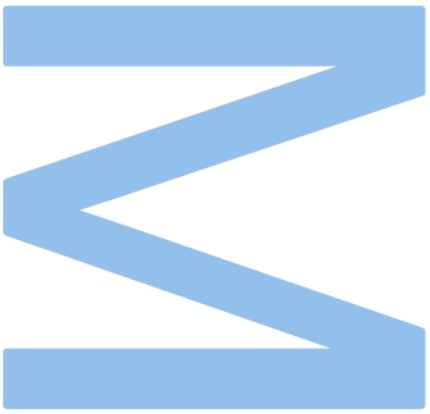
Genetic diversity of human specific X-STRs in humans and chimpanzees: a comparative study and proposal for nomenclature

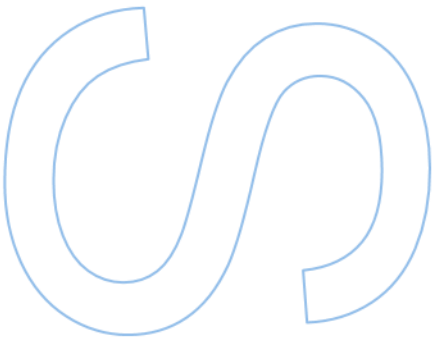
Bruno Miguel Alves Mendes Teixeira Moutinho
MSc in Forensic Genetics
Department of Biology
2021/2022

Supervisor
Iva Gomes, Associate researcher, IPATIMUP/i3S

Co-supervisor
Leonor Gusmão, State University of Rio de Janeiro- UERJ


U.PORTO
FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO





Sworn Statement

I, Bruno Miguel Alves Mendes Teixeira Moutinho, enrolled in the Master Degree in Genetic Forensics at the Faculty of Sciences of the University of Porto hereby declare, in accordance with the provisions of paragraph a) of Article 14 of the Code of Ethical Conduct of the University of Porto, that the content of this dissertation reflects perspectives, research work and my own interpretations at the time of its submission.

By submitting this dissertation, I also declare that it contains the results of my own research work and contributions that have not been previously submitted to this or any other institution.

I further declare that all references to other authors fully comply with the rules of attribution and are referenced in the text by citation and identified in the bibliographic references section. This dissertation does not include any content whose reproduction is protected by copyright laws.

I am aware that the practice of plagiarism and self-plagiarism constitute a form of academic offense.

Bruno Miguel Alves Mendes Teixeira Moutinho

29 de outubro, 2022

Acknowledgements

Primeiramente, quero agradecer à minha orientadora Iva Gomes. Obrigado por ter aceitado orientar-me no projeto e dissertação de mestrado. Apesar de todas as dificuldades durante este ano, a sua paciência, disponibilidade e ajuda estiveram sempre presentes e foram cruciais. Agradeço também todo o conhecimento ao qual me proporcionou a nível, tanto teórico como prático. Toda esta ajuda, auxiliou na construção desta dissertação, mas também o meu desenvolvimento como investigador. Um enorme obrigado.

Quero agradecer à minha coorientadora Leonor Gusmão. Apesar da distância, as suas opiniões e dicas foram de enorme ajuda para a escrita desta dissertação.

Ao Professor António Amorim, grupo leader da Genética Populacional e Evolução do i3s, por me ter recebido e aceite no seu grupo para a realização do projeto de mestrado.

A todo o grupo de Genética Populacional e Evolução, que esteve sempre disponível para ajudar. Não podia ter encontrado melhor grupo de pessoas. Um grande obrigado.

À minha melhor amiga, Mafalda. Obrigado por todo o apoio, por sempre ouvires os meus desabafos e as minhas patetices e por sempre me fazeres sorrir quando é necessário. Por todas as “pep talks”, encorajamento e força que me deste. Sem toda a tua ajuda, esta experiência tinha sido um pouco mais difícil, e por isso um obrigado do fundo do coração.

A um grupo de pessoas muito especiais ao qual me aproximei bastante durante a pandemia. Sofia, Tomás, Pinto, Paulo, Gustavo e Alex, obrigado por estarem quase sempre presentes nesta jornada. Obrigado por me distraírem um pouco, por todos os risos e “loucuras” e me apoiarem sempre que necessário. A vossa ajuda e força foi incondicional no desenvolvimento desta dissertação. Um enorme obrigado. Não podia ter conhecido melhores pessoas.

Ao meu irmão Nuno. Obrigado por sempre me mostrares como enfrentar os meus desafios e ser corajoso. Não podia ter pedido melhor irmão. A tua força e apoio foram incondicionais para o término desta grande etapa. Obrigado por nunca me deixares desistir dos meus sonhos. Obrigado por tudo.

E por último, mas o mais importante, aos meus pais. Sem o vosso esforço e trabalho, eu não estaria a vivenciar esta grande experiência, por isso um obrigado do fundo do coração. Vocês são um pilar de apoio, força e uma grande inspiração desde sempre.

Estou orgulhoso de poder chamar-vos de pais. Estou-vos eternamente grato por tudo. Adoro-vos.

Resumo

Em procedimentos forenses, a análise de marcadores de DNA na identificação humana é o processo estabelecido devido às características apresentadas por este ácido nucleico. A utilização de marcadores genéticos como método tem como base a análise de repetições variáveis de nucleótidos em tandem (VNTRs) encontradas no DNA extraído de uma amostra biológica e a sua comparação com uma amostra referência, uma vez que a *match probability* entre dois indivíduos é baixa. Dentro da categoria dos VNTRs, os Short Tandem Repeats (STRs), nomeadamente os localizados nos cromossomas autossómicos, são vastamente utilizados na área forense uma vez que são altamente informativos além de apresentarem um grande número e disponibilidade de marcadores. Apesar destas características, os STRs autossomais apresentam limitações, em casos particulares, e por isso a sua informação pode ser complementada por STRs localizados nos cromossomas sexuais, nomeadamente no cromossoma X. Devido aos processos evolutivos e as características de herança genética do cromossoma X, este reúne propriedades peculiares que permite a sua utilização na área forense, sendo extremamente úteis em casos de análise de parentesco e na análise de restos humanos (por exemplo em desastres naturais ou não naturais) e na análise de manchas. Apesar da sua utilidade, existem problemas que podem impedir a sua utilização, em particular, a falta de bases de dados populacionais estabelecidas e possíveis problemas de nomenclaturas dos X-STRs devido, em parte, à falta de estudos minuciosos que englobem distintos grupos populacionais. A importância de estabelecer uma nomenclatura global e inclusiva é altíssima, uma vez que irá facilitar a comunicação de resultados obtidos e a sua precisão entre laboratórios. Todos estes problemas já foram realçados em alguns estudos científicos, no entanto, ainda existe uma falha neste tema e em particular para os X-STRs.

De acordo com as regras existentes e estabelecidas para a atribuição da nomenclatura alélica de um STR, o procedimento mais correto passa pela análise da sequência do DNA de um determinado locus em distintos grupos populacionais (por exemplo, Europeus, Asiáticos e Africanos) de modo a englobar a maior variação possível para esse STR. Além disso, o alinhamento das sequências repetitivas com as sequências do genoma do chimpanzé, sendo que este é o parente mais próximo do ser humano, apresenta um complemento à atribuição de uma nomenclatura de um STR uma vez que tiveram menos tempo de evolução e por isso estão mais próximos da estrutura ancestral.

O trabalho apresentado nesta dissertação teve duas grandes finalidades: a categorização da diversidade genética dos X-STRs em humanos, em particular nos três principais grupos populacionais: Europeus, Africanos e Asiáticos e a comparação com o genoma dos chimpanzés para vários marcadores integrados no kit comercial mais utilizado “The Investigator X-12 QS” (Qiagen) tais como DXS10074, DXS10134, DXS10135, DXS10146 e DXS10148. Para a realização do trabalho prático, amostras de portugueses foram utilizadas como representantes do grupo populacional Europeu, amostras de Macaenses foram utilizadas como representantes do grupo Asiático e, por fim, foram reunidos dados de trabalhos publicados e na literatura maioritariamente para o grupo populacional Africano.

Nos resultados obtidos neste trabalho, foram observadas várias inconsistências entre as nomenclaturas encontradas através da análise de resultados de sequenciação das amostras neste estudo, e as previamente estabelecidas, especialmente, para os marcadores DXS10135 e DXS10148, sendo que uma revisão da nomenclatura destes marcadores foi proposta. Este fato realça a importância de que para estabelecer uma nomenclatura inclusiva, várias análises a nível populacional têm de ser realizadas. Os alinhamentos com sequências de chimpanzé também mostraram um pouco das diferenças de evolução que os marcadores estudados sofreram em relação à sua estrutura ancestral e estabilidade.

Palavras-chave: DNA, forense, STRs, cromossoma X, X-STRs, nomenclatura, chimpanzé, humano, marcadores genéticos

Abstract

In forensic procedures, analysis of DNA markers is the established methodology due to the characteristics that this nucleic acid presents. This methodology is based on the analysis of variable nucleotide tandem repeats (VNTRs) found in the DNA extracted from a biological sample and its comparison to a reference sample since the match probability between two individuals is low. Included on the VNTRs category, Short Tandem Repeats (STRs), namely present on the autosomal chromosomes are vastly used in forensic studies due to the high quantity, high polymorphism and having a big availability in terms of markers throughout the genome. Even though, autosomal genetic markers possess some limitations and its information can be complemented by STRs located on the sex chromosomes, such as the X-chromosome. Due to the evolutionary processes and history that the X-chromosome underwent, it gathers peculiar characteristics that allow the usage in the forensic field, being extremely useful in parenthood cases, analyses of human remains in mass destructions, and analyses of stains. Apart from these good qualities, X-chromosome STRs (X-STRs) have some negative points that can be a barrier to its usage in some cases. Lack of an established database and nomenclature due to, in part, the lack of studies that analyze various population groups are somewhat prejudicial to the usage of X-chromosome genetic markers in the area. The importance of establishing a global and inclusive nomenclature is high, since it will facilitate the reports obtained in studies making these results more precise. These problems have been emphasized in many studies about X-STRs. The alignment of these human repetitive sequences with the sequences of the chimpanzee (*Pan troglodytes*), the closest living relative of human species, presents a solution on solving the nomenclature issues of X-STRs since the chimpanzee sequences had less time to evolve therefore it is closer to the ancestral structure.

The work presented in this dissertation has two big finalities: the categorization of genetic diversity of X-STRs in humans, in three population groups, namely Europeans, Asians and Africans, and the comparison with the chimpanzee genome, for various markers integrated in the most used commercial kit “The Investigator X-12 QS” (QIAGEN) such as DXS10074, DXS10134, DXS10135, DXS10146 and DXS10148.

To perform this practical work, samples from the Portuguese population were used as representation for the Europeans, samples from Macanese population were used as representation for Asians and for the African population group, data was mostly gathered from the literature.

The sequencing results obtained in this work revealed several discrepancies between some of the nomenclatures of the X-STRs analyzed when compared with the results and the previously established nomenclatures. Most inconsistencies were encountered for the markers DXS10135 and DXS10148. A revision of the established nomenclature for these two markers is proposed. These findings highlight the importance of performing studies on several, different population groups for X chromosome genetic markers. The alignment results showed the differences between both species that the markers analyzed underwent and its potential ancestral structure and stability that support the inference on the most accurate nomenclature.

Keywords: DNA, forensics, STRs, X-chromosome, X-STRs, nomenclature, chimpanzee, human, genetic markers

Table of contents

List of Tables

List of Figures

List of Abbreviations

CE	Capillary electrophoresis
DNA	Deoxyribonucleic acid
INDELS	Insertions or deletions
LD	Linkage disequilibrium
MPS	Massively Parallel Sequencing
NGS	Next Generation Sequencing
PAR	Pseudoautossomal region
PAR1	Pseudoautossomal region 1
PAR2	Pseudoautossomal region 2
PCR	Polymerase Chain Reaction
SRY	Sex-determining region Y protein
STRs	Short Tandem Repeats
VNTRs	Variable number of Tandem Repeats
XAR	X-added Region
X-Chr	X-Chromosome
XCI	X-Chromosome Inactivation
XCR	X-Conserved Region
XIST	X-Inactive Transcript
X-STRs	X-Chromosomal Short Tandem Repeats

1. Introduction

1.1. DNA in forensic genetics

The set of nucleic acid sequences encoded as deoxyribonucleic acid (DNA) molecules located in the nucleus of the cell, and in the mitochondria make up our complete genetic information - the human genome. Genetic variation is present in the form of DNA sequence or length polymorphisms. Therefore, human identification is possible due to the unique genetic profiles obtained from DNA and has consequently a central role in the forensic genetics field. Genetic variability is key for this process and the discovery of variable number of tandem repeats (VNTRs) allowed this method to take place since the match probabilities are so low that only monozygotic twins share the same DNA fingerprinting (Udogadi et al., 2020). VNTRs, also known as minisatellites, are repetitive sequences of 8 to 100 bps, and Short Tandem Repeats (STRs) are microsatellites generated by a sequence (in tandem) of copies of small DNA segments (ranging from 2 to 6 base pairs) (Jeffreys et al., 1985). Since its introduction in the early 80's by Sir Alec Jeffreys (Jeffreys et al., 1985) the analysis of DNA markers - known as DNA fingerprinting - was a breath of fresh air in forensic investigations. Up until the modern days, DNA fingerprinting is the stapled technique used in this area of study. Obviously, throughout the decades, many advancements were introduced involving discrimination power, speed and sensitivity of DNA profiling methods (Haddrill, 2021; Oberacher & Parson, 2007). This analysis relies on the usage of STR markers from the DNA extracted from the questioned biological sample to find a match with a reference sample (Dash et al., 2020). Usually, the source of genetic material for DNA analyses can be, for example, blood (or other body fluids, for example saliva), bone, hair and even soft tissue (Ziętkiewicz et al., 2011). The most informative STRs may present more than a dozen alleles and thousands of these polymorphisms have already been identified in humans. Some estimations point to the existence of about one million STRs distributed among the human genome (Marano & Fridman, 2019). Due to the vast availability of markers in the human genome, many polymorphisms are well known and the most suitable are chosen for different types of identification cases.

In the context of casework applications, DNA can help convict a guilty person but also exonerate an innocent. It can also help identify the remains of missing persons and victims of mass disasters or to analyze genetic relationships, including paternity testing

(Butler, 2015; Ziętkiewicz et al., 2011). DNA markers may allow for identification of perpetrators when biological material is transferred between the aggressor and the victim or to the crime scene itself. Another advantageous characteristic of the used of DNA in forensic investigations is its inheritance mode since it comes from both parents. In cases of missing persons or mass disaster victims, family members such as mother, father or even brothers can provide samples to help compare the genetic material obtained from the remains (Butler, 2015).

Although numerous polymorphisms are available throughout the genome that are suitable for forensic genetic applications such as autosomal, X and Y chromosomal and mitochondrial markers, this study concentrates on the study of STRs located on the X chromosome (X-chromosomal Short Tandem Repeats, X-STRs) and, therefore, the remaining introduction of this dissertation will specifically focus on its characteristics.

1.2. The X chromosome

1.2.1. Main structural characteristics

Many structural properties of the X chromosome are unique when compared to its sex chromosome equivalent, the Y chromosome, or even when compared to the autosomes. The eutherian X chromosome has three main regions: the PAR (pseudoautossomal) regions, the X-added region (XAR) that comprises an autosomal segment added at the time of the eutherian-metatherian divergence and the X conserved region (XCR) that maintained almost all ancestral genes (MN & JMA, 2018).

One of the main characteristics of the X chromosome is its highly preserved autosomal character and the fact this it is one of the most stable nuclear chromosomes possessing one of the biggest and most conserved gene arrangement and gene content across placental mammals (Brashear et al., 2021; Sangrithi & Turner, 2018). Considering structural properties, the X chromosome has low CG content, being 39%, in comparison with the genome average that is 41% (Ross et al., 2005). The X chromosome has a size of approximately 155 million base pairs, which represents 5% of the estimated human genome size (Ross et al., 2005). The analysis of sequence data performed by Ross and his coworkers in 2005 revealed that the X chromosome has a low concentration of genes and small gene length since only 1.7% of the chromosome are exons which leads to the transcription of only 33% of the chromosome (Ross et al., 2005). Another interesting characteristic linked to the X chromosome is the inactivation it undergoes in females, due to the extra dosage present, since females have two X chromosomes.

1.2.2. Origin and evolution of the sex chromosomes

The evolutionary process of the sex chromosomes originates the characteristics that make the X chromosome today so unique. Approximately, between 200 and 300 million years ago, the mammalian sex chromosomes started to differentiate from a pair of autosomes, originating the so-called proto-sex chromosomes (Ross et al., 2005). These primitive chromosomes began to differ from each other, physiologically and biologically, due to the appearance of sex determining genes, for example, the sex-determining region Y protein (SYR), on the proto-Y-chromosome (Iwase et al., 2010). In 1999, Lahn and Page proposed a model that tries to explain the evolutionary steps that the sex chromosomes underwent (Lahn & Page, 1999). This model suggests four main events, namely four inversions, on the Y chromosome that contributed to the cessation of X-Y recombination (Lahn & Page, 1999). Each inversion increased the differences among each chromosome as they evolved, i.e., each portion of the chromosomes that inverted added to the length of DNA that could no longer align and recombine. Each stratum resulting from each inversion event consists of four regions with genomic characteristics that can be organized into different groups (Furman et al., 2020). These groups will then have different ages, the oldest stratum being the first inversion, which means the first recombination suppression and so forth. In addition, the initial stratum will have a higher genetic divergence than the most recent ones (Furman et al., 2020). It is known that inversions tend to stop recombination, which supports this model of evolution, because after each inversion that happened on the Y chromosome the regions that were affected became non-recombining, evolved and diverged independently (Bachtrog & Charlesworth, 2001; Pandey et al., 2013). These events, on top of accumulated mutations on each chromosome, lead to the creation of specific (non-recombining) Y and X chromosome regions (Iwase et al., 2010). There are two regions on the sex chromosomes that escaped this lack of recombination, which are the PAR1 and PAR2 regions (pseudoautossomal regions). These locations are very interesting since, for example, the recombination rate in the major pseudoautossomal region is twenty times the genome average. When comparing both sex chromosomes, the Y chromosome lost a big portion of its genes and sequences and developed a unique pattern of repeated sequences. As for the X chromosome is the complete opposite, since this chromosome retained its autosomal character in females and it is the most stable nuclear chromosome amongst placental mammals (Schaffner, 2004).

1.2.3. The X chromosome in females and males: inheritance features

The evolution that the sex chromosomes went through not only shaped them structurally but also gave them gender specific characteristics. Males transmit the X chromosome to their female offspring unchanged (unless mutation occurs) due to the lack of recombination in meiosis, except for the PAR regions (see Figure 1). Females transmit the X chromosome as a changed piece because they have two copies of the X chromosome, which allows for full recombination (Figure 1) and therefore exchange and complete rearrangement of information occurs (Figure 1) (Diegoli, 2015; Gomes et al., 2020; Schaffner, 2004). This promotes variation in new generations, increasing diversity (Gomes et al., 2020). The second example is the imbalance of dosage of X chromosomes between males and females (Heard & Turner, 2011). This imbalance of dosage creates a need for compensation. Therefore, one of the chromosomes is inactivated in early female cell development by a process called XCI (X chromosome inactivation). This inactivation was discovered by Mary Lyon (Lyon, 1961) and is controlled by the XIST (X-inactive transcript). Xist produces a noncoding RNA that involves the targeted X chromosome, which attracts factors, and will modify the chromatin and inactivate that chromosome (Brockdorff et al., 2020).

These specific characteristics, one single copy in males and two recombining portions in females, make the X chromosome a particular interesting tool for forensic and population genetic studies (Gomes et al., 2020). Clinical genetics is another great area of study where the X chromosome has a big impact due to of its mode of inheritance and link to genetic diseases such as hemophilia, Duchenne muscular dystrophy and many more (Szibor, 2007). Since these diseases are usually linked to recessive genes, males will express the traits and symptoms due to the X-linked recessive allele that is passed on by the mother. On that note, if a male has female descendants, these will 100% carry that allele, even though, they might not express the phenotype of that disease. These female descendants will pass on the gene with 50% chance to their offspring. Having knowledge of this transmission mode and contexts, the study of X-chromosomal markers is really useful for clinic genetics, forensic and kinship analysis (Szibor, 2007).

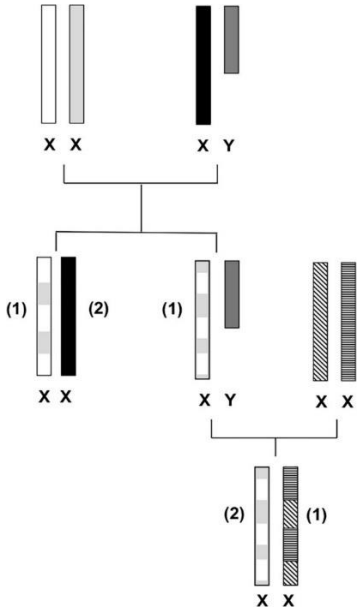


Figure 1. Inheritance of the X chromosome, withdrawn from "Twenty years later: A comprehensive review of the X chromosome use in Forensic Genetics" (Gomes et al., 2020). After recombination, females transmit a new, reshuffled chromosome (1) while males a transmit copy of the X-chromosome as an unchanged block to female descendants (2).

1.3. X-chromosomal Short Tandem Repeats

As time passed, improvements of methods have taken place and advances in technology allowed the detailed study of STRs. These markers are very well studied and therefore, are regularly used as the gold standard tool in forensic investigations. As mentioned, STRs are a class of repetitive sequences that have repeat units between two and seven nucleotides in length and are tandemly repeated from approximately a half a dozen to several dozen times. This repetitive structure makes STRs tremendously variable, in part due to slippage during DNA replication, which leads to alleles with different numbers of repeat units. Repeat length polymorphisms such as STRs are only one of the several types of sequence variation throughout the human genome: bi- and multi-allelic polymorphisms, such as single nucleotide polymorphisms (SNPs) and insertion-deletions (INDELs), are examples of other commonly used types of genetic variations in population and forensic genetics. However, due to the content of this dissertation (X-STR sequence variation), the focus will be on STR type of markers only.

There is a considerable amount of STRs present in the human genome but only a few have been selected to be used in forensic studies (Butler, 2007; Cao et al., 2015; Gomes et al., 2020). The main reasons behind the choice for STRs are: 1) they have a lot of variation, so they are highly polymorphic allowing a good discrimination power; 2) the

techniques used for STR analysis are easy to apply to forensics due to the sensitive nature of PCR and capillary electrophoresis (CE) and finally 3) the ability to generate STR multiplexes with small amplicon lengths for degraded DNA. The STRs used in forensic investigations come from autosomal chromosomes and are usually the source of genetic information for these investigations, which have been in use in the field for a long time with great results and are widely accepted and validated (Amorim & Pereira, 2005). Although autosomal STRs are very informative and useful, in the last couple of decades there's been a huge interest on the STRs located on the sex chromosomes, such as X-STRs. In fact, X-STRs have been developed into an established method used in forensics laboratories to provide extra information alongside the results from the autosomal STRs in highly specific cases (Diegoli, 2015). There is a plethora of studies using the X-chromosome genetic markers (Bottinelli et al., 2022; Chen et al., 2018; Veselinović et al., 2017).

In resume, X-STRs gained interest in the fields of population and forensic genetics due to their characteristics of inheritance that allow complementing the information obtained from autosomal markers or even exclusively provide the information needed to answer certain questions.

1.4. Applications of X-STRs in forensic and population genetics

X-STRs have many, but particular applications in population and forensic genetics such as, for example, in kinship analysis. For instance, in father/daughter relationships and mother/son relationships, X chromosome genetic markers are a great help because of its transmission mode presented above (Figure 1). Another great example of the contribution of X-STRs are cases where potential paternal half-sisters' relationships are tested and, the putative father is unavailable, also known as "deficiency cases" (Figure 2). In these cases, the X-STRs are useful since the half-sisters must share an entire, unchanged (except when mutations occur) X chromosome copy, which they got from their putative father (Gomes et al., 2008; Gomes & Pereira, et al., 2009). Due to the lack of recombination in males, this leads to direct haplotyping and exclusion of sisterhood if they do not share the same paternal alleles. This can also be said for granddaughter/grandmother relationships (Gomes et al., 2008) (Figure 2).

