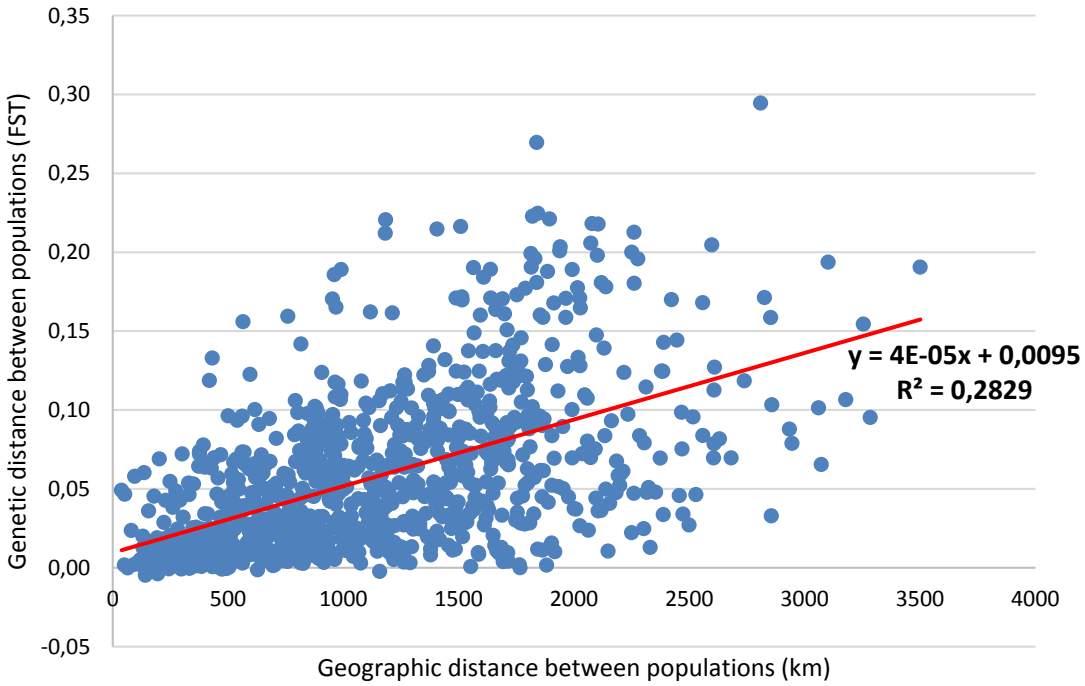

Fig. 28 – Scatter plot and linear regression of the genetic (F_{ST} values based on Yfiler™ haplotypes) and geographic distances of the European populations. P-value equals to 0.000.

5 CONCLUSION

The present work represents an effort to evaluate the potential contribution of the PowerPlex® Y23 kit to Forensic Genetics in the Portuguese population, as well as to other applications in Population Genetics.

The typing of 250 Portuguese male samples with PowerPlex® Y23 showed discrepancies in two samples (involving in each a unique, but distinct Y-STR) comparatively to the results previously obtained with the Yfiler™ kit. A silent allele was detected in the DYS385 locus and an apparent silent allele that was in fact a deletion was found in GATA H4 locus. These type of differences between multiplex kits happen due to the different primer sets used by the manufacturer of each kit and can be potentially prejudicial to inter or even intra-laboratorial reproducibility.

The addition of 6 Y-STRs to the 17 present in Yfiler™ kit, allows an increase from 0.9993 to 0.9996 in the HD in our sample, which indicates that PowerPlex® Y23 may be the best option for forensic analysis in the Portuguese Population. From the 6 new loci, DYS576 and DYS481 are the ones which contribute more to the increase of HD, whereas on the other hand, DYS643 did not produce any change of HD. From the total set of 23 Y-STRs, DYS570 and DYS76 (two of the 6 new loci) have the highest values of gene diversity and DYS393 has the lowest. 5 from the 6 new loci have gene diversity values higher than the average value given by the entire set of 23 STRs.

The results obtained in the Haplogroup Predictor were accurate for both kits in 80 of the 91 haplotypes tested. In 6 cases, the previsions obtained with Yfiler™-haplotypes were correct whereas they turned to be wrong with the extended haplotypes provided by PowerPlex® Y23. Of note, however, that all these discrepancies involved correct haplogroups that are low-frequent or moderately represented in Western European populations. In 5 cases the results were wrong for both multiplex kits, but 4 of them can be explained by the fact that the Haplogroup Predictor doesn't consider haplogroups with scarce phylogenetic resolution. All these results indicate that currently the set of Yfiler™ markers is better alone to infer Y-chromosome haplogroups than complemented with the additional markers contained in PowerPlex® Y23. This may be likely due to the still very incipient coverage of data for the new markers available on the Haplogroup Predictor database.

The population analysis performed using F_{ST} values based on 23 Y-STRs haplotypes were able to afford a broad panorama on the global structure of human genetic diversity. The MDS plotted based on F_{ST} values illustrates the sharp clustering

of three major groups of populations: European, African and Eastern Asian. At the European level, geography accounts considerably for the pattern of population substructuring, although it is not rare to observe consistencies in the relationship between geographical and genetic distance.

The results obtained from the population analysis performed using F_{ST} values based on 17 Y-STR haplotypes are very similar to the results based on 23 Y-STR haplotypes. F_{ST} values tended to be higher when obtained with 17 Y-STR haplotypes than with 23 Y-STRs, which may be caused by the fact that the 6 additional loci are fast mutating and very diverse in most populations. As a consequence, when added to the 17 Y-STRs of Yfiler™ the combined panel loses some ability to retain inter-population diversity, lowering the level afforded by the restricted panel of 17 Y-STRs, since on average these are typically less diverse and mutable.

6 REFERENCES

[1] Carracedo A. Forensic Genetics: History. In: Editors-in-Chief: Jay AS, Pekka JS, editors. Encyclopedia of Forensic Sciences. Waltham: Academic Press; 2013. p. 206-10.

[2] Goodwin W, Linacre A, Hadi S. An Introduction to Forensic Genetics: John Wiley & Sons; 2007.

[3] Crow JF. Felix Bernstein and the First Human Marker Locus. Genetics. 1993;133:4-7.

[4] Butler JM. Chapter 1 - Overview and History of DNA Typing. Fundamentals of Forensic DNA Typing. San Diego: Academic Press; 2010. p. 1-18.

[5] Noordam MJ, Repping S. The Human Y Chromosome: a Masculine Chromosome. Current opinion in genetics & development. 2006;16:225-32.

[6] Navarro-Costa P. Sex, Rebellion and Decadence: the Scandalous Evolutionary History of the Human Y Chromosome. Biochimia et Biophysica Acta. 2012;1822:1851-63.

[7] Quintana-Murci Ls, Krausz C, McElreavey K. The Human Y Chromosome: Function, Evolution and Disease. Forensic science international. 2001;118:169-81.

[8] Gusmão L, Brion M, González-Neira A, Lareu M, Carracedo A. Y Chromosome Specific Polymorphisms in Forensic Analysis. Legal Medicine. 1999;1:55-60.

[9] Iida R, Kishi K. Identification, Characterization and Forensic Application of Novel Y-STRs. Legal Medicine. 2005;7:255-8.

[10] Graves JA. Sex Chromosome Specialization and Degeneration in Mammals. Cell. 2006;124:901-14.

[11] Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, et al. The Male-Specific Region of the Human Y Chromosome is a Mosaic of Discrete Sequence Classes. Nature. 2003;423:825-37.

[12] Butler JM. Chapter 13 - Y-Chromosome DNA Testing. Advanced Topics in Forensic DNA Typing. San Diego: Academic Press; 2012. p. 371-403.

[13] Waters PD, Wallis MC, Marshall Graves JA. Mammalian Sex--Origin and Evolution of the Y Chromosome and SRY. Seminars in cell & developmental biology. 2007;18:389-400.

[14] Schlegel PN. The Y Chromosome. Reproductive BioMedicine Online. 2002;5:22-5.

[15] Ohno S. Sex Chromosomes and Sex-linked Genes. Annals of Internal Medicine. 1968;68:1375-.

[16] Prinz M, Sansone M. Y Chromosome-Specific Short Tandem Repeats in Forensic Casework. Croatian medical journal. 2001;42:288-91.

[17] Alshamali F, Alkhayat AQ, Budowle B, Watson N. Y Chromosome in Forensic Casework and Paternity Testing. International Congress Series. 2004;1261:353-6.

[18] de Knijff P. Messages Through Bottlenecks: on the Combined Use of Slow and Fast Evolving Polymorphic Markers on the Human Y Chromosome. American journal of human genetics. 2000;67:1055-61.

[19] Consortium YC. A Nomenclature System for the Tree of Human Y-Chromosomal Binary Haplogroups. Genome research. 2002;12:339-48.

[20] Genealogy ISoG. Y-DNA Haplogroup Tree 2013. http://www.isogg.org/tree/. 8.75 ed2012.