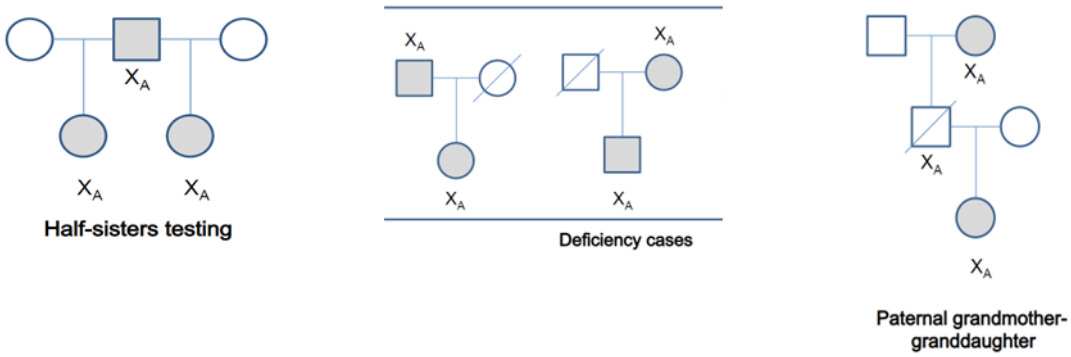


Figure 2. Some example pedigrees where X-chromosome testing can be useful. Image obtained from Advanced topics in Forensic DNA typing: Methodology, John M. Butler (2012) (Butler, 2012).

In stain analysis, the X chromosome can also be useful in some instances. For example, to identify female traces in male contaminations, X-STRs are advantageous compared to autosomal ones because female alleles can only be completely included in the male counterpart only if the female is homozygous for all the alleles. In male traces on female contaminations, this does not happen simply because there is a higher chance of the male alleles to be included on the female component therefore presents a limitation for the X chromosome usage (Szibor, 2007).

A particular subject that comes to mind on any forensic genetic-based investigation is having access to a well-developed population genetic (frequency) database that allows for access of relevant population data parameters or even for comparison with other population groups. Population databases regarding autosome and Y chromosome genetic markers are well stablished and are in constant growth (Roewer, 2013). On the contrary, for X-STRs, compiled population information is scarce and no solid database exists (Gomes et al., 2020), which limits their use in forensic investigations. Attempts have been made in the past and an X chromosome database was created (www.ChrX-STR.org). However, the objectives were not achieved, and for several years now, it has not been updated.

Since the increased interest on the use of X-STRs, a high number of populations have been genetically characterized. In the study by Gomes and co-authors (Gomes et al., 2020), a detailed compilation of population X-STR data was performed. Regarding the number of loci described in the literature retrieved from this study, 85 X-chromosomal loci were reported (Attachment 1- Supplementary Table 1). In addition, the 236 papers analyzed in this work provided the forensic genetics community an overview of what is

available on this topic. Those researchers concluded that throughout 1999 until 2008, the number of papers regarding population data on the X-STRs saw a huge boost but it slowly became stagnant in the last years (Gomes et al., 2020). It was also pointed out that the number of markers included in the studies varied a lot partially due to the use of commercial kits (Table 1), which had a development throughout the years, the first consisting of four markers, and the most recent one, the Investigator Argus X-12 (Qiagen, Hilden, Germany), expanded to 12 markers. Alleles of linked markers form haplotypes if markers are found to be in linkage disequilibrium and, in these cases, these should be analyzed together (Figure 3). X chromosome haplotype analyses has been demonstrated to be a valuable tool in pedigree kindship testing therefore the Investigator Argus X-12 can support these cases. This relates to another focal point in the forensic field and in population genetics, which is linkage and linkage disequilibrium. This is particularly important when the X chromosome is the topic since all genetic markers are on the same chromosome. Linkage means that if more than 50% of the gametes are expected to have the same segment as the parental chromosome the two markers are linked. Linkage disequilibrium (LD), on the contrary, is a phenomenon that occurs due to the non-random association of alleles from different markers at a population level.

Regarding worldwide information about X-STRs, the data is scarce for some populations like those from sub–Saharan Africa and America (with some exceptions). On the contrary, China, for example, is well represented with the highest number of populations studied. This shows an imbalance representation of worldwide population data and further studies must be performed to have a highly comprehensive and representative human X-STR database (Gomes et al., 2020).

Studies have been performed using the standardized kit, the Investigator Argus X-12 (Qiagen, Hilden, Germany), and several other multiplex systems have grown along the years (Table 1). These systems are adding information on allelic frequencies to the databases of the specific countries where the study was completed (Bini et al., 2020; Gomes et al., 2020; Perera et al., 2021; Tao et al., 2018).

Table 1. Most used multiplex PCR assays targeting X-STRs. References do not necessarily refer to the original development papers (Gomes et al., 2020).

Name	References	Nr. and STR loci
Goldeneye 17X kit	Gao et al. (2019)	16 (DXS6795, DXS9902, DXS8378, HPRTB, GATA165B12, DXS7132, DXS7424, DXS6807, DXS6803, GATA172D05, DXS6800, DXS10134, GATA31E08, DXS10159, DXS6789, and DXS6810)
Investigator® Argus X-12 QS (Qiagen) kit	Elakkary et al. (2014)	12 (DXS7132, DXS7423, DXS8378, DXS10074, DXS10079, DXS10101, DXS10103, DXS10134, DXS10135, DXS10146, DXS10148, and HPRTB)
Microreader™ 19X ID System kit	Lin et al. (2020)	19 (DXS6795, DXS6803, DXS6807, DXS9907, DXS7423, GATA172D05, DXS101, DXS9902, DXS7133, DXS6810, GATA31E08, DXS6800, DXS981, DXS10162, DXS6809, GATA165B12, DXS10079, DXS10135, and HPRTB)
AGCU X19 STR Kit	Li et al. (2017)	19 (DXS8378, DXS7423, DXS10148, DXS10159, DXS10134, DXS7424, DXS10164, DXS10162, DXS7132, DXS10079, DXS6789, DXS101, DXS10103, DXS10101, HPRTB, DXS6809, DXS10075, DXS10074, and DXS10135)
–	Deng et al. (2017)	19 (DXS8378, DXS9898, DXS7133, GATA31E08, GATA172D05, DXS7423, DXS6809,DXS7132, DXS9902, DXS6789, DXS8378, DXS7423, DXS7132, DXS10079, DXS6801, DXS6799, DXS6800, DXS10075, DXS6807, and DXS6803)
–	Prieto-Fernández et al. (2016)	17 (DXS9895, GATA144D04, DXS10077, DXS10078, DXS10161, DXS10160, DXS981, DXS6800, DXS6803, DXS9898, DXS6801, DXS6799, DXS6797, DXS7133, DXS6804, GATA172D05, DXS8377, DXS10146, and DXS10147)
GHEP-ISFG decaplex	Gusmão et al. (2009)	10 (DXS8378, DXS9898, DXS7133, GATA31E08, GATA172D05, DXS7423, DXS6809, DXS7132, DXS9902, and DXS6789)
–	Zhang et al. (2017b)	15 (DXS6807, DXS8378, DXS6795, DXS10164, DXS7132, DXS10074, DXS6803, DXS6801, DXS101, DXS7133, GATA165B12, DXS10103, HPRTB, GATA31E08, and DXS7423)
MiSeq FGx™ Forensic Genomics	Jäger et al. (2017)	7 (HPRTB, DXS7132, DXS7423, DXS8378, DXS10074, DXS10103, and DXS10135)

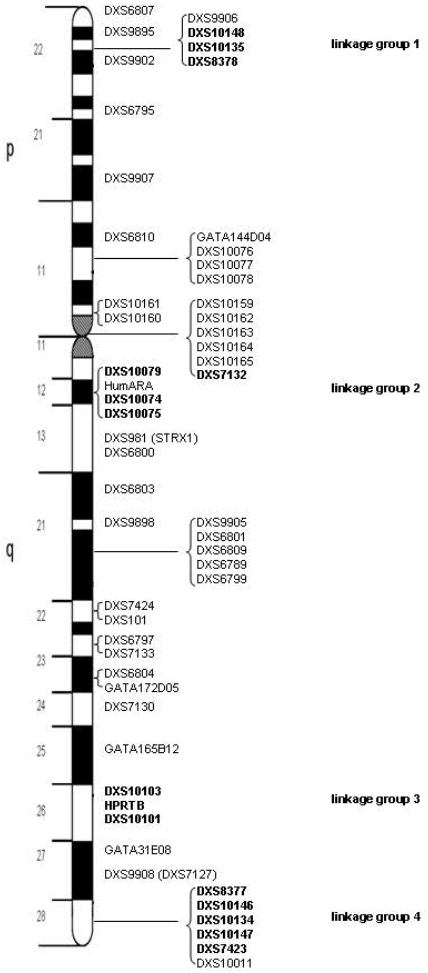


Figure 3. Location of some commonly used X-chromosomal STR markers for forensic and population genetic studies. Image obtained from the chrx-str.org database. The four linkage groups of the Investigator Argus X-12 markers as well as the complete 12 X-STRs (in bold) that compose this commercial kit are represented in the image, except for loci DXS10175, DXS8377, DXS10147 which are not part of this system.

There has clearly been a step forward and great attempt to expand the information on X-STRs but more studies about other relevant population genetic topics such as haplotype frequencies, mutation rates and linkage disequilibrium need to be made to build a reliable database about X-chromosomal genetic markers (Gomes et al., 2020)

1.5. Sequence variation, repeat structure and nomenclature of X-STRs

1.5.1. Applications in forensic genetics

Creating X-STR population-specific databases for identification purposes as well as for population genetic studies can only be useful if a worldwide well accepted nomenclature is established. This is very important to avoid misinterpretation of obtained results such as comparing allele frequency amongst different populations and no misinformation between laboratories when these are exchanging data. This problem has already been pointed out in many studies (for example, in (Gusmão et al., 2002; Gomes et al., 2009; Gomes et al., 2009)). Due to the substantial increase of data over the years, not only these studies generated a good amount of sequencing data for X-STRs but they also made it easier to find differences among nomenclature reports making it hard to compare the data between laboratories (Diegoli, 2015; Gomes et al., 2020).

As stated in the International Society for Forensic Genetics (ISFG) guidelines (Tillmar et al., 2017), initially, the X-STRs tests and information were mostly gathered in individuals of European ancestry. This led to a lot of discrepancies in the repeat structures when studies from populations around the world started to be made, with these structures being more variable than what initially was seen (Tillmar et al., 2017). This not only shows that, for an inclusive nomenclature, worldwide research needs to be made but, in these studies, a more detailed look at the variation and the structure of the repeat is highly needed. This problem has also been looked upon and pointed out in some studies (Gomes et al., 2008; Gomes et al., 2009). In fact, this problem is so impactful that it led to a complete revamp of the primers initially used in the first version of the Investigator Argus X-12 kit (Qiagen, Hilden, Germany) that was substituted by the Investigator Argus X-12 QS kit (Qiagen, Hilden, Germany). As already mentioned, as most of the DNA sequence data mainly came from subjects with European ancestry, a lot of silent alleles were detected due to the mismatch of primer binding sites when comparisons with other different populations groups began, such as in Africans (Gomes et al., 2020).

This is a concerning point for the future applications of X-STRs in forensic and population studies since a good nomenclature, a well explored sequence variation and repeat structure is crucial for this area of study.

1.5.2. *Pan troglodytes* genome

To establish the most accurate allele or locus nomenclature, the DNA sequence of a microsatellite should be analyzed in several diverse human population groups and not be restricted to only one population group, which is usually the case. In order to detect for possible hidden DNA sequence variation distinct population genetic groups such as Europeans, Asians and Africans should be included for the detection of possible hidden sequence variation. On the other hand, comparisons with the chimpanzee (*Pan troglodytes*) genome may provide insights into the discernment of the most likely ancestral sequence. In addition, being the closest living animal to humans (together with bonobos – *Pan paniscus*), the chimpanzees can help us infer crucial information about our genome. For example, comparisons can help infer on mutation rates of certain sequences, structural and functional differences of certain mutations and discover evolutionary forces that have lead to our species being so different from the rest (Feuk et al., 2005; Mikkelsen et al., 2005; Newman et al., 2005.).

The divergence between the two species happened about 5 to 7 million years ago but it is still interesting to find the genetic elements and evolutionary forces that made us, humans, so different from the chimpanzees (Suntsova & Buzdin, 2020; Varki & Altheide, 2005). This fact, combined with biomedical interests and general evolutionary principles make the chimpanzee a great resource to be used (Varki & Altheide, 2005).

In the medical research field, even though the chimpanzee is not the most used model for most human diseases due to the inherent technical and ethical implications , comparing our genome to theirs can broad our knowledge of certain mechanisms and processes that happen when, for example, an infection occurs (Varki & Altheide, 2005). In fact, the chimpanzee is used as a model to help us better perceive and understand diseases that affect humanity, as for example, in the case of malaria. Studying chimpanzees may help scientists understand the illness and how it evolved. This also happened for the case of human trisomy 21 that in the chimpanzees is known as trisomy 22. Comparing similar pathologies between these two species is a great aid to understand and help expand the expertise in the medical field (https://microbewiki.kenyon.edu/index.php/Chimpanzee_Evolution).

Humans and chimpanzees share 98% similarity in their DNA. This is true when single nucleotide divergences are being considered, since the divergence is said to be approximately 1.23% (Varki & Altheide, 2005). When more than just a single nucleotide divergence is considered, a different perspective and number is presented, especially for repetitive genomic DNA and transposable elements. When, for example, INDELs are

analyzed, approximately 90MB difference between humans and chimpanzees is found which leads to a 3% divergence rate in this specific category (Varki & Altheide, 2005, https://answersingenesis.org/genetics/dna-similarities/greater-than-98-chimp-human-dna-similarity-not-any-more/).

For the sex chromosomes, comparison between humans and chimpanzees showed that the X chromosome present the lowest divergence amongst all chromosomes, noted as 0.94% (Kehrer-Sawatzki & Cooper, 2006). On the contrary, the Y chromosome has the largest divergence between the two species, noted as 1.78%. Impressively, although the X chromosome shows low divergence, the differences at a structural level are very significant as it has the highest numbers of paracentric inversions of all chromosomes (Kehrer-Sawatzki & Cooper, 2006).

With all of this in mind, we can learn about our own genome, which is especially useful in forensics and tied in with the topic above for creating a better nomenclature for STRs. Comparing our repeat structure with the one from chimpanzee (*Pan troglodytes)* allows for a view on what the ancestral gene sequence would look like, giving a better analysis on the sequence structure and possible mutations of a specific genetic marker (Gomes et al., 2009). In fact, this has been explored and applied not only to X-STRs (Gomes et al., 2008; Gomes et al., 2009), but also to Y-STRs (Gusmão et al., 2002; Gusmão, Alves, et al., 2002).

Another use of the comparison between humans and chimpanzees is the validation of cross-species human-specific STR systems, since there are cases where human antiserums showed cross reactivity with non-human blood. Therefore, it is important to explore this field of cross-species amplification to better improve the quality of these kits (Singh et al., 2019). Cross amplification systems are important for genetic divergence and evolution studies since a similar set of markers can be used to test these cases (Thakur et al., 2018).

The comparison of human and chimpanzee genomes is a great tool and can be used for future studies that address sequence variations and repeat structures. This approach when establishing a new nomenclature gives us a way to be more precise and accurate when reporting a repeat of a studied STR, ensuring an accurate representation and nomenclature. It also allows us to learn more about our own history and evolution

1.5.3. Types of STR allele structures

STR DNA sequences diverge in the length of the repeat unit and the number of repeats and are always read in the 5' to 3' direction. STR allele sequence structures can be divided into three classes: "simple", "compound" and "complex" repeats (Urquhart et al., 1994) (Table 2):

i) "Simple" repeats only show variation in the number of repeats without additional sequence variation (i.e., repeat units of identical length and sequence);

ii) "Compound" repeats include several adjacent simple repeats of the same repeat unit length;

iii) "Complex" repeats are repeats of variable length as well as sequences.

Table 2. Examples of STR alleles categorized in three different classes based on the core repeat pattern. Table in Department of Biostatistics, School of public health, University of Washington. (https://si.biostat.washington.edu/sites/default/files/modules/2).

Class	Locus	Allele sequence
Simple	CSF1PO	$[TCTA]_8$
Simple	Penta D	$[AAAGA]_{12}$
Compound	vWA	$[TCTA][TCTG]_4[TCTA]_{13}$
Compound	D22S1045	$[ATT]_7ACT[ATT]_2$
Complex	FGA	$[TTTC]_3TTTTTT[CTTT]_{11}CTCC[TTCC]_2$
Complex	D1S1656	$[TAGA]_4TGA[TAGA]_{13}TAGG[TG]_5$

1.5.4. Implications of X-STR nomenclature studies

As proposed about 25 years ago by the DNA recommendations of the International Society for Forensic Genetics (Olaisen et al., 1997), a common STR nomenclature is a prerequisite for inter-laboratory reproducibility, exchange and comparison of data. Initially, when the interest on X-STRs begun to rise, a need for studies exploring not only allele frequencies in populations but also the sequence variation at the allele and STR sequence level also began to increase. Studies such as the ones by Edelmann et al., (2002, 2008) focused on describing the repeat structure and determining a nomenclature of usable X-STRs in forensic investigations. These studies paved the way and gave a better insight on the structure of the repeats but had a minor limitation (as it was already mentioned) that the population analyzed was mainly from European ancestry. Studies involving more populations with diverse ancestries and comparisons with the chimpanzee genome such as demonstrated, for example, in the studies of Gomes et al. (2008, 2009, 2016) showed new variations on the previously determined structures of the alleles that had gone undetected. These mentioned studies gave a better perspective

on the possible existence of silent alleles or possible repeats in the flanking regions of the alleles that should be included, and a new nomenclature proposed, revamping the already established one. These types of studies where the repeat structure is looked upon are fundamental since a well-established nomenclature is necessary to better insure a solid reporting of results and communication between the scientific communities.

1.6. Forensic and population genetics: future perspectives of X-STRs

DNA genetic markers such as X-STRs have a lot of applications in forensic and population genetics, which were already mentioned above, such as solving kinship relationships and help solving criminal cases through specific cases of human identification when, for example, autossomal markers are not enough. X chromosome markers may also be applied to clinical analysis due to their link to certain diseases and even in worldwide population genetic studies aiding in finding variation among populations and understanding and our species own history.

Although DNA profiling techniques are well established (for example, detection of DNA fragments though capillary eletrophoresis are viable and provide great results), it can present some limitations especially when the quantity of the DNA sample is low or too degraded to be analysed (Aly & Sabri, 2015; Haddrill, 2021). On top of that problem, capilary eletrophoresis has a limited multiplex capability and cannot identify sequence variation in the STR core repeat regions neither in its flanking regions if the size of the fragment or amplification itself is not affected (for example, null alleles or Indels) (Simayijiang et al., 2022).

In recent years there has been many and significant advancements in tecnology also in the scientifc field. A particular technique has gained a lot of attraction which is Next Generation Sequencing (NGS) or Massively parallel sequencing (MPS). This is defined as a method that sequences a portion of DNA sequence of an individual's genome (Bruijns et al., 2018) or the entire genome in a relatively quick time frame. Being a technique with such great throughput it has the ability to generate a lot of data (millions of data) in a single run (Haddrill, 2021). This is one of the major advantages of this technique since there is no limit on the number of analised STRs in one single run (Dash et al., 2021). Generating a huge number of data is very useful but that is not the only good point about the NGS. In contrast to fragment length analysis by PCR-CE, sequencing reveals the true variation of STR loci and recognize a difference in alleles

that have similar length which CE method can not. This will lead to a better analysis of mixtures, and increase the efficiency of many legal cases (Aly & Sabri, 2015; Li, 2018). MPS can also improve results for samples where DNA has degraded or samples where DNA quantity is low (Haddrill, 2021). These are positive points that come from NGS based methods but there are many possible applications of these techniques besides from routine forensics such as age estimation, body fluid identification, detection of geographic origin and ancestry and many more (Dash et al., 2021).

In fact, there are already some studies available using NGS to infer genetic structure and provide some genetic insights from a population wise perspective (Chen et al., 2021; Dash et al., 2021; Lan et al., 2022.). Also advances have been made in the analysis of human remains (Ambers et al., 2020; Ambers et al., 2016) which keeps showing the potential use of this technique to this field.

In 2020, Gomes I and coworkers (Gomes et al., 2020) made an overall review on the status of the use of X chromosome markers, in particular, of X-STRs in forensic population genetics. This work came to conclusion that the X chromosome is being underused and being overlooked in terms of potential since the numbers of studies and publications are declining and starting to become stagnant (Gomes et al., 2020). This is a worrying fact but it gets even more concerning when the available data for the X-STRs in some parts of the world is still low and unexplored (as seen in Figure 4) (Gomes et al., 2020).

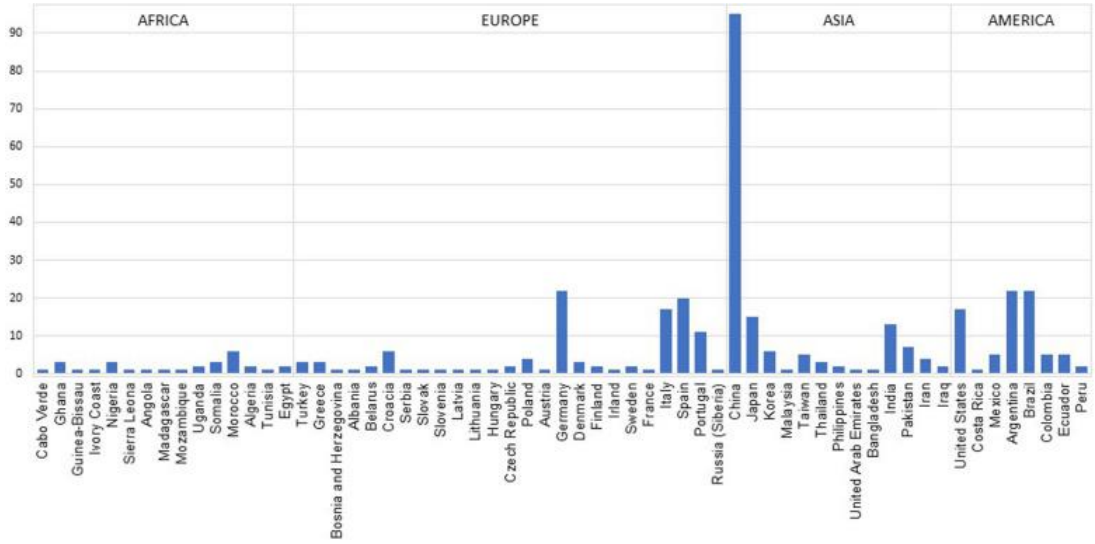


Figure 4. Available data of X-STRs by country. Image obtained from "Twenty years later: A Comprehensive Review of the X Chromosome Use in Forensic Genetics. (Gomes et al., 2020).

This is not a good look on the usage of X chromosome in this field, but this sex chromosome has shown a lot of potential and is in fact a good complement to the information obtained from the autosomal markers. Hopefully, in the years to come the X-STRs will gain more attention and the new technologies being developed will allow to create a larger database of X chromosome genetic markers and population data so some of the problems that exist are solved. In general, there are problems in the analytical and statistical analysis of X chromosome markers by the forensic community mostly due to the genetic model of inheritance (haploid state in males versus the diploid state in females) because different estimation methods are needed for each sex (Gomes et al., 2020). This is probably one of the main reasons why the X chromosome fell in terms on publications (Gomes et al., 2020). Nevertheless, in the mentioned study (Gomes et al., 2020) it was also pointed-out that much effort has till date been put in this topic (validation and development of X-STR multiplex systems, characterization of populations, statistical kinship-base testing solutions) and thus it is important to revive the interest in X chromosome markers for population and forensic genetics applications.

2. Objectives

The analysis of autosomal and gonosomal genetic markers, especially STRs, proved to be an essential tool not only in forensic and population genetic studies but in diverse fields of study as well. Even though sex chromosome markers are used mainly to complement the information of the autosomal ones, the X chromosome is yet to be fully explored. Having a general view on the genetic variation, specifically regarding the repeat structure variation of the X-STRs on several population groups is essential to establish an accurate nomenclature. To access this variation, the knowledge of the ancestral STR structure is also essential which is obtained by performing a comparison with the genome of the chimpanzees (*Pan troglodytes*). Gathering information about the various markers is valuable and a great help to build a concrete and rich database, setting up the usage of these genetic tools in forensic investigations. This study was done with the purpose of categorizing the genetic diversity of X-STRs in humans, in particular of European, Asian and African populations, and to compare with the chimpanzee genome in order to propose a nomenclature that better describes the variation at the DNA sequence.

These are the points underlined for this study and therefore the aims established are:

1) To analyze DNA sequence variation by amplification and sequencing of the allele and flanking regions of the X-STRs: DXS10074, DXS10134, DXS10148 and DXS10146.

2) To detect as much variation as possible by comparing distinct samples from European (Portugal), Asian (Macau) and African populations (published data).

3) To analyze the general structure and stability for the DNA sequence of the markers cited in point 1) by comparing the chimpanzee (*Pan troglodytes*) and human (*Homo sapiens)* genomes.

3. Material and methods

3.1. Strategy

In order to achieve the proposed aims of the present study, due to several limitations of financial and experimental nature, two different strategies were adapted and combined: experimental laboratory and *in silico* work (note that *in silico* in this study referes to literature search, compilation and comparions as well as to the chimpanzee genome alignments using informatic tools). Table 3 ilustrates the information for the procedure used for the samples that were sequenced (wet-lab) and for the samples that were analysed by comparing available published data (*in silico*). The studies used are referenced in the "Results and Discussion" section. For the African data, as considerable sequence data is already publically available for the majority of the X-STRs, and also due to financial constraints, it was decided to used the available information in the literature

Table 3. Compiled information on the strategies for both the practical (laboratory) and computer work adapted.

Marker		Europeans	Asians	African
DXS10074	***Samples***	n= 20	n= 5	n= 13
	Population	Portugal	Macau	Guinea-Bissau and other West African groups
	Method	Sanger Sequencing	Sanger Sequencing	*In silico*
DXS10134	***samples***	n=20	n= 3	n= 11
	population	Portugal	Macau	Guinea-Bissau
	method	Sanger Sequencing	Sanger Sequencing	*In silico*
DXS10135	***samples***	n=20	n= 3	n= 19
	population	Portugal	Macau	Guinea-Bissau
	method	Sanger Sequencing	Sanger Sequencing	*In silico*
DXS10148	***samples***		n=19	n= 71
	population	n.a.	Korea/Malasia	Guinea-Bissau and other West African groups
	method		*In silico*	*In silico*
DXS10146	***samples***	n= 6		n= 26
	population	Portugal	n.a	Guinea-Bissau
	method	Sanger Sequencing		*In silico*

3.2. Samples and DNA extraction

Samples from Portugal were used for the DNA sequencing of the European group (Table 3). Blood samples were collected using a sterile, single-use lancing device (Accu-Chek, Roche, Germany) from healthy and unrelated male volunteers, under informed consent. The samples were collected in the scope of a previous study (Faustino M., 2021) for population genetic research purposes. Therefore, DNA extracts were already available for use. Nevertheless, even though the samples were not extracted in this study, the method used to extract the DNA from the blood samples is provided, which was dependent on the card type used for the blood stain preservation, the card sample age and dryness level. For the Whatman FTA cards, DNA was extracted using a standard Chelex method (Walsh et al., 1991) or a modified version for dry and old bloodstains. For the GenReleaz cards an elute methodology was used (Fast and efficient recovery of DNA, Ahlstrom-Munksjõ, 2018).

3.3. Amplification of the STRs

1- Amplification of the X-STRs was done through a PCR reaction, in singleplex, for each marker, namely DXS10074, DXS10134, DXS10135, DXS10148 and DXS10146. The primers used were newly design using the website primer3 (https://primer3.ut.ee/). Primer information and fragment sizes are shown in Table 4.

Table 4. Primer sequences (5'-3') and additional information on the studied loci.

X-STRs	Primer sequences (5'-3')	Reference	Frag Size	Tm ºC
DXS10135	F- GCCAAGAACATTTTTCAGTCA	Gomes et al. (2017)	254	58.52
	R- TAGCCAAGCCTGAAGACCAC			58.94
DXS10146	F- GATCTGCCTTGCCCTTCCTA	new this study	327	59.16
	R- TGGGGTGACAGACTGAGAGA	Gomes et al. (2017)		59.52
DXS10148	F-TGCATGACAGAGGGAGATTCT	new this study	209	58.52
	R- CTGGTTTTGGCAGTGGTGTT	new this study		58.18
DXS10074	F- TGCCATTTGTTGTGATCTCC	Gomes et al. (2017)	227	
	R- CCCTCAGAGAGCTGACACAG			
DXS10134	F- GCCTGGGTGACATAGAGAGA	new this study	284	58.22
	R- TGTGAACAAAACTGCTATGAACA	new this study		57.17

Tm= melting temperature (annealing temperature)

2- Initially, a test was performed using the Master mix and reagents concentration and volumes presented in Table 5 with a few samples only, and including a negative and positive control to test for possible contamination and to see how the reaction was carried out, respectively. The thermocycling conditions used are listed in Table 6.

Table 5. Master Mix used in both standard and "Touchdown" PCR reactions of 10µl final volume.

Reagents	**Vol per Sample (µL)**	**Concentration in PCR**
Hot start QIAGEN master mix	5	2.5 units HotStarTaq DNA polymerase 1x PCR buffer w/ 1.5 mM $MgCl_2$ 200µM
Primer forward	0.8	0.4 µM
Primer reverse	0.8	0.4µM
H20	0.9	-
DNA	2.5	-

Table 6. Initial thermocycling conditions for standard PCR.

Phase	**Temperature**	**Time**	**Cycles**
Initial denaturation	95ºC	15 Min	-
Denaturation	94ºC	30 sec	30x
Annealing	58ºC	90 sec	
Extension	72ºC	60 sec	
Final extension	72ºC	30 min	-

3- Amplification results were checked using a polyacrilamide gel (detailed protocol given below). After obtaining the results (Figure 5), unspecified amplification products appeared alongside the product of interest that would most likely interfere with the DNA sequencing reaction. This lead to a change in the amplification conditions and it was decided that a touchdown PCR (TD-PCR) could possibly be more efficient and overall

better for the amplification of our samples. TD-PCR is a technique that increases the specificity of PCRs due to the annealing step being initially set at 5ºC to 10ºC higher than the Tm of the primers (Green & Sambrook, 2018). The thermocycling conditions of TD-PCR are described in Table 7.

Table 7. Thermocycling conditions of TD-PCR reactions.

Phase	**Temperature**	**Time**	**Cycles**
Initial denaturation	95ºC	15 min	-
Denaturation	95ºC	30 sec	10x
Annealing	*70ºC	45 sec	
Extension	72ºC	60 sec	
Denaturation	95ºC	30 sec	25x
Annealing	60ºC	45 sec	
Extension	72ºC	60sec	
Final extension	72ºC	30 min	-

*-1ºC/cycle

4- With that said, a master mix was made for each reaction with the same quantities and reagents as the standard PCR described in Table 5. When preparing the master mix every quantity of each reagent was then multiplied by the number of samples done in that specific run. Then this master mix was distributed by all tubes used in the experiment leading to a total volume of 10µL per tube (7.5µL of Master mix + 2.5µL of DNA). DNA concentrations were not available, however, intensitites of the bands seen in the polyacrylamide gels were used to infer on the quantity of amplified product to be used in the sequencing reactions. In addition, for each PCR reaction, a negative control was included by replacing the DNA by sterile water to trace possible contaminations.

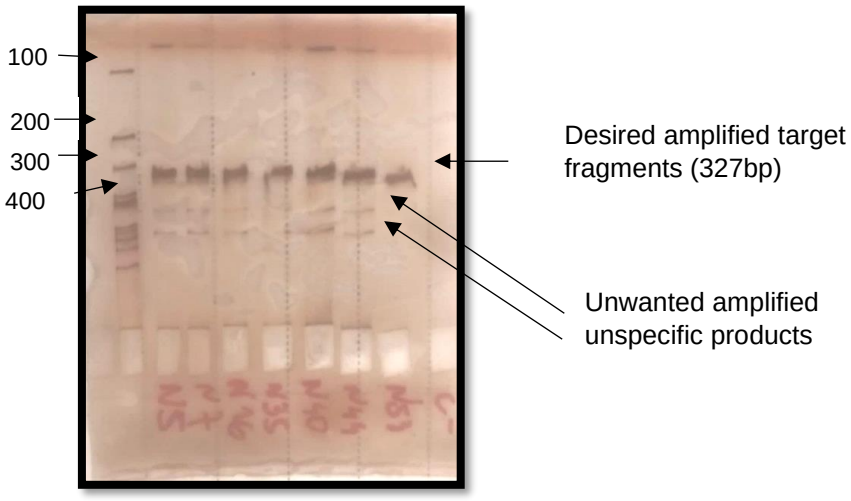


Figure 5. Polyacrylamide gel revealing amplified PCR products for DXS10146 with undesirable bands. Sample well layout (from left to right): DNA GRS 100bp ladder (Grisp, Portugal).

3.4. Detection of PCR amplification products

To make sure the PCR reactions were successful and without any problem prior to sequencing reaction, a Polyacrylamide gel eletrophoresis (PAGE) was performed to confirm amplified product. This polyacrylamide solution was prepared using 25ml of gel buffer solution (1.5M Tris-HCL), 20ml of acrylamide (Bis acrylamide 19:1 40%), 7ml of glycerol and 43ml of deionized water.

The gel solution was made combining 6ml of polyacrylamide solution, 340µL of APS and 14µL of Temed. Then this solution was poured out to a frame where it was left to polimerise for 15/20 minutes.

The eletrophoresis was performed using an Amersham Biosciences Multitemp III thermostatic circulator and an consort EV 243 power supply. The run was performed at 195V, 45Ma and 20W using a 100bp ladder.

After the run, the gel was stained using the silver staining method to reveal the amplified products. This method follows the respective steps:

1) 10 minutes in Ethanol 10%
2) 5 minutes in Nitric Acid 1% (in a shaker)
3) Quick wash with destiled water
4) 20 minutes in Silver nitrate 0.2% (in obscurity and in a shaker)
5) Quick wash with destiled water

6) Sodium carbonate 0.28M and formaldehyde 0.02% solution (6g of Sodium carbonate + 1ml of formaldehyde + 200ml of destiled water) prepared and used to visualize the amplified fragments.
7) 2 minutes in Acetic acid 10%
8) Final wash with destiled water

3.5. Sequencing reaction

After the confirmation of successful amplification of the TD-PCR, the post PCR products needed to go through a couple of final proceses:

First purification: PCR product

Firstly, a initial purification reaction was performed containing 1µL of ExoFastap (1:4) + 2µL of post TD-PCR DNA. This solution was then submitted to two different temperature steps in a termocycler (conditions described in Table 8). Initially, only 15 minutes were used at the 37ºC, but in other experiments performed at the laboratory it was found that using 30 minutes at 37ºC had better results.

Table 8. Thermocycling conditions used for the ExoFastap purification.

Temperature	Time
37ºC	30 min
80ºC	15 min

After this purification, the sequencing reaction was done for both DNA strands using the forward and reverse primers to access which better primer ensured better results, in this case, a better sequence. Once again a master mix containing the reagents described in Table 9 was made. The mix was divided by the tubes, for a total reaction volume of 5µL (3µL of the mix + 2µL of the purified PCR product). Termocycling conditions used are described in Table 10.

Table 9. Reagents used for the sequencing reaction mix.

Reagent	**Volume per tube (µL)**
Big Dye Terminator v3.1	0.5
Primer (Forward or Reverse) 5µM	1
Buffer	1
H20	0.5

Table 10. Thermocycling conditions used for sequencing.

Phase	**Temperature**	**Time**	**Cycles**
Initial denaturation	96ºC	2 min	-
Denaturation	96ºC	10 sec	30x
Annealing	55ºC	10 sec	
Extension	60ºC	1 min	
Final extension	60ºC	10 min	-

Second purification: Sequencing reaction products

Lastly, after the sequencing reaction a second purification was performed using sephadex columns, which is based on the following protocol:

1) Columns are put in 2.0 ml microtubes;
2) Put 750µL of Sephadex in the columns;
3) Centrifuge at 4400 rpm for 4 minutes;
4) The columns are transferred to 1.5ml tubes marked with the sample code;
5) Total amplified products are then pippeted to the center of the column;
6) Another centrifugation at 4400 rpm for 4 minutes;
7) Store the reusable columns after throwing away the sephadex;

After this process, 8µL of HiDi formamide is added to each sample and submitted to capillary electrophoresis using the system ABI 3130 Genetic Analyser for product detection (this step is performed at the Genomics Core Facility-GenCore- available in-house). Results were analysed with the Sequencing analysis software v.5.2 (Life Technologies).

3.6. Alignment of human (*Homo sapiens*) and chimpanzee (*Pan troglodytes*) genome sequences

The human and chimpanzee sequences were obtained using the website “The National Center for Biotechnology Information” (https://www.ncbi.nlm.nih.gov). The human sequences for each X-STR were retrieved using the assembly of Dec. 2013 (GRCh38/hg38) reference sequence. For the reference chimpanzee sequences the assembly of Jan. 2018 (Clint_PTRv2/panTro6) was used. After obtaining the sequences, the alignments were performed by using the Emboss Needle Sequence Alignment tool from EMBL-EBI (https://www.ebi.ac.uk/Tools/psa/emboss_needle/).

4. Results and discussion

4.1. Selection of samples for Sanger sequencing

Samples used for the Portuguese population group were already genotyped due to a previous study performed *in-house* (data not published). The kit used for genotyping of those samples was the Investigator Argus X-12 QS Kit (QIAGEN, Hilden, Germany), and the genotypes obtained are presented in Attachment 2- Supplementary table 2. For this study, based on the genotypes, n=20 Portuguese male samples were randomly chosen, but with an attempt of having a large representation of different alleles without trying to skew the results. Having different alleles allows for a better understanding of the repeat structure and variation that might happen in the DNA sequence of the STR. For quality control, the genotypes obtained by fragment analysis were compared with those obtained in this study by sequencing.

For the Macau population samples, no information was available on the genotypes for the X-STRs chosen for this study (DXS10074, DXS10134, DXS10135, DXS10146 and DXS10148). Due to the quality of the DNA extracts, only five male samples analysed in this study for Macau were sucessfully sequenced after a trial and error method (i.e., until satisfatory amplification results were obtained and financial constraints allowed for).

4.2. Singleplex amplification of X-STRs: technical challenges

All X-STRs were amplified by singleplex and results were checked using a polyacrylamide eletrophoresis gel prior to sequencing (complete methodolgy is explained in Material & Methods).

Initial tests in the amplification of the samples from Portugal for DXS10146 using a standard PCR showed some unwanted unspecific amplicons (Figure 5 in Material & Methods). After changing from a standard PCR to a TD-PCR protocol (explained in Material & Methods section), the results highly improved (Figure 6).

Figure 6. Polyacrylamide gel obtained after TD-PCR (unwanted unspecific problems are no longer visible).

Even though, PCR amplification of this marker still had some issues. Some unwanted amplicons started to appear in the polyacrilamide gel (Figure 7). Due to this problem, DNA contained in the amplified bands was extracted from the polyacrilamide gel and re-amplified (since it is the targeted fragment, in this case, the DXS10146 allele). The method consisted in the following:

- Cut, with a scalpel, the bands with the targeted fragment (allele);

- Put the band in a tube with 50µL of TE solution;

- Heat the tubes at 56ºC for 15 minutes and freeze them for 30 minutes (this process is performed 3 times);

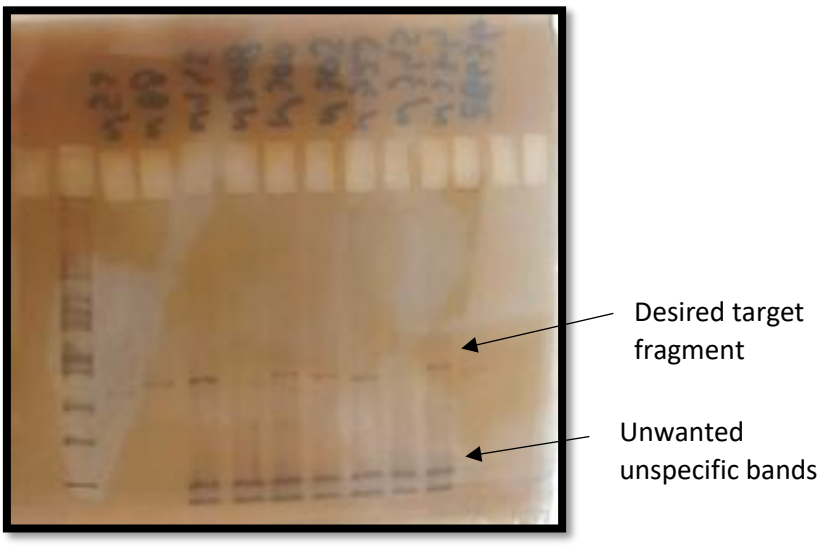


Figure 7. Polyacrylamide gel with unwanted unspecific bands after TD-PCR amplification for DXS10146.

After this method another PCR was performed but still the same unwanted bands appeared. It was concluded that the primers used probably amplified more than just the targeted fragment and, therefore, new primers were designed (final primers in Table 4).

Since the TD-PCR technique worked well for the DXS10146 marker overall, it was adapted for all the markers studied in this dissertation. Using new primers for most markers, the TD-PCR showed no problems or issues (Figure 8).

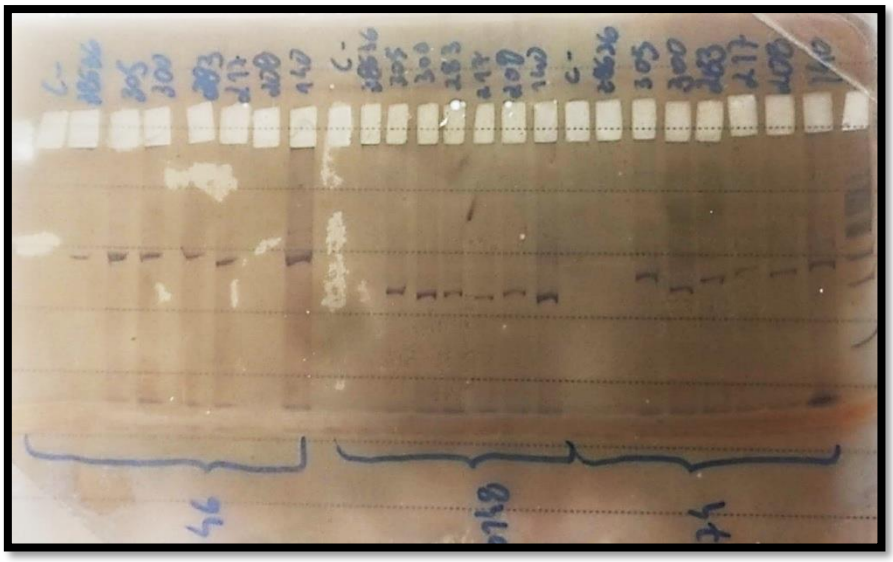


Figure 8. Polyacrylamide gel obtained for DXS10146, DXS10148 and DXS10074 (from left to right, respectively) after using new designed primers for amplification.

However, for the samples from Macau, although the same primers and conditions were used to perform the PCR, most amplifications for the majority of X-STRs were not successful. For example, no bands appeared in the eletrophoresis gel or some bands were barely visible (Figure 9). In addition, some contamination started to appear (even after new stocks were made for all reagents) (Figure 10).

Figure 9.Polyacrylamide gel obtained from amplification of Macanese samples (barely visible band for one sample and no amplification of other samples).

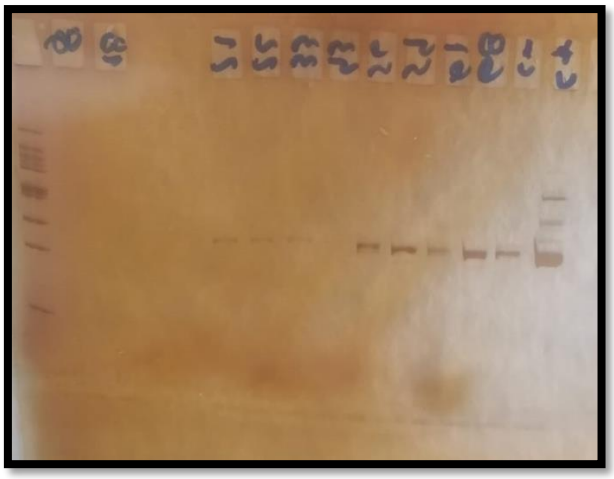


Figure 10. Polyacrylamide gel obtained from amplification of Macanese samples where some contamination is seen (due to negative control showing a band).

These DNA extracts were quite old and it was not their first use (at least 20 yrs of storage). Nevertheless, to make sure that the results obtained were not due to possible contaminants in the extract that inhibit the PCR process, quantification of DNA and contaminants were estimated using the NanoDrop One microvolume UV-Vis spectrophotometer (Thermo Scientific Scientific, USA). The obtained results (Attachment 3- Supplementary Table 3) showed that the quantity of DNA in the samples is acceptable. As for the quality control parameters, the A260/280 ratio were below (1.0-1.4) the 1.8 value recommended for high quality DNA denoting the presence of some proteins, most likely not enough to justify the non-amplification of the targeted loci. On the other hand,

the A260/230 were very low compared to the accepted 1.8-2.0 quality parameters, with values ranging from 0.1 to 0.3. This possibly indicates a high concentration and presence of inhibitors in the DNA extracts resulting from carry-over of reagents from the extraction procedure. This could be one of the reasons why the TD-PCR was failing. Another possibility is high degradation due to the long storage of samples as already mentioned (over 20 yrs). Despite this, the try and error method mentioned above was used by applying changes to the protocol, such as dilution of samples. The eletrophoresis results is shown in Figure 11. As ilustrated in Figure 11, the amplification failed even after some dilution of the samples, or showed some unwanted bands.

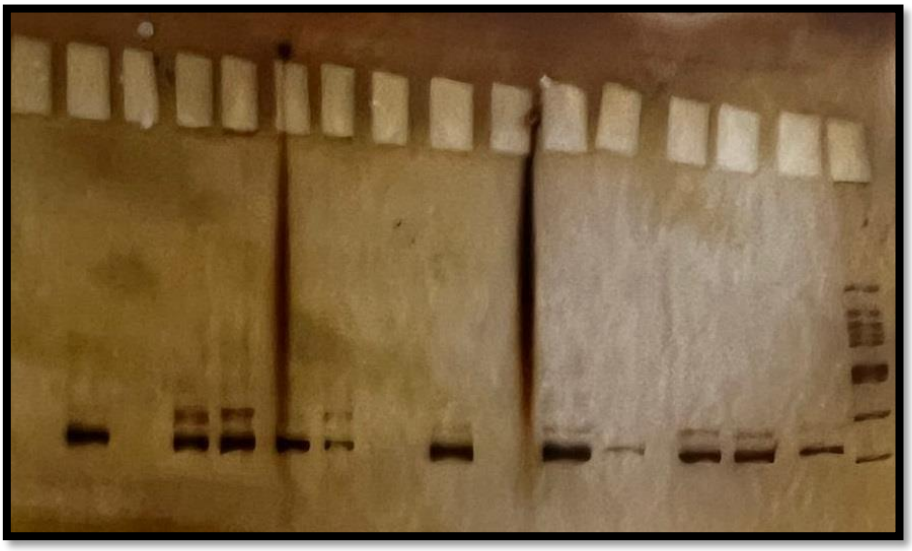


Figure 11. Polyacrylamide gel obtained after amplification of diluted samples.

Therefore, the strategy was going back to the original samples and try the amplification of the samples that were in good conditions for one genetic marker (DXS10174) until some results were obtained. This showed some results as seen in Figure 12. After some samples worked, those same samples were used for the amplification of the rest of the markers DXS10134, DXS10135 with some positive results, excluding the DXS10146 and DXS10148 (due to reasons explained on the next point of this dissertation). Even so, two of the samples that worked for the DXS10174 marker still failed for DXS10134 and DXS10135 markers. Due to this problem, few results were obtained for the Macanese population group (i.e., DNA sequences with good, readable quality, five for marker DXS10074, three for DXS10134 and DXS10135).