[21] Lessig R, Zoledziewska M, Fahr K, Edelmann J, Kostrzewa M, Dobosz T, et al. Y-SNP-Genotyping – a New Approach in Forensic Analysis. Forensic science international. 2005;154:128-36.

[22] Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, Hammer MF. New Binary Polymorphisms Reshape and Increase Resolution of the Human Y Chromosomal Haplogroup Tree. Genome research. 2008;18:830-8.

[23] Amorim A, Pereira L. Pros and Cons in the Use of SNPs in Forensic Kinship Investigation: a Comparative Analysis with STRs. Forensic science international. 2005;150:17-21.

[24] Butler J, Coble M, Vallone P. STRs vs. SNPs: Thoughts on the Future of Forensic DNA Testing. Forensic science, medicine, and pathology. 2007;3:200-5.

[25] Kayser M, Kittler R, Erler A, Hedman M, Lee AC, Mohyuddin A, et al. A Comprehensive Survey of Human Y-Chromosomal Microsatellites. American journal of human genetics. 2004;74:1183-97.

[26] Dupuy BM, Stenersen M, Egeland T, Olaisen B. Y-Chromosomal Microsatellite Mutation Rates: Differences in Mutation Rate Between and Within loci. Human mutation. 2004;23:117-24.

[27] Bao W, Zhu S, Pandya A, Zerjal T, Xu J, Shu Q, et al. MSY2: a Slowly Evolving Minisatellite on the Human Y Chromosome Which Provides a Useful Polymorphic Marker in Chinese Populations. Gene. 2000;244:29-33.

[28] Jobling MA, Heyer E, Dieltjes P, de Knijff P. Y-Chromosome-Specific Microsatellite Mutation Rates Re-examined Using a Minisatellite, MSY1. Human molecular genetics. 1999;8:2117-20.

[29] Gusmao L, Carracedo A. Y Chromosome Specific STRs. Profiles in DNA. 2003;6:4.

[30] Prinz M, Boll K, Baum H, Shaler B. Multiplexing of Y Chromosome Specific STRs and Performance for Mixed Samples. Forensic science international. 1997;85:209-18.

[31] Leat N, Ehrenreich L, Benjeddou M, Cloete K, Davison S. Properties of Novel and Widely Studied Y-STR Loci in Three South African Populations. Forensic science international. 2007;168:154-61.

[32] Kayser M, Caglia A, Corach D, Fretwell N, Gehrig C, Graziosi G, et al. Evaluation of Y-chromosomal STRs: a Multicenter Study. International journal of legal medicine. 1997;110:125-33, 41-9.

[33] Pascali VL, Dobosz M, Brinkmann B. Coordinating Y-Chromosomal STR Research for the Courts. International journal of legal medicine. 1999;112:1.

[34] Ayub Q, Mohyuddin A, Qamar R, Mazhar K, Zerjal T, Mehdi SQ, et al. Identification and Characterisation of Novel Human Y-Chromosomal Microsatellites From Sequence Database Information. Nucleic acids research. 2000;28:e8.

[35] Butler JM. Forensic DNA Typing: Biology, Technology, and Genetics of STR Markers (2nd Edition). New York: Elsevier Academic Press; 2005.

[36] Thompson JM, Ewing MM, Frank WE, Pogemiller JJ, Nolde CA, Koehler DJ, et al. Developmental Validation of the PowerPlex® Y23 System: A Single Multiplex Y-STR Analysis System for Casework and Database Samples. Forensic Science International: Genetics. 2013;7:240-50.

[37] Butler JM. Chapter 7 - Quality Assurance and Validation. Advanced Topics in Forensic DNA Typing. San Diego: Academic Press; 2012. p. 167-211.

[38] Butler JM. Chapter 5 - Short Tandem Repeat (STR) Loci and Kits. Advanced Topics in Forensic DNA Typing. San Diego: Academic Press; 2012. p. 99-139.

[39] SWGDAM SWGoDAM. Validation Guidelines for DNA Analysis Methods. http://swgdam.org/SWGDAM_Validation_Guidelines_APPROVED_Dec_2012.pdf. 2012.

[40] FBI. Quality Assurance Standards for Forensic DNA Testing Laboratories. http://www.fbi.gov/about-us/lab/biometric-analysis/codis/qas-standards-for-forensic-dna-testing-laboratories-effective-9-1-2011. 2011.

[41] Corporation P. Internal Validation Guide for Y-STR Systems in Forensic Laboratories. Promega Corporation; 2012.

[42] Alves C, Gusmão L, Pereira L, Amorim A. Multiplex STR Genotyping: Comparison Study, Population Data and New Sequence Information. International Congress Series. 2003;1239:131-5.

[43] Kline MC, Jenkins B, Rodgers.S. Non-Amplification of a vWA Allele. Journal of Forensic Sciences. 1998;43.

[44] Davis C, Ge J, Sprecher C, Chidambaram A, Thompson J, Ewing M, et al. Prototype PowerPlex® Y23 System: A Concordance Study. Forensic Science International: Genetics. 2013;7:204-8.

[45] Budowle B, Masibay A, Anderson SJ, Barna C, Biega L, Brenneke S, et al. STR Primer Concordance Study. Forensic science international. 2001;124:47-54.

[46] Van Nieuwerburgh F, Goetghebeur E, Vandewoestyne M, Deforce D. Impact of Allelic Dropout on Evidential Value of Forensic DNA Profiles Using RMNE. Bioinformatics. 2009;25:225-9.

[47] National Forensic Science Technology Center website. http://www.nfstc.org/.

[48] Willuweit S, Roewer L. Y Chromosome Haplotype Reference Database (YHRD). http://www.yhrd.org/. 2000.

[49] Roewer L, Krawczak M, Willuweit S, Nagy M, Alves C, Amorim A, et al. Online Reference Database of European Y-Chromosomal Short Tandem Repeat (STR) Haplotypes. Forensic science international. 2001;118:106-13.

[50] Willuweit S, Roewer L. Y Chromosome Haplotype Reference Database (YHRD): Update. Forensic Science International: Genetics. 2007;1:83-7.

[51] Holland MM, Parsons TJ. Mithocondrial DNA Sequence Analysis - Validation and Use for Forensic Casework. Forensic Science Review. 1999;11.

[52] Buckleton JS. Forensic DNA Evidence Interpretation: CRC Press; 2004.

[53] Corporation P. Promega New PowerPlex® Y23 STR System Reveals More Y-STR Loci in Half the Time. http://worldwide.promega.com/aboutus/press-releases/2012/20120712-powerplex-y23?__utma=1.558124223.1365090555.1365090555.1365090555.1&__utmb=1.4.10.1365090555&__utmc=1&__utmx=-&__utmz=1.1365090555.1.1.utmcsr=google|utmccn=(organic)|utmcmd=organic|utmctr=(not%20provided)&__utmv=-&__utmk=186243881. Madison, Wisconsin2012.

[54] Corporation P. PowerPlex® Y23 System. http://worldwide.promega.com/products/pm/genetic-identity/powerplex-y23/?origUrl=http%3a%2f%2fwww.promega.com%2fproducts%2fpm%2fgenetic-identity%2fpowerplex-y23%2f. 2012.

[55] Butler JM, Hill CR, Coble MD. Variability of New STR Loci and Kits in US Population Groups. Promega Corporation 2012.

[56] Vermeulen M, Wollstein A, van der Gaag K, Lao O, Xue Y, Wang Q, et al. Improving Global and Regional Resolution of Male Lineage Differentiation by Simple Single-Copy Y-Chromosomal Short Tandem Repeat Polymorphisms. Forensic Science International: Genetics. 2009;3:205-13.

[57] Lim SK, Xue Y, Parkin EJ, Tyler-Smith C. Variation of 52 new Y-STR loci in the Y Chromosome Consortium Worldwide Panel of 76 Diverse Individuals. International journal of legal medicine. 2007;121:124-7.

[58] D'Amato ME, Ehrenreich L, Cloete K, Benjeddou M, Davison S. Characterization of the Highly Discriminatory Loci DYS449, DYS481, DYS518, DYS612, DYS626, DYS644 and DYS710. Forensic Science International: Genetics. 2010;4:104-10.

[59] Hanson EK, Ballantyne J. Comprehensive Annotated STR Physical Map of the Human Y Chromosome: Forensic Implications. Legal Medicine. 2006;8:110-20.

[60] Ballantyne KN, Goedbloed M, Fang R, Schaap O, Lao O, Wollstein A, et al. Mutability of Y-Chromosomal Microsatellites: Rates, Characteristics, Molecular bases, and Forensic Implications. American journal of human genetics. 2010;87:341-53.

[61] Geppert M, Edelmann J, Lessig R. The Y-Chromosomal STRs DYS481, DYS570, DYS576 and DYS643. Legal Medicine. 2009;11 Suppl 1:S109-10.

[62] Ballantyne KN, Keerl V, Wollstein A, Choi Y, Zuniga SB, Ralf A, et al. A New Future of Forensic Y-Chromosome Analysis: Rapidly Mutating Y-STRs for Differentiating Male Relatives and Paternal Lineages. Forensic Science International: Genetics. 2012;6:208-18.