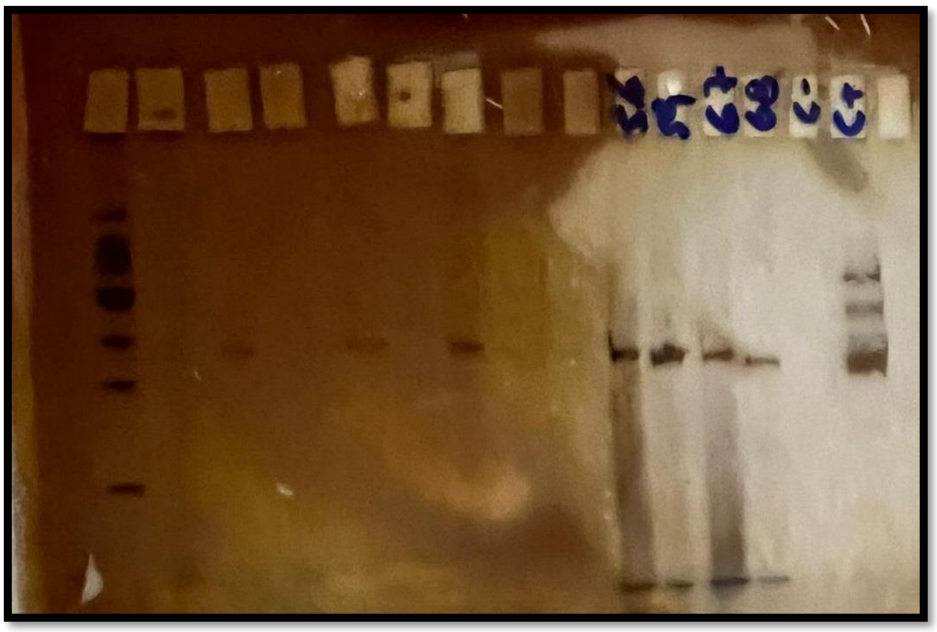


Figure 12. Good results obtained by trial and method error (for marker DXS10074).

After designing new primers (Table 4, in Material & Methods), the problems encountered for the marker DXS10146 in the Portuguese samples were solved, as the polyacrylamide gel obtained after amplification with these new primers did not show any unwanted amplicons and appeared clean (Figure 8). For the remaining markers DXS10074, DXS10134, DXS10135 and DXS10148 no obstacles were encountered regarding the amplification of the samples as the polyacrilamide gel obtained showed great results for subsequent Sanger sequencing (Figure 8,13 and 14).

Figure 13. Polyacrylamide gel obtained after using the new primers for DXS10135.

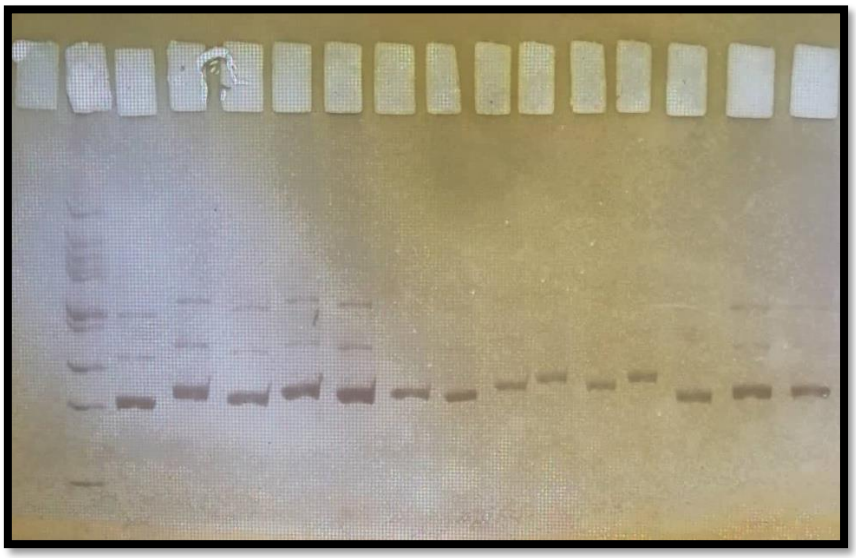


Figure 14. Polyacrylamide gel obtained for DXS10134 after using the new designed primers.

4.3. Sanger sequencing of DXS10074, DXS10134, DXS10135, DXS10146 and DXS10148

The first eletropherograms obtained from sequencing were for the marker DXS10146 (these tests were before the new primers were designed). The results were not great

because either it showed loss of the amplified product and no sequencing results or a lot of interference that was present which would make the reading of the DNA sequence hard. Using the new designed primers, the sequencing results showed no initial improvement. To eliminate the possibility of a technical user error, an attempt was done by another colleague, regarding the full protocol for the sequencing reaction. The same results were obtained, i.e., no sequence was visible in the eletropherograms. Then, it was suspected that the columns used for the second purification process using a Sephadex based *in-house* protocol were not in good conditions so, to eliminate this doubt, this purification process was done in the Genomics Core Facility-GenCore-available *in house*. The quality of the results obtained in this attempt confirmed that the bad sequencing observed previously was a consequence of the columns used in the laboratory. After using new columns for the purification by Sephadex, the markers DXS10074, DXS10134 and DXS10135 showed highly improved results where the eletropherograms obtained were clean. As for the marker DXS10146, a few Portuguese samples worked perfectly and good sequencing results were obtained. However, most samples still showed some problems with the eletropherograms obtained. For the marker DXS10148, the eletropherograms obtained for all samples were not readable. This pointed to the need of performing *in silico* tests for these two markers, DXS10146 and DXS10148.

The analyses of the **final allele nomenclature results** where, therefore, performed using two different sources of DNA sequences for the studied loci (as mentioned in the "Material and Methods" section):

1) *Wet-lab*: Direct analyses of sequencing data obtained for the European (Portugal) and Asian (Macau) samples;

2) *In silico (literature research, compilation and comparisons):* For the non-sucessful amplification of samples by Sanger Sequencing, sequence data for the European and Asian groups were selected from the literature (when available). For the African group, sequence data was always selected from the literature (when available).

After the compilation of data for each locus and for each population group (Table 3 in Material and methods), comparisons of allele nomenclature among groups were performed and the final human allele nomenclatures were assessed ("**Comparison between the three population groups: Europeans, Africans and Asians**"). These were subsequently compared with the chimpanzee genome and final inferences on the more accurate allele structures were proposed ("**Comparison with *Pan troglodytes* genome**"). Although no specific ISFG guidelines on the use of X chromosome markers

in forensics analysis are available, the existing recommendations on locus structure and allele nomenclature designation for autosomal STRs and Y-STRs can be extended to X-STRs (Bar et al., 1997; Gusmão et al., 2006) and were considered to discuss results. Some of the main suggestions for nomenclature designation should be taken into consideration (according to the ISFG guidelines):

- the identification of regions that (may likely) vary should be supported by sequence analysis of individuals from different human populations and with comparison with chimpanzee sequences;
- the first repetitive motif in a simple or complex sequence structure that encloses the highest number of repeats should be defined;
- alleles must be named in agreement with the total number of adjacent variant and non-variant repeats obtained from sequence data analysis;
- non-variant repeats should be included in the allele nomenclature if the number of nucleotides that separate these structures from the main variable repeat is equal or less to the number of nucleotides of the repeat unit;
- if a previously established nomenclature of an STR is not in accordance with the ISFG guidelines, but has been widely used, the nomenclature should not be changed to avoid unnecessary confusion.

4.4. Comparison between the three population groups : Europeans, Africans and Asians

4.4.1. DXS10074

Regarding the allele nomenclature of DXS10074, acomplex type of structure was initially described in the study of Hering and co-authors (Hering et al., 2005), with two different types of alleles:

- Short alleles are composed by the following repeat structure $(ACAC)_2$-$(AGAG)_4$-AA-AAAG-$(AAGA)_n$-AAGG-AAGA (n varies between 7 and 10)

- Longer alleles described as: $(ACAC)_2$-$(AGAG)_4$-AA-AAAG-$(AAGA)_n$-<u>AAGG-$(AAGA)_2$</u>-AAGG-AAGA (n varies between 10 and 18)

Due to the insertion of the motif AAGG-$(AAGA)_2$ (underlined above), the alleles that contain the motif mentioned should be caracterized as n+3. In a later study done by

Gomes and co-workers (Gomes et al., 2016) on African populations, a new variation of this genetic marker was found. In some intermediate alleles, additional AG nucleotides were detected outside the core repeat in the upstream flanking region, meaning this region was composed by two stretches of dinucleotide repeats AC and AG motifs. The new variation reported in this study was : $(AC)_{3-4}$-$(AG)_{8-9}$-AAAAAG-$(AAA)_{0-1}$-$(AAGA)_2$-$(AGA)_{0-1}$-$(AAGA)_n$-$[AAGG-(AAGA)_2]_{0-1}$-AAGG-AAGA. Thus, for the purpose of this study the latter nomenclature is going to be considered and used for comparison because it is the one that embraces more variation. In addition, this variation also serves for the African population since the data analysed in this study showed the same conclusion.

After analysing the sequences obtained in the current study for this marker, the general structure obtained for the European samples (Table 11) is $(AC)_4$-$(AG)_8$-N_6-$(AAGA)_n$-$[AAGG-(AAGA)_2]_{0-1}$-AAGG-AAGA. No significant deviation was seen in the alleles sequenced among the different individuals. In fact, for the European population represented by Portugal, the most common dinucleotide stretch structure observed, upstream to the repeat was $(AC)_4$-$(AG)_8$, supporting the findings in the study by Gomes et al. (Gomes et al., 2016). For these samples, there is not a lot of intra-population variation. For the Macanese samples (Table 12), although the number of alleles sequenced was low due to the quality of the samples, the structure observed was $(AC)_4$-$(AG)_8$-N_6-$(AAGA)_n$-$[(AAGG)-(AAGA)_2]_1$-AAGG-AAGA. This is also very similar to the ones observed for the Portuguese population group. Comparing all three structures found for the three populations, there is no significant variation among them.

Table 11. Sequenced data for marker DXS10074 obtained from Portuguese samples.

Allele sequenced	n	sequence structure of the repeat
8	2	(AC)4-(AG)7-N6-(AAGA)8-AAGG-AAGA
	3	(AC)4-(AG)8-N6-(AAGA)8-AAGG-AAGA
13	1	(AC)4-(AG)8-N6-(AAGA)10-[AAGG-(AAGA)2]1-AAGG-AAGA
14	2	(AC)4-(AG)8-N6-(AAGA)11-[AAGG-(AAGA)2]1-AAGG-AAGA
15	2	(AC)4-(AG)8-N6-(AAGA)12-[AAGG-(AAGA)2]1-AAGG-AAGA
	1	(AC)4-(AG)8-N6-(AAGA)15-AAGG-AAAGA
16	3	(AC)4-(AG)8-N6-(AAGA)13-[AAGG-(AAGA)2]1-AAGG-AAGA
17	2	(AC)4-(AG)8-N6-(AAGA)14-[AAGG-(AAGA)2]1-AAGG-AAGA
18	2	(AC)4-(AG)8-N6-(AAGA)15-[AAGG-(AAGA)2]1-AAGG-AAGA
19	2	(AC)4-(AG)8-N6-(AAGA)16-[AAGG-(AAGA)2]1-AAGG-AAGA

Table 12. Sequenced data for marker DXS10074 obtained from Macanese samples.

Allele sequenced	n	sequence structure of the repeat
15	1	(AC)4-(AC)8-N6-(AAGA)12-(AAGG)-(AAGA)2-AAGG-AAGA
16	1	(AC)4-(AC)8-N6-(AAGA)13-(AAGG)-(AAGA)2-AAGG-AAGA
17	1	(AC)4-(AG)8-N6-(AAGA)14-(AAGG)-(AAGA)2-AAGG-AAGA
18	2	(AC)4-(AG)8-N6-(AAGA)15-(AAGG)-(AAGA)2-AAGG-AAGA

4.4.2. DXS10134

According to Edelmann and co-workers (Edelmann et al., 2008), the DXS10134 is composed by a highly complex repeat structure described as: $(GAAA)_3$-GAGA-$(GAAA)_4$-AA-(GAAA)-GAGA-$(GAAA)_4$-GAGA-$(GACAGA)_3$-(GAAA)-GTAA-$(GAAA)_3$-AAA-$[(GAAA)_4$-$AAA]_{1-2}$-$(GAAA)_n$. In a later study done by Gomes and co-workers (Gomes et al., 2016) on African populations, new variants in the repeat structure with some considerable differences, when compared to the initial nomenclature, were found. Some alleles showed a simpler strucuture, with a loss of the $[(GACAGA)_3$-(GAAA)-GTAA-$(GAAA)_3$-AAA] part of the repeat structure initially described. So in this article (Gomes et al., 2016) the reported structure that outcame from the analyzed African population for DXS10134 was $(GAAA)_3$-GAGA-$(GAAA)_4$-AA-(GAAA)-GAGA-$(GAAA)_{3–4}$-[GAGA-$(GACAGA)_3$-(GAAA)-GTAA-$(GAAA)_3$-AAA-$[(GAAA)_4$-$AAA]_{1,3}]_{0–1}$-(GAAA)5-$(A)_{0–1}$-$(GAAA)_n$. This allele structure nomenclature considers higher variation in the DNA sequence when compared with the original sequence. For the purpose of this dissertation, the two reported structures will serve as a point of comparison.

After analysis of the sequenced samples from Portugal (Table 13), the general structure of the repeat obtained was as follows: $(GAAA)_3$-GAGA-$(GAAA)_4$-AA-(GAA)-GAGA-$(GAAA)_4$-GAGA-$(GACAGA)_3$-(GAAA)-GTAA-$(GAA)_3$-$[[(GAAA)_4$-$AAA]_{1-2}]_{0-1}$-$(GAAA)_5$-A_{0-1}-$(GAAA)_n$.

Allele 33 sequenced in this study, showed a loss of the $[(GAAA)_4$-AAA] part of the repeat, which was not reported in the first nomenclature proposed by Edelmann and co-workers but is present in the reports done by Gomes and co-workers (Table 15). In alleles 42.1 and 35.2, an insertion of one and two adenines, respectively, is located in the final portion of the repeat interrupting the $(GAAA)_n$ repeat (Table 13). For the Asian samples from Macau (Table 14), the structure observed was $(GAAA)_3$-GAGA-$(GAAA)_4$-AA-(GAAA)-GAGA-$(GAAA)_4$-GAGA-$(GACAGA)_3$-(GAAA)-GTAA-$(GAAA)_3$-$[(GAAA)4$-$AAA]_1$-$(GAAA)_n$. Although, the number of samples analysed is low, the structure obtained is, once again, very similar to the published reports. If we compare this structure obtained

with the one obtained from European samples, there is no significant differences. As for the African structure (Table 15), the analysis showed the same structure reported in Gomes et al. (Gomes et al., 2016) since the data gathered was from that work. In summary, there is no significant variation between all three structures found for the three populations and the more complete sequence reported by Gomes et al. (Gomes et al., 2016).

Table 13. Sequenced data for DXS10134 obtained from Portugueses samples.

Allele sequenced	n	sequenced structure of the repeat
33	1	(GAAA)3-GAGA-(GAAA)4-AA-(GAAA)-GAGA-(GAAA)4-GAGA-(GACAGA)3-(GAAA)-GTAA-(GAAA)3-AAA-(GAAA)14
34	4	(GAAA)3-GAGA-(GAAA)4-AA-(GAAA)-GAGA-(GAAA)4-GAGA-(GACAGA)3-(GAAA)-GTAA-(GAAA)3-AAA-[(GAAA)4-AAA]1-(GAAA)14
35	2	(GAAA)3-GAGA-(GAAA)4-AA-(GAAA)-GAGA-(GAAA)4-GAGA-(GACAGA)3-(GAAA)-GTAA-(GAAA)3-AAA-[(GAAA)4-AAA]1-(GAAA)15
35.2	1	(GAAA)3-GAGA-(GAAA)4-AA-(GAAA)-GAGA-(GAAA)4-GAGA-(GACAGA)3-(GAAA)-GTAA-(GAAA)3-AAA-[(GAAA)4-AAA]1-(GAAA)4-AA-(GAAA)11
36	3	(GAAA)3-GAGA-(GAAA)4-AA-(GAAA)-GAGA-(GAAA)4-GAGA-(GACAGA)3-(GAAA)-GTAA-(GAAA)3-AAA-[(GAAA)4-AAA]2-(GAAA)15
	2	(GAAA)3-GAGA-(GAAA)4-AA-(GAAA)-GAGA-(GAAA)4-GAGA-(GACAGA)3-(GAAA)-GTAA-(GAAA)3-AAA-[(GAAA)4-AAA]1-(GAAA)16
37	2	(GAAA)3-GAGA-(GAAA)4-AA-(GAAA)-GAGA-(GAAA)4-GAGA-(GACAGA)3-(GAAA)-GTAA-(GAAA)3-AAA-[(GAAA)4-AAA]1-(GAAA)17
38	1	(GAAA)3-GAGA-(GAAA)4-AA-(GAAA)-GAGA-(GAAA)4-GAGA-(GACAGA)3-(GAAA)-GTAA-(GAAA)3-AAA-[(GAAA)4-AAA]1-(GAAA)18
39	2	(GAAA)3-GAGA-(GAAA)4-AA-(GAAA)-GAGA-(GAAA)4-GAGA-(GACAGA)3-(GAAA)-GTAA-(GAAA)3-AAA-[(GAAA)4-AAA]2-(GAAA)18
	1	(GAAA)3-GAGA-(GAAA)4-AA-(GAAA)-GAGA-(GAAA)4-GAGA-(GACAGA)3-(GAAA)-GTAA-(GAAA)3-AAA-[(GAAA)4-AAA]1-(GAAA)19
42.1	1	(GAAA)3-GAGA-(GAAA)4-AA-(GAAA)-GAGA-(GAAA)4-GAGA-(GACAGA)3-(GAAA)-GTAA-(GAAA)3-AAA-[(GAAA)4-AAA]1-(GAAA)5-A-(GAAA)17

Table 14. Sequenced data for DXS10134 obtained from Macanese samples.

Allele sequenced	n	sequenced structure of the repeat
32	1	(GAAA)3-GAGA-(GAAA)4-AA-(GAAA)-GAGA-(GAAA)4-GAGA-(GACAGA)3-(GAAA)-GTAA-(GAAA)3-AAA-[(GAAA)4-AAA]1-(GAAA)12
33	1	(GAAA)3-GAGA-(GAAA)4-AA-(GAAA)-GAGA-(GAAA)4-GAGA-(GACAGA)3-(GAAA)-GTAA-(GAAA)3-AAA-[(GAAA)4-AAA]1-(GAAA)13
36	1	(GAAA)3-GAGA-(GAAA)4-AA-(GAAA)-GAGA-(GAAA)4-GAGA-(GACAGA)3-(GAAA)-GTAA-(GAAA)3-AAA-[(GAAA)4-AAA]1-(GAAA)16

Table 15. Published data for marker DXS10134 obtained from "New sequence variants detected at DXS10148, DXS10074 and DXS10134 loci" (Gomes et al., 2016).

Allele Genotyped	Allele Sequenced	n	Sequence structure of the repeat
34	34	1	(GAAA)3-GAGA-(GAAA)4-AA-(GAAA)-GAGA-(GAAA)4-GAGA-(GACAGA)3-(GAAA)-GTAA-(GAAA)3-AAA-[(GAAA)4-AAA]1-(GAAA)14
35	35	1	(GAAA)3-GAGA-(GAAA)4-AA-(GAAA)-GAGA-(GAAA)4-GAGA-(GACAGA)3-(GAAA)-GTAA-(GAAA)3-AAA-[(GAAA)4-AAA]1-(GAAA)15
43.2	43.2	1	(GAAA)3-GAGA-(GAAA)4-AA-(GAAA)-GAGA-(GAAA)3-GAGA-(GACAGA)3-(GAAA)-GTAA-(GAAA)3-AAA-[(GAAA)4-AAA]3-(GAAA)15
39.1	39.1	1	(GAAA)3-GAGA-(GAAA)4-AA-(GAAA)-GAGA-(GAAA)4-GAGA-(GACAGA)3-(GAAA)-GTAA-(GAAA)3-AAA-[(GAAA)4-AAA]1-(GAAA)5-A-(GAAA)14
15	23	5	GAAA)3-GAGA-(GAAA)4-AA-(GAAA)-GAGA-(GAAA)15
16	24	1	(GAAA)3-GAGA-(GAAA)4-AA-(GAAA)-GAGA-(GAAA)16

4.4.3. DXS10135

The most common and established nomenclature for this X-STR is a compound type of structure initially described by Becker et al. (Becker et al., 2008) and later supplemented by additional data from Gomes et al. (Gomes et al., 2017). Different structures according to the different alleles were described most commonly as:

Alleles16-25: $(TTTC)_{13-22}$ CTTTC (TCTT)3

Alleles 19-24: TTTC-TTCC $(TTTC)_{14-16}$ CTTTC $(TCTT)_3$

Alleles 25-35: TTTC-TTCC (TTTC)3 TTCC $(TTTC)_{16-26}$ CTTTC $(TCTT)_3$

Alleles 12.1-23.1: **del (TCT)** …. $(TTTC)_{8-21}$ CTTTC $(TCTT)_3$ (Deletion seen in Figure 15)

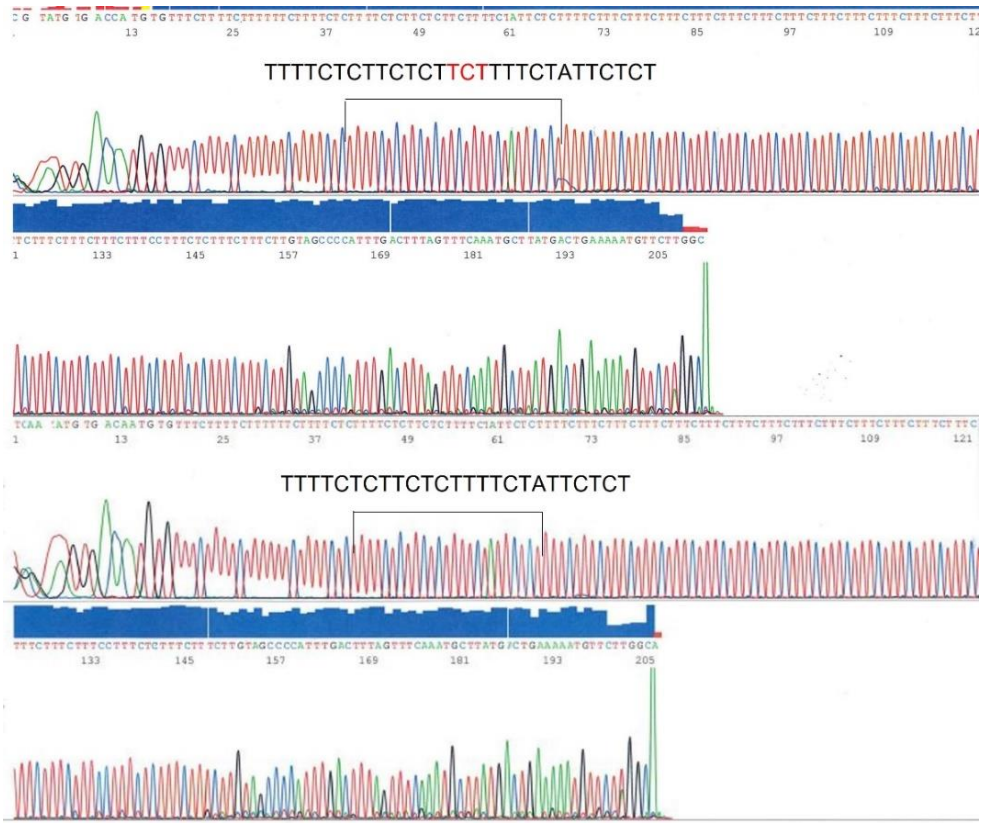


Figure 15. Deletion of TCT motif on intermediate alleles found after sequencing of samples for DXS10135 (top electropherogram is a normal allele, down electropherogram is an intermediate allele).

Regarding samples from Portugal that were sequenced and studied (Table 16), the allele structure obtained was similar to the one described above. Several interruptions of the TTTC main repeat motif can be observed which defines different allele ranges (TTTC)-$(TTCC)_{0-4}$-$(TTTC)_3$-$(TTCC)_{0-1}$-$(TTTC)_n$-CTTTC-$(TCTT)_3$.

Table 16. Sequenced data for marker DXS10135 from portuguese samples.

Allele sequenced	n	sequenced structure of the repeat
17	1	(TTTC)14-CTTTC-(TCTT)3
17.1	1	del (TCT)….(TTTC)15-(CTTTC)-(TCTT)3
19	3	(TTTC)16-CTTTC-(TCTT)3
19.1	3	del (TCT)….(TTTC)17-CTTTC-(TCTT)3
21	2	(TTTC)18-CTTTC-(TCTT)3
22	1	(TTTC)1-(TTCC)3-(TTTC)15-CTTTC-(TCTT)3
23	1	(TTTC)20-CTTTC-(TCTT)3
24	1	(TTTC)1-(TTCC)3-(TTTC)17-CTTTC-(TCTT)3
	1	(TTTC)-TTCC-(TTTC)3-TTCC-(TTTC)15-CTTTC-(TCTT)3
27	1	(TTTC)1-TTCC-(TTTC)3-TTCC-(TTTC)18-CTTTC-(TCTT)3
28	1	(TTTC)-(TTCC)4-(TTTC)20-CTTTC-(TCTT)3
29	2	(TTTC)1-TTCC-(TTTC)3-TTCC-(TTTC)20-CTTTC-(TCTT)3

In alleles 22, 24, 27, 28 and 29, a transition from a T to C in the main repeat motif $(TTTC)_n$, causes an interruption of this repeat, creating a more complex structure when compared to the previous established nomenclature.