[63] Thompson J. PowerPlex Y23 Developmental Validation. 2012.

[64] Alves C, Gomes V, Prata MJ, Amorim A, Gusmao L. Population Data for Y-Chromosome Haplotypes Defined by 17 STRs (AmpFlSTR YFiler) in Portugal. Forensic science international. 2007;171:250-5.

[65] Carvalho M, Anjos MJ, Andrade L, Lopes V, Santos MV, Gamero JJ, et al. Y-Chromosome STR Haplotypes in Two Population Samples: Azores Islands and Central Portugal. Forensic science international. 2003;134:29-35.

[66] Gonzalez-Neira A, Gusmao L, Brion M, Lareu MV, Amorim A, Carracedo A. Distribution of Y-Chromosome STR Defined Haplotypes in Iberia. Forensic science international. 2000;110:117-26.

[67] Fernandes AT, Brehm A, Gusmao L, Amorim A. Y-Chromosome STR Haplotypes in the Madeira Archipelago Population. Forensic science international. 2001;122:178-80.

[68] Walsh PS, Metzger DA, Higuchi R. Chelex 100 as a Medium for Simple Extraction of DNA for PCR-Based Typing From Forensic Material. BioTechniques. 1991;10:506-13.

[69] Butler JM, Reeder DJ. Short Tandem Repeat DNA Internet DataBase. http://www.cstl.nist.gov/strbase/index.htm. 1997.

[70] Budowle B, Chakraborty R, Giusti AM, Eisenberg AJ, Allen RC. Analysis of the VNTR Locus D1S80 by the PCR Followed by High-Resolution PAGE. American journal of human genetics. 1991;48:137-44.

[71] Alves C, Amorim A, Gusmao L, Pereira L. VWA STR Genotyping: Further Inconsistencies Between Perkin-Elmer and Promega Kits. International journal of legal medicine. 2001;115:97-9.

[72] Porath J, Flodin P. Gel Filtration: a Method for Desalting and Group Separation. Nature. 1959;183:1657-9.

[73] Alves C, Gusmão L, Meirinhos J, Amorim A. Making the Most of Y-STR Haplotypes: The HapYDive. International Congress Series. 2006;1288:201-3.

[74] Excoffier L, Lischer HEL. Arlequin Suite ver 3.5: a New Series of Programs to Perform Population Genetics Analyses Under Linux and Windows. Molecular Ecology Resources. 2010;10:564-7.

[75] Corp. I. IBM SPSS Statistics for Windows. 21.0 ed. Armonk, NY: IBM Corp.; 2012.

[76] Athey W. Haplogroup Predictor. http://www.hprg.com/. 2004.

[77] Coble MD, Hill CR, Butler JM. Haplotype Data for 23 Y-Chromosome Markers in Four U.S. Population Groups. Forensic Science International: Genetics. 2013;7:e66-e8.

[78] Pereira V, Gomes V, Amorim A, Gusmao L, Joao Prata M. Genetic Characterization of Uniparental Lineages in Populations from Southwest Iberia with Past Malaria Endemicity. Am Journal of Human Biology. 2010;22:588-95.

[79] Lao O, Lu TT, Nothnagel M, Junge O, Freitag-Wolf S, Caliebe A, et al. Correlation Between Genetic and Geographic Structure in Europe. Current Biology: CB. 2008;18:1241-8.

[80] Seldin MF, Shigeta R, Villoslada P, Selmi C, Tuomilehto J, Silva G, et al. European Population Sbstructure: Clustering of Northern and Southern Populations. PLoS genetics. 2006;2:e143.

[81] Zerjal T, Beckman L, Beckman G, Mikelsaar AV, Krumina A, Kucinskas V, et al. Geographical, Linguistic, and Cultural Influences on Genetic Diversity: Y-Chromosomal Distribution in Northern European Populations. Molecular biology and evolution. 2001;18:1077-87.

[82] Norio R. Finnish Disease Heritage II: Population Prehistory and Genetic Roots of Finns. Human genetics. 2003;112:457-69.

[83] Sajantila A, Salem AH, Savolainen P, Bauer K, Gierig C, Paabo S. Paternal and Maternal DNA Lineages Reveal a Bottleneck in the Founding of the Finnish Population. Proceedings of the National Academy of Sciences of the United States of America. 1996;93:12035-9.

[84] Weale ME, Weiss DA, Jager RF, Bradman N, Thomas MG. Y Chromosome Evidence for Anglo-Saxon Mass Migration. Molecular biology and evolution. 2002;19:1008-21.

[85] Capelli C, Redhead N, Abernethy JK, Gratrix F, Wilson JF, Moen T, et al. A Y chromosome Census of the British Isles. Current biology : CB. 2003;13:979-84.

[86] Gusmao A, Gusmao L, Gomes V, Alves C, Calafell F, Amorim A, et al. A Perspective on the History of the Iberian Gypsies Provided by Phylogeographic Analysis of Y-Chromosome Lineages. Annals of human genetics. 2008;72:215-27.

[87] Gresham D, Morar B, Underhill PA, Passarino G, Lin AA, Wise C, et al. Origins and Divergence of the Roma (Gypsies). The American Journal of Human Genetics. 2001;69:1314-31.

[88] Regueiro M, Rivera L, Chennakrishnaiah S, Popovic B, Andjus S, Milasin J, et al. Ancestral Modal Y-STR Haplotype Shared Among Romani and South Indian Populations. Gene. 2012;504:296-302.

7 APPENDIX

ID	Population	Continent	Group	Group ID	N
1	Portugal	Europe	Southwestern Europe	SWE	248
2	Spain	Europe	Southwestern Europe	SWE	200
3	Calabria Italy	Europe	Southern Europe	SE	30
4	Brescia Italy	Europe	Southern Europe	SE	124
5	Liguria Italy	Europe	Southern Europe	SE	45
6	Puglia Italy	Europe	Southern Europe	SE	160
7	Sicily Italy	Europe	Southern Europe	SE	157
8	Milano Italy	Europe	Southern Europe	SE	70
9	Tuscany Italy	Europe	Southern Europe	SE	59
10	Northeast Italy	Europe	Southern Europe	SE	200
11	Ravenna Italy	Europe	Southern Europe	SE	200
12	Greece	Europe	Southern Europe	SE	200
13	Croatia	Europe	Southern Europe	SE	200
14	Macedonia	Europe	Southern Europe	SE	101
15	Bosnia and Herzegovina	Europe	Southern Europe	SE	100
16	Belgium	Europe	Central Europe	CE	200
17	Netherlands	Europe	Central Europe	CE	200
18	Germany	Europe	Central Europe	CE	200
19	Mecklenburg-Vorpommen Germany	Europe	Central Europe	CE	173
20	Stuttgart Germany	Europe	Central Europe	CE	118
21	Freiburg Germany	Europe	Central Europe	CE	200
22	Switzerland	Europe	Central Europe	CE	150
23	Basel Switzterland	Europe	Central Europe	CE	138
24	Lausanne Switzerland	Europe	Central Europe	CE	100
25	Aarau Switzerland	Europe	Central Europe	CE	200
26	Bialystock Poland	Europe	Central Europe	CE	150
27	Poznan Poland	Europe	Central Europe	CE	150
28	Poland	Europe	Central Europe	CE	200
29	Austria	Europe	Central Europe	CE	200
30	Czech Republic	Europe	Central Europe	CE	114
31	Hungary	Europe	Central Europe	CE	143
32	Budapest Hungary	Europe	Central Europe	CE	100
33	Slovenia	Europe	Central Europe	CE	104
34	Wales	Europe	Northwestern Europe	NWE	118
35	United Kingdom	Europe	Northwestern Europe	NWE	200
36	London UK (European)	Europe	Northwestern Europe	NWE	162
37	Latvia	Europe	Northeastern Europe	NEE	139
38	Lithuania	Europe	Northeastern Europe	NEE	84
39	Estonia	Europe	Northeastern Europe	NEE	125
40	Denmark	Europe	Northeastern Europe	NEE	185
41	Sweden	Europe	Northeastern Europe	NEE	54
42	Vaster Sweden	Europe	Northeastern Europe	NEE	41
43	Finland	Europe	Northeastern Europe	NEE	200