Interestingly, several intermediate alleles (17.1 and 19.1 in our Portuguese sample set) genotyped by fragment analyses revealed after sequencing no interruption of the core repeat DNA sequence. The mutations responsible for the – 3 bp were detected in the upstream flanking region of the repeat caused by a TCT deletion (Table 16). This deletion had already been described in other studies as well (Gomes et al., 2017). There are some intra-population variation due to this mutation, i.e., alleles with the same size have different sequence compositions. As for the rest of the alleles analysed, the structure is the same as the established one. For Macau samples (Table 17) the general structure seen is: (TTCC)-$(TTCC)_{0-1}$-$(TTTC)_3$-$(TTCC)_{0-1}$,$(TTTC)_n$-(CTTTC)-$(TCTT)_3$. Pretty similar to the one observed in Portuguese samples, with no significant observable variation. As for African samples, obtained from the literature (Gomes et al., 2017), the obtained allele structure is $(TTTC)_n$-T_{0-1}-$(TC)_{0-1}$-$(TTTC)_n$-CTTTC-$(TCTT)_3$. The intermediate alleles studied showed an insertion of either a single T base or a dinucleotide insertion TC. These findings show us that an update of the established nomenclature should be performed for DXS10135 due to the variations found in this study where some more complex structures appear regarding some alleles.

Table 17. Sequenced data for marker DXS10135 obtained from Macanese samples.

allele sequenced	N	sequenced structure of the repeat
19	1	(TTTC)16-CTTTC-(TCTT)3
20	1	(TTTC)17-CTTTC-(TCTT)3
30	1	(TTTC)1-(TTCC)-(TTTC)3-(TTCC)-(TTTC)21-CTTTC-(TCTT)3

4.4.4. DXS10146

This STR was initially described as having a compound type of structure containing two variable simple, repeat blocks $(TTCC)_x$ and $(CTTT)_y$ with 10 non-variable 4 bp repeats (Edelmann et al.,2007). The established and most used nomenclature for this marker proposed in the article mentioned is: $(TTCC)_x$-T-$(TTCC)_4$-N_8-$(TCCC)_2$-N_{10}-$(TTCC)_2$-N_6-$(CTTT)_Y$-$(CTTT)_2$. Longer alleles showed more variation in the first variable motif and the second variable block showed variation as a result of a two base pair insertion (CT or TT) (Edelmann et al.,2007). For this dissertation, the nomenclature mentioned above will

be used as a point of comparison with the chimpanzee genome. The structure found on the Portuguese samples (Table 18) analyzed in this study was the same structure as the established and most commonly used one, with no variation. The African data analised obtained from (Gomes et al., 2017) showed the same structure encountered in the Portuguese samples.

Table 18. Sequenced data for marker DXS10146 obtained from portuguese samples.

Allele sequenced	n	sequenced structure of the repeat
27	2	(TTCC)3-T-(TTCC)4-TCCCTTCC-(TCCC)2-TTCTTCTTTC-(TTCC)2-TTTCTT-(CTTT)14-T-(CTTT)2
28	2	(TTCC3-T-(TTCC)4-TCCCTTCC-(TCCC)2-TTCTTCTTTC-(TTCC)2-TTTCTT-(CTTT)15-T-(CTTT)2
29	1	(TTCC)3-T-(TTCC)4-TCCCTTCC-(TCCC)2-TTCTTCTTTC-(TTCC)2-TTTCTT-(CTTT)16-T-(CTTT)2
31	1	(TTCC)3-T-(TTCC)4-TCCCTTCC-(TCCC)2-TTCTTCTTTC-(TTCC)2-TTTCTT-(CTTT)28-T-(CTTT)2

4.4.5. DXS10148

This repeat was described as $(GGAA)_4$-$(AAGA)_X$-$(AAAG)_Y$-N_8-$(AAGG)_2$ with all tetrameric repeats being considered for allele assignment (Hundertmark et al., 2008). In a study done by Gomes and co-workers (Gomes et al., 2016), additional eight bases samples with three copies of (AAGG-AAAG) instead of two [$(AAGG\text{-}AAAG)_2$] were identified at the beginning of the repeat motif in some samples when compared to the original nomenclature (Hundertmark et al., 2008). Also some null alleles where detected due to a mutation that corresponds to the second base of the 3' of the forward primer (Gomes et al., 2016), which potentially lead to a innacurate nomenclature. In another paper (Gomes et al., 2017) this upstream variation was also found.

With that said, the reported new allele structure found and suggested as being the most accurate (Gomes et al., 2016) is: $(GGAA)_{2,4}$ $(AGAA\text{-}GGAA)_{0–1}$ $(AAGA)_X$ $(AAAG)_{0,Y}$ $(AAAG\text{-}A)_{0–2}(AAAG)_4$-$N_8$-$(AAGG)_2AAAG(AAGG\text{-}AAAG)_{1–3}$. For the purpose of this dissertation, the newly reported nomenclature will be used as reference.

The analyses of the African data presented in Table 19 shows a high number of variation regarding the allelic structure of this STR, in specific high intra-variation for the non-consensus alleles were found. In Asian samples obtained from published data (Samejima et al., 2012; Sim et al., 2009), even though the data does not contain information on the upstream region of the repeat, there is an inconsistency with the findings of this study and the ones by Gomes and co-workers (Gomes et al., 2016, 2017),

so the general structure obtained was: $(GGAA)_4$-$(AAGA)_Y$-A_{0-1}-(AAGA)-$(AAAG)_X$-N8-$(AAGG)_2$.

The results obtained on the genetic variation of this STR shows that the nomenclature most commonly used for forensic casework lacks the variation in the upstream flanking region of the repeat and as a result an update containing the new variation should be added.

Table 19. Published data for marker DXS10148 obtained from "Genetic characterization of Guinea-Bissau using a 12 X-chromosomal STR system: Inferences from a multiethnic population" (Gomes et al., 2017).

Allele Sequenced	n	Sequence structure of the repeat
35.1	2	(CTTTCCTT)2-CTTT-(CCTT)2-CTTT-CCTT-(CTTT)4 - (TCTT)13-T-(TCTT)12-(TTCC)-(TTCT)-(TTCC)2
	2	(CTTTCCTT)2-CTTT-(CCTT)2-CTTT-CCTT-(CTTT)4 - (TCTT)12-T-(TCTT)13-(TTCC)-(TTCT)-(TTCC)2
37.1	1	(CTTTCCTT)2-CTTT-(CCTT)2-CTTT-CCTT-(CTTT)4 - (TCTT)11-T-(TCTT)16-(TTCC)-(TTCT)-(TTCC)2
	1	(CTTTCCTT)2-CTTT-(CCTT)2-CTTT-CCTT-(CTTT)4 - (TCTT)12-T-(TCTT)15-(TTCC)-(TTCT)-(TTCC)2
	1	(CTTTCCTT)2-CTTT-(CCTT)2-CTTT-CCTT-(CTTT)4 - (TCTT)13-T-(TCTT)14-(TTCC)-(TTCT)-(TTCC)2
	3	(CTTTCCTT)2-CTTT-(CCTT)2-CTTT-CCTT-(CTTT)4 - (TCTT)14-T-(TCTT)13-(TTCC)-(TTCT)-(TTCC)2
	1	(CTTTCCTT)2-CTTT-(CCTT)2-CTTT-CCTT-(CTTT)4 - (TCTT)15-T-(TCTT)12-(TTCC)-(TTCT)-(TTCC)2
38.1	3	(CTTTCCTT)2-CTTT-(CCTT)2-CTTT-CCTT-(CTTT)4 - (TCTT)13-T-(TCTT)15-(TTCC)-(TTCT)-(TTCC)2
	2	(CTTTCCTT)2-CTTT-(CCTT)2-CTTT-CCTT-(CTTT)4 - (TCTT)14-T-(TCTT)14-(TTCC)-(TTCT)-(TTCC)2
	1	(CTTTCCTT)2-CTTT-(CCTT)2-CTTT-CCTT-(CTTT)4 - (TCTT)15-T-(TCTT)13-(TTCC)-(TTCT)-(TTCC)2
39.1	1	(CTTTCCTT)2-CTTT-(CCTT)2-CTTT-CCTT-(CTTT)4 - (TCTT)12-T-(TCTT)17-(TTCC)-(TTCT)-(TTCC)2
	3	(CTTTCCTT)2-CTTT-(CCTT)2-CTTT-CCTT-(CTTT)4 - (TCTT)13-T-(TCTT)16-(TTCC)-(TTCT)-(TTCC)2
	1	(CTTTCCTT)2-CTTT-(CCTT)2-CTTT-CCTT-(CTTT)4 - (TCTT)16-T-(TCTT)13-(TTCC)-(TTCT)-(TTCC)2
	2	(CTTTCCTT)2-CTTT-(CCTT)2-CTTT-CCTT-(CTTT)4 - (TCTT)14-T-(TCTT)15-(TTCC)-(TTCT)-(TTCC)2
	3	(CTTTCCTT)2-CTTT-(CCTT)2-CTTT-CCTT-(CTTT)4 - (TCTT)15-T-(TCTT)14-(TTCC)-(TTCT)-(TTCC)2
	1	(CTTTCCTT)2-CTTT-(CCTT)2-CTTT-CCTT-(CTTT)4 - (TCTT)17-T-(TCTT)12-(TTCC)-(TTCT)-(TTCC)2
40.1	1	(CTTTCCTT)2-CTTT-(CCTT)2-CTTT-CCTT-(CTTT)4 - (TCTT)11-T-(TCTT)19-(TTCC)-(TTCT)-(TTCC)2

	1	(CTTTCCTT)2-CTTT-(CCTT)2-CTTT-CCTT-(CTTT)4 - (TCTT)12-T-(TCTT)18-(TTCC)-(TTCT)-(TTCC)2
	8	(CTTTCCTT)2-CTTT-(CCTT)2-CTTT-CCTT-(CTTT)4 - (TCTT)14-T-(TCTT)16-(TTCC)-(TTCT)-(TTCC)2
	4	(CTTTCCTT)2-CTTT-(CCTT)2-CTTT-CCTT-(CTTT)4 -(TCTT)15-T-(TCTT)15-(TTCC)-(TTCT)-(TTCC)2
	2	(CTTTCCTT)2-CTTT-(CCTT)2-CTTT-CCTT-(CTTT)4 - (TCTT)10-(TCTA)-(TCTT)4-T-(TCTT)15-(TTCC)-(TTCT)-(TTCC)2
	1	(CTTTCCTT)2-CTTT-(CCTT)2-CTTT-CCTT-(CTTT)4 - (TCTT)16-T-(TCTT)14-(TTCC)-(TTCT)-(TTCC)2
41.1	2	(CTTTCCTT)2-CTTT-(CCTT)2-CTTT-CCTT-(CTTT)4 - (TCTT)14-T-(TCTT)17-(TTCC)-(TTCT)-(TTCC)2
	1	(CTTTCCTT)2-CTTT-(CCTT)2-CTTT-CCTT-(CTTT)4 - (TCTT)9-(TCTA)-(TCTT)4-T-(TCTT)17-(TTCC)-(TTCT)-(TTCC)2
	5	(CTTTCCTT)2-CTTT-(CCTT)2-CTTT-CCTT-(CTTT)4 - (TCTT)15-T-(TCTT)16-(TTCC)-(TTCT)-(TTCC)2
	1	(CTTTCCTT)1 -CTTT- (CCTT-CTTT)-CTTT-CCTT-(CTTT)4 - (TCTT)15-T-(TCTT)16-(TTCC)-(TTCT)-(TTCC)2
	1	(CTTTCCTT)2-CTTT-(CCTT)2-CTTT-CCTT-(CTTT)4 - (TCTT)16-T-(TCTT)15-(TTCC)-(TTCT)-(TTCC)2
42.1	3	(CTTTCCTT)2-CTTT-(CCTT)2-CTTT-CCTT-(CTTT)4 - (TCTT)16-T-(TCTT)16-(TTCC)-(TTCT)-(TTCC)2
	2	(CTTTCCTT)2 -CTTT-(CCTT)2-CTTT-CCTT-(CTTT)4 - (TCTT)15-T-(TCTT)17-(TTCC)-(TTCT)-(TTCC)2
43.1	1	(CTTTCCTT)2-CTTT-(CCTT)2-CTTT-CCTT-(CTTT)4 - (TCTT)16-T-(TCTT)17-(TTCC)-(TTCT)-(TTCC)2
44.1	1	(CTTTCCTT)2-CTTT-(CCTT)2-CTTT-CCTT-(CTTT)4 - (TCTT)16-T-(TCTT)18-(TTCC)-(TTCT)-(TTCC)2
45.1	1	(CTTTCCTT)2-CTTT-(CCTT)2-CTTT-CCTT-(CTTT)4 - (TCTT)17-T-(TCTT)18-(TTCC)-(TTCT)-(TTCC)2

Table 20. Sequenced data for DXS10148 obtained in West African countries (obtained from (Gomes et al., 2016)).

Al. seq.	n	Individual allele structures detected:
21*	1	(CTTTCCTT)3-CTTT-(CCTT)2-N8-(CTTT)4-(TCTT)11-(TTCC)4
27*	1	(CTTTCCTT)3-CTTT-(CCTT)2-N8-(CTTT)4-(TCTT)17-(TTCC)4
30*	1	(CTTTCCTT)3-CTTT-(CCTT)2-N8-(CTTT)4-(TCTT)20-(TTCC)4
37.1	1	(CTTTCCTT)3-CTTT-(CCTT)2-N8-(CTTT)4-A-(CTTT)12-(TCTT)15-TTCCTTCT-(TTCC)2
38.1	1	(CTTTCCTT)3-CTTT-(CCTT)2-N8-(CTTT)4-A-(CTTT)13-(TCTT)15-TTCCTTCT-(TTCC)2
39.1	2	(CTTTCCTT)2-CTTT-(CCTT)2-N8-(CTTT)4-A-(CTTT)14-(TCTT)15-TTCCTTCT-(TTCC)2
41.1	1	(CTTTCCTT)-CTTT-(CCTT)2-N8-(CTTT)4-A-(CTTT)15-(TCTT)16-TTCCTTCCT-(TTCC)2

4.5. Comparison with *Pan troglodytes* genome

4.5.1. DXS10074

Some differences were revealed when we compared and aligned the human DNA sequence of this genetic marker with the chimpanzee genome (Figure 16). On a first note, no differences were encountered in the flanking regions of the repeat region or on the primer binding sites (primers used are underlined in bold in Figure 16) between the two sequences, which means that the primers used for humans can be applied to amplify both human and chimpanzee samples. The allelic structure found in chimpanzee was $(AC)_5$-$(AG)_7$-AA-(AG)-AAAAAG-A-$(AAGA)_9$-GA-$(AAGA)_4$-GA. The chimpanzee structure contains an extra dinucleotide AC but, on the other hand, it has one less dinucleotide AG when compared to the human DNA sequence, as almost as a "compensation" for the fragment size. It also contains an insertion of a dinucleotide AA that interruptes the motif (AG) of the struture. Interestingly, the *Pan troglodytes* sequence shows an insertion of a dinucleotide GA that interrupts the main repeat motif $(AAGA)_n$ when compared to the *Homo sapiens* allele repeat structure, which shows a much more continuous repeat core of AAGA. The final part of the repeat shows again some differences between both species where more variation is seen in the chimpanzee compared to the AAGG-AAGA structure that seems stable among humans according to the sequences obtained (Tables 11 and 12). Apart from the differences mentioned at the repeat region, the structure of the STR remained quite stable as no differences between both species was detected for the analyzed amplicons.

```
Human TGCCATTTGTTGTGATCTCCAGTACATGGACTTCCTACTGCCCCACCTTT
Chimp ..................................................

Human ATTGTGTGTGTGCATGCAT
Chimp ...................

Human  (AC)4 (AG)8 AAAAAG (AAGA)14 AAGG-AAGA
Chimp  (AC)5 (AG)7 AA (AG) AAAAAG A (AAGA)9 GA (AAGA)4 GA

Human AAATAGAACAAATCAGCTTATATTCAGTATTT
Chimp ............................

Human TTTTAGTATTTTCTGTGTCAGCTCTCTGAGGG
Chimp ............................
```

Figure 16. Aligned sequences of human and chimpanzee for marker DXS10074 (Primers used are underlined bold, and repeat is in bold).

4.5.2. DXS10134

Alignment between human and chimpanzee reference sequences for DXS10134 (Figure 17) showed some differences. However, in the primer binding sites (in bold and underlined in Figure 17) there is no variation observed therefore we can conclude that the same primers used in this study for amplification of the human sequence can be used to amplify *Pan troglodytes* samples.

The chimpanzee structure observed as a result of the alignment is: $(GAAA)_3$-$(GAGA)_2$-$(GAAA)_4$-AA-(GAAA)-(GAGA)-$(GAAA)_2$-$(GAGA)_2$-GA-$(GACAGA)_2$-GAAA-GTAA-$(GAAA)_3$-AAA-$(GAAA)_3$-A-(GAAA)-AA-$(GAAA)_3$-A-(GAAA)-GAA-$(GAAA)_n$. Just as in humans this is a highly complex allele sequence full of interruptions of the possibly main GAAA repeat motif.

As for the allelic repeat sequence, there is one copy of the GAGA motif in the initial part of the human allele sequence leading to a difference on the number of this motif because chimpanzee possess two copies. Further, the *Pan troglodytes* sequence has a possible deletion of two GAAA motifs (or several single nucleotide base substituitions occurred interrupting this motif) when compared with the human sequence leading to a discrepancy of the number of this stretch being therefore $(GAAA)_2$. In fact, humans have $(GAAA)_{3-4}$ demonstrating variation at this level. However, it is not possible to know if variation occurs also among different chimpanzees as we have only access to the genome reference of *Pan troglodytes*. In this particular case, the established nomenclature has this part of the repeat labeled as stable $(GAAA)_4$ but the proposed nomenclature by Gomes and co-workers (Gomes et al., 2016) has a varying $(GAAA)_{3-4}$. The sequences obtained in this study for Portugal and Macau samples presented the established nomenclature, but in data gathered from African populations, the $(GAAA)_{3-4}$ variation was present (Table 15). Since variation is occuring in the human genomes, an update at this specific region regarding the repeat should be made. This would allow to have a better and inclusive nomenclature for the DXS10134.

Continuing on the analysis of the STR structure, there is either a deletion of an (GAGA) on the human sequence and also an insertion on the chimpanzee genome or a deletion on the human counterpart of a dinucleotide GA before the (GACAGA) part of the repeat. In the (GACAGA) part of the STR, the chimpanzee shows a loss of one full (GACAGA) motif, as humans have three repeat motifs (GACAGA) and *Pan troglodytes* only have two. To conclude the analysis there are some single and dinucleotide base insertions at the final part of the repeat in the chimpanzee sequence.

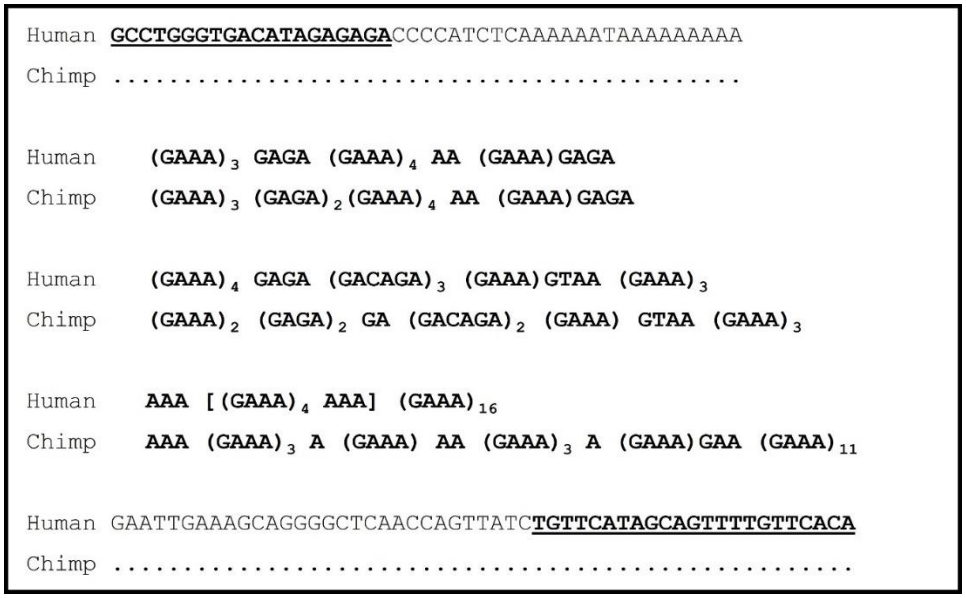


Figure 17. Aligned sequences of human and chimpanzee for marker DXS10134 (primers used are in underline bold and repeat structure in bold).

4.5.3. DXS10135

When both human and chimpanzee sequences for this genetic marker were aligned (Figure 18), the structure found for the chimpanzee was: $(AAGA)_3$-GAAA-G-$(GAAA)_4$-AA-$(GAAA)_n$. As for the repeat structure, no differences were detected besides an insertion on the chimpanzee sequence of a dinucleotide AA that interrupts the main repeat motif (GAAA) that is not present in the human sequences. What is interesting about this alignment is that the downstream flanking region of the repeat shows several differences among species: a single base substituion is observed (highlighted in blue in Figure 18) and a 5 bp deletion, on the chimpanzee sequence, followed by two single base substituitions (noted in blue in Figure 18) is also present. This means that the primers used for the amplification of the human sequences in this study would most likely not amplify samples from chimpanzees due to the considerable mismatch of the binding primer reverse site.

Figure 18. Aligned sequences of human and chimpanzee genomes for marker DXS10135 (Primers used in underlined bold, repeat is in bold and differences encountered are in blue).

4.5.4. DXS10146

Resulting from the alignment of both sequences for this genetic marker, the chimpanzee structure obtained was: $(TTCC)_7$-T-$(TTCC)_4$-$(TCCCTTCC)_2$-$(TCCC)_2$-TTCTT-TC-$(TTCC)_2$-TTTCTT-$(CTTT)_2$-TTCTT-$(CTTT)_3$-CCTT-$(CTTT)_3$. It seems that several substitutions/deletions have masked the possible ancestral TTCC and CTTT varying motifs creating a more complex structure.

Going into details, compared to the established nomenclature for the human X-STR, the chimpanzee allelic structure has an extra (TCCCTTCC) motif and it also has a loss of a trinucleotide CTT in the TTCTTCTTTC part of the repeat (not considered for allele counting) when compared to the human sequence. The chimpanzee also possesses a possible transition of a T to a C interrupting the $(CTTT)_y$ repeat motif (seen in Figure 19).

Regarding the flanking regions, there is a transition from a T to a C (or vice versa) in the forward primer and no mismatches at the reverse primer binding sites. Possibly, these primers could also be used to amplify this locus in chimpanzees due to the location of the mutation (middle of the forward sequence) that would not interfer with amplification.
A TCTT deletion downstream to the repeat is observed in chimpanzees when compared to the human sequence, while a large CTTCTCTT deletion in the human sequence is present also further downstream to the repeat region.

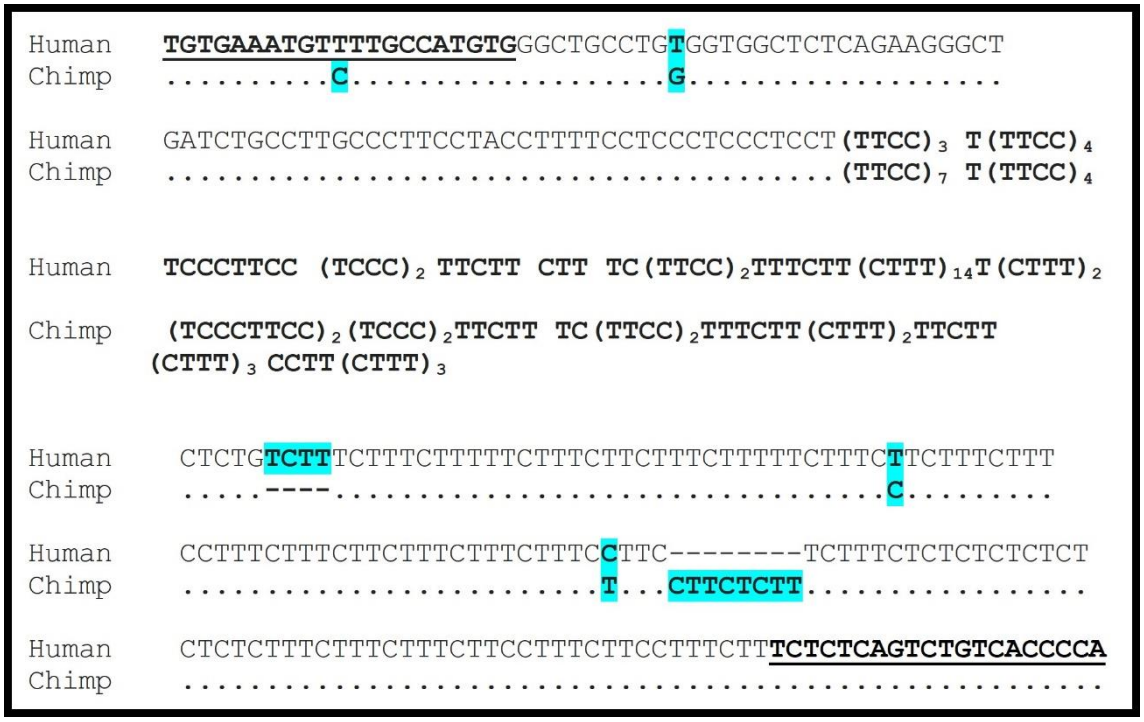


Figure 19. Aligned sequences of human and chimpanzee genomes for marker DXS10146 (Primers used in underlined bold, repeat is in bold and differences are noted in blue).

4.5.5. DXS10148

For this genetic marker, the main differences between the established nomenclature [$(GGAA)_4$-$(AAGA)_X$-$(AAAG)_Y$-(AAGG-AAAG)-$(AAGG)_2$AAAG(AAGG-$AAAG)_2$] (Hundertmark et al., 2008) and the one encountered by Gomes and co-workers, after a thorough analyses of West African samples [$(GGAA)_{2,4}$-$(AGAA\text{-}GGAA)_{0\text{-}1}$-$(AAGA)_X(AAAG)_{0,Y}$ $(AAAG\text{-}A)_{0\text{-}2}(AAAG)_4$-(AAGG-AAAG)-$(AAGG)_2$ AAAG (AAGG-$AAAG)_{1\text{-}3}$] (Gomes et al., 2016), was seen at the beginning and end of the repeat region where considerable variation was observed. When aligned, the human and chimpanzee genomes (Figure 20), revealed also interesting differences as the human and chimpanzee sequences both have the three copies of (AAGG- AAAG). Curiously, these findings were only observed when analyzing samples of West African origin (Gomes et al., 2016) and also found in this study when we analysed data gathered from the literature (Table 20). The chimpanzee reference sample here used also has the three copies (AAGG- AAAG) and therefore together with the findings in West Africa support an ancestral state of this portion of the repeat sequence.

On top of that huge difference, the repeat structure overall appears to have slightly different flanking regions to the beginning of the repeat since the (GGAA), first motif is

flanked by a long stretch of 14 poly A' and three G's (AAAAAAAAAAAAAAGGG) on the human sequence while in chimpanzee's only four A's flankthe repeat and thus shows a del (AAAAGGG) (Figure 20).

No differences other differences in the STR sequence is seen, specifically, that do not interfere with the primer binding sites meaning that the primers used for this study would amplify chimpanzee samples. The presence of varying copies of (AAGG-AAAG) when comparing different populations as well the chimpanzee sequence urges a must needed update on the established and most commonly used nomenclature for this genetic marker since a global and accurate nomenclature must be established to better report and analyse the data.

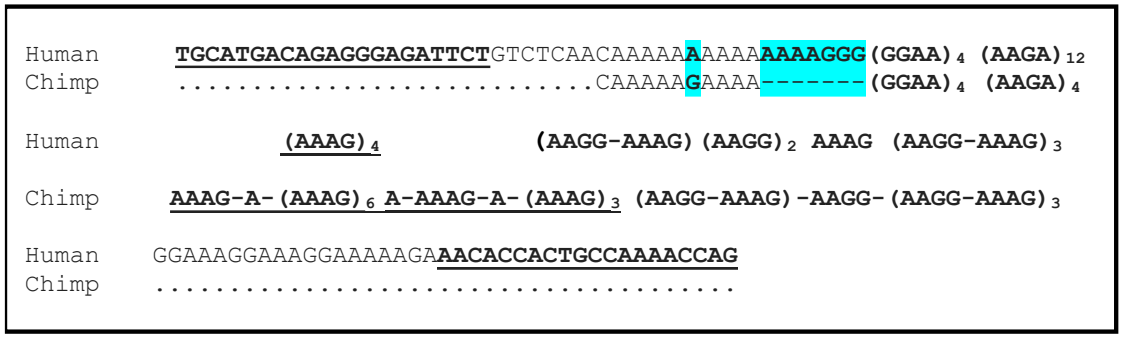


Figure 20. Aligned sequences of human and chimpanzee genomes for marker DXS10148 (Primers used are underline bold, differences between the 2 sequences are in blue).

5. Final Conclusions

The importance of an accurate and common nomenclature are very important to secure a clear communication, accurate data exchange and data comparisons among laboratories. STR nomenclature, independently of the marker (autosomal, X or Y chromosomal), has been pointed out by several studies as well as by the ISFG and other DNA groups (as has been mentioned in the introduction section of this dissertation).

The main objective of this work was to analyse the possible hidden variation that may exist at the DNA sequence allelic structure in some X-chromosome STRs and suggest new nomenclatures, if that is the case. The markers selected for this study were some of the common markers that are used in the fields of population and forensic genetics and that are also included in the main commercial X-STR kit (The Investigator X-12 QS, Qiagen). The results obtained on the X-STRs DXS10074, DXS10134, DXS10135, DXS10146 and DXS10148 throughthout this project are insightful and shed a light on the problems that exist in the genetic forensics field, mainly about the allele nomenclature of some of these loci.

The analysis of the DNA sequence structures of the three major population groups, namely Europeans, Asians and Africans obtained from sequenced samples and the ones gathered from published data allowed for the detection of possible discrepancies. These could include undetected mutations or silent alleles or variants previously not found by comparing the commonly used allele nomenclatures and the analysed samples for the studied markers, which answers one of the proposed aims for this work. One of the novelties that this work presents is the sequencing of Portuguese samples for the mentioned X-STRs, which helped have a better understanding on the selected loci in that particular population.

Additionally, comparing the chimpanzee (*Pan troglodytes*) and human (*Homo sapiens*) genomes, by aligning both sequences, which until to date have not been yet reported, helped perceive and infer the ancestral STR structure and stability. The results obtained from this alignment also allows us to see the evolution that the STRs underwent.

Population and species comparisons at the DNA sequence level with the allele nomenclatures currently used for DXS10135 and DXS10148 are not the most accurate and representative. The findings of this study, suport that a review of the established

nomenclature, at least for these two markers, should be done in order to be more accurate and inclusive due to possible silent alleles or mutations not found or reported when the most used nomenclature was established.

Hopefully the data obtained in this work elucidates the need of these type of studies to improve the vast field of forensic genetics by looking upon the established nomenclature for the commonly used genetic markers, to better report the findings and future works by following the recommended guideline and therefore being accurate and inclusive.

6. References

Aly, S. M., & Sabri, D. M. (2015). Sekwencjonowanie nowej generacji (NGS): “złote narzędzie” w arsenale metod kryminalistycznych Next generation sequencing (NGS): a golden tool in forensic toolkit. *Arch Med Sąd Kryminol*, *65*(4), 260–271. https://doi.org/10.5114/amsik.2015.61029

Ambers, A., Bus, M. M., King, J. L., Jones, B., Durst, J., Bruseth, J. E., Gill-King, H., & Budowle, B. (2020). Forensic genetic investigation of human skeletal remains recovered from the La Belle shipwreck. *Forensic Science International*, *306*. https://doi.org/10.1016/J.FORSCIINT.2019.110050

Ambers, A. D., Churchill, J. D., King, J. L., Stoljarova, M., Gill-King, H., Assidi, M., Abu-Elmagd, M., Buhmeida, A., & Budowle, B. (2016). *More comprehensive forensic genetic marker analyses for accurate human remains identification using massively parallel DNA sequencing*. https://doi.org/10.1186/s12864-016-3087-2

Amorim, A., & Pereira, L. (2005). Pros and cons in the use of SNPs in forensic kinship investigation: a comparative analysis with STRs. *Forensic Science International*, *150*(1), 17–21. https://doi.org/10.1016/J.FORSCIINT.2004.06.018

Bachtrog, D., & Charlesworth, B. (2001). *Towards a complete sequence of the human Y chromosome*.

Bar, W., Brinkmann, B., Budowle, B., Carracedo, A., Gill, P., Lincoln, P., Mayr, W., & Olaisen, B. (1997). DNA recommendations. Further report of the DNA commission of the ISFH regarding the use of short tandem repeat systems. International Journal of Legal Medicine, 110(4), 175–176. https://doi.org/10.1007/S004140050061/METRICS

Becker, D., Rodig, H., Augustin, C., Edelmann, J., Götz, F., Hering, S., Szibor, R., & Brabetz, W. (2008). Population genetic evaluation of eight X-chromosomal short tandem repeat loci using Mentype Argus X-8 PCR amplification kit. *Forensic Science International: Genetics*, *2*(1), 69–74. https://doi.org/10.1016/j.fsigen.2007.08.013

Bini, C., Sarno, S., Tangorra, E., Iuvaro, A., De Fanti, S., Yohannes, &, Tseghereda, G., Pelotti, S., Luiselli, D., & Tseghereda, Y. G. (2020). Haplotype data and forensic evaluation of 23 Y-STR and 12 X-STR loci in eight ethnic groups from Eritrea. *International Journal of Legal Medicine*, *135*, 449–453. https://doi.org/10.1007/s00414-020-02446-2

Bottinelli, M., Gouy, A., Utz, S., & Zieger, M. (2022). Population genetic analysis of 12 X-chromosomal STRs in a Swiss sample. *International Journal of Legal Medicine*, *136*(2), 561–563. https://doi.org/10.1007/S00414-021-02684-Y

Brashear, W. A., Bredemeyer, K. R., & Murphy, W. J. (2021). *Genomic architecture constrained placental mammal X Chromosome evolution*. https://doi.org/10.1101/gr.275274.121

Brockdorff, N., Bowness, J. S., & Wei, G. (2020). Progress toward understanding chromosome silencing by Xist RNA. *Genes & Development*, *34*(11–12), 733. https://doi.org/10.1101/GAD.337196.120

Bruijns, B., Tiggelaar, R., & Gardeniers, H. (2018). Massively parallel sequencing techniques for forensics: A review. *Electrophoresis*, *39*(21), 2642. https://doi.org/10.1002/ELPS.201800082

Butler, J. M. (2007). Short tandem repeat typing technologies used in human identity testing. *BioTechniques*, *43*(4). https://doi.org/10.2144/000112582

Butler, J. M. (2015). The future of forensic DNA analysis. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*(1674). https://doi.org/10.1098/rstb.2014.0252

Cao, M. D., Balasubramanian, S., & Bodén, M. (2015). Sequencing technologies and tools for short tandem repeat variation detection. *Briefings in Bioinformatics*, *16*(2), 193–204. https://doi.org/10.1093/BIB/BBU001

Chen, C., Jin, X., Zhang, X., Zhang, W., Guo, Y., Tao, R., Chen, A., Xu, Q., Li, M., Yang, Y., Zhu, B., Zhabagin, M., & Muthukrishnan, E. (2021). Comprehensive Insights Into Forensic Features and Genetic Background of Chinese Northwest Hui Group Using Six Distinct Categories of 231 Molecular Markers. *Front. Genet*, *12*, 705753. https://doi.org/10.3389/fgene.2021.705753

Chen, L., Guo, Y., Xiao, C., Wu, W., Lan, Q., Fang, Y., Chen, J., & Zhu, B. (2018). Genetic polymorphisms and forensic efficiency of 19 X-chromosomal STR loci for Xinjiang Mongolian population. *PeerJ*, *6*(7). https://doi.org/10.7717/PEERJ.5117

Dash, H. R., Kaitholia, K., Kumawat, R. K., Singh, A. K., Shrivastava, P., Chaubey, G., & Das, S. (2021). Sequence variations, flanking region mutations, and allele frequency at 31 autosomal STRs in the central Indian population by next generation sequencing (NGS). *Scientific Reports*, *11*(1), 23238. https://doi.org/10.1038/S41598-021-02690-5

Dash, H. R., Rawat, N., & Das, S. (2020). Alternatives to amelogenin markers for sex determination in humans and their forensic relevance. *Molecular Biology Reports*, *47*(3), 2347–2360. https://doi.org/10.1007/S11033-020-05268-Y

Diegoli, T. M. (2015). *Forensic typing of short tandem repeat markers on the X and Y chromosomes*. https://doi.org/10.1016/j.fsigen.2015.03.013

Edelmann, J., Deichsel, D., Hering, S., Plate, I., & Szibor, R. (2002). Sequence variation and allele nomenclature for the X-linked STRs DXS9895, DXS8378, DXS7132, DXS6800, DXS7133, GATA172D05, DXS7423 and DXS8377. *Forensic Science International*, *129*(2), 99–103. https://doi.org/10.1016/S0379-0738(02)00230-X

Edelmann, J., Hering, S., Augustin, C., & Szibor, R. (n.d.). *Characterisation of the STR markers DXS10146, DXS10134 and DXS10147 located within a 79.1 kb region at Xq28*. https://doi.org/10.1016/j.fsigen.2007.08.001

Edelmann, J., Hering, S., Augustin, C., & Szibor, R. (2008). Characterisation of the STR markers DXS10146, DXS10134 and DXS10147 located within a 79.1 kb region at Xq28. *Forensic Science International: Genetics*, *2*(1), 41–46. https://doi.org/10.1016/j.fsigen.2007.08.001

Faustino, M. (2021). Polymorphisms and mutation rate estimates of Y-STRs and X-STRs in the Portuguese population [MSc thesis]. Faculdade de Ciências da Universidade do Porto (FCUP).

Feuk, L., MacDonald, J. R., Tang, T., Carson, A. R., Li, M., Rao, G., Khaja, R., & Scherer, S. W. (2005). Discovery of Human Inversion Polymorphisms by Comparative Analysis of Human and Chimpanzee DNA Sequence Assemblies. *PLoS Genetics*, *1*(4), 0489–0498. https://doi.org/10.1371/JOURNAL.PGEN.0010056

Furman, B. L. S., Metzger, D. C. H., Darolti, I., Wright, A. E., Sandkam, B. A., Almeida, P., Shu, J. J., & Mank, J. E. (n.d.). *Sex Chromosome Evolution: So Many Exceptions to the Rules*. https://doi.org/10.1093/gbe/evaa081

Gomes, I., Brehm, A., Gusmão, L., & Schneider, P. M. (2016). New sequence variants detected at DXS10148, DXS10074 and DXS10134 loci. *Forensic Science International: Genetics*, *20*, 112–116. https://doi.org/10.1016/j.fsigen.2015.10.005

Gomes, I., Pereira, P. J. P., Harms, S., Oliveira, A. M., Schneider, P. M., & Brehm, A. (2017). Genetic characterization of Guinea-Bissau using a 12 X-chromosomal STR system: Inferences from a multiethnic population. *Forensic Science International: Genetics*, *31*, 89–94. https://doi.org/10.1016/j.fsigen.2017.08.016

Gomes, I., Pereira, R., Mayr, W. R., Amorim, A., Carracedo, A., & Gusmão, L. (2009). Evaluation of DXS9902, DXS7132, DXS6809, DXS7133, and DXS7423 in humans and chimpanzees: sequence variation, repeat structure, and nomenclature. *International Journal of Legal Medicine 2009 123:5*, *123*(5), 403–412. https://doi.org/10.1007/S00414-009-0357-4

Gomes, I., Pinto, N., Antão-Sousa, S., Gomes, V., Gusmão, L., & Amorim, A. (2020). Twenty Years Later: A Comprehensive Review of the X Chromosome Use in Forensic Genetics. *Frontiers in Genetics*, *11*. https://doi.org/10.3389/FGENE.2020.00926

Gomes, I., Prinz, M., Pereira, R., Bieschke, E., Mayr, W. R., Amorim, A., Carracedo, A., & Gusmão, L. (2008). X-chromosome STR sequence variation, repeat structure, and nomenclature in humans and chimpanzees. *International Journal of Legal Medicine 2008 123:2*, *123*(2), 143–149. https://doi.org/10.1007/S00414-008-0303-X

Gomes, I; Brehm, A.; Gusmão, L. ;Schneider, P.M. (2016). New sequence variants detected at DXS10148, DXS10074 and DXS10134 loci. *Forensic Science International: Genetics*, *20*, 112–116. https://doi.org/10.1016/J.FSIGEN.2015.10.005

Green, M. R., & Sambrook, J. (2018). Touchdown polymerase chain reaction (PCR). *Cold Spring Harbor Protocols*, *2018*(5), 350–353. https://doi.org/10.1101/PDB.PROT095133

Gusmão, L., Alves, C., Costa, S., Amorim, A., Brion, M., González-Neira, A., Sánchez-Diz, P., & Carracedo, A. (2002). Point mutations in the flanking regions of the Y-chromosome specific STRs DYS391, DYS437 and DYS438. *International Journal of Legal Medicine 2002 116:6*, *116*(6), 322–326. https://doi.org/10.1007/S00414-002-0310-2

Gusmão, L., Butler, J. M., Carracedo, A., Gill, P., Kayser, M., Mayr, W. R., Morling, N., Prinz, M., Roewer, L., Tyler-Smith, C., & Schneider, P. M. (2006). DNA Commission of the International Society of Forensic Genetics (ISFG): An update of the recommendations on the use of Y-STRs in forensic analysis. International Journal of Legal Medicine, 120(4), 191–200. https://doi.org/10.1007/S00414-005-0026-1/METRICS

Gusmão, L., González-Neira, A., Alves, C., Lareu, M., Costa, S., Amorim, A., & Carracedo, A. (2002). Chimpanzee homologous of human Y specific STRs: A comparative study and a proposal for nomenclature. *Forensic Science International*, *126*(2), 129–136. https://doi.org/10.1016/S0379-0738(02)00046-4

Gusmão, L., González-Neira, A., Alves, C., Sánchez-Diz, P., Dauber, E. M., Amorim, A., & Carracedo, A. (2002). Genetic diversity of Y-specific STRs in chimpanzees (Pan troglodytes). *American Journal of Primatology*, *57*(1), 21–29. https://doi.org/10.1002/AJP.1084

Haddrill, P. R. (2021). Developments in forensic DNA analysis. *Emerging Topics in Life Sciences*, *5*(3), 381. https://doi.org/10.1042/ETLS20200304

Heard, E., & Turner, J. (2011). Function of the sex chromosomes in mammalian fertility. *Cold Spring Harbor Perspectives in Biology*, *3*(10), 1–17. https://doi.org/10.1101/cshperspect.a002675

Hering, S., Augustin, C., Edelmann, J., Heidel, M., Dressler, J., Rodig, H., Kuhlisch, E., & Szibor, R. (2005). DXS10079, DXS10074 and DXS10075 are STRs located within a 280-kb region of Xq12 and provide stable haplotypes useful for complex kinship cases. *International Journal of Legal Medicine 2005 120:6*, *120*(6), 337–345. https://doi.org/10.1007/S00414-005-0061-Y

Hundertmark, T., Hering, S., Edelmann, J., Augustin, C., Plate, I., & Szibor, R. (2008). The STR cluster DXS10148–DXS8378–DXS10135 provides a powerful tool for X-chromosomal haplotyping at Xp22. *International Journal of Legal Medicine 2008 122:6*, *122*(6), 489–492. https://doi.org/10.1007/S00414-008-0277-8

Iwase, M., Satta, Y., Hirai, H., Hirai, Y., & Takahata, N. (2010). Frequent gene conversion events between the X and Y homologous chromosomal regions in primates. *BMC Evolutionary Biology 2010 10:1*, *10*(1), 1–11. https://doi.org/10.1186/1471-2148-10-225

Jeffreys, A. J., Wilson, V., & Thein, S. L. (1985). Hypervariable 'minisatellite' regions in human DNA. *Nature 1985 314:6006*, *314*(6006), 67–73. https://doi.org/10.1038/314067a0

Kehrer-Sawatzki, H., & Cooper, D. N. (2006). Structural divergence between the human and chimpanzee genomes. *Human Genetics 2006 120:6*, *120*(6), 759–778. https://doi.org/10.1007/S00439-006-0270-6

Lahn, B. T., & Page, D. C. (1999a). Four evolutionary strata on the human X chromosome. *Science (New York, N.Y.)*, *286*(5441), 964–967. https://doi.org/10.1126/SCIENCE.286.5441.964

Lahn, B. T., & Page, D. C. (1999b). Four evolutionary strata on the human X chromosome. *Science (New York, N.Y.)*, *286*(5441), 964–967. https://doi.org/10.1126/SCIENCE.286.5441.964

Lan, Q., Zhao, C., Chen, C., Xu, H., Fang, Y., Yao, H., & Zhu, B. (n.d.). *Forensic Feature Exploration and Comprehensive Genetic Insights Into Yugu Ethnic Minority and Northern Han Population via a Novel NGS-Based Marker Set*. https://doi.org/10.3389/fgene.2022.816737

Li, C. (2018). Forensic genetics. *Forensic Sciences Research*, *3*(2), 103–104. https://doi.org/10.1080/20961790.2018.1489445

Lyon, M. F. (1961). Gene Action in the X-chromosome of the Mouse (Mus musculus L.). *Nature 1961 190:4773*, *190*(4773), 372–373. https://doi.org/10.1038/190372a0

Marano, L. A., & Fridman, C. (2019). DNA phenotyping: current application in forensic science . *Research and Reports in Forensic Medical Science*, *Volume 9*, 1–8. https://doi.org/10.2147/RRFMS.S164090

Mikkelsen, T. S., Hillier, L. W., Eichler, E. E., Zody, M. C., Jaffe, D. B., Yang, S. P., Enard, W., Hellmann, I., Lindblad-Toh, K., Altheide, T. K., Archidiacono, N., Bork, P., Butler, J., Chang, J. L., Cheng, Z., Chinwalla, A. T., Dejong, P., Delehaunty, K. D., Fronick, C. C., … Waterston, R. H. (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature 2005 437:7055*, *437*(7055), 69–87. https://doi.org/10.1038/nature04072

MN, S., & JMA, T. (2018). Mammalian X Chromosome Dosage Compensation: Perspectives From the Germ Line. *BioEssays : News and Reviews in Molecular, Cellular and Developmental Biology*, *40*(6). https://doi.org/10.1002/BIES.201800024

Newman, T. L., Tuzun, E., Morrison, V. A., Hayden, K. E., Ventura, M., Mcgrath, S. D., Rocchi, M., & Eichler, E. E. (n.d.). *A genome-wide survey of structural variation between human and chimpanzee*. https://doi.org/10.1101/gr.4338005

Oberacher, H., & Parson, W. (2007). Forensic DNA fingerprinting by liquid chromatography-electrospray ionization mass spectrometry. *BioTechniques*, *43*(4). https://doi.org/10.2144/000112581/ASSET/IMAGES/LARGE/FIGURE2.JPEG

Olaisen, B., Bär, W., Brinkmann, B., Budowle, B., Carracedo, A., Gill, P., Lincoln, P., Mayr, W. R., & Rand, S. (1997). *DNA recommendations 1997*.