ID	Population	Continent	Group	Group ID	N
44	Turku Finland	Europe	Northeastern Europe	NEE	162
45	Hungary Roma	Europe	Hungary Roma	HR	53
46	Hungary Romani	Europe	Hungary Roma	HR	101
47	United Kingdom (British African)	Europe	Admixed European	AE	171
48	United Kingdom (British Asian)	Europe	Admixed European	AE	142
49	North Benin	Africa	Africa	Af	51
50	Webuye Kenya (Luhya)	Africa	Africa	Af	44
51	Kinyawa Kenya (Maasai)	Africa	Africa	Af	100
52	Ibadan Nigeria (Yoruba)	Africa	Africa	Af	81
53	South Africa (Xhosa)	Africa	Africa	Af	114
54	Zimbabwe	Africa	Africa	Af	55
55	Bolivia (Native American)	America	America	Am	56
56	Brazil (Native American)	America	America	Am	61
57	USA (Gujarati Indians)	America	America	Am	66
58	Argentina (European)	America	Admixed American	AA	200
59	Bolivia (Mestizo)	America	Admixed American	AA	44
60	Brazil (Admixed)	America	Admixed American	AA	200
61	Costa Rica (Mestizo)	America	Admixed American	AA	166
62	Jamaica	America	Admixed American	AA	66
63	Panama	America	Admixed American	AA	100
64	Peru	America	Admixed American	AA	83
65	USA (African American)	America	Admixed American	AA	169
66	USA (Chinese)	America	Admixed American	AA	60
67	USA (European American)	America	Admixed American	AA	186
68	USA (Hispanic American)	America	Admixed American	AA	41
69	USA (Mexican)	America	Admixed American	AA	34
70	Iraq	Asia	Middle East	ME	124
71	Lebanon	Asia	Middle East	ME	200
72	East Timor	Asia	Southeastern Asia	SEA	100
73	Philippines	Asia	Southeastern Asia	SEA	169
74	Singapore (Indian)	Asia	Southeastern Asia	SEA	106
75	Singapore (Han)	Asia	Southeastern Asia	SEA	104
76	Singapore (Malay)	Asia	Southeastern Asia	SEA	104
77	Vietnam	Asia	Southeastern Asia	SEA	45
78	Yunnan China (Bai)	Asia	Eastern Asia	EA	101
79	Xishuangbanna China (Dai)	Asia	Eastern Asia	EA	92
80	China (Han)	Asia	Eastern Asia	EA	200
81	Southern China (Han)	Asia	Eastern Asia	EA	30
82	Xuanwei China (Han)	Asia	Eastern Asia	EA	145
83	Shantou China (Minnan Han)	Asia	Eastern Asia	EA	108
84	Japan	Asia	Eastern Asia	EA	200
85	South Korea	Asia	Eastern Asia	EA	200
86	India (Tamil)	Asia	India	In	126

Supplementary Table 1 – Information on the 86 worldwide population haplotypes provided by the YHRD/Promega colaboration.

ID	Pop	DYS576	DYS389I	DYS448	DYS389II	DYS19	DYS391	DYS481	DYS549	DYS533	DYS438	DYS437	DYS570	DYS635	DYS390	DYS439	DYS392	DYS643	DYS393	DYS458	DYS385	DYS456	GATA H4
4875	NP	17	13	19	27	14	11	22	11	11	12	15	19	23	23	13	13	10	13	19	11-15	17	12
4894	SP	15	13	21	31	15	10	28	11	11	11	14	21	21	22	11	11	14	13	16	16-19	15	12
4905	NP	19	13	19	29	14	11	22	12	12	12	15	18	23	24	12	13	10	13	18	11-13	16	12
4913	CP	19	13	19	29	14	11	22	14	12	12	14	17	23	24	12	13	10	16	18	12-15	16	12
4912	NP	17	12	20	29	14	10	24	12	11	10	16	22	22	23	11	11	12	13	15	14-14	15	11
4930	NP	18	13	21	30	14	10	21	12	12	9	14	16	22	22	12	11	8	12	15	12-15	15	11
4939	CP	17	14	20	30	14	9	26	11	11	10	14	20	21	24	10	11	12	13	18	13-14	16	12
4962	NP	18	13	19	29	15	10	22	13	12	12	15	17	23	23	13	13	10	13	17	12-14	17	12
4967	NP	18	14	19	30	14	11	22	12	13	12	15	17	23	23	12	13	10	13	17	11-14	15	12
4955	CP	16	12	22	28	16	11	21	11	10	10	16	18	23	21	11	11	12	14	16	15-15	16	11
4965	NP	18	13	19	29	14	11	22	12	12	12	15	19	23	24	13	13	9	14	17	11-14	16	11
4977	CP	17	14	21	29	17	10	22	10	12	10	15	20	22	23	11	11	11	13	17	12-12	15	12
4994	NP	18	13	19	29	14	10	22	12	12	12	15	16	23	24	12	13	10	13	17	11-14	16	11
5002	NP	19	13	19	29	14	11	22	12	12	12	15	17	24	24	13	13	10	13	17	11-14	15	10
5010	CP	19	13	19	29	14	11	22	14	12	12	15	16	23	24	13	13	10	13	18	12-14	15	11
5017	SP	15	13	19	30	13	10	24	12	12	10	14	18	21	23	13	11	13	13	16	15-16	17	11
5020	SP	16	12	21	28	15	10	22	13	11	10	16	19	23	21	12	11	13	14	17	11-13	15	11
5024	SP	17	12	18	28	14	10	22	12	12	12	14	17	23	24	12	13	11	13	17	11-15	16	11
5051	NP	18	14	18	31	14	11	22	13	12	12	14	17	23	25	13	13	10	13	17	12-14	15	11
4993	NP	17	13	19	29	14	10	22	14	12	12	15	17	23	24	12	13	10	13	16	11-14	15	12
5058	NP	18	13	18	31	15	11	22	13	11	12	14	16	23	24	11	13	10	13	18	11-13	16	12
5073	NP	17	12	22	28	14	10	24	13	10	10	16	19	21	23	12	12	11	13	15	14-14	15	12

ID	Pop	DYS576	DYS389I	DYS448	DYS389II	DYS19	DYS391	DYS481	DYS549	DYS533	DYS438	DYS437	DYS570	DYS635	DYS390	DYS439	DYS392	DYS643	DYS393	DYS458	DYS385	DYS456	GATA H4
5077	SP	19	13	20	31	13	10	23	12	12	10	14	19	22	24	12	11	12	13	14	16-18	16	10
5080	SP	17	12	22	28	15	10	22	12	10	11	16	18	23	21	11	11	13	14	17	13-15	14	11
5092	SP	17	12	19	28	14	11	22	12	12	13	14	18	23	23	14	13	10	14	17	11-14	15	12
5099	NP	17	13	18	29	15	11	28	12	12	10	14	17	24	22	13	11	12	12	13	12-12	15	11
5107	NP	18	14	18	30	14	11	22	11	12	12	15	17	23	23	12	13	10	13	19	12-12	16	12
5119	NP	18	13	19	29	15	11	23	12	12	12	15	17	23	24	12	14	10	13	16	11-14	16	12
5076	NP	19	13	19	29	14	11	23	13	12	12	15	18	23	24	12	13	10	13	16	11-13	16	13
5143	NP	17	12	20	29	14	10	24	12	11	10	16	22	22	23	12	11	12	13	15	15-15	14	11
5170	SP	14	12	23	29	15	10	23	13	10	10	16	17	22	21	11	11	12	15	17	13-17	15	11
5260	SP	20	12	20	28	14	10	23	13	11	11	15	17	20	25	13	14	12	12	17	13-17	15	12
5268	SP	16	12	20	28	14	10	25	13	11	10	15	19	22	22	11	11	13	13	16	13-14	14	11
5301	SP	17	13	18	29	14	11	22	13	12	11	14	18	23	24	14	13	9	13	18	11-13	16	12
5318	NP	17	13	19	29	15	11	23	13	11	12	15	21	23	23	13	14	12	13	16	11-13	16	13
5376	NP	18	12	21	30	15	10	21	12	9	10	15	18	21	22	12	11	12	13	19	12-13	14	11
5385	CP	16	12	21	29	16	10	19	12	9	10	16	16	20	22	10	11	11	14	18	14-15	15	12
5389	SP	17	13	18	29	14	10	22	14	12	12	14	17	24	24	11	13	10	13	17	11-14	16	12
5395	SP	18	14	18	30	14	11	22	12	12	12	15	17	23	24	12	13	9	13	18	11-14	16	12
5398	CP	18	14	20	31	14	10	22	12	11	9	15	14	23	23	12	11	10	13	17	13-15	15	11
5692	NP	18	13	17	29	15	11	22	13	12	12	15	17	23	23	13	13	10	13	18	11-14	16	12
5415	NP	17	12	21	29	15	10	22	11	9	10	16	19	22	22	13	11	11	14	17	15-15	18	12
5421	NP	17	13	19	29	14	11	23	13	13	12	15	17	23	24	13	13	10	12	16	11-14	15	12
5737	NP	18	13	20	29	13	9	26	11	11	10	14	21	21	24	10	11	12	13	18	13-14	15	12