Pandey, R. S., Wilson Sayres, M. A., & Azad, R. K. (2013). Detecting evolutionary strata on the human X chromosome in the absence of gametologous Y-linked sequences. *Genome Biology and Evolution*, *5*(10), 1863–1871. https://doi.org/10.1093/gbe/evt139

Perera, N., Galhena, G., & Ranawaka, G. (2021). X-chromosomal STR based genetic polymorphisms and demographic history of Sri Lankan ethnicities and their relationship with global populations. *Scientific Reports*, *11*(1), 12748. https://doi.org/10.1038/S41598-021-92314-9

Roewer, L. (2013). *DNA fingerprinting in forensics: past, present, future*. https://doi.org/10.1186/2041-2223-4-22

Ross et al. (2005). *The DNA sequence of the human X chromosome*. http://vega.sanger.ac.uk/

Samejima, M., Nakamura, Y., Nambiar, P., & Minaguchi, K. (2012). Genetic study of 12 X-STRs in Malay population living in and around Kuala Lumpur using Investigator Argus X-12 kit. *International Journal of Legal Medicine 2012 126:4*, *126*(4), 677–683. https://doi.org/10.1007/S00414-012-0705-7

Schaffner, S. F. (2004). The X chromosome in population genetics. *Nature Reviews Genetics 2004 5:1*, *5*(1), 43–51. https://doi.org/10.1038/nrg1247

Sim, J. E., Hwan, A. E., Lee, Y., Woo, A. E., Yang, I., Kyoung, A. E., & Shin, J. (n.d.). *Population genetic study of four closely-linked X-STR trios in Koreans*. https://doi.org/10.1007/s11033-009-9733-x

Simayijiang, H., Morling, N., & Børsting, C. (2022). Sequencing of human identification markers in an Uyghur population using the MiSeq FGxTM Forensic Genomics System. *Forensic Sciences Research*, *7*(2), 154. https://doi.org/10.1080/20961790.2020.1779967

Singh, A., Thakur, M., Sahajpal, V., Singh, S. K., Chandra, K., Sharma, A., Devi, N., & Bernot, A. (2019). Cross-species validation of human specific STR system, SureID ® 21G and SureID ® 23comp (Health Gene Technologies) in Chimpanzee (Pan Troglodytes). *BMC Research Notes*, *12*(1), 1–5. https://doi.org/10.1186/S13104-019-4801-3/TABLES/3

Suntsova, M. V., & Buzdin, A. A. (2020). Differences between human and chimpanzee genomes and their implications in gene expression, protein functions and biochemical properties of the two species. *BMC Genomics*, *21*(7), 1–12. https://doi.org/10.1186/S12864-020-06962-8/TABLES/1

Szibor, R. (2007). X-chromosomal markers: past, present and future. *Forensic Science International. Genetics*, *1*(2), 93–99. https://doi.org/10.1016/J.FSIGEN.2007.03.003

Tao, R., Zhang, J., Bian, Y., Dong, R., Liu, X., Jin, C., Zhu, R., Zhang, S., & Li, C. (2018). Investigation of 12 X-STR loci in Mongolian and Eastern Han populations of China with comparison to other populations. *Scientific Reports*, *8*(1), 4287. https://doi.org/10.1038/S41598-018-22665-3

Thakur, M., Chandra, K., Sahajpal, V., Samanta, A., Sharma, A., & Mitra, A. (2018). Functional validation of human-specific PowerPlex® 21 System (Promega, USA) in chimpanzee (Pan troglodytes). *BMC Research Notes*, *11*(1), 1–4. https://doi.org/10.1186/S13104-018-3803-X/TABLES/2

Tillmar, A. O., Kling, D., Butler, J. M., Parson, W., Prinz, M., Schneider, P. M., Egeland, T., & Gusmão, L. (2017). DNA Commission of the International Society for Forensic Genetics (ISFG): Guidelines on the use of X-STRs in kinship analysis. *Forensic Science International. Genetics*, *29*, 269–275. https://doi.org/10.1016/J.FSIGEN.2017.05.005

Udogadi, N. S., Abdullahi, M. K., Bukola, A. T., Imose, O. P., & Esewi, A. D. (2020). Forensic DNA Profiling: Autosomal Short Tandem Repeat as a Prominent Marker in Crime Investigation. *The Malaysian Journal of Medical Sciences : MJMS*, *27*(4), 22. https://doi.org/10.21315/MJMS2020.27.4.3

Urquhart, A., Kimpton, C. P., Downes, T. J., & Gill, P. (1994). *Variation in Short Tandem Repeat sequences-a survey of twelve microsatellite loci for use as forensic identification markers*.

Varki, A., & Altheide, T. K. (2005). Comparing the human and chimpanzee genomes: searching for needles in a haystack. *Genome Research*, *15*(12), 1746–1758. https://doi.org/10.1101/GR.3737405

Veselinović, I., Petrić, G., Vapa, D., Djan, M., Veličković, N., & Veljović, T. (2017). Genetic analysis of 12 X-STR loci in the Serbian population from Vojvodina Province. *International Journal of Legal Medicine 2017 132:2*, *132*(2), 405–408. https://doi.org/10.1007/S00414-017-1677-4

Walsh, P. S., Metzger, D. A., & Higuchi, R. (1991). Chelex® 100 as a medium for simple extraction of DNA for PCR-based typing from forensic material. *BioTechniques*, *10*(4), 506–513. https://doi.org/10.2144/000114018

Ziętkiewicz, E., Witt, M., Daca, P., Żebracka-Gala, J., Goniewicz, M., Jarząb, B., & Witt, M. (n.d.). *Current genetic methodologies in the identification of disaster victims and in forensic analysis*. https://doi.org/10.1007/s13353-011-0068-7

https://answersingenesis.org/genetics/dna-similarities/greater-than-98-chimp-human-dna-similarity-not-any-more/

https://microbewiki.kenyon.edu/index.php/Chimpanzee_Evolution

https://si.biostat.washington.edu/sites/default/files/modules/2.1%20STR%20Introduction.pdf

Attachments

Attachment 1

Supplementary Table 1. Described loci in the literature. Adapted from “Twenty Years Later: A Comprehensive Review of the X Chromosome Use in Forensic Genetics” (Gomes et al., 2020).

Locus	Reference	Aliases/ comments
DXS6807	Lin, L., Li, J., Hu, Y. et al. Genetic characterization of 19 X-STRs in Sierra Leone population from Freetown. Int J Legal Med (2020).	DXS9910; GATA52B03; RH28271; RH63404; RH7712
DXS9895	Deng, C., Song, F., Li, J., Li, Y., Hou, Y., Luo, H. Multiplex PCR for 19 X-chromosomal STRs in Chinese population. Forensic Science International: Genetics Supplement Series December (2017) (6) e24-e26	P35270; GATA124B04
DXS10148	He G, Zou X, Wang M et al. Population genetics, diversity, forensic characteristics of four Chinese populations inferred from X-chromosomal short tandem repeats. Legal Medicine (2020) 43:101677	-
DXS10135	Pinto, N, Pereira, V, Tomas, C et al. Paternal and maternal mutations in X-STRs: A GHEP-ISFG collaborative study. Forensic Science International: Genetics (2020) 46:102258	-
DXS8378	Gomes C, Amorim A, Okolie, VO et al. Genetic insight into Nigerian population groups using an X-chromosome decaplex system. Forensic Science International: Genetics Supplement Series (2019) 7(1)501-503	GATA119E07
DXS9902	Gomes C, Amorim A, Okolie, VO et al. Genetic insight into Nigerian population groups using an X-chromosome decaplex system. Forensic Science International: Genetics Supplement Series (2019) 7(1)501-503	GATA175D03
DXS6795	Yang Z, Chen C, Zhang J et al. Genetic polymorphisms in 16 X-STR loci analyzed in the She population from Zhejiang Province, China Legal Medicine (2019) 39, 25-28	ATA28C05
DXS9907	Hering S, Edelmann J, Augustin C, Szibor R, Immel UD. Chromosome X markers DXS6795, DXS9907 and GATA144D04: Repeat structure and allele distribution in a German population. Forensic Science International: Genetics Supplement Series (2011) 3(1), e321-e322	GATA186D06
DXS6810	Yang Z, Chen C, Zhang J et al. Genetic polymorphisms in 16 X-STR loci analyzed in the She population from Zhejiang Province, China Legal Medicine (2019) 39, 25-28	GATA69C12; RH28272; RH30556; RH63417; RH7726; SHGC-18105
GATA144D04	Deng C, Song F, Li J et al. Multiplex PCR for 19 X-chromosomal STRs in Chinese population. Forensic Science International: Genetics Supplement Series December (2017) (6) e24-e26	-
DXS10076	Fukuta, M., Gaballah, M., Takada, K et al. Genetic polymorphism of 27 X-chromosomal short tandem repeats in an Egyptian population Legal Medicine (2019) 37:64-66	-
DXS10077	Fukuta, M., Gaballah, M., Takada, K et al. Genetic polymorphism of 27 X-chromosomal short tandem repeats in an Egyptian population Legal Medicine (2019) 37:64-66	-
DXS10078	Fukuta, M., Gaballah, M., Takada, K et al. Genetic polymorphism of 27 X-chromosomal short tandem repeats in an Egyptian population Legal Medicine (2019) 37:64-66	-
DXS10161	Deng C, Song F, Li J et al. Multiplex PCR for 19 X-chromosomal STRs in Chinese population. Forensic Science International: Genetics Supplement Series December (2017) (6) e24-e26	-
DXS10160	Deng C, Song F, Li J et al. Multiplex PCR for 19 X-chromosomal STRs in Chinese population. Forensic Science International: Genetics Supplement Series December (2017) (6) e24-e26	-
DXS10159	Yang Z, Chen C, Zhang J et al. Genetic polymorphisms in 16 X-STR loci analyzed in the She population from Zhejiang Province, China Legal Medicine (2019) 39, 25-28	-

DXS10162	Lin L, Li J, Hu Y et al. Genetic characterization of 19 X-STRs in Sierra Leone population from Freetown. Int J Legal Med (2020). doi.org/10.1007/s00414-019-02243-6	-
DXS10163	Fukuta, M., Gaballah, M., Takada, K et al. Genetic polymorphism of 27 X-chromosomal short tandem repeats in an Egyptian population Legal Medicine (2019) 37:64-66	-
DXS10164	Fukuta, M., Gaballah, M., Takada, K et al. Genetic polymorphism of 27 X-chromosomal short tandem repeats in an Egyptian population Legal Medicine (2019) 37:64-66	-
DXS10165	Elakkary S, Hoffmeister-Ullerich S, Schulze C. Genetic polymorphisms of twelve X-STRs of the investigator Argus X-12 kit and additional six X-STR centromere region loci in an Egyptian population sampleForensic Science International: Genetics (2014) 11(1) 26-30	-
DXS7132	Gomes C, Amorim A, Okolie, VO et al. Genetic insight into Nigerian population groups using an X-chromosome decaplex system. Forensic Science International: Genetics Supplement Series (2019) 7(1)501-503	GATA72E05; RH28273; RH30545; RH63248; RH7563; SHGC-18086; sWXD2228
DXS10079	Jiménez- Moreno S, García C, Lobato E, Baeta M, Bañón E. M.de Pancorbo M. Genetic analysis of 12 X-chromosomal STRs an autochthonous population of Southeast Spain. Forensic Science International: Genetics Supplement Series (2019) 7(1) 482-484	-
DXS10074	Jiménez- Moreno S, García C, Lobato E, Baeta M, Bañón E. M.de Pancorbo M. Genetic analysis of 12 X-chromosomal STRs an autochthonous population of Southeast Spain. Forensic Science International: Genetics Supplement Series (2019) 7(1) 482-484	-
DXS10075	Ferragut, J.F., Bassitta, M., Torrens, V et al. Analysis of 21 X-chromosome polymorphisms in urban and rural populations in Salta province (north-western Argentina) International Journal of Legal Medicine (2019) 133(4) 1043-1047	-
DXS981	Fukuta, M., Gaballah, M., Takada, K et al. Genetic polymorphism of 27 X-chromosomal short tandem repeats in an Egyptian population Legal Medicine (2019) 37:64-66	sWXD528
DXS6800	Yang Z, Chen C, Zhang J et al. Genetic polymorphisms in 16 X-STR loci analyzed in the She population from Zhejiang Province, China Legal Medicine (2019) 39, 25-28	GATA31D10; RH28269; RH63257; RH7569
DXS6803	Yang Z, Chen C, Zhang J et al. Genetic polymorphisms in 16 X-STR loci analyzed in the She population from Zhejiang Province, China Legal Medicine (2019) 39, 25-28	GATA45H11; RH63401; RH7706
DXS9898	Zambrano AK, Vaca-Pólit M, Boada, L et al. A X-STR decaplex study in the population of Imbabura-Ecuador. Forensic Science International: Genetics Supplement Series (2019) 7(1)288-290	GATA126G01
DXS6801	He G, Zou X, Wang M et al. Population genetics, diversity, forensic characteristics of four Chinese populations inferred from X-chromosomal short tandem repeats. Legal Medicine (2020) 43:101677	GATA41B11
DXS6809	He G, Zou X, Wang M et al. Population genetics, diversity, forensic characteristics of four Chinese populations inferred from X-chromosomal short tandem repeats. Legal Medicine (2020) 43:101677	GATA69B12; RH30555; RH63263; RH7574; SHGC-18104
DXS6789	He G, Zou X, Wang M et al. Population genetics, diversity, forensic characteristics of four Chinese populations inferred from X-chromosomal short tandem repeats. Legal Medicine (2020) 43:101677	CHLC.32036; GATA31F01; RH28266; RH34061; SHGC-4389

DXS6799	Fukuta, M., Gaballah, M., Takada, K et al. Genetic polymorphism of 27 X-chromosomal short tandem repeats in an Egyptian population Legal Medicine (2019) 37:64-66	GATA29G07; RH28268; RH30550; RH63603; RH7691; SHGC-18097
DXS7424	He G, Zou X, Wang M et al. Population genetics, diversity, forensic characteristics of four Chinese populations inferred from X-chromosomal short tandem repeats. Legal Medicine (2020) 43:101677	-
DXS101	Fukuta, M., Gaballah, M., Takada, K et al. Genetic polymorphism of 27 X-chromosomal short tandem repeats in an Egyptian population Legal Medicine (2019) 37:64-66	sWXD1232;
DXS6797	Deng C, Song F, Li J et al. Multiplex PCR for 19 X-chromosomal STRs in Chinese population. Forensic Science International: Genetics Supplement Series December (2017) (6) e24-e26	GATA10C11; RH28267; RH63601; RH7688
DXS7133	Zambrano AK, Vaca-Pólit M, Boada, L et al. A X-STR decaplex study in the population of Imbabura-Ecuador. Forensic Science International: Genetics Supplement Series (2019) 7(1)288-290	GATA81B07; RH30546; RH63329; RH7636; RH95783; SHGC-18089
DXS6804	Deng C, Song F, Li J et al. Multiplex PCR for 19 X-chromosomal STRs in Chinese population. Forensic Science International: Genetics Supplement Series December (2017) (6) e24-e26	GATA46G10; RH28270; RH30553; RH63403; RH7709; SHGC-18101
GATA172D05	Yang Z, Chen C, Zhang J et al. Genetic polymorphisms in 16 X-STR loci analyzed in the She population from Zhejiang Province, China Legal Medicine (2019) 39, 25-28	-
DXS7130	Fukuta, M., Gaballah, M., Takada, K et al. Genetic polymorphism of 27 X-chromosomal short tandem repeats in an Egyptian population Legal Medicine (2019) 37:64-66	GATA119G08 ; sWXD2629
GATA165B12	Yang Z, Chen C, Zhang J et al. Genetic polymorphisms in 16 X-STR loci analyzed in the She population from Zhejiang Province, China Legal Medicine (2019) 39: 25-28	-
DXS10103	Jiménez- Moreno S, García C, Lobato E, Baeta M, Bañón E. M.de Pancorbo M. Genetic analysis of 12 X-chromosomal STRs an autochthonous population of Southeast Spain. Forensic Science International: Genetics Supplement Series (2019) 7(1) 482-484	-
HPRTB	Chen M, Ren H, Liu Z et al. Genetic polymorphisms and mutation rates of 16 X-STRs in a Han Chinese population of Beijing and application examples in second-degree kinship cases. International Journal of Legal Medicine (2020) 134(1)163-168	-
DXS10101	Jiménez- Moreno S, García C, Lobato E, Baeta M, Bañón E. M.de Pancorbo M. Genetic analysis of 12 X-chromosomal STRs an autochthonous population of Southeast Spain. Forensic Science International: Genetics Supplement Series (2019) 7(1) 482-484	-
GATA31E08	Zambrano AK, Vaca-Pólit M, Boada, L et al. A X-STR decaplex study in the population of Imbabura-Ecuador. Forensic Science International: Genetics Supplement Series (2019) 7(1) 288-290	CHLC.31776; GATA-D3S2390; RH28260
DXS9908	Edelmann J, Lessig R, Willenberg A et al. Forensic validation of the X-chromosomal STR-markers GATA165B12, GATA164A09, DXS9908 and DXS7127 in German population. International Congress Series (2006) 1288: 298-300	GATA182E04

DXS7127	Edelmann J, Lessig R, Willenberg A et al. Forensic validation of the X-chromosomal STR-markers GATA165B12, GATA164A09, DXS9908 and DXS7127 in German population. International Congress Series (2006) 1288: 298-300	GATA100G03 ;DXS9894; GATA-P33734; RH30548; RH63677; RH7770; SHGC-18092
DXS8377	Fukuta, M., Gaballah, M., Takada, K et al. Genetic polymorphism of 27 X-chromosomal short tandem repeats in an Egyptian population Legal Medicine (2019) 37:64-66	-
DXS10146	Jiménez- Moreno S, García C, Lobato E, Baeta M, Bañón E. M.de Pancorbo M. Genetic analysis of 12 X-chromosomal STRs an autochthonous population of Southeast Spain. Forensic Science International: Genetics Supplement Series (2019) 7(1) 482-484	-
DXS10134	Jiménez- Moreno S, García C, Lobato E, Baeta M, Bañón E. M.de Pancorbo M. Genetic analysis of 12 X-chromosomal STRs an autochthonous population of Southeast Spain. Forensic Science International: Genetics Supplement Series (2019) 7(1) 482-484	-
DXS10147	Ferragut, J.F., Bassitta, M., Torrens, V et al. Analysis of 21 X-chromosome polymorphisms in urban and rural populations in Salta province (north-western Argentina) International Journal of Legal Medicine (2019) 133(4) 1043-1047	-
DXS7423	Zambrano AK, Vaca-Pólit M, Boada, L et al. A X-STR decaplex study in the population of Imbabura-Ecuador. Forensic Science International: Genetics Supplement Series (2019) 7(1) 288-290	sWXD2772
DXS6854	Liu Q-L, Li Z-D, Li C-T et al. X chromosomal recombination - A family study analyzing 26 X-STR Loci in Chinese Han three-generation pedigrees Electrophoresis (2013) 34(20-21) 3016-3022	-
DXS6808	X-chromosome genetic variation in São Paulo State (Brazil) population Martins, J.A., Silva, R.H.A., Freschi, A., (...), Oliveira, R.N., Cicarelli, R.M.B. 2010 Annals of Human Biology 37(4), pp. 598-603	GATA68H10
DXS6793	Israr M, Shahid AA, Rahman Z et al. Development and characterization of a new 12-plex ChrX miniSTR system. International Journal of Legal Medicine 128 (2014) (4) 595-598	ATA14A02;
LC149476	Nishi T, Fukui K, Iwadate K. Genetic polymorphism analyses of three novel X chromosomal short tandem repeat loci in the Xp22.3 region. Legal Medicine (2020) 45:101709	-
LC149480	Nishi T, Fukui K, Iwadate K. Genetic polymorphism analyses of three novel X chromosomal short tandem repeat loci in the Xp22.3 region. Legal Medicine (2020) 45:101709	-
LC149479	Nishi T, Fukui K, Iwadate K. Genetic polymorphism analyses of three novel X chromosomal short tandem repeat loci in the Xp22.3 region. Legal Medicine (2020) 45:101709	-
LC149484	Nishi T, Fukui K, Iwadate K. Genetic polymorphism analyses of three novel X chromosomal short tandem repeat loci in the Xp22.3 region. Legal Medicine (2020) 45:101709	-
LC317283	Nishi T, Fukui K, Iwadate K. Genetic polymorphism analyses of three novel X chromosomal short tandem repeat loci in the Xp22.3 region. Legal Medicine (2020) 45:101709	-
LC317284	Nishi T, Fukui K, Iwadate K. Genetic polymorphism analyses of three novel X chromosomal short tandem repeat loci in the Xp22.3 region. Legal Medicine (2020) 45:101709	-
LC317285	Nishi T, Fukui K, Iwadate K. Genetic polymorphism analyses of three novel X chromosomal short tandem repeat loci in the Xp22.3 region. Legal Medicine (2020) 45:101709	-

DXS7129	Dong C, Fu L, Zhang X et al. Development of three X-linked tetrameric microsatellite markers for forensic purposes. Molecular Biology Reports (2014) 41(10), pp. 6429-6432	GATA117G02 ; GATA-P34273; sWXD2877
DXS2500	Dong C, Fu L, Zhang X et al. Development of three X-linked tetrameric microsatellite markers for forensic purposes. Molecular Biology Reports (2014) 41(10), pp. 6429-6432	HUMUT1595; sWXD1688
G10583	Dong C, Fu L, Zhang X et al. Development of three X-linked tetrameric microsatellite markers for forensic purposes. Molecular Biology Reports (2014) 41(10), pp. 6429-6432	GATA151A05
DXS10102	Samejima M, Nakamura Y, Minaguchi K. Population genetic study of six closely linked groups of X-STRs in a Japanese population. Int J Legal Med. 2011 Nov;125(6):895-900.	-
DXS10106	Samejima M, Nakamura Y, Minaguchi K. Population genetic study of six closely linked groups of X-STRs in a Japanese population. Int J Legal Med. 2011 Nov;125(6):895-900.	-
DXS680705	Lee J C-I, Lin C-Y, Tsai L-C et al. Establishment of 11 linked X-STR loci within 1.1 Mb to assist with kinship testing. International Journal of Legal Medicine (2017) 132(4): 967–973.	-
DXS680707	Lee J C-I, Lin C-Y, Tsai L-C et al. Establishment of 11 linked X-STR loci within 1.1 Mb to assist with kinship testing. International Journal of Legal Medicine (2017) 132(4): 967–973.	-
DXS680708	Lee J C-I, Lin C-Y, Tsai L-C et al. Establishment of 11 linked X-STR loci within 1.1 Mb to assist with kinship testing. International Journal of Legal Medicine (2017) 132(4): 967–973.	-
DXS680709	Lee J C-I, Lin C-Y, Tsai L-C et al. Establishment of 11 linked X-STR loci within 1.1 Mb to assist with kinship testing. International Journal of Legal Medicine (2017) 132(4): 967–973.	-
DXS680704	Lee J C-I, Lin C-Y, Tsai L-C et al. Establishment of 11 linked X-STR loci within 1.1 Mb to assist with kinship testing. International Journal of Legal Medicine (2017) 132(4): 967–973.	-
DXS680710	Lee J C-I, Lin C-Y, Tsai L-C et al. Establishment of 11 linked X-STR loci within 1.1 Mb to assist with kinship testing. International Journal of Legal Medicine (2017) 132(4): 967–973.	-
DXS680701	Lee J C-I, Lin C-Y, Tsai L-C et al. Establishment of 11 linked X-STR loci within 1.1 Mb to assist with kinship testing. International Journal of Legal Medicine (2017) 132(4): 967–973.	-
DXS680703	Lee J C-I, Lin C-Y, Tsai L-C et al. Establishment of 11 linked X-STR loci within 1.1 Mb to assist with kinship testing. International Journal of Legal Medicine (2017) 132(4): 967–973.	-
DXS680706	Lee J C-I, Lin C-Y, Tsai L-C et al. Establishment of 11 linked X-STR loci within 1.1 Mb to assist with kinship testing. International Journal of Legal Medicine (2017) 132(4): 967–973.	-
DXS680702	Lee J C-I, Lin C-Y, Tsai L-C et al. Establishment of 11 linked X-STR loci within 1.1 Mb to assist with kinship testing. International Journal of Legal Medicine (2017) 132(4): 967–973.	-
DXS-p11.3	Zhu YS, Wu H, Lai JH. Structure and polymorphism of novel X-chromosome short tandem repeat loci in a Chinese Han population. Genet Mol Res. 2015;14(4):15044-15049	-
DXS-q12	Zhu YS, Wu H, Lai JH. Structure and polymorphism of novel X-chromosome short tandem repeat loci in a Chinese Han population. Genet Mol Res. 2015;14(4):15044-15049	-