ID	Pop	DYS576	DYS389I	DYS448	DYS389II	DYS19	DYS391	DYS481	DYS549	DYS533	DYS438	DYS437	DYS570	DYS635	DYS390	DYS439	DYS392	DYS643	DYS393	DYS458	DYS385	DYS456	GATA H4
5432	CP	18	13	22	29	14	10	23	15	11	9	15	15	21	23	13	11	11	12	15	14-18	16	11
5429	NP	19	13	19	29	14	11	21	14	12	13	15	16	24	24	13	13	10	12	16	11-12	13	13
5451	NP	17	12	20	29	16	10	26	11	14	10	15	16	24	22	11	11	12	12	16	12-14	14	11
5466	SP	18	13	20	29	14	11	22	13	12	12	15	17	23	24	11	13	10	13	17	11-14	15	13
5484	NP	18	13	20	31	14	9	27	13	11	10	14	16	21	25	11	11	10	12	18.2	14-14	16	10
5507	SP	18	14	20	30	13	9	27	11	11	10	14	23	22	24	10	11	12	13	20	13-14	15	12
5509	NP	18	12	19	28	14	11	23	13	12	12	15	16	23	24	12	13	10	13	16	11-14	16	12
5515	SP	18	14	19	31	14	11	21	14	12	12	15	18	23	25	13	13	10	13	17	11-14	18	12
5558	CP	15	13	21	30	14	10	20	13	11	9	15	19	24	24	12	11	10	12	17	14-16	17	12
5559	NP	19	12	19	30	13	10	24	12	11	10	14	21	23	24	13	11	12	13	18	16-16	16	11
5561	NP	15	13	21	30	15	10	22	12	13	9	14	18	22	22	12	11	8	12	15	13-15	15	11
5572	NP	17	13	19	29	14	11	21	13	11	12	15	17	23	23	13	13	10	13	19	11-14	15	12
5588	NP	17	14	21	31	14	10	22	11	11	9	14	17	23	23	11	11	10	13	17	14-15	17	11
5591	NP	22	13	19	29	14	11	22	13	12	12	15	16	23	25	12	13	8	13	17	11-13	16	12
5599	CP	17	14	20	31	14	11	22	12	11	9	15	14	23	23	12	11	10	13	17	13-15	15	11
5623	NP	18	13	20	29	15	10	22	12	12	12	15	17	23	24	13	13	10	13	17	11-15	15	13
5615	CP	20	13	20	29	15	10	22	12	12	12	15	18	23	24	11	13	10	13	17	11-15	15	13
5627	SP	16	13	21	29	13	10	22	12	12	9	15	17	22	23	11	11	10	12	16	13-16	16	12
5643	SP	17	13	19	29	14	11	22	12	12	12	14	17	24	24	11	13	10	14	17	11-14	15	12
5645	SP	17	13	19	29	14	11	22	13	12	13	15	18	23	24	12	14	10	13	18	11-14	16	11
5648	NP	18	13	20	29	15	10	23	11	11	9	14	17	21	23	10	11	10	12	16	13-16	17	11
5651	SP	17	12	19	29	15	10	23	13	11	12	15	18	23	24	11	13	11	13	18	11-14	17	11

ID	Pop	DYS576	DYS389I	DYS448	DYS389II	DYS19	DYS391	DYS481	DYS549	DYS533	DYS438	DYS437	DYS570	DYS635	DYS390	DYS439	DYS392	DYS643	DYS393	DYS458	DYS385	DYS456	GATA H4
5654	SP	16	13	21	31	15	10	28	11	11	11	14	19	21	21	12	11	13	13	16	15-19	16	12
6088	SP	16	13	20	29	14	10	23	11	11	9	14	17	21	22	11	11	8	12	16	12-15	15	11
5662	NP	16	13	22	30	15	11	22	13	12	9	15	17	24	22	11	11	11	12	16	12-16	16	12
5666	NP	19	13	19	30	14	11	22	13	12	12	16	15	25	23	12	13	10	13	18	11-15	15	11
5686	NP	18	13	18	28	15	11	22	11	12	12	15	16	23	23	11	13	10	13	17	11-14	17	11
5733	NP	18	12	22	29	16	11	21	11	10	10	16	18	23	21	12	11	12	14	16	15-15	16	11
5695	NP	18	13	19	29	14	10	22	12	12	12	14	17	23	23	12	13	10	13	17	11-14	17	11
5746	NP	18	13	19	29	15	11	21	13	13	12	15	17	23	22	11	13	10	12	17	12-14	16	12
5709	NP	15	13	21	29	14	10	22	12	11	9	15	17	24	22	11	11	8	12	16	13-15	15	10
5712	NP	18	14	20	30	14	10	22	12	12	12	15	16	24	24	12	13	10	13	16	11-14	16	12
5726	SP	14	13	19	29	15	9	23	12	12	9	14	17	21	23	12	13	10	13	17	15-15	14	11
5713	NP	19	14	19	30	13	10	28	12	11	10	14	23	22	24	9	11	12	13	17	13-14	16	12
5752	NP	17	12	19	27	15	10	25	12	12	9	14	17	21	25	12	14	9	13	16	13-16	15	11
5740	NP	19	14	20	31	15	11	24	11	12	10	15	17	21	22	11	12	12	13	17	15-16	13	11
5766	Ma	18	13	19	29	14	10	22	12	12	12	14	18	23	24	11	14	10	13	16	11-14	15	11
6013	NP	18	12	22	30	16	10	24	11	12	10	15	17	24	25	11	11	12	12	18	12-16	17	12
5763	NP	18	14	20	31	15	10	24	12	11	10	14	18	22	23	10	12	13	14	15	14-15	15	11
5775	NP	17	13	19	29	13	10	22	13	13	12	15	17	24	24	12	13	10	13	18	12-14	17	11
5934	SP	18	13	19	29	15	11	22	11	12	12	14	16	23	24	13	13	10	13	18	11-14	16	11
5789	NP	19	13	19	29	14	10	22	12	11	11	16	17	23	24	13	13	11	13	16	11-13	16	12
5797	SP	18	13	19	29	14	11	22	12	12	12	15	15	23	24	13	13	10	13	18	11-15	15	12
5806	NP	17	12	22	28	15	10	22	12	10	11	16	18	23	21	11	11	13	14	17	13-15	14	11

ID	Pop	DYS576	DYS389I	DYS448	DYS389II	DYS19	DYS391	DYS481	DYS549	DYS533	DYS438	DYS437	DYS570	DYS635	DYS390	DYS439	DYS392	DYS643	DYS393	DYS458	DYS385	DYS456	GATA H4
5753	CP	17	14	20	30	13	10	27	11	11	10	14	22	21	24	11	11	12	13	18	13-14	15	12
5812	NP	17	14	19	30	15	11	23	12	12	12	15	17	23	24	12	14	10	13	16	11-14	16	13
5822	SP	17	12	20	28	14	10	25	13	11	10	16	19	23	22	11	11	12	13	16	13-14	14	11
5825	NP	18	13	19	29	14	11	22	13	12	12	15	19	23	24	12	13	10	13	19	8-14	15	12
5832	SP	18	12	19	29	15	10	23	13	12	12	15	18	23	23	13	13	10	13	17	11-14	16	13
5834	SP	20	13	20	29	14	11	22	13	13	13	15	17	23	24	13	13	10	13	18	11-15	18	12
5837	NP	18	13	21	28	17	10	23	10	12	10	15	19	22	23	11	11	12	13	16	11-12	14	12
5843	SP	16	12	18	28	15	10	23	13	12	9	16	19	23	23	12	11	9	12	16	13-17	13	11
5871	NP	17	14	21	30	13	9	26	11	12	10	14	22	21	24	10	11	12	13	20	14-14	16	11
5873	NP	18	13	19	29	14	10	22	12	13	13	15	19	23	25	11	13	10	12	16	11-14	16	13
5879	CP	16	13	20	30	15	10	23	12	12	9	14	16	22	22	12	11	8	13	18	14-14	16	11
5881	NP	17	13	20	30	13	10	22	12	12	10	14	19	21	25	13	11	12	13	15	16-18	17	12
5906	NP	17	12	20	28	14	10	26	12	11	10	16	19	22	22	11	11	12	13	15	13-14	14	11
6132	SP	17	13	19	29	14	10	23	13	12	13	15	18	23	24	11	15	10	13	16	11-14	15	12
5887	CP	15	13	20	31	15	10	29	11	11	11	14	21	21	21	11	11	14	13	15	17-19	15	12
5989	SP	18	13	19	29	14	10	23	12	12	12	15	17	24	24	12	13	10	13	17	11-14	16	12
5916	Az	17	12	21	29	15	10	22	13	9	10	16	18	21	22	12	11	11	14	17	12-16	18	12
5923	NP	17	13	19	30	14	11	21	12	12	12	15	17	23	23	11	13	10	13	19	11-14	16	12
5930	NP	18	14	20	30	13	9	26	11	11	10	14	24	21	25	10	11	12	13	17	13-14	16	11
5919	CP	21	13	19	29	14	11	21	13	12	12	15	17	23	24	11	13	10	13	17	11-14	16	12
5922	NP	18	13	19	29	14	11	22	13	12	11	14	17	24	24	12	13	10	13	17	11-14	14	12
5962	CP	18	13	19	29	15	11	23	13	12	12	15	16	23	24	12	13	10	13	16	11-14	15	12

ID	Pop	DYS576	DYS389I	DYS448	DYS389II	DYS19	DYS391	DYS481	DYS549	DYS533	DYS438	DYS437	DYS570	DYS635	DYS390	DYS439	DYS392	DYS643	DYS393	DYS458	DYS385	DYS456	GATA H4
5943	NP	17	13	20	29	14	10	23	13	11	9	15	18	23	23	13	11	10	12	17	14-15	15	12
5956	NP	19	13	20	29	14	11	22	12	12	12	15	18	23	24	11	13	10	13	18	11-15	18	12
5960	NP	18	13	19	29	14	10	22	13	12	12	15	17	23	24	12	13	10	13	16	11-14	15	11
5969	NP	18	13	20	29	14	11	22	15	12	12	14	17	24	24	12	13	11	13	16	11-14	15	11
5959	NP	17	11	19	26	14	11	20	13	12	12	15	18	24	24	14	13	10	13	17	11-14	17	12
5973	NP	18	14	21	32	15	10	21	12	11	10	16	18	21	22	13	11	11	14	19	13-13	13	11
5982	NP	17	13	19	29	14	11	23	12	12	11	14	17	23	24	11	13	10	13	17	12-14	15	12
5987	SP	16	12	19	28	15	10	23	13	12	9	16	17	21	24	11	11	9	13	16	13-17	13	11
5990	NP	14	12	20	28	14	10	25	12	11	10	16	19	21	23	11	11	12	13	15	13-15	15	11
6087	NP	19	13	19	29	14	11	22	12	12	12	15	17	23	23	12	13	10	13	17	11-14	15	12
5998	CP	17	12	21	29	15	10	21	14	9	11	16	18	20	22	11	11	11	14	16	14-14	15	11
5993	NP	20	13	19	29	14	10	23	12	12	13	16	16	24	24	12	13	10	13	16	11-14	15	12
6037	NP	16	13	20	28	14	10	23	12	12	12	15	17	23	23	11	13	10	13	17	11-14	16	12
5809	NP	18	13	22	29	13	10	24	12	12	9	15	15	23	26	11	11	11	12	20	13-14	15	11
6008	NP	20	13	19	30	14	11	22	14	12	12	15	18	23	24	12	13	10	13	17	11-11	16	12
6011	SP	18	13	20	30	15	11	22	13	12	12	15	18	23	24	11	13	10	14	17	11-14	14	12
6040	NP	17	13	20	29	14	10	23	13	11	9	15	18	23	24	12	11	10	12	17	14-15	17	12
6043	CP	18	14	20	30	13	9	29	12	11	10	14	22	21	24	10	11	12	13	18	13-15	15	12
6049	SP	18	13	20	29	13	9	26	11	11	10	14	23	21	24	10	11	12	13	19	13-14	16	12
6055	NP	18	13	19	32	13	10	24	12	11	10	14	19	24	24	13	11	12	13	17	16-16	15	11
6073	NP	17	13	19	29	14	10	22	12	12	12	14	19	23	24	11	13	10	13	16	11-15	16	12
6071	NP	17	13	20	29	14	10	22	13	13	12	15	17	23	24	12	14	10	13	18	11-14	16	12

ID	Pop	DYS576	DYS389I	DYS448	DYS389II	DYS19	DYS391	DYS481	DYS549	DYS533	DYS438	DYS437	DYS570	DYS635	DYS390	DYS439	DYS392	DYS643	DYS393	DYS458	DYS385	DYS456	GATA H4
6080	CP	17	13	18	29	15	11	21	12	12	12	14	18	23	24	12	13	10	13	17	11-13	17	11
5402	NP	16	13	18	29	14	11	23	13	11	12	14	16	23	24	12	13	10	13	17	11-15	16	11
6122	NP	16	13	21	29	15	10	23	13	12	9	15	17	23	23	11	11	12	12	16	13-16	15	12
6102	NP	17	14	20	30	13	9	27	11	11	10	14	22	22	24	10	11	12	13	18	13-14	15	12
6117	NP	19	13	20	31	16	11	23	12	12	11	14	20	23	25	10	11	10	13	17	11-14	15	12
6116	NP	19	13	18	29	14	10	22	12	12	12	14	17	23	24	12	13	10	13	16	12-14	15	13
6123	NP	17	12	20	29	14	12	22	13	12	12	15	17	23	24	11	13	10	15	16	11-15	15	12
6133	NP	18	14	18	30	14	10	22	12	12	12	14	17	23	24	11	13	10	13	17	11-14	16	11
CC1	CP	19	13	19	30	15	12	21	13	12	12	15	16	23	24	12	14	10	13	18	11-15	16	12
CC2	CP	17	13	21	29	15	10	23	12	11	9	15	17	21	22	13	11	8	12	15	13-15	15	11
CC4	CP	17	13	21	29	15	10	23	12	11	9	15	17	21	22	12	11	8	12	15	13-15	15	11
CC6	CP	18	12	19	29	15	10	23	13	11	12	15	18	23	24	12	13	11	13	18	11-14	16	11
CC7	CP	17	13	19	29	14	11	22	13	12	12	15	17	23	24	12	13	10	13	17	12-14	16	12
CC8	CP	21	13	19	29	14	11	22	12	13	12	15	17	23	24	12	13	10	13	18	11-14	15	11
CC9	CP	14	12	22	29	15	10	21	12	10	10	16	17	21	21	11	11	12	14	17	13-16	16	11
CC10	CP	19	13	20	29	15	10	23	12	12	12	15	17	23	24	12	13	10	13	17	11-15	15	13
CC11	CP	18	13	19	30	14	10	22	12	12	12	15	18	23	24	11	13	11	14	18	11-14	15	12
CC12	CP	20	12	18	28	14	11	21	14	13	12	14	17	23	23	12	14	10	13	18	11-14	15	11
CC13	CP	17	13	21	29	15	10	23	12	11	9	15	17	21	22	12	11	8	12	15	13-15	15	11
CC14	CP	18	13	19	29	14	11	22	13	13	12	15	17	23	24	12	13	10	13	17	12-13	16	12
CC15	CP	19	13	21	28	16	10	23	10	12	10	14	18	22	23	12	11	13	13	17	12-12	15	12
CC16	CP	19	13	19	29	14	11	22	12	12	12	15	19	23	24	12	13	11	12	16	11-14	15	13

ID	Pop	DYS576	DYS389I	DYS448	DYS389II	DYS19	DYS391	DYS481	DYS549	DYS533	DYS438	DYS437	DYS570	DYS635	DYS390	DYS439	DYS392	DYS643	DYS393	DYS458	DYS385	DYS456	GATA H4
CC17	CP	21	13	19	29	14	11	22	12	12	12	15	15	23	26	12	13	11	12	16	11-14	15	12
CC18	CP	18	14	20	30	13	9	27	11	11	10	14	22	21	25	10	11	11	13	18	14-14	16	12
CC19	CP	20	14	19	30	13	10	28	12	11	10	14	23	22	24	9	11	12	13	18	13-14	17	12
CC20	CP	17	13	20	28	16	11	22	10	12	10	15	19	21	22	12	11	12	13	17	11-12	14	12
CC21	CP	17	13	19	29	15	10	23	13	12	12	14	17	23	24	12	13	9	13	17	11-14	15	12
CC22	CP	21	13	19	29	14	11	22	12	13	12	15	17	23	24	12	13	10	13	18	11-14	15	11
CC24	CP	18	13	20	29	13	9	27	12	12	10	14	21	21	24	10	11	12	13	18	13-14	16	12
CC26	CP	14	13	19	29	14	10	22	12	12	12	14	18	23	25	11	13	10	14	17	11-14	16	11
CC34	CP	18	13	20	29	13	9	28	11	11	10	14	22	21	24	10	11	12	13	18	13-14	16	12
CC45	CP	19	13	19	30	13	10	27	13	11	9	13	19	22	25	13	14	11	12	18	15-17	15	11
CC50	CP	18	13	19	30	14	10	21	13	11	12	15	19	23	24	13	13	10	12	16	11-14	12	11
CC53	CP	20	13	21	31	13	11	22	12	12	10	14	19	23	23	12	11	11	13	15	16-17	18	12
CC55	CP	19	13	20	29	14	10	28	13	12	10	14	19	20	24	12	11	9	12	21.2	12-19	15	11
CC111	CP	18	13	19	29	14	11	23	13	12	12	15	17	23	24	12	13	10	13	18	12-14	17	11
CC115	CP	16	13	19	29	14	11	22	13	12	12	15	17	23	24	12	13	11	13	17	11-14	15	13
CC116	CP	17	13	19	29	14	11	22	12	13	12	15	17	23	24	12	13	10	13	17	11-14	16	11
CC117	CP	18	13	19	28	14	11	22	13	12	12	15	16	24	24	12	13	11	13	17	11-13	15	12
CC118	CP	17	14	20	30	13	9	27	11	11	10	14	22	21	25	10	11	11	13	18	14-14	16	12
CC119	CP	17	13	19	30	14	11	22	13	12	12	15	17	23	24	12	14	10	13	18	12-14	15	11
CC120	CP	16	14	19	29	14	10	22	13	12	12	15	19	23	24	12	14	11	13	18	11-14	16	12
CC124	CP	20	13	18	29	14	10	22	12	12	12	15	17	23	24	12	13	10	13	17	12-14	15	12
CC125	CP	19	14	19	30	14	10	21	11	12	12	14	18	23	23	11	13	10	13	17	11-15	16	12