DXS-q13.3	Zhu YS, Wu H, Lai JH. Structure and polymorphism of novel X-chromosome short tandem repeat loci in a Chinese Han population. Genet Mol Res. 2015;14(4):15044-15049	-
DXS-q22.1	Zhu YS, Wu H, Lai JH. Structure and polymorphism of novel X-chromosome short tandem repeat loci in a Chinese Han population. Genet Mol Res. 2015;14(4):15044-15049	-
DXS-q25	Zhu YS, Wu H, Lai JH. Structure and polymorphism of novel X-chromosome short tandem repeat loci in a Chinese Han population. Genet Mol Res. 2015;14(4):15044-15049	-
DXS10011	Yang X, Zhang X, Zhu J et al. Genetic analysis of 19 X chromosome STR loci for forensic purposes in four Chinese ethnic groups. Sci Rep. (2017) 7:42782.	DYS384; HUMUT413
DXS337	Valle Y, Padilla-Gutiérrez JR, Rodarte K et al. Five X-chromosome short tandem repeats in a Western Mexican population. Clin Chem Lab Med. (2008) 46(10):1388-1390.	sWXD1239
DXS1214	Yu B, Zhang H, Li S. X-chromosome STRs polymorphisms of Han ethnic group from Northwest China. Forensic Sci Int. (2005) 153(2-3):269-271.	-
GATA164A09	Edelmann J, Lessig R, Willenberg A et al. Forensic validation of the X-chromosomal STR-markers GATA165B12, GATA164A09, DXS9908 and DXS7127 in German population. International Congress Series (2006) 1288 298-300	-

Attachment 2

Supplementary Table 2. Genotypes obtained for the X-STRs markers on the Portuguese samples studied which was used for comparison. (Obtained in house)

Original code	DXS10134	DXS10074	DXS10135	DXS10146	DXS10148
N7	37	15	27	28	24,1
N5	37	8	19	28	18
N13	42,3	15	19,1	31	26,1
N16	35	16	24	28	24,1
N19	39,3	17	24	39,2*	19
N35	34	17	24	27	29,1
N40	35	14	22	27	24,1
N44	36	17	21	29	25,1
N53	36	16	24	25	18
N57	33	13	22	30	27,1
N65	37	16	33	27	23,1
N71	35	18	34	30	22,1
N75	39,3	8	21	26	27,1
N84	34	8	19,1	28	24,1
N88	35,2	18	19,1	40,2	40,1
N91	38	18	22	27	24
N97	36	8	21	25	24
N104	34	8	22	29	27,1
N109	42,1	8	30	26	25,1
N114	38	7	23	27	24
N33	34	18	27	29	25
N122	35	17	24	27	26
N130	39	17	24	31	23
N133	36	8	23	28	27
N137	34	18	24	28	27
N140	42	18	21	40	22
N145	39	16	19	41	25
N46	37	8	24	27	24
N152	37	17	24	27	24
N154	39	13	23	29	22
N147	35	18	21	31	26
N158	35	8	26	31	27
N160	37	17	27	31	29
N170	37	17	19	30	26
N173	35	13	24	27	29

N177	35	8	18	27	18
N181	35	8	22	42	26
N189	36	17	18	30	23
N192	36	16	20	26	26
N195	36	8	23	30	28
N197	34	14	24	26	23
N202	37	7	18	27	26
N208	36	16	19	33	26
N212	36	16	28	30	25
N217	35	15	19	43	23
N221	33	8	20	28	23
N231	43	14	22	30	23
N238	34	16	28	29	26
N268	37	19	26	30	24
N275	37	16	20	29	27
N276	35	16	24	29	24
N281	36	15	29	28	24
N283	36	14	29	40	25
N290	39	10	21	29	18
N296	38	18	19	26	27
N300	39	8	29	45	25
N303	35	15	29	41	18
N305	37	16	17	44	25
N312	35	8	22	29	24
N321	33	15	24	29	25
N323	41	19	19	38	23
N327	37	17	15	28	26
N332	35	17	23	41	23
N337	35	18	23	28	23
N342	36	8	25	26	26
N345	35	16	23	39,2*	25
N348	35	16	21	28	20
N351	34	8	20	29	30
N356	41	16	26	43	25
N359	33	7	28	47	25
N364	37	18	24	42	23
N368	36	8	21	30	23
N374	37	16	29	44	23
N379	37	18	21	29	24
N380	33	14	29	30	13
N384	37	7	27	26	25
N388	34	17	21	30	27

N389	37	8	21	27	26
N394	43	17	28	30	23
N399	37	18	32	29	23
N403	37	18	22	27	24
N411	36	16	27	29	18
N423	36	18	28	28	26
N429	37	15	21	30	27
N181	35	8	22	42	26
N192	36	16	20	26	26
N5	37	8	19	28	18
N5	37	8	19	28	18
N97	36	8	21	25	24
N97	36	8	21	25	24
Z9043	35	18	19	30	18
Z8762	38	18	20	25	25
Z8671	37	17	19	43	26
Z8636	38	16	25	39,2*	18
Z8587	34	18	20	27	30
Z8542	33	18	20	29	24
Z8519	36	8	21	42	18
Z8432	35	15	21	29	28
Z8384	36	17	19	28	26
Z8336	36	16	26	28	24
Z8290	36	16	22	27	29
Z8285	36	18	17	27	23
Z8317	35	19	25	43	17
Z8168	36	15	23	28	18
Z8030	37	8	25	27	18
Z7955	34	8	22	28	27
Z7964	37	8	18	29	27
Z7879	37	19	25	26	22
Z7863	37	18	34	27	24
Z7803	38	8	30	40	25
Z7802	36	17	18	41	27
Z7762	35	18	26	28	25
Z7748	39	17	31	27	23
Z7737	37	8	22	41	25
Z7713	29	15	17	24	30
Z7712	39	19	28	39,2*	27
Z7676	35	15	24	40	19
Z7623	36	18	17	26	27

Z7617	35	13	21	25	26
Z7505	36	7	17	26	24
Z7498	37	7	18	43	25
Z7476	34	16	23	30	27
Z7501	36	15	21	40	25
Z7376	39	17	24	41	23
Z7362	35	17	15	31	27
Z7314	35	17	22	29	24
Z7320	36	8	19	32	23
Z7315	37	19	23	40	25
Z7250	37	8	21	39,2*	18
Z7202	34	20	21	27	24
Z7152	32	8	17	28	26
Z7137	37	19	23	27	20
Z7133	38	17	26	26	23
Z7077	33	7	22	28	22
Z7017	35	8	19	31	23
Z7012	37	19	17	28	28
Z6945	40	17	19	41	27

Attachment 3

Supplementary table 3. Quantification of Macanese samples with NanoDrop One microvolume UV-Vis spectrophotometer (Thermo Scientific Scientific, USA).

Sample Name	(ng/uL)	A260/A280	A260/A230
M2 1	292,442	1,239	0,270
M7 1	61,896	1,424	0,174
M7 2	62,097	1,430	0,173
M8 1	78,097	1,344	0,207
m9 1	287,860	1,170	0,259
M 10 1	49,169	1,233	0,231
M 18 1	78,940	1,219	0,238
M 19 1	34,845	1,216	0,226
M 21 1	68,705	1,189	0,238
M 22 1	62,577	1,222	0,253
M 23 1	233,841	1,206	0,242
M 24 1	151,096	1,269	0,218
M 25 1	112,356	1,258	0,190
M 27 1	154,988	1,187	0,268
M 31 1	80,909	1,201	0,155
M 33 1	75,897	1,204	0,217
M 35 1	115,252	1,220	0,149
M 38 1	32,522	1,144	0,237
M 40 1	149,600	1,209	0,195
M 42 1	71,390	1,153	0,176
M 43 1	53,644	1,011	0,234
M 44 1	56,802	1,146	0,157
M 46 1	27,939	1,390	0,169
M 47 1	105,133	1,107	0,170
M 48 1	441,325	1,103	0,297
M 49 1	368,870	1,133	0,283
M 50 1	182,063	1,248	0,323
M 51 1	472,740	1,107	0,230
M 52 1	226,071	1,057	0,212
M 53 1	193,695	1,106	0,267
M 54 1	95,070	1,153	0,223
M 55 1	152,258	1,093	0,239
M 56 1	205,543	1,062	0,233
M 57 1	193,704	1,165	0,283
M 59 1	134,284	1,175	0,279
M 63 1	118,249	1,117	0,221
M 67 1	135,180	1,106	0,220

M 68 1	80,982	1,206	0,292
M 69 1	129,593	1,217	0,312
M 70 1	202,751	1,089	0,266
M 71 1	170,521	1,152	0,213
M 72 1	175,927	1,138	0,270
M 73 1	194,233	1,235	0,305
M 74 1	220,539	1,123	0,219
M 75 1	444,517	1,143	0,274
M 76 1	80,589	1,152	0,245
M 77 1	114,183	1,151	0,241
M 78 1	118,398	1,147	0,266

Attachment 4

Poster done by analyzing African data obtained from the literature and published data in prowl of the work in this dissertation

Genetic variation of human X-STRs: a comparative study

Moutinho B[1,2], Gusmão L [3], Gomes I [1,4]

[1] i3S, Instituto de Investigação e Inovação em Saúde, University of Porto, Portugal; [2] Department of Biology, Faculty of Science, University of Porto, Portugal; [3] DNA Diagnostic Laboratory (LDD), State University of Rio de Janeiro (UERJ), Brazil; [4] IPATIMUP, Institute of Molecular Pathology and Immunology of the University of Porto, Portugal

INTRODUCTION & AIMS

Forensic genetic tools have been a major field in solving several legal cases. Autosomal short tandem repeats (STRs) are the prime choice for these investigations but they present some limitations that can be complemented by STRs located on the sex chromosomes. Besides their importance in solving complex kinship cases [1], X-chromosome STRs (X-STRs), may also provide valuable information about the genetic evolution of populations [2, 3]. The existence of an accurate database regarding X-STRs, on top of an accepted nomenclature, is of an enormous importance so there is no misinterpretation of data such as when comparing allele frequency distributions among different populations or data exchange between laboratories [4-7]. Thus, to have the most precise STR kits in forensics, studies comparing the genetic markers in different populations must be made to reinsure that these are inclusive. For this aim, we present preliminary results by analyzing sequencing data from African population groups for the DXS10148 locus. We performed comparisons with the chimpanzee (*Pan troglodytes*) genome to infer on the ancestral STR and allelic structure. This will allow us to access the differences in the selected markers and to propose a more accurate nomenclature encompassing the differences encountered.

MATERIAL & METHODS

X-STR: DXS10148; **Samples**: n= 63 Guinea-Bissau [2]; n= 8 West African [4], n= 36 Algeria [3]; **Nomenclature**: most commonly used and initially described **(CCTT)2- N8 -(CTTT)x-(TCTT)y-(TTCC)4** [1]; **Genome assemblies:** *Pan troglodytes* (Jan 2018, Clint_PTRv2/panTro6) *and Homo sapiens* (Dec. 2013, GRCh38/hg38) sequence alignment using the Genome Browser, BLAT tool (http://genome.ucsc.edu).

```
Human ctttccttctttccttctttccttctttccttccttctttccTtctttc
                                                  ||||||||||||||||||
Chimp ctttccttctttccttctttcctt.....ccttctttccTtctttc

Human tttctttctttttctttctttctttctttctttccttccttcctcc
                                                  ||||||||||||||||||
Chimp tttcttt....tctttctttctttctttctttctttccttccttcctcc
```

Figure 1. Sequence alignment results for DXS10148 repeat region between *Homo Sapiens* and *Pan Troglodytes*.

RESULTS & DISCUSSION

Table 1. DXS10148 sequence data obtained in n=107 individuals from African population groups.
ALG: Algeria; AF= West Africa; GB= Guinea Bissau.

Allele Seq	n	Sequence structure of the repeat	Pop
14	1	[CCTT]-[CTTT]-[CTTT]2-N8-[CTTT]6-[TTCC]4	ALG
17	1	[CCTT]2-N8-[CTTT]4-[TCTT]7-[TTCC]4	ALG
18	2	[CCTT]2-N8-[CTTT]4-[TCTT]8-[TTCC]4	ALG
19	1	[CCTT]2-N8-[CTTT]4-[TCTT]9-[TTCC]4	ALG
20.1	1	[CCTT]2-N8-[CTTT]3-A-[TCTT]11-[TTCC]4	ALG
21*	1	[CTTTCCTT]3-CTTT-[CCTT]2-N8-[CTTT]4-[TCTT]11-[TTCC]4	WAF
21	1	[CCTT]2-N8-[CTTT]4-[TCTT]11-[TTCC]4	ALG
22.1	2	[CCTT]2-N8-[CTTT]4-[TCTT]11-[TTCC]4	ALG
23	1	[CCTT]-[CTTT]-[CCTT]2-N8-[CTTT]4-[TCTT]11-[TTCC]4	ALG
23.1	1	[CCTT]2-N8-[CTTT]4-[TCTT]-A-[TCTT]12-[TTCC]4	ALG
24.1	1	[CCTT]2-N8-[CTTT]3-A-[TCTT]15-[TTCC]4	ALG
24.1	1	[CCTT]2-N8-[CTTT]4-[TCTT]-A-[TCTT]13-[TTCC]4	ALG
25.1	2	[CCTT]2-N8-[CTTT]3-A-[TCTT]16-[TTCC]4	ALG
26.1	3	[CCTT]2-N8-[CTTT]3-A-[TCTT]17-[TTCC]4	ALG
26.1	1	[CCTT]2-N8-[CTTT]4-[TCTT]-A-[TCTT]15-[TTCC]4	ALG
27*	1	[CTTTCCTT]3-CTTT-[CCTT]2-N8-[CTTT]4-[TCTT]17-[TTCC]4	WAF
27.1	1	[CCTT]2-N8-[CTTT]3-A-[TCTT]18-[TTCC]4	ALG
28	1	[CCTT]-[CTTT]-[CCTT]2-N8-[CTTT]4-[TCTT]16-[TTCC]4	ALG
28.1	2	[CCTT]2-N8-[CTTT]4-[TCTT]-A-[TCTT]17-[TTCC]4	ALG
28.1	1	[CCTT]2-N8-[CTTT]3-A-[TCTT]19-[TTCC]4	ALG
29	1	[CCTT]-[CTTT]-[CCTT]2-N8-[CTTT]4-[TCTT]17-[TTCC]4	ALG
29.1	2	[CCTT]2-N8-[CTTT]4-[TCTT]-A-[TCTT]18-[TTCC]4	ALG
30*	1	[CTTTCCTT]3-CTTT-[CCTT]2-N8-[CTTT]4-[TCTT]20-[TTCC]4	WAF
31.1	1	[CCTT]2-N8-[CTTT]4-[TCTT]-A-[TCTT]20-[TTCC]4	ALG
31.1	1	[CCTT]4-N8-[CTTT]-[CTCT]3-[TCCT]16-A-[CTCC]3-[TTCC]4	ALG
35.1	2	[CTTTCCTT]2-CTTT-[CCTT]2-N8-[CTTT]4 - [TCTT]13-T-[TCTT]12-[TTCC]-[TTCT]-[TTCC]2	GB
	2	[CTTTCCTT]2-CTTT-[CCT]2-N8-[CTTT]4 - [TCTT]12-T-[TCTT]13-[TTCC]-[TTCT]-[TTCC]2	GB
36.1	1	[CCTT]4-N8-[CTTT]-[CTCT]3-[TCCT]16-A-[CTCC]3-[TTCC]4	ALG
37.1	1	[CTTTCCTT]2-CTTT-[CCT]2-N8-[CTTT]4 - [TCTT]13-T-[TCTT]16-[TTCC]-[TTCT]-[TTCC]2	GB
	1	[CTTTCCTT]2-CTTT-[CCT]2-N8-[CTTT]4 - [TCTT]12-T-[TCTT]15-[TTCC]-[TTCT]-[TTCC]2	GB
	1	[CTTTCCTT]2-CTTT-[CCT]2-N8-[CTTT]4 - [TCTT]13-T-[TCTT]14-[TTCC]-[TTCT]-[TTCC]2	GB
	3	[CTTTCCTT]2-CTTT-[CCT]2-N8-[CTTT]4 - [TCTT]14-T-[TCTT]13-[TTCC]-[TTCT]-[TTCC]2	GB
	1	[CTTTCCTT]2-CTTT-[CCT]2-N8-[CTTT]4 - [TCTT]15-T-[TCTT]12-[TTCC]-[TTCT]-[TTCC]2	GB
37.1	1	[CTTTCCTT]3-CTTT-[CCT]2-N8-[CTTT]4-A-[CTTT]12-[TCTT]15-TTCCTTCT-[TTCC]2	WAF
37.1	1	[CCTT]2-N8-[CTTT]4-[TCTT]16-A-[TCT]13-[TTCC]4	ALG
38.1	3	[CTTTCCTT]2-CTTT-[CCT]2-N8-[CTTT]4 - [TCTT]13-T-[TCTT]15-[TTCC]-[TTCT]-[TTCC]2	GB
	2	[CTTTCCTT]2-CTTT-[CCT]2-N8-[CTTT]4 - [TCTT]14-T-[TCTT]14-[TTCC]-[TTCT]-[TTCC]2	GB
	1	[CTTTCCTT]2-CTTT-[CCT]2-N8-[CTTT]4 - [TCTT]15-T-[TCTT]13-[TTCC]-[TTCT]-[TTCC]2	GB
38.1	1	[CTTTCCTT]3-CTTT-[CCT]2-N8-[CTTT]4-A-[CTTT]13-[TCTT]15-TTCCTTCT-[TTCC]2	WAF
38.1	2	[CCTT]2-N8-[CTTT]4-[TCTT]15-A-[TCT]13-[TTCC]4	ALG
38.1	1	[CCTT]2-N8-[CTTT]4-[TCTT]13-[TCCT]3-A-[TCT]12-[TTCC]4	ALG
39.1	1	[CTTTCCTT]2-CTTT-[CCT]2-N8-[CTTT]4 - [TCTT]12-T-[TCTT]17-[TTCC]-[TTCT]-[TTCC]2	GB
	3	[CTTTCCTT]2-CTTT-[CCT]2-N8-[CTTT]4 - [TCTT]13-T-[TCTT]16-[TTCC]-[TTCT]-[TTCC]2	GB
	1	[CTTTCCTT]2-CTTT-[CCT]2-N8-[CTTT]4 - [TCTT]16-T-[TCTT]13-[TTCC]-[TTCT]-[TTCC]2	GB
	2	[CTTTCCTT]2-CTTT-[CCT]2-N8-[CTTT]4 - [TCTT]14-T-[TCTT]15-[TTCC]-[TTCT]-[TTCC]2	GB
	3	[CTTTCCTT]2-CTTT-[CCT]2-N8-[CTTT]4 - [TCTT]15-T-[TCTT]14-[TTCC]-[TTCT]-[TTCC]2	GB
	1	[CTTTCCTT]2-CTTT-[CCT]2-N8-[CTTT]4 - [TCTT]17-T-[TCTT]12-[TTCC]-[TTCT]-[TTCC]2	GB
39.1	2	[CTTTCCTT]2-CTTT-[CCT]2-N8-[CTTT]4-A-[CTTT]14-[TCTT]15-TTCCTTCT-[TTCC]2	WAF
40.1	1	[CTTTCCTT]2-CTTT-[CCT]2-N8-[CTTT]4 - [TCTT]11-T-[TCTT]19-[TTCC]-[TTCT]-[TTCC]2	GB
	1	[CTTTCCTT]2-CTTT-[CCT]2-N8-[CTTT]4 - [TCTT]12-T-[TCTT]18-[TTCC]-[TTCT]-[TTCC]2	GB
	9	[CTTTCCTT]2-CTTT-[CCT]2-N8-[CTTT]4 - [TCTT]14-T-[TCTT]16-[TTCC]-[TTCT]-[TTCC]2	GB
	4	[CTTTCCTT]2-CTTT-[CCT]2-N8-[CTTT]4 -[TCTT]15-T-[TCTT]15-[TTCC]-[TTCT]-[TTCC]2	GB
	2	[CTTTCCTT]2-CTTT-[CCT]2-N8-[CTTT]4 - [TCTT]10-[TCTA]-[TCTT]4-T-[TCTT]15-[TTCC]-[TTCT]-[TTCC]2	GB
	1	[CTTTCCTT]2-CTTT-[CCT]2-N8-[CTTT]4 - [TCTT]16-T-[TCTT]14-[TTCC]-[TTCT]-[TTCC]2	GB
41.1	2	[CTTTCCTT]2-CTTT-[CCT]2-N8-[CTTT]4 - [TCTT]14-T-[TCTT]17-[TTCC]-[TTCT]-[TTCC]2	GB
	1	[CTTTCCTT]2-CTTT-[CCT]2-N8-[CTTT]4 - [TCTT]9-[TCTA]-[TCTT]4-T-[TCTT]17-[TTCC]-[TTCT]-[TTCC]2	GB
	5	[CTTTCCTT]2-CTTT-[CCT]2-N8-[CTTT]4 - [TCTT]15-T-[TCTT]16-[TTCC]-[TTCT]-[TTCC]2	GB
	1	[CTTTCCTT]3 -CTTT- [CCTT-CTTT]-N8-[CTTT]4 - [TCTT]15-T-[TCTT]16-[TTCC]-[TTCT]-[TTCC]2	GB
	1	[CTTTCCTT]2-CTTT-[CCT]2-N8-[CTTT]4 - [TCTT]16-T-[TCTT]15-[TTCC]-[TTCT]-[TTCC]2	GB
41.1	1	[CTTTCCTT]-CTTT-[CCTT]2-N8-[CTTT]4-A-[CTTT]15-[TCTT]15-TTCCTTCCT-[TTCC]2	WAF
41.1	2	[CCTT]2-N8-[CTTT]4-[TCTT]3-[TCTA]-[TCTT]9-A-[TCTT]17-[TTCC]4	ALG
42.1	3	[CTTTCCTT]2-CTTT-[CCT]2-N8-[CTTT]4 - [TCTT]16-T-[TCTT]16-[TTCC]-[TTCT]-[TTCC]2	GB
	2	[CTTTCCTT]2 -CTTT- [CCTT]2-N8-[CTTT]4 - [TCTT]15-T-[TCTT]17-[TTCC]-[TTCT]-[TTCC]2	GB
43.1	1	[CTTTCCTT]2-CTTT-[CCT]2-N8-[CTTT]4 - [TCTT]16-T-[TCTT]17-[TTCC]-[TTCT]-[TTCC]2	GB
44.1	1	[CTTTCCTT]2-CTTT-[CCT]2-N8-[CTTT]4 - [TCTT]16-T-[TCTT]18-[TTCC]-[TTCT]-[TTCC]2	GB
45.1	1	[CTTTCCTT]2-CTTT-[CCT]2-N8-[CTTT]4 - [TCTT]17-T-[TCTT]18-[TTCC]-[TTCT]-[TTCC]2	GB

- Qualitative information regarding the allele and locus sequence structure of the X-STR DXS10148 in African populations was gathered from the literature (Table 1).
- Analyses of the data (Table 1) reveals a lot of variation regarding the allelic structure of the DXS10148 genetic marker, in particular, high intra-variation for the non-consensus alleles were found (i.e., same alleles with different structures).
- Sequence data published in two different studies [2, 4] detected variation in the upstream flanking region of the DXS10148 locus (reverse strand) in a population sample of Guinea-Bissau and other West African regions (Table 1) as following **(CTTT-CCTT)1-3 -*CTTT*-(CCTT)2-*CTTTCCTT*-(CTTT)x-(TCTT)y-T*-(TTCC)4.**
- In the nomenclature suggested in the study of Hundertmark and co-authors [1] only two copies of this structures is observed **(CTTT-CCTT)2-*CTTT*-(CCTT)2-*CTTTCCTT*-(CTTT)x-T*-(TCTT)y-(TTCC)4.**
- In the study for the Algerian group [3] results for the repeat region are presented only, therefore it is not clear how many copies of the upstream flanking sequence **CTTT-CCTT** were detected (Table 1).
- Results of sequence alignments for the human DXS10148 repeat region and the chimpanzee genome (Figure 1) also revealed 3 copies of CTTT-CCTT with the following allele structure: **(CTTT-CCTT)3- (CCTT)1-*CTTTCCTT*-(CTTT)3-TCTT-T- (TCTT)4-(TTCC)4.**

(*T= insertion in non-consensus alleles; italic bases not considered for allele nomenclature.)

STR consensus structure for DXS10148 taking into account the locus variation observed in this study between humans and chimpanzees and allele nomenclatures [based on the ISFG guidelines (7-9)]: **(CTTT-CCTT)1-4 *CCTT CTTT CCTT*- (CTTT)x- (TCTT)y- (TCTT)4.**

The preliminary results obtained on the genetic variation of X-STRs suggest that the nomenclature in use for **DXS10148 [1] misses variation at the upstream flanking region of the repeat**. Therefore the new variation should have been included in the (original) nomenclature, in accordance to the ISFG guidelines for STR nomenclature [7-9].

These results reinforce the need to establish a globally accepted nomenclature to avoid misinterpretation among population genetic data and ultimately allow for the most precise X-STR kits in forensic genetics.

References: [1] ... [2] ... [3] ... [4] ... [5] ... [6] ... [7] ... [8] ... [9] ...

FCT I. Gomes is funded by CEECIND/02509/2017