ID	Pop	DYS576	DYS389I	DYS448	DYS389II	DYS19	DYS391	DYS481	DYS549	DYS533	DYS438	DYS437	DYS570	DYS635	DYS390	DYS439	DYS392	DYS643	DYS393	DYS458	DYS385	DYS456	GATA H4
CC126	CP	18	13	19	29	15	10	22	14	12	12	15	17	23	23	12	13	10	13	17	12-14	16	12
CC127	CP	18	13	20	30	15	10	25	11	10	10	15	18	19	24	11	12	12	15	16	14-14	14	11
CC128	CP	18	13	21	30	14	10	25	12	11	10	14	18	22	23	12	11	9	12	17.2	13-15	15	11
CC132	CP	16	13	19	29	15	11	20	11	11	10	14	20	22	23	10	14	11	14	17	11-14	15	12
CC133	CP	16	13	20	30	13	10	22	12	12	10	14	20	21	24	13	11	12	13	16	16-21	17	11
CC134	CP	16	13	20	30	13	10	22	12	12	10	14	19	24	24	12	11	12	13	15	16-18	16	11
CC138	CP	17	13	19	28	14	10	22	13	12	12	15	17	23	25	12	13	10	13	17	11-14	15	12
CC139	CP	17	13	20	30	14	10	25	13	11	10	14	17	20	22	11	11	9	12	17.2	13-19	15	11
CC140	CP	18	13	20	29	15	10	25	11	13	10	14	20	21	23	12	12	12	15	17	15-17	14	11
CC141	CP	18	13	19	29	14	11	21	12	12	13	15	17	23	23	12	13	10	12	16	11-14	15	13
CC145	CP	17	13	19	29	15	10	23	13	13	12	15	17	23	24	12	13	9	13	17	11-14	15	12
CC146	CP	19	13	19	29	14	11	22	12	12	12	15	19	23	23	13	13	10	13	17	11-14	15	11
CC147	CP	18	12	20	29	14	10	26	12	11	10	16	19	22	22	11	11	12	13	15	13-14	14	11
CC148	CP	20	13	18	29	14	10	22	12	12	12	15	17	23	24	12	13	10	13	17	12-14	15	12
CC150	CP	19	14	18	30	14	11	22	13	12	12	14	18	23	24	13	13	10	13	17	11-14	16	11
CC151	CP	19	13	19	29	14	11	23	12	12	12	15	17	23	24	12	13	10	13	17	11-14	16	12
CC152	CP	18	13	18	30	14	10	22	14	12	12	15	18	23	24	12	13	10	13	18	11-15	16	12
CC153	CP	18	12	19	27	15	10	23	13	10	9	14	18	20	22	11	14	9	13	15	14-16	15	10
CC154	CP	17	14	20	32	15	10	27	12	12	10	14	19	22	23	11	12	13	14	15	15-16	15	11
CC155	CP	16	12	20	28	14	10	24	12	11	10	15	21	21	22	11	11	12	13	15	14-14	14	11
CC156	CP	20	13	19	29	14	11	22	13	12	13	15	18	23	24	11	13	10	13	17	11-14	15	12
CC157	CP	19	13	19	29	14	11	22	12	12	12	15	19	23	24	12	13	11	12	16	11-14	15	13

ID	Pop	DYS576	DYS389I	DYS448	DYS389II	DYS19	DYS391	DYS481	DYS549	DYS533	DYS438	DYS437	DYS570	DYS635	DYS390	DYS439	DYS392	DYS643	DYS393	DYS458	DYS385	DYS456	GATA H4
CC158	CP	20	13	20	31	13	10	25	12	11	10	14	18	23	24	12	11	12	12	18	16-16	15	12
CC159	CP	18	13	19	29	14	11	22	13	13	12	15	17	23	24	12	13	10	13	17	12-13	16	12
CC160	CP	19	13	19	30	15	11	22	14	12	12	15	19	23	25	12	13	10	13	15	11-14	16	12
CC161	CP	16	14	20	31	15	10	23	12	11	9	14	15	22	22	12	11	8	13	16	13-14	15	11
CC163	CP	19	13	20	29	15	10	23	12	12	12	15	18	23	24	12	13	10	13	17	11-15	15	13
CC164	CP	20	13	20	31	13	10	25	12	11	10	14	18	23	24	12	11	12	12	18	16-16	15	12
SH02	SP	19	13	19	29	14	11	23	13	12	12	15	16	24	24	11	13	10	13	16	11-14	15	12
SH05	SP	21	13	20	29	14	10	25	11	11	10	14	19	21	23	12	11	9	12	18.2	12-15	16	11
SC1	SP	20	13	19	29	14	12	22	13	12	12	15	16	23	24	12	13	10	13	17	11-14	15	12
SC2	SP	18	13	19	29	14	11	21	13	12	12	15	17	23	24	12	13	10	13	17	12-14	16	12
SC3	SP	19	13	20	30	14	10	22	12	12	12	15	17	23	25	12	13	10	13	18	11-14	15	12
SC4	SP	18	13	19	29	14	11	22	13	12	12	15	17	23	24	12	13	10	13	17	12-14	16	12
SC5	SP	18	12	22	29	15	10	21	12	9	10	15	18	21	22	11	11	12	13	17	12-13	14	11
SC7	SP	19	13	19	29	14	11	23	12	12	12	15	17	23	24	12	13	10	13	17	12-14	15	12
SC8	SP	18	14	18	31	14	11	22	13	13	12	15	19	23	24	13	13	10	13	22	11-14	15	12
SC9	SP	17	13	20	31	17	10	23	12	13	11	14	18	23	25	10	11	10	14	15	11-14	15	13
SC11	SP	17	13	19	30	14	11	21	12	12	12	15	17	23	23	12	13	10	13	18	11-14	16	12
SC13	SP	18	12	21	28	16	10	23	13	9	11	16	17	20	22	11	11	11	14	16	14-16	15	13
SC14	SP	17	13	19	29	14	11	22	11	12	12	15	16	23	24	11	13	9	14	20	11-14	18	12
SC15	SP	20	14	19	30	14	11	22	12	12	12	15	18	23	23	12	13	10	13	18	11-14	16	12
SC18	SP	16	13	18	29	15	11	23	12	12	12	15	17	24	23	12	13	11	13	17	11-13	16	12
SC19	SP	16	14	19	30	14	10	22	12	12	12	15	17	23	23	12	13	10	13	19	11-11	16	12

ID	Pop	DYS576	DYS389I	DYS448	DYS389II	DYS19	DYS391	DYS481	DYS549	DYS533	DYS438	DYS437	DYS570	DYS635	DYS390	DYS439	DYS392	DYS643	DYS393	DYS458	DYS385	DYS456	GATA H4
SC20	SP	18	14	20	30	14	11	23	13	11	12	15	17	23	24	13	13	10	13	18	11-14	15	12
SC23	SP	17	13	18	29	14	11	22	13	12	12	14	17	23	24	12	13	10	13	17	11-14	16	12
SC24	SP	15	13	21	30	13	10	23	12	12	10	14	18	23	25	12	11	13	14	16	16-17	17	12
SC25	SP	18	13	19	29	14	11	22	13	12	12	15	18	23	25	12	14	10	13	18	11-14	15	12
SC27	SP	21	13	20	29	14	10	25	11	11	10	14	19	21	23	12	11	9	12	18.2	12-15	16	11
SC28	SP	18	12	22	29	15	10	21	12	9	10	15	18	21	22	11	11	12	13	17	12-13	14	11
SC29	SP	18	13	20	31	13	11	26	12	10	10	14	20	22	24	12	11	12	13	18	17-17	17	11
SC31	SP	18	14	21	30	13	9	26	11	12	10	14	23	21	24	10	11	12	13	19	14-14	16	12
SC32	SP	17	13	19	29	14	11	22	12	12	12	15	16	23	24	12	13	9	14	19	11-14	18	12
SC34	SP	16	14	21	30	15	9	23	12	12	9	14	18	22	23	12	11	10	12	14	13-15	16	12
SC35	SP	19	13	20	29	14	10	24	11	12	9	14	16	23	23	10	11	10	12	14	13-16	14	11
SC38	SP	16	13	19	29	14	10	22	13	12	12	15	17	23	23	12	13	10	13	19	11-12	16	12
SC39	SP	18	13	19	29	14	11	22	13	12	12	15	15	23	24	12	14	10	13	17	11-14	15	12
SC45	SP	18	14	18	30	14	11	22	14	12	12	14	18	23	24	11	13	11	13	18	11-14	16	11
SC46	SP	15	13	21	30	13	10	23	12	12	10	14	18	23	25	12	11	13	14	16	16-17	17	12
SC47	SP	19	13	19	29	13	11	22	13	12	12	15	18	23	24	12	13	10	13	17	12-14	16	11
SC48	SP	18	14	18	30	14	11	22	14	12	12	14	17	23	25	12	13	10	13	18	11-15	16	11
SC49	SP	16	13	19	30	15	10	22	13	12	9	14	19	21	23	11	13	10	13	18	14-16	15	11
SC50	SP	18	13	18	29	15	11	21	12	12	12	14	18	23	24	12	13	10	13	18	11-13	18	11
SC51	SP	17	14	20	31	14	10	22	12	11	9	15	15	23	23	10	11	11	12	16	15-15	18	11
SC52	SP	16	12	19	28	15	10	23	13	12	9	16	17	21	24	11	11	9	13	16	13-17	13	11
SC53	SP	17	13	20	31	17	10	23	12	14	11	14	18	23	25	10	11	10	14	15	11-14	15	13

ID	Pop	DYS576	DYS389I	DYS448	DYS389II	DYS19	DYS391	DYS481	DYS549	DYS533	DYS438	DYS437	DYS570	DYS635	DYS390	DYS439	DYS392	DYS643	DYS393	DYS458	DYS385	DYS456	GATA H4
SC55	SP	17	14	20	31	14	10	22	12	11	9	15	15	23	23	10	11	11	12	16	15-15	18	11
SC57	SP	18	13	19	29	14	10	22	12	11	12	15	16	23	24	12	13	10	14	18	11-14	15	12
SC58	SP	18	13	19	29	14	11	22	12	12	12	15	17	23	24	11	13	11	13	17	11-15	16	12
SC60	SP	19	13	19	29	14	11	23	13	12	12	15	16	24	24	11	13	10	13	16	11-14	15	12
SC61	SP	17	13	20	29	14	11	24	14	12	12	15	17	23	24	12	13	10	13	17	11-15	15	12
SC62	SP	17	14	20	32	14	10	22	12	11	9	15	15	23	23	10	11	11	12	16	15-15	18	11
SC63	SP	17	13	19	29	14	11	22	13	12	12	14	17	24	24	12	13	11	14	17	11-14	15	12
SC66	SP	20	13	19	29	14	12	22	13	12	12	15	16	23	24	12	13	10	13	17	11-14	15	12

Supplementary Table 2 – 250 haplotypes obtained from the Portuguese population using the PowerPlex® Y23 System. NP: Northern Portugal; CP: Central Portugal; SP: Southern Portugal Az: Azores; Ma: Madeira.

ID	PowerPlex® Y23	Yfiler™	SNP
CC1	R1b (100%)	R1b (100%)	R1b1
CC2	G2a (89.5%)	J2a1 (62.3%)	J2
CC4	G2a (91.5%)	J2a1 (54.3%)	J2
CC6	R1b (100%)	R1b (100%)	R1b1
CC7	R1b (100%)	R1b (100%)	R1b1
CC8	R1b (100%)	R1b (100%)	R1b1
CC10	R1b (100%)	R1b (100%)	R1b1
CC11	R1b (100%)	R1b (100%)	R1b1
CC13	G2a (91.5%)	J2a1 (54.3%)	J2
CC14	R1b (100%)	R1b (100%)	R1b1
CC15	I2a1 (94.8%)	I2a1 (100%)	I2a1a
CC16	R1b (100%)	R1b (100%)	R1b1
CC17	R1b (100%)	R1b (100%)	R1b1
CC18	E1b1b (100%)	E1b1b (100%)	E1b1b
CC19	E1b1b (100%)	E1b1b (100%)	E1b1b
CC22	R1b (100%)	R1b (100%)	R1b1
CC24	E1b1b (100%)	E1b1b (100%)	E1b1b
CC34	E1b1b (100%)	E1b1b (100%)	E1b1b
CC45	Q (100%)	Q (98.2%)	P(xR1)
CC53	E1b1b (100%)	E1b1b (100%)	E1b1b
CC55	J1 (100%)	J1 (99.6%)	J*(xJ1a,2)
CC111	R1b (100%)	R1b (100%)	R1
CC115	R1b (100%)	R1b (100%)	R1b1
CC116	R1b (100%)	R1b (100%)	R1b1
CC117	R1b (100%)	R1b (100%)	R1b1
CC118	E1b1b (100%)	E1b1b (100%)	E1b1b
CC119	R1b (100%)	R1b (100%)	R1b1
CC120	R1b (100%)	R1b (100%)	R1b1
CC124	R1b (100%)	R1b (100%)	R1b1
CC125	R1b (100%)	R1b (100%)	R1b1
CC126	R1b (100%)	R1b (100%)	R1b1
CC128	J1 (100%)	J1 (99.2%)	J*(xJ1a,2)
CC132	Q (61.1%)	N (100%)	N3
CC133	E1b1b (100%)	E1b1b (100%)	E1b1b
CC134	E1b1b (100%)	E1b1b (100%)	E1b1b
CC138	R1b (100%)	R1b (100%)	R1b1
CC139	J1 (100%)	J1 (99.8%)	J*(xJ1a,2)
CC140	I2b1 (100%)	I2b1 (100%)	I*(xI1b1b)
CC141	R1b (100%)	R1b (100%)	R1b1

ID	PowerPlex® Y23	Yfiler™	SNP
CC145	R1b (100%)	R1b (100%)	R1b1
CC147	I1 (100%)	I1 (99.7%)	I*(xI1b1b)
CC148	R1b (100%)	R1b (100%)	R1b1
CC150	R1b (100%)	R1b (100%)	R1b1
CC151	R1b (100%)	R1b (100%)	R1b1
CC152	R1b (100%)	R1b (100%)	R1b1
CC153	L (68.7%)	T (100%)	T1a
CC154	I2b1 (100%)	I2b1 (100%)	I*(xI1b1b)
CC156	R1b (100%)	R1b (100%)	R1b1
CC157	R1b (100%)	R1b (100%)	R1b1
CC158	E1b1b (98.2%)	E1b1b (99.8%)	E1b1b
CC159	R1b (100%)	R1b (100%)	R1b1
CC160	R1b (100%)	R1b (100%)	R1b1
SH02	R1b (100%)	R1b (100%)	R1b1
SH05	J1 (100%)	J1 (97.9%)	J*(xJ1a,2)
SC1	R1b (100%)	R1b (100%)	R1b1
SC2	R1b (100%)	R1b (100%)	R1b1
SC3	R1b (100%)	R1b (100%)	R1b1
SC4	R1b (100%)	R1b (100%)	R1b1
SC5	G2a (100%)	G2a (99.5%)	G
SC7	R1b (100%)	R1b (100%)	R1b1
SC8	R1b (100%)	R1b (100%)	R1a
SC9	R1a (100%)	R1a (100%)	R1b1
SC11	R1b (100%)	R1b (100%)	K
SC13	G2a (100%)	G2a (100%)	G
SC14	R1b (100%)	R1b (100%)	R1b1
SC15	R1b (100%)	R1b (100%)	R1b1
SC18	R1b (100%)	R1b (100%)	R1b1
SC19	R1b (100%)	R1b (100%)	R1b1
SC20	R1b (100%)	R1b (100%)	R1b1
SC23	R1b (100%)	R1b (100%)	R1b1
SC24	E1b1b (100%)	E1b1b (100%)	E1b1b
SC25	R1b (100%)	R1b (100%)	R1b1
SC27	J1 (100%)	J1 (97.9%)	J*(xJ1a,2)
SC28	G2a (100%)	G2a (99.5%)	G
SC29	E1b1b (100%)	E1b1b (100%)	E1b1b
SC31	E1b1b (100%)	E1b1b (100%)	E1b1b
SC32	R1b (100%)	R1b (100%)	R1b1
SC34	J1 (70.7%)	J2a1h (100%)	J2

ID	PowerPlex® Y23	Yfiler™	SNP
SC35	J1 (99.2%)	J2a1b (72.2%)	P(xR1)
SC38	R1b (100%)	R1b (100%)	R1b1
SC39	R1b (100%)	R1b (100%)	R1b1
SC45	R1b (100%)	R1b (100%)	R1b1
SC46	E1b1b (100%)	E1b1b (100%)	E1b1b
SC47	R1b (100%)	R1b (100%)	R1b1
SC48	R1b (100%)	R1b (100%)	R1b1
SC49	R1b (70%)	T (100%)	K
SC50	R1b (100%)	R1b (100%)	R1b1
SC60	R1b (100%)	R1b (100%)	R1b1
SC61	R1b (100%)	R1b (100%)	R1
SC63	R1b (100%)	R1b (100%)	R1b1
SC66	R1b (100%)	R1b (100%)	R1b1

Supplementary Table 3 – 91 samples, their haplogroup prevision made by Haplogroup Predictor (with probability) based on 23 Y-STRs from PowerPlex® Y23 and on 17 Y-STRs from Yfiler™, and their actual haplogroups defined by Y-SNPs